Colluding RF Fingerprint Impersonation Attack Based on Generative Adversarial Network

Yuxuan Xu*, Ming Liu*, Linning Peng†, Junqing Zhang‡, Yawen Zheng*
*School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China
†School of Cyber Science and Engineering, Purple Mountain Laboratories, Southeast University, Nanjing, China
‡ Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, United Kingdom
E-mail: mingliu@bjtu.edu.cn, pengln@seu.edu.cn, junqing.zhang@liverpool.ac.uk

***Abstract*—Radio frequency fingerprint (RFF) is an effective way to improve the security of wireless communications. Existing research mainly focused on the classification capability and the robustness of RFFs but overlooked malicious attacks. In this paper, a colluding impersonation attack framework is proposed to emulate the RFF of legitimate users. A colluding attacker is introduced to observe the signal features of the impersonation attacker and the legitimate user and compare their difference. The difference is fed back to the impersonation attacker to help improve its RFF impersonation method. With this idea, the impersonation attack is realized by the Generative Adversarial Network (GAN) structure. The RFF impersonation is formulated as the generator whose objective is to output the signal with RFF similar to the legitimate user, viewed from the colluding attacker's perspective. Simulation results show that the proposed method can effectively impersonate the legitimate user's RFF under the dynamic block fading channel.**

***Index Terms*—Radio frequency fingerprint (RFF), impersonation attack, general adversarial network (GAN), adversarial machine learning**

I. Introduction

Wireless communications face increasing challenges from malicious attackers due to the broadcasting nature of radio propagation. Radio frequency fingerprint (RFF) consists of a series of unique signal features caused by the hardware impairments. It can be exploited to identify the transmitter device and is widely recognized as a promising means for device authentication [1]–[5].

However, like all other authentication approaches, RFF also faces impersonation attacks from adversaries. With the rapid development of software-defined radio, malicious devices have stronger agility and capability to reform their signal characteristics by re-configuring their signal parameters. Whereas, the potential risks that RFF is tampered by the impersonation attack have not drawn broad attention yet. A generative adversarial network (GAN)-based wireless signal spoofing attack is proposed in [6], achieving the signal spoofing with the assumption of channel similarity between a legitimate user and the attacker. Yet, this assumption is not realistic in most application scenarios. More importantly, it is unclear whether the features that the attacker learned are the RFF of the transmitter or the channel characteristics. In addition, the success rate of spoofing attacks is not satisfactory ($\sim$ 76% in the favorable condition). A policy-based reinforcement learning was proposed for the RF fingerprint spoofing attacks in [7]. However, the proposed spoofing attack cannot target a specific device when there exist multiple active legitimate devices. It does not learn the RFF of a legitimate device but it tries to find the weakness of the discriminator to break the RFF-based authentication as any identity of the legitimate device. This attack becomes ineffective when the higher layer authentication measures are also employed, because the RFF impersonated by the attacker may not match the identity encoded in the higher layer authentication protocol.

In the existing work, the spoofed RFFs can only emulate the features of the legitimate users on a randomly generated signal, but cannot send tampered information at the same time [6], [7]. It can only be used as a denial-of-service (DoS) attack to the physical layer authentication but cannot perform identity impersonation attacks. In contrast, this paper investigates the targeted impersonation attack that can learn the RFF associated with the specific transmitter irrespective of the channel condition and can emulate the RFF features with arbitrary information that the signal conveys. The contribution of the paper are summarized as follows.

- A colluding impersonation attack framework based on the GAN structure is proposed so that the attacker can learn the RFF features of the targeted legitimate transmitter.
- With the proposed RFF impersonation, the attacker can send arbitrary information with the RFF features of the targeted transmitter.
- The proposed attack is adapted to the channel variation with a high success rate, which makes it a realistic solution.

The rest part of the paper is organized as follows. The concept of hardware impairments is introduced in Section II. The motivation and a colluding attack strategy is introduced in Section III. A colluding impersonation attack framework is proposed in Section IV. The performance of the proposed impersonation attack is presented in Section V and Section VI concludes the paper.

