2024

Explore Security and Machine Learning Applications in Next Generation Wireless Networks

Haolin Tang
Virginia Commonwealth University

EXPLORE SECURITY AND MACHINE LEARNING APPLICATIONS IN NEXT GENERATION WIRELESS NETWORKS

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

by

HAOLIN TANG

B.S. in Computer Science and Technology, Yunnan Normal University, China

Advisor: Yanxiao Zhao,

Associate Professor, Department of Electrical and Computer Engineering

Virginia Commonwealth University

Richmond, Virginia

August, 2024

Acknowledgements

I would like to express my sincere gratitude to Dr. Yanxiao Zhao for her patient guidance, insightful suggestions, and continuous encouragement over the past five years. I have been impressed by her expertise and dedication. I also want to thank my dissertation committee members for their time and advice. I would like to extend my appreciation to the Department of Electrical and Computer Engineering faculty and staff members at VCU. They provided me with a professional and friendly educational environment.

I also want to express my appreciation to my undergraduate advisor, Dr. Yang Yang, at Yunnan Normal University, China. He inspired my interest in scientific research and encouraged me to pursue a doctoral degree in the USA.

My heartfelt thanks go to my family for their unconditional love, patience, and encouragement. To my parents, thank you for believing in me and supporting me through every step of my life journey. To my wife, Yuting Yang, thank you for your unwavering support, understanding, and sacrifices. It's great to have you by my side all the time. There are no words to express my thanks and love to you.

It has been a long and challenging journey to complete my Ph.D. program. This achievement would not have been possible without their support.

Thank you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abstract

EXPLORE SECURITY AND MACHINE LEARNING APPLICATIONS IN NEXT GENERATION WIRELESS NETWORKS

By Haolin Tang

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

Virginia Commonwealth University, 2024.

Advisor: Yanxiao Zhao,

Associate Professor, Department of Electrical and Computer Engineering

Next-generation (NextG) or Beyond-Fifth-Generation (B5G) wireless networks have become a prominent focus in academic and industry circles. This is driven by the increasing demand for cutting-edge applications such as mobile health, self-driving cars, the metaverse, digital twins, virtual reality, and more. These diverse applications typically require high communication network performance, including spectrum utilization, data speed, and latency. New technologies are emerging to meet the communication requirements of various applications. Intelligent Reflecting Surface (IRS) and Artificial Intelligence (AI) are two representatives that have been demonstrated as promising and powerful technologies in NextG communications. While new technologies significantly enhance communication performance, they also introduce new security concerns. Therefore, security remains a top priority within the communication community. This dissertation studies innovative solutions for the security in NextG networks. Specifically, we will investigate the security applications of IRS

and machine learning-based solutions to enhance security using wireless signal denoising and signal modulation recognition methods. Also, we will explore defending AI-powered communication systems against adversarial attacks.

IRS can be utilized to flexibly re-configure the fundamental communication environment to realize low-cost, energy-saving, and low-interference wireless communications. However, one concern of the IRS is that malicious users may manipulate it to their advantage, which presents significant security issues. We will propose security scenarios of IRS communication systems and investigate how the IRS affects the Signal-to-Noise Ratio (SNR) with different experimental settings. Furthermore, the SNR plays a critical role in wireless security. SNR is constantly degraded during transmission in a practical communication environment due to interference from malicious attackers or surrounding noise. To address the problems, we will develop a Generative Adversarial Networks (GAN)-based signal denoising method to improve signal quality. Moreover, active malicious physical layer attacks such as spoofing and jamming can disrupt communications and bring unpredictable security risks. Automatic modulation Recognition (AMR), which identifies the modulation types of active attack signals, plays a crucial role in the physical-layer security of wireless communication. A new AMR method based on parallel neural networks will be proposed. In addition, while AI technologies have recently been integrated into NextG networks, the security threats and mitigation methods for AI-powered communication systems in NextG networks have not been thoroughly investigated. Therefore, we will explore the performance of AI-powered communication systems under machine learning adversarial attacks.

The main objectives of this dissertation on AI-driven innovations to security in NextG networks are summarized as follows. We first introduce recent works on using IRS in wireless communications and present basic security scenarios from construc-

tive and adversarial perspectives. Second, a GAN-based wireless signal-denoising method will be developed and compared with traditional algorithms. Third, a deep learning-based AMR method will be proposed and tested on various signal modulation schemes. Last, we investigate the vulnerability of AI-driven NextG communication systems under adversarial attacks and propose the defensive distillation mitigation method to improve its robustness.

CHAPTER 1

INTRODUCTION

In the last decade, we have witnessed the rapid evolution of wireless communication networks, which have been widely applied to our daily lives, including autonomous driving, real-time surveillance, smart cities, and intelligent transportation. Beyond doubt, Fifth-generation (5G) and NextG wireless networks will continue significantly contributing to our crucial societal infrastructure. While we increasingly rely on such critical and essential services provided by the network infrastructure, it also becomes the target of cybercriminals. Numerous cyber attacks are launched every day, and the victims include high-profile companies and government agencies such as the U.S. Office of Personnel Management, Google, and Capital One [1]. To this end, security has become critically important to the network infrastructure and has attracted significant attention recently.

1.1 Motivations

New emerging technologies such as the IRS have opened up new horizons and could be integrated into developing secure intelligent 5G and NextG network systems. It has been demonstrated as a promising solution to provide sustainable, cost-effective, and resilient capabilities for NextG networks. IRS is an enabling technology to engineer radio signal propagation and is capable of dynamically changing the wireless channels to improve the performance of wireless communications. On the other hand, malicious users may also utilize the powerful capability of the IRS to manipulate the communication environment and launch security attacks such as eavesdropping and

jamming. Therefore, while the integration of IRS into wireless communications brings promising and new opportunities, it also raises significant concerns from the security perspective. However, the security perspective of IRS-assisted systems have not been thoroughly examined. The literature indicates that security research in IRS systems is still in its early stages, with only a few studies currently available.

Moreover, wireless signal strength plays a critical role in wireless security. For instance, to prevent eavesdropping for security purposes, we can intentionally reduce the signal power at a transmitter so that eavesdroppers have difficulty receiving the signal due to weak strengths. Also, signals always deteriorate due to surrounding noise and interference during transmission. Because of the above concern, a signal denoising or enhancement method is needed to improve the signal strength. Furthermore, active physical layer threats such as spoofing and jamming can significantly degrade the communication performance. AMR has been recognized as an essential part of identifying attacks. Hence, a new AMR method is desirable to alleviate the physical layer attacks in NextG. Last but not least, AI has achieved tremendous success in many fields, such as computer vision, natural language processing, and so on. This breakthrough also motivated researchers to apply AI to NextG wireless communications. However, limited studies have investigated the security threats and mitigation methods for AI-powered NextG networks for several reasons, e.g., being new and multidisciplinary.

1.2 Contributions

The research goal of this proposal is to secure NextG wireless communication by leveraging new and AI technologies to address the challenges mentioned above. First of all, this study provides a comprehensive literature review of the applications of the IRS in NextG networks and groups them into two categories: 1) securing com-

munication via the IRS and 2) launching attacks using the IRS. Then, we present four typical scenarios of utilizing the IRS for security or threats to wireless communications. Second, we apply AI technology to achieve signal denoising to improve the SNR of communications. Then, we conduct studies on AMR and develop an AI-based wireless signal modulation recognition method. Last, we investigate the vulnerability of AI-powered systems in NextG networks under adversarial machine learning attacks. The defensive distillation mitigation method will also be discussed to defend and improve the robustness of the systems. The above-mentioned research has been published in three conference proceedings and two journals.

Our main contributions to this thesis can be summarized as follows.

1. We study the potential security impact of IRS on wireless communication systems from two categories: improving secure communication and launching security attacks using IRS. Simulations are conducted for four typical scenarios and the impact of IRS on communication performance is examined by measuring SNR to fully understand the capability of IRS and its impact on wireless networks. This work has been published in a conference paper [1].

2. We develop an adversarial learning-based approach for wireless signal denoising, which will correspondingly enhance signal strength. Specifically, we design a conditional Generative Adversarial Network (cGAN) at the receiver to establish an adversarial game between a generator and a discriminator. Unlike traditional signal denoising methods that estimate the noise or interference in the noisy signals, our proposed method estimates and learns the features of real noise-free signals, which is more adaptive to dynamic wireless communication environments. We also conduct simulations on signals with four different modulations to evaluate the performance. This work has been published in a

conference paper [2].

3. We propose a parallel neural network architecture for AMR, which extracts spatial features using CNN (Convolutional Neural Networks) layers and temporal features using LSTM (Long Short-Term Memory) layers, respectively, in two parallel routes. Afterward, the extracted features will be combined to predict the modulation scheme of a signal. Extensive simulations are conducted on signals with 11 different modulation methods at a wide range of SNR levels to evaluate our proposed parallel feature extraction architecture. In addition, we compare our solution against three other deep learning-based methods, and the results verify that our method outperforms other methods regarding recognition accuracy. This work has been published in a conference paper [3].

4. We explore the vulnerabilities of AI-driven NextG communication systems under widely used adversarial attack methods in terms of attack success ratio and identify the potential weakness in the communication system. We also propose a defensive distillation mitigation method to train a more robust model to improve the robustness of the systems. This work has been published in two journal papers [4, 5].

1.3 Dissertation Outline

In Chapter 1, the backgrounds, motivations, and contributions have been introduced. The the rest of this dissertation is organized as follows. Chapter 2 introduces IRS applications in NextG wireless networks and summarizes research on IRS-aided wireless communication systems in security. Then, we propose four typical scenarios of utilizing the IRS for security or threats to wireless communications and evaluate the IRS-assisted system performance in terms of SNR affected by the IRS in those

four scenarios. Chapter 3 proposes a conditional generative adversarial network-based signal denoising approach and compares it to traditional algorithms to demonstrate the superiority of our method. In chapter 4, we develop a deep learning-based automatic modulation recognition method that can identify the modulation scheme of wireless signals. In chapter 5, we investigate the vulnerability of AI-driven NextG communication systems against different adversarial attacks and analyze the robustness of the undefended systems. The mitigation method will be applied to enhance the robustness and reduce the vulnerability of systems. Chapter 6 summarizes all research works in this dissertation and sheds lights on future research directions.

CHAPTER 2

SECURITY AND THREATS IN IRS-AIDED WIRELESS COMMUNICATION SYSTEM

2.1 Introduction

Although we have just entered the 5G wireless communication era, academia, industries, and governments are enthusiastically looking into the future beyond 5G, such as the NextG wireless networks that target meeting more stringent requirements than 5G, e.g., ultra-high data rate, energy efficiency, high reliability, and connectivity. To achieve the above goals, IRS has recently been proposed as a new emerging and promising technology to enhance the performance of 5G and NextG networks. IRS has received substantial attention in the literature due to its constructive or disruptive capability of 'reprogramming' wireless communication environments. An IRS consists of a large number of low-cost passive reflecting elements that can cause a phase change for incident signals. All reflecting elements can jointly adjust their phase shifts to make the reflected signals constructively or destructively added at the receiver [6]. As a result, a wireless communication environment could be dynamically programmed to enhance or degrade the communication performance via IRS.

Researchers have applied IRS-aided wireless communications to enhance communication performance in various areas, including cellular networks, Non-Orthogonal Multiple Access (NOMA) systems, and Unmanned Aerial Vehicle (UAV) [7, 8, 9, 10]. In [11], the authors propose a power-efficient scheme to design the transmit power allocation and phase shift of the reflecting surface to ensure secure communications. In [12], IRS is applied to resource allocation for vehicular communications to maximize

the sum capacity of vehicle-to-infrastructure (V2I) links. The work in [13] studies an IRS-aided single-cell wireless system to minimize the total transmit power at the Access Point (AP) by jointly optimizing the transmit beamforming at the AP and reflected beamforming at the IRS. IRS has also been studied in NOMA systems for various situations [14, 15, 16, 17]. In [16], the throughput and energy efficiency of non-orthogonal users are studied both in delay-limited and delay-tolerant transmission modes. In [14], the downlink communications of IRS-assisted NOMA systems is investigated to maximize the system throughput. A design of IRS-assisted NOMA downlink transmission is proposed in [15]. A symbiotic UVA-assisted IRS radio system is introduced in [18], where the UAV is leveraged to help the IRS reflect its own signals to the base station and meanwhile enhance the UAV transmission.

IRS-assisted networks have been demonstrated as promising solutions to provide sustainable, cost-effective, and resilient capabilities for 5G and beyond networks. Nevertheless, the security perspectives of IRS-assisted systems have not been thoroughly examined. The literature indicates that security research via IRS is still in its infancy. This chapter investigates the full potential in security when the IRS is introduced to 5G and NextG systems to narrow the gap. In particular, we will examine the scenarios for using IRS to secure or threaten communication. One benefit of integrating IRS into communication networks is to increase the quality of communication between legitimate transmitters and receivers through intelligently programming a large number of reflecting elements. Further, IRS can also be used to mitigate the negative effects of malicious attackers (e.g., eavesdroppers) by producing artificial noise. On the other hand, malicious users may leverage IRS to introduce security threats and threaten critical wireless communication systems as well. This chapter will introduce and discuss these above-mentioned security scenarios.

The SNR is a critical performance metric to assess wireless communication sys-

tems, including IRS-assisted communication systems. To fully understand the capability of the IRS and its impact on IRS-assisted wireless networks, in this chapter, we mathematically derive the SNR for IRS-aided wireless communications. Simulations are conducted for four typical security/threat scenarios, and the impact of IRS on communication performance is examined by measuring SNR. Simulation results verify that IRS can considerably increase or decrease SNR in a wireless communication system by controlling the position and phase shifts of the reflecting elements to either enhance or degrade the security in wireless communication systems.

The remainder of this chapter is organized as follows. Section 2.2 reviews recent works of IRS on security. Section 2.3 presents an IRS-assisted communication system. Four basic security scenarios in IRS system are discussed in section 2.4. Simulation results are presented in section 2.5. Section 2.6 provides a summary of the study in this chapter.

2.2 Related Work

In this section, we summarize the research of IRS-aided wireless communications systems in security. IRS is commonly proposed to enhance SNR and, hence, to improve the quality of wireless signals. However, the use of IRS by malicious users could introduce security threats as well. Therefore, there are two major categories of research for using IRS in wireless communication in terms of security. The first category secures wireless communications by increasing the secrecy rate or SNR, which is a major performance metric for the legitimate receiver. The second category of IRS research with regard to security is to implement an attack (e.g., jamming) to degrade the receivers' ability to get the intended signal or plays as an accomplice to assist the eavesdropper [19]. The following surveys recent research in both categories.

2.2.1 Secure Wireless Communications Using IRS

The first category of security research on IRS enhances the secrecy rate or SNR of the communication between a legitimate transmitter and a legitimate receiver in the presence of an eavesdropper. The ability of IRS to dynamically alter the wireless channel is studied to enhance the secrecy rate with increased physical layer security in [20]. [21] presents an IRS-aided wireless communication system for security where an Access Point sends confidential messages to a user in the presence of a highly correlated eavesdropper in space and when the eavesdropping channel is stronger than the legitimate communication channel particularly. To maximize the secrecy rate, an optimization algorithm is proposed by jointly designing the AP's transmit beamforming and the IRS's reflecting beamforming.

An IRS-assisted Gaussian Multiple-Input Multiple-Output (MIMO) wiretap channel is carefully considered in [22]. The transmitter, receiver and eavesdropper are equipped with multiple antennas. They propose an alternating joint optimization algorithm to optimize the transmit covariance ($\mathbf{R}$) at the transmitter and phase shift coefficient ($\mathbf{Q}$) at IRS to maximize the secrecy rate. Specifically, a numerical algorithm is used to achieve the global optimal $\mathbf{R}$ when $\mathbf{Q}$ is fixed and the minorization-maximization algorithm is proposed to find the local optimal $\mathbf{Q}$ when $\mathbf{R}$ is fixed. [23] investigates an IRS-aided NOMA network and proposes a robust beamforming scheme using artificial noise to defend against a multi-antenna eavesdropper. The designed optimization scheme, which has two alternating steps is similar to [22]. [24] studies transmission optimization for IRS-aided multi-antenna systems where the source transmit power is limited and the unit modulus imposed on phase shifts at the IRS is constrained. Since this optimization problem is non-convex, they advocate an alternating algorithm as well to the transmit covariance of the transmitter and then build

a bisection search-based semi-closed form solution to the phase shift matrix of the IRS. In [25], authors explore a double IRS-assisted system, the inter-surface signal reflection is considered, to enhance the secrecy performance of the wireless transmission. Moreover, a product Riemannian manifold-based alternating algorithm is built to optimize the beamformer at the transmitter and phase shift coefficients at the double IRS.

Driven by its tremendous success in various fields such as image/speech recognition, online recommendation, and drug discovery, AI has also become a popular powerful tool in research on wireless communications in recent years, including IRS-assisted communications. In wireless communications, AI-based intelligent signal detection and user classification can promote better awareness of the communication channel. Furthermore, they will help to handle the complexities of wireless signals. The Intelligent Spectrum Learning (ISL) is used to tackle the interfering signals by dynamically controlling the IRS elements in [26]. The ISL algorithm, which consists of convolutional neural networks and fully connected layers, provides a multi-class classification for the incident signals. Then, the IRS elements can be turned on/off depending on the class of that signal using an IRS binary control. In the IRS system, a dynamic 'think-and-decide' function allows the reflection of incident signals to be blocked or passed based on the state of the IRS element block [27]. The use of online training and properly designed data could improve the SNR and reduce the overall system workload. In [28], IRS is deployed to prevent the communications of multiple legitimate users from eavesdropping in presence of multiple eavesdroppers. They propose a novel deep reinforcement learning-based secure beamforming algorithm to obtain the optimal beamforming policy since the system is highly dynamic and complex. Overall, those studies all demonstrate that by adjusting phase shifts of IRS elements, the signals reflected by IRS can be added constructively to the directed

path to enhance the desired signal power at the receiver.

2.2.2 Security Attacks Utilizing IRS

As mentioned previously, the use of IRS by malicious users could introduce security threats to wireless communications as well. An attack (e.g., jamming) can be realized by blocking the receivers' ability to get the intended signal. Therefore, one adverse application of IRS is that an IRS could be used as a green jammer to interfere with the communication between two legitimate devices and, hence, to degrade the SNR at the legitimate receiver [29]. Another adverse application of IRS is to help eavesdroppers (Eve) successfully steal information from legitimate users. Furthermore, after Eve steals statistical channel state information from the legitimate pair, it can establish synchronization with legitimate users and thus further conduct spoofing attacks [30]. In general, to launch the above-mentioned security attacks utilizing IRS, the main approach is to jointly optimize some critical factors related to IRS and, hence, to decrease the SNR at the legitimate receiver or increase SNR at the attacker.

2.3 Typical IRS Communication System

This section introduces an IRS-assisted wireless communication system. We then derive its SNR, one of the major performance metrics for evaluating a wireless communication system.

An IRS is composed of a vast number of re-configurable reflective elements and a microcontroller used to control the electromagnetic response of each reflector. All reflective elements can jointly adjust their phase shifts to make the reflected signals constructively or destructively added at the receiver. As illustrated in Fig. 1, an IRS with N elements is deployed between the transmitter Tx and the receiver Rx. Considering the line of sight (LOS) communication between Tx and Rx as well as reflected

signals by IRS, the received signal at the receiver, denoted by $r(t)$, is represented as:

$$r(t) = \underbrace{\sum_{n=1}^{N} f_n(\alpha_n e^{j\phi_n}) h_n^* \cdot m(t)}_{\text{N reflected path}} + \underbrace{h_{LOS}^* \cdot m(t)}_{\text{direct path}} + n(t), \tag{2.1}$$

where $m(t)$ is the transmitted signal from Tx and $n(t)$ is noise; α_n and ϕ_n are controllable magnitude and phase shift of the n^{th} element; f_n is the channel gain between Tx and IRS for the n^{th} reflected path, while h_n^* is the channel gain between IRS and Rx. h_{LOS}^* is the channel gain for the direct path between Tx and Rx.

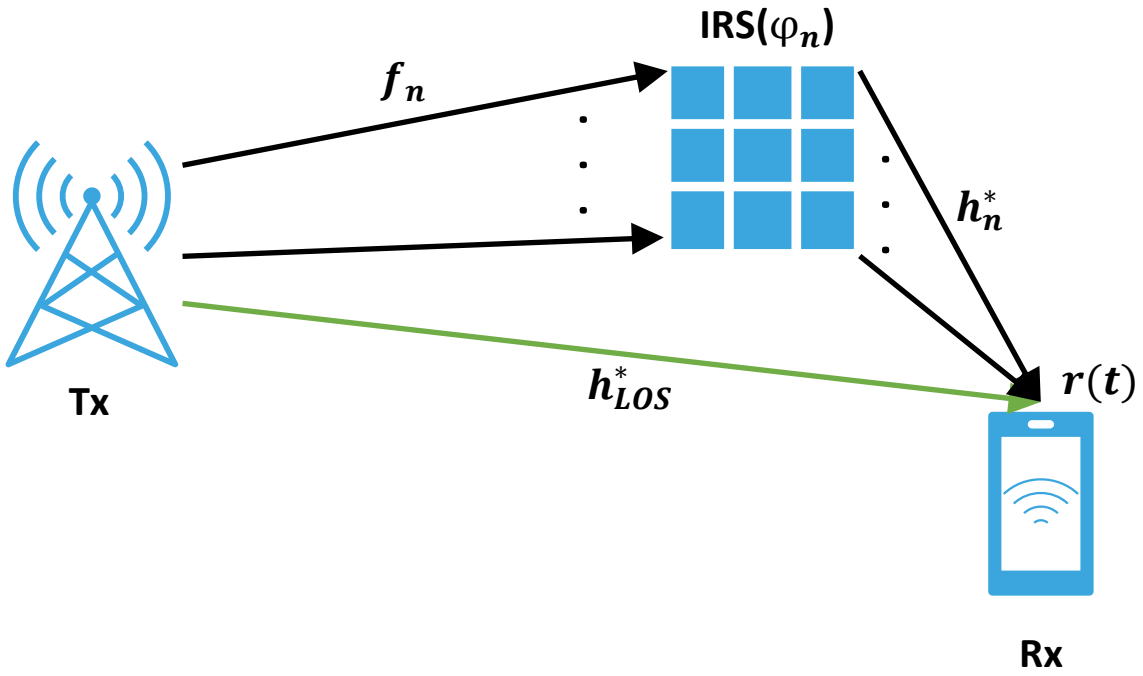

Fig. 1. IRS-assisted communication with one LOS path and N reflected paths.

We employ a channel model in the literature which provides the channel gains $\mathbf{g}^T$, $\overrightarrow{\mathbf{g}}$ and h_{LOS}^* [31]. Based on Eq. (2.1), the achievable SNR can be derived as:

$$\text{SNR} = \frac{P_t|\mathbf{g}^T\Phi\mathbf{h} + h_{LOS}^*|^2}{\sigma^2}, \tag{2.2}$$

where $\mathbf{g}^T = \left[\ f_1, f_2, ..., f_N\ \right]$, $\mathbf{h} = \left[\ h_1^*, h_2^*, ..., h_N^*\ \right]^T$, and P_t is the transmission power from Tx and σ^2 is the additive white Gaussian noise (AWGN) power (i.e., $\sigma^2 = E|n(t)|^2$), and $\Phi = diag\left\{\ \alpha_1 e^{j\phi_1}, \alpha_2 e^{j\phi_2}, ..., \alpha_N e^{j\phi_N}, ...,\right\}$. In terms of phase shift, the discrete phase-shift ϕ_n will be considered, which is more practical with less hardware requirement (e.g., $\phi_n = \{0, 2\pi/N..... 2\pi(N-1)/N\}$) [31].

From Eq. (2.2), it can be seen that the SNR is affected by several factors, including $\mathbf{g}^T$ (channel gain vector between Tx and IRS), $\overrightarrow{\mathbf{g}}$ (channel gain vector between IRS and Rx), h_{LOS}^* (channel gain for the direct path between Tx and Rx), the transmission power P_t and phase shift matrix Φ for IRS. By adjusting these factors, different SNRs can be obtained as needed, and hence, further improve wireless communications by increasing SNR or degrade communications by decreasing SNR.

2.4 Security and Threats Scenarios in IRS-Aided System

IRS can adaptively adjust the phase shifts of its reflecting elements to enhance the intended signals or attenuate the undesired signals to secure wireless communication. Therefore, we propose four basic security scenarios in the IRS-aided system from the constructive and adversarial aspects. In the first two scenarios, IRS is used by legitimate users and hence to protect the legitimate receiver from eavesdropping and attacks. In the last two scenarios, IRS is manipulated by malicious users and used for conducting hostile attacks such as eavesdropping and interfering with legitimate users. The following presents four scenarios mentioned above.

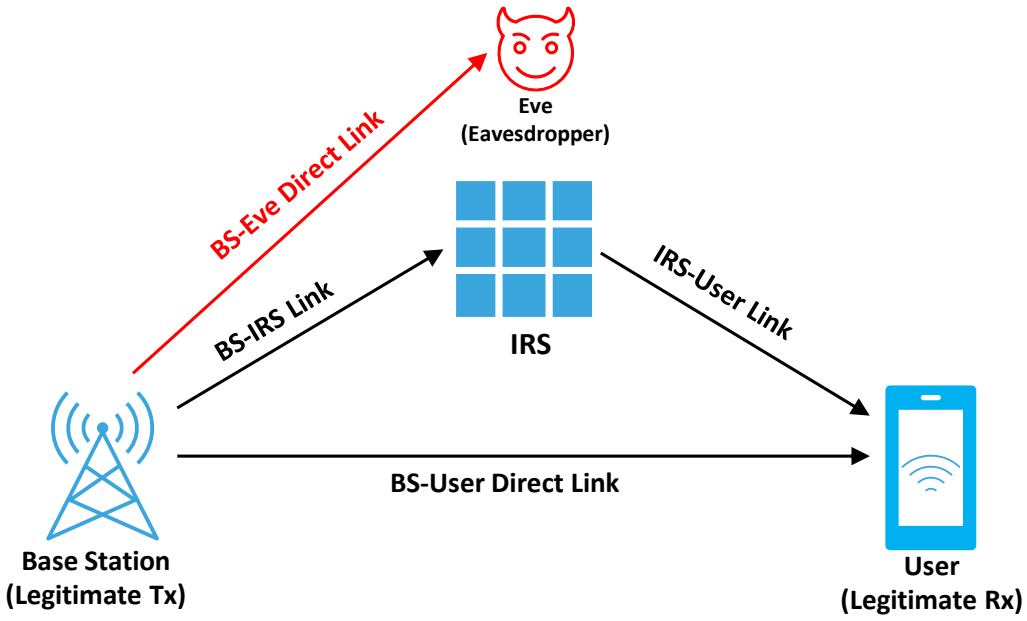

Fig. 2. IRS enhances the communication between legitimate transmitter and receiver by increasing the SNR between them.

***Scenario A**: IRS Protecting Legitimate Receiver from Eavesdropping*

In this scenario, IRS is designed to prevent eavesdropping and enhancing the SNR of the communication between a legitimate transmitter, e.g., a Base Station (BS), and a legitimate receiver, e.g., a User, in the presence of an eavesdropper (i.e., Eve), as depicted in Fig. 2. To prevent eavesdropping, the BS transmits signals with a lower power so that the eavesdropper cannot receive the signal from the BS successfully. Meanwhile, an IRS is programmed to constructively reflect the signal from the BS to the legitimate receiver. Hence, the user receives the direct signal from the BS through BS-User direct link as well as constructive reflected signals via IRS through IRS-User link at the same time. As a result, the SNR of the communication between the BS and the legitimate user could be significantly increased while the SNR between the BS and eavesdropper is sufficiently low.

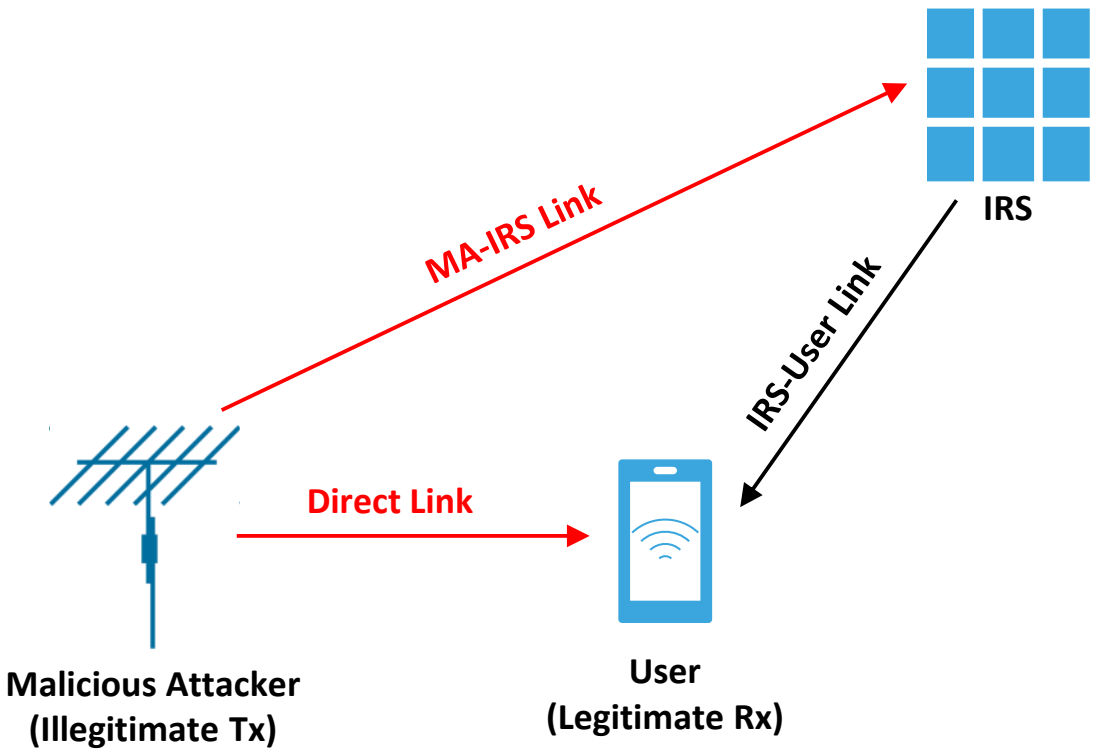

Fig. 3. IRS mitigates attack from the illegitimate transmitter for the legitimate receiver by decreasing the SNR between them.

***Scenario B**: IRS Mitigating Attack from Illegitimate Transmitter*

In this scenario, IRS is implemented to assist the receiver in mitigating attacks from a malicious attacker, as shown in Fig. 3. The malicious attacker sends the signal to the user through the direct link. An IRS is programmed to reflect the malicious signal destructively to the user through the Attacker-IRS-User link (i.e., MA-IRS link and the IRS-User link in Fig. 3) by adjusting their reflection coefficients and phase shift. Consequently, the SNR between the malicious attacker and user can be significantly decreased. This efficiently mitigates the negative impact of malicious signals and enhances the resilience and security of communication.

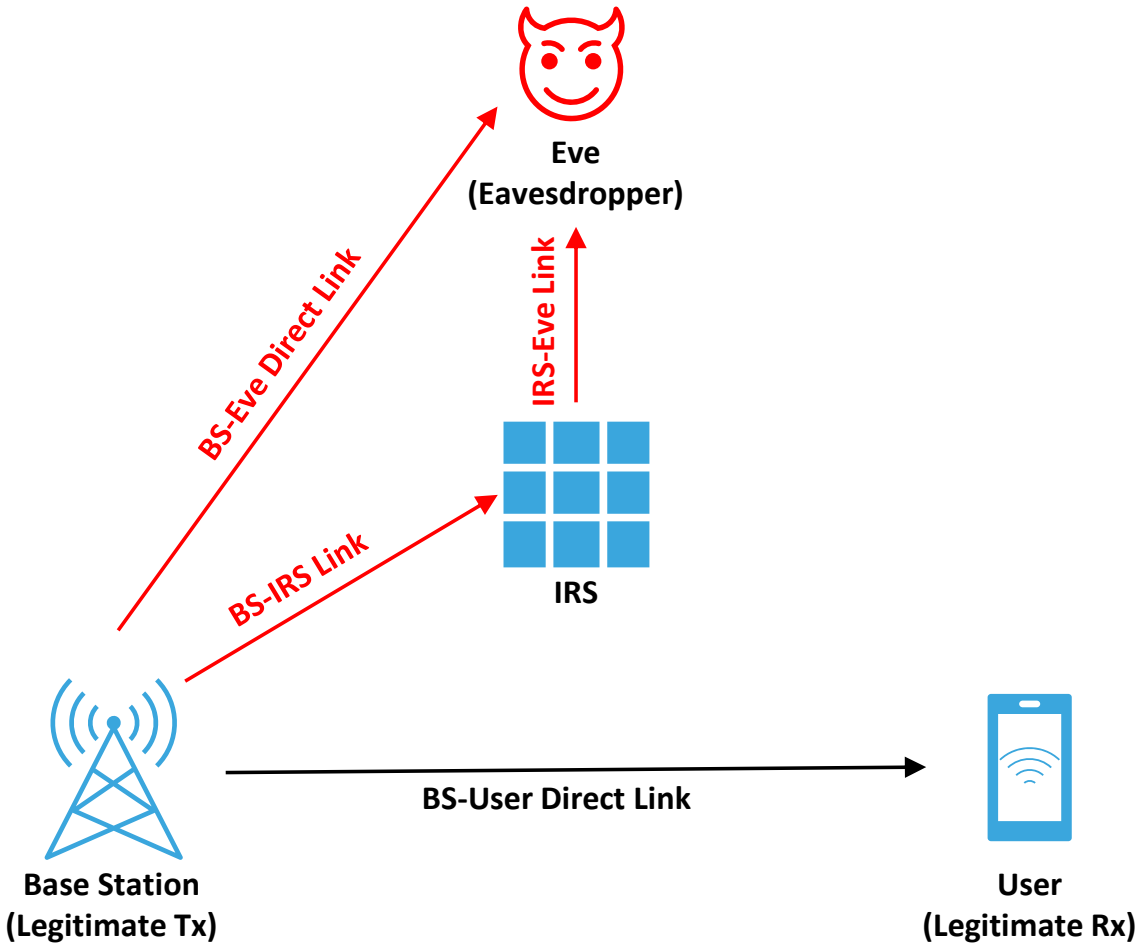

Fig. 4. IRS enhances the eavesdropping for the eavesdropper by increasing the SNR between the legitimate transmitter and eavesdropper.

***Scenario C**: IRS Enhancing Eavesdropping for an Eavesdropper*

In this scenario, we consider one possible adverse application of IRS that improves the eavesdropping performance for the eavesdropper, as depicted in Fig. 4. A BS sends signals to the user through the BS-User direct link, while Eve receives the signals by the BS-Eve direct link. The eavesdropper controls IRS and makes the reflected signals constructive from BS to Eve via the BS-IRS-Eve link (i.e., BS-IRS

link and the IRS-Eve link in Fig. 4). Thus, the SNR of the communication between the BS and the eavesdropper is increased. Consequently, Eve can successfully steal information sent from the legitimate transmitter.

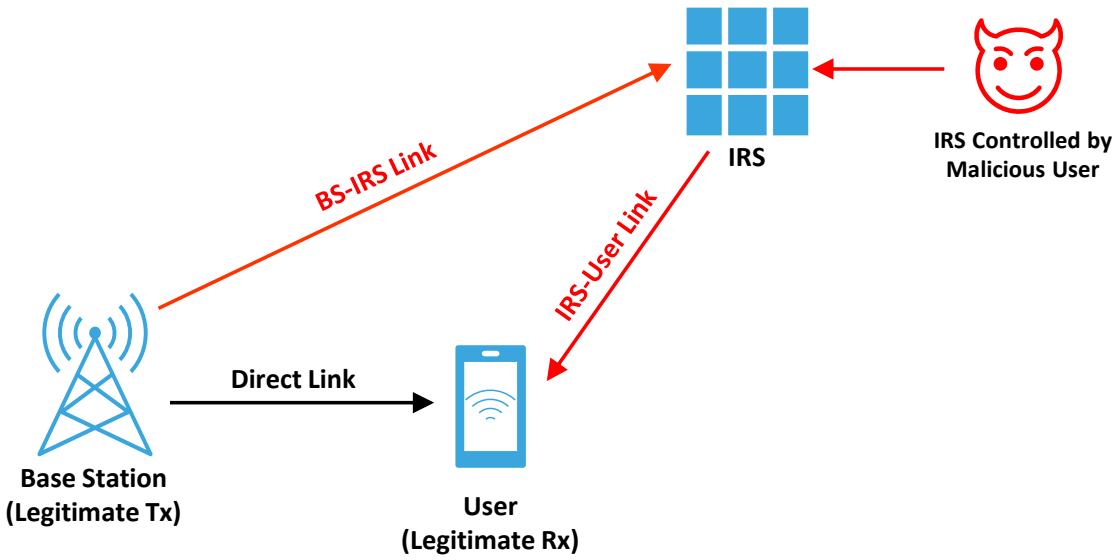

Fig. 5. IRS interferes with the communication between the legitimate transmitter and receiver.

***Scenario D**: IRS Interfering Communication between a Legitimate Pair of Transmitter and Receiver*

In this scenario, IRS plays the role of a green jammer to interfere with the communication between the legitimate transmitter and receiver, as illustrated in Fig. 5. The BS transmits signals to the user through a direct link. IRS is controlled by the jammer to reflect the signals from the BS to the user destructively through the reflected links (i.e., BS-IRS direct link and the IRS-User link in Fig. 5), so that the SNR between BS and the receiver is decreased. IRS-based jammers, unlike conventional jamming attacks, use legitimate signals directly by changing their reflection coeffi-

cients and phase shifts instead of relying on internal energy to send strong signals to a victim system. Due to its ability to interfere with a system without leaving any energy footprint, IRS-based jamming is practically impossible to detect.

2.5 Performance Evaluation

As we discussed in Section 2.4, there are four major scenarios for IRS-aided communications to enhance security or launch attacks. Those four scenarios can be divided into two categories in terms of the impact of IRS on SNR. The first category is the SNR improvement, in which the SNR is increased for a legitimate receiver to improve the communication performance, as in Scenario A , or for an eavesdropper to assist in eavesdropping, as in Scenario C. Another category is the SNR reduction via IRS. We decrease the SNR for a legitimate receiver to mitigate attacks from an illegitimate transmitter as in Scenario B or for malicious users to interfere with legitimate communication as in Scenario D.

In this section, we will conduct simulations and measure the SNR of receiver, as derived in Eq. (2.2), for both two categories. Some parameters in Eq. (2.2) can be obtained using the SimRIS channel simulator [31]. For the simulation setting, an outdoor environment is considered with an IRS of N elements. The transmission power from the Transmitter (Tx) is denoted as P_t and the noise is assumed as $-100dBm$. For each category, SNR is examined with different IRS positions.

2.5.1 SNR Improvement by IRS

In this subsection, we will evaluate the performance for the first category, including Scenario A and Scenario C. Specifically, we will conduct simulations and examine SNR improvement by controlling the phase shift and positions of IRS. The positions of transmitter (Tx) and receiver (Rx) are fixed at $(x_1 = 0, y_1 = 25, z_1 =$

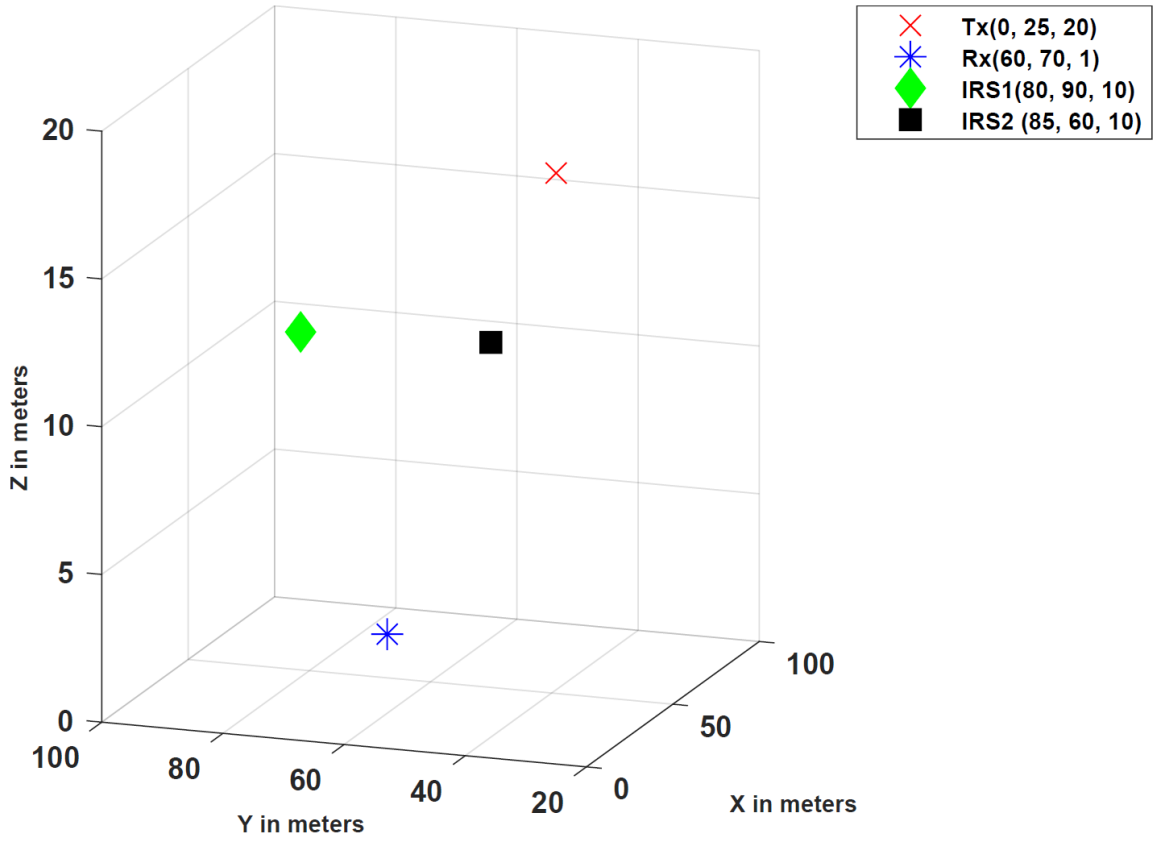

Fig. 6. Tx, Rx and two IRS positions for simulations to improve SNR between Tx and Rx.

20) and $(x_2 = 60, y_2 = 70, z_2 = 1)$ in meters. The coordinates of two IRS are $IRS1(x_3 = 80, y_3 = 90, z_3 = 10)$ and $IRS2(x_3 = 85, y_3 = 60, z_3 = 10)$ in meters, respectively. The total received signal at the receiver (Rx) consists of the direct signal from Tx to Rx and N reflected signals. All IRS elements are configured to have the same phase shift, e.g., $\phi_n = \frac{2\pi}{\lambda}\big(\sqrt{(x_1 - x_3)^2 + (y_1 - y_3)^2 + (z_2 - z_3)^2} + \sqrt{(x_3 - x_1)^2 + (y_3 - y_1)^2 + (z_3 - z_1)^2}\big)$.

With the above configurations, we evaluate SNR between Tx and RX when IRS is deployed at two different positions to explore SNR improvement, respectively, as shown in Fig. 6. The resulting SNR values for the two positions are plotted in Fig. 7 and Fig. 8, respectively. SNR is assessed versus P_t, without IRS as well as with

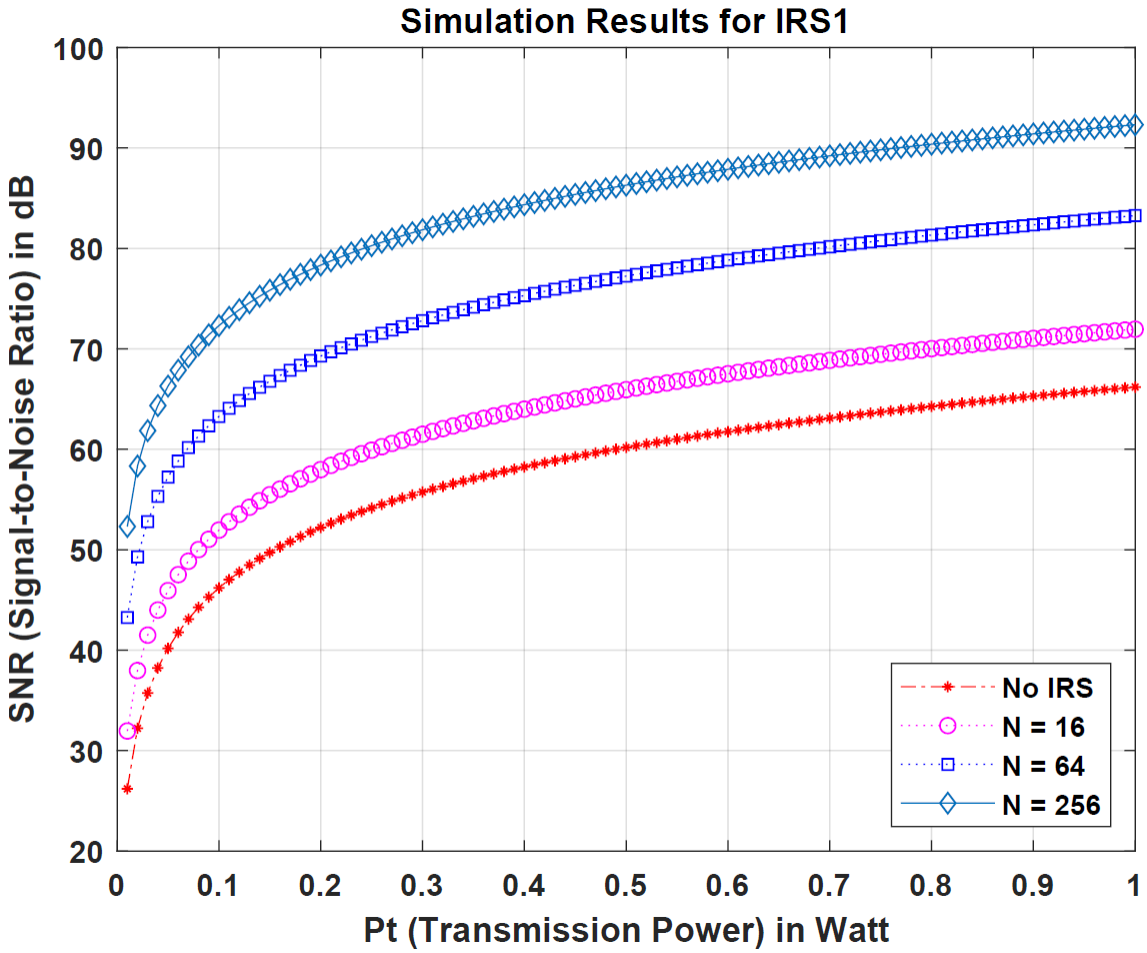

Fig. 7. SNR improvement with IRS at position 1 (IRS1).

IRS consisting of N number of reflecting elements. These two figures reveal a trend: (1) SNR at Rx is significantly improved thanks to IRS; and (2) SNR clearly improves from 0dBm to 30dBm while the transmission power P_t increases. In addition, the difference of SNR results between Fig. 7 and Fig. 8 suggests that the position of IRS matters for SNR. Our results verify that IRS could be an important tool to improve SNR with appropriate phase shifts of reflective elements.

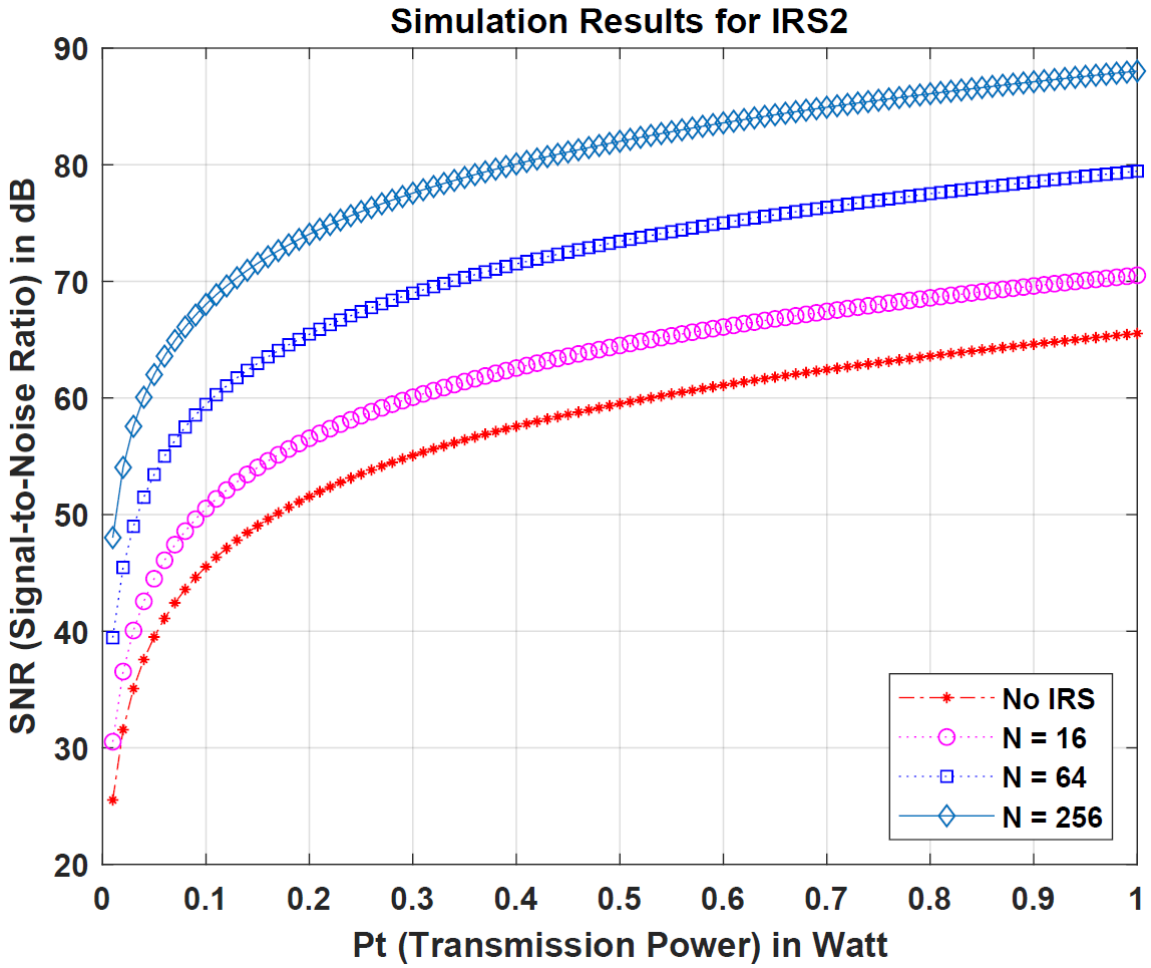

Fig. 8. SNR improvement with IRS at position 2 (IRS2).

2.5.2 SNR Reduction by IRS

In this subsection, we evaluate the performance for the second category, including Scenario B and Scenario D. We examine the SNR reduction between Tx and Rx via controlling IRS. The use of IRS to decrease the SNR at Rx is examined with the following setting. The positions of Tx and Rx are fixed at $(x_1 = 0, y_1 = 50, z_1 = 6)$ and $(x_2 = 0, y_2 = 30, z_2 = 2)$ in meters. The positions of two IRS are $(x_3 = 0, y_3 = 10, z_3 = 4)$ and $(x_3 = 0, y_3 = 5, z_3 = 3)$ in meters, respectively. The positions of Tx, Rx, and two positions of IRS are shown in Fig. 9.

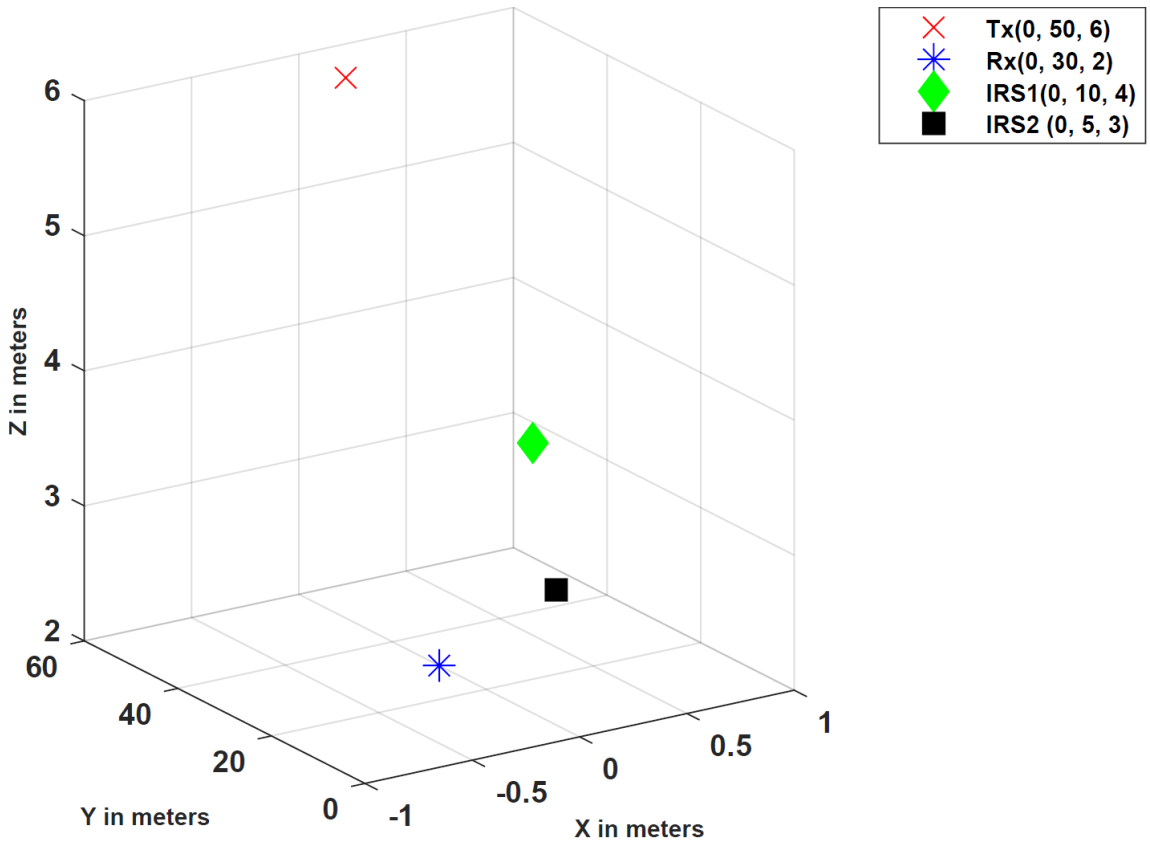

Fig. 9. Tx, Rx and two IRS positions for simulations to decrease SNR between Tx and Rx.

The SNR between Tx and Rx is assessed in a similar manner for each IRS position. Results of SNR evaluation for the two positions are plotted in Fig. 10 and Fig. 11, respectively. The SNR is plotted versus P_t, the transmit power, without IRS as well as with IRS consisting of N number of elements for both IRS positions. SNR in this figure follows a similar trend: (1) SNR is significantly decreased by a malicious user that controls IRS; and indeed (2) using IRS for malicious intent, SNR clearly improves while the transmit power P_t increases. SNR changes between two positions of IRS suggest that the location of the IRS affects the wireless communication channel. These results reveal that the use of the IRS in a harmful way could degrade communication considerably by controlling phase shifts of reflective elements and the

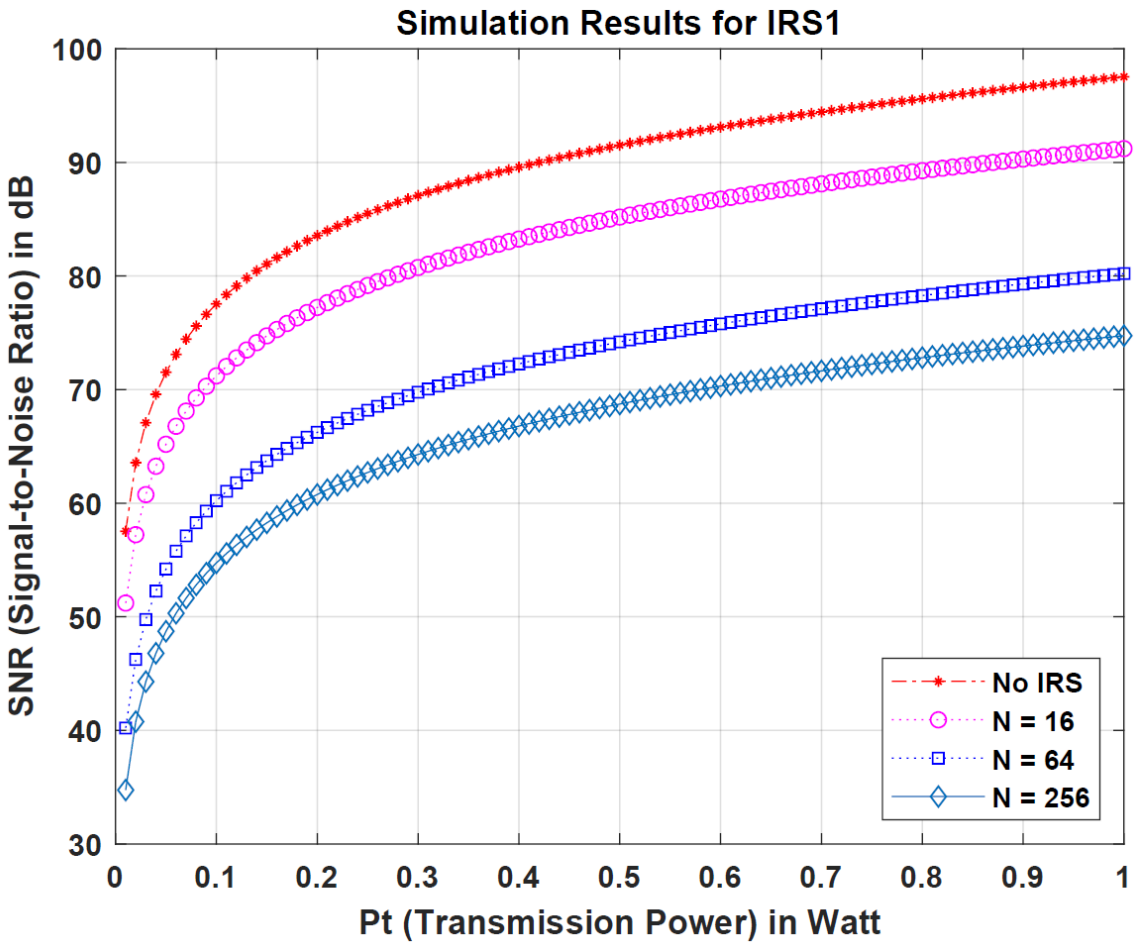

Fig. 10. SNR reduction with IRS at position 1 (IRS1).

position of the IRS.

