Explainable Adversarial Learning Framework on Physical Layer Secret Keys Combating Malicious Reconfigurable Intelligent Surface

Zhuangkun Wei, Wenxiu Hu, Weisi Guo

***Abstract*—The development of reconfigurable intelligent surfaces (RIS) is a double-edged sword to physical layer security (PLS). Whilst a legitimate RIS can yield beneficial impacts including increased channel randomness to enhance physical layer secret key generation (PL-SKG), malicious RIS can poison legitimate channels and crack most of existing PL-SKGs. In this work, we propose an adversarial learning framework between legitimate parties (namely Alice and Bob) to address this Man-in-the-middle malicious RIS (MITM-RIS) eavesdropping. First, the theoretical mutual information gap between legitimate pairs and MITM-RIS is deduced. Then, Alice and Bob leverage generative adversarial networks (GANs) to learn to achieve a common feature surface that does not have mutual information overlap with MITM-RIS. Next, we aid signal processing interpretation of black-box neural networks by using a symbolic explainable AI (xAI) representation. These symbolic terms of dominant neurons aid feature engineering-based validation and future design of PLS common feature space. Simulation results show that our proposed GAN-based and symbolic-based PL-SKGs can achieve high key agreement rates between legitimate users, and is even resistant to MITM-RIS Eve with the knowledge of legitimate feature generation (NNs or formulas). This therefore paves the way to secure wireless communications with untrusted reflective devices in future 6G.**

***Index Terms*—Adversarial Learning, Symbolic representation, Man-in-the-middle attack, Reconfigurable intelligent surface, Physical layer secret key.**

I. Introduction

Wireless communication systems are susceptible to various attack vectors due to their broadcasting nature. Traditional cryptography leverages the advantages in complexities of designed mathematical problems [1], which, however, renders it less attractive in securing wireless communications for future lightweight and massive Internet-of-Everything (IoE) devices in 6G. In the last 10 years, the concept and techniques of physical layer security (PLS) have been proposed and widely studied to secure the legitimate wireless channels.

A. Literature Review

From research perspectives, PLS are categorized as key-less PLS and physical layer secret key generation (PL-SKG).

This work is supported by the Engineering and Physical Sciences Research Council [grant number: EP/V026763/1].

Zhuangkun Wei, Weisi Guo is with the School of Aerospace, Transport, and Manufacturing, Cranfield University, MK43 0AL, UK.
Wenxiu Hu is with the Optoelectronics Research Centre, University of Southampton, SO17 1BJ, UK.
Corresponding author: zhuangkun.wei@cranfield.ac.uk

1) Key-Less PLS: Key-less PLS aims to maintain the superiority of legitimate channels by minimizing information leakage to potential eavesdroppers (Eves). The methods include the optimizations of beamforming vector [2], the anti-jamming artificial noise [3], the trajectory of autonomous systems [4], the spin modulation, etc. The challenge lies in the lack of guarantee of a feasible solution, especially when combined with real-world constraints from other layers (e.g., control and mission).

2) PL-SKG: Another category is PL-SKG, whereby legitimate Alice and Bob leverage the reciprocal and random channel state information (CSI) between them as common features to generate symmetrical secret keys [5]–[11]. The exploited reciprocal CSIs in existing works include the received signal strength (RSS) [6], the channel phases [12], and the channel frequency response [13]. In these cases, Alice and Bob are required to send public pilot sequences to each other and pursue channel estimations to acquire these common CSI, which will then be passed to the key quantization [14], [15], key reconciliation [16], and privacy amplification [17] modules for key generation.

To further improve the secret key rate (SKR), one-way random signal-based [18]–[20] and two-way random signal-based [21]–[24] PL-SKG have been proposed. The designs include the popular two-way cross-multiplication PL-SKG, where Alice and Bob send random pilot sequences to each other and crossly multiply their sent and received signals as the common feature for further standard key generation process. In this view, the randomness of common features (i.e., the entropy in terms of bit per sample) not only involves the random CSI but is enhanced by Alice's and Bob's sent random signal, thereby improving the SKR.

3) When PLS meets RIS: Reconfigurable intelligent surface (RIS) has been recently proposed and designed to reconfigure the wireless channels to improve communication quality of services (QoS) [25]–[28]. In key-less PLS, RIS phase provides a new degree-of-freedom (DoF) to minimize the information leakage [29], [30]. For PL-SKG, RIS can be used to increase channel randomness by randomly assigning the RIS phase [31]–[34], which enables a fast generation of the physical layer secret keys. Based on this idea, the work in [31] computes the theoretical SKR of RIS-secured low-entropy channel, and the work in [35], [36] further designs an optimal RIS phase set by maximizing the theoretical SKR.

The advance of RIS also introduces new attacking and eavesdropping threats. This can be categorized as attackers

whether to destroy or to maintain the channel reciprocity between Alice and Bob. The first category aims to ruin the reciprocal channel-based PL-SKG [37]. The second category tries to intercept Alice's and Bob's messages and reconstruct their physical layer secret keys. This physical layer man-in-the-middle (MITM) attack was first proposed by [38] and validated via hardware experiments in [39]. Then, our previous work in [40] further proposes two MITM-RIS Eve schemes to crack the exiting CSI-based and two-way randomness-based PL-SKGs. What is worse, MITM-RIS is hardly to be detected or countermeasured by standard authentication methods (e.g., hash-based signatures), given the inability of its reflective elements to actively send pilots for protocol [41], [42] and authentication purposes.

B. Contributions & Paper Structure

In this work, we aim to propose an explainable adversarial learning-based physical layer common feature generators for Alice and Bob, to generate symmetrical secret keys that are resistant to the MITM-RIS Eve. Indeed, several machine learning-based PL-SKGs have been designed recently [43], including deep neural network (NN) [44], [45], long short-term memory (LSTM) [46], auto-encoder [47], [48], etc. However, none of them considers the harsh physical layer MITM eavesdropping, and all of them result in black-box NNs that lack interpretation and explicit formulas to present trustworthiness. The detailed contributions are provided in the following.

(1) We deduce a theoretical mutual information gap between legitimate Alice-Bob and MITM-RIS Eve. This serves as the theory of the existence of a common feature space that cannot be reconstructed via the leaked signals to MITM-RIS.

(2) Based on the deduced mutual information gap, we design a generative adversarial network (GAN) based framework, for Alice and Bob to learn to reach the common feature space. Here, Alice and Bob are assigned two NNs as feature generators, and an NN of an assumed MITM-RIS Eve is set as the discriminator. Cross-correlation is used in the generator's and discriminator's loss functions to measure the similarity of features as well as maintain the randomness.

(3) To explain the trained common feature generator NNs, we deploy Meijer G function-based symbolic metamodelling to identify the dominant special terms of dominant neurons. The results then lead to the finding of one common feature space, and the designs of explicit formulas-based common feature generators. As such, we provide a more interpretable and transparent approach for generating physical layer common features and the secret keys relied upon.

(4) We evaluate our proposed GAN-based and explicit formula-based PL-SKG. The results show high key agreement rates between Alice and Bob, and that even the MITM-RIS Eve with the full knowledge of Alice's and Bob's common feature generator (NNs or formulas) still cannot reconstruct the generated physical layer secret keys. This promising outcome signifies a significant step toward establishing secure wireless communications in the context of untrusted reflective devices, laying the groundwork for future 6G cybersecurity.

The rest of this work is structured as follows. In Section II, we describe channel models and the MITM-RIS Eve threats. In Section III, we elaborate on our design of GAN-based common feature generators, and the symbolic metamodelling process to derive the explicit formula-based common feature generators. In Section VI, we show our simulation results. We finally conclude this work in Section V.

In this work, we use bold lower-case letters for vectors. We use $\|\cdot\|_2$ to denote the 2-norm, and $diag(\cdot)$ to diagonalize a vector. $|\cdot|$ and $(\cdot)^*$ represent the absolute value and the conjugate of a complex value. We denote $mod(\cdot,\cdot)$ as the modulus operator and $\lfloor\cdot\rfloor$ is to truncate the argument. The matrix transpose is denoted as $(\cdot)^T$. $\mathbb{E}(\cdot)$ and $\mathbb{D}(\cdot)$ represent the expectation and variance. $\mathcal{CN}(\cdot,\cdot)$ is to represent the complex Gaussian distribution with mean and variance.

II. System Model & Problem Formulation

In this work, three nodes, namely Alice, Bob and an untrusted RIS are considered, where two legitimate users Alice and Bob are to generate secret keys leveraging channel reciprocity. Here, we omit the direct channel between Alice and Bob, but assume the worst-case that only the channels between Alice to RIS, and Bob to RIS are existed. As such, the untrusted RIS, equipped with RF chains [49], can pursue the MITM to reconstruct the legitimate common features and secret keys of Alice and Bob [40].

A. Channel and Signal Model

The legitimate users, Alice and Bob are considered as single-antenna. The untrusted RIS is modelled as a uniform planar array (UPA) of size $M = M_x \times M_y$ (see Fig. 1(a)). The direct channels from Alice to RIS and Bob to RIS, denoted as $\mathbf{g}_{aR} \in \mathbb{C}^{M\times 1}$ $a \in \{A, B\}$, are modelled as [50]:

$$\begin{aligned}\mathbf{g}_{aR} &= \sum_{n=1}^{L} \frac{\rho_{aR,n}}{\sqrt{\iota}} \cdot \mathbf{u}\left(\alpha_{aR,n}, \beta_{aR,n}\right) \\ \rho_{aR,n} &\sim \mathcal{CN}\left(0, C_0 \cdot d_{aR}^{-\eta}\right)\end{aligned} \tag{1}$$

In Eq. (1), C_0 is the path loss at the reference distance (i.e., $1m$). d_{aR} is the line-of-sight (LoS) distance between a to RIS, and η is the exponential non-line-of-sight (NLoS) loss factor. L is the number of NLoS Rayleigh paths and $\rho_{aR,n}$ is the gain for nth path. $\mathbf{u}(\alpha,\beta) \triangleq [\exp(j\mathbf{a}(\alpha,\beta)\mathbf{l}_1), \cdots, \exp(j\mathbf{a}(\alpha,\beta)\mathbf{l}_M)]^T$, with $\mathbf{a}(\alpha,\beta) \triangleq \frac{2\pi}{\lambda}[\sin(\alpha)\cos(\beta), \sin(\alpha)\sin(\beta), \cos(\alpha)]$ and $\mathbf{l}_m \triangleq [0, mod(m-1, M_x)d, \lfloor (m-1)/M_y \rfloor d]^T$. $\alpha_{aR,n}, \beta_{aR,n} \in [-\pi/2, \pi/2]$ are the half-space elevation and azimuth angles of nth path. For the structure of RIS, a square shape element is used with the size as $d \times d$, where $d = \lambda/8$ is set (i.e., less than half-wavelength $\lambda/2$ [50]–[52]).