II. Preliminary: Hardware Impairments

There are inherent hardware impairments within various components of wireless transceivers, including both transmitters and receivers, due to the inevitable variations in the manufacturing process. As modeled in [8], a transmitter is subject to imperfections including oscillator drift, in-phase (I) and

quadrature (Q) imperfection at the mixer, and power amplifier nonlinearity, while receiver impairments involve oscillator drift and IQ imbalance. This paper adopts IQ imbalance at both transmitter and receiver as a case study.

The equivalent baseband signal at transmitter with IQ imbalance can be given as [8], [9]

$$s_{BB}(t) = s_I(t) + j s_Q(t), \tag{1}$$

where

$$\begin{aligned} s_I(t) &= g_I^{tx} x_I(t) \cos(\theta^{tx}) + g_Q^{tx} x_Q(t) \sin(\theta^{tx}), \\ s_Q(t) &= g_I^{tx} x_I(t) \sin(\theta^{tx}) + g_Q^{tx} x_Q(t) \cos(\theta^{tx}), \end{aligned} \tag{2}$$

and $x_I(t)$ ($x_Q(t)$) and g_I^{tx} (g_Q^{tx}) are the modulated signal and gain of the I (Q) branch, respectively, θ^{tx} is half of the phase mismatch between I and Q branches.

At the receiver side, the received baseband signal with IQ imbalance is

$$y(t) = K_1^{rx} h(t) s_{BB}(t) + K_2^{rx} \left(h(t) s_{BB}(t)\right)^*, \tag{3}$$

where $h(t)$ is the channel effect, and

$$K_1^{rx} = \frac{g_I^{rx} e^{-j\theta^{rx}} + g_Q^{rx} e^{j\theta^{rx}}}{2}, \tag{4}$$

$$K_2^{rx} = \frac{g_I^{rx} e^{j\theta^{rx}} - g_Q^{rx} e^{-j\theta^{rx}}}{2}, \tag{5}$$

with g_I^{rx} (g_Q^{rx}) the gain of I (Q) branch and θ^{rx} the half of phase mismatch between I and Q branches at receiver, respectively. The received signal in (3) can be noted as

$$y(t) = \mathcal{R}\big(h(t)\mathcal{T}_k(x(t))\big), \tag{6}$$

where $\mathcal{T}_k(\cdot)$ and $\mathcal{R}(\cdot)$ represent the response of the hardware impairments of the kth transmitter and the receiver, respectively. There hardware impairments are unique and stable hence can be exploited as the RFF of devices.

III. Colluding Impersonation Attack Strategy

A. Motivation

Impersonation attack based on passive eavesdropping and signal replay perfectly replicates the authentication information of the legitimate transmitter and can compromise the traditional password or encryption based authentication mechanisms. In fact, the attacker will inevitably leave its "fingerprint" in the captured and replayed signal due to the hardware impairments of the RF front-end circuits. When the RFF is adopted as the authentication measure, the attacker's fingerprint can be detected by the legitimate receiver's physical layer security mechanism. The impersonation attack can thus be detected and resolved, even though the authentication information carried by the replayed signal is absolutely correct. Therefore, the attacker needs to effectively imitate the legitimate user's RFF in order to achieve a successful impersonation attack.

However, imitating the legitimate user's RFF is challenging. The main reason is that the transmission and reception front-ends of the attacker also have their own hardware impairments. These imperfections will cause additional distortion to the received signal. Therefore, the RFF of the legitimate transmitter observed by the attacker may not be the same as the one observed by the legitimate receiver. In the meantime, it is not easy for the attacker to calibrate its transmission RF circuit. Hence, the attacker can hardly conceal its RFF in the transmitted signal. Apparently, it is difficult for the attacker to perform the impersonation attack to the RFF identification mechanism of the legitimate users.

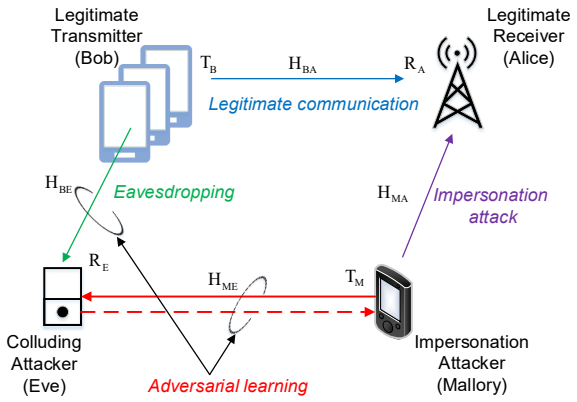


Fig. 1. RFF impersonation attack by a colluding attacker.

B. Colluding Impersonation Attack Strategy

In this work, we propose to introduce a colluding attacker (Eve) to help an impersonation attacker (Mallory) improve its attack quality. More specifically, we consider a wireless communication environment where a legitimate transmitter (Bob) is communicating with a legitimate receiver (Alice), as shown in Fig. 1. A pre-trained deep learning-based identifier is used at Alice to determine whether a signal transmission is from the intended legitimate transmitter.

The attacker Mallory aims to launch an impersonation attack such that its transmissions are identified as the targeted legitimate ones. More importantly, Mallory intends to transmit the tampered information with the RFF features of the targeted legitimate transmitter. Thanks to the RFF carried by the signal, Mallory's transmission can be detected as a malicious transmission by the RFF-based identification. Thus, Mallory needs to learn the unique signal features of Bob and tune its own signal characteristics accordingly to enable the impersonation attack. To this end, Mallory performs some delicate distortion to its baseband signal. This distortion modifies the RFF feature that the signal carries.

A colluding attacker (Eve) fully works with Mallory to help improve the way that Mallory disguises its RFF. Eve observes the disguised signal sent by Mallory, and compares it to the legitimate signal coming from Bob. Eve evaluates the similarity between the two signals using the same principle as the way that Alice judges the legitimacy of the signal's RFF. Then Eve sends the difference to Mallory. With the feedback, Mallory knows better how to improve the way it disguises its baseband signal so that the RFF features shown in the RF signal look like Bob's. Through an iterative process, Mallory can gradually achieve a

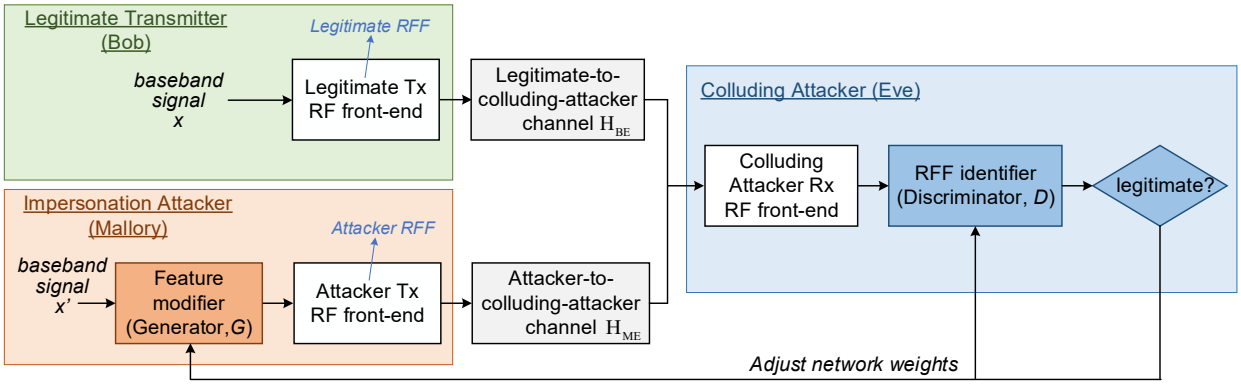


Fig. 2. Colluding RFF impersonation attack framework based on GAN.

high RFF similarity to Bob. As Eve evaluates the RFF similarity from an objective perspective, Alice will probably accept the spoofed RFF once it passes the similarity evaluation of Eve.

In fact, the colluding impersonation attack strategy mentioned above coincides with the adversarial learning process of GAN where Eve and Mallory exchange information to help Mallory disguise its RFF. The detailed attack procedure will be elaborated in the following section. It is worth noting that the proposed colluding attack scheme does not restrict the locations of different parties, nor the similarity among the channels, as other existing work did [6], [7]. The four channels (e.g. H_{BA}, H_{BE}, H_{ME} and H_{MA}) in our attack model are assumed to be independent.

IV. GAN-based Colluding Impersonation Attack Framework And Implementation

In this section, we propose a GAN-based impersonation attack framework that implements the colluding attack presented in the previous section, as shown in Fig. 2. In contrast to the existing RFF spoofing attack techniques where the main target is to emulate the RFF features [6], [7], this work aims to accomplish a more sophisticated attack that can send arbitrary messages with the RFF features of a targeted legitimate user. With this basis, a real impersonate attack can be carried out once the higher layer authentication measure is stolen or cracked.

A. Framework Design

Mallory uses a generator network (G) to modify the baseband signal X' and sends the spoofed signal to Eve. Since Mallory and Eve are fully cooperated, Eve knows perfectly whether a signal is from Mallory or Bob. Eve uses this knowledge as the ground truth label to train the discriminator network (D) which determines the received signal as "true" or "false". If the disguised RFF is similar enough to that of Bob, Eve cannot detect that the signal is from an attacker. This similarity suggests that Alice cannot detect it either. The identification results are then fed back to Mallory. Mallory will update the weights of the generator network according to the gradients' feedback from Eve.

The generator and discriminator networks are trained in a min-max manner. Mallory optimizes G to fool D at Eve. The discriminator, on the other hand, optimizes D to discover the RFF disguised by G. The training of G and D are carried out iteratively until the convergence. When the training converges, G in Mallory can transform its signal features into Bob's type, and can be used for the impersonation attack.

It is worth noting that GAN is an unsupervised learning approach, and depends on the similarity between true and generated samples by nature. The GAN-based attack framework breaks through the limitation of existing attack algorithms that require a large amount of a priori information including the structure and weights of the network, and can be applied to various attack scenarios.

B. Proposed GAN Structure

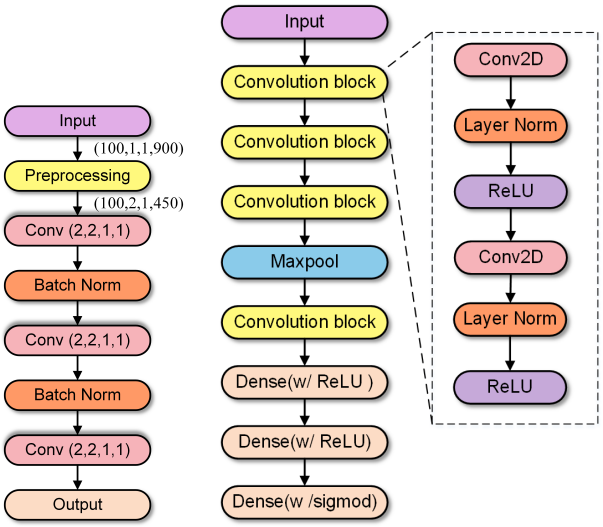
We employ an improved Wasserstein GAN with gradient penalty (WGAN-GP) [10] to the targeted impersonate attack. WGAN-GP alleviates the mode collapse problem of the original GAN and can diversify the generated results to be adaptive to the more general transmission conditions.

1) Generator Network: We use a generator network for tuning the signal's fingerprint features (i.e., the "Feature modifier" module in Fig. 2). According to (1) and (2), most RFF features can be expressed in a generic form as

$$s_{BB}(t) = Ax_I(t) + Bx_Q(t) + j\left(Cx_I(t) + Dx_Q(t)\right), \quad (7)$$

where A, B, C, and D are the coefficients. The structure of the generator G is illustrated in Fig. 3(a). The attacker performs a targeted impersonation attack that learns the RFF associated with a specific transmitter and can emulate the RFF characteristics with the signal carrying arbitrary information. We use the convolutional neural network to realize G to modify the waveform characteristics of the signal while maintaining the physical meaning that the signal conveys. A layer of preprocessing is employed in G, which divides the complex-valued input data into its real part $x_I(t)$ and imaginary part $x_Q(t)$, forming a dual-channel[1] aligned data. These two data channels

[1]To be distinguished from the sense in signal transmission, a "channel" in neural network is referred to a separate data path in the forward pass.



(a) Generator (G). (b) Discriminator (D).

Fig. 3. Proposed WGAN-GP network for the impersonate attack. The parameters in the parentheses (a, b, c, d) indicate *the number of input channels*, *the number of output channels*, *kernel size*, and *stride*, respectively.

are fed into the following 1-dimensional (1-D) convolutional layer for the signal feature transformation as shown in (7). This expression of G is generic and can represent most of the fingerprint features.

Adding multiple convolutional layers can transfer information hierarchically. That is to say, by deepening the network, the problems to be learned at each layer can be decomposed into simple problems that are easy to solve. The batch norm layer is added to improve the speed of model training, and can effectively avoid the disappearance and explosion of gradients. In the meantime, the generalization ability is improved to avoid overfitting in the training. In addition, as all the signal and fine-tuning distortion are centrosymmetric with respect to the origin, the bias of all convolution layers is set to zero to ensure the output still being centrosymmetric and improve the stability in the network training.

2) Discriminator Network: The structure of the discriminator is shown in Fig. 3(b). Increasing the network depth can effectively reduce the impact of noise and interference [11]. The batch normalization layer is not used in the discriminator because it affects the calculation of gradient penalty and leads to inaccurate gradient update [10]. In contrast, layer normalization is used in the discriminator. The discriminator network only needs to identify whether the signal is emitted by the target legitimate transmitter or the attacker. Hence, a binary classification will be applied. This is different from the identification network at the legitimate receiver, which needs to classify multiple legitimate transmitters. We reduce the size of the last few fully connected layers in the discriminator (D) to prevent overfitting. The activation function is also changed to Sigmoid in the last layer for discriminator (D).

TABLE I
IQ IMBALANCE FOR ALL DEVICES. $g_Q = 1$

Device	$(g_I, 2\theta)$	Device	$(g_I, 2\theta)$
legitimate Tx 0	$(-0.3, -15°)$	legitimate Tx 5	$(-0.2, -15°)$
legitimate Tx 1	$(-0.1, 5°)$	legitimate Tx 6	$(0.2, 10°)$
legitimate Tx 2	$(0.3, 5°)$	legitimate Tx 7	$(0.1, 15°)$
legitimate Tx 3	$(0.1, 10°)$	legitimate Tx 8	$(0.2, 5°)$
legitimate Tx 4	$(0.3, -5°)$	legitimate Tx 9	$(0.2, -10°)$
legitimate Rx	$(-0.1, -5°)$		
attacker	$(0.3, -15°)$	colluding attacker	$(-0.3, 15°)$

V. PERFORMANCE EVALUATION

A. Simulation Setup

1) Device Configuration: There are ten legitimate transmitters, one legitimate receiver, one impersonation attacker and a colluding attacker. As discussed in Section II, all the devices are impaired by IQ imbalance. Their configurations are given in Table I. The parameters are chosen according to the model in [12]. Without loss of generality, set $g_Q = 1$ for all cases.

The RFF identification network at the legitimate receiver has the same structure as given in Fig. 3(b) but with different number of classes. This paper considers the colluding attacker only knows the structure of the identification network of legitimate receivers, but not the network weights. The impersonation and colluding attackers are configured with the generative network and the discriminator, respectively, as shown in Fig. 3.

2) Channel Configuration: This paper considers a block fading Rayleigh channel. A single tap channel is assumed. Specifically, we consider two types of channels.

- Dynamic Channel I: This channel model assumes perfect time and frequency synchronization. The channel coefficient can be given as

$$h = a \cdot e^{j\theta_h}, \tag{8}$$

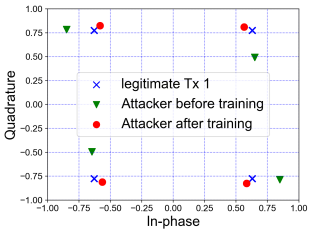
  where a is the real-valued channel amplitude, θ_h is the phase offset caused by the channel fading.
- Dynamic Channel II: This channel model further involves timing error and frequency jitter. The channel coefficient is written as

$$h = a \cdot e^{j(\theta_h + \theta_T + \Delta f t)}, \tag{9}$$

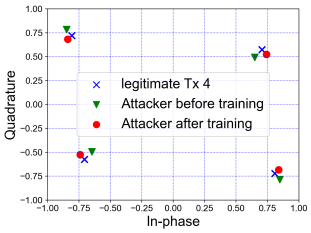
  where the timing error θ_T is a random phase offset uniformly distributed in $(-10°, 10°)$ and the frequency jitter Δf is randomly chosen from a Gaussian distribution with zero mean and standard deviation of 1 KHz.

3) Dataset Description: For all the legitimate transmitters and the impersonation attacker, the training dataset is generated as follows:

- Each device generates a packet of 450 complex RF symbols.
- Every 100 packets form a data segment, which share a common channel realization.
- Each device generates 100 data segments.

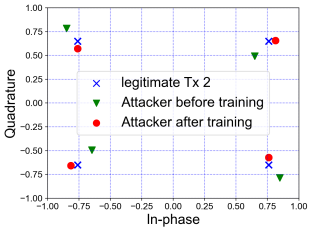


(a) Target on legitimate Tx 1.

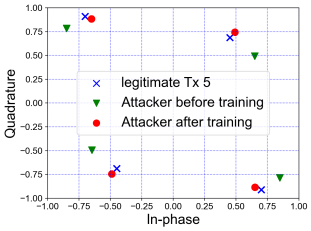


(b) Target on legitimate Tx 4.

Fig. 4. Spoofed signal received by Alice using the network trained in dynamic channel I.



(a) Target on legitimate Tx 2.



(b) Target on legitimate Tx 5.

Fig. 5. Spoofed signal received by Alice using the network trained in dynamic channel II.

With the block fading assumption, the channel parameters remain the same within one data segment and vary among segments.

There are two neural network models, namely the identification model of the legitimate receiver and the WGAN-GP model.

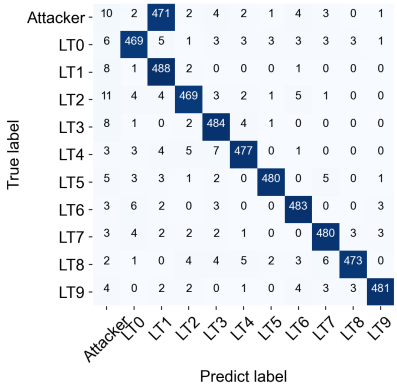
- Identification model of the legitimate receiver: The training dataset consists of 110,000 samples for ten legitimate transmitters and the attacker.
- WGAN-GP model: Depending on the legitimate transmitter that the attacker aims to impersonate, the training dataset consists of 20,000 samples from the transmitted data of that particular transmitter and the attacker.

For both datasets, the ratio of training set to validation set is 8:2. The data from the training set is randomly selected in batch and put into the network for training. Different neural network models are trained depending on the channel model and the transmitter to be impersonated.

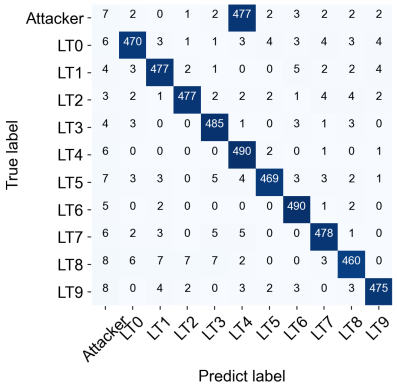
There are another 5,500 new samples generated for testing with 500 samples for each device. The attack model is independent of the modulation method (here QPSK is used as an example). The four channels H_{BA}, H_{BE}, H_{ME}, and H_{MA} in our considered attack scenario are mutually independent and follow two types of distribution in the evaluation.

B. Simulation Results and Analysis

Figures 4 and 5 present the constellation of the spoofed signal received by Alice using the network models trained in the dynamic channel I and dynamic channel II, respectively. For the sake of presentation clarity, the channel distortion is not involved in the examples shown in the figures. As can



(a) Target on legitimate transmitter 1.

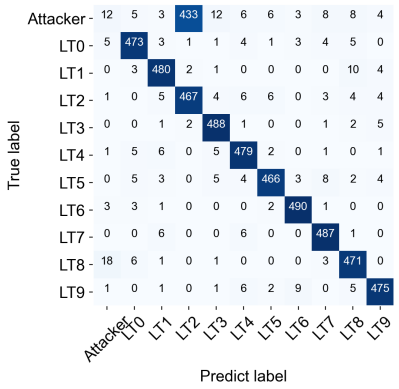


(b) Target on legitimate transmitter 4.

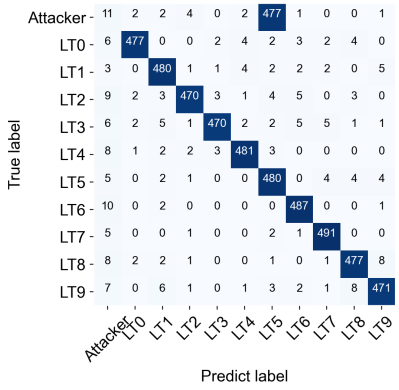
Fig. 6. Identification results under impersonation attack in dynamic channel I.

be observed, the constellation points of the attacker Mallory become much closer to the targeted legitimate transmitter. This means even though Mallory does not have any prior information on the RFF of the legitimate transmitters, it can adjust its signal distribution to approach the characteristics of the targeted transmitter. The signal of Mallory is independent of Bob, which means that Mallory can send arbitrary information with the RFF features of the targeted legitimate transmitter by employing the proposed WGAN-GP. This provides Mallory the ability to perform the RFF impersonation attack.

Then, we evaluate the effectiveness of the proposed attack through the possibility that Mallory's spoofed signal can fool Alice's RFF identification measures. We randomly choose the legitimate Tx 1 and 4 as the targets of impersonation attack in dynamic channel I and Tx 2 and 5 as the targets in dynamic channel II. The identification network of the legitimate receiver is trained with separate data samples and can achieve an identification accuracy of 99.7% under dynamic channel I and 98.0% under dynamic channel II. This network will be used by the legitimate receiver Alice to identify legitimate users and discover attacker. The results of the identification network under impersonation attack are given in Fig. 6 and Fig. 7. From the figures, it can be observed that using the proposed WGAN-GP, Mallory can adjust its signal to possess the RFF features of specific legitimate transmitters under different channel condi-



(a) Target on legitimate transmitter 2.



(b) Target on legitimate transmitter 5.

Fig. 7. Identification results under impersonation attack in dynamic channel II.

tions and is recognized as the targeted legitimate transmitters by the RFF identification mechanism in most of the cases. This suggests that the attack is very likely to be successful.

We calculate the *attack success rate* which is defined as the ratio of the number of signal that are misjudged as the target legitimate to the number of signal that are sent by the attacker. The results over all legitimate users and under different channel conditions are shown in Fig. 8. Though facing different channel conditions, the proposed WGAN-GP-based impersonation attack is effective to spoof the RFF features of all the legitimate transmitters considered in the evaluation. The attack success rate is at least 91.3% in dynamic channel I and 93.3% in dynamic channel II. The slight difference in the success rate is caused by the fact that the RFF identification network is trained independently with the data samples under the two channel conditions. The dynamic channel II introduces more uncertainty in terms of timing error and frequency jitter. Therefore, the acceptable signal spread under dynamic channel II is more significant than dynamic channel I, which leaves more room for the attacker to "sneak in".

VI. Conclusion

In this paper, we proposed a GAN-based colluding attack framework to realize the targeted RFF impersonation attack. The proposed attack mechanism can learn the RFF features

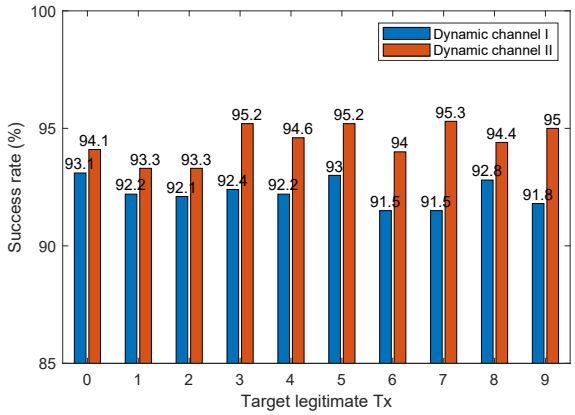


Fig. 8. Impersonation attack success rate under different channel conditions.

associated with the specific transmitter, irrespective of the channel condition or the information that the signal conveys. A WGAN-GP based network was proposed to realize the attack accordingly. The attacker can adapt to more flexible channel conditions, and send arbitrary information with the RFF features of the legitimate transmitter. Simulation results show that the spoofed signal can fool RFF identification measures at a high success rate in varying channel conditions, which suggests the effectiveness of the proposed impersonation attack in realistic scenarios.

References

[1] S. Riyaz, K. Sankhe, S. Ioannidis, *et al.* "'Deep learning convolutional neural networks for radio identification," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 146–-152, 2018.

[2] L. Peng, J. Zhang, M. Liu, and A. Hu, "Deep learning based RF fingerprint identification using differential constellation trace figure," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 1091–1095, 2020.

[3] P. Liu, P. Yang, W.-Z. Song, *et al.* "Real-time identification of rogue WiFi connections using environment-independent physical features," in *Proc. IEEE INFOCOM*, 2019, pp. 190–-198.

[4] M. Ramasubramanian, C. Banerjee, D. Roy, *et al.* "Exploiting spatio-temporal properties of I/Q signal data using 3D convolution for RF transmitter identification," *IEEE J. Radio Freq. Identif.*, vol. 5, no. 2, pp. 113–127, 2021.

[5] G. Shen, J. Zhang, A. Marshall, *et al.* "Radio frequency fingerprint identification for LoRa using spectrogram and CNN," in *Proc. IEEE INFOCOM*, 2021, pp. 1–10.

[6] Y. Shi, K. Davaslioglu, and Y. E. Sagduyu, "Generative adversarial network for wireless signal spoofing," in *Proc. ACM WiseML*, 2019, pp. 55–60.

[7] S. Karunaratne, E. Krijestorac, and D. Cabric, "Penetrating RF fingerprinting-based authentication with a generative adversarial attack," in *Proc. IEEE ICC*, 2021, pp. 1-6.

[8] J. Zhang, R. Woods, M. Sandell, *et al.* "Radio frequency fingerprint identification for narrowband systems, modelling and classification," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 3974-3987, 2021.

[9] B. He and F. Wang, "Cooperative specific emitter identification via multiple distorted receivers," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3791-3806, 2020.

[10] I. Gulrajani, F. Ahmed, M. Arjovsky, *et al.* "Improved training of Wasserstein GANs," *arXiv preprint arXiv:1704.00028*, 2017.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[12] M. Cekic, S. Gopalakrishnan, and U. Madhow, "Wireless fingerprinting via deep learning: The impact of confounding factors," *arXiv preprint-arXiv:2002.10791*, 2020.