IEEE ComSoc IEEE Communications Society | IEEE COMPUTER SOCIETY | IEEE Signal Processing Society | VTS | IEEE Transactions on Machine Learning in Communications and Networking

Received 19 December 2024; revised 25 March 2025; accepted 22 April 2025.
Date of publication 28 April 2025; date of current version 7 May 2025.

The associate editor coordinating the review of this article and approving it for publication was Y. Liu.

Digital Object Identifier 10.1109/TMLCN.2025.3564907

Deep Joint Source Channel Coding for Privacy-Aware End-to-End Image Transmission

MEHDI LETAFATI [1] (Student Member, IEEE), SEYYED AMIRHOSSEIN AMELI KALKHORAN [2], ECENAZ ERDEMIR [3] (Student Member, IEEE), BABAK HOSSEIN KHALAJ [1,4,5] (Senior Member, IEEE), HAMID BEHROOZI [1] (Member, IEEE), AND DENIZ GÜNDÜZ [3] (Fellow, IEEE)

[1]Department of Electrical Engineering, Sharif University of Technology, Tehran 14588-89694, Iran
[2]Computer Engineering Department, Sharif University of Technology, Tehran 14588-89694, Iran
[3]Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ London, U.K.
[4]Sharif Center for Information Systems and Data Science, Sharif University of Technology, Tehran 14588-89694, Iran
[5]School of Computer Science, Institute for Research in Fundamental Sciences, Tehran 19538-33511, Iran
CORRESPONDING AUTHOR: M. LETAFATI (mletafati@ee.sharif.edu)

The work of Babak Hossein Khalaj was supported in part by the Institute for Research in Fundamental Sciences (IPM).

ABSTRACT Deep neural network (DNN)-based joint source and channel coding is proposed for privacy-aware end-to-end image transmission against multiple eavesdroppers. Both scenarios of colluding and non-colluding eavesdroppers are considered. Unlike prior works that assume perfectly known and independent identically distributed (i.i.d.) source and channel statistics, the proposed scheme operates under unknown and non-i.i.d. conditions, making it more applicable to real-world scenarios. The goal is to transmit images with minimum distortion, while simultaneously preventing eavesdroppers from inferring certain private attributes of images. Simultaneously generalizing the ideas of *privacy funnel* and wiretap coding, a multi-objective optimization framework is expressed that characterizes the trade-off between image reconstruction quality and information leakage to eavesdroppers, taking into account the structural similarity index (SSIM) for improving the perceptual quality of image reconstruction. Extensive experiments on the CIFAR-10 and CelebA, along with ablation studies, demonstrate significant performance improvements in terms of SSIM, adversarial accuracy, and the mutual information leakage compared to benchmarks. Experiments show that the proposed scheme restrains the adversarially-trained eavesdroppers from intercepting privatized data for both cases of eavesdropping a common secret, as well as the case in which eavesdroppers are interested in different secrets. Furthermore, useful insights on the *privacy-utility trade-off* are also provided.

INDEX TERMS DeepJSCC, secure image transmission, end-to-end learning, privacy-utility trade-off, adversarial neural networks, deep learning.

I. INTRODUCTION

THE sixth generation (6G) of wireless networks is envisioned to realize *connected intelligence*, supporting an overabundance of disruptive technologies such as autonomous driving and haptic communications for extended reality (XR) and metaverse [2]. Recently, intelligent *multimedia transmission* is receiving much attention due to its various applications in XR, metaverse, and surveillance systems [3]. In this regard, if the transmission of source images is not properly secured, the performance of such services are no longer reliable.

Although different radio protocols have been proposed for the security of core network, the wireless edge of 6G networks seem to face ever-increasing privacy and security threats, like man-in-the-middle attacks [4], [5], [6], spoofing [7], and eavesdropping [8], [9]. Fundamentally, this stems from the intrinsic open nature of the wireless medium, which makes transmitted signals susceptible to security and

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

privacy risks. In this regard, deep neural network (DNN)-based approaches have shown to be capable of propelling the performance of wireless systems such that context-aware security can be achieved [10], [11], [12].

A. MOTIVATION AND BACKGROUND

The communication of images from a source node, Alice, to a legitimate node, Bob, over a noisy wireless channel, is a joint source channel coding (JSCC) problem. DNN-aided JSCC design, known as DeepJSCC, has received significant attention in recent years thanks to their superior performance, particularly their lack of reliance on accurate channel state information [13], [14], [15], [16], [17], [18].

DeepJSCC frameworks are becoming appealing for practical deployments in real-world wireless systems [13]. It is becoming increasingly evident that conventional separate network architectures struggle to meet real-world performance demands, often overlooking signal processing complexities like compression, which can dominate end-to-end latency. In contrast, DeepJSCC integrates source and channel coding, jointly optimizing them for better overall performance. Moreover, it is seen as a promising solution for reducing bandwidth requirements, saving radio resources by transmitting only the most relevant information [13], [18].

In JSCC, unlike in separable source and channel coding, the transmitted channel codeword is correlated with the underlying source signal. While this benefits the legitimate encoder by providing robustness against channel noise, it also creates additional vulnerability in terms of leakage to eavesdroppers. Also note that classical encryption methods are not applicable here as they would destroy the correlation between the source and the channel input. The DNN-based design of JSCC is particularly attractive for the design of secure content delivery schemes, since they can learn the security sensitivity of different parts of the contents, and adopt their transmission strategy accordingly. Inspired by [16] and [17], this paper studies a generalization of the DeepJSCC approach for privacy-aware end-to-end image transmission against multiple eavesdroppers, with both colluding and non-colluding eavesdropping strategies, and over AWGN as well as fading channels.

B. PRIOR ARTS

In the context of secure end-to-end (E2E) communications, autoencoders were proposed in [19], [20], and [21] for communication over AWGN wiretap channel. Feedforward neural networks composed of linear layers were employed as the encoder-decoder pair, and a weighted sum of block error rate and approximated information leakage was used as the loss function. The input data to the autoencoder was combined with additional non-informative random bits to confuse Eve, while such redundant bits can negatively affect the data rate. However, these works only focused on learning secure *channel coding* via DNNs rather than taking the source and channel coding jointly into account, i.e., undifferentiated with respect to the *content* of the message being delivered.

In the context of DNN-based secure wireless image transmission, [22] proposes cryptography-based approaches (public key cryptography) for secure transmission against one eavesdropper. It incorporates DeepJSCC scheme with encryption coding, proposing deep joint source-channel and encryption coding. The authors in [23] propose an image "protection" scheme with two additional DNN modules (so-called protection and de-protection), parameterized via the U-Net architecture added to the DeepJSCC pipeline. The authors are mainly concerned about "application layer" digital content protection, and they do not take into account the privacy information leakage to eavesdroppers over wireless links. In [24], secure image transmission with DeepJSCC as the neural model backbone for E2E communication is studied considering only the effect of one eavesdropper. This letter assumes the eavesdropper has access to a similar DNN model as of the legitimate nodes, and aims to reconstruct images. To combat the image reconstruction at the eavesdropper, the loss function for training the DeepJSCC pipeline is simply modified by incorporating the mean-squared-error (MSE) between the source images and eavesdropper's reconstruction, which is called "SecureMSE." Applying SecureMSE for model training requires the statistical characteristics of the eavesdropper's channel to be known by legitimate parties, which seems not to be feasible in practice, especially in the context of "totally passive" eavesdroppers. Recently, a preprint [25] studies the security of wireless image transmission against multiple eavesdroppers. The authors approach security from a different perspective, i.e., signal steganography. They propose two additional neural modules to be added to the encoder/decoder, calling it steganography modules. Finally, the security of DeepJSCC-based wireless image transmission against active adversaries and backdoor attacks is studied in [26], which falls into the category of adversarial machine learning (AML) algorithms.

C. OUR CONTRIBUTIONS

We propose a DeepJSCC-based solution for secure E2E wireless image transmission against *multiple eavesdroppers*. By employing DeepJSCC pipeline, idealistic assumptions of perfectly known and i.i.d. source and channel distributions are dropped, and the proposed scheme assumes unknown source and channel statistics. Generalizing the ideas of privacy funnel and wiretap coding, the goal is to transmit images with minimum distortion, while simultaneously preventing eavesdroppers from inferring private image attributes, taking into account the perceptual quality metrics for image transmission and reconstruction. We consider both eavesdropping strategies, i.e., the colluding setup, in which the eavesdroppers cooperate with each other to extract sensitive/private parts of images, and, the non-colluding setup, where the eavesdroppers act alone. We also study both cases of eavesdroppers being interested in a common secret, and the

case in which every eavesdropper is interested in a different secret. Notably, previous works [16] and [17] only considered a single eavesdropper with a single-antenna transmitter and multiple parallel channels, respectively. In addition, both [16] and [17] are limited to static channels, while we assume two well-established time-varying wireless channel models. That is, our proposed scheme is trained over complex-valued Rayleigh fading channels, and tested over Nakagami-*m* and AWGN channels in addition to Rayleigh-fading. Our proposed scheme prevents adversarially-trained eavesdroppers from accurately inferring sensitive attributes of the transmitted images. Moreover, (different from [19], [20]) no additional redundant bits are required to be added to the source image. Unlike the methods in [23] and [25], our scheme integrates seamlessly into the standard DeepJSCC pipeline without requiring additional processing modules. Instead, we modify the loss function to come up with a *privacy-aware* training and inference strategy for the communication entities, taking the information leakage and perceptual quality of images into account.

Our scheme relies on a *data-driven* approach, i.e., we do not consider any specific assumption on the underlying distribution of the data source or the sensitive/private part. Instead, we have access to datasets to facilitate the process of learning secure encoder-decoder pairs. We provide extensive numerical experiments over different datasets, i.e., CIFAR-10[1] and CelebFaces Attributes (CelebA).[2] We demonstrate the performance gain of our approach, in terms of structural similarity index (SSIM), information leakage, adversarial accuracy, and reconstruction quality, through extensive experimental studies in different communication scenarios. We also conduct *ablation studies* to further study the effect of different hyperparameters of the implemented DNNs and the loss function on the E2E performance.

Finally, we would like to highlight the main differences between our paper and the prior work of [16]. 1) Our proposed scheme considers both AWGN and fading channels, while [16] only considers an AWGN channel. 2) Our model is designed for privacy-aware transmission against multiple eavesdroppers, with different eavesdropping strategies, while [16] simply considers one Eve. 3) Our neural architecture differs from [16]. We do not simply employ "fully-convolutional" DNNs, but also utilize generalized normalization transformation (GDN) blocks, and their inverse counterpart (IGDN), [27] at the hidden layers. This has been shown to be able to provide significant improvements in capturing image statistics of natural images. 4) The model is trained using a mixed loss function that combines pixel-wise reconstruction (via MSE) with *perceptual quality optimization* (via the SSIM metric), whereas [16] only considers the pixel-wise metric of peak signal-to-noise ratio (PSNR). Unlike PSNR, the SSIM metric is a perception-based metric, which captures pixel inter-dependencies, and hence, the semantics of the source image.