With the modelling of the direct channels to RIS, the combined channels from Bob to Alice (Alice to Bob), denoted as g_{BA} (g_{AB}), are expressed as:

$$g = g_{AB} = g_{BA} = \mathbf{g}_{BR}^T \cdot diag(\mathbf{w}) \cdot \mathbf{g}_{AR}, \tag{2}$$

In Eq. (2), $\mathbf{w} = [\exp(j\vartheta_1), \cdots, \exp(j\vartheta_M)]^T$ is the phase vector of the RIS, where $\vartheta_m \in [0, 2\pi)$ with $m \in \{1, \cdots, M\}$ is the phase of mth RIS element.

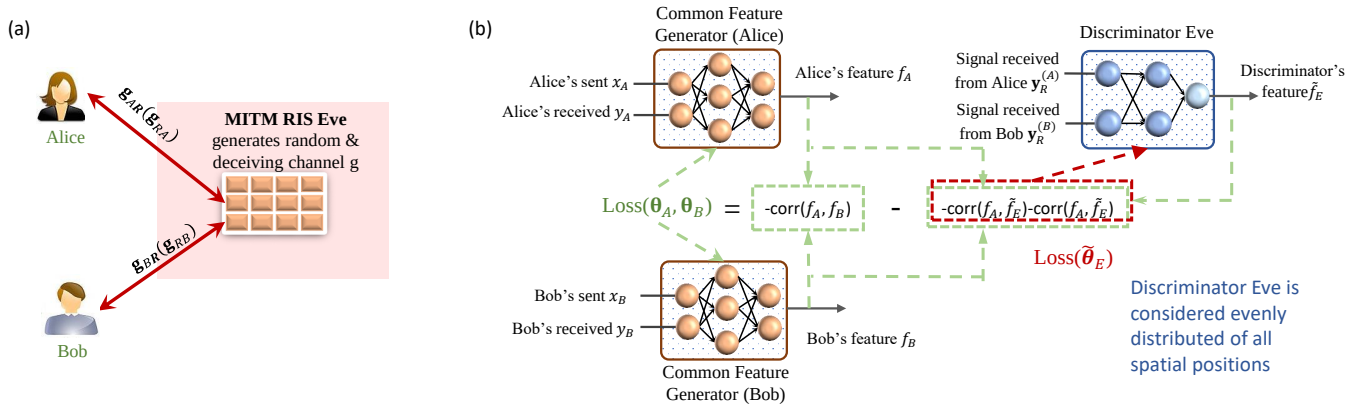

Fig. 1. Sketch of adversarial learning framework against MITM-RIS Eve: (a) MITM-RIS Eve that cracks the existing PL-SKGs; (b) the GAN-based structure with Alice's and Bob's common feature generator NNs and a discriminator Eve.

B. Man-In-The-Middle RIS Eavesdropping

In our previous work [40], two MITM-RIS Eve schemes are proposed. This designed RIS Eve creates a reciprocal random channel via manipulating its RIS phase $\vartheta_m \sim [0, 2\pi)$, and inserts it into Alice and Bob. This therefore enables RIS Eve to reconstruct the legitimate common features using reciprocal CSI or two-way cross-multiplication-based PL-SKGs.

To be specific, for both PL-SKGs, Alice and Bob are required to send signals in time-division duplex (TDD) mode, where the CSI during the two consecutive time slots are reciprocal [31]–[34]. In this context, we simplify the notations by reducing the transmitted/received time series to single time step scalars in the description of the PL-SKG process, where the vector version can be found in [40]. We denote Alice's and Bob's sent signals as $x_A, x_B \in \{\sqrt{E_t}\exp(j\phi)|\phi \in [0, 2\pi]\}$, and their received signals as:

$$\begin{aligned} y_A =& g_{BA} \cdot x_B + n_A \\ y_B =& g_{AB} \cdot x_A + n_B, \end{aligned} \tag{3}$$

with n_A, n_B the received noise.

1) Eavesdropping CSI-based Secret Keys: In CSI-based PL-SKG, Alice and Bob send pre-defined pilot sequences and estimate the reciprocal CSI in two consecutive time slots. Their estimated CSIs are [31]–[34]:

$$\begin{aligned} \hat{h}_A &= y_A \cdot x_B^*/E_t \approx g_{BA}, \\ \hat{h}_B &= y_B \cdot x_A^*/E_t \approx g_{AB}. \end{aligned} \tag{4}$$

As such, leveraging the reciprocal CSIs as the common features, i.e., $\hat{h}_A \approx \hat{h}_B$, secret keys can be generated at Alice and Bob individually, via the standard key quantization [14], [15], reconciliation [16] and privacy amplification [17].

MITM-RIS Eve aims at recovering the reciprocal CSIs of g_{AB}, g_{BA}. To create the worst-case scenario, we assume that all RIS's elements are equipped with RF chains for receiving and base-band channel estimation. This extreme setting provides a more accurate channel estimation and therefore serves as a more powerful Eve than the design in [40], where a sparse RF chain placement and compressed sensing method is used. As such, RIS will receive signals from Alice and Bob at two consecutive time slots, i.e.,

$$\begin{aligned} \mathbf{y}_R^{(A)} &= \mathbf{g}_{AR} \cdot x_A + \mathbf{n}_R^{(A)}, \\ \mathbf{y}_R^{(B)} &= \mathbf{g}_{BR} \cdot x_B + \mathbf{n}_R^{(B)}. \end{aligned} \tag{5}$$

This therefore can be used to estimate the channels of Alice-RIS and Bob-RIS as:

$$\begin{aligned} \hat{\mathbf{g}}_{AR} &= \mathbf{y}_R^{(A)} \cdot x_A^*/E_t \approx \mathbf{g}_{AR} \\ \hat{\mathbf{g}}_{BR} &= \mathbf{y}_R^{(B)} \cdot x_B^*/E_t \approx \mathbf{g}_{BR}. \end{aligned} \tag{6}$$

Then, MITM-RIS Eve can reconstruct the legitimate CSI by taking the estimated results into Eq. (2). The legitimate secret key using this CSI is therefore cracked by this MITM-RIS.

2) Eavesdropping Two-Way Cross-Multiplication-based Secret Keys: In two-way cross-multiplication PL-SKG, Alice and Bob send random sequences in two consecutive time slots, and multiply their sent and received signals as common features, i.e., [23], [53]

$$\begin{aligned} \phi_A &= x_B \cdot y_B \approx g \cdot x_B \cdot x_A \\ \phi_B &= x_A \cdot y_A \approx g \cdot x_B \cdot x_A \end{aligned} \tag{7}$$

Then, secret keys can be generated at Alice and Bob via their common features $\phi_A \approx \phi_B$.

MITM-RIS Eve here aims to reconstruct the cross-multiplication common features. Similar to eavesdropping the CSI-based case, we consider the worst-case scenarios where all RIS's elements are equipped with RF chains. According to our previous work [40], the legitimate common features can be reconstructed by directly multiplying the received signals from Alice and Bob, i.e.,

$$\phi_E = \left(\mathbf{y}_R^{(B)}\right)^T \cdot diag(\mathbf{w}) \cdot \mathbf{y}_R^{(A)} \approx g \cdot x_B \cdot x_A. \tag{8}$$

As such, MITM-RIS Eve can then obtain the legitimate secret keys generated by two-way cross-multiplication features.

C. Aim of this Work

As CSI-based and two-way cross-multiplication-based PL-SKGs are compromised by the existing MITM-RIS Eve, the rest of this work aims to design the adversarial learning-based PL-SKG where the MITM-RIS cannot reconstruct the common features and the secret keys relied upon.

III. Adversarial Learning based PL-SKG

In this section, we will elaborate on (i) our design of GAN-based physical layer common feature generators, and (ii) Meijer-G function-based symbolic metamodeling for explicit formula-based common feature generators. In the following part, the feature construction leverages the aforementioned two-way signals, where Alice and Bob will send random pilot sequences in TDD mode, and then leverage their sent and received signals to generate common features.

The theory underlies the existence of common feature space that is resistant to MITM-Eve comes from the positive mutual information difference between Alice's and Bob's two-way signals and MITM-RIS Eve received signals, which reads:

$$I\left(\begin{bmatrix}x_A\\y_A\end{bmatrix},\begin{bmatrix}x_B\\y_B\end{bmatrix}\right) - \max\left\{I\left(\begin{bmatrix}x_A\\y_A\end{bmatrix},\begin{bmatrix}\mathbf{y}_R^{(A)}\\\mathbf{y}_R^{(B)}\\\mathbf{w}\end{bmatrix}\right), I\left(\begin{bmatrix}x_B\\y_B\end{bmatrix},\begin{bmatrix}\mathbf{y}_R^{(A)}\\\mathbf{y}_R^{(B)}\\\mathbf{w}\end{bmatrix}\right)\right\} > 0. \tag{9}$$

Here, we omit all the receiving noises, as we aim to show the pure mutual information gap between legitimate nodes and MITM Eve. A detailed explanation is provided in Appendix A.

A. GAN-based Common Feature Generators

Leveraging the theory showing the existence of the common feature space that is resistant to MITM Eve, the GAN-based framework is designed, which includes Alice's and Bob's common feature generators and a discriminator MITM-RIS Eve that is designed to cover all the spatial positions between Alice and Bob.

1) Legitimate Feature Generator NNs: As is shown in Fig. 1(b), Alice and Bob deploy two fully connected NNs as the common feature generators, denoted as $\Psi_{\boldsymbol{\theta}_A}(\cdot)$ and $\Psi_{\boldsymbol{\theta}_B}(\cdot)$ with $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$ the parameters. Here, the two NNs have the same structures. The input layer has 4 neurons, which are the real and imaginary of one's sent and received signals (i.e., x_A, y_A for Alice, and x_B, y_B for Bob). Each NN contains 3 hidden layers with 1024, 512, and 128 neurons respectively. The activation function is ReLU. Given the complex form of wireless channels and signals, the output layer is designed as 2 neurons. These neurons correspond to the real and imaginary components, and the angle they form contributes to the final constructed features. As such, the feature generator NNs of Alice and Bob are expressed as:

$$\begin{aligned} f_A &= \Psi_{\boldsymbol{\theta_A}}\Big(\mathrm{Re}\{x_A\},\mathrm{Im}\{x_A\},\mathrm{Re}\{y_A\},\mathrm{Im}\{y_A\}\Big)\\ f_B &= \Psi_{\boldsymbol{\theta_B}}\Big(\mathrm{Re}\{x_B\},\mathrm{Im}\{x_B\},\mathrm{Re}\{y_B\},\mathrm{Im}\{y_B\}\Big) \end{aligned} \tag{10}$$

2) Discriminator Eve NN: To prevent MITM-RIS from reconstructing the legitimate features, Alice and Bob further assume a discriminator MITM-RIS Eve NN. The inputs are the simulated RIS's (real and imaginary parts of) phase and received signals from Alice and Bob, i.e., $\tilde{\mathbf{g}}_{AR}\cdot x_A$ and $\tilde{\mathbf{g}}_{BR}\cdot x_B$, and the RIS phase $\tilde{\mathbf{w}}$. Then, 3 hidden layers with 1024, 512, and 128 neurons are assigned, with ReLU activation function. Similar to Alice's and Bob's NNs, the output layer contains 2 neurons to represent the real and imaginary components, which form the angles as the constructed feature. By denoting the parameters of the discriminator NN as $\tilde{\theta}_E$, its output is expressed as:

$$\tilde{f}_E = \Psi_{\tilde{\boldsymbol{\theta}}_E}\Big(\mathrm{Re}\{\tilde{\mathbf{g}}_{AR}x_A\},\mathrm{Im}\{\tilde{\mathbf{g}}_{AR}x_A\},\mathrm{Re}\{\tilde{\mathbf{g}}_{BR}x_B\},\mathrm{Im}\{\tilde{\mathbf{g}}_{BR}x_B\},\mathrm{Re}\{\tilde{\mathbf{w}}\},\mathrm{Im}\{\tilde{\mathbf{w}}\}\Big) \tag{11}$$

It is noteworthy that this assumed discriminator MITM-RIS Eve NN is not the real MITM-RIS Eve. The training data for this discriminator should encompass all the spatial positions of Alice and Bob to this assumed RIS and cover all the channel distributions. This approach ensures the incorporation of potential Man-in-the-Middle (MITM) scenarios in real-world situations.

3) Data & Loss Function: For the data, Alice, Bob, and RIS are modelling following the description in Section II-A. From Eq. (1), the distribution of channels between one legitimate user to RIS is fully determined by the elevation and azimuth angles and the LoS distance. In this view, to encompass comprehensive scenarios, we evenly split (i) the angle half-space $[-\pi/2, \pi/2]$ by 100, and (ii) the distance from 0m to 200m by 1000. Then, a total of $N = 10^6$ pairs of $(\tilde{\mathbf{g}}_{AR}, \tilde{\mathbf{g}}_{BR}, \tilde{\mathbf{w}}, x_A, x_B, y_A, y_B)$ is sampled as the training data from the channel and signal models.

In the training stage, the batch size is assigned as $K = 64$ and the Adam optimizer is utilized with a learning rate of 10^{-5}. The loss functions of generators and the discriminator have opposite objectives. The loss function of the discriminator Eve NN is:

$$Loss(\tilde{\boldsymbol{\theta}}_E) = -|corr(f_A, \tilde{f}_E)| - |corr(f_B, \tilde{f}_E)|. \tag{12}$$

The aim is to let discriminator Eve learn the legitimate features. This is designed by maximizing the absolute correlation coefficients between Alice's (Bob's) and discriminator Eve's features in one batch.

The loss function of Alice and Bob contains (i) the aim to maximize the absolute correlation coefficient between Alice's and Bob's features among one batch, and (ii) the aim that the discriminator Eve cannot reconstruct the legitimate features from their received signals, i.e.,

$$\begin{aligned} Loss(\boldsymbol{\theta}_A, \boldsymbol{\theta}_B) = &- |corr(f_A, f_B)| \\ &+ \lambda \cdot \Big[|corr(f_A, \tilde{f}_E)| + |corr(f_B, \tilde{f}_E)|\Big]. \end{aligned} \tag{13}$$

where $\lambda = 0.8$ is the assigned coefficient.

It is noteworthy that, in these loss functions, we characterize the commonality of features via correlation coefficients other than the basic mean squared error (MSE). The reasons are two-fold. First, MSE only accounts for the difference between two features, but is unable to characterize the mutual information. For example, if Alice's and Bob's features are with all 1 elements, their MSE will approach 0 but their features lack

Algorithm 1: Training of Adversarial Feature Construction NNs

Input: number of Training data N, learning rate $r = 10^{-5}$, episode number $E_{\max} = 10^4$, batch size $B = 64$

1 Training data: Sampling a total of N tuples of $(\tilde{\mathbf{g}}_{AR}, \tilde{\mathbf{g}}_{BR}, \tilde{\mathbf{w}}, g_{AB}, x_A, x_B, y_A, y_B)$ according to Eqs. (1)-(3), where the elevation and azimuth angles and LoS distances are evenly selected given the split grids of angle and distance spaces;

2 Initialize Alice's and Bob's feature generator NNs and discriminator Eve NN as $\Psi_{\boldsymbol{\theta}_A}(\cdot)$, $\Psi_{\boldsymbol{\theta}_B}(\cdot)$, $\Psi_{\tilde{\boldsymbol{\theta}}_E}(\cdot)$;

3 **for** *episode* $= 1, \cdots, E_{max}$ **do**

4 Sampling a batch $B = 64$ of training data ;

5 **for** *each* $i = 1, \cdots, B$ *batch* **do**

6 Compute $f_A[i]$ and $f_B[i]$ via Eq. (10);

7 Compute $\tilde{f}_E[i]$ via Eq. (11);

8 Update Discriminator Eve NN by $Loss(\tilde{\boldsymbol{\theta}}_E)$ in Eq. (12);

9 Update Alice's and Bob's feature generator NNs by $Loss(\boldsymbol{\theta}_A, \boldsymbol{\theta}_B)$ in Eq. (13);

10 **end**

11 **end**

Output: Alice's and Bob's feature generator NNs $\Psi_{\boldsymbol{\theta}_A}(\cdot)$ and $\Psi_{\boldsymbol{\theta}_B}(\cdot)$.

randomness and mutual information. In contrast, the correlation coefficient expressed by:

$$corr(X,Y) = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\cdot\mathbb{E}(Y)}{\sqrt{\mathbb{D}(X)\cdot\mathbb{D}(Y)}}, \tag{14}$$

involves not only the similarity (in the numerator) but also the variances (in the denominator) to consider the randomness, which thereby serves as the better way to characterize the (first-order) mutual information. Secondly, achieving the MSE criterion is possible by directly scaling down the outputs of f_A and f_B. However, it's important to note that whilst the MSE may decrease, e.g., $MSE(0.1\cdot f_A, 0.1\cdot f_B) < MSE(f_A, f_B)$, the mutual information remains unchanged, i.e., $I(0.1\cdot f_A, 0.1\cdot f_B) = I(f_A, f_B)$.

4) Algorithm Flow of Training Process: The overall training process is detailed in Algorithm 1. Given the number of training data, learning rate, maximum episode number and batch size, step 1 is to generate the training data for Alice's and Bob's generator NNs and the discriminator Eve NN. Step 2 is to initialize these 3 NNs. The training starts from steps 3 to 11, by sampling a batch of training data in step 4, computing the outputs of NNs in steps 5-6, and updating the discriminator NN (step 8) and the generator NNs (step 9) respectively. The outputs are the trained Alice's and Bob's feature generator NNs which can (i) generate common features from the two-way signals, and (ii) prevent MITM-RIS Eve from reconstructing the common features via RIS's inserted channels (we will further test this in Section IV and Simulations).

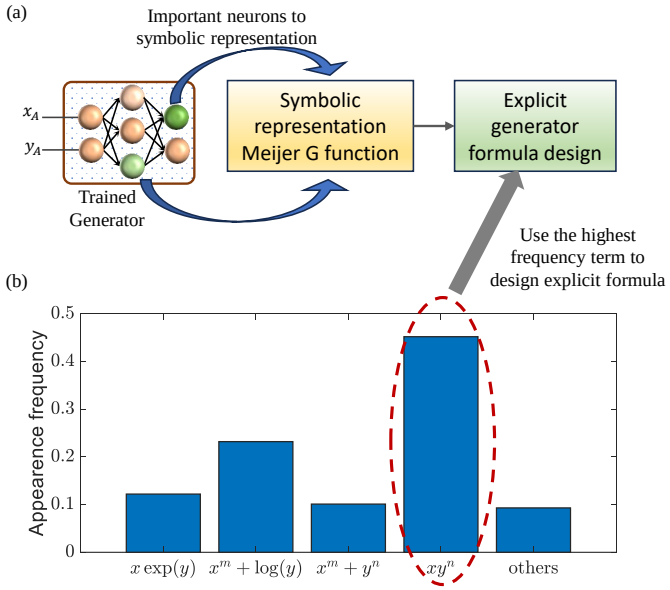

Fig. 2. Symbolic representation of NN for explicit formula design: (a) using Meijer G function for symbolic representation of dominant neurons of trained legitimate generator NNs; (b) shows the appearance frequency of special terms derived from fitted Meijer G functions, which then are used to design explicit formula-based common feature generators as Eqs. (19)-(21)

B. Explainability to Explicit Feature Generator Formula

Explaining black-box neural networks is important to understand and validate the work, to interpret findings of key signal features to researchers and engineers, and to draw lessons, as outlined in our previous vision for xAI for 6G [54].

After the training of feature generator NNs, Alice and Bob can generate binary secret keys individually via their generated common features. However, the opaque black-box NNs may not be able to present trustworthiness in both academic and real applications. To address this, we rely on Meijer G function to represent our trained common feature generator NNs. Then, based on the representation results, we deduce a concise and explicit formula of common feature generator, serving as the white-box to defend the MITM-RIS eavesdropping. The schematic flow is shown in Fig. 2(a).

1) Symbolic Representation via Meijer G Functions: In essence, the Meijer G-function is a family of univariate functions, each of which corresponds to a linear combination of certain special functions (e.g., $\exp(x)$, $1/(1+x)$, $\log_2(x)$, etc.). The equation of Meijer G function is expressed as [55], [56]:

$$\begin{aligned} &G_{p,q}^{m,n}\left(\begin{matrix}a_1,\cdots,a_p\\ b_1,\cdots,b_q\end{matrix}\middle| c\cdot x\right)\\ =&\frac{1}{2\pi j}\int_{\mathcal{L}}\frac{\prod_{i=1}^{m}\Gamma(b_i-s)\prod_{i=1}^{n}\Gamma(1-a_i+s)}{\prod_{i=m+1}^{q}\Gamma(1-b_i+s)\prod_{i=n+1}^{p}\Gamma(a_i-s)}(c\cdot x)^s ds, \end{aligned} \tag{15}$$

where $0 \le m \le q, 0 \le n \le p$ are predefined integer numbers, $a_1 \cdots a_p$ and $b_1, \cdots, b_q$ are continuous real parameters with relations $a_k - b_i \notin \mathbb{Z}^+$ for $k = 1, \cdots, n$ and $i = 1, \cdots, m$ and $x \neq 0$. Here $\Gamma(\cdot)$ is Gamma function extended on complex variable s. The integral path $\mathcal{L}$ is from $-i\infty$ to $i\infty$ separating the poles of the factors $b_i - s$ from those of factors $1 - a_k + s$

[57]. Some of the equivalent form of Meijer G functions are [56]:

$$\begin{aligned} &G_{3,1}^{0,1}\left(_{1}^{2,2,2}|c\cdot x\right) \equiv c\cdot x \\ &G_{0,1}^{1,0}\left(_{0}^{-}|c\cdot x\right) \equiv \exp(cx) \\ &G_{1,2}^{2,2}\left(_{1,0}^{1,1}|c\cdot x\right) \equiv \log(1+cx). \end{aligned} \tag{16}$$

When fitting an NN, the Kolmogorov superposition theorem [58] states that every multivaruate continuous functions can be written as a finite two-layer composition of univariate continuous functions. As such, by stacking the input of Alice's feature generator NN as $\mathbf{x} = [x_1, x_2, x_3, x_4] = [\text{Re}\{x_A\}, \text{Im}\{x_A\}, \text{Re}\{y_A\}, \text{Im}\{y_A\}]$, we create our Meijer G function-based representation framework as:

$$G(\mathbf{x}, \boldsymbol{\theta}_G) = \sum_{l=0}^{1} G_{1,l}\left(\boldsymbol{\theta}_{1,l}\Big|\sum_{i=0}^{3} G_{0,i}(\boldsymbol{\theta}_{0,i}|x_i)\right). \tag{17}$$

In Eq. (17), the function has 2 layers, where the first layers are 4 univariate Meijer G functions, i.e., $G_{0,i} \triangleq G_{2,2}^{1,2}(\boldsymbol{\theta}_{0,i}|x_i)$, and the second layers are 2 univariate Meijer G functions, i.e., $G_{1,l} \triangleq G_{0,1}^{1,0}(\boldsymbol{\theta}_{1,l}|\cdot)$ with the inputs as the summation of first layer outputs. Here, $\boldsymbol{\theta}_G = [\boldsymbol{\theta}_{1,0}, \boldsymbol{\theta}_{1,1}, \boldsymbol{\theta}_{0,0}, \boldsymbol{\theta}_{0,1}, \boldsymbol{\theta}_{0,2}, \boldsymbol{\theta}_{0,3}]$ are the parameters that use to fit the Alice's feature generator NN. As such, the loss function is designed as:

$$Loss_G(\boldsymbol{\theta}_G) = \left\| G(\mathbf{x}, \boldsymbol{\theta}_G) - f_A^{(m,n)}(\mathbf{x}) \right\|_2^2, \tag{18}$$

where $f_A^{(m,n)}(\mathbf{x})$ represents the output of the nth neurons of mth layer. It is worth noting that there are many potential fittings in the function space [59].

To implement the Meijer G function-based fitting, we first select 50 neurons from 2nd and 3rd hidden layers of Alice's trained common feature generator NN, whose activated values are larger than the threshold $thr = 1$ in the feature generation process with test data. Then, each selected neurons is fitted using the representative framework in Eq. (17) and the loss function in Eq. (18). Fig. 2(b) shows the frequency of appearance of some special forms (e.g., xy^n, $x\log y$, etc.). It is seen that the percentages of the form xy^n is obviously higher than others, which hints its importance in preserving the mutual information of Alice and Bob and countering the MITM Eve. Inspired by this, we design an explicit formula of common feature generators in the next part.

2) Explicit Formula of Common Feature Generator: Given the result of Fig. 2(b), we simplify and break down the learned common feature generator NNs to only embrace the linear combinations of forms $x^m y^n$. Here, the term $x^m y^n$ is simplified by the logarithm terms as $\log(1+x^m) + \log(1+y^n) = \log((1+x^m)(1+y^n)) \approx x^m + y^n + x^m y^n$, where the approximation holds as the transmitted power of Alice and Bob are set less than 1W. As such, we define $\mathbf{z}_A$ by the following expression:

$$\begin{aligned} \mathbf{z}_a = [z_a, z_a^n, \log(1+z_a^n)]^T,\ n \in \{2,3,4\}, \\ a \in \{Alice, Bob\} \end{aligned} \tag{19}$$

where z_a is traversing the set, each composed by the inputs of common feature generator NN, i.e., $\forall z_a \in \{\text{Re}\{x_a\}, \text{Im}\{x_a\}, \text{Re}\{y_a\}, \text{Im}\{y_a\}\}$. Then, we solve the least square problem to obtain the linear coefficients $\boldsymbol{\varrho}$, i.e.,

$$\boldsymbol{\varrho} = \text{argmin} \left\| f_A - \boldsymbol{\varrho}^T \cdot \mathbf{z}_A \right\|_2^2. \tag{20}$$

Following the derivation of linear coefficients $\boldsymbol{\varrho}$, the explicit formula-based common features can be generated at Alice and Bob as:

$$\begin{aligned} \Upsilon_A &= \boldsymbol{\varrho}^T \cdot \mathbf{z}_A \\ \Upsilon_B &= \boldsymbol{\varrho}^T \cdot \mathbf{z}_B \end{aligned} \tag{21}$$

C. NN-based MITM-RIS Eve

In this work, we test our designed PL-SKG in the face of a strong NN-based MITM-RIS Eve. We consider the worst scenarios where this MITM-RIS Eve knows the trained common feature generator NNs, or the explicit form in Eq. (21). Here, with the knowledge of Alice's and Bob's feature generator NNs, Eve cannot directly reconstruct their generated common features, as Eve does not know the exact sent and received signals of Alice and Bob. However, via the MITM-RIS, Eve can intercept the signals from Alice and Bob, which thereby poses the threat of recovering the legitimate common features via a well-trained NN. In this part, we elaborate on the build-up of such Eve NN, which will then be used in the simulation part to test the security of our designed PL-SKG.

It is also noteworthy that this MITM-RIS Eve is conceptually different from the discriminator Eve in our adversarial learning framework, as the latter is employed by legitimate Alice and Bob to ensure their features cannot be estimated by potential Eves.

The structure of MITM-RIS Eve NN consists of one input layer, 3 hidden layers, and an output layer. The input is RIS Eve's received signals from Alice and Bob, and the RIS phase vector. To further strengthen this Eve, the numbers of neurons of 3 hidden layers are assigned as 2048 512 and 128 respectively, where the activation functions are ReLU. The output layer is designed according to the legitimate feature generator NNs. Here 2 neurons correspond to the real and imaginary components, and the angle they form is the final Eve's reconstructed features, i.e.,

$$\begin{aligned} f_E = \Psi_{\boldsymbol{\theta}_E}\Big(&\text{Re}\{\mathbf{y}_R^{(A)}\}, \text{Im}\{\mathbf{y}_R^{(A)}\}, \text{Re}\{\mathbf{y}_R^{(B)}\}, \\ &\text{Im}\{\mathbf{y}_R^{(B)}\}, \text{Re}\{\mathbf{w}\}, \text{Im}\{\mathbf{w}\}\Big) \end{aligned} \tag{22}$$

where $\boldsymbol{\theta}_E$ is the parameters to be learnt.

The training data is generated by the simulations of this MITM-RIS Eve, where an assumed Alice and Bob are modelled with split angle half-space $[-\pi/2, \pi/2]$ by 100 and distance from 0m to 200m by 1000, according to the model description in Section II-A. As this MITM-RIS Eve knows Alice's and Bob's trained feature generator NNs, it is able to generate Alice's and Bob's common features f_A and f_B given the simulated sent and received signals of Alice and Bob. As such, to train this Eve NN to generate the same features of Alice and Bob, the loss function is designed as:

$$Loss_E(\boldsymbol{\theta}_E) = -|corr(f_E, f_A)| - |corr(f_E, f_B)| \tag{23}$$

which is minimized by the Adam optimizer with a learning rate of 10^{-5}. We will show in the Simulation section that

even the MITM-RIS Eve with exact knowledge of legitimate feature generator NNs, it is still impossible to reconstruct the legitimate common features for secret key generation.

IV. Simulation Results

In this section, we evaluate our trained and designed legitimate feature generators, in the face of the trained MITM-RIS Eve NN. The model configuration is provided in the following. In a 3D space, the MITM-RIS is located at $(0, 0, 0)$ with unit m, and the positions of Alice and Bob are within the 100m from RIS. The direct channels from Alice and Bob to MITM-RIS are modelled in Eq. (1) according to [50], where a square structure of RIS is considered with $M = 100$ elements. Here, the referenced path loss is set as $C_0 = -30$dB at the reference distance (i.e., 1m), and the NLoS path loss exponents are $\alpha_N = 3$. The number of multi-paths is $L = 5$ [60] with random half-space elevation and azimuth angles independently and randomly distributed over $\mathcal{U}[-\pi/2, \pi/2]$. The power of sent signals from Alice and Bob is set as $E_t = 0.1$W.

A. Performance of designed GAN-based Common Feature Generators

We first show the training and testing performance of the designed GAN-based framework. It is shown in Fig. 3 that with the increase of the training epoch, the loss function of legitimate feature generator NNs converges very fast (within 2000 epochs). According to Eq. (13), this loss function contains to (i) maximize the absolute correlation coefficient between Alice's and Bob's features, and (ii) minimize the absolute correlation coefficient between Alice's (Bob's) and discriminator Eve's features. These are shown to be achieved in Fig. 4, where the absolute correlation coefficient between Alice's and Bob's features converges to 1 and that between Alice's (Bob's) and discriminator Eve's features converges to 0. It is noteworthy that during the training stage, some spikes appear in both Fig. 3-4 (e.g., from 2000-4000 epochs). This is due to the adversarial characteristics to jump out of the local optima area to better satisfy both positive and discriminator aims in the generators' loss function. Such spikes then gradually disappear as the increase of epoch, which indicates the finding of the stable optimal area. As such, by combining the results from Figs. 3-4, we demonstrate the successful training of the designed adversarial learning framework.

We next show the training and testing results of the NN-based MITM-RIS Eve, which is described in Section III-C. Here, we evaluate the performances of MITM-RIS Eve NN to learn the common features generated from both trained legitimate generator NNs and explicit formula-based generators in Eqs. (19)-(21). In Fig. 5, it is seen that the NN of MITM-RIS Eve cannot converge to a minimal value when trying to learn either the legitimate generator NN-based or the explicit formula-based legitimate common feature generators. For example, these Eve NNs are trained for 1.5×10^6 epochs but are still stuck at very high loss values (-0.24 and -0.3 compared with the beginning value around -0.2).

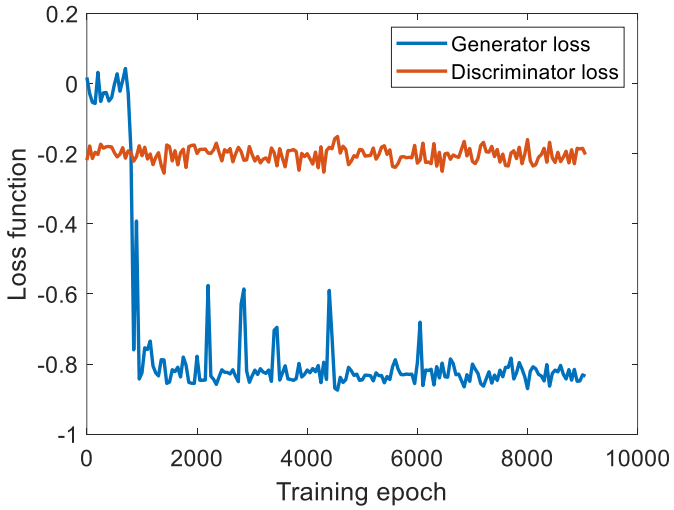

Fig. 3. Converge of loss functions under training epochs of GAN-based design of legitimate common feature generator NNs and discriminator NN.

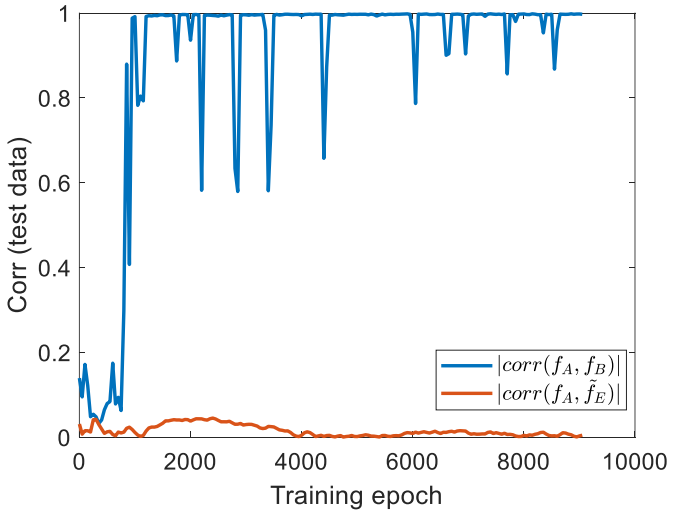

Fig. 4. Converge (to 1) of absolute correlation coefficients of common features generated by Alice's and Bob's feature generator NNs.

Then, in Fig. 6, we show the results of absolute correlation coefficients between Eve's feature and Alice's features (with generator NN or explicit formula), i.e., $|corr(f_A, f_E)|$ and $|corr(\Upsilon_A, f_E)|$. It is seen that even with 1.5×10^6 epochs, NN-based Eve still cannot learn the legitimate features, as the absolute correlation coefficients never exceed 0.1. Combined the results from Figs. 5-6, we demonstrate the capability of the proposed PL-SKG to defend the deep learning-based MITM-RIS Eve.

B. Performance of proposed PL-SKG with Noise

In this part, we evaluate our proposed PL-SKG under different levels of the receiving noise. Here, for a fair comparison, the receiving signal-to-noise ratio (SNR) of Alice, Bob, and MITM-RIS Eve are assigned as same from 5dB to 30dB. In Fig. 7, the absolute correlation coefficients of features between (i) Alice-Bob with feature generator NNs, and (ii)

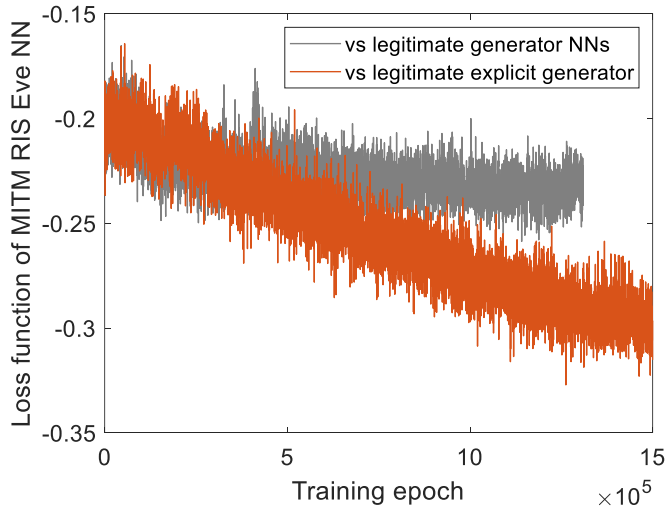

Fig. 5. Training performance of NN-based MITM RIS Eve

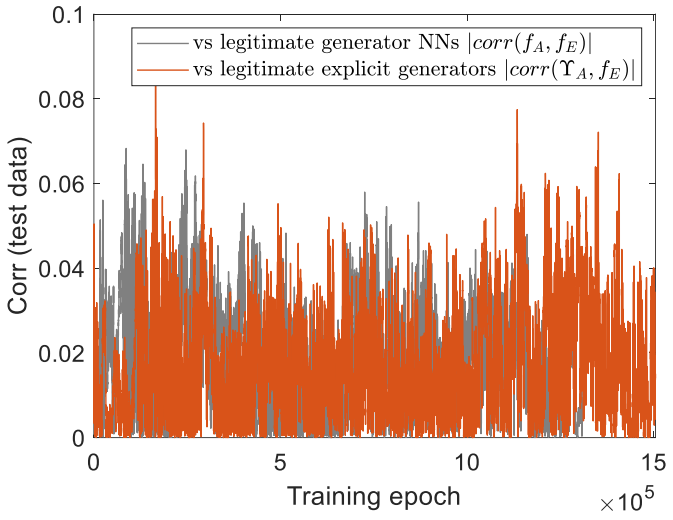

Fig. 6. The absolute correlation coefficients of features generated by MITM-RIS Eve NN and Alice during the training stage.

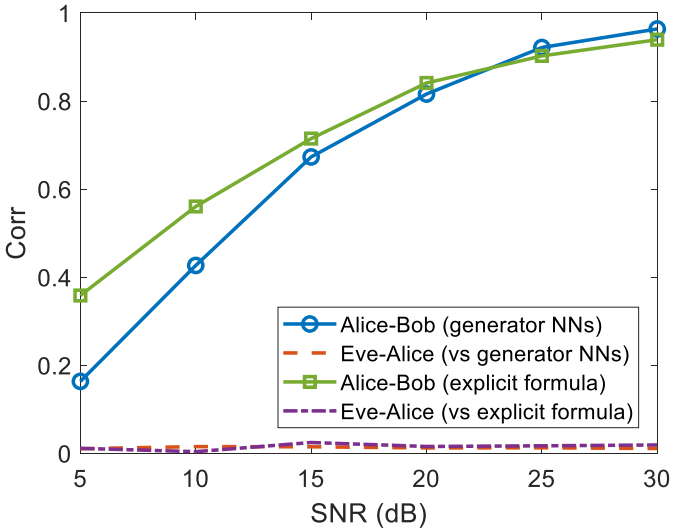

Fig. 7. The absolute correlation coefficients of features between Alice-Bob, and Alice-Eve, under different levels of receiving SNR.

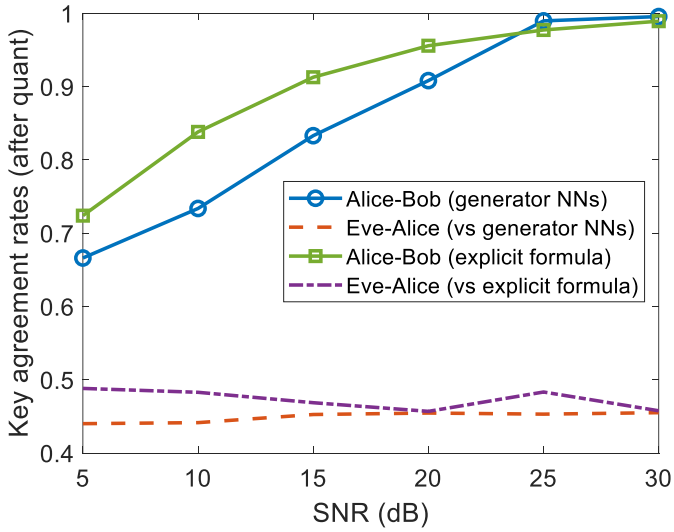

Fig. 8. Key agreement rate after key quantization from Alice's and Bob's common features.

Alice-Bob with explicit formula-based feature generator are provided. These correlation coefficients increase (from 0.2 to 0.95) with the growth of the receiving SNR (from 5dB to 30dB), as higher SNRs provide less noisy two-way signals for legitimate common feature generation. Then, it is seen that NN-based feature generators yield less correlated Alice-Bob features in the low SNR area (e.g., SNR<15dB), while exhibiting higher correlations in the high SNR region (e.g, SNR>20dB), as opposed to the explicit formula-based method. This indicates that NN-based feature generators can learn the latent and nonlinear dependency from Alice's and Bob's two-way signals, but are sensitive to noisy inputs, given the cascaded compounded coupling of nonlinear effects in a deep NN structure. In contrast, the feature generator leveraging the explicit formula, as presented in Eqs. (19)-(21), while conceding a slight decrease in performance in high SNR regions, provides a more interpretable and transparent approach for generating physical layer common features and secret keys relied upon. This attribute is of significance in the establishment of a reliable and trustworthy framework for securing wireless communications.

We finally provide the secret key agreement rates under different levels of receiving SNR. Note that in this paper, the aim is to provide physical layer common feature generation against MITM-RIS Eve. These common features further require the standard key quantization, reconciliation, and privacy amplification to generate secret keys, which are not within the scope and are provided in Appendix B. In Fig. 8, the key agreement rates between Alice and Bob and between Alice and Eve are provided, which corresponds to the correlation coefficient results in Fig. 7. As such, combining the illustrations in Fig. 7-8, our proposed PL-SKG shows a promising capability in countermeasuring the untrusted RIS, which therefore renders a cybersecurity approach in securing wireless communications

for future 6G.

V. Conclusion

In future 6G, the integrated sensing and communication scenarios require the amounts of RISs and passive sensors. These devices raises new security issues, which (i) are hard to be authenticated due to their inabilities to actively send authenticated messages, and (ii) pose the threat to pursue Man-in-the-middle physical layer eavesdropping. Existing physical layer secret key generations, either directly using reciprocal CSI, or leveraging the cross-multiplication of two-way signals (one's sent and received), have been shown to be easily countermeasures by such MITM-RIS Eve [40].

This work proposed the adversarial learning framework to address the security issue raised by this MITM-RIS Eve. The proposed approach leveraged GAN to generate common features through the two-way signals involving reciprocal CSIs. By designing the adversarial loss function to embrace both the correlations between Alice's and Bob's generator outputs, as well as correlations between Alice's outputs and a discriminator Eve NN, Alice and Bob are able to learn common feature generators to prevent MITM-RIS from reconstructing common features. Furthermore, the use of Meijer G function-based symbolic representation facilitated the design of explicit formula-based legitimate common feature generators. Simulation results demonstrate the efficacy of the proposed GAN-based and symbolic-based PL-SKGs, showcasing a high key agreement rate between legitimate users. Notably, these solutions exhibit resilience against MITM-RIS Eve, even when possessing knowledge of legitimate feature generation NNs. This promising outcome signifies a significant step toward establishing secure wireless communications in the context of untrusted reflective devices, laying the groundwork for future 6G networks.

Appendix A
Deduction of noise-released version of Eq. (9)

To simplify the notation, we denote the reciprocal channel between Alice and Bob as $g = g_{AB} = g_{BA}$. Here, we omit all receiving noises in Eq. (9), and have $y_A = gx_B$, $y_B = gx_A$, $\mathbf{y}_R^{(A)} = \mathbf{g}_{AR}x_A$, and $\mathbf{y}_R^{(B)} = \mathbf{g}_{BR}x_B$. In the following, we show the proof of the first part of Eq. (9), i.e.,

$$I\left(\begin{bmatrix}x_A\\y_A\end{bmatrix},\begin{bmatrix}x_B\\y_B\end{bmatrix}\right) - I\left(\begin{bmatrix}x_A\\y_A\end{bmatrix},\begin{bmatrix}\mathbf{g}_{AR}x_A\\\mathbf{g}_{BR}x_B\\\mathbf{w}\end{bmatrix}\right) > 0, \quad (24)$$

which is equivalent by replacing x_A, y_A by x_B, y_B.

The first mutual information term of Eq. (24) can be expressed in terms of the entropy $h(x) = -\int p(x)\log_2 p(x)dx$, i.e.,

$$I\left(\begin{bmatrix}x_A\\y_A\end{bmatrix},\begin{bmatrix}x_B\\y_B\end{bmatrix}\right) = h\left(\begin{bmatrix}x_A\\y_A\end{bmatrix}\right) + h\left(\begin{bmatrix}x_B\\y_B\end{bmatrix}\right) - h\left(\begin{bmatrix}x_A\\y_A\\x_B\\y_B\end{bmatrix}\right) \quad (25)$$

In Eq. (25), given the independent pairs of $(x_A,\ y_A = g\cdot x_B)$ and $(x_B,\ y_B = g\cdot x_A)$, we have

$$h\left(\begin{bmatrix}x_A\\y_A\end{bmatrix}\right) = h\left(\begin{bmatrix}x_A\\g\cdot x_B\end{bmatrix}\right) = h(x_A) + h(g\cdot x_B) \quad (26)$$

$$h\left(\begin{bmatrix}x_B\\y_B\end{bmatrix}\right) = h\left(\begin{bmatrix}x_B\\g\cdot x_B\end{bmatrix}\right) = h(x_B) + h(g\cdot x_A) \quad (27)$$

Also, note that the joint probability density function (PDF) of x_A, y_A, x_B, y_B can be re-written as the joint PDF of g, x_A, x_B,

$$\begin{aligned} &p(x_A, y_A, x_B, y_B) = p(x_A, gx_B, x_B, gx_A)\\ =&p(gx_B, gx_A|x_A, x_B)\cdot p(x_A, x_B)\\ =&p(g)\cdot p(x_A, x_B)\\ =&p(g, x_A, x_B)\end{aligned} \quad (28)$$

where the last equation is because the independence between g, x_A, x_B. This helps to simplify the last term in Eq. (25) as:

$$h\left(\begin{bmatrix}x_A\\y_A\\x_B\\y_B\end{bmatrix}\right) = h\left(\begin{bmatrix}x_A\\x_B\\g\end{bmatrix}\right) = h(x_A) + h(x_B) + h(g) \quad (29)$$

As such, by taking Eqs. (26)-(29) back into Eq. (25), the mutual information of Alice's and Bob's sent and received pairs is:

$$I\left(\begin{bmatrix}x_A\\y_A\end{bmatrix},\begin{bmatrix}x_B\\y_B\end{bmatrix}\right) = h(g\cdot x_A) + h(g\cdot x_B) - h(g). \quad (30)$$

The second mutual information term of Eq. (24) is expressed by entropy as:

$$I\left(\begin{bmatrix}x_A\\y_A\end{bmatrix},\begin{bmatrix}\mathbf{g}_{AR}x_A\\\mathbf{g}_{BR}x_B\\\mathbf{w}\end{bmatrix}\right) = h\left(\begin{bmatrix}x_A\\y_A\end{bmatrix}\right) + h\left(\begin{bmatrix}\mathbf{g}_{AR}x_A\\\mathbf{g}_{BR}x_B\\\mathbf{w}\end{bmatrix}\right) - h\left(\begin{bmatrix}x_A\\y_A\\\mathbf{g}_{AR}x_A\\\mathbf{g}_{BR}x_B\\\mathbf{w}\end{bmatrix}\right). \quad (31)$$

In Eq. (31), the first term of the right-hand side is provided in Eq. (26). The second term of the right-hand side can be simplified given the independence between $\mathbf{g}_{AR}x_A, \mathbf{g}_{BR}x_A, \mathbf{w}$, i.e.,

$$h\left(\begin{bmatrix}\mathbf{g}_{AR}x_A\\\mathbf{g}_{BR}x_B\\\mathbf{w}\end{bmatrix}\right) = h(\mathbf{g}_{AR}x_A) + h(\mathbf{g}_{BR}x_B) + h(\mathbf{w}). \quad (32)$$

The last term of the right-hand-side is expressed as:

$$\begin{aligned} &h\left(\begin{bmatrix}x_A\\y_A\\\mathbf{g}_{AR}x_A\\\mathbf{g}_{BR}x_B\\\mathbf{w}\end{bmatrix}\right) = h\left(\begin{bmatrix}x_A\\\mathbf{g}_{AR}\\\mathbf{g}_{BR}x_B\\\mathbf{w}\end{bmatrix}\right)\\ =&h(x_A) + h(\mathbf{g}_{AR}) + h(\mathbf{g}_{BR}x_B) + h(\mathbf{w}),\end{aligned} \quad (33)$$

which is because:

$$\begin{aligned}
&p(x_A, y_A, \mathbf{g}_{AR}x_A, \mathbf{g}_{BR}x_B, \mathbf{w}) \\
=&p(gx_B, \mathbf{g}_{AR}, \mathbf{g}_{BR}x_B, \mathbf{w}|x_A) \cdot p(x_A) \\
=&p(gx_B, \mathbf{g}_{BR}x_B, \mathbf{w}|\mathbf{g}_{AR}, x_A) \cdot p(x_A, \mathbf{g}_{AR}) \\
=&p(gx_B, \mathbf{w}|\mathbf{g}_{BR}x_B, \mathbf{g}_{AR}, x_A) \cdot p(x_A, \mathbf{g}_{AR}, \mathbf{g}_{BR}x_B) \\
=&p(gx_B|\mathbf{w}, \mathbf{g}_{BR}x_B, \mathbf{g}_{AR}, x_A) \cdot p(x_A, \mathbf{g}_{AR}, \mathbf{g}_{BR}x_B, \mathbf{w}) \\
=&p\left(\mathbf{g}_{AR}^T \cdot diag(\mathbf{w}) \cdot \mathbf{g}_{BR} \cdot x_B|\mathbf{w}, \mathbf{g}_{BR}x_B, \mathbf{g}_{AR}, x_A\right) \\
&\cdot p(x_A, \mathbf{g}_{AR}, \mathbf{g}_{BR}x_B, \mathbf{w}) \\
=&p(x_A, \mathbf{g}_{AR}, \mathbf{g}_{BR}x_B, \mathbf{w})
\end{aligned} \tag{34}$$

Then, by taking Eq. (26) and Eqs. (32)-(33) back to Eq. (31), the mutual information between MITM RIS Eve's received signals and Alice's sent and received pair are:

$$I\left(\begin{bmatrix} x_A \\ y_A \end{bmatrix}, \begin{bmatrix} \mathbf{g}_{AR}x_A \\ \mathbf{g}_{BR}x_B \\ \mathbf{w} \end{bmatrix}\right) = h(g \cdot x_B) + h(\mathbf{g}_{AR}x_A) - h(\mathbf{g}_{AR}) \tag{35}$$

We take Eq. (30) and Eq. (35) back to Eq. (24), the mutual information difference is derived as:

$$\begin{aligned}
&I\left(\begin{bmatrix} x_A \\ y_A \end{bmatrix}, \begin{bmatrix} x_B \\ y_B \end{bmatrix}\right) - I\left(\begin{bmatrix} x_A \\ y_A \end{bmatrix}, \begin{bmatrix} \mathbf{g}_{AR}x_A \\ \mathbf{g}_{BR}x_B \\ \mathbf{w} \end{bmatrix}\right) \\
=&h(g \cdot x_A) + h(\mathbf{g}_{AR}) - h(g) - h(\mathbf{g}_{AR} \cdot x_A).
\end{aligned} \tag{36}$$

Recalling that $g = \mathbf{g}_{BR}^T \cdot diag(\mathbf{w}) \cdot \mathbf{g}_{AR}$ is a summation of random variables, which, according to [40], can be expressed by complex Gaussian distribution, i.e., $g \sim \mathcal{CN}(0, 2\sigma^2)$, where [40]

$$\sigma^2 = 0.5 \cdot M \cdot C_0^2 \cdot d_{AR}^{-\alpha_L} \cdot d_{BR}^{-\alpha_L}. \tag{37}$$

$\mathbf{g}_{AR}$ is the direct channel from Alice to RIS, which follows the complex Gaussian distribution as $\mathbf{g}_{AR} \sim \mathcal{CN}(0, 2\mathbf{\Sigma}_{AR})$ [50]. x_A is the sent signal with form $x_A = \sqrt{E_t}\exp(j\phi), \phi \sim [0, 2\pi)$. As such, we have $g \cdot x_A \sim \mathcal{CN}(0, 2E_t\sigma^2)$ and $\mathbf{g}_{AR} \cdot x_A \sim \mathcal{CN}(0, 2E_t\mathbf{\Sigma}_{AR})$. The entropy terms in the right-hand side of Eq. (36) are:

$$h(g) = 0.5\log_2\left(2\pi e \cdot \sigma^2\right), \tag{38}$$

$$h(g \cdot x_A) = 0.5\log_2\left(2\pi e \cdot E_t \cdot \sigma^2\right), \tag{39}$$

$$h(\mathbf{g}_{AR}) = 0.5\log_2\left((2\pi e)^M \cdot \det\left(\mathbf{\Sigma}_{AR}\right)\right), \tag{40}$$

$$h(\mathbf{g}_{AR}x_A) = 0.5\log_2\left((2\pi e)^M \cdot E_t^M \cdot \det\left(\mathbf{\Sigma}_{AR}\right)\right). \tag{41}$$

By taking Eqs. (38)-(41) into Eq. (36), we have

$$\begin{aligned}
&I\left(\begin{bmatrix} x_A \\ y_A \end{bmatrix}, \begin{bmatrix} x_B \\ y_B \end{bmatrix}\right) - I\left(\begin{bmatrix} x_A \\ y_A \end{bmatrix}, \begin{bmatrix} \mathbf{g}_{AR}x_A \\ \mathbf{g}_{BR}x_B \\ \mathbf{w} \end{bmatrix}\right) \\
=&0.5(1-M)\log_2 E_t,
\end{aligned} \tag{42}$$

which is positive when the transmitting power $E_t < 1$.

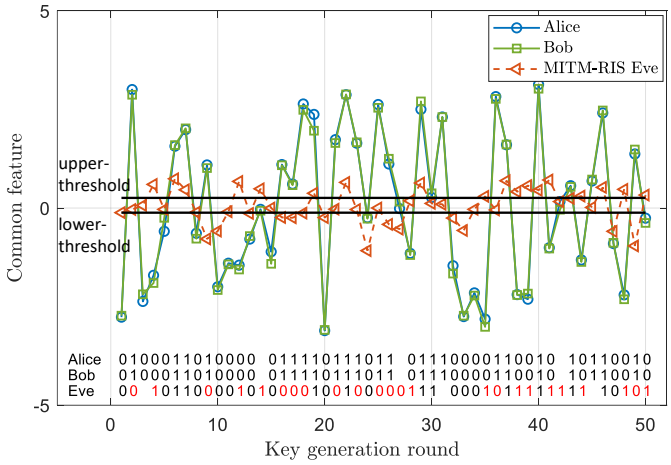

Fig. 9. Secret key quantization from common features by upper and lower thresholds

Appendix B
Secret Key Generation after Common Feature Generation

After the generation of common features at Alice and Bob, they will individually quantize their features into binary keys, which is illustrated in Fig. 9. Here, two threshold quantization method is used. In each key generation round, the key will be quantized as 1 or 0 if the current feature is larger than a predefined upper-threshold or lower than a predefined lower-threshold, or will be discarded if the feature is within the region between two thresholds.

After the key quantization process, key reconciliation will be pursued. In this process, one legitimate user (say Alice) will generate check-bits by channel coding scheme (e.g., polar codes). Then, Alice will send check-bits to Bob, who will use these to reconcile the mismatched bits by decoding processes.

Finally, privacy amplification will be employed, with the aim of enlarging the secret key length to meet the secrecy capacity requirement.

References

[1] H.-M. Wang, X. Zhang, and J.-C. Jiang, "UAV-Involved Wireless Physical-Layer Secure Communications: Overview and Research Directions," *IEEE Wireless Communications*, vol. 26, no. 5, pp. 32–39, 2019.

[2] W. Wang, X. Liu, J. Tang, N. Zhao, Y. Chen, Z. Ding, and X. Wang, "Beamforming and Jamming Optimization for IRS-Aided Secure NOMA Networks," *IEEE Transactions on Wireless Communications*, vol. 21, no. 3, pp. 1557–1569, 2022.

[3] Y. Zhou, P. L. Yeoh, H. Chen, Y. Li, R. Schober, L. Zhuo, and B. Vucetic, "Improving Physical Layer Security via a UAV Friendly Jammer for Unknown Eavesdropper Location," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 11, pp. 11 280–11 284, 2018.

[4] X. Pang, N. Zhao, J. Tang, C. Wu, D. Niyato, and K.-K. Wong, "IRS-Assisted Secure UAV Transmission via Joint Trajectory and Beamforming Design," *IEEE Transactions on Communications*, vol. 70, no. 2, pp. 1140–1152, 2022.

[5] A. Mukherjee, S. A. A. Fakoorian, J. Huang, and A. L. Swindlehurst, "Principles of Physical Layer Security in Multiuser Wireless Networks: A Survey," *IEEE Communications Surveys Tutorials*, vol. 16, no. 3, pp. 1550–1573, 2014.

[6] J. Zhang, T. Q. Duong, A. Marshall, and R. Woods, "Key Generation From Wireless Channels: A Review," *IEEE Access*, vol. 4, pp. 614–626, 2016.

[7] C. Ye, A. Reznik, and Y. Shah, "Extracting Secrecy from Jointly Gaussian Random Variables," in *2006 IEEE International Symposium on Information Theory*, 2006, pp. 2593–2597.

[8] C. Ye, S. Mathur, A. Reznik, Y. Shah, W. Trappe, and N. B. Mandayam, "Information-Theoretically Secret Key Generation for Fading Wireless Channels," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 2, pp. 240–254, 2010.

[9] X. Wu, Y. Song, C. Zhao, and X. You, "Secrecy extraction from correlated fading channels: An upper bound," in *2009 International Conference on Wireless Communications & Signal Processing*, 2009, pp. 1–3.

[10] J. Li, P. Wang, L. Jiao, Z. Yan, K. Zeng, and Y. Yang, "Security analysis of triangle channel-based physical layer key generation in wireless backscatter communications," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 948–964, 2023.

[11] J. Li, P. Wang, Z. Yan, Y. Yang, and K. Zeng, "Bgkey: Group key generation for backscatter communications among multiple devices," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 2470–2486, 2024.

[12] Y. Peng, P. Wang, W. Xiang, and Y. Li, "Secret Key Generation Based on Estimated Channel State Information for TDD-OFDM Systems Over Fading Channels," *IEEE Transactions on Wireless Communications*, vol. 16, no. 8, pp. 5176–5186, 2017.

[13] K. Moara-Nkwe, Q. Shi, G. M. Lee, and M. H. Eiza, "A Novel Physical Layer Secure Key Generation and Refreshment Scheme for Wireless Sensor Networks," *IEEE Access*, vol. 6, pp. 11 374–11 387, 2018.

[14] S. N. Premnath, S. Jana, J. Croft, P. L. Gowda, M. Clark, S. K. Kasera, N. Patwari, and S. V. Krishnamurthy, "Secret Key Extraction from Wireless Signal Strength in Real Environments," *IEEE Transactions on Mobile Computing*, vol. 12, no. 5, pp. 917–930, 2013.

[15] S. Mathur, W. Trappe, N. Mandayam, C. Ye, and A. Reznik, "Radio-Telepathy: Extracting a Secret Key from an Unauthenticated Wireless Channel," in *Proceedings of the 14th ACM international conference on Mobile computing and networking*, 2008, pp. 128–139.

[16] G. Brassard and L. Salvail, "Secret-Key Reconciliation by Public Discussion," in *Advances in Cryptology — EUROCRYPT '93*, T. Helleseth, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 1994, pp. 410–423.

[17] R. Impagliazzo, L. A. Levin, and M. Luby, "Pseudo-Random Generation from One-Way Functions," in *Proceedings of the twenty-first annual ACM symposium on Theory of computing*, 1989, pp. 12–24.

[18] G. Li, A. Hu, J. Zhang, and B. Xiao, "Security Analysis of a Novel Artificial Randomness Approach for Fast Key Generation," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, 2017, pp. 1–6.

[19] Y. Lou, L. Jin, Z. Zhong, K. Huang, and S. Zhang, "Secret Key Generation Scheme based on MIMO Received Signal Spaces," *Scientia Sinica Informationis*, vol. 47, no. 3, pp. 362–373, 2017.

[20] H. Taha and E. Alsusa, "Secret Key Exchange Using Private Random Precoding in MIMO FDD and TDD Systems," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 6, pp. 4823–4833, 2017.

[21] A. Khisti, "Secret-Key Agreement over Non-Coherent Block-Fading Channels with Public Discussion," *IEEE Transactions on Information Theory*, vol. 62, no. 12, pp. 7164–7178, 2016.

[22] S. Sharifian, F. Lin, and R. Safavi-Naini, "Secret key agreement using a virtual wiretap channel," in *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, 2017, pp. 1–9.

[23] S. Zhang, L. Jin, Y. Lou, and Z. Zhong, "Secret Key Generation based on Two-Way Randomness for TDD-SISO System," *China Communications*, vol. 15, no. 7, pp. 202–216, 2018.

[24] G. Wunder, R. Fritschek, and K. Reaz, "RECiP: Wireless Channel Reciprocity Restoration Method for Varying Transmission Power," in *2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2016, pp. 1–5.

[25] Q. Wu, S. Zhang, B. Zheng, C. You, and R. Zhang, "Intelligent Reflecting Surface-Aided Wireless Communications: A Tutorial," *IEEE Transactions on Communications*, vol. 69, no. 5, pp. 3313–3351, 2021.

[26] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE Communications Magazine*, vol. 58, no. 1, pp. 106–112, 2020.

[27] ——, "Intelligent Reflecting Surface Enhanced Wireless Network via Joint Active and Passive Beamforming," *IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5394–5409, 2019.

[28] Q. Ma, L. Chen, H. B. Jing, Q. R. Hong, H. Y. Cui, Y. Liu, L. Li, and T. J. Cui, "Controllable and Programmable Nonreciprocity based on Detachable Digital Coding Metasurface," *Advanced Optical Materials*, vol. 7, no. 24, p. 1901285, 2019.

[29] W. Jiang, B. Chen, J. Zhao, Z. Xiong, and Z. Ding, "Joint Active and Passive Beamforming Design for the IRS-Assisted MIMOME-OFDM Secure Communications," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 10, pp. 10 369–10 381, 2021.

[30] S. Hong, C. Pan, H. Ren, K. Wang, and A. Nallanathan, "Artificial-Noise-Aided Secure MIMO Wireless Communications via Intelligent Reflecting Surface," *IEEE Transactions on Communications*, vol. 68, no. 12, pp. 7851–7866, 2020.

[31] X. Hu, L. Jin, K. Huang, X. Sun, Y. Zhou, and J. Qu, "Intelligent Reflecting Surface-Assisted Secret Key Generation With Discrete Phase Shifts in Static Environment," *IEEE Wireless Communications Letters*, vol. 10, no. 9, pp. 1867–1870, 2021.

[32] X. Lu, J. Lei, Y. Shi, and W. Li, "Intelligent Reflecting Surface Assisted Secret Key Generation," *IEEE Signal Processing Letters*, vol. 28, pp. 1036–1040, 2021.

[33] Z. Ji, P. L. Yeoh, G. Chen, C. Pan, Y. Zhang, Z. He, H. Yin, and Y. Li, "Random Shifting Intelligent Reflecting Surface for OTP Encrypted Data Transmission," *IEEE Wireless Communications Letters*, vol. 10, no. 6, pp. 1192–1196, 2021.

[34] P. Staat, H. Elders-Boll, M. Heinrichs, R. Kronberger, C. Zenger, and C. Paar, "Intelligent Reflecting Surface-Assisted Wireless Key Generation for Low-Entropy Environments," in *2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2021, pp. 745–751.

[35] Z. Ji, P. L. Yeoh, D. Zhang, G. Chen, Y. Zhang, Z. He, H. Yin, and Y. li, "Secret Key Generation for Intelligent Reflecting Surface Assisted Wireless Communication Networks," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 1, pp. 1030–1034, 2021.

[36] G. Li, C. Sun, W. Xu, M. D. Renzo, and A. Hu, "On maximizing the sum secret key rate for reconfigurable intelligent surface-assisted multiuser systems," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 211–225, 2022.

[37] L. Hu, G. Li, H. Luo, and A. Hu, "On the RIS Manipulating Attack and Its Countermeasures in Physical-layer Key Generation," in *2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*, 2021, pp. 1–5.

[38] G. Li, L. Hu, P. Staat, H. Elders-Boll, C. Zenger, C. Paar, and A. Hu, "Reconfigurable Intelligent Surface for Physical Layer Key Generation: Constructive or Destructive?" *IEEE Wireless Communications*, vol. 29, no. 4, pp. 146–153, 2022.

[39] M. Wei, H. Zhao, V. Galdi, L. Li, and T. J. Cui, "Metasurface-enabled smart wireless attacks at the physical layer," *Nature Electronics*, vol. 6, no. 8, pp. 610–618, 2023.

[40] Z. Wei, B. Li, and W. Guo, "Adversarial reconfigurable intelligent surface against physical layer key generation," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 2368–2381, 2023.

[41] C. D. T. Thai, J. Lee, and T. Q. Quek, "Physical-Layer Secret Key Generation with Colluding Untrusted Relays," *IEEE Transactions on Wireless Communications*, vol. 15, no. 2, pp. 1517–1530, 2015.

[42] M. Letafati, A. Kuhestani, D. W. K. Ng, and H. Behroozi, "A New Frequency Hopping-Aided Secure Communication in the Presence of an Adversary Jammer and an Untrusted Relay," in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2020, pp. 1–7.

[43] A. K. Kamboj, P. Jindal, and P. Verma, "Machine learning-based physical layer security: techniques, open challenges, and applications," *Wireless Networks*, vol. 27, pp. 5351–5383, 2021.

[44] S. Im, H. Jeon, J. Choi, and J. Ha, "Secret Key Agreement with Large Antenna Arrays under the Pilot Contamination Attack," *IEEE Transactions on Wireless Communications*, vol. 14, no. 12, pp. 6579–6594, 2015.

[45] X. Zhang, G. Li, J. Zhang, A. Hu, Z. Hou, and B. Xiao, "Deep-learning-based physical-layer secret key generation for fdd systems," *IEEE Internet of Things Journal*, vol. 9, no. 8, pp. 6081–6094, 2021.

[46] Z. Wang, Y. Liu, J. Wang, Z. Li, Z. Li, X. Yang, F. Qi, and H. Jia, "A reliable physical layer key generation scheme based on rss and lstm network in vanet," *IEEE Internet of Things Journal*, 2023.

[47] J. Han, X. Zeng, X. Xue, and J. Ma, "Physical layer secret key generation based on autoencoder for weakly correlated channels," in *2020 IEEE/CIC International Conference on Communications in China (ICCC)*. IEEE, 2020, pp. 1220–1225.

[48] J. Zhou and X. Zeng, "Physical-layer secret key generation based on domain-adversarial training of autoencoder for spatial correlated channels," *Applied Intelligence*, vol. 53, no. 5, pp. 5304–5319, 2023.

[49] A. Taha, M. Alrabeiah, and A. Alkhateeb, "Enabling Large Intelligent Surfaces with Compressive Sensing and Deep Learning," *IEEE Access*, vol. 9, pp. 44 304–44 321, 2021.

[50] E. Björnson and L. Sanguinetti, "Rayleigh Fading Modeling and Channel Hardening for Reconfigurable Intelligent Surfaces," *IEEE Wireless Communications Letters*, vol. 10, no. 4, pp. 830–834, 2021.

[51] O. Tsilipakos, A. C. Tasolamprou, A. Pitilakis, F. Liu, X. Wang, M. S. Mirmoosa, D. C. Tzarouchis, S. Abadal, H. Taghvaee, C. Liaskos *et al.*, "Toward Intelligent Metasurfaces: The Progress from Globally Tunable Metasurfaces to Software-Defined Metasurfaces with an Embedded Network of Controllers," *Advanced optical materials*, vol. 8, no. 17, p. 2000783, 2020.

[52] O. Özdogan, E. Björnson, and E. G. Larsson, "Intelligent Reflecting Surfaces: Physics, Propagation, and Pathloss Modeling," *IEEE Wireless Communications Letters*, vol. 9, no. 5, pp. 581–585, 2020.

[53] L. Jin, S. Zhang, Y. Lou, X. Xu, and Z. Zhong, "Secret Key Generation With Cross Multiplication of Two-Way Random Signals," *IEEE Access*, vol. 7, pp. 113 065–113 080, 2019.

[54] W. Guo, "Explainable artificial intelligence for 6g: Improving trust between human and machine," *IEEE Communications Magazine*, vol. 58, no. 6, pp. 39–45, 2020.

[55] A. Pishkoo and M. Darus, "On meijer's g-functions (mgfs) and its applications," *Reviews in Theoretical Science*, vol. 3, no. 2, pp. 216–223, 2015.

[56] A. M. Alaa and M. van der Schaar, "Demystifying black-box models with symbolic metamodels," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[57] H. Chergui, M. Benjillali, and S. Saoudi, "Performance analysis of project-and-forward relaying in mixed mimo-pinhole and rayleigh dual-hop channel," *IEEE Communications Letters*, vol. 20, no. 3, pp. 610–613, 2016.

[58] A. N. Kolmogorov, "On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition," in *Doklady Akademii Nauk*, vol. 114, no. 5. Russian Academy of Sciences, 1957, pp. 953–956.

[59] S. C. Sun and W. Guo, "Approximate symbolic explanation for neural network enabled water-filling power allocation," in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, 2020, pp. 1–4.

[60] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Channel Estimation and Hybrid Precoding for Millimeter Wave Cellular Systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 831–846, 2014.