2.6 Conclusion

The security of 5G and NextG networks has been a constant concern. The potential of integration of recent emerging technologies such as IRS into 5G and NextG may provide promising solutions from the security perspective. IRS has been widely studied to support higher throughput and performance, whereas research on the security of the IRS-aided system still lags. In this chapter, we have studied the potential secu-

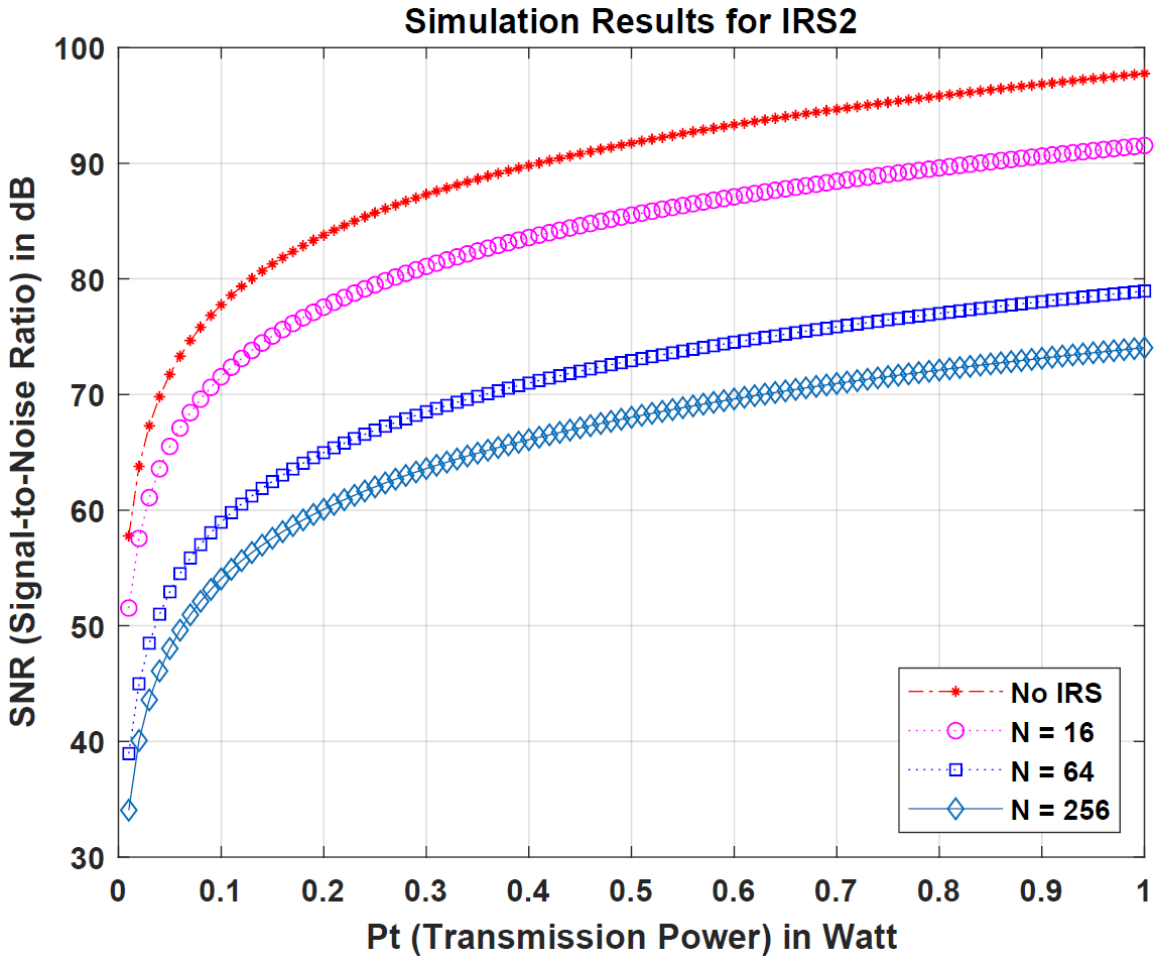

Fig. 11. SNR reduction with IRS at position 2 (IRS2).

rity impact of IRS on wireless communication systems from two categories: improving secure communication and launching security attacks using IRS. Both categories benefit from the flexibility of reprogramming a communication environment by adjusting its position and phase shifts of reflecting elements. Simulations have been conducted for four major scenarios. Our study is expected to shed light on a deep understanding of IRS-assisted networks concerning security.

CHAPTER 3

WIRELESS SIGNAL DENOISING USING CONDITIONAL GAN

3.1 Introduction

5G and beyond technologies are promising to unlock the potential benefits of wireless networks, including higher capacity, faster data rate, low latency, massive device connectivity, and reduced cost compared with traditional networks [1, 27]. Wireless security becomes one of the top concerns as 5G and beyond technologies become more widely available [4]. Wireless signal strength plays a critical role in wireless security. For instance, to prevent eavesdropping for security purposes, we can intentionally reduce the signal power at a transmitter so that eavesdroppers have difficulties of receiving the signal due to weak strengths. Although legitimate receivers will also experience noise and interference during transmission, signal denoising techniques can be used at the receiver side to boost the SNR.

Wireless signal denoising remains a challenge, and various methods have been developed to tackle this challenge. There are two types of traditional methods: linear and nonlinear methods [32]. It is not satisfactory to use linear algorithms since they assume signals are stationary, while real signals are typically non-stationary. Nonlinear methods are then developed and widely used, while the performance could be limited if the basis function and threshold are not selected appropriately [33]. In the last decade, deep learning has achieved tremendous success in various areas, such as computer vision, natural language processing, etc [34]. This breakthrough motivated researchers to implement signal denoising via a learning manner. As an example of machine learning-based solutions, Deep Neural Network (DNN) has been

applied to signal strength improvement, which outperforms traditional methods [35, 36]. In [35], the authors demonstrated DNN could learn and analyze the noise and interference characteristics of the wireless channels. Deep learning can perform better than traditional methods when the wireless communication environment is complex, where the signals are mixed with substantial noise.

In conclusion, both the traditional and DNN-based signal denoising methods are mainly used to estimate the features of noise or interference and attempt to remove them from the time or frequency domains. However, the wireless communication environment changes dynamically, and the impact of noise or interference on the signals is unpredictable, which makes it difficult to estimate and filter out the noise or interference precisely. Therefore, a novel method is desired to estimate the features of the wireless signals, instead of the noise or interference, to deal with the dynamic environment.

To tackle the above-mentioned challenges in wireless signal enhancement, this chapter proposes a novel machine learning-based solution. We will explore the potential of GAN [37] to estimate the features of the noise-free signals in an adversarial manner and then remove the dynamic noise and interference by generating the denoised signals directly [38]. Specifically, we propose to apply a conditional Generative Adversarial Network (cGAN) for signal denoising [39] at the receiver, which establishes a min-max problem between a generator and a discriminator. The generator attempts to generate a denoised signal without noise and interference. Meanwhile, the discriminator aims to supervise the generation process to ensure the denoised signal is from its corresponding noisy signal. After effective training, the well-trained generator will be deployed to denoise incoming signals automatically at the receiver.

The main contributions of this chapter are summarized as follows.

- We develop a conditional generative adversarial architecture that consists of two crucial components: a generator and a discriminator for signal denoising. A loss function specialized for wireless signal denoising is designed for training both components. The proposed signal denoising method is an end-to-end model, where the denoised signal can be generated automatically from the input noisy signal without any manually designed feature extraction.

- The generator of an original GAN may randomly generate denoised signals that do not correspond to the input noisy signals. The reason is that the generator only learns the distribution of real noise-free signals and no control is imposed to force the generated denoised signals related to the input noisy signals. To overcome this shortcoming, we use the noisy signals as conditional information for both the generator and discriminator. This will ensure that the resulting denoised signal is from its noisy signal rather than an uncorrelated one.

- Simulations are conducted for four wireless signal modulation types (i.e., Binary Phase-shift Keying (BPSK), Quadrature Phase Shift Keying (QPSK), Quadrature Amplitude Modulation (QAM16), and QAM64) to evaluate the denoising performance of our proposed method. The results verify the effectiveness and feasibility of our method and show superior performance across all comparisons.

The rest of this chapter is organized as follows. Section 3.2 describes the related work of signal denoising and GAN. Section 3.3 proposes the framework of the cGAN-based wireless signal denoising method. We present the architecture of the cGAN-based signal denoising method, the dataset preparation, and the training procedure in section 3.4. In Section 3.5, we test the performance of our proposed method on different types of noisy signals with comparisons to other approaches. Section 3.6 draw conclusions and future works.

3.2 Related Work

In this section, we first summarized signal denoising methods into two groups: traditional and Deep Learning (DL)-based methods to provide an overview of the development of signal denoising. We then introduce the GAN and cGAN briefly.

3.2.1 Traditional Signal Denoising Methods

Traditional signal denoising methods can be classified into two groups: linear methods and nonlinear methods [32]. Linear signal denoising methods have been widely applied to remove noise due to their relative simplicity. Typical linear approaches of signal enhancement for received signals are based on the Least Mean Square (LMS) or its variants [40]. However, the performance of these algorithms on nonlinear signals is not satisfactory. It cannot provide an optimal solution for noise and interference removal since these algorithms assume the signals are stationary, while the real signals generally have non-stationary features. Nonlinear methods are proposed and commonly used because they can simultaneously interpret the spectral and temporal features of signals. The nonlinear threshold wavelet transform method is widely used and adopted due to its strong local time-frequency analysis ability [41]. The basic idea is to separate signals into various segments according to frequency range and filter out noise and interference by a reasonable threshold [33]. However, those nonlinear methods lack the characteristics of self-adaptation. The denoising performance could be limited if the selected wavelet basis function and threshold are inappropriate. Thus, it is challenging for nonlinear denoising methods to tackle the dynamic impacts on signals in the changing wireless communication environment.

3.2.2 DL-based Signal Denoising Methods

DL has recently been applied to signal denoising. Specifically, DNN is proposed for signal strength improvement, which outperforms traditional methods [35, 36, 42]. In [35], the authors demonstrated DNN could learn and analyze the noise and interference characteristics of the wireless channels. Moreover, [36] proposed a signal denoising method based on a convolutional neural network. The model can estimate the noise in the signals, and the high-quality denoised signal can be obtained by subtracting the estimated noise from the raw signal.

3.2.3 Conditional Generative Adversarial Network

GAN [37] is a promising tool for learning data distribution and has been used for image denoising [43], and speech enhancement [38]. GAN typically comprises two adversarial components: a generator G and a discriminator D. After they complete a minimax game, the generator G can generate samples that are very close to the real data. Both G and D could be non-linear mapping functions, such as a multi-layer perception. This unique ability of GAN inspired us to investigate the feasibility of applying GAN to signal denoising. However, there is no control over the modes of data being generated in a typical GAN. To address this issue, cGAN [39] proposes to add conditional information regarding the expected outputs of the model to control the generation process.

3.3 cGAN-based Method for Signal Denoising

This section introduces our proposed approach to applying a cGAN for wireless signal denoising. First, we describe two critical phases, the training and deployment phases, and then discuss the tailored objective function for signal denoising.

3.3.1 Wireless Signal Denoising Using cGAN

We adopt cGAN, instead of an original GAN, as our solution to wireless signal denoising because the generator of an original GAN may randomly generate signals based on the distribution learned from the real noise-free signals. In other words, the generated signal from an original GAN does not match the corresponding noisy signal well. Our proposed cGAN-based wireless signal denoising approach consists of two phases, namely the training and deployment phases, which are presented below.

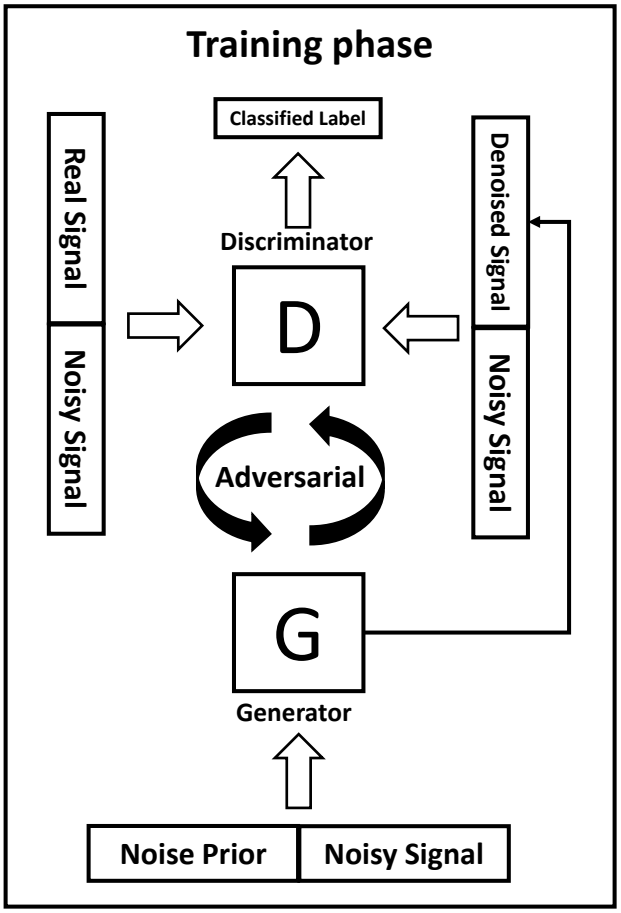

Fig. 12. The training phase of our proposed method for wireless signal denoising. We train the generator G and discriminator D in an adversarial manner using the noisy signal S_N as the condition.

Training phase: Our goal is to learn the potential features in real noise-free signal S_R and then remove noise and interference in the noisy signal S_N by generating the denoised signal $G(S_N)$, which is expected to be very close to the real noise-free signal S_R. To achieve this goal, we treat the wireless signal denoising problem as a conditional adversarial problem, as illustrated in Figure 12. Specifically, the noisy signal S_N is considered as the condition in cGAN. The generator G receives the condition along with a random noise prior z [37], which has the same dimension with real signal S_R and then generates the denoised signal $G(S_N)$. The adversarial component discriminator D is a typical binary classifier that receives the real noise-free signal S_R concatenated with the condition S_N or the generated denoised signal $G(S_N)$ concatenated with the condition S_N. The goal of the discriminator D is to distinguish the real noise-free signal S_R and the generated denoised signal $G(S_N)$. We train the generator G and discriminator D simultaneously in an adversarial manner. During the training, the generated denoised signal $G(S_N)$ will gradually become extremely close to the corresponding real noise-free signal S_R. The training phase is completed offline.

Deployment phase: Once the generated signal $G(S_N)$ is indistinguishable by the discriminator D during the training, we obtain the desired generator G that can be deployed at the receiver for wireless signal denoising as illustrated in Figure 13. Given a noisy signal not included in the training dataset, we concatenate it with a random noise prior z and feed them into the well-trained generator G. The generator G is able to subsequently output the denoised signal $G(S_N)$. It is noticeable that G has learned the data distribution of the real signal S_R rather than memorizing the input-output pairs.

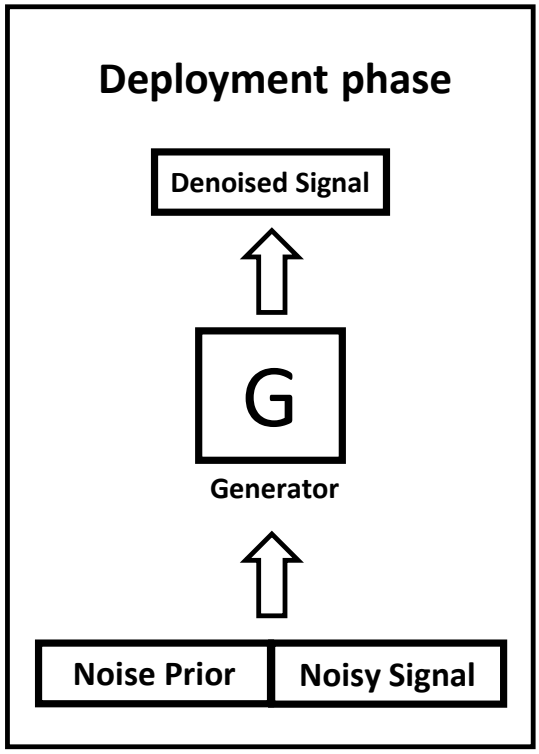

Fig. 13. The deployment phase of our proposed method for wireless signal denoising. We deploy the generator G once the generated denoised signal $G(S_N)$ is indistinguishable by the discriminator D.

3.3.2 Objective Function Design

In this subsection, we will design a specialized objective function tailored for wireless signal denoising. The objective function of the original cGAN does not work effectively for signal denoising. We first point out its weakness and then propose a new objective function considering two aspects.

We denote the real signal data distribution as $S_R \sim p_{data}(S_R)$ and the noise prior distribution as $z \sim p_z(z)$. For an original cGAN-based signal denoising approach, the objective function of the min-max game between the generator G and discriminator D is expressed as:

$$\min_G \max_D V(D,G) = \mathbb{E}_{S_R \sim p_{data}(S_R)}[\log D(S_R|S_N)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z|S_N)))], \tag{3.1}$$

where $D(S_R|S_N)$ indicates the discriminator D labels the real signal and outputs the expected value of 1. $D(G(z|S_N))$ represents the discriminator D, which distinguishes the generated denoised signal and outputs the expected value of 0. During the training phase, G tries to minimize $\log(1 - D(G(z|S_N)))$ and D aims to maximize $\log D(S_R|S_N)$.

However, the cross entropy loss function in the Equation (20) may lead to the vanishing gradients problem during the training phase. To address this problem, we adopt the least square loss function [44] to the objective function, which is improved as:

$$\begin{aligned} \min_G V(G) &= \frac{1}{2}\mathbb{E}_{z \sim p_z(z)}[D(G(z|S_N)) - 1]^2, \\ \min_D V(D) &= \frac{1}{2}\mathbb{E}_{S_R \sim p_{data}(S_R)}[D(S_R|S_N) - 1]^2 \\ &\quad + \frac{1}{2}\mathbb{E}_{z \sim p_z(z)}[D(G(z|S_N))]^2. \end{aligned} \tag{3.2}$$

In practice, the least square loss penalizes the samples located at a long distance from the decision boundary, which has two benefits. First, it makes the samples generated by the generator close to the decision boundary, and thus the generator can generate denoised signals with higher quality. Second, it generates more gradients and thus makes the training phase more stable than the original GAN. To make the generated denoised signals $G(z|S_N)$ close to the real noise-free signals further, we consider mixing the objective function in Equation (22) with L1 and L2 loss. Therefore, the newly designed objective function becomes:

$$O = \min_G V(G) + \min_D V(D) + \mathcal{L}_1(G) + \lambda \mathcal{L}_2(G), \tag{3.3}$$

where $\mathcal{L}_1(G)$ and $\mathcal{L}_2(G)$ represent the L1 and L2 loss, as shown in the following two formulas:

$$\mathcal{L}_1(G) = \mathbb{E}_{S_N,S_R,z}[\|S_R - G(z|S_N)\|_1], \tag{3.4}$$

$$\mathcal{L}_2(G) = \mathbb{E}_{S_N,S_R,z}[\|S_R - G(z|S_N)\|_2], \tag{3.5}$$

where $\|\cdot\|_p$ stands for the vector $p-$norm. L2 loss penalizes the objective function heavier than L1 loss, which allows L2 loss to force the generated denoised signals related to the noisy signals and have fewer outliers. However, L2 loss is too sensitive to outliers in the early training stage. Therefore, λ in Equation (4.4) is used to strike a balance between the L1 and L2 loss. In the early training stage, λ will be set to a small value to encourage sparsity and make the model less sensitive to outliers. Then, λ increases gradually as the model is being trained. We will discuss the details in Section 4.4.2.

3.4 Implementation

In this section, we first describe the network structures of the generator and discriminator. Then, the dataset preparation used for training and testing is introduced. The training details are discussed as well.

3.4.1 Network Architecture

By conducting extensive preliminary experiments, we concluded the network architectures for the generator G and the discriminator D as follows.

Generator: The network structure of the generator G is a five-layer neural network composed of three convolutional layers and two deconvolutional layers, as shown in Figure 14. The convolutional layers are designed to extract the features

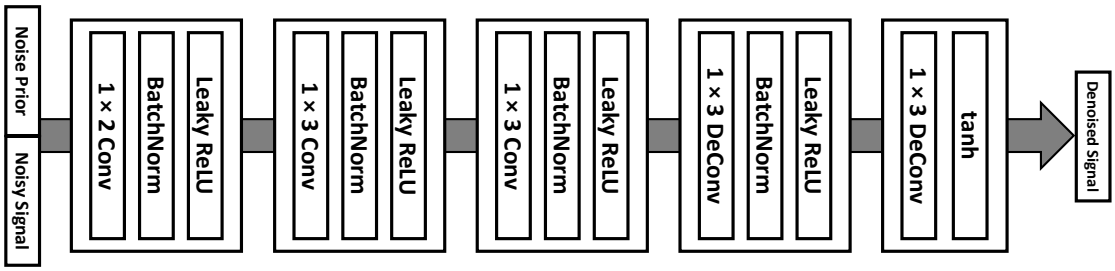

Fig. 14. The network structure of generator G. G is a five-layer neural network that consists of three convolutional layers and two deconvolutional layers.

in noisy signals and map the noisy signals into feature maps. The deconvolutional layers are applied to recall the extracted features and generate the denoised signal based on the extracted feature maps. The convolutional layers and deconvolutional layers play the roles of encoder and decoder, respectively. 1×2 filters are applied in the first convolutional and the stride is set to 2. The 1×3 filters are used and the stride is set to 1 in the last four layers. There are no padding operations in any layers. The Leaky-ReLU [45] activation function and batch normalization [46] are applied for each layer except the last layer to speed up the training and make the training process more stable. The tanh activation function is adopted to output the denoised signals in the last layer.

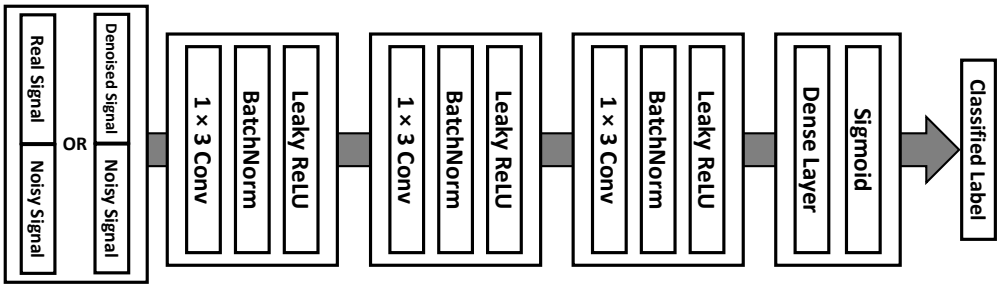

Fig. 15. The network structure of discriminator D. D is a four-layer neural network that consists of three convolutional layers and one dense layer.

Discriminator: The network architecture of discriminator D is a four-layer neural network consisting of three convolutional layers and one dense layer, as illustrated in Figure 15. The convolutional layers are used to extract the feature maps of the input signals, and the dense layer is used to distinguish whether the input signals are real noise-free signals or generated denoised signals. We use 1×3 filters in all convolutional layers and set the stride to 2. We also use the Leaky-ReLU activation function and batch normalization for each layer except the last layer. The sigmoid function is used for binary classification in the last layer.

3.4.2 Dataset Preparation

To train and test our cGAN-based denoising method, we need to generate a wireless signal dataset that consists of the original real signals S_R and the corresponding noisy signals S_N. We first generate the real signals and each signal sample is associated with one specific modulation at a particular SNR. In our experiments, real signals with four different modulations, i.e. BPSK, QPSK, QAM16, and QAM64, at different SNR levels from -8 dB to 8 dB are generated in the time domain. Then the corresponding noisy signals S_N are constructed by adding additive Gaussian white noise (AWGN) to S_R and overlapping other types of signals with small amplitudes over S_R. We denote a pair of data as (S_R, S_N). A total of $200,000$ pairs denoted as $D_{data} = \{(S_R, S_N)\}$, were generated and the generation process is implemented in Matlab. The generated data set D_{data} is split into 90% training dataset and 10% testing dataset.

3.4.3 Training Details

In the training phase, we sample batch data $\{(S_R, S_N)\}$ from the training dataset, and the batch size is set to 200. We first use the batch data to train the discriminator

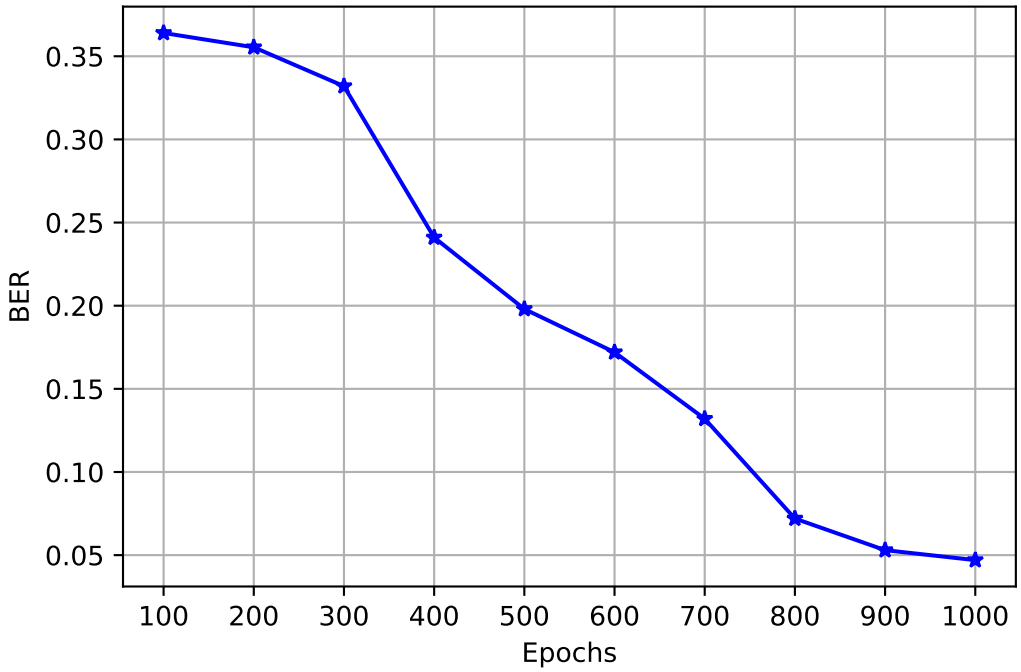

Fig. 16. The BERs of the denoised signals generated by the generator at every 100 epochs in the training phase.

t times to optimize the parameters of the discriminator. The reason is that the discriminator D is too weak to distinguish the real signals and the generated denoised signals at the beginning of the training phase, and thus, it can not provide useful feedback to the generator G. t is suggested to be 5. Then, we train G one time, and the learning rate is set as 10^{-5} with Adam optimizer for both D and G. We use the alternative training manner above to train D and G in n epochs. Our preliminary experiments observed that the generator could not significantly reduce the Bit Error Rates (BERs) of noisy signals after 1000 training epochs. Therefore, the training epochs n is set to 1000, and the BERs of the generated denoised signals at every 100 epochs during the training are shown in Figure 16. λ in the final object function Equation (4.4) is set to 0 before 800 training epochs. In the last 200 epochs, λ

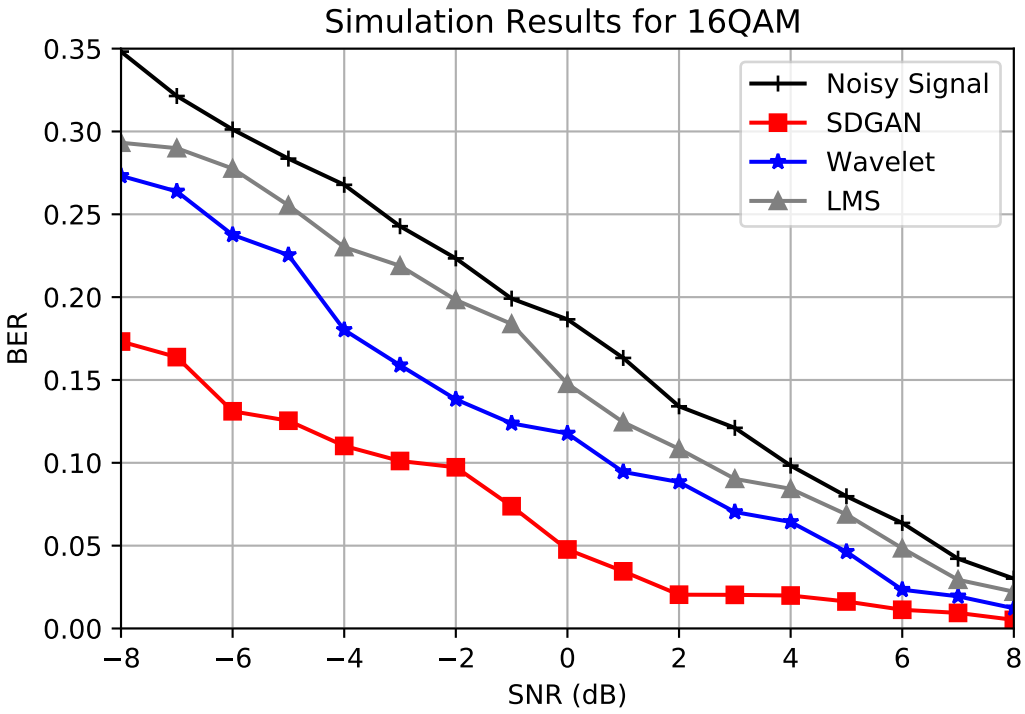

Fig. 17. The denoising performance comparison of our SDGAN against Wavelet and LMS on noisy signals with 16QAM at different SNR levels.

increases to 1 linearly.

3.5 Simulation Results

After the training phase, the generator is capable of reducing the BERs of noisy signals and ready for deployment at the receiver. To evaluate the performance of our proposed wireless signal denoising method (referred as Signal Denosing GAN (SDGAN), we conduct simulations using the testing dataset. We also compare SDGAN with the classical linear algorithm LMS and the nonlinear threshold wavelet transform (referred as Wavelet) method. All simulations are implemented using Nvidia GTX 1080Ti GPU with Pytorch.

We first choose noisy signals with 16QAM and 64QAM at different SNR levels for

simulations and also compare it with the two traditional signal denoising algorithms. The simulation results are shown in Figure 17 and Figure 18. The black line indicates the BERs of the noisy signals before the denoising processing. The grey, blue, and red lines represent the denoising results of LMS, Wavelet, and SDGAN, respectively. It can be seen that all three methods are capable of reducing the BERs of noisy signals at all SNR levels. However, SDGAN can reduce the BERs of noisy signals by over 50% and provide the best denoising performance across the simulations.

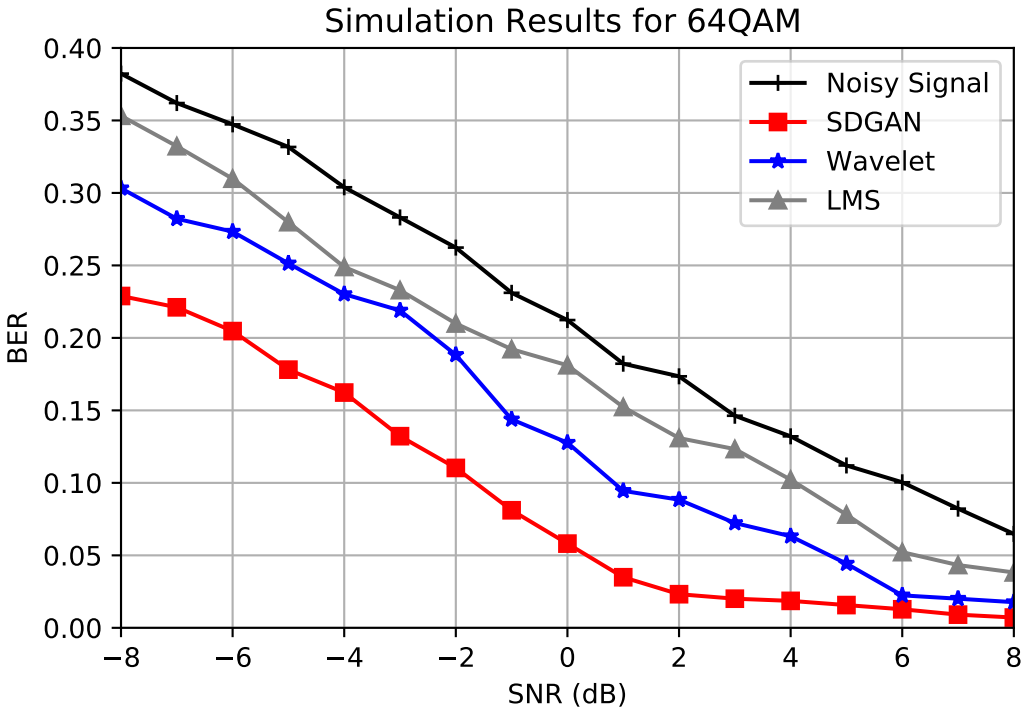

Fig. 18. The denoising performance comparison of our SDGAN against Wavelet and LMS on noisy signals with 64QAM at different SNR levels.

In addition, we also evaluate the denoising performance of SDGAN on noisy signals with BPSK or QPSK modulations to verify the general effectiveness of SDGAN on different modulation types. As exhibited in Figure 19, the two blue lines with different shapes represent the BERs of noisy signals with BPSK and QPSK modulations.

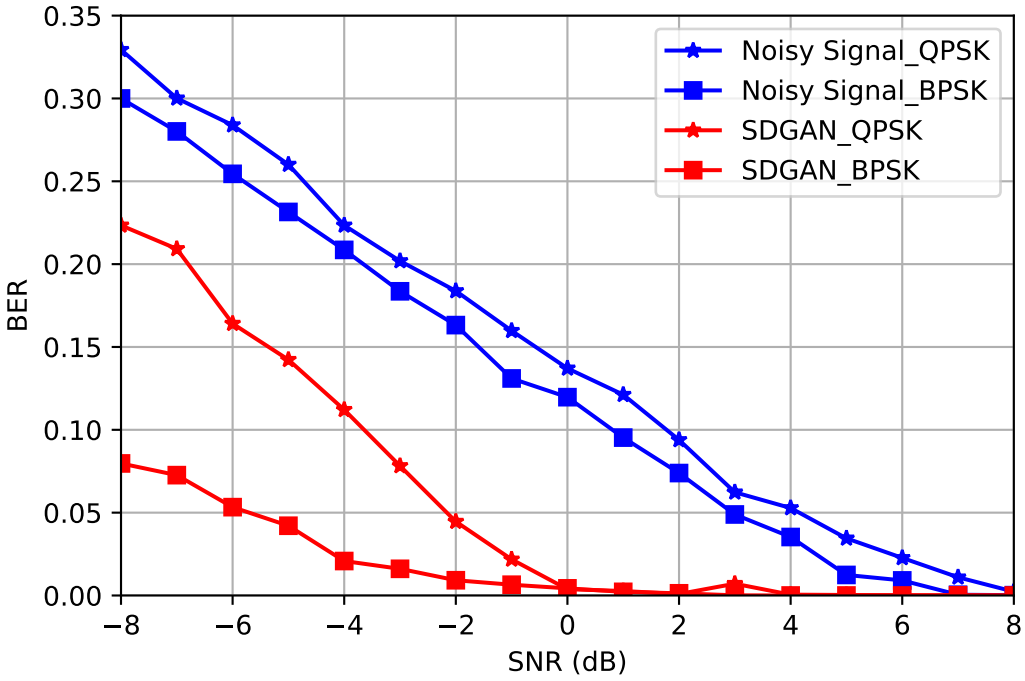

Fig. 19. The denoising performance of our SDGAN on noisy signals with BPSK and QPSK at different SNR levels.

The two red lines with different shapes indicate the denoising results from SDGAN. We can see that SDGAN is still effective on noisy signals with different modulations at all SNR levels. Moreover, SDGAN can reduce the BERs of noisy signals to nearly zero at SNR levels from 0 to 8. All these simulation results verify the effectiveness and potential of the cGAN-based solution to wireless signal denoising.

3.6 Conclusion

In this chapter, we have formulated the wireless signal denoising process as an adversarial learning problem and presented a novel approach based on conditional generative adversarial networks. We designed an applicable loss function and the architecture of networks to make the networks capable of estimating and learning the

features of the wireless signals. To evaluate the effectiveness of our proposed method, we conducted denoising simulations on noisy signals with four different modulations at different SNR levels. The results demonstrated that our proposed method can complete the signal denoising task in a learning manner and also achieved better performance compared with two traditional signal denoising algorithms. The network architecture will be improved further to unlock the full potential and capability of GAN for wireless signal denoising.

CHAPTER 4

DL-BASED SIGNAL MODULATION RECOGNITION

4.1 Introduction

Over the past few decades, there has been a significant surge in the advancement of wireless communication systems, virtually impacting every sector of our daily lives [47], [48]. Wireless transmission has diversified exponentially to meet a wide range of application scenarios [27]. To transmit data efficiently, various modulation schemes are adopted to modulate signals before transmission. Automatic modulation recognition refers to the intermediary process of automatically detecting and classifying the modulation schemes of received signals without prior knowledge between signal detection and signal demodulation [49]. The importance of AMR arises from its ability to optimize spectrum utilization, detect interference, enhance wireless security, and enable cognitive radios in wireless communication systems. The rapid and exceptional evolution of wireless communication technology has led to the development of diverse modulation schemes to cater to increasingly complex communication scenarios. Consequently, various AMR techniques have been developed to achieve effective modulation recognition.

Generally, traditional AMR methods fall into two categories: likelihood-based and feature-based [49]. However, the likelihood-based methods suffer from high computational complexity, while feature-based methods cannot achieve optimal recognition. Motivated by the extraordinary development of DL, different DL-based AMR approaches are proposed due to the powerful feature extraction capability of artificial neural networks [49]. In terms of the architecture of networks, DL-based methods

can be mainly classified into two groups: pure and hybrid models. The pure models are constructed by only one type of neural network, e.g., Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN) [50], [51]. By contrast, the hybrid models combine the advantages of CNN and RNN by integrating them into a single network architecture [52] to achieve better performance than the pure models.

On the one hand, the hybrid DL-based AMR models improve recognition performance by stacking CNN and RNN layers together into a new network structure. On the other hand, such hybrid AMR models bring a concern, i.e., the two attributes of the signal could not be extracted efficiently. To be specific, the temporal features of the signal could be lost after several CNN layers. Similarly, the spatial feature might be degraded after stacking RNN layers. To overcome the weakness of these hybrid AMR models, it is desirable to develop a new model that can extract spatial and temporal features effectively and independently. This chapter proposes a novel and parallel neural network architecture that extracts the spatial and temporal features in two parallel routes. The first route is designed to learn the spatial feature map of the modulated signal by applying CNN layers, while the second route aims to construct the temporal feature map via LSTM (Long Short-Term Memory) layers. Afterward, these two feature maps will be concatenated to identify the modulation type of the modulated signal. Finally, to verify the effectiveness and evaluate the performance of our proposed method, signals with the 11 most representative modulation schemes in modern wireless communications are adopted for simulations. The main contributions of our study are summarized as follows.

- To overcome the weakness of the hybrid DL-based AMR methods, we propose a new neural network architecture consisting of two parallel routes to extract the spatial and temporal characteristics independently. In this way, the two

extraction routes do not interfere with each other.

- A CNN-based network dedicated to spatial feature extraction is designed, and an LSTM-based network is built for temporal feature learning. The proposed AMR method is an end-to-end model that learns the features of modulated signals by CNN and LSTM layers and outputs the predicted label without any manually designed feature extraction process.

- Extensive recognition simulations of our proposed method are conducted with various modulated signals at different SNR levels. We also compare it to other typical DL-based methods. The simulation results show that our method outperforms others in terms of recognition accuracy.

The rest of this chapter is organized as follows. Section 4.2 introduces the related work of both traditional and DL-based AMR methods. Section 4.3 formulates the AMR problem as a multi-class classification and proposes the architecture of the parallel architecture-based neural network for signal modulation recognition. The adopted training dataset and the training procedure are discussed in section 4.4. In section 4.5, we test the performance of our proposed method on different types of modulated signals with comparisons to other approaches. Section 4.6 concludes and sheds light on future studies.

4.2 Related Work

4.2.1 Traditional AMR methods

Traditional AMR methods are typically divided into two categories: likelihood-based and feature-based. In likelihood-based AMR approaches, AMR is perceived as a multi-hypothesis test and the likelihood function of a received signal is compared

with a threshold under the assumption of the known probability density function of the signal [53]. Likelihood-based AMR can achieve optimal performance in the sense of Bayesian estimation but at the cost of high computational complexity or the requirement for prior knowledge. In contrast, feature-based AMR approaches can provide sub-optimal performance but with low computational complexity. Feature extraction and classification are two major steps in feature-based AMR methods. First, instantaneous features and/or statistical features are extracted from the received signals, and then decision-making methods are employed to classify the received signals based on the extracted features. However, extensive domain knowledge and engineering expertise are required to design the handcrafted features in feature-based AMR methods.

4.2.2 DL-based AMR methods

Over the past decade, remarkable advancements have been made through the utilization of DL techniques in tackling various applications [34]. It is known that DL is capable of solving applications that are challenging for conventional methods, including computer vision and natural language processing. Inspired by its exceptional performance, DL has also been applied extensively in the AMR field. It is worth noting that neural networks can automatically extract complex features without manually designed features. AMR is defined as a multi-class classification problem in DL implementation. DL-based AMR models can typically provide better classification performance than traditional approaches.

One of the earliest DL studies in the AMR field was published in [54]. This paper developed a CNN-based DL model to extract spatial features of modulated signals for modulation classification. The simulation results demonstrated that the proposed CNN-based model provides higher accuracy compared to traditional methods. In [51],

an RNN-based model has been developed, which showed the capability of exploiting the temporal features of received modulated signals. In [55], an adopted CNN and RNN model classifies six types of signal modulations under different channel conditions, such as Additive White Gaussian Noise (AWGN) and Rayleigh fading. Simulation results verified that DL-AMR methods outperform traditional algorithms under two fading channels. However, the above DL-based AMR methods have overlooked the feature interaction because these methods rely only on the spatial or temporal features of modulated signals.

To address this problem, the study [56] introduced a Convolutional Long Short-term Deep Neural Network (CLDNN [57]) that includes four convolutional layers and one LSTM layer to achieve an accuracy of approximately 88.5 % at high SNR level. Furthermore, [58] proposed a hybrid architecture that combines the advantages of CNN and LSTM to extract the spatial and temporal features of signals, respectively, by using two signal representations. The experimental results showed that the performance of DL-based AMR can be improved significantly by utilizing both modulated signal attributes. The recent hybrid neural network architecture for AMR was proposed in [59]. This model is built upon stacking Gated Recurrent Units (GRUs) and CNN layers.

4.2.3 Convolutional Neural Networks

CNN represent a pivotal advancement in machine learning and artificial intelligence and have demonstrated superior performance across a wide range of applications, from object detection and facial recognition to image analysis and autonomous driving [60]. CNN leverages a specialized architecture designed to automatically and adaptively learn spatial hierarchies of features from input data [61]. Specifically, CNN applies a series of learnable filters to the input, producing feature maps highlighting

various aspects of the data, such as edges, textures, and patterns [62]. In short, CNN excels in spatial feature extraction and has succeeded tremendously in classification and detection tasks [63].

4.2.4 Long Short Term Memory

RNN are commonly applied to learn persistent features of sequence data. LSTM is a particular type of RNN that is efficient in learning long-term dependencies and is heavily used for natural language processing and signal processing [64]. The major components in an LSTM cell are three gates, namely the input gate, the forget gate, and the output gate, which are used to control how the information propagates in the network. The gating mechanism allows LSTM cells to memorize information for extended periods, thus realizing continuous feature learning. The key equations of an LSTM cell are listed below:

$$\begin{aligned} i_t &= \sigma(x_t U^i + h_{t-1} W^i + b_i) \\ f_t &= \sigma(x_t U^f + h_{t-1} W^f + b_f) \\ o_t &= \sigma(x_t U^o + h_{t-1} W^o + b_o) \\ \hat{C}_t &= tanh(x_t U^g + h_{t-1} W^g + b_c) \\ C_t &= f_t \odot C_{t-1} + i_t \odot \hat{C}_t \\ h_t &= o_t \odot tanh(C_t U^o) \end{aligned} \tag{4.1}$$

where x_t is input vector, i_t is input gate vector, f_t is forget gate vector, o_t is output gate vector, c_t is cell state vector, h_t is hidden state vector, b_i, b_f, b_o, b_c are bias vectors, U, W is parameter matrices, and σ, $tanh$ are activation functions. $\odot$ denotes the Hadamard product for the element-wise product of matrices.

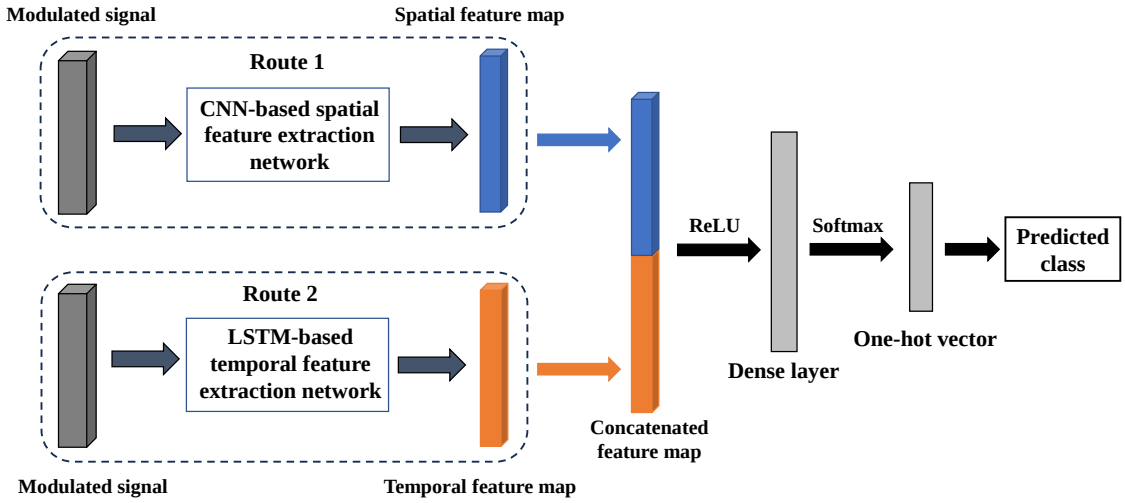

Fig. 20. The proposed parallel architecture includes two parallel routes to extract the signal features in AMR. The two routes extract the spatial and temporal features using CNN-based and LSTM-based networks, respectively.

4.3 System Model

In this section, the AMR problem is first formulated as a multi-class classification task. Then, we propose a parallel feature extraction neural network architecture to solve the ARM problem.

4.3.1 Problem Statement

In wireless communication, the received signal is often represented as a waveform or mathematical expression that carries the transmitted information. The expression essentially depends on various factors, including the modulation scheme used, channel conditions, noise, and interference. For simplicity, the received signal can be modeled as:

$$y(t) = x(t) * h(t) + n(t), \tag{4.2}$$

where $x(t)$ represents the modulated signal with a certain modulation method from the transmitter at time t. $h(t)$ is the channel impulse response of the transmitted wireless channel. $n(t)$ denotes the AWGN. $y(t)$ denotes the received signal, which generally consists of two components: In-phase and Quadrature (I/Q) for flexible hardware design and efficient mathematical operations [65].

The goal of the AMR task is to identify the modulation scheme of $x(t)$ based on the received signal $y(t)$. In the DL field, this task can be described as a multi-classes classification problem, which can be formulated as the two equations below:

$$\mathcal{P}\big(x(t) \in C_i \big| y(t)\big) = \mathcal{F}\big(y(t); \Theta\big), \tag{4.3}$$

$$\hat{C}_i = \arg\max \mathcal{P}\big(x(t) \in C_i \big| y(t)\big), \tag{4.4}$$

where $\mathcal{F}(\cdot)$ is the model function with parameters Θ. The output of the model is the conditional probability $\mathcal{P}$ that the modulated signal $x(t)$ belongs to the i^{th} modulation scheme C_i based on the observation of the received signal $y(t)$. The Equation 4.4 outputs the predicted label $\hat{C}_i$ of the modulation scheme of $x(t)$ by maximizing $\mathcal{P}$.

Table 1. The parameters of CNN layers

CNN layer / Parameters	#1	#2	#3
Filter size	1×3	2×3	1×3
Number of Filters	256	256	80
Padding	(0×2)	(0×2)	(0×2)
Stride	1	1	1
Dropout rate	0.25	0.25	0.5

4.3.2 Model Architecture

In this subsection, we propose the parallel neural network architecture, as illustrated in Figure 20, to solve the AMR problem. The architecture is concluded

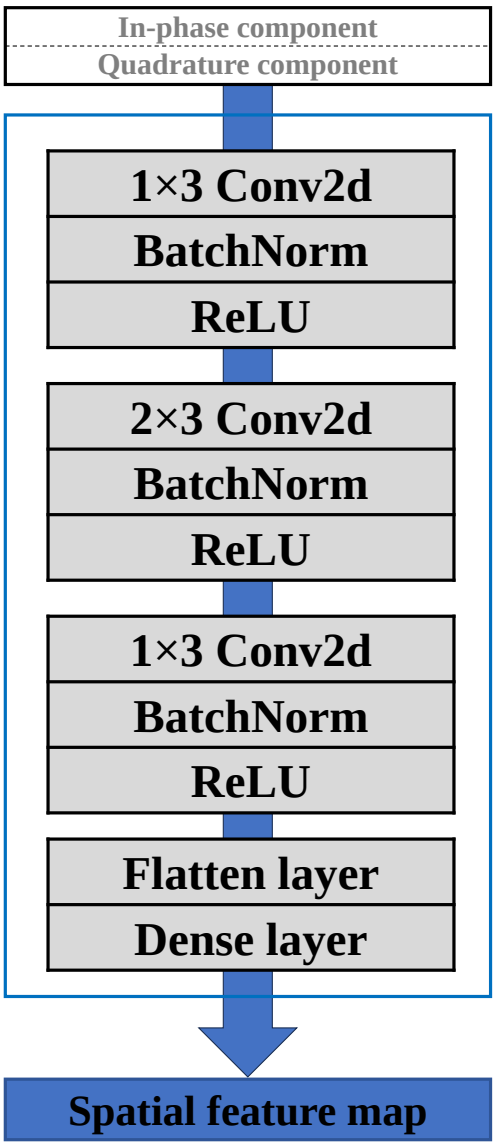

Fig. 21. The CNN-based neural network for spatial feature extraction of the modulated signal in Route 1.

empirically by preliminary simulations. It has two parallel routes to extract the two types of features independently.

In the first route, the modulated signal in I/Q format is fed into a CNN-based network to extract the spatial feature, as shown in Figure 22 (a). By conducting

extensive comparative experiments, we selected the architecture, which consists of three convolutional layers followed by a flattened layer and a dense layer. The number of filters used in the three 2-dimensional CNN layers is 256, 256, and 80 with sizes 1×3, 2×3, and 1×3, respectively. The stride of each CNN layer is set to 1, and the zero-padding operation is used before each CNN layer. The Batch Normalization (BatchNorm) [46] and Rectified Linear Unit (ReLU) activation function [66] are applied after each CNN layer. In addition, to avoid overfitting and improve the generalization of our model, the dropout technique [67] is adopted, and the dropout rate p is 0.25 for the first two CNN layers. For the third CNN layer and the last two layers, p is set to 0.5. Table 1 presents the settings of CNN layers more intuitively. The final output of this model is a 128×1 dimensional spatial feature map.

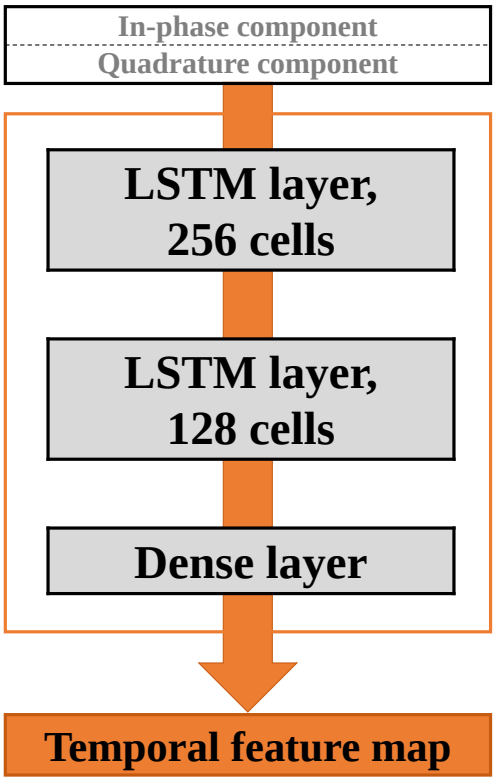

Fig. 22. The LSTM-based neural network for temporal feature extraction of the modulated signal in Route 2.

In Route 2, an LSTM-based neural network is adopted for exploiting the temporal feature of the signal, as exhibited in Figure 22 (b). Essentially, the intuition behind applying LSTM for temporal feature extraction is that the modulated signal has continuous temporal attributes and LSTM can learn these temporal features effectively. This network consists of two LSTM layers followed by a dense layer. The first LSTM layer has 256 cells, while the second LSTM includes 128 cells. This model takes the modulated signal with I/Q representation as the input, and the dense layer outputs the temporal feature map that is a 128×1 dimensional vector. Furthermore, the dropout method with $p = 0.5$ is applied on both LSTMs and dense layers.

After extracting spatial and temporal features in two parallel routes, both feature maps are concatenated to a 256×1 vector which is passed to a dense layer with the ReLU activation function. Finally, the output layer generates the probabilities of each modulated scheme using the Softmax function. The predicted class corresponds to the modulated scheme with the highest probability.

4.4 Implementation

In this section, we first describe the dataset used for simulations and then discuss the training details.

4.4.1 Dataset Description

To train our model and evaluate the recognition performance of our proposed approach, the radio ML dataset (RML2016.10a) [68] is adopted for simulations, which is publicly available and widely used in AMR research as the benchmark. The dataset is synthetically generated by utilizing GNU Radio with practical modulation parameters. It has 220,000 signal samples and each sample is associated with one modulation at a specific SNR. A sample is composed of a 256-dimensional vector, including 128

in-phase and 128 quadrature components. The data samples are generated at 20 different SNR levels from -20 dB and 18 dB with an interval of 2 dB. There are 11 different modulations, including BPSK, QPSK, 8PSK, QAM16, QAM64, CPFSK, GFSK, PAM4, WBFM, AM-SSB, and AM-DSB, which are prevalent in modern wireless communication systems. Notably, RML2016.10a also considers realistic channel imperfections, e.g., channel frequency offset, sample rate offset, and noise, which result in challenging recognition. More specific generation and parameters of this dataset are available at [68].

4.4.2 Model Training

The RML2016.10a dataset is split into 60% training set, 20% validation set, and 20% testing set. In the training phase, the adopted loss function is categorical cross-entropy loss expressed as

$$\mathcal{L}(\Theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij} log(p_{ij}), \tag{4.5}$$

where N indicates the mini-batch size set to 400, and C is the number of modulation methods. For each signal sample i, we have the predicted probabilities p_{i1}, p_{i2}, ..., p_{iC}, which represent the estimated probability of the signal belonging to each modulation scheme. y_{ij} denotes the true modulation class labels. The Adam optimizer with the initial learning rate of 0.001 is used to optimize Θ of our model for loss function minimization. The learning rate is halved if the validation loss does not decrease within five (5) epochs and the training phase is stopped if the validation loss remains stable within 50 epochs. The training process and the simulations in Section 4.5 are conducted on a laptop with an Nvidia GTX 1080Ti GPU and Keras with Tensorflow as the backend.

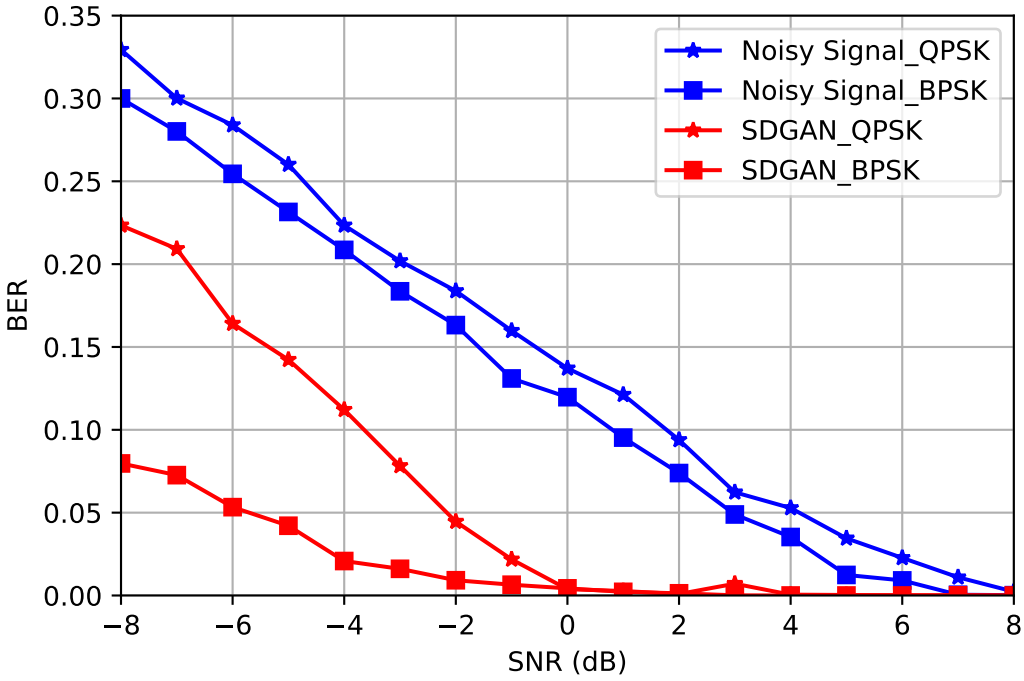

Fig. 23. Recognition accuracy comparison of our proposed method with three DL-based AMR methods on the RadioML dataset.

4.5 Simulations

After the training procedure, our proposed model can classify the modulation schemes of signals. To evaluate the performance of our method, we have conducted extensive simulations. In this section, we first introduce the AMR models chosen for comparisons and then discuss the simulation results.

4.5.1 AMR Models for Comparisons

To ensure fair and diverse comparisons, three typical DL-based AMR models are adopted. The first AMR model is based on two CNN layers (referred to as CNN-AMR) which have been widely accepted as the baseline model in AMR research [68]. The second AMR model is an LSTM-based model (referred to as LSTM-AMR) [65],

and the last AMR model is CLDNN [56], which is composed of four CNN layers and one LSTM layer. Specifically, the CNN-AMR and LSTM-AMR models are pure models, while CLDNN is a hybrid model.

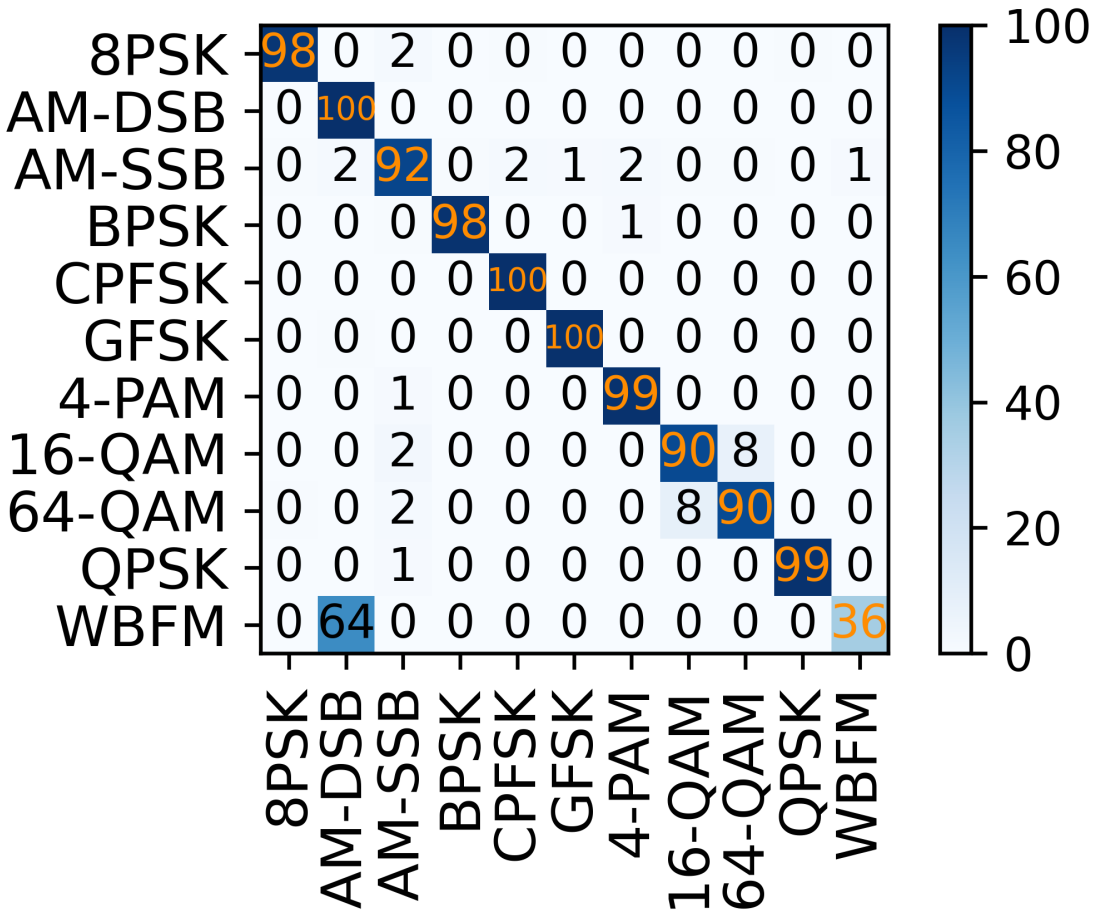

Fig. 24. Confusion matrix of our proposed method on 100 random RadioML signal samples at 6dB SNR.

4.5.2 Results and Discussion

The recognition accuracy of all four models on the testing set is shown in Figure 23. The blue, red, black, and purple lines represent the results of our method, CLDNN, CNN-AMR, and LSTM-AMR, respectively. Our method provides the best

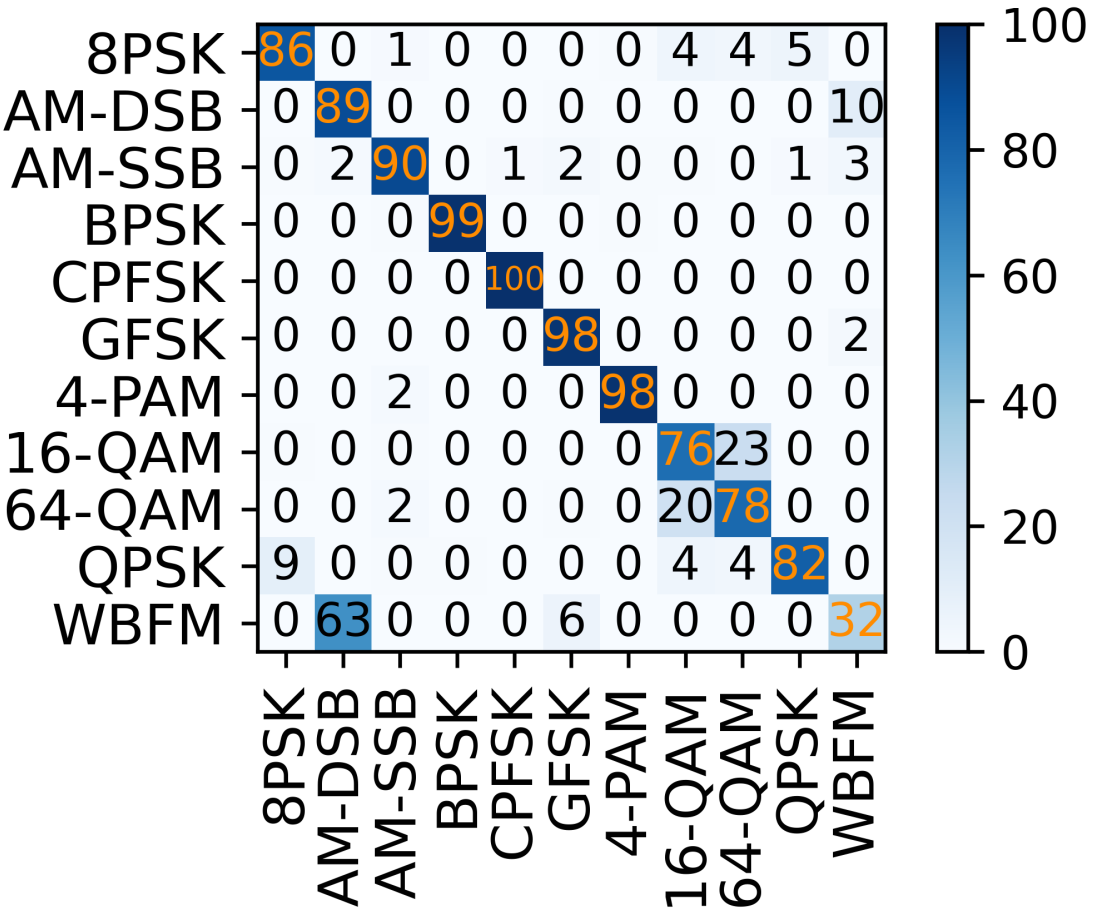

Fig. 25. Confusion matrix of our proposed method on 100 random RadioML signal samples at 0dB SNR.

average accuracy across all SNR levels. It recognizes modulated signals at an accuracy of over 90% in SNR ranging from 0dB to 18dB, demonstrating that our method can combine the advantages of CNN and LSTM and outperform them. Although CLDNN is a hybrid model consisting of the CNN layer and LSTM layer, CLDNN cannot outperform LSTM-AMR at SNR higher than 0dB. It verifies that our method with two parallel routes can extract and learn the features of modulated signals more efficiently than CLDNN.

To describe the recognition performance of our method in detail, we also provide

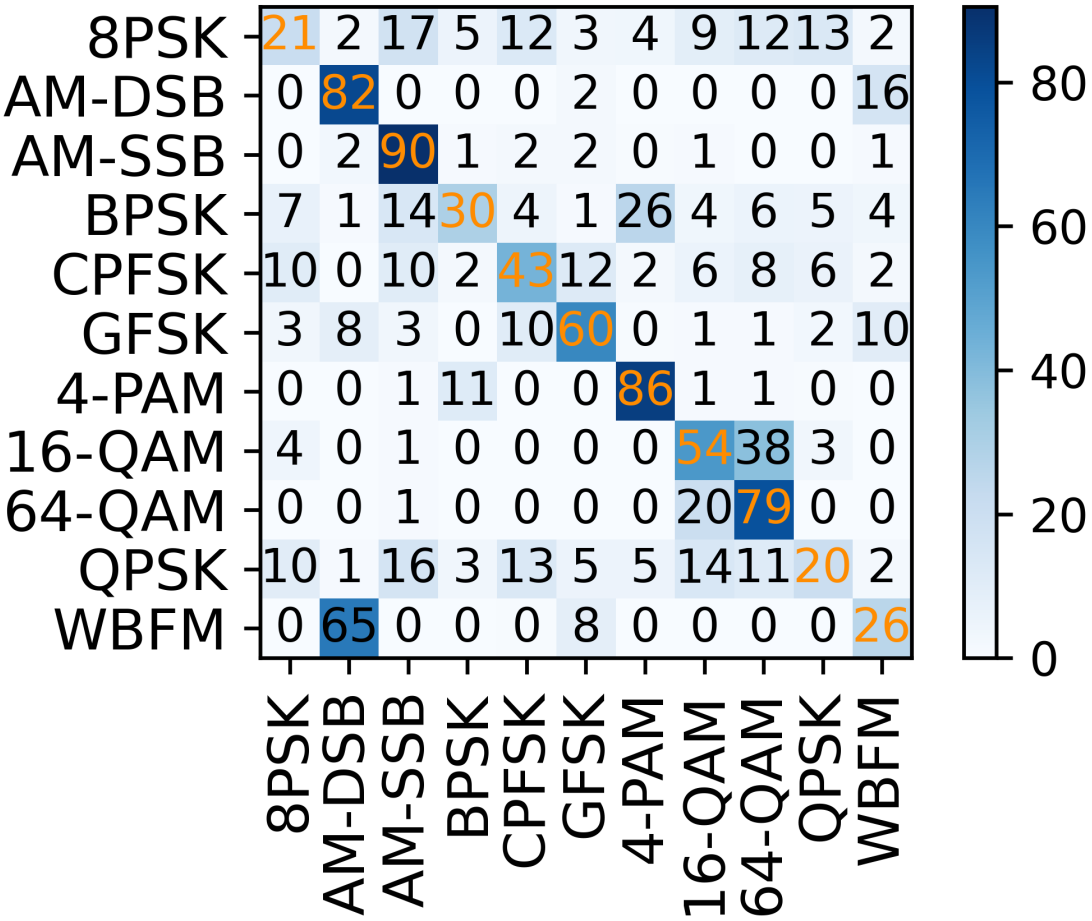

Fig. 26. Confusion matrix of our proposed method on 100 random RadioML signal samples at -6dB SNR.

the confusion matrices at different SNR levels. The vertical axis of each matrix represents the true label, while the horizontal axis represents the predicted label. As shown in Figure 24, almost all modulation schemes can be recognized accurately, excluding WBFM which could be classified as AM-DSB, when SNR is 6dB. The reason behind this is that WBFM and AM-DSB include some similar attributes, e.g., interludes and off-time [58]. Similarly, it becomes more difficult to separate QAM16 and QAM 64 signals at the SNR of 0dB, as presented in Figure 25. This is

caused by the factor that QAM16 is the subset of QAM64 and they have overlapping constellation points.

Figure 26 shows that the overall misclassifications increase with the decrease in SNR. Nevertheless, AM-DSB, AM-SSB, and PAM4 can still be identified successfully with relatively acceptable accuracy. The discussions above demonstrate that our proposed method can recognize most modulated signals with SNR higher than 0 dB with higher accuracy.

4.6 Conclusion

In this chapter, we have formulated the recognition of wireless modulated signals as a multi-class classification problem in DL and proposed a parallel feature extraction neural network architecture. Compared to most current hybrid AMR models that extract both features step by step via CNN-RNN-like architectures, our model adopts a parallel manner utilizing both spatial and temporal features of the signal extracted by a CNN-based and LSTM-based network, respectively. Therefore, the two feature extraction processes can be performed independently and will not interfere with each other. Extensive simulations are conducted to demonstrate that our proposed parallel model can not only complete the AMR task but also outperform pure CNN or RNN models as well as a typical hybrid model. In future work, the potential research directions can be summarized in the following aspects. First, we will evaluate our model by using different modulated signal datasets to investigate the generalization of our model. Second, the recognition accuracy of modulated signal with SNR lower than 0dB is unsatisfactory. Hence, it is necessary to improve the model by integrating signal enhancement algorithms into AMR. Last, DL-based AMR models usually require a large number of labeled modulated signal samples during the training, which is not feasible in realistic wireless communications. Therefore, the exploration of utilizing

smaller data sizes for training models without compromising performance will be considered.

CHAPTER 5

AI-POWERED WIRELESS COMMUNICATIONS UNDER ADVERSARIAL ATTACKS

5.1 Introduction

In recent years, NextG or 5G and beyond has been paying more attention to academia and industry, along with high demand and new ways of communication needed from consumers. According to the report released by the International Telecommunication Union (ITU), the mobile data traffic based on NextG will constantly increase yearly and reach thousands of exabytes in 2030 [69]. NextG networks aim to connect billions of devices, systems, and applications to meet high data rates and low latency requirements to support new applications. Fortunately, NextG networks can satisfy these requirements and support these applications with advanced communication, computing, and AI technologies [70]. AI is an extraordinary contributor among them in NextG networks [4]. Recent studies have demonstrated that AI and Machine Learning (ML) based solutions outperform across all aspects of next-generation networks, from the physical layer to the application layer [71, 72, 73]. A conceptual model for 6G has been presented in [74], which emphasizes the importance of AI/ML-powered solutions at each layer of the model to meet the requirements of next-generation wireless networks in terms of latency, power allocation, privacy, security, and more.

Inspired by the tremendous achievements of AI, AI-powered models have also been applied to IRS-driven wireless communication in NextG wireless networks to improve performance [75, 76, 27, 77]and many AI-based AMR automatic modulation

recognition methods have been proposed. However, the security threats (e.g., model poisoning or adversarial machine learning attacks) and mitigation methods (e.g., adversarial training or defensive distillation) have not been investigated in AI-powered applications of NextG networks due to being new, complicated, and multi-disciplinary topics (e.g., next-generation communications, cybersecurity, and AI) [4, 78].

To fill this gap, this chapter will focus on an AI-powered IRS model in 5G and beyond networks and their vulnerabilities, which have received limited attention. The vulnerabilities of an AI-powered model are among the top security concerns and deserve a thorough investigation. For example, a trained AI model might be manipulated by adding noise to the data, i.e., targeted and non-targeted adversarial attacks. The adversarial attacks are generated by adding a perturbation to a legitimate data point, i.e., an adversarial example, to fool the AI-powered models. The mitigation methods will be provided to improve the robustness of the model as well. In addition, the robustness of the deep learning-based AMR method proposed in Chapter 4 will also be analyzed.

The remainder of this chapter is organized as follows. Section 5.2 provides the background information about the common adversarial attacks. Section 5.3 evaluates the performance of an AI-powered IRS system under adversarial attacks. Section 5.4 investigates an AI-based AMR model under adversarial attacks in both SISO (Single-Input and Single-Output) and MIMO (Multiple-Input and Multiple-Output) scenarios. Section 5.5 concludes this chapter.

5.2 Related Work

ML-based models are trained to automatically learn the underlying patterns and correlations in data using algorithms. Once an ML-based model is trained, it can be used to predict the patterns in new data. The accuracy of the trained model is

essential to achieving a high performance, which can also be called a generalization. However, the trained model can be manipulated by targeted and non-targeted adversarial ML attacks to fool the models. There are various kinds of adversarial ML attacks, such as evasion attacks, data poisoning attacks, and model inversion attacks.

Liu et al. [79] conducted a comprehensive survey on adversarial ML for wireless and mobile systems. Adversarial ML approaches can be used to generate and detect adversarial samples, which are samples that have been specifically designed to deceive a machine-learning model. These samples can fool a model into misclassifying an input and can be used to exploit certain blind spots in image classifiers. The article reviews the state-of-the-art adversarial ML approaches to generating and detecting adversarial samples. It provides detailed discussions highlighting the open issues and challenges these approaches face.

An evasion attack aims to cause the ML-based models to misclassify the adversarial examples as legitimate data points, i.e., targeted and non-targeted evasion attacks. Targeted attacks aim to force the models to classify the adversarial example as a specific target class. Non-targeted attacks aim to push the models to classify the adversarial example as any class other than the ground truth. Data poisoning aims to generate malicious data points to train the ML-based models to find the desired output. It can be applied to the training data, which causes the ML-based models to produce the desired outcome. Model inversion aims to generate new data points close to the original data points to find the sensitive information of the specific data points.

These adversarial attack types are given as follows.

5.2.1 Fast Gradient Sign Method (FGSM)

FGSM is one of the most popular and straightforward approaches to constructing adversarial examples. It is called one-step gradient-based attack. It is used to compute the gradient of the loss function with respect to the input, $\mathbf{x}$, and then the attacker creates the adversarial example by adding the sign of the gradient to the input data. It was first introduced by Goodfellow et al. in 2014 [80]. The gradient sign is computed using the backpropagation algorithm. The steps are summarized as follows:

- Compute the gradient of loss function, $\nabla_{\mathbf{x}}\ell(\mathbf{x}, \mathbf{y})$
- Add the gradient to the input data, $\mathbf{x}_{adv} = \mathbf{x} + \epsilon \times sign(\nabla_{\mathbf{x}}\ell)$,

where ϵ is the budget. FGSM attack has been used in [81] to attack models.

5.2.2 Basic Iterative Method (BIM)

BIM is one of the most popular attacks called an iterative gradient-based attack. This attack is derived from the FGSM attack. It is used to compute the gradient of the loss function with respect to the input, $\mathbf{x}$, and then the attacker creates the adversarial example by adding the sign of the gradient to the input data. The gradient sign is computed using the backpropagation algorithm. The steps are summarized as follows:

- Initialize the adversarial example as $\mathbf{x}_{adv} = \mathbf{x}$
- Iterate i times, where $i = 0, 1, 2, 3, ..., N$
  - Compute the gradient of loss function, $\nabla_{\mathbf{x}}\ell(\mathbf{x}_{adv}, \mathbf{y})$

- Add the gradient to the input data,

  $\mathbf{x}_{adv} = \mathbf{x}_{adv} + \epsilon \times sign(\nabla_{\mathbf{x}} \ell)$,

where ϵ is the budget, and N is the number of iterations. The BIM attack has been used in [81] to attack models.

5.2.3 Projected Gradient Descent (PGD)

PGD is one of the most popular and powerful attacks [82]. It is used to compute the gradient of the loss function with respect to the input, $\mathbf{x}$, and then the attacker creates the adversarial example by adding the sign of the gradient to the input data. The gradient sign is computed using the backpropagation algorithm. The steps are summarized as follows:

- Initialize the adversarial example as $\mathbf{x}_{adv} = \mathbf{x}$
- Iterate i times, where $i = 0, 1, 2, 3, ..., N$
  - Compute the gradient of loss function, $\nabla_{\mathbf{x}} \ell(\mathbf{x}_{adv}, \mathbf{y})$
  - Add random noise to the gradient,

    $\hat{\nabla}_{\mathbf{x}} \ell(\mathbf{x}_{adv}, \mathbf{y}) = \nabla_{\mathbf{x}} \ell(\mathbf{x}_{adv}, \mathbf{y}) + \mathcal{U}(\epsilon)$
  - Add the gradient to the input data,

    $\mathbf{x}_{adv} = \mathbf{x}_{adv} + \alpha \times sign(\hat{\nabla}_{\mathbf{x}} \ell)$,

where ϵ is the budget, N is the number of iterations, and α is the step size. PGD can generate stronger attacks than FGSM and BIM.

5.2.4 Momentum Iterative Method (MIM)

MIM is a variant of the BIM adversarial attack, introducing momentum and integrating it into iterative attacks [83]. It is used to compute the gradient of the loss

function with respect to the input, $\mathbf{x}$, and then the attacker creates the adversarial example by adding the sign of the gradient to the input data. The gradient sign is computed using the backpropagation algorithm. The steps are summarized as follows:

- Initialize the adversarial example $\mathbf{x}_{adv} = \mathbf{x}$ and the momentum, $\mu = 0$
- Iterate i times, where $i = 0, 1, 2, 3, ..., N$
  - Compute the gradient of loss function,

    $\nabla_{\mathbf{x}} \ell(\mathbf{x}_{adv}, \mathbf{y})$
  - Update the momentum,

    $\mu = \mu + \frac{\eta}{\epsilon} \times \nabla_{\mathbf{x}} \ell(\mathbf{x}_{adv}, \mathbf{y})$
  - Add random noise to the gradient,

    $\hat{\nabla}_{\mathbf{x}} \ell(\mathbf{x}_{adv}, \mathbf{y}) = \nabla_{\mathbf{x}} \ell(\mathbf{x}_{adv}, \mathbf{y}) + \mathcal{U}(\epsilon)$
  - Add the gradient to the input data,

    $\mathbf{x}_{adv} = \mathbf{x}_{adv} + \alpha \times sign(\hat{\nabla}_{\mathbf{x}} \ell)$,

where ϵ is the budget, N is the number of iterations, η is the momentum rate, and α is the step size.

Note that there are many types of adversarial attacks and defenses. The existing defenses and adversarial attacks for images can be applied to attack and defend on intelligent reflecting surfaces and other fields [84, 85, 86, 87]. The cleverly designed adversarial examples can fool the deep neural networks with high success rates on the test images. The adversarial examples can also be transferred from one model to another model. In our experiments, we generated the adversarial inputs with untargeted attacks.

5.3 AI-powered IRS Communication System Under Adversarial Attacks

5.3.1 System Model Overview

It is challenging to acquire channel knowledge to estimate the Tx-IRS and IRS-Rx channel link in an IRS-assisted system since all the reflecting elements are expected to be nearly passive. Authors in [88] propose a new IRS architecture where all elements are passive except for a few active sensing elements and adopt a deep learning technique to assist the IRS in addressing this problem. Specifically, the transmitter and receiver first transmit two orthogonal uplink pilots to the active elements of IRS, and the active elements estimate the sampled channel vectors to construct the multi-path signature as the environment descriptors. Motivated by recent advances in deep learning, that paper proposes to train a neural network to observe the environment descriptors and predict the achievable rate with each IRS interaction vector. Based on the predictions, the IRS interaction vector corresponding to the highest predicted achievable rate will be used to reflect the transmitted data from the transmitter to the receiver. In this study, we refer to the model above as the AI-powered IRS model and will investigate and examine the vulnerability of this model and apply the defensive distillation mitigation method. As we briefly discussed above, a neural network is designed to map the observed environment descriptors to the predicted achievable rate in the AI-powered IRS model. This subsection below introduces the neural network architecture, dataset, and training details.

5.3.1.1 Neural Network Architecture

The input of the neural network model is defined as a stack of environment descriptors (i.e., uplink pilot signals) received from both transmitter and receiver. Since the training process is designed to build function mapping descriptors for reflection

vectors, the output target of the neural network is to be a set of predictions on the achievable rates of every possible reflection beamforming vector. The neural network is built as a Multi-Layer Perceptron (MLP) network, which is well-demonstrated as an effective universal approximator. The MLP is adopted to establish the connection between the environment descriptors and the predicted achievable rates using reflection beamforming vectors, as shown in Figure 27. The MLP is composed of four fully connected layers. ReLU activation function is adopted, and a dropout layer is added after the activation function for every layer except for the last layer. The MLP consists of the following dimensions: M(Input), [M, $2M$](Layer1), [$2M$, $4M$](Layer2), [$4M$, $4M$](Layer3), [$4M$, M](Layer4), where M is the number of the antenna elements on IRS.

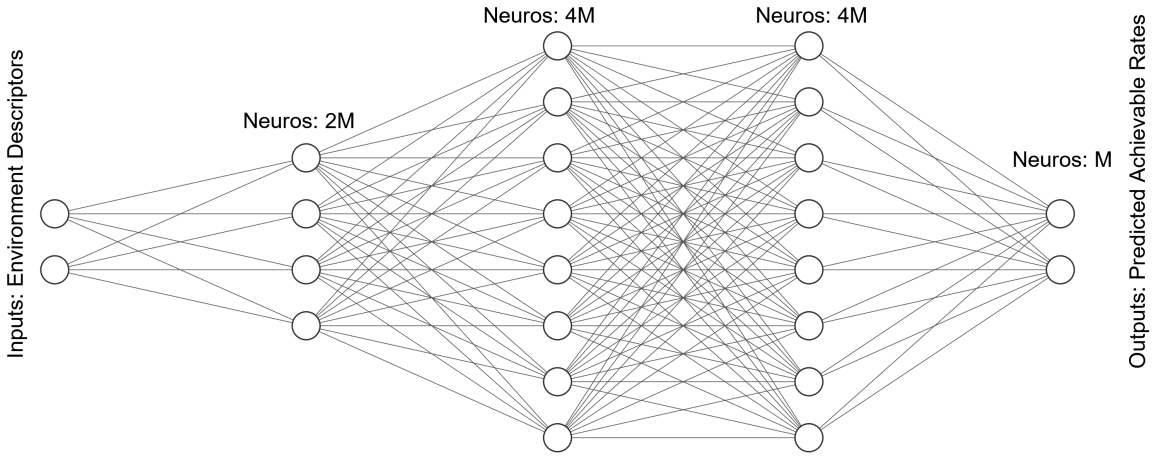

Fig. 27. The adopted neural network architecture is composed of four fully connected layers. The number of the neurons of the four layers is ($2M$, $4M$, $4M$, M), where M indicates the number of the antenna elements on IRS.

5.3.1.2 Dataset Preparation

To examine the performance of the AI-powered IRS model, a publicly available ray-tracing-based DeepMIMO dataset [89] is adopted to generate the training dataset.

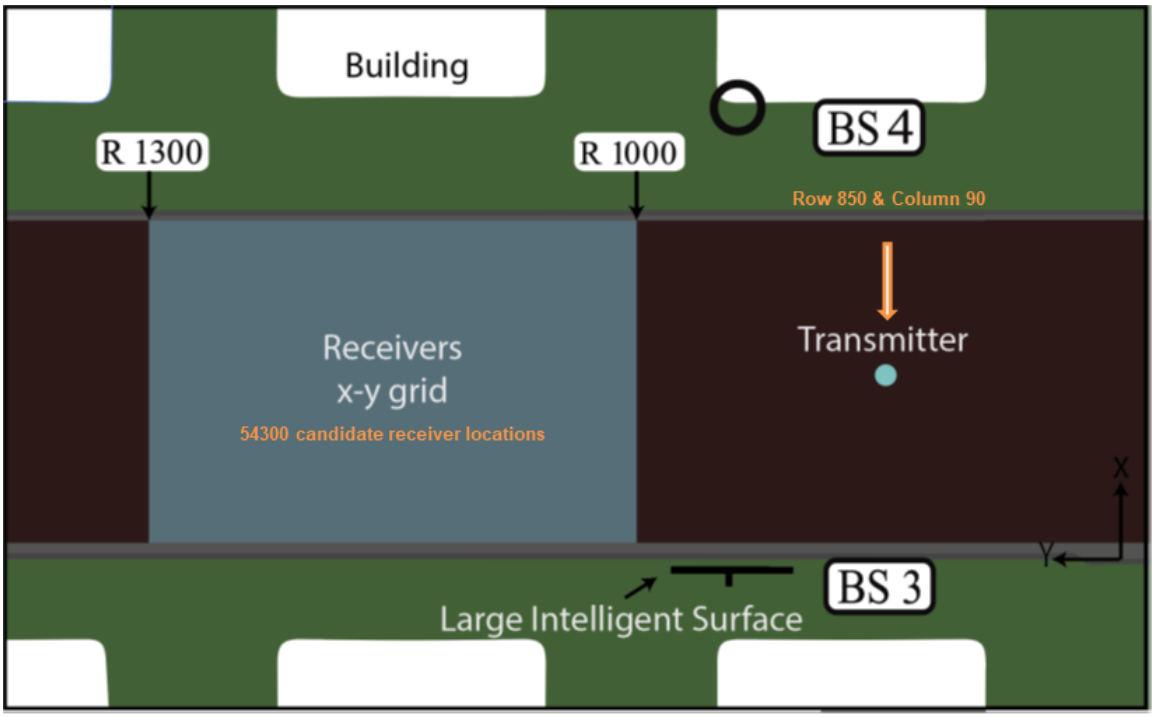

Fig. 28. The adopted ray-tracing scenario where the Large Intelligent Surface (i.e., IRS) is deployed to reflect the signal from the fixed transmitter to the candidate receivers.

The DeepMIMO dataset is a parameterized dataset designed for constructing the MIMO channels based on ray-tracing data obtained from the accurate ray-tracing scenario simulation. Similar to the simulation setup in [88], the outdoor ray-tracing scenario 'O1' is selected as shown in Figure 28. Base Station 3 (BS 3) is set as an IRS, which is equipped with a UPA (Uniform Planar Array) with 32×32 ($M = 1024$) or 64×64 ($M = 4096$) antennas at the mmWave 28GHz setup. The transmitter is fixed in row R850 and column 90, and the candidate receiver locations are in the uniform x-y grid from row R1000 to R1300 (i.e., 54300 points). Both the transmitter and receiver are assumed to have a single antenna. The antenna elements have a gain of 3dBi and a transmit power of 35dBm. Table 2 summarizes the adopted parameters in the DeepMIMO dataset. The generated DeepMIMO dataset includes the channel

vectors between the IRS and the transmitter/receiver of the specified subcarriers for all candidate user locations in the x-y grid. With these channel vectors and given the randomly selected active elements, we can construct the sampled active channel vectors between the active elements of IRS and the transmitter/receiver. Note that the channel vectors depend on the various elements of the surrounding environment [88]. Therefore, the sampled active channel vectors (i.e., environment descriptors) can be used to describe the wireless environment and fed into the deep neural networks described earlier.

Table 2. The adopted DeepMIMO dataset parameters

DeepMIMO Dataset Parameter	**Value**
Frequency band	28GHz
Active BSs	3
Number of Antennas	$(M_x, M_y, M_z) \in \{(1, 32, 32); (1, 64, 64)\}$
Active users (receivers)	From row R1000 to R1300
Active transmitter	row R850 column 90
System bandwidth	100MHz
Number of OFDM subcarriers	512
OFDM sampling factor	1
OFDM limit	64
Number of channel paths	1
Antenna spacing	0.5λ

5.3.1.3 Training Details

The training dataset has 54300 data samples since the candidate receiver locations contain 54300 points. The dataset is split into two sets, namely a training set and a testing set with 85% and 15% of the points, respectively. To measure the quality of the predictions and make the predicted achievable rates close to the real achievable rates in the dataset, we define the loss function with Mean-Squared-Error (MSE) between them. In the training process, the batch size is set to 500 samples, and the training epochs are set to 20. The dropout rate is set to 50%, and a L_2 regularization term with the factor of 10^{-4} is added to the loss function. The learning rate decreases by 50% every three epochs, starting at 0.1 with the Stochastic Gradient Descent (SGD) optimizer.

5.3.2 Defensive Distillation

As mentioned previously, in this study, we leverage the defensive distillation mitigation method to improve the robustness of our AI-powered IRS model. Defensive distillation is a method that applies defensive knowledge distillation to train a more robust model [90]. Knowledge distillation was previously introduced by Hinton et al. [91] to compress the knowledge of a large, densely connected neural network (the teacher) into a smaller, sparsely connected neural network (the student). It has been shown that the student could achieve a similar performance as the teacher by mimicking the output of the teacher, and the teacher would be used as a soft label to train the student. Furthermore, the student could be trained to be more resistant to adversarial attacks than the teacher by using the label of the teacher as the label of the student [92].

The architecture of the defensive distillation consists of the following steps:

- **Step 1:** Train a model with cross-entropy loss as the classification task-based model (teacher).
- **Step 2:** Train the same model (teacher) with defensive distillation loss (soft label + cross-entropy) to generate the respective soft label.
- **Step 3:** Train a model with the soft label generated in step 2 as the label (student) to obtain the robust model.

The defensive distillation loss function is defined as

$$\mathcal{L}_D(\theta) = (1-\lambda)\,\mathcal{L}_{CE}(\theta) + \lambda \mathcal{L}_{KL}\left(P_T(y|\theta), P_T(y)\right), \tag{5.1}$$

where $\mathcal{L}_{CE}(\theta)$ and $\mathcal{L}_{KL}\left(P_T(y|\theta), P_T(y)\right)$ denote the cross entropy and Kullback Leibler (KL) divergence losses, respectively. $P_T(y|\theta)$ is the output of the teacher model with parameters θ. $P_T(y)$ is the output of the soft label. λ is a trade-off parameter between cross entropy and KL divergence losses. Algorithm 1 shows the pseudocode.

5.3.3 Performance Metric

This study evaluates the AI-powered IRS model through the Mean Squared Error (MSE) performance metric. MSE scores are utilized to analyze the model vulnerabilities under undefended and defended conditions. The equation regarding the MSE score is given below.

$$MSE = \frac{\sum \left(Y_t - \hat{Y}_t\right)^2}{n} \tag{5.2}$$

where :

- Y_t : The actual t^{th} instance,

Algorithm 1 Training the defensive distillation.

Input: Training data set $\mathcal{D}$, base model M_T, λ, α, ϵ, number of iterations N

Output: Defensive distillation model M_D

Train the base model M_T by minimizing the cross entropy loss $\mathcal{L}_{CE}$ on $\mathcal{D}$

Initialize the defensive distillation model $M_D = M_T$

while $iter < N$ **do**

  Get a batch of samples X and labels Y from $\mathcal{D}$

  Calculate the cross entropy loss $\mathcal{L}_{CE}$ and KL divergence loss $\mathcal{L}_{KL}$ of X

  Calculate the defensive distillation loss $\mathcal{L}_D$ using Eq. 5.1

  Calculate the adversarial samples X_{adv} by FGSM, BIM, MIM and PGD with ϵ

  Calculate the new loss $\mathcal{L}_D^{'}$ with the adversarial samples X_{adv}

  Update the weights of the defensive distillation model M_D by minimizing the new loss $\mathcal{L}_D^{'}$

  $iter \leftarrow iter + 1$

end while

return M_D =0

- $\hat{Y}_t$: The forecasted t^{th} instance,
- n: The total number of instance

MSE score measures the average squared difference between the actual and predicted values. A high MSE score represents a high prediction error.

5.3.4 Experimental Results

This section analyses the results obtained from the experiments related to AI-powered IRS models against adversarial machine learning attacks. Results are repre-

sented in three ways: (1) bar plots showing the impact of each adversarial machine learning attack on the performance of undefended and defended models, i.e., MSE, (2) histogram plots showing the MSE metric values for each attack of defended and undefended models, and (3) the table showing the prediction performance results of defended and undefended models for each adversarial attack. Figure 29-30 show the bar plots, while Figure 31-37 show the histogram plots. Table 3 shows the prediction performance results of the defended and undefended AI-powered IRS models against the attacks.

The trained AI-powered IRS model is implemented using Python 3.7.13 and the TensorFlow 2.8.2 framework running on Google Colab Tesla T4 GPU with 16GB of memory. Adversarial inputs are generated using Cleverhans 4.0.0. library.

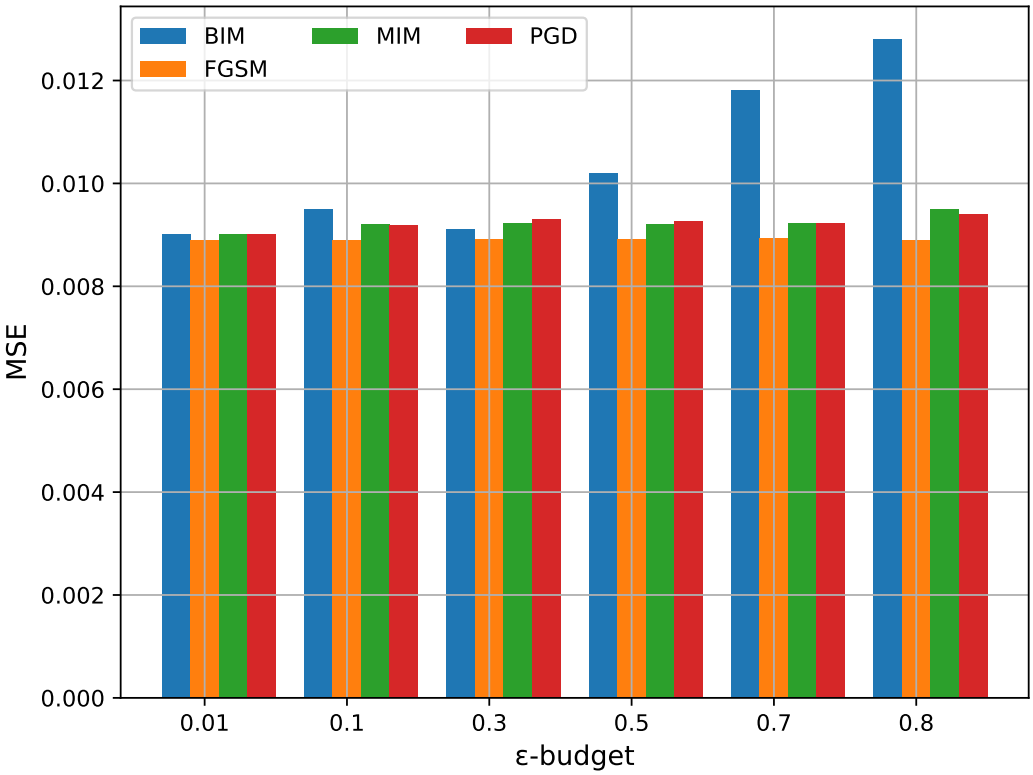

Fig. 29. MSE values of the undefended models for each adversarial machine learning attack under different attack powers (ϵ)

The adversarial attack on AI-powered models has become more popular with various attack methods. This study uses FGSM, MIM, BIM, and PGD methods to generate adversarial examples. The performance of each model is evaluated through the MSE metric.

Figure 29 shows MSE values for the selected attack methods under attack powers from $\epsilon = 0.01$ to $\epsilon = 0.8$. MSE values look similar for MIM, BIM, and PGD methods, i.e., around 0.09, for all attack powers. On the other hand, MSE values increase along with a higher attack power ($\epsilon > 0.5$) for BIM attacks and go from 0.009 to 0.0128. The results also indicate that AI-powered models are dramatically vulnerable to adversarial attacks.

Mitigation methods have been widely used to increase the robustness of the AI-powered model against adversarial attacks. In this study, the defensive distillation method is applied to the model to reduce vulnerability against adversarial attacks. The performance of the AI-powered model is evaluated in terms of MSE after applying the mitigation method. Figure 30 shows the performance of models, i.e., MSE values, against adversarial attacks from $\epsilon = 0.01$ to $\epsilon = 0.8$ after applying the selected mitigation method. The figure shows that the AI-powered model is still sensitive to adversarial attacks. However, the robustness is better against adversarial attacks. According to the figure, the model can resist any attack under low attack power ($\epsilon < 0.3$). The MSE values increase along with a high attack power ($\epsilon > 0.3$) as expected. However, the impact of the mitigation method on the performance is not the same for all attacks. For example, the MSE values can increase up to 0.006 and 0.008 under the PGD and MIM attack, respectively, while only going up to 0.003 under the BIM attack with a very high attack power ($\epsilon = 0.8$). It is very interesting that there is no impact on the attack power under the FGSM attack if the mitigation method is applied to the model. The results also indicate that

the defensive distillation method significantly contributes to the model's robustness against adversarial attacks.

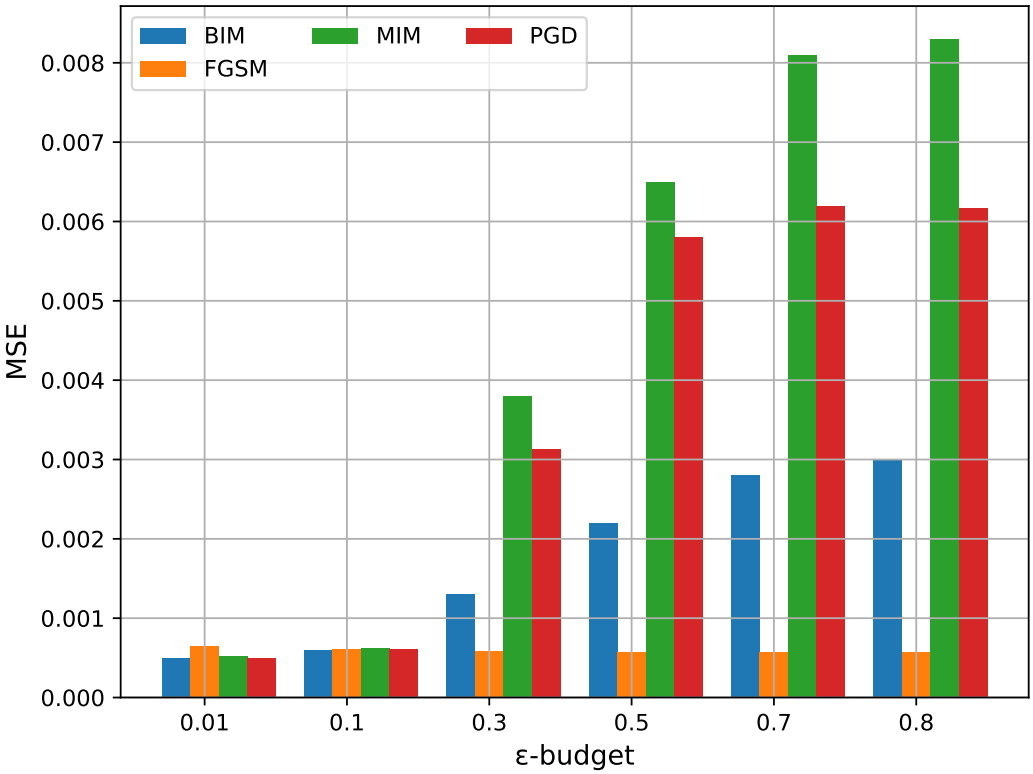

Fig. 30. MSE values of the defended models for each adversarial machine learning attack under different attack powers (ϵ)

The histogram plots investigate the distribution of MSE values for undefended and defended models under adversarial attacks. In Figure 31-37, (a) represents the undefended models, while (b) represents defended models for each attack, i.e., FGSM, BIM, MIM, and PGD, respectively. According to the results, the undefended models, i.e., (a), represent a little right-skewed distribution, which has a peak to the left of the distribution and data values that taper off to the right. MSE values vary from 0.005 to 0.025 for all attack types, and around 50% percent of MSE values are between 0.006 and 0.009. It is compatible with Figure 29-30. On the other hand, it is difficult to define the histogram plots for defended models, i.e., (b). According to the results,

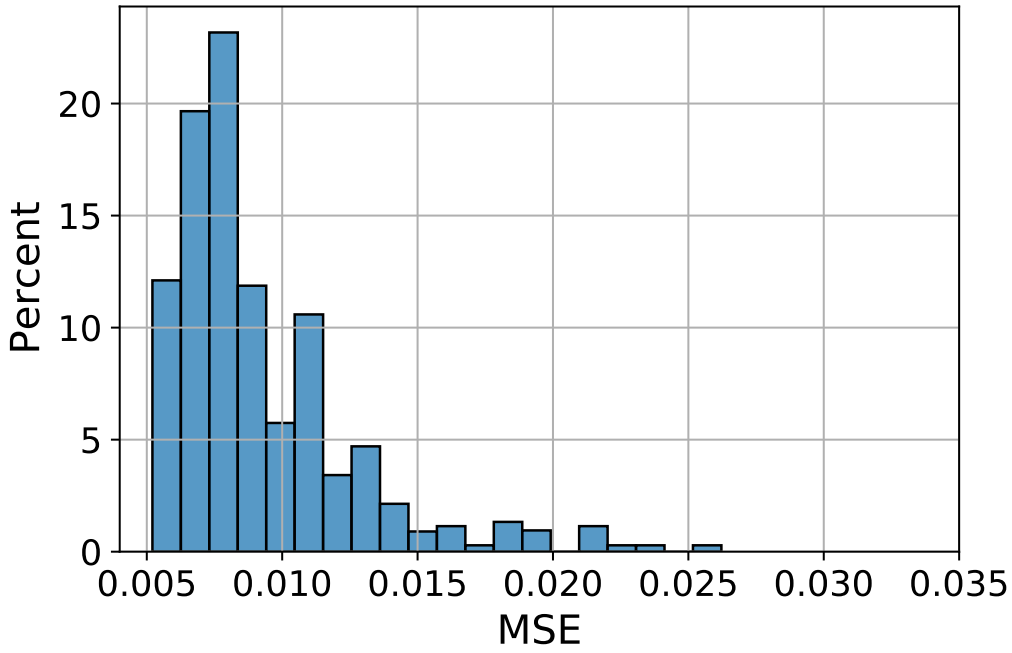

Fig. 31. Distribution of MSE values for undefended models under the FGSM attack

Figure 31, 33, 37 (b) represents a little right-skewed distribution like the undefended model ones, while Figure 35 (b) does not represent any distribution. The most MSE values are clustered around 0.0, i.e., 30% - 60%. This means the AI-powered model can correctly predict the target values. It is also clear that the percent of the high MSE values (< 0.015) is much lower than the undefended model. The defended models are more effective against FGSM and BIM attacks, as shown in Figure 31 and 33. It is obvious that the mitigation methods can dramatically improve the model robustness under FGSM attacks, i.e., 90% of MSE values are less than 0.005. On the other hand, the defended models are not successful against MIM and PGD attacks compared to FGSM and BIM, as shown in Figure 35 and 37. Although low MSE values, i.e., < 0.005, are clustered around 50%, the MSE values still go up to 0.015 for MIM and PGD attacks.

Table 3 shows the impact of a specific ϵ value on the MSE performance met-

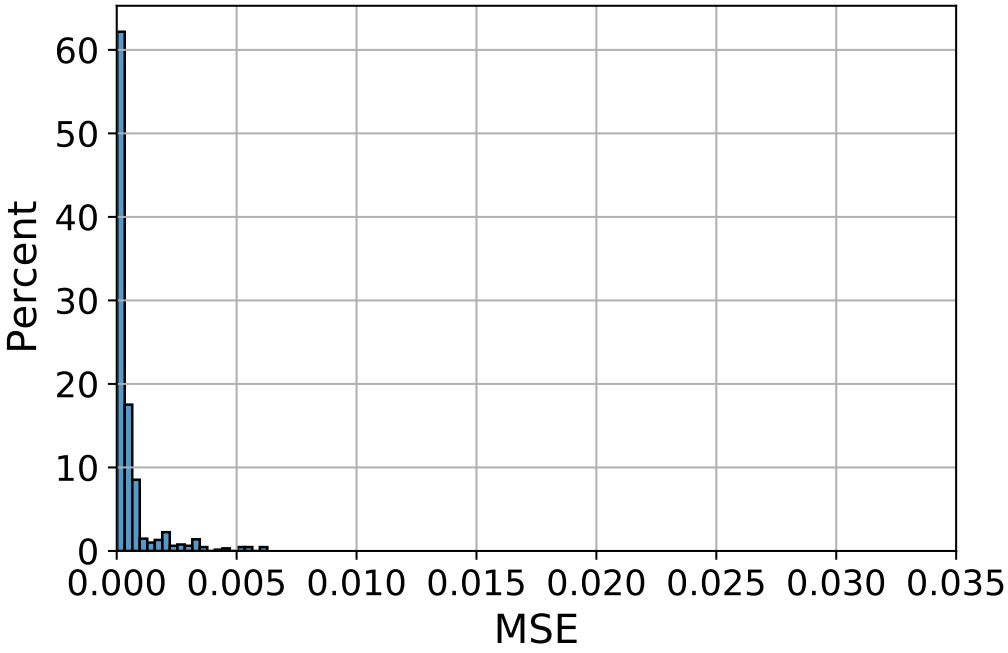

Fig. 32. Distribution of MSE values for defended models under the FGSM attack

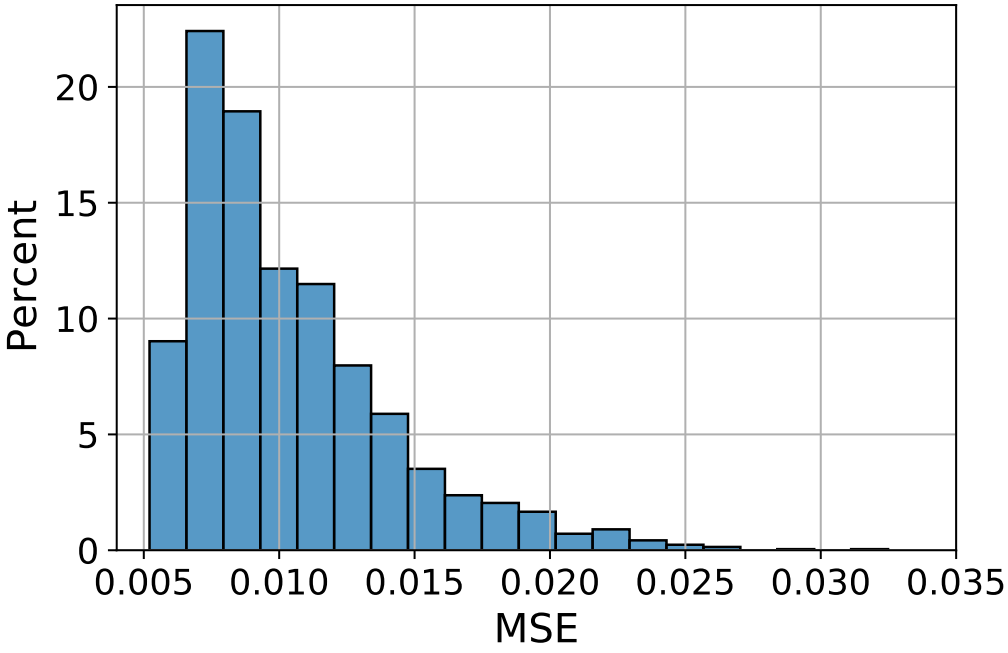

Fig. 33. Distribution of MSE values for undefended models under the BIM attack

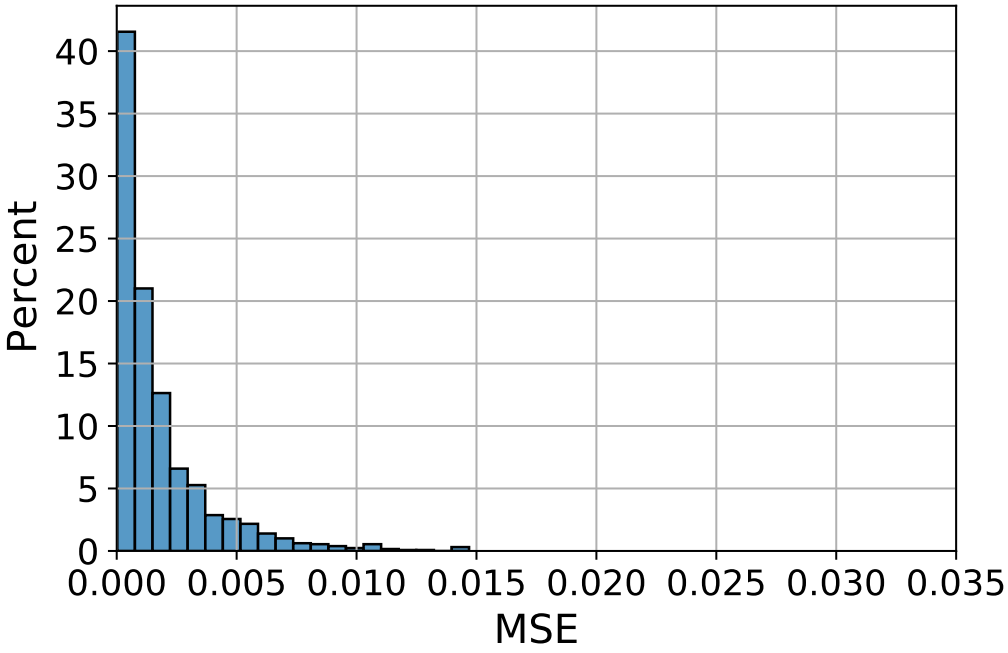

Fig. 34. Distribution of MSE values for defended models under the BIM attack

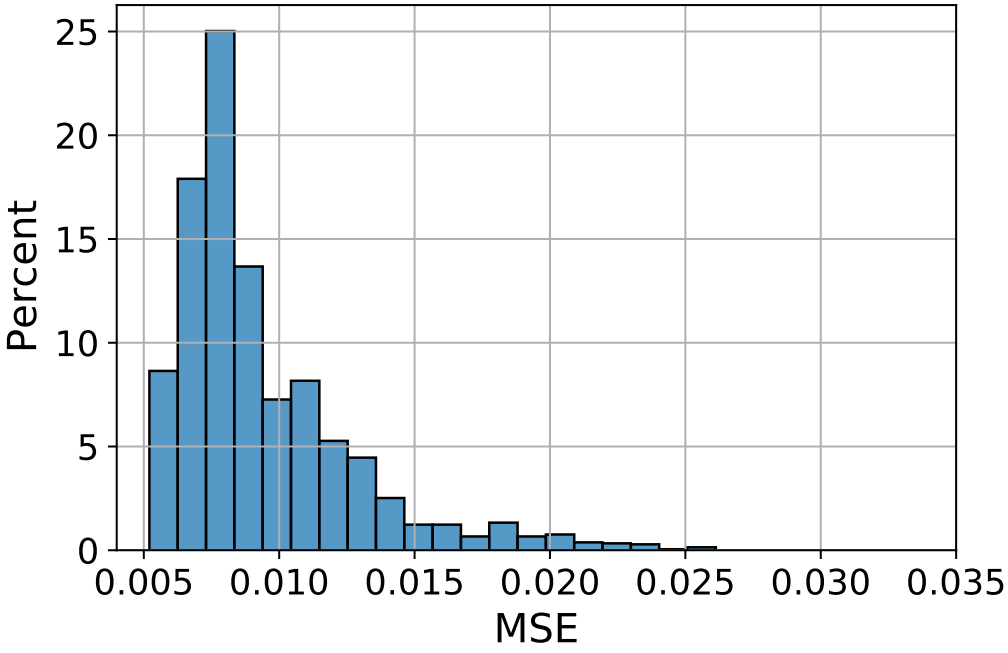

Fig. 35. Distribution of MSE values for undefended models under the MIM attack

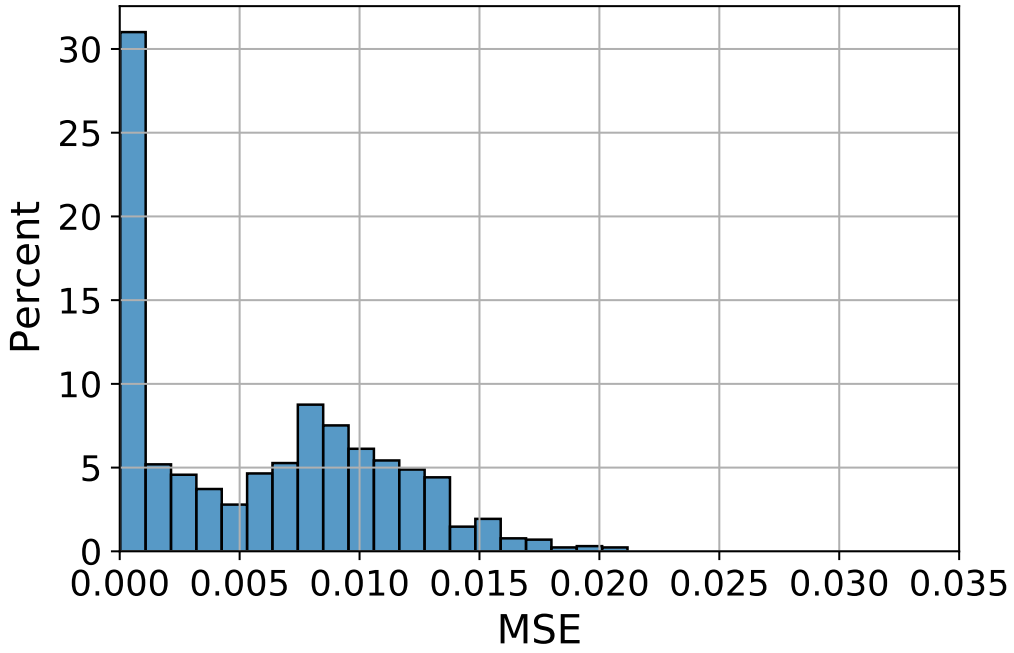

Fig. 36. Distribution of MSE values for defended models under the MIM attack

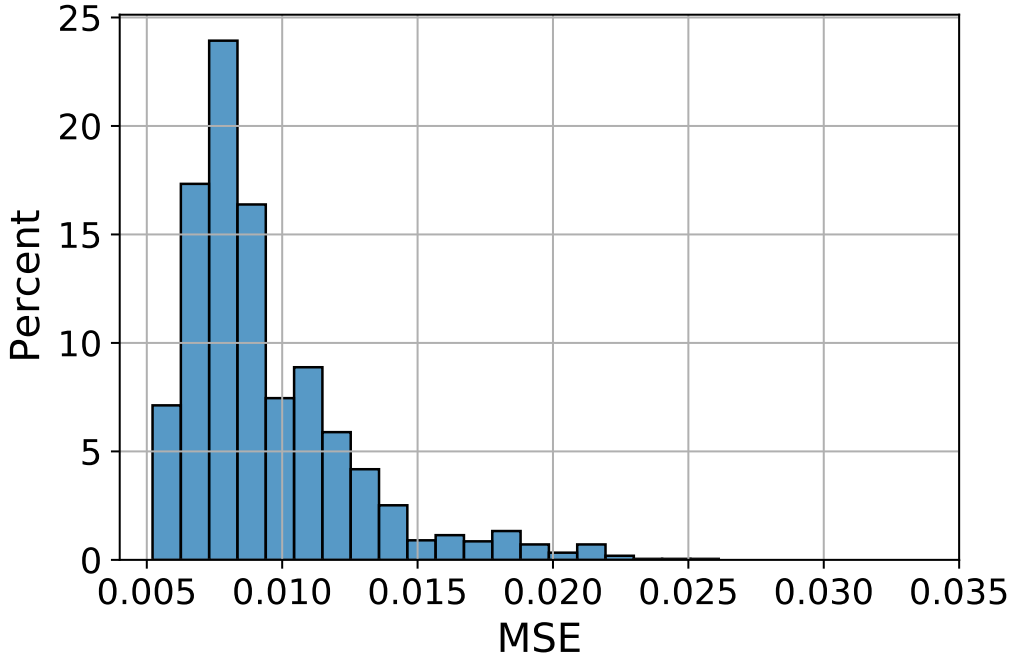

Fig. 37. Distribution of MSE values for undefended models under the MIM attack

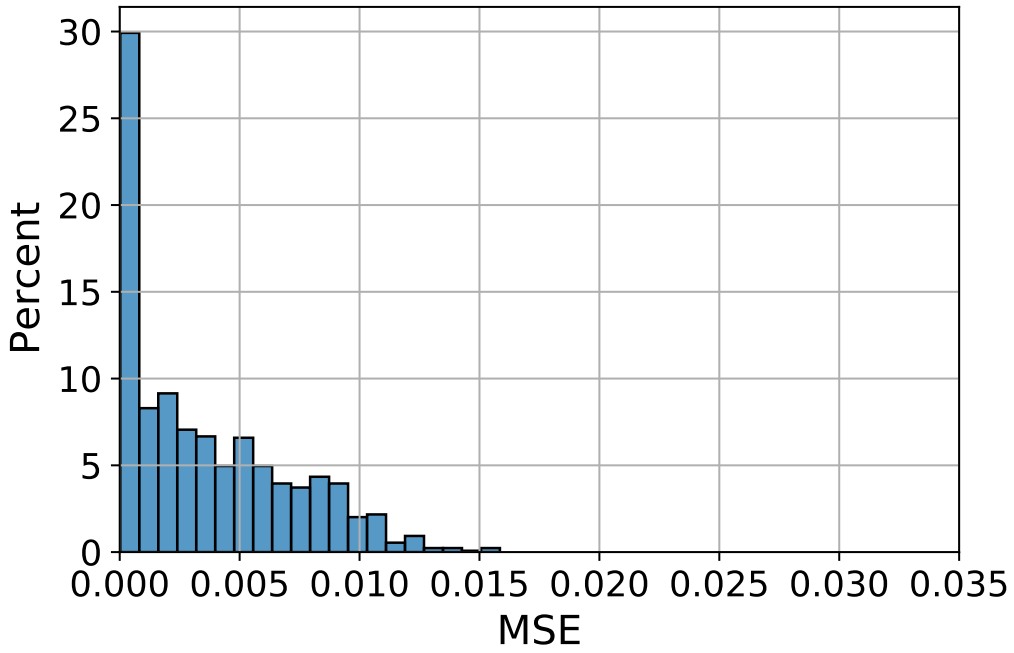

Fig. 38. Distribution of MSE values for defended models under the MIM attack

Table 3. Prediction performance results in terms of the MSE metric.

		ϵ **values**					
		0.01	**0.1**	**0.3**	**0.5**	**0.7**	**0.8**
FGSM	Undef.	0.009161	0.009162	0.009169	0.009177	0.009187	0.009193
	Distil.	0.000632	0.000631	0.00063	0.000628	0.000628	0.000629
BIM	Undef.	0.009205	0.009308	0.009191	0.010089	0.011761	0.012642
	Distil.	0.000555	0.000625	0.001321	0.002208	0.002895	0.002957
MIM	Undef.	0.009206	0.009402	0.009402	0.009269	0.009438	0.009539
	Distil.	0.00057	0.000663	0.003606	0.006696	0.008069	0.00836
PGD	Undef.	0.009206	0.009398	0.009582	0.009382	0.00937	0.009389
	Distil.	0.000555	0.00065	0.00294	0.005439	0.00618	0.006191

rics of the AI-powered IRS model for each adversarial attack in detail. The value of ϵ ranges from 0.01 to 0.8. The higher the value of ϵ means, the more powerful attack on the AI-powered model is expected. Except for BIM, the MSE values are usually around 0.0092-0.0095 for undefended models under any attack power and type. It reaches up to 0.012 under a high attack power (BIM and $\epsilon = 0.8$). However, MSE values dramatically decrease, e.g., from 0.0091/0.0092 to 0.0005/0.0006 for FGSM/BIM/MIM/PGD, once the mitigation method is applied. It is clear that the mitigation method significantly affects the robustness of the model, but not for all types of attacks. For example, MSE values are the same for the defended model under the FGSM attack at all attack powers. The mitigation method can handle FGSM-type attacks because of its simplicity. However, MSE values increase for the defended model under the other types of attacks at a high attack power level. For example, MSE values go from 0.0005 to 0.002, 0.0005 to 0.008, and 0.0005 to 0.006 for BIM, MIM, and PGD attacks, respectively. MSE values are the highest under the MIM attack (0.008 at 0.8 of the attack power). The MIM is the most effective adversarial attack type among the selected attacks.

5.3.5 Discussion

This study investigates AI-powered IRS models in NextG networks and their vulnerabilities against adversarial attacks and the contribution of mitigation methods to the model's robustness. The models' vulnerabilities are studied for various adversarial attacks, i.e., FGSM, BIM, MIM, and PGD, as well as the mitigation method, i.e., defensive distillation. The results show that AI-powered IRS models are vulnerable to adversarial attacks. On the other hand, the mitigation methods can significantly improve the model's robustness under adversarial attacks. According to the results, adversarial attacks on AI-powered IRS models and the use of the proposed mitigation

method can be summarized as:

Observation 1: AI-powered IRS models are vulnerable to adversarial attacks, especially BIM with a high attack power ($\epsilon > 0.5$).

Observation 2: There is no significant impact of the attack power (ϵ) on some adversarial attacks, i.e., FGSM.

Observation 3: The defensive distillation mitigation method significantly increases the model robustness, especially under FGSM and BIM attacks.

Observation 4: The MSE values histogram usually represents a smaller right-skewed distribution, especially for the undefended models.

Observation 5: Around 50% percent of MSE values are between 0.006 and 0.009 for the undefended models.

Observation 6: The most MSE values are clustered around 0.0, i.e., 30% - 60% for the defended model.

Observation 7: The most effective adversarial attack types are BIM and MIM for undefended and defended models, respectively.

5.4 Defending AI-based AMR models Against Adversarial Attacks

AMR is one of the critical steps in the signal-processing chain of wireless networks, which can significantly improve communication performance. AMR detects the modulation scheme of the received signal without any prior information. Recently, many AI-based AMR methods have been proposed, inspired by the considerable progress of AI methods in various fields. On the one hand, AI-based AMR methods can outperform traditional methods in terms of accuracy and efficiency. On the other hand, they are susceptible to new types of cyberattacks, such as model poisoning or adversarial attacks. This section explores the vulnerabilities of an AI-based AMR model to adversarial attacks in both single-input-single-output and multiple-

input-multiple-output scenarios.

5.4.1 System Model Overview

In this subsection, we first introduce the dataset preparation for both SISO and MIMO scenarios and then describe the adopted LSTM-based AMR model in the simulations. SISO refers to a communication system with only one antenna at the transmitter and one at the receiver while MIMO represents a communication system with multiple antennas at both the transmitter and the receiver. MIMO systems are used in modern wireless communication standards, such as 4G LTE, 5G, and beyond, to improve the data throughput, increase the range, and enhance the reliability of the communication link. The critical differences between SISO and MIMO systems are antenna configuration, channel capacity, complexity, and performance. MIMO systems typically provide higher data rates, longer ranges, and better reliability than SISO systems, especially in environments with multipath propagation and interference.

5.4.1.1 Dataset Preparation for SISO Scenario

To investigate the performance and vulnerability of the AI-based (i.e., LSTM-based) AMR model in a SISO scenario, the GNU radio ML dataset RML2016.10a [68] is adopted for simulations since this dataset is publicly available and widely used in research as the benchmark. There are 220,000 signal samples in the GNU radio ML dataset RML2016.10a, and each sample is associated with one modulation at a specific SNR level. Each sample consists of a 256-dimensional vector comprising 128 in-phase and 128 quadrature components. There are 11 different modulations, including BPSK, QPSK, 8PSK, QAM16, QAM64, CPFSK, GFSK, PAM4, WBFM, AM-SSB, and AM-DSB. The data samples are constructed at 20 different SNR levels

from -20 dB and 18 dB with an interval of 2 dB.

5.4.1.2 Dataset Preparation for MIMO Scenario

MIMO system with precoding is adopted from [49]. It is a common MIMO system that consists of a transmitter with N_t antennas and a receiver with N_r antennas. The transmitter and receiver are assumed to have full knowledge of the channel, and the transmission is over a flat fading channel. With the MIMO system above, we generate the dataset with three different antenna setting groups: ($N_t = 4$, $N_r = 2$), ($N_t = 16$, $N_r = 4$) and ($N_t = 64$, $N_r = 16$). The signal samples are modulated with six different modulations, i.e., 2PSK, QPSK, 8PSK, 16QAM, 64QAM, and 128QAM, at different SNR levels from -10 dB to 20 dB. 500 samples are prepared per SNR for each modulation, and the number of symbols transmitted per signal sample is 128.

5.4.1.3 Model Description

This subsection explains the LSTM-based AMR model adopted from [65]. RNN is commonly applied to learn persistent features of sequence data. LSTM is a particular type of RNN that is efficient in learning long-term dependencies and is heavily used for natural language processing and signal processing [64]. The major components in an LSTM cell are three gates, namely the input gate, the forget gate, and the output gate, which are used to control how the information propagates in the network. The gating mechanism allows LSTM cells to memorize information for extended periods, thus realizing continuous feature learning.

The adopted LSTM-based AMR model consists of two LSTM layers followed by a fully connected layer and a softmax layer, as shown in Figure 39. The in-phase and quadrature components of modulated signals are fed to the model as a two-dimensional vector. The first two LSTM layers have 128 LSTM units each, and the

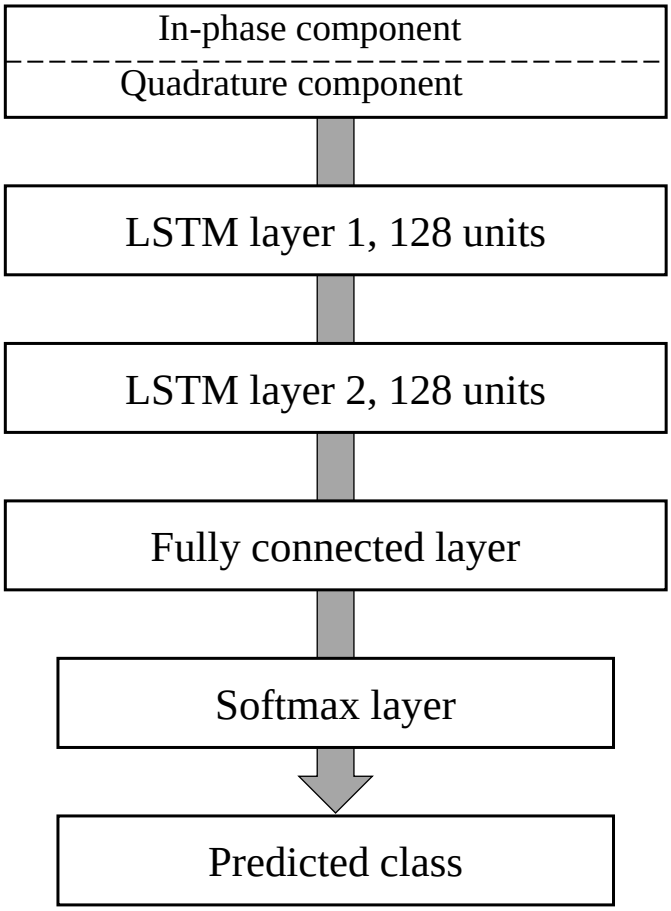

Fig. 39. The architecture of the LSTM-based AMR model. This model is trained for signal modulation recognition using the amplitude-phase signal.

output of the last LSTM layer is a 128-dimensional vector, which is passed to the following fully connected linear layer and softmax layer. In the SISO scenario, the softmax layer maps the features learned from previous layers to one of 11 output classes indicating the 11 modulation schemes. In the MIMO scenario, the softmax layer maps the features learned from previous layers to one of 6 output classes since there are six different modulations in the MIMO dataset. Essentially, the reason for using an LSTM model for signal classification is that signals with different modulation schemes contain different amplitude and phase features, and the LSTM model is

capable of learning these temporal features effectively.

To train the LSTM model for modulation recognition, first, the SISO and MIMO datasets are split into training, validation, and test sets at a ratio of 6:2:2 for SISO and MIMO scenarios, respectively. The loss function used is categorical cross-entropy, and the initial learning rate is set to 0.001 with the Adam optimizer. The learning rate will be halved if the validation loss does not decrease within 5 epochs, and the training process will be stopped if the validation loss remains stable for 50 epochs. The batch size is set to 400, and the training process is conducted using an Nvidia GTX 1080Ti GPU and Keras with Tensorflow as the backend.

5.4.2 Experiments

This section provides the experimental results for the SISO and MIMO scenarios using LSTM-based AMR undefended and defended models, with the attack success ratio. The attack success ratio refers to the ratio of successfully transmitted malicious data or signals to the total amount of data or signals transmitted. It is also widely used in communication systems to assess their security and measure how vulnerable the system is to different types of attacks. In this study, the experimental results are obtained by averaging across multiple iterations, i.e., 30 times. The analysis focuses on the attack success ratio of four different adversarial attack methods (BIM, FGSM, MIM, and PGD) with and without applying a mitigation method (defensive distillation). The objective is to identify potential weaknesses in the communication system.

A grid search approach is employed to determine the optimal parameters for defensive distillation-based adversarial ML attack mitigation. The grid search involved systematically exploring a predefined parameter grid to find the parameter combination that yielded the best performance. The parameters considered in the grid

search included the temperature parameter for defensive distillation, the regularization strength, and the learning rate. The combination that resulted in the highest model robustness against adversarial attacks was identified by exhaustively searching the parameter grid. This grid search methodology ensures a comprehensive exploration of parameter space, leading to an informed selection of the optimal parameters for defensive distillation in the context of adversarial ML attack mitigation.

5.4.3 Simulation Results in SISO Scenario

In a SISO scenario, the transmitter sends a single signal, which is received by the receiver over a single channel. This type of system is commonly used in simple point-to-point communication links, such as those between a mobile phone and a base station. Fig. 40 illustrates the attack success ratio of the undefended SISO model for each adversarial attack, i.e., BIM, FGSM, MIM, and PGD. According to the figure, the developed model is not robust under BIM, MIM, and PGD attacks, i.e., the attack success ratio can go up to 1.0 even under attack powers $\epsilon < 0.06$. However, the FGSM attack has a low success ratio compared to other attack methods, i.e., the maximum attack success ratio is 0.6 under a heavy attack power $\epsilon = 1.0$. It means the developed AMR model is robust against FGSM attacks. In some cases, FGSM attacks may be less effective than other more sophisticated attacks, such as BIM, MIM, and PGD attacks. Therefore, it is important to carefully consider the threat model and evaluate the effectiveness of different attack methods under different scenarios.

Table 4 shows the attack success ratio of different types of attacks along with different levels of attack strength for the undefended SISO model in detail. The first row shows the attack strength, ranging from 0.01 to 1.0, and the first column shows the names of the attack types, i.e., BIM, FGSM, MIM, and PGD. According to the table, BIM, MIM, and PGD have a high success ratio for most attack powers (ϵ),

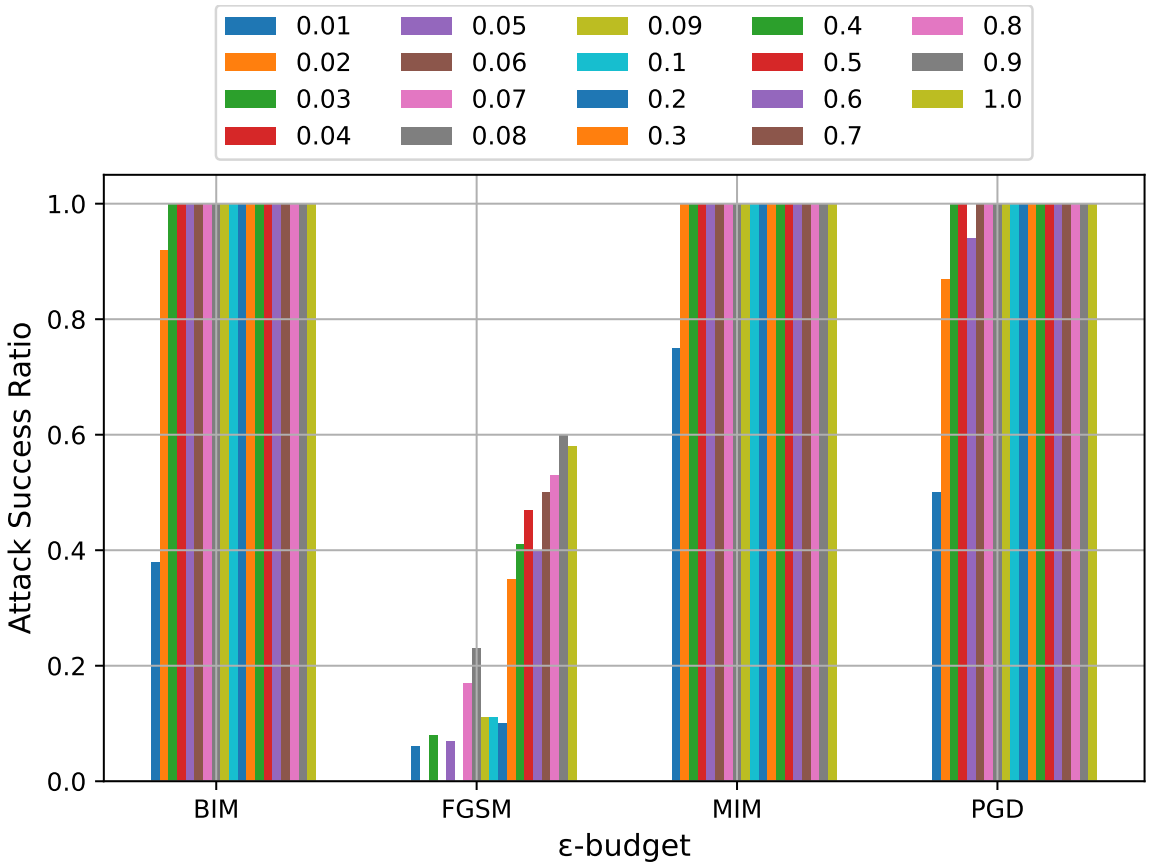

Fig. 40. Attack success ratio of the undefended SISO model.

while FGSM has a lower success ratio. This indicates that the other three attacks may be more effective than FGSM in generating adversarial attacks. For example, the BIM attack had a success ratio of 0.38, the FGSM attack had a success ratio of 0.06, the MIM attack had a success ratio of 0.75, and the PGD attack had a success ratio of 0.50 at an attack strength of 0.01. On the other hand, the values of the success ratio go up to 1.00, 0.58, 1.00, and 1.00 for BIM, FGSM, MIM, and PGD attacks at the highest attack power ($\epsilon = 1.0$), respectively.

Fig. 41 shows the attack success ratio of the defended SISO model under the selected attacks, i.e., BIM, FGSM, MIM, and PGD. According to the figure, all attack success ratio values decrease under all attack types compared to the undefended

Attack	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
BIM	0.38	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FGSM	0.06	0.00	0.08	0.00	0.07	0.00	0.17	0.23	0.11	0.11	0.10	0.35	0.41	0.47	0.40	0.50	0.53	0.60	0.58
MIM	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PGD	0.50	0.87	1.00	1.00	0.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 4. Attack success ratio of the undefended SISO model.

model. BIM, MIM, and PGD show similar trends. However, the attack success ratio values vary between around 0.1 to 0.6 under light ($\epsilon = 0.01$) and heavy attack powers ($\epsilon = 1.0$). As expected, the FGSM attack has a low success ratio compared to other attack methods; i.e., the maximum ratio is around 0.1 under all attack powers ϵ, and the developed model is more robust against FGSM attacks.

Table 5 provides detailed information about the performance of different attack methods on a machine learning model in terms of attack success ratio at different levels of attack power. The table is organized in a grid format, with the rows indicating the attack methods (BIM, FGSM, MIM, and PGD) and the columns indicating the strength of the attack (ranging from 0.01 to 1.0). Each cell in the table represents the success ratio of the corresponding attack method at the corresponding level of attack power. For instance, the success ratio of the BIM at 0.01 attack power is 0.11, while the attack success ratio at 1.0 attack power is 0.55. Similarly, the attack success ratio of MIM attack at 0.1 power is 0.04, while its attack success ratio at 1.0 attack power is 0.58. Note that the FGSM attack method has the least impact on the machine learning model, as its success ratio is consistently low at all levels of attack power. The minimum success ratio for FGSM is 0.00, while the maximum success ratio is 0.13 (at attack power of 0.9 and 1.0). On the other hand, the BIM, MIM, and PGD attack methods are more effective at compromising the model's performance. For BIM/MIM/PGD, the minimum success ratios are 0.11, 0.04, and 0.12 (at an attack power of 0.01), respectively. The maximum success ratios for these methods are 0.61, 0.60, and 0.67 at a high attack power, respectively.

5.4.4 Simulation Results in MIMO Scenario

In a MIMO scenario, multiple signals are transmitted simultaneously over multiple channels, and the receiver uses advanced signal processing techniques to sep-

Attack	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
BIM	0.11	0.28	0.19	0.30	0.33	0.40	0.52	0.42	0.45	0.45	0.55	0.57	0.60	0.56	0.57	0.60	0.57	0.61	0.55
FGSM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.04	0.04	0.02	0.10	0.09	0.09	0.13	0.13
MIM	0.04	0.22	0.32	0.36	0.29	0.32	0.44	0.53	0.60	0.54	0.55	0.52	0.51	0.52	0.53	0.60	0.55	0.56	0.58
PGD	0.12	0.15	0.25	0.35	0.35	0.31	0.37	0.59	0.50	0.45	0.50	0.59	0.55	0.67	0.60	0.63	0.63	0.60	0.59

Table 5. Attack success ratio of the defended SISO model.

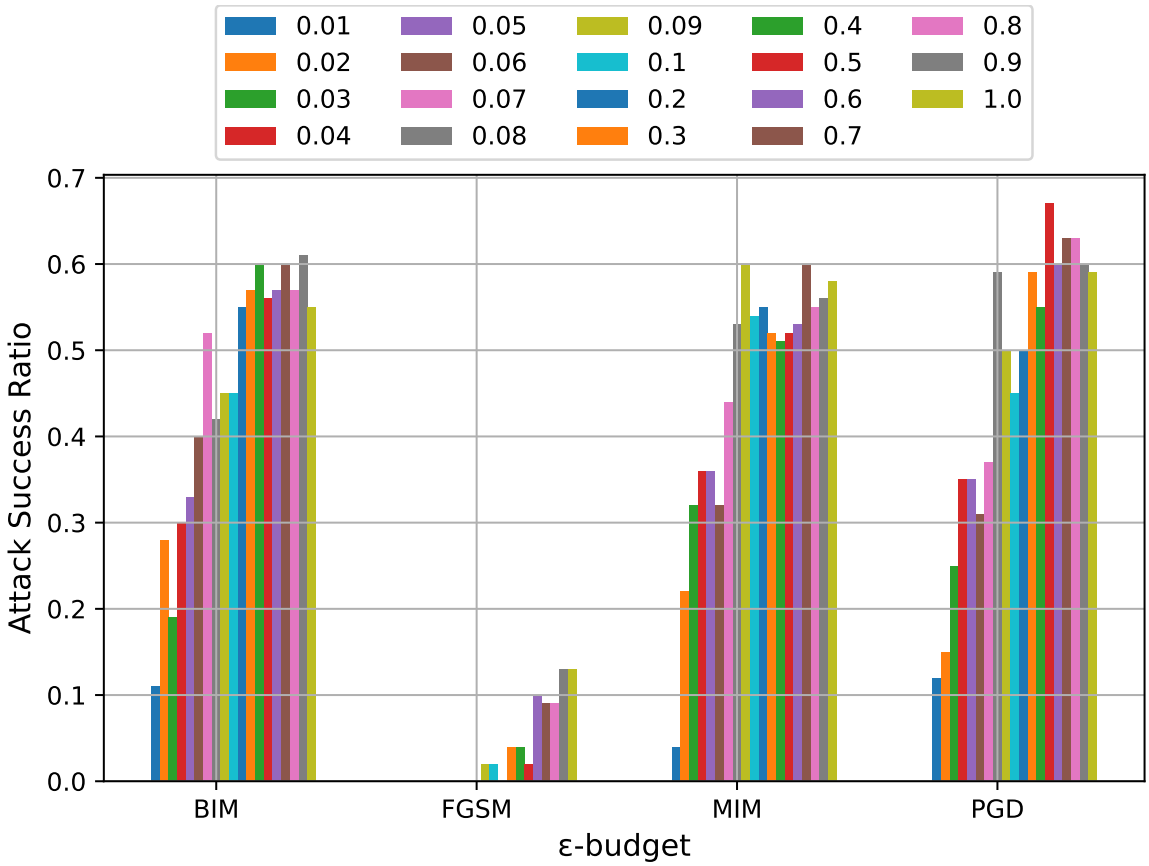

Fig. 41. Attack success ratio of the defended SISO model.

arate and decode the signals. Fig. 42 shows the attack success ratio of the defended MIMO model under the selected adversarial attacks. According to the figure, BIM/MIM/PFG attacks are very effective, and the attack success ratio values can achieve 1.0 (i.e., 100%) even at mid-level attack powers, $\epsilon >= 0.5$. As in the previous scenario, the FGSM attack has a low attack success ratio compared to other attack methods, i.e., the maximum attack success ratio is around 0.4 under heavy attack powers $\epsilon = 1.0$. It is obvious that the attack success ratio increases with the attack power in parallel. The details will be investigated in the following table.

Table 6 presents the attack success ratio for the selected four adversarial attacks (FGSM, BIM, MIM, and PGD) on the developed undefended MIMO model at differ-

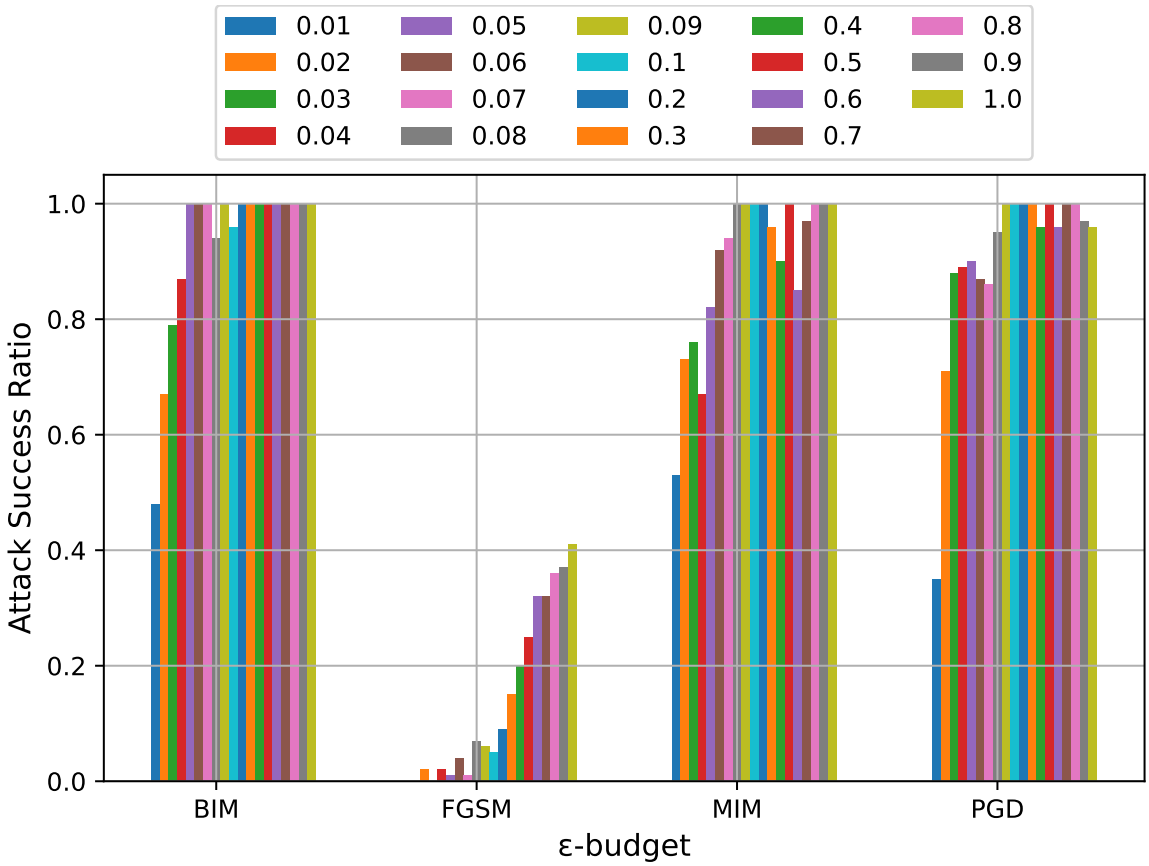

Fig. 42. Attack success ratio of the undefended MIMO model.

ent levels of attack powers (from 0.01 to 1.0). According to the table, all attack types except FGSM seem very effective, as they achieve a 100% attack success ratio on several high attack powers. Among them, BIM and MIM are the most effective attack methods against the model, as they achieve a high success ratio across a wide range of strength levels. On the other hand, FGSM is not very effective at lower strength levels (0.01 and 0.1) but becomes more effective as the strength level increases.

Fig. 43 illustrates the attack success ratio of the defended MIMO model for the same attacks and attack powers as in the previous scenario. The figure shows that the attack success ratio values significantly decrease for the defended MIMO model, especially for mid-level attack power. BIM/MIM/PGD attacks exhibit similar trends,

Attack	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
BIM	0.48	0.67	0.79	0.87	1.00	1.00	1.00	0.94	1.00	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FGSM	0.00	0.02	0.00	0.02	0.01	0.04	0.01	0.07	0.06	0.05	0.09	0.15	0.20	0.25	0.32	0.32	0.36	0.37	0.41
MIM	0.53	0.73	0.76	0.67	0.82	0.92	0.94	1.00	1.00	1.00	1.00	0.96	0.90	1.00	0.85	0.97	1.00	1.00	1.00
PGD	0.35	0.71	0.88	0.89	0.90	0.87	0.86	0.95	1.00	1.00	1.00	1.00	0.96	1.00	0.96	1.00	1.00	0.97	0.96

Table 6. Attack success ratio of the undefended MIMO model.

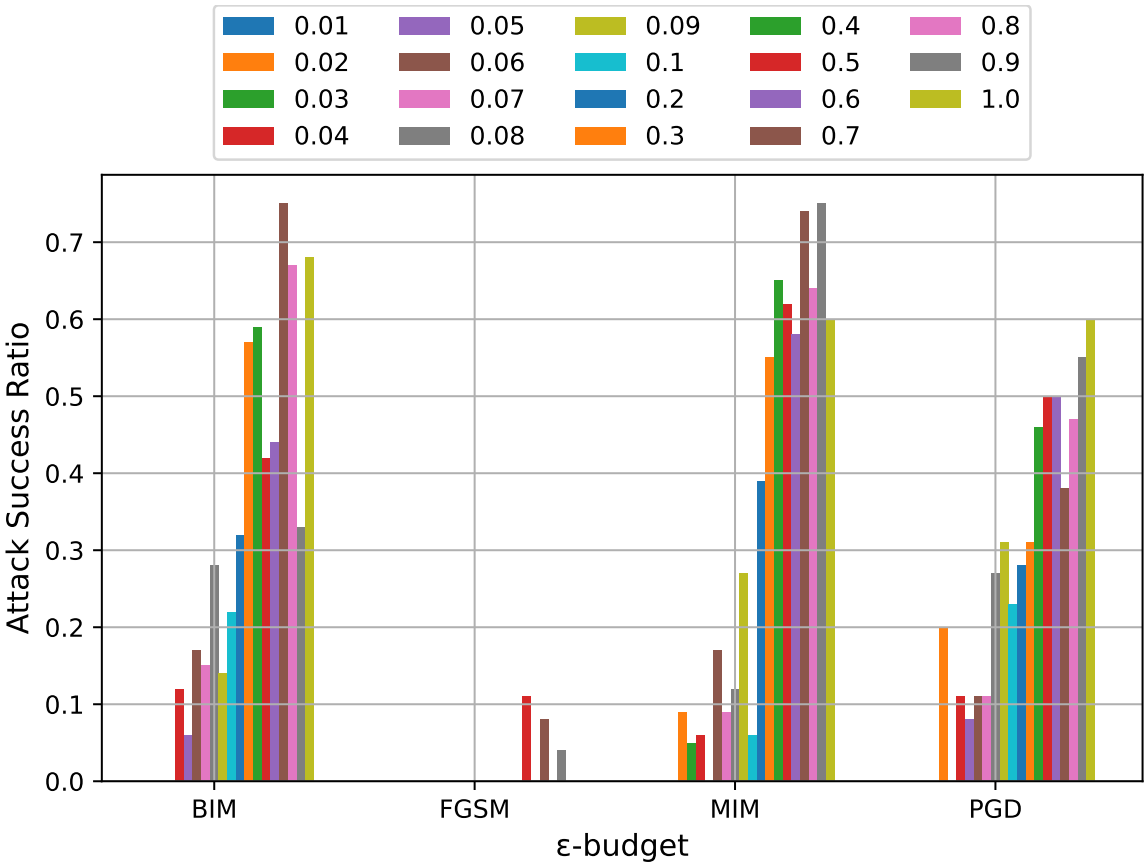

Fig. 43. Attack success ratio of the defended MIMO model.

i.e., having a low attack success ratio at low attack power and a high attack success ratio at high attack powers. As expected, the FGSM attack method has the least impact, i.e., almost none, as its success ratio is consistently low at all levels of attack power, i.e., around 0.1. Some results show a zero (0) attack success ratio, meaning the attack success ratio is very low or almost 0.

Table 7 provides more detailed information regarding the attack success ratio of different adversarial attack methods on the defended MIMO model at different attack powers. According to the table, the FGSM attack has almost no impact on the defended MIMO model at all attack powers, i.e., the maximum attack success ratio is 0.11. Other attack types (BIM/MIM/PGD) still impact the defended model.

For example, in the BIM attack, at 0.01 attack power, the success ratio is 0.0, meaning the attack was not successful. However, at 0.7 attack power, the success ratio jumps to 0.75. For the MIM attack, the success ratio for 0.01 attack power is 0.0, but it increases to 0.09 for 0.02 attack power. The success ratio remains low for the following attack powers but increases substantially for higher attack powers, reaching a maximum success ratio of 0.75 for 0.9 attack power. For the PGD attack, the success ratio remains 0.0 for 0.01 attack power, but it increases to 0.20 for 0.02 attack power. The success ratio then varies between 0.0 and 0.5 for different attack powers and reaches a maximum value of 0.60 for 1.0 attack power.

5.4.5 Discussion

This study aims to investigate the performance and vulnerabilities of AI-based AMR models under popular adversarial attacks, such as FGSM, BIM, MIM, and PGD, as well as the impact of the selected mitigation method (defensive distillation) on performance improvement. The simulation results indicate that AI-based AMR models are vulnerable to model poisoning attacks, but the impact can be reduced or eliminated with mitigation methods. Based on the findings, the following observations can be made:

Observation 1: Adversarial attacks are effective in compromising the accuracy of deep learning models, with attack success ratios ranging from 0% to over 100% depending on the attack method and power.

Observation 2: The attack success ratio of adversarial attacks tends to increase with the attack power. In most cases, attack success ratios increase rapidly as the attack power goes from 0.01 to 0.1 but then plateau or increase more slowly for larger attack powers.

Observation 3: Mitigation methods can reduce the attack success ratios of adversarial

Attack	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
BIM	0.0	0.00	0.00	0.12	0.06	0.17	0.15	0.28	0.14	0.22	0.32	0.57	0.59	0.42	0.44	0.75	0.67	0.33	0.68
FGSM	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.08	0.00	0.04	0.00
MIM	0.0	0.09	0.05	0.06	0.00	0.17	0.09	0.12	0.27	0.06	0.39	0.55	0.65	0.62	0.58	0.74	0.64	0.75	0.60
PGD	0.0	0.20	0.00	0.11	0.08	0.11	0.11	0.27	0.31	0.23	0.28	0.31	0.46	0.50	0.50	0.38	0.47	0.55	0.60

Table 7. Attack success ratio of the defended SISO model.

attacks, but their effectiveness varies depending on the attack method and power.

Observation 4: MIMO models can provide better defense against adversarial attacks compared to single-input single-output (SISO) models.

Observation 5: Adversarial attacks significantly impact both undefended and defended SISO/MIMO models in terms of attack success ratio, particularly for BIM, MIM, and PGD attacks.

Observation 6: FGSM attack method has the least impact on models, as its success ratio is consistently low at all levels of attack power.

Observation 7: PGD is the most effective attack against the defended SISO model, with an attack success ratio of 0.67.

Observation 8: MIM is the most effective attack against the defended MIMO model, with an attack success ratio of 0.75.

5.5 Conclusion

The next-generation networks, i.e., NextG or 5G and beyond, have dramatically enhanced along with advanced communication, computing, and AI technologies in the last decade. AI is the most important contributor to the improvement of NextG in terms of performance. This chapter investigates the vulnerability of AI-powered IRS models against adversarial attacks (i.e., FGSM, BIM, PGD, and MIM) and the impact of the proposed mitigation method, i.e., defensive distillation, on improving the robustness of AI models in NextG networks. The results indicate that the AI-powered NextG networks are vulnerable to adversarial attacks. On the other hand, mitigation methods can make the models more robust against adversarial attacks. According to the overall results, the most effective adversarial attack types are BIM and MIM for undefended and defended models, respectively. The proposed mitigation

method can provide better results for the attacks, including FGSM, BIM, MIM, and PGD, in terms of increasing the model robustness and reducing the vulnerability.

In addition, we evaluated the performance of AI-based AMR models and their robustness against various adversarial attacks (i.e., FGSM, BIM, PGD, and MIM) with and without the selected mitigation method (defensive distillation). The experimental results demonstrate that both undefended and defended SISO/MIMO models are vulnerable to adversarial attacks, with attack success ratio values significantly increasing at high attack power. In the defended SISO model, the PGD attack has the highest success ratio, followed by BIM and MIM attacks. In the defended MIMO model, the MIM attack has the highest success ratio, followed by BIM and PGD attacks. The FGSM attack had minimal impact on the attack success ratio for both undefended and defended SISO/MIMO models compared to other adversarial attack types due to its simplicity and limitations, i.e., linear approximation, limited perturbation strength, and knowledge of the model. The experimental results also reveal that mitigating methods significantly impact model robustness, reducing the attack success ratio of all attacks. These findings highlight the need to develop more secure and robust AI-based models for next-generation communication technologies to protect against adversarial attacks.

In future work, we will focus on adversarial attack detection in AI-based communications models, which is the necessary step before attack mitigation. Furthermore, we will attempt to develop better defense mechanisms against adversarial attacks for the AI-based communication models, improving the security of machine learning systems. While the current study provides valuable insights into the effectiveness of defensive distillation for defending AI-based models against adversarial attacks, it is acknowledged that further comparisons and sensitivity/stability analyses are warranted. These additional analyses, planned as part of future work, will enable a more

comprehensive evaluation of the proposed approach, including comparisons with alternative mitigation methods and assessment of the sensitivity of models to different attack scenarios and stability over varying conditions. This will provide a more robust and convincing evaluation of the AI-based communication models.

CHAPTER 6

CONCLUSIONS AND FUTURE WORKS

The preceding chapters have explored the security and machine learning applications in NextG networks. Specifically, the security scenarios for IRS-aided communication systems are studied. Two deep learning-based models for wireless signal denoising and signal modulation recognition are developed. The vulnerability of two AI-powered communication systems under machine learning adversarial attacks is also investigated. This chapter summarizes the main contributions of this thesis in Section 6.1 and discusses potential future directions in Section 6.2.

6.1 Conclusions

The main contributions of this dissertation can be summarized in the following four aspects.

1. We have studied the potential security impact of the IRS on wireless communication systems from two perspectives: improving the security of communications and committing security threats using the IRS in four scenarios. Both perspectives benefit from the flexibility of reprogramming a communication environment by adjusting its position and phase shifts of reflecting elements. Extensive simulations have been conducted for all scenarios.

2. Wireless signals are not only weakened by the communication environment but also disrupted by malicious attackers. To enhance the signal quality, we developed a generative adversarial networks-based wireless signal denoising method. Simulations on different types of wireless signals have demonstrated that our

proposed method outperforms traditional signal denoising techniques in a dynamically changing environment.

3. AMR is an essential component in wireless communications and can be applied to various scenarios. We defined AMR as a multi-class classification task in deep learning and proposed a model that extracts the spatial and temporal features from wireless signals in parallel. Simulations are conducted using wireless signals with 11 different modulation methods. Our proposed method delivers a more convincing performance compared to typical deep learning-based AMR techniques.

4. Although integrating AI into NextG wireless communications brings new opportunities and improves performance in various aspects, it leads to a significant security concern, i.e., adversarial attacks. We evaluated the performances of two AI-powered communication systems under four common machine-learning adversarial attacks and discussed their vulnerabilities. In addition, a defensive distillation method is introduced to protect AI-powered models and mitigate attacks.

6.2 Future Works

NextG promises ultra-fast speeds, extremely low latency, and increased capacity to support a massive number of devices. This will enable advancements in various fields, including autonomous vehicles, smart cities, and advanced healthcare solutions. IRS and AI have been recognized as crucial components in contributing to advanced networks. Researchers have been actively integrating IRS and AI into NextG to improve communication performance from different aspects. However, several concerns and challenges remain to be addressed.

- Recent research on IRS-aided wireless communications primarily focuses on optimizing the phase shift coefficients at the IRS while considering various constraints such as total transmit power and the presence of eavesdroppers. However, these works are based on a strong assumption that the IRS is able to identify the signal sources so that the IRS can be controlled to strengthen or weaken the total received signal strength. Therefore, an efficient signal identification method is desirable to assist the operations of IRS-aided communication systems.

- It is extremely challenging to acquire the full channel state information (CSI) in IRS-aided communication systems due to the large number of IRS elements. In many research works, it is generally assumed to be known. However, having full CSI can significantly enhance the performance and security of wireless communication systems by enabling adaptive techniques, improving resource management, and reducing interference. Therefore, a scalable approach to obtaining the CSI for IRS needs to be developed.

- As more devices connect to NextG networks, the communication scenarios will become increasingly complex. AI is adaptive and intelligent, with immense potential to transform wireless communication. It can be applied to allocate resources and configurations for efficient network management. Additionally, AI can enhance channel estimation accuracy and develop dynamic channel models, especially in complex and dynamic environments. In addition to adversarial attacks, AI-driven communication systems are also vulnerable to other threats, including model manipulation, inference attacks, and backdoor attacks. Consequently, a thorough investigation into the robustness of AI-driven communication systems is necessary.

REFERENCES

[1] Haolin Tang et al. "Security and Threats of Intelligent Reflecting Surface Assisted Wireless Communications". In: *2022 International Conference on Computer Communications and Networks (ICCCN)*. IEEE. 2022, pp. 1–9.

[2] Haolin Tang et al. "Wireless Signal Denoising Using Conditional Generative Adversarial Networks". In: *IEEE INFOCOM 2023-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE. 2023, pp. 1–6.

[3] Haolin Tang et al. "Automatic Modulation Recognition Using Parallel Feature Extraction Architecture". In: *Wireless Algorithms, Systems, and Applications: 18th International Conference, WASA 2024*. Springer. 2024.

[4] Ferhat Ozgur Catak et al. "Security Hardening of Intelligent Reflecting Surfaces Against Adversarial Machine Learning Attacks". In: *IEEE Access* 10 (2022), pp. 100267–100275.

[5] Haolin Tang et al. "Defending AI-Based Automatic Modulation Recognition Models Against Adversarial Attacks". In: *IEEE Access* (2023).

[6] Qingqing Wu and Rui Zhang. "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network". In: *IEEE Communications Magazine* 58.1 (2019), pp. 106–112.

[7] Jiancheng An et al. "Optimal Pilot Power Based Channel Estimation Improves the Throughput of Intelligent Reflective Surface Assisted Systems". In: *IEEE Transactions on Vehicular Technology* (2020).

[8] Zhiqing Tang et al. “Physical Layer Security of Intelligent Reflective Surface Aided NOMA Networks”. In: *arXiv preprint arXiv:2011.03417* (2020).

[9] Shree Prasad Maruthi, Trilochan Panigrahi, and Mahbub Hassan. “Improving the Reliability of Pulse-Based Terahertz Communication using Intelligent Reflective Surface”. In: *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE. 2020, pp. 1–6.

[10] Dong Ma, Ming Ding, and Mahbub Hassan. “Enhancing cellular communications for UAVs via intelligent reflective surface”. In: *2020 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE. 2020, pp. 1–6.

[11] Zheng Chu et al. “Intelligent reflecting surface aided multi-antenna secure transmission”. In: *IEEE Wireless Communications Letters* 9.1 (2019), pp. 108–112.

[12] Yuanbin Chen et al. “Resource Allocation for Intelligent Reflecting Surface Aided Vehicular Communications”. In: *IEEE Transactions on Vehicular Technology* 69.10 (2020), pp. 12321–12326.

[13] Qingqing Wu and Rui Zhang. “Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming”. In: *IEEE Transactions on Wireless Communications* 18.11 (2019), pp. 5394–5409.

[14] Jiakuo Zuo et al. “Resource allocation in intelligent reflecting surface assisted NOMA systems”. In: *arXiv preprint arXiv:2002.01765* (2020).

[15] Zhiguo Ding and H Vincent Poor. “A simple design of IRS-NOMA transmission”. In: *IEEE Communications Letters* 24.5 (2020), pp. 1119–1123.

[16] Xinwei Yue and Yuanwei Liu. "Performance analysis of intelligent reflecting surface assisted NOMA networks". In: *arXiv preprint arXiv:2002.09907* (2020).

[17] Xidong Mu et al. "Exploiting intelligent reflecting surfaces in NOMA networks: Joint beamforming optimization". In: *IEEE Transactions on Wireless Communications* 19.10 (2020), pp. 6884–6898.

[18] Meng Hua et al. "UAV-assisted intelligent reflecting surface symbiotic radio system". In: *IEEE Transactions on Wireless Communications* (2021).

[19] Ertugrul Basar et al. "Wireless communications through reconfigurable intelligent surfaces". In: *IEEE access* 7 (2019), pp. 116753–116773.

[20] Qingqing Wu et al. "Intelligent reflecting surface aided wireless communications: A tutorial". In: *IEEE Transactions on Communications* (2021).

[21] Miao Cui, Guangchi Zhang, and Rui Zhang. "Secure wireless communication via intelligent reflecting surface". In: *IEEE Wireless Communications Letters* 8.5 (2019), pp. 1410–1414.

[22] Limeng Dong and Hui-Ming Wang. "Secure MIMO transmission via intelligent reflecting surface". In: *IEEE Wireless Communications Letters* 9.6 (2020), pp. 787–790.

[23] Zheng Zhang et al. "Robust and secure communications in intelligent reflecting surface assisted NOMA networks". In: *IEEE Communications Letters* 25.3 (2020), pp. 739–743.

[24] Hong Shen et al. "Secrecy rate maximization for intelligent reflecting surface assisted multi-antenna communications". In: *IEEE Communications Letters* 23.9 (2019), pp. 1488–1492.

[25] Limeng Dong et al. “Double intelligent reflecting surface for secure transmission with inter-surface signal reflection”. In: *IEEE Transactions on Vehicular Technology* 70.3 (2021), pp. 2912–2916.

[26] Bo Yang et al. “Intelligent Spectrum Learning for Wireless Networks with Reconfigurable Intelligent Surfaces”. In: *IEEE Transactions on Vehicular Technology* (2021).

[27] Salih Sarp, Haolin Tang, and Yanxiao Zhao. “Use of Intelligent Reflecting Surfaces For and Against Wireless Communication Security”. In: *2021 IEEE 4th 5G World Forum (5GWF)*. IEEE. 2021, pp. 374–377.

[28] Helin Yang et al. “Deep reinforcement learning based intelligent reflecting surface for secure wireless communications”. In: *IEEE Transactions on Wireless Communications* (2020).

[29] Bin Lyu et al. “IRS-based wireless jamming attacks: When jammers can attack without power”. In: *IEEE Wireless Communications Letters* 9.10 (2020), pp. 1663–1667.

[30] Jie Yang et al. “A Novel Pilot Spoofing Scheme via Intelligent Reflecting Surface Based On Statistical CSI”. In: *IEEE Transactions on Vehicular Technology* 70.12 (2021), pp. 12847–12857.

[31] Ertugrul Basar and Ibrahim Yildirim. “SimRIS channel simulator for reconfigurable intelligent surface-empowered communication systems”. In: *2020 IEEE Latin-American Conference on Communications (LATINCOM)*. IEEE. 2020, pp. 1–6.

[32] Saeed V Vaseghi. *Advanced digital signal processing and noise reduction*. John Wiley & Sons, 2008.

[33] Jing Li, Jingyuan Wang, and Zhang Xiong. "Wavelet-based stacked denoising autoencoders for cell phone base station user number prediction". In: *2016 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*. IEEE. 2016, pp. 833–838.

[34] Haolin Tang and Yanxiao Zhao. "A Conditional Generative Adversarial Network for Non-rigid Point Set Registration". In: *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*. IEEE. 2021, pp. 1–6.

[35] Hao Ye, Geoffrey Ye Li, and Biing-Hwang Juang. "Power of deep learning for channel estimation and signal detection in OFDM systems". In: *IEEE Wireless Communications Letters* 7.1 (2017), pp. 114–117.

[36] Yongshi Wang et al. "Residual learning based RF signal denoising". In: *2018 IEEE International Conference on Applied System Invention (ICASI)*. IEEE. 2018, pp. 15–18.

[37] Ian Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (2014).

[38] Meet H Soni, Neil Shah, and Hemant A Patil. "Time-frequency masking-based speech enhancement using generative adversarial network". In: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2018, pp. 5039–5043.

[39] Mehdi Mirza and Simon Osindero. "Conditional generative adversarial nets". In: *arXiv preprint arXiv:1411.1784* (2014).

[40] Laid Chergui and Saad Bouguezel. “A new pre-whitening transform domain LMS algorithm and its application to speech denoising”. In: *Signal processing* 130 (2017), pp. 118–128.

[41] Jun Wan, Xiaohui Zhang, and Jionghui Rao. “Research and application of denoising method based on wavelet threshold”. In: *2010 2nd International Conference on Information Engineering and Computer Science*. IEEE. 2010, pp. 1–4.

[42] Ebtesam Almazrouei et al. “Using autoencoders for radio signal denoising”. In: *Proceedings of the 15th ACM International Symposium on QoS and Security for Wireless and Mobile Networks*. 2019, pp. 11–17.

[43] Linh Duy Tran, Son Minh Nguyen, and Masayuki Arai. “GAN-based noise model for denoising real images”. In: *Proceedings of the Asian Conference on Computer Vision*. 2020.

[44] Xudong Mao et al. “Least squares generative adversarial networks”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2794–2802.

[45] Bing Xu et al. “Empirical evaluation of rectified activations in convolutional network”. In: *arXiv preprint arXiv:1505.00853* (2015).

[46] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *International conference on machine learning*. PMLR. 2015, pp. 448–456.

[47] Jun Yan Dai et al. “Realization of multi-modulation schemes for wireless communication by time-domain digital coding metasurface”. In: *IEEE transactions on antennas and propagation* 68.3 (2019), pp. 1618–1627.

[48] Haolin Tang et al. “Security and Threats of Intelligent Reflecting Surface Assisted Wireless Communications”. In: *2022 International Conference on Computer Communications and Networks (ICCCN)*. IEEE. 2022, pp. 1–9.

[49] Fuxin Zhang et al. “Deep learning based automatic modulation recognition: Models, datasets, and challenges”. In: *Digital Signal Processing* (2022), p. 103650.

[50] Ade Pitra Hermawan et al. “CNN-based automatic modulation classification for beyond 5G communications”. In: *IEEE Communications Letters* 24.5 (2020), pp. 1038–1041.

[51] Dehua Hong, Zilong Zhang, and Xiaodong Xu. “Automatic modulation classification using recurrent neural networks”. In: *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*. IEEE. 2017, pp. 695–700.

[52] Judith Nkechinyere Njoku, Manuel Eugenio Morocho-Cayamcela, and Wansu Lim. “CGDNet: Efficient hybrid deep learning model for robust automatic modulation recognition”. In: *IEEE Networking Letters* 3.2 (2021), pp. 47–51.

[53] Godwin Brown Tunze et al. “Sparsely connected CNN for efficient automatic modulation recognition”. In: *IEEE Transactions on Vehicular Technology* 69.12 (2020), pp. 15557–15568.

[54] Timothy O’shea and Jakob Hoydis. “An introduction to deep learning for the physical layer”. In: *IEEE Transactions on Cognitive Communications and Networking* 3.4 (2017), pp. 563–575.

[55] Cheng Yang et al. “Deep Learning Aided Method for Automatic Modulation Recognition”. In: *IEEE Access* 7 (2019), pp. 109063–109068. DOI: 10.1109/ACCESS.2019.2933448.

[56] Xiaoyu Liu, Diyu Yang, and Aly El Gamal. "Deep neural network architectures for modulation classification". In: *2017 51st Asilomar Conference on Signals, Systems, and Computers*. IEEE. 2017, pp. 915–919.

[57] Tara N Sainath et al. "Convolutional, long short-term memory, fully connected deep neural networks". In: *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. Ieee. 2015, pp. 4580–4584.

[58] Zufan Zhang et al. "Automatic modulation classification using CNN-LSTM based dual-stream structure". In: *IEEE Transactions on Vehicular Technology* 69.11 (2020), pp. 13521–13531.

[59] Pejman Ghasemzadeh, Michael Hempel, and Hamid Sharif. "GS-QRNN: A high-efficiency automatic modulation classifier for cognitive radio IoT". In: *IEEE Internet of Things Journal* 9.12 (2022), pp. 9467–9477.

[60] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[61] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[62] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012).

[63] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[64] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[65] Sreeraj Rajendran et al. “Deep learning models for wireless signal classification with distributed low-cost spectrum sensors”. In: *IEEE Transactions on Cognitive Communications and Networking* 4.3 (2018), pp. 433–445.

[66] Abien Fred Agarap. “Deep learning using rectified linear units (relu)”. In: *arXiv preprint arXiv:1803.08375* (2018).

[67] Nitish Srivastava et al. “Dropout: a simple way to prevent neural networks from overfitting”. In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958.

[68] Timothy J O’shea and Nathan West. “Radio machine learning dataset generation with gnu radio”. In: *Proceedings of the GNU Radio Conference*. Vol. 1. 1. 2016.

[69] International Telecommunication Union. *IMT Traffic Estimates for the Years 2020 to 2030*. 2015.

[70] Dajie Jiang and Guangyi Liu. “An overview of 5G requirements”. In: *5G Mobile Communications* (2017), pp. 3–26.

[71] Mohammed H Alsharif et al. “Sixth generation (6G) wireless networks: Vision, research activities, challenges and potential solutions”. In: *Symmetry* 12.4 (2020), p. 676.

[72] Yulei Wu et al. “A survey of intelligent network slicing management for industrial IoT: integrated approaches for smart transportation, smart energy, and smart factory”. In: *IEEE Communications Surveys & Tutorials* 24.2 (2022), pp. 1175–1211.

[73] Vasileios P Rekkas et al. “Machine Learning in Beyond 5G/6G Networks—State-of-the-Art and Future Trends”. In: *Electronics* 10.22 (2021), p. 2786.

[74] Jasneet Kaur et al. “Machine Learning Techniques for 5G and Beyond”. In: *IEEE Access* 9 (2021), pp. 23472–23488. DOI: 10.1109/ACCESS.2021.3051557.

[75] Qianqian Pan et al. “Leveraging ai and intelligent reflecting surface for energy-efficient communication in 6g iot”. In: *arXiv preprint arXiv:2012.14716* (2020).

[76] Shimaa A Abdel Hakeem, Hanan H Hussein, and HyungWon Kim. “Security Requirements and Challenges of 6G Technologies and Applications”. In: *Sensors* 22.5 (2022), p. 1969.

[77] Yi Shi, Yalin E Sagduyu, and Tugba Erpek. “Federated Learning for Distributed Spectrum Sensing in NextG Communication Networks”. In: *arXiv preprint arXiv:2204.03027* (2022).

[78] Ferhat Ozgur Catak et al. “Defensive Distillation based Adversarial Attacks Mitigation Method for Channel Estimation using Deep Learning Models in Next-Generation Wireless Networks”. In: *arXiv preprint arXiv:2208.10279* (2022).

[79] Jinxin Liu et al. “Adversarial machine learning: A multilayer review of the state-of-the-art and challenges for wireless and mobile systems”. In: *IEEE Communications Surveys & Tutorials* 24.1 (2021), pp. 123–159.

[80] Felix Michels et al. “On the vulnerability of capsule networks to adversarial attacks”. In: *arXiv preprint arXiv:1906.03612* (2019).

[81] Omer Faruk Tuna, Ferhat Ozgur Catak, and M. Taner Eskil. “Exploiting epistemic uncertainty of the deep learning models to generate adversarial samples”. In: *arXiv e-prints*, arXiv:2102.04150 (Feb. 2021), arXiv:2102.04150. arXiv: 2102.04150 [cs.LG].

[82] Yan Jiang et al. “Project Gradient Descent Adversarial Attack against Multi-source Remote Sensing Image Scene Classification”. In: *Security and Communication Networks* 2021 (2021).

[83] Iordanis Fostiropoulos, Basel Shbita, and Myrl Marmarelis. “Robust Defense Against Lp-Norm-Based Attacks by Learning Robust Representations”. In: ().

[84] Samuel G. Finlayson et al. “Adversarial attacks on medical machine learning”. In: *Science* 363.6433 (2019), pp. 1287–1289. DOI: 10.1126/science.aaw4399. URL: https://www.science.org/doi/abs/10.1126/science.aaw4399.

[85] Wei Emma Zhang et al. “Adversarial Attacks on Deep-Learning Models in Natural Language Processing: A Survey”. In: 11.3 (2020). ISSN: 2157-6904. URL: https://doi.org/10.1145/3374217.

[86] Piotr Żelasko et al. “Adversarial Attacks and Defenses for Speech Recognition Systems”. In: *arXiv e-prints*, arXiv:2103.17122 (Mar. 2021), arXiv:2103.17122. arXiv: 2103.17122 [eess.AS].

[87] Jiliang Zhang and Chen Li. “Adversarial examples: Opportunities and challenges”. In: *IEEE transactions on neural networks and learning systems* 31.7 (2019), pp. 2578–2593.

[88] Abdelrahman Taha, Muhammad Alrabeiah, and Ahmed Alkhateeb. “Enabling large intelligent surfaces with compressive sensing and deep learning”. In: *IEEE Access* 9 (2021), pp. 44304–44321.

[89] Ahmed Alkhateeb. “DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications”. In: *arXiv preprint arXiv:1902.06435* (2019).

[90] Nicolas Papernot et al. “Distillation as a defense to adversarial perturbations against deep neural networks”. In: *2016 IEEE symposium on security and privacy (SP)*. IEEE. 2016, pp. 582–597.

[91] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. “Distilling the knowledge in a neural network”. In: *arXiv preprint arXiv:1503.02531* (2015).

[92] Xinchi Chen et al. “Adversarial Multi-Criteria Learning for Chinese Word Segmentation”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. 2017, pp. 1193–1203. URL: http://aclweb.org/anthology/P/P17/P17-1110.pdf.

VITA

Haolin Tang

Research Interests

Next-generation wireless communication, cyber security, computer vision, image processing, and machine learning.

Education

Virginia Commonwealth University — Richmond, VA

Ph.D. in Electrical and Computer Engineering — *May 2019 – Aug 2024*

Yunnan Normal University — Kunming, China

B.S. in Computer Science and Technology — *Sep 2014 – Jun 2018*

Publications

1 **Haolin Tang**, Yang Yang, Kun Yang, Yi Luo, Yaying Zhang and Fangyu Zhang, Non-rigid point set registration with mixed features, *Acta Automatica Sinica*, 2016, 42(11):1732-1743.

2 **Haolin Tang** and Yang Yang, Non-Rigid Point Set Registration with Multiple Feature, *International Conference on Audio, Language and Image Processing*, 07/2016.

3 Yang Yang, Yifan Hu, **Haolin Tang**, Su Zhang and Ling He, Image-Based Biomechanical Relationship Estimation between Maximum Jaw Opening and Masticatory Muscle Activities, *2017 2nd International Conference on Image, Vision and Computing*, 06/2017.

4 Kun Yang, Anning Pan, Yang Yang, Su Zhang, Sim Heng Ong and **Haolin Tang**, Remote Sensing Image Registration Using Multiple Image Features, *Remote Sensing*, 2017, 9(6):581.

5 **Haolin Tang**, Anning Pan, Yang Yang, Kun Yang, Yi Luo, Su Zhang and Sim Heng Ong, Retinal Image Registration Based on Robust Non-rigid Point Matching Method, *Journal of Medical Imaging and Health Informatics*, 2018, 8(2):240-249.

6 **Haolin Tang** and Yanxiao Zhao, A Conditional Generative Adversarial Network for Non-rigid Point Set Registration, *IEEE Asia-Pacific Conference on Computer Science and Data Engineering*, 12/2021.

7 Sarp, Salih, **Haolin Tang**, and Yanxiao Zhao, Use of Intelligent Reflecting Surfaces For and Against Wireless Communication Security, *2021 IEEE 4th 5G World Forum (5GWF)*, 10/2021.

8 **Haolin Tang**, Sarp, Salih, Yanxiao Zhao, Wei Wang and Chunsheng Xin, Security and Threats of Intelligent Reflecting Surface Assisted Wireless Communications, *International Conference on Computer Communications and Networks (ICCCN)*, 07/2022.

9 Ferhat Ozgur Catak, Murat Kuzlu, **Haolin Tang**, Evren Catak, and Yanxiao Zhao, Security Hardening of Intelligent Reflecting Surfaces Against Adversarial Machine Learning Attacks, *IEEE Access*, 2022.

10 **Haolin Tang**, Yanxiao Zhao, Guodong Wang, Changqing Luo and Wei Wang Wireless Signal Denoising Using Conditional Generative Adversarial Networks, *2023 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 05/2023.

11 **Haolin Tang**, Ferhat Ozgur Catak, Evren Catak, Murat Kuzlu and Yanxiao

Zhao, Defending AI-Based Automatic Modulation Recognition Models Against Adversarial Attacks, *IEEE Access*, 2023.

12 **Haolin Tang**, Yanxiao Zhao and Wei Wang Angle of Arrival Based Signal Classification in Intelligent Reflecting Surface-Aided Wireless Communications, *MobiMedia 2022: Mobile Multimedia Communications* , 07/2022.

13 **Haolin Tang**, Yanxiao Zhao, Murat Kuzlu, Changqing Luo, Ferhat Ozgur Catak and Wei Wang, Automatic Modulation Recognition Using Parallel Feature Extraction Architecture, *Wireless Algorithms, Systems, and Applications: 18th International Conference (WASA)*, 06/2024.