[1]https://www.cs.toronto.edu/ kriz/cifar.html
[2]https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

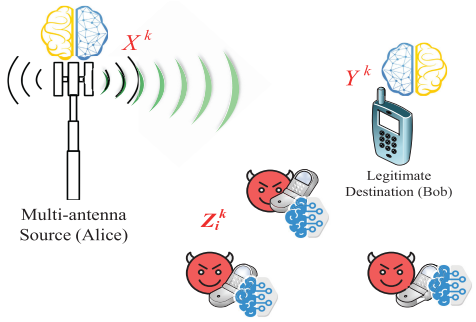


FIGURE 1. Proposed learning-based system model for end-to-end image transmission againstmultiple eavesdroppers.

D. PAPER ORGANIZATION AND NOTATIONS

The remainder of this paper is organized as follows. In Section II, our proposed system model together with the main assumptions and intuitions are provided. We also formulate our problem to characterize the privacy-funnel-like framework. In Section III we propose our learning-based E2E approach to solve our problem in a data-driven manner. In Section IV, we investigate the performance of our scheme through extensive numerical experiments. We also compare our learning-based scheme with different benchmarks. Finally, Section V concludes our paper.

Notations: We denote the transpose, the conjugate transpose, and ℓ^2 norm of a vector by $(\cdot)^{\mathsf{T}}$, $(\cdot)^{\dagger}$, and $||\cdot||$, respectively. Vectors are represented by bold lowercase letters. The zero and the identity matrices are shown by $\mathbf{0}$ and $\boldsymbol{I}$, respectively. $\mathcal{CN}(\mu, \sigma^2)$ and $\mathcal{G}(k, \theta)$ represent a complex random variable (RV) with mean μ and standard deviation σ, and gamma distribution with shape parameter k and scale parameter θ, respectively. The expected value and the probability density function (pdf) of RV X are denoted by $\mathbb{E}[X]$ and $p_X(x)$, respectively. The mutual information between RVs X and Y is denoted by $I(X;Y)$, while the cross-entropy of distributions p relative to distribution q over a given set is defined as $H(p,q) \triangleq -\mathbb{E}_p[\log q]$. Moreover, $\mathrm{KL}(\cdot||\cdot)$ denotes the Kullback-Leibler (KL) divergence.

II. SYSTEM MODEL AND PROBLEM STATEMENT

A. MODEL DESCRIPTION

Consider the communication scenario demonstrated in Fig. 1, where a multi-antenna source node, Alice ($\mathcal{A}$), aims to transmit images $U^n \in \mathcal{U}^n$ to a destination node, Bob ($\mathcal{B}$), over k uses of the wireless channel, where $\mathcal{U}$ denotes the alphabet of source images. According to the JSCC literature [14], we refer to the image dimension, n, as the source bandwidth. The channel dimension k characterizes the channel bandwidth. We usually have $k < n$, which reflects the concept of bandwidth compression (please see [14] and references therein for more details). The image transmission service

should be kept secret from multiple eavesdroppers, which overhear the communication and are equipped with adversarial DNNs, aiming to extract private data regarding the source image. $\mathcal{A}$ and $\mathcal{B}$ employ DNNs and perform *secure Depp-JSCC* by leveraging the concept of autoencoders. Details of the proposed DNN architectures, together with the training strategies of legitimate autoencoder and adversarial neural decoders are given in the next sections.

Alice aims to convey the source information U^n to Bob with minimum distortion, while preventing the information of sensitive part $S_i \in \mathcal{S}_i$, with a discrete alphabet $\mathcal{S}_i$, to be leaked to the i-th eavesdropper, $i = 1, \cdots, M$, where M denotes the number of eavesdroppers. We note that the sensitive (private) parts are correlated with U^n with distribution $p_{U^n,S_1,S_2\cdots,S_M}$. Toward this end, Alice maps the source information U^n into a channel input codeword $X^k \in \mathcal{X}^k$ via implementing a potentially non-trivial function $f_{\mathcal{A}} : \mathcal{U}^n \to \mathcal{X}^k$, where $X^k = f_{\mathcal{A}}(U^n)$. Transmitted codeword is subject to an average power constraint, $\frac{1}{k}\mathbb{E}[(X^k)^* X^k] \leq P$. Channel outputs at Bob and the i'th eavesdropper are denoted, respectively, by Y^k and Z_i^k, $i \in \{1, \cdots, M\}$. Bob then applies a decoding function $f_{\mathcal{B}}$ to obtain $\hat{U}^n = f_{\mathcal{B}}(Y^k)$. Meanwhile, each eavesdropper tries to extract the sensitive information S_i, that he is interested in, from his observations Z_i^k. We consider the trade-off between delivering images to Bob with the highest fidelity and controlling the information leakage to each adversary, which is theoretically measured by the mutual information metric $I(S_i; Z_i^k)$. In the case that all eavesdroppers are interested in the same secret we have $S_i = S, \forall i$.

Transmission of data-streams over the air experiences independent realizations of communication channels. Generally speaking, the channels impose random corruption on any transmitted symbol vector $\boldsymbol{\tau} \in \mathbb{C}^k$, which we model through the transfer function $\eta(\boldsymbol{\tau})$. We consider both the AWGN and slow fading channels, where for the slow fading, we adopt two widely-used assumptions of Rayleigh fading and Nakagami-m channels. The transfer function of the Gaussian channel can be formulated by $\eta_v(\boldsymbol{\tau}) = \boldsymbol{\tau} + \boldsymbol{v}$, where the vector $\boldsymbol{v} \in \mathbb{C}^k$ is comprised of independent identically distributed (i.i.d.) samples from a circularly symmetric complex Gaussian distribution, i.e., $\boldsymbol{v} \sim \mathcal{CN}(\mathbf{0}, \sigma^2 \boldsymbol{I}_k)$, where σ^2 is the average power of additive noise. For the fading scenario, the multiplicative effect of the communication channel is modeled by $\eta_h(\boldsymbol{z}) = h\boldsymbol{\tau}$, where $h \sim \mathcal{CN}(0, \delta_h^2)$ for the Rayleigh fading scenario, and $|h|^2 \sim \mathcal{G}(m, \delta_h^2/m)$ for the Nakagami-m scenario with $m > 1$. Here, δ_h represents the standard deviation (std) of the channel distribution, which reflects the large-scale effects of channel fading. The joint effect of channel fading and Gaussian noise can be treated by the composition of transfer functions as $\eta(\boldsymbol{\tau}) = \eta_v(\eta_h(\boldsymbol{\tau})) = h\boldsymbol{\tau} + \boldsymbol{v}$.

Fundamentally, the trade-off between the minimum achievable distortion at the legitimate receiver and the minimum leakage to Eve is asymptotically characterized in [28] under the idealistic assumption of i.i.d. source and channels with known distributions. However, this result is not applicable to practical systems with finite block-lengths and unknown source and channel statistics. Hence, in our system model we consider a *data-driven* approach for the practical non-asymptotic regime. That is, we do not consider any underlying assumption on the distribution of the data source or the latent private variable. We remark that while we do assume a known channel model, we use this model to generate samples from conditional channel distribution. We could easily drop this assumption if we had data collected from a particular channel with unknown statistics.

B. PROBLEM FORMULATION

Inspired by the key idea of *privacy funnel* [29], we first formulate an optimization framework to characterize the trade-off between successful image recovery at Bob and the information leakage to the eavesdroppers. The formulated funnel-like framework is then solved in a data-driven manner, utilizing DeepJSCC architecture. That is, we implement DNNs, based on the concept of autoencoders, to learn the encoding-decoding functions, while restraining the adversarially-trained eavesdroppers from intercepting private information.

We aim to simultaneously minimize both the distortion $d(U^n, \hat{U}^n)$ at Bob and the information leakage, $I(S_i; Z_i^k)$, $i \in [M]$, about the secrets S_i. This corresponds to a multi-objective programming (MOP) that can be solved using the scalarization approach [30]. That is, we minimize a weighted sum of the multiple objectives with weights $w_i \in \mathbb{R}_{>0}$ as follows

$$\underset{p_{X^k|U^n} f_{\mathcal{B}}}{\text{minimize}} \quad \mathbb{E}\Big[d(U^n, \hat{U}^n)\Big] + \frac{1}{M}\sum_{i\in[M]} w_i I(S_i; Z_i^k). \tag{1}$$

Notably, w_i in (1) adjusts the trade-off between the information leakages and the distortion at the legitimate receiver. To make (1) tractable, we further need to estimate the mutual information term. In particular, it is still a major challenge to optimize DNNs, while having an objective function comprised of expressions about the mutual information between two (or more) data distributions. It is also challenging to directly estimate the mutual information metric from samples according to [16], [17], [19], and [20]. Hence, an approximation of the mutual information metric is needed, in a way that it is also "interpretable" for the training of DNNs.

In the following, inspired by [16] and [17], we apply a *variational approximation* for the mutual information metric. First, we can write

$$\begin{aligned} I(S_i; Z_i^k) &= H(S_i) - H(s_i|Z_i^k) \\ &\overset{(a)}{=} H(S_i) + \mathrm{KL}(p_{S_i|Z_i^k} || q_{S_i|Z_i^k}) + \mathbb{E}\Big[\log q_{S_i|Z_i^k}(s_i|z_i)\Big] \\ &\overset{(b)}{=} H(S_i) + \max_{q_{S_i|Z_i^k}} \mathbb{E}\Big[\log q_{S_i|Z_i^k}(s_i|z_i)\Big], \end{aligned} \tag{2}$$

where (a) follows from the fact that $-H(s_i|Z_i^k) = \mathrm{KL}(p_{S_i|Z_i^k} || q_{S_i|Z_i^k}) + \mathbb{E}\Big[\log q_{S_i|Z_i^k}(s_i|z_i)\Big]$, and (b) follows from the fact that KL divergence is nonnegative, where the maximum is attained when the variational posterior $q_{S_i|Z_i^k}$ equals

the true posterior $p_{S_i|Z_i^k}$, i.e., $q_{S_i|Z_i^k} = p_{S_i|Z_i^k}$. Due to the intractability of the true posterior distribution, $q_{S_i|Z_i^k}(s_i|z_i)$ can be characterized as the likelihood estimation of the i'th adversary's with respect to the correct distribution of S_i, given the observation Z_i^k, which will be realized via adversarial DNNs employed at eavesdroppers. Therefore, dropping the constant term $H(S_i)$ in (2), our objective function can be rewritten as

$$\underset{p_{X^k|U^n} f_{\mathcal{B}}}{\text{minimize}} \ \mathbb{E}\Big[d(U^n, \hat{U}^n)\Big] + \frac{1}{M}\sum_{i\in[M]} w_i \max_{q_{S_i|Z_i^k}} \mathbb{E}\Big[\log q_{S_i|Z_i^k}(s_i|z_i)\Big]. \tag{3}$$

The benefit of this approximation is that we can interpret the variational lower bound on information leakage as the sample-wise cross-entropy between the distribution over adversaries' predictions, and the true distribution of private attributes [31], [32].

Invoking (3), one can infer that it can be viewed as a minimax game between Alice-Bob pair, and the set of adversaries. To elaborate, while adversarial nodes wish to maximize their individual information leakage by choosing the posterior likelihood distributions $q_{S_i|Z_i^k}$, Alice and Bob should jointly determine the optimum encoding-decoding functions, $f_{\mathcal{A}}$ and $f_{\mathcal{B}}$, that minimizes the weighted sum of the distortion and leakage. Nevertheless, we emphasize that we do not assume to have any knowledge about the underlying distribution of image source and the sensitive attributes, which is aligned with the real-world scenarios. Our scheme relies upon a *data-driven* approach, i.e., we leverage datasets to "learn" the optimized privacy-aware encoding-decoding functionalities.

To solve the proposed MOP of (3) in a data-driven manner, we implement DNNs to parameterize the optimized encoding-decoding functionalities at Alice and Bob, while the adversaries also employ optimized adversarially-trained DNNs and try to infer the sensitive attribute of the transmitted image. Let $\Omega_{\mathcal{A}}$ and $\Omega_{\mathcal{B}}$ stand for the set of (trainable) parameters of the autoencoder pair at Alice and Bob, respectively. In addition, let $\Theta_{E,i}$ parameterize the adversarial network of the i'th adversary. We can formulate the following loss function to be minimized.

$$\begin{aligned}\mathcal{L}(\Omega_{\mathcal{A}}, \Omega_{\mathcal{B}}, \Theta_{E,1}, \cdots, \Theta_{E,M}) &= \mathbb{E}\left[d(U^n, f_{\Omega_{\mathcal{B}}}(Y^k))\right] \\ &\quad + \frac{1}{M}\sum_{i\in[M]} w_i \max_{\Theta_{E,i}} \mathbb{E}\left[\log q_{\Theta_{E,i}}(s_i|z_i)\right],\end{aligned} \tag{4}$$

where $f_{\Omega_{\mathcal{A}}}$ and $f_{\Omega_{\mathcal{B}}}$ represent the encoder and decoder functionality of Alice's DNN and Bob's DNN, respectively, for which we have $X^k = f_{\Omega_{\mathcal{A}}}(U^n)$. Moreover, $q_{\Theta_{E,i}}(s_i|z_i)$ formulates the approximated adversarial likelihood about the correct value $S_i = s_i$, estimated at the DNN of the i'th adversary with parameters $\Theta_{E,i}$ and based on the observation $Z_i^k = z_i$. We emphasize that similarly to [14], [15], and [16], the legitimate and wiretap channels are treated as non-trainable layers. However, since the considered channels are differentiable, we incorporate them as part of our proposed E2E transmission scheme.

In the following, we leverage the novel concept utilized in generative adversarial networks (GANs) and iteratively train the DNNs based on the loss function proposed in (4). To elaborate, each joint training phase of the autoencoder pair $(\Omega_{\mathcal{A}}, \Omega_{\mathcal{B}})$ is followed by a training step for the adversaries' parameters $(\Theta_{E,1}, \Theta_{E,2}, \cdots, \Theta_{E,M})$, where more details are given in the next section. In what follows, we also address context-aware interpretations for the distortion measure and the likelihood distribution of sensitive attributes, which leads us toward a machine learning-interpretable approach for E2E privacy-aware communication.

III. PROPOSED DEEP JSCC-BASED SOLUTION

Following the DeepJSCC concept, we employ DNNs to directly map the pixel values to the complex-valued samples, which are sent over the air. Consider image input files with dimensions $n = H \times W \times C$, where H, W, and C stand for the image height, width, and the number of channels (3 for colored images and 1 for grayscale), respectively. Alice maps each realization of the source data U^n, denoted by $\boldsymbol{u} \in \mathbb{R}^n$, to a vector of channel input $\boldsymbol{x} \in \mathbb{C}^k$, which can be viewed as a realization of X^k. $\boldsymbol{x}$ should be securely encoded at Alice and decoded by Bob based on the objective introduced in (4). This is done via our proposed DNN-based solution, where the block diagram of the DNNs employed for the neural encoder and decoder components of legitimate parties is illustrated in Fig. 2. In addition, the DNN employed by each of the adversaries is demonstrated in Fig. 3. The structure of each DNN component is described next.[3]

A. LEGITIMATE NEURAL ENCODER-DECODER PAIR

At the legitimate encoder, we implement convolutional layers, followed by normalization invoked by the generalized normalization transformations (GDN) block [27], which is then followed by a parametric ReLU (PReLU) [33] activation function. At the output of the last PReLU layer, which consisting of $2k \times n_{\mathsf{T}}$ elements, we employ a flattening layer for each of the antennas, to reshape the encoded tensor to a data-stream from the latent space, resulting in a data sequence of the form $\tilde{\boldsymbol{x}}_i \in \mathbb{C}^k$ for each antenna $1 \le i \le n_{\mathsf{T}}$. The encoded latent sequence is further normalized, such that the channel input $\boldsymbol{X} \triangleq [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{n_{\mathsf{T}}}] \in \mathbb{C}^{k\times n_{\mathsf{T}}}$ satisfies the average transmit power constraint P. Mathematically speaking, we have

$$\boldsymbol{x}_i = \sqrt{kP}\frac{\tilde{\boldsymbol{x}}_i}{\sqrt{\tilde{\boldsymbol{x}}_i^{\dagger}\tilde{\boldsymbol{x}}_i}}, \quad 1 \le i \le n_{\mathsf{T}}. \tag{5}$$

Each column of the channel input data matrix $\boldsymbol{X}$ is then distributed over each of n_{T} antennas to exploit the available degrees of freedom.

[3]We state that the considered hyper-parameters for the legitimate DNNs, as demonstrated in Fig. 2, are selected based on comprehensive experiments and numerous trials, while the general architecture is inspired by [15].

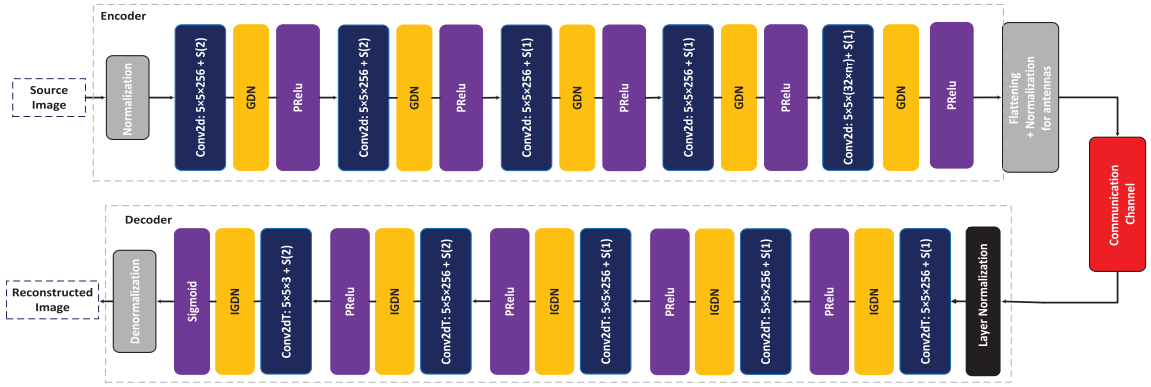


FIGURE 2. **Proposed deep neural networks at Alice (encoder) and Bob (decoder).The notation $w \times w \times f$ denotes a convolutional layer with f filters (channel outputs) of spatial extent$w$.The notation $w \times w \times (f \times n_{\mathrm{T}})$ at the last convolutional layer of encoder indicates that we require $f$filters for each of n_{T} antennas at the encoder output.Moreover, $s(\bullet)$ denotes the stride, which can be downsampling (at the encoder) or upsampling (at the decoder).**

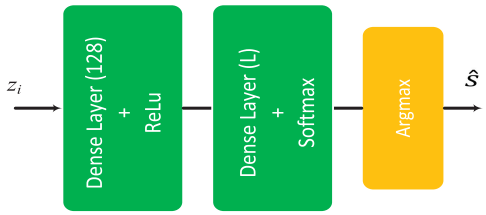


FIGURE 3. **Implemented DNN at each of the adversaries for extracting sensitive part of image files.**

The joint source-channel coded output of Alice's DNN is sent over the communication channel, where the quality of the information-bearing sequence will be corrupted due to additive noise, fading, or other channel impairments. The channel distorted version $\boldsymbol{y} \in \mathbb{C}^k$ is observed by Bob, where $\boldsymbol{y}$ is considered as a realization of Y^k, i.e., the output of the communication channel. Accordingly, $\boldsymbol{y}$ is fed to Bob's DNN, which tries to recover the image by estimating $\hat{\boldsymbol{x}} \in \mathbb{R}^n$ from $\boldsymbol{y}$, resulting in an approximate reconstructed version of the original data. Specifically, the real and imaginary parts of the k (complex-valued) channel output samples form $2k$ real-valued elements, which are fed into the convolutional layers. Based on the encoding conducted by Alice, Bob's DNN inverts the operations implemented at Alice by passing the observed data-stream $\boldsymbol{y}$ through a series of transposed convolutional layers followed by the inverse of GDN (IGDN) blocks and PReLU activation functions.

The intuition behind the proposed architecture can be explained as follows. Convolutional layers are capable of extracting image features. After that the convolutional layers have extracted the features, the GDNs (which perform differentiable and invertible operations) conduct local divisive normalization, which has been shown to be effective in density modeling of data [27]. Moreover, the GDNs offer significant improvements in capturing image statistics through Gaussianizing the data of natural images. In the next step, the activation functions facilitate the learning process of non-linear mappings from the source image space to the latent space, i.e., the channel input, and vice versa. Finally, the hyperparameters of Bob's DNN mirror the corresponding functionalities performed by the encoder layers of Alice's DNN.

Remark 1: We remark that before running the proposed DNNs, the input images are normalized by the maximum pixel value, e.g. 255 in this paper, to produce values in the range of [0, 1]. This operation will be inverted at the decoder (as demonstrated in Fig. 2) to reconstruct (de-normalized) pixel values within [0, 255] range. The rationale behind this reformulation is that the dependence of DNNs on the maximum pixel values is relaxed, since the statistics of data are unknown at the decoder. This also facilitates the training process of our proposed DNNs [11]. Notably, the sigmoid activation function applied to the output of Bob's decoder is for producing normalized approximation ($\hat{\boldsymbol{x}}$) for the original data in the range of [0, 1] before de-normalizing.

Remark 2: Prior to performing DNN-based decoding operations, a layer normalization (LN) block [34] is implemented at Bob's DNN. The LN block is aimed to realize data normalization via taking into account all of the summed inputs to the Bob's DNN during *every single training or inference sample*. Notably, unlike batch normalization which considers the distribution of the summed input to a neuron over *mini-batches* of training set, LN performs the same computation at training and inference phases *per data sample*. It is also shown in [34] that compared with previous techniques, LN substantially reduces the training time.

B. ADVERSARIAL NEURAL NETWORKS FOR EAVESDROPPERS

The structure of adversarial DNNs, which are parameterized by $\Theta_{E,i}$, $i \in [M]$, is presented in Fig. 3. According to the proposed system model in Section II-A, the adversaries utilize

DNNs to facilitate the extraction of sensitive information S_i from their received signals. The sensitive information S_i can be assumed to be the class to which the images belong [16], [17]. For instance, the identity of patients within medical imaging in e-health applications, or the locations of critical infrastructures in a cyber-physical system. To extract sensitive information (e.g., privatized data) from images, each adversary employs the DNN architecture illustrated in Fig. 3 as a class predictor, which is comprised of a dense layer (fully-connected network with 128 neurons) with rectified linear unit (ReLU) activation function.[4] The dense layer is concatenated by another dense layer with softmax activation, where the dimension of the output neurons, L, equals the cardinality of the secret $|S_i|$. The output of the softmax layer produces an adversarial likelihood estimation regarding the posterior distribution $q_{\Theta_{E,i}}(s_i|z_i)$ of sensitive attribute.[5] Specifically, considering the output $\boldsymbol{z}_i \in \mathbb{C}^k$ of the i'th wiretap channel (a realization of Z_i^k) as the observation of the i'th eavesdropper, it is fed to the adversarial DNN to obtain an approximated prediction $q_{\Theta_{E,i}}(s_i|z_i)$ regarding the sensitive attribute.

Invoking (4), each adversary tries to minimize its cross-entropy between the adversarially-estimated likelihood distribution and the ground-truth, where the ground-truth is represented by the one-hot encoded vector of S_i, as $\boldsymbol{\epsilon}_{s_i} \in \{0, 1\}^L$. Having lower cross-entropy values results in higher similarity between the posterior adversarial likelihood $q_{\Theta_{E,i}}(s_i|z_i)$ and the ground-truth $\boldsymbol{\epsilon}_{s_i}$, which can be interpreted as increase in the information leakage. Meanwhile, Alice and Bob jointly try to minimize the reconstruction distortion $d(\cdot, \cdot)$ and the information leakage (approximated by the negative cross-entropy). Exploiting the above-mentioned learning-based interpretation, our sample-wise framework can be reformulated as

$$\begin{aligned}\text{minimize}&\mathcal{L}(\Omega_{\mathcal{A}}, \Omega_{\mathcal{B}}, \Theta_{E,1}, \Theta_{E,2}, \cdots, \Theta_{E,M})\\ &= \mathbb{E}_{p(\boldsymbol{u},\hat{\boldsymbol{u}})}\left[d\left(\boldsymbol{u}, f_{\Omega_{\mathcal{B}}}(\boldsymbol{y})\right)\right]\\ &\quad + \frac{1}{M}\sum_{i\in[M]} w_i \max_{\Theta_{E,i}} \left(-H\big(q_{\Theta_{E,i}}(s_i|z_i), \boldsymbol{\epsilon}_{s_i}\big)\right), \end{aligned} \tag{6}$$

where $\hat{\boldsymbol{u}} = f_{\Omega_{\mathcal{B}}}(\boldsymbol{y})$, and $p(\boldsymbol{u}, \hat{\boldsymbol{u}})$ stands for the joint probability distribution of the original and the reconstructed image.

Remark 3: The choice of cross-entropy as the privacy leakage metric has some advantages in terms of computational efficiency, interpretability, and more stable training. Notably, cross-entropy offers several practical advantages in scalable neural network-based environments, compared to other privacy measure such as mutual information and differential privacy (DP). Cross-entropy naturally integrates into DNN training as a stable, differentiable loss function, unlike mutual information which cannot be optimized directly, whilst mutual information is often approximated from cross-entropy to enable empirical calculation in the DNN training process. Moreover, a small DP noise can cause unpredictable utility loss, which can de-stabilize training.

Due to the fact that the true distribution of data $p(\boldsymbol{u})$ is often unknown, obtaining an analytical (closed-form) solution for (6) is not tractable [14], [15], [16], [20]. Thus, we estimate the expected distortion measure using samples $\boldsymbol{u}_j$ of available datasets by computing $\mathbb{E}_{p(\boldsymbol{u},\hat{\boldsymbol{u}})}\left[d\left(\boldsymbol{u}, f_{\Omega_{\mathcal{B}}}(\boldsymbol{y})\right)\right] \approx \frac{1}{N_{\mathcal{T}}}\sum_{j=1}^{N_{\mathcal{T}}} d(\boldsymbol{u}_j, \hat{\boldsymbol{u}}_j)$, where $N_{\mathcal{T}}$ stands for the number of training samples. We note that, in this framework, we are assuming that we know the sensitive attribute each eavesdropper is interested in and their channel models, both of which are common assumptions in the privacy [16], [17], [29] and wiretap channel [20], [28] literature.

C. TRAINING PROCEDURE

In order to train our learning-based approach based on the objective function proposed in (6), we follow an iterative procedure. Inspired by the key idea of GANs, each joint training phase of the autoencoder pair $(\Omega_{\mathcal{A}}, \Omega_{\mathcal{B}})$ is followed by a training step for the adversarial networks $(\Theta_{E,1}, \Theta_{E,2}, \cdots, \Theta_{E,M})$. Intuitively, we are faced with a minimax game, i.e., the competition between legitimate autoencoder and the adversarial DNNs. Therefore, the following strategy is run through our proposed E2E system.

(i) Alice and Bob try to jointly minimize their loss function,

$$\mathcal{L}_{\mathcal{AB}} = \frac{1}{N_{\mathcal{T}}}\sum_{j=1}^{N_{\mathcal{T}}}\left(d(\boldsymbol{u}_j, \hat{\boldsymbol{u}}_j) - \frac{1}{M}\sum_{i\in[M]} w_i H\Big(q_{\Theta_{E,i}}(s_i^{(j)}|z_i^{(j)}), \boldsymbol{\epsilon}_{s_i}^{(j)}\Big)\right) \tag{7}$$

To reinforce system's security against adversaries, we take one further step within the training process of legitimate nodes. We perform *adversarial likelihood compensation* (ALC), which has been shown to be more effective in confusing an adversary than the data-dependent one-hot encoding approach [16]. The main idea behind ALC is that we want to make the estimated likelihood of adversaries imitate a uniform distribution $\bar{p}_L = [\frac{1}{L}, \cdots, \frac{1}{L}]^{\mathsf{T}}$. By doing so, Alice and Bob try to jointly maximize the uncertainty of adversarial predictions by maximizing the similarity (minimizing the cross-entropy) between the adversarial likelihood and the uniform distribution $\bar{p}_L$, instead of minimizing the likelihood corresponding to the correct prediction $\boldsymbol{\epsilon}_{s_i}$. Hence, a revised objective function for Alice-Bob pair is formulated as

$$\mathcal{L}_{\mathcal{AB}}^{\mathsf{ALC}} = \frac{1}{N_{\mathcal{T}}}\sum_{j=1}^{N_{\mathcal{T}}}\left(d(\boldsymbol{u}_j, \hat{\boldsymbol{u}}_j) + \frac{1}{M}\sum_{i\in[M]} w_i H\Big(q_{\Theta_{E,i}}(s_i^{(j)}|z_i^{(j)}), \bar{p}_L\Big)\right) \tag{8}$$

[4] While our evaluation focused on specific DNN architectures, the framework is model-agnostic and can incorporate more advanced neural architectures, such as transformers. Model-agnostic theoretical bounds on information leakage against advanced architectures are left for future works.

[5] An argmax layer is added over the obtained distribution to take a guess about the predicted secret class $\hat{s}$. This also helps assess the performance of adversaries via measuring their accuracy as the fraction of correct guesses.

(ii) Each step of the joint training for $\mathcal{A}$-to-$\mathcal{B}$ autoencoder $(\Omega_{\mathcal{A}}, \Omega_{\mathcal{B}})$ is followed by a training step for the adversarial DNNs. Eavesdroppers seek to minimize the cross-entropy between their estimated likelihood $q_{\Theta_{E,i}}(s_i|z_i)$ and the one-hot vector $\boldsymbol{\epsilon}_{s_i}$ corresponding to S_i. That is, we have the following loss function for training the adversarial networks.

$$\mathcal{L}_E = \frac{1}{N_{\mathcal{T}}} \sum_{j=1}^{N_{\mathcal{T}}} H\Big(q_{\Theta_{E,i}}(s_i^{(j)}|z_i^{(j)}), \boldsymbol{\epsilon}_{s_i}^{(j)}\Big), \quad i \in [M] \tag{9}$$

Note that all the adversarial networks can be trained in parallel. The distortion measure we consider for our legitimate loss function $\mathcal{L}_{\mathcal{AB}}$ is a mixture of the average MSE, denoted by Δ^{MSE}, and SSIM, Δ^{SSIM}, between the input image $\boldsymbol{u}$ and the recovered pixels $\hat{\boldsymbol{u}}$ at the output of Bobs' DNN. Therefore, we assume $d(\cdot,\cdot)$ to be measured as follows

$$d(\boldsymbol{u}, \hat{\boldsymbol{u}}) = \Delta^{\mathsf{MSE}} + \alpha \Delta^{\mathsf{SSIM}}, \tag{10}$$

where

$$\Delta^{\mathsf{MSE}} = \frac{1}{n}||\boldsymbol{u} - \hat{\boldsymbol{u}}||^2, \quad \Delta^{\mathsf{SSIM}} = 1 - \mathsf{SSIM}(\boldsymbol{u}, \hat{\boldsymbol{u}}), \tag{11}$$

with n denoting the size of the original image vector $\boldsymbol{u} \in \mathbb{R}^n$. Moreover, α is a tuning parameter representing the contribution of SSIM metric in the distortion measure $d(\cdot,\cdot)$. The SSIM measure between two images I and K is defined as

$$\mathsf{SSIM}(I, K) \triangleq \frac{(2\mu_I\mu_K + c_1)(2\sigma_{IK} + c_2)}{(\mu_I^2 + \mu_K^2 + c_1)(\sigma_I^2 + \sigma_K^2 + c_2)}, \tag{12}$$

where μ_I, μ_K, σ_I, σ_K, and σ_{IK} are the local means, standard deviations, and cross-covariance for images I and K. c_1 and c_2 are two adjustable constants [35]. We note that the SSIM index is a *perception-based metric*, which is able to capture pixel inter-dependencies even for spatially close pixels. The rationale behind assuming the proposed distortion metric in (10) is that we not only aim to recover every pixel of a source image with minimum error (realized via MSE measure), but also want to obtain a good-quality reconstruction from the human visual system point of view. We also remark that SSIM calculation is relatively simple, and its derivatives can be easily computed within the gradient descent-based algorithms [36]. Training procedure can be carried out using off-the-shelf stochastic gradient descent algorithms such as the widely-adopted adaptive moment estimation (Adam) optimizer [40].

Remark 4: The main bottleneck for memory and optimization of the proposed framework lies in step (i) of the training procedure. This is due to the fact that all neural networks (Alice, Bob, and all Eves) are involved in training, and the entire computational graph needs to be updated through backpropagation. To mitigate the computational bottleneck of joint training in case the number of adversaries is large, one can define a maximum number of Eves to participate in each step of backpropagation. This can be modeled as a scheduling problem, based on memory storage constraints. For example, if a single adversary can be chosen at each step, each batch of image samples is passed through Alice, Bob, and the chosen Eve every time, realizing "sequential adversarial training."

TABLE 1. Parameters for training the proposed system.

Learning Parameters	Values
Batch size (m)	128
Maximum number of training episodes (N_{episode})	200
Number of warm-up epochs ($N_{\text{warm-up}}$)	50
Number of legitimate training epochs per episode (N_{L})	5
Number of adversarial training epochs per episode (N_{E})	5
Training SNR of legitimate link (Γ_L^{train})	20
Training SNR of adversarial links (Γ_E^{train})	15
Learning rate	10^{-4}
Learning rate drop factor	0.9
Optimizer	Adam [40]

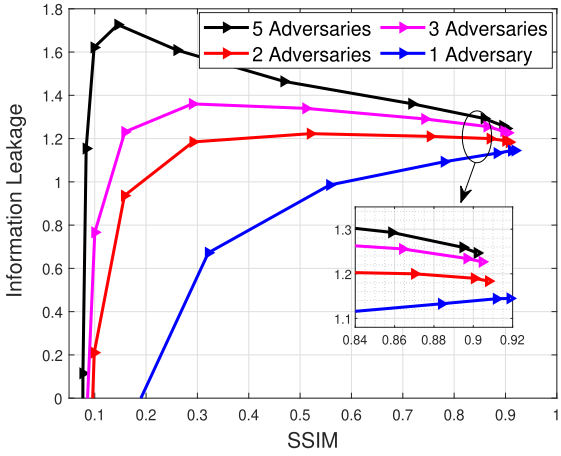
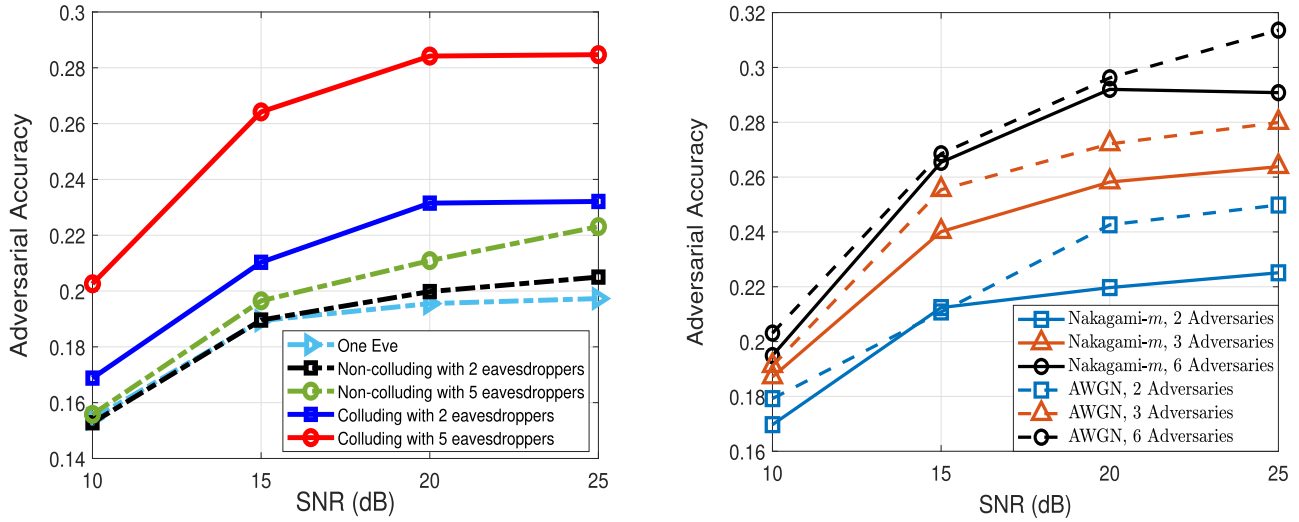


FIGURE 4. Privacy-utility trade-off over CIFAR-10.

IV. IMPLEMENTATION AND EVALUATION

In this section, we present different experiments to demonstrate the performance of our proposed deep learning-based approach in various scenarios. Different benchmarks and ablation studies are also provided. In our experiments, we consider both scenarios of colluding and non-colluding eavesdroppers. In addition, we study both cases of eavesdroppers being interested in a common secret S, and the case in which every eavesdropper is interested in a different secret S_i, using two different datasets of CIFAR-10 [38] and CelebA [39], respectively. We examine the performance over both AWGN and complex fading (Rayleigh and Nakagami-m) communication channels for different channel SNRs and 4 antennas (unless otherwise stated). Moreover, we address the generalization capability of our proposed scheme for different communication scenarios and over a wide range of channel SNRs. We also study the *privacy-utility trade-off* for the proposed learning-based approach, which provides useful insights. To quantify the level of privacy, we evaluate the information leakage characterized via the mutual information metric. Accordingly, with a similar approach to what proposed in (2), we use the following approximation for mutual



(a) Adversarial accuracy (Rayleigh)

(b) Adversarial accuracy (Nakagami-m)

FIGURE 5. Adversarial accuracy vs. SNR (Γ_E) for Rayleigh fading, AWGN, and Nakagami-m (with $m = 3$) channels.

information leakage

$$I(S_i, Z_i^k) \approx H(S_i) - H\big(q_{\Theta_{E,i}}(s_i|z_i), \boldsymbol{\varepsilon}_{si}\big)\,, \tag{13}$$

where the entropy of private attributes $H(S_i)$ are numerically calculated for each of the datasets according to the empirical distribution of CIFAR-10 and CelebA classes.

The SNRs of communication links are defined as:

$$\Gamma_L = 10\log_{10}\frac{P}{\sigma_L^2}\ \text{(dB)}, \quad \Gamma_E = 10\log_{10}\frac{P}{\sigma_E^2}\ \text{(dB)}, \tag{14}$$

which represent the ratio of the average power of the encoded data-stream of the latent space (i.e., the channel input) to the average noise power of legitimate σ_L^2 and adversarial nodes σ_E^2, respectively. Without loss of generality, we set the average signal power to $P = 1$ for all experiments, while varying the SNR by setting the noise variances σ_L^2 and σ_E^2. The std of the eavesdropping channels are set to $\{0.04, 0.16, 0.36, 0.64, 1, 1.44\}$ for $i \in [6]$, for $M = 6$ simulated eavesdroppers in total. The bandwidth compression ratio is set to $\frac{k}{n} = \frac{1}{3}$. For the training, we have assumed the general case of complex Rayleigh fading channel model which is treated by the joint effect of channel fading and additive noise, i.e., $\eta(\boldsymbol{\tau}) = \eta_v(\eta_h(\boldsymbol{\tau})) = h\boldsymbol{\tau} + \boldsymbol{v}$ with $h \sim \mathcal{CN}(0, 1)$. Nevertheless, during the inference phase we study the performance of our proposed scheme in different scenarios of AWGN and Nakagami-m channels $|h|^2 \sim \mathcal{G}(m, \delta_h^2/m)$.

For simplicity, we consider a single weight in the loss function, that is, $w_i = w, \forall i$. We set $w = 5$ and $\alpha = 0.1$ during training. The values of w and α were chosen empirically. We first defined a set of plausible values for α and w, trained the system for all combinations of (α, w) pairs, and evaluated the performance metrics on a validation set, and finally chose the best desired parameters. We used TensorBoard[6] for tracking and visualizing performance metrics during training, and we ran multiple combinations of α and w, and constantly monitored the loss function and other performance metrics such as SSIM, MSE, and cross-entropy over epochs. In Appendix A, the evolution of performance metrics over the training steps are visualized. Our observations have shown that for a given SNR, the performance is not too sensitive to the choice of these parameters, (although there is no guarantee that a selected set of parameters will be optimal for all ranges of SNRs).

We have implemented our proposed network architecture (shown in Figs. 2 and 3) using Python3 with Tensorflow [37]. The codes were run on Intel(R) Xeon(R) Silver 4114 CPU running at 2.20 GHz with GeForce RTX 2080 Ti GPU. To minimize the legitimate and adversarial loss functions (LF), we have chosen the widely-adopted Adam optimizer algorithm [40]. Similarly to [16], we choose to fix the number of training episodes N_{episode} in advance, where the value of $N_{\text{episode}} = 200$ is set according to our experiments. The learning parameters used during the training process are summarized in Table 1. The detailed description of training process can be stated as follows. i) During a ***warm-up phase***, both legitimate pairs and adversarial networks are independently trained in order to reach a plausible initial state of reconstructing the image. To elaborate, for $N_{\text{warm-up}} = 50$ epochs, the encoder-decoder pair of Alice and Bob $(\Omega_{\mathcal{A}}, \Omega_{\mathcal{B}})$ are trained according to the reconstruction objective of minimizing $\mathcal{L}_{\mathcal{AB}}^{\text{warm-up}} = \frac{1}{N_{\mathcal{T}}}\sum_{i=1}^{N_{\mathcal{T}}} d(\boldsymbol{u}_i, \hat{\boldsymbol{u}}_i)$, where $d(\cdot,\cdot)$ is given in (10). Following the legitimate warm-up, the adversarial networks $(\Theta_{E,1}, \cdots, \Theta_{E,M})$ are trained for $N_{\text{warm-up}} = 50$ epochs, while the legitimate encoder $\Omega_{\mathcal{A}}$ is frozen. ii) After performing the warm-up, the main

[6]https://www.tensorflow.org/tensorboard

phase of ***minimax training*** starts. In this phase, for each of the $N_{\text{episode}} = 200$ episodes, Alice and Bob aim to learn the optimal mapping for the latent space and the decoding neural structure so that the reconstruction distortion and information leakage are minimized. Hence, the legitimate encoder-decoder counterparts $(\Omega_{\mathcal{A}}, \Omega_{\mathcal{B}})$ are trained for $N_{\text{L}} = 5$ epochs based on minimizing the LF $\mathcal{L}^{\text{ALC}}_{\mathcal{AB}}$ given in (8). Since the neural encoder $\Omega_{\mathcal{A}}$ at Alice has learned a new embedding for the original data, the adversaries are then trained for $N_{\text{E}} = 5$ epochs to learn and adjust their adversarial role in extracting sensitive information from the newly-learned latent space based on the LF, $\mathcal{L}_E$, given in (9). In this manner, the concept of minimax game between $(\mathcal{A}, \mathcal{B})$ and the set of eavesdroppers is realized, while maintaining the fairness for the learning process of legitimate and adversarial nodes.

A. EVALUATION ON CIFAR-10 DATASET

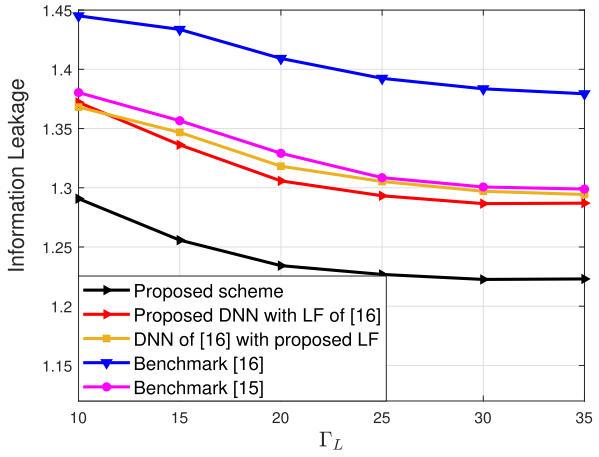
We evaluate our proposed framework using images with dimension $32 \times 32 \times 3$ (height, width, channels) from the CIFAR-10 dataset [38]. The dataset consists of 60000 colored images of size 32×32 pixels. The training and evaluation sets are two completely separated sets of images, containing 50000 and 10000 images, respectively, associated with 10 classes. Adversaries wish to infer a common secret S from the received noisy encoded signals. Hence, the common secret S here is considered as the class of images with $|S| = L = 10$.

Fig. 4 demonstrates the *privacy-utility trade-off* for our proposed scheme. The figure is obtained for the legitimate SNR from $\Gamma_L = -20$dB to 30dB with steps of 5 dB. The figure addresses the fact that by improving the quality of the reconstructed image in SSIM, it is inevitable to have a certain level of information leakage in general. This phenomena, which is actually in accordance with one's intuition, also verifies our generic mathematical framework proposed in (1). The point of this figure is that by increasing the SNR and gradually improving the quality of data reconstruction (increasing the SSIM at Bob), the information leakage soon saturates. After that, we can achieve higher SSIMs without any significant increase in the leakage. For instance, having $M = 3$ adversaries, by setting $\Gamma_L \geq 5$ dB, we can enhance the image reconstruction at Bob's decoder, without any further decrease in information. In this case, 15 dB increase in channel SNR can improve the SSIM about 56%. Therefore, such trade-off curves can help network designers adjust the system parameters to achieve desired levels of privacy and utility. The (small) negative impact of adversaries on the SSIM values could be explained by the minimax game invoked in Sections III-C and IV, which mutually affects the performance of legitimate counterparts as well.

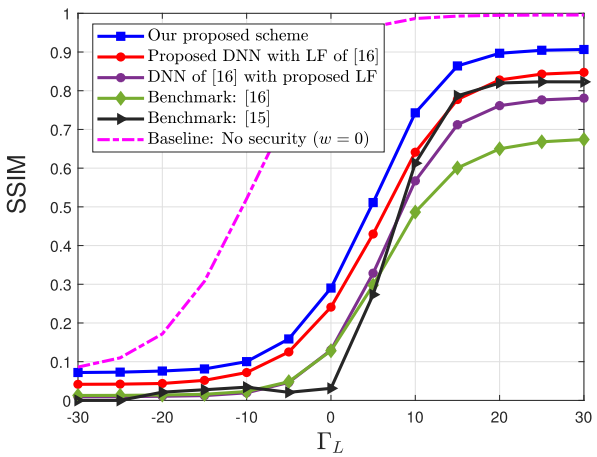
1) ADVERSARIAL STRATEGIES AND EAVESDROPPING PERFORMANCE EVALUATION

We study the performance of our proposed scheme under two different adversarial strategies: Adversaries may infer private attributes either individually (non-colluding case), or by learning from the combination of the individual logits obtained by each eavesdropper (colluding setup). Mathematically speaking, for the non-colluding setup, the training is performed based on (8)–(10), and the eavesdroppers try to extract the common S or individual secret S_i based on their own loss function in (9). Meanwhile, for the colluding setup (with a common secret S) an extra level of "knowledge combination" is performed, where adversaries employ *ensemble learning* method and combine the individually-extracted logits to infer the secret S. That is, a weighted sum in the form of $\sum_{i\in[M]} \varphi_i q_{\Theta_{E,i}}(s_i|z_i)$ is considered in the training, where ω_i's are learned by the set of colluding eavesdroppers via ensemble learning.

Figs. 5-(a) and 5-(b) demonstrate the adversarial accuracy of our proposed scheme for the scenario of colluding eavesdroppers. For this experiment, we study the performance of *colluding adversarial inference,* and compare it with the non-colluding setup. We note that for the non-colluding setup, the training is performed based on (8)–(10), and the eavesdroppers try to extract the common or individual secret S_i based on their own loss functions given in (9). Meanwhile, for the colluding setup (with a common secret S) an extra level of "knowledge combination" is performed afterward—we combine the individually-extracted logits of adversaries, and leverage the concept of *ensemble learning* to infer the secret S. For the colluding setup, the adversarial performance is measured by the overall accuracy of adversaries in correctly finding the ground-truth $\boldsymbol{\varepsilon}_s$ from their aggregated logits, while for the non-colluding benchmarks, the mean accuracy across eavesdroppers is plotted for the sake of comparison. The overall accuracy of adversaries is plotted versus channel SNR of the wiretap links, Γ_E, for different number of eavesdroppers and different types of communication channels. We can observe from the figures that increasing the number of eavesdroppers results in achieving higher accuracy for the adversaries, which is aligned with one's intuition. The increase in adversarial accuracy is more significant in the colluding case, due to the collaboration and knowledge combination among eavesdroppers which helps them learn the secret more accurately. Figs. 5-(a) and 5-(b) indicate that by increasing the quality of adversarial links, i.e., increasing Γ_E, the accuracy of adversaries increases by at most 10%. This can be observed from (14), where higher values for Γ_E results in having less-distorted (less noisy) observations $Z_i^k = z_i$, $i \in [M]$, at the adversarial DNNs, which results in more accurate estimations about the posterior adversarial distribution $q_{\Theta_{E,i}}(s|z_i)$, with respect to the ground-truth $\boldsymbol{\varepsilon}_s$. Interestingly, the amount of increase in the adversarial accuracy reduces with the increase in Γ_E, which highlights the limitation of adversarial nodes based on our proposed scheme. Fig. 5-(b) also addresses the generalization capability of our learning-based framework extended to the AWGN communication model and Nakagami-m scenario $|h|^2 \sim \mathcal{G}(m, 1/m)$. This experiment verifies that we can achieve almost similar performance in other channel scenarios, despite being trained for



(a) Privacy performance in terms of information leakage $I(S; Z_i^k)$



(b) Reconstruction performance in terms of SSIM

FIGURE 6. Performance evaluation during inference.

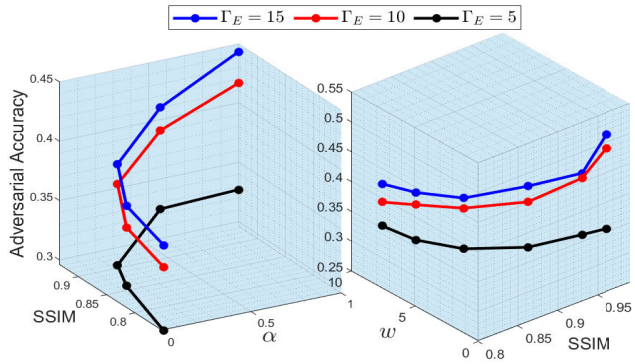


FIGURE 7. Effect of α and w on the privacy-utility trade-off.

the Rayleigh fading case, which highlights the *robustness* and *generalizability* of our proposed scheme.

Fig. 6-(a) demonstrates the information leakage for our proposed scheme compared with different benchmarks. Notably, this figure highlights the superiority of our approach compared to the benchmarks of [15] and [16] in terms of the mutual information. This figure also implies that increasing the SNR of legitimate link Γ_L, results in having lower information leakage This is because increasing Γ_L (given in (14)) results in less noisy observations at Bob than the adversarial DNNs. Therefore, Alice can better hide the sensitive attributes in noise, making it harder for the adversaries to extract the sensitive information from distorted received data.

Fig. 6-(b) illustrates the data reconstruction performance of our proposed framework compared with other benchmarks. One can infer from this figure and Fig. 6-(a) that our proposed system outperforms the benchmarks in terms of both information leakage and utility. Accordingly, 20% and 10% performance gain is achieved by our proposed scheme compared with [15] and [16], respectively. Fig. 6-(b) also implies that increasing Γ_L results in having higher SSIM values. This is because increasing Γ_L results in less distorted observations at Bob, which facilitates the image reconstruction performance of Bob's neural decoder. Notably, Fig. 6-(b) highlights that if we neglect the existence of adversaries and set $w = 0$, our proposed scheme can achieve almost perfect ($\mathsf{SSIM} = 1$) data recovery. The ablation-like examinations conducted in Figs. 6-(a) and 6-(b) imply that both the DNN structure and the employed LF for optimizing the framework contribute to the overall privacy and reconstruction performance. Specifically, our proposed DNNs at Alice and Bob, together with the proposed objective function elaborated in (6) and (8)–(10) have resulted in achieving the best performance compared with other benchmarks and baselines.

Fig. 7 studies the impact of tuning parameters α and w on the adversarial performance of our proposed system. These tuning parameters are the coefficients associated with utility and privacy adjustment within our training LF in (10) and (8), respectively. For this experiment, the adversarial performance is captured by investigating the accuracy of adversaries in correctly finding the ground-truth label $\boldsymbol{\varepsilon}_s$ (representing the sensitive information S) out of $L = 10$ labels of CIFAR-10 dataset. For the accuracy measurement in this experiment, we consider the pessimistic scenario in which the correct guess of true labels by any adversary $i \in [M]$ increments the total accuracy of adversaries. The figure indicates that by increasing α, higher values of SSIM can be achieved, since more emphasis on the SSIM error Δ^{SSIM} is put based on (10). However, the accuracy of adversaries in extracting the private information also increases, which verifies the privacy-utility trade-off. Notably, by increasing α to values more than 0.1, a jump in the adversarial accuracy can be observed, which lead us choosing $\alpha = 0.1$ for our network. Similarly, by increasing w, the emphasis goes toward the secrecy criteria introduced in (1), (6), and (8), which leads to the reduction in adversarial accuracy and achieved SSIM, verifying the privacy-utility trade-off as well. The figure also

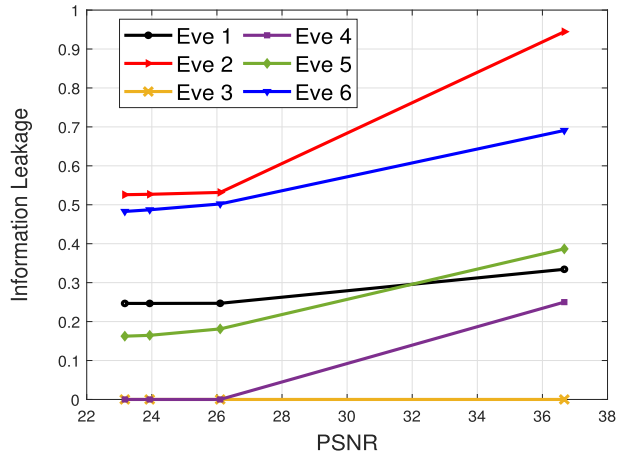


FIGURE 8. **Privacy-utility trade-off over CelebA dataset.**

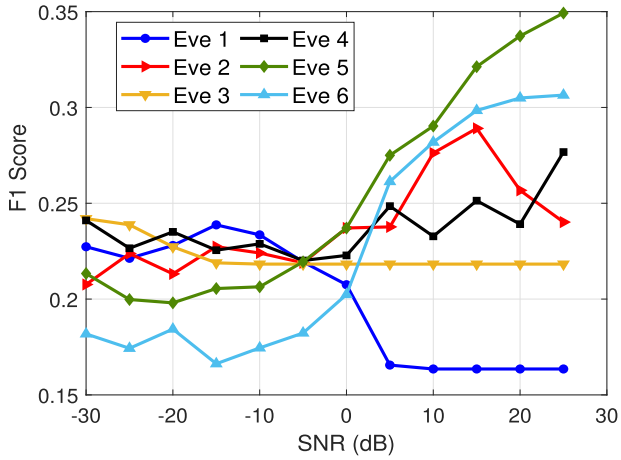


FIGURE 9. **Adversarial accuracy in terms of F1 score over CelebA dataset.**

TABLE 2. **Privatized CelebA attributes for each eavesdropper.**

Eavesdropper	Private Attributes
Eve 1	`Wavy_Hair` and `Black_Hair`
Eve 2	`Wearing_Lipstick` and `Smiling`
Eve 3	`Double_Chin` and `Wearing_Necklace`
Eve 4	`No_Beard` and `5_o_Clock_Shadow`
Eve 5	`Bags_Under_Eyes` and `Arched_Eyebrows`
Eve 6	`High_Cheekbones` and `Pointy_Nose`

shows that increasing the adversary channels SNR Γ_E can improve the adversarial accuracy in finding the sensitive data $\boldsymbol{\epsilon}_s$. Interestingly, the amount of increase in the adversarial accuracy reduces with the increase in Γ_E which highlights the limitation of adversarial nodes based on our proposed scheme.

B. EVALUATION ON CELEBA DATASET

In this subsection, we evaluate our proposed scheme on CelebA dataset [39]. This is a large-scale face attributes dataset with more than 200k images of celebrities with 40 attribute annotations. For the following experiments, we consider the non-colluding scenario, in which eavesdroppers are interested in *different secrets* $S_i \in \mathcal{S}_i$, while we assume $|S_i| = 4$ for all eavesdroppers. Details of the privatized attributes for each eavesdropper $i \in [M]$ are provided in Table 2.

Fig. 8 demonstrates the privacy-utility trade-off over CelebA dataset with SNR $\Gamma_E = 25$dB. Since the eavesdroppers are interested in different secrets, the results are plotted for each individual eavesdropper $i \in [6]$. The figure is obtained by changing the hyperparameter $w \in \{0, 10, 50, 100\}$. In this figure, the extreme case of $w = 100$ has the minimum leakage while having the minimum utility (measured by the PSNR in this experiment). This is because larger w implies that DNN encoder-decoder pair gives more importance to the secrecy criteria in (6)–(8). However, as we gradually decrease w, higher PSNRs can be achieved with a slight increase in the information leakage. The figure also implies that for the baseline of no security design, i.e., $w = 0$, the information leakage increases by about 30%, which highlights the importance of designing a secure neural encoding-decoding pair.

Remark 5: We remark that the privacy-utility trade-off curves enable users to dynamically select the *operation point* that matched their specific requirements, balancing between utility and privacy. We quantified this balance with hyperparameter w, where our experiments demonstrate how different w values affect this balance - higher values favor utility (image quality) while lower values prioritize privacy.

Fig. 9 shows the accuracy of eavesdroppers versus Γ_E for $w = 10$. Since CelebA is a highly *imbalanced dataset* with respect to most of the attributes, we have chosen the F-score metric with macro averaging to measure the adversarial accuracy in this scenario. Note that due to the imbalance in CelebA dataset attributes, we employed a penalty mechanism. However, we still observe the effect of the imbalance on the eavesdroppers in terms of F-scores. The figure indicates that our privacy-aware neural encoding-decoding pair has been successful in confusing Eves about the private attributes, as a maximum of 25% accuracy is achieved by adversaries for most of the SNR values. Similar to the results over CIFAR-10, this figure also indicates that increasing the quality of adversarial links can increase the accuracy of eavesdroppers by at most 15%.

Fig. 10 illustrates the reconstruction quality (for different values of w) at Bob and each of the eavesdroppers. We can observe that the bigger w, the worse would be the reconstruction quality at Eves (and also at Bob), as the emphasis shifts toward the secrecy criteria. The figure also implies that the eavesdroppers are strongly restrained from reconstructing any meaningful (high perceptual quality) images, due to very low adversarial SSIM values.

To gain visual insights into the reconstruction performance of Bob, and what the reconstructed images at eavesdroppers would be like, Fig. 11 is provided. Please note that the goal of eavesdroppers is to infer the secrets rather than reconstruct

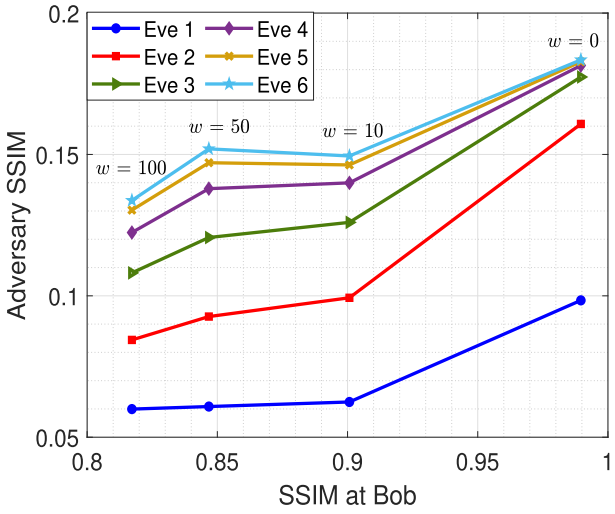


FIGURE 10. **Reconstruction quality of the legitimate destination vs. the eavesdroppers over CelebA dataset.**

the entire image. Here, we show the reconstructions at the eavesdroppers only for illustration purposes. To this end, the legitimate DNN encoder is frozen after being trained for 180 epochs, and then, a similar neural structure to that of Bob is considered for each of the eavesdroppers, which gets trained for image reconstruction.[7] The reconstructed images in Fig. 11 correspond to the default case of $w = 10$, and the extreme case of $w = 100$. Interestingly, one can observe that our proposed framework has been successful in hiding the corresponding private information against each eavesdropper, e.g., the hair style, smiling, facial attributes, etc., while recovering the entire image at Bob such that it can be distinguished clearly. Notably, although the main objective of the legitimate encoder was to prevent adversaries from extracting the privatized attributes, we observe that the eavesdroppers cannot successfully reconstruct the images either.

V. CONCLUSION AND FUTURE DIRECTIONS

In this paper, we proposed a privacy-aware DeepJSCC framework for end-to-end image transmission in the presence of multiple eavesdroppers. DeepJSCC pipeline was designed to transmit images over the communication channel with the highest reconstruction fidelity, while controlling the privacy leakage theoretically measured by the mutual information metric. Autoencoders (with convolutional neural networks as backbone) were implemented for neural encoder-decoder pair, along with an information-theoretic "privacy funnel" framework as the training loss function, realizing the privacy-utility trade-off while taking perceptual quality into account. Extensive numerical experiments were carried out over CIFAR-10 and CelebA dataset. Ablation studies and data visualizations have been proposed to validate the performance gain of our scheme compared with related benchmarks, both numerically and visually. While datasets like CIFAR and CelebA seem to be useful

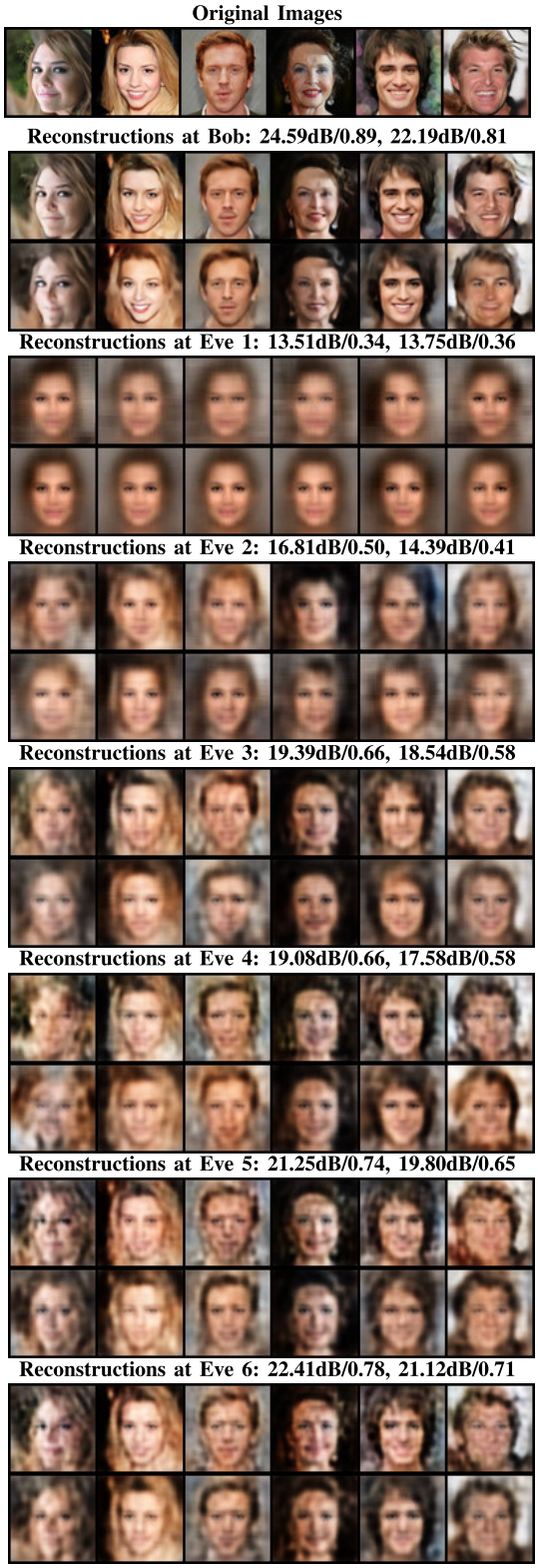


FIGURE 11. **Reconstructed CelebA images at Bob and each Eve.For each set of images, the first and second rows correspond to $w = 10$ and $w = 100$, respectively.PSNR(dB)/SSIM values are provided above each set of images.**

[7]During the evaluation, we transmit each image 10 times and the performance metrics are averaged over these realizations.

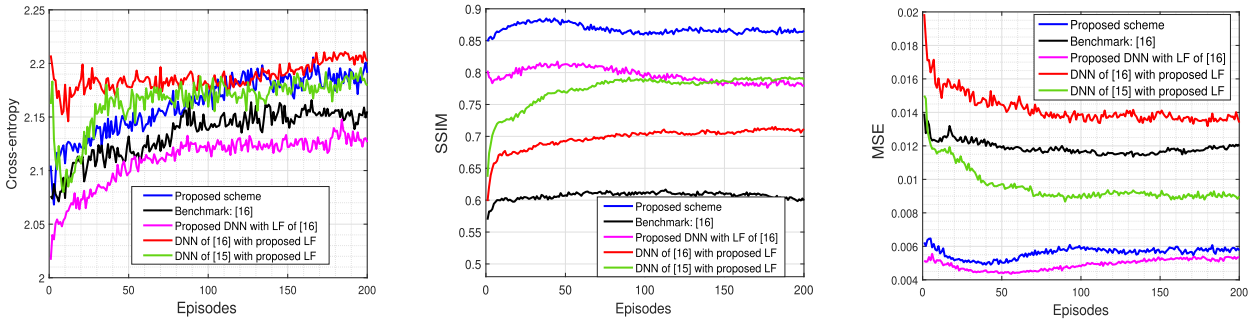


FIGURE 12. The training process and ablation studies.The cross-entropy, SSIM, and MSE metrics are visualized from left to right.

for testing specific downstream tasks like classification or retrieval, in order to ensure out-of-distribution (OOD) generalization, the system might need to be trained on richer datasets suited specifically for image compression and decompression tasks. Studying OOD performance under general-purpose datasets can be studied in the future works.

Some of the key future directions in the context of DeepJSCC and information-theoretic privacy can be outlined below.

1) *Training over real-world measurement data:* In this paper, we considered known channel models to generate samples from conditional channel distributions. One could easily drop this assumption if having enough channel state information (CSI) measurements collected from a particular wireless channel with unknown statistics. However, training wireless systems over real-world wireless channels is the relatively long coherence time compared to the rate at which data samples can be processed for training. Thus, a few channel realizations are observed over every mini-batch which may lead to poor sample efficiency and slow convergence. This could be addressed in future works.
2) *Robustness against advanced adversarial strategies:* In this paper we used adversarial likelihood compensation to push adversaries toward uniform predictions. Another direction to study would be a scenario in which adversaries employ more sophisticated strategies like reinforcement learning or meta-learning. Such scenarios reflect the ‘‘asymmetric information’’ case from game-theoretic perspective, while in this paper we considered the ‘‘complete information’’ case, where both sides know each other’s strategy. Our approach leads to Nash Equilibrium, and both parties can optimize their best responses resulting in more predictable and stable outcomes. However, asymmetric information scenario may lead to more dynamic and adaptive game playing and creates an uncertainty which can be studied in the future works.
3) *The role of novel generative models:* Integrating novel generative models, such as denoising diffusion probabilistic models, into DeepJSCC frameworks can help enhance the robustness of wireless system against adversarial attacks. Diffusion models, with their underlying capability to learn complicated distributions and generate content out of noisy signals, have been shown to be effective in multimedia transmission [41], removing unwanted signals in multi-user systems [42], while taking practical assumptions of wireless systems into account [43]. An interesting direction would be to study the potential roles of such generative models at the encoder side, and also the decoder side, investigating whether such generative foundation models can substitute encoder-dependent decoders, and realize a more generalized model-agnostic multimedia transmission framework.

APPENDIX A
TRAINING PROCESS AND ABLATION STUDY

Fig. 12 demonstrates the training process. Three main performance metrics for training, i.e., the cross-entropy (averaged over all training Eves), SSIM, and MSE are examined.

The cross-entropy results show that during our *fair minimax training* between the legitimate neural pair and the adversarial DNNs, our proposed scheme gradually learns to minimize the sensitive information leakage, as the average cross-entropy is increasing over episodes. The figure also shows that the training process has been successful in restricting adversaries, with respect to information leakage, as the training steps go on. Our ablation studies in Fig. 12 highlight the importance of loss function (LF) in controlling the secrecy performance against adversarial DNNs. This can be seen by investigating the red, green and blue curves, which show almost similar cross-entropy values for the same LF but different DNNs.

Fig. 12 also demonstrates data reconstruction performance during training, highlighting the outperformance of our approach. This is because we take into account the perceptual characteristics of image both in the implemented DNN architecture ({Conv2d + GDN} units), as well as the proposed LF (considering the SSIM metric). The figure highlights the importance of DNN architecture in the achievable

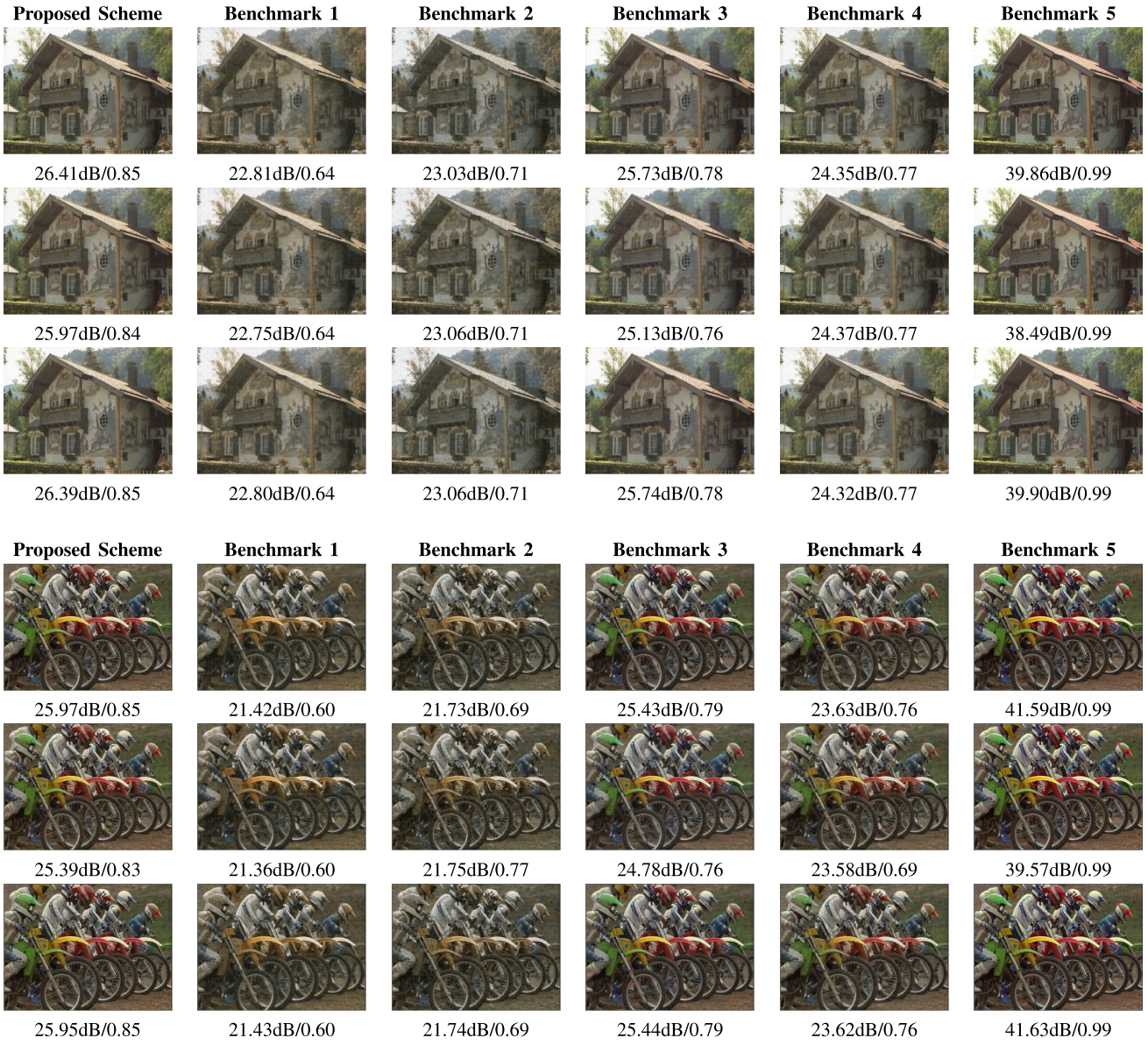


FIGURE 13. Generalization performance of our scheme (pre-trained on CIFAR-10 dataset) evaluated over Kodak dataset and compared with different benchmarks.PSNR/SSIM metrics are reported below each image.

reconstruction performance. Notably, by implementing the neural architecture of [16] while considering our proposed LF, the reduction in reconstruction quality in terms of SSIM is three times more than the case in which we use our proposed DNN architecture with the LF of [16].

APPENDIX B
FURTHER EXPERIMENTS ON KODAK DATASET

Fig. 13 provides a visual comparison on the generalization performance of our system over Kodak dateset and with different benchmarks. For this experiment, channel SNRs are set to $\Gamma_E = 15$ dB and $\Gamma_L = 20$ dB. From top to bottom, each row corresponds to the Rayleigh-fading, AWGN, and Nakagami-m channel models, respectively. Benchmark 1 corresponds to the scheme proposed in [16]. Benchmark 2 corresponds to the DNN proposed in [16], while utilizing our proposed loss function. Benchmark 3 visualizes the performance of a system in which our proposed neural encoder-decoder are trained via the training framework proposed of [16]. Benchmark 4 corresponds to the scheme proposed in [15]. Benchmark 5 reflects the case in which the security requirement is neglected by setting $w_i = 0$ and training the encoder-decoder pair without any security constraint. The last benchmark indicates that if there is no security requirement, our scheme can reconstruct images with almost lossless reconstruction over different channel models. It also highlights the security-utility trade-off, showing that for achieving security/privacy-aware

deep-JSCC, around 15% decrease in reconstruction performance (in terms of SSIM) can be imposed to the model. Fig. 13 validates the outperformance of our scheme compared with other benchmarks, both visually and numerically. This can not only be seen from the PSNR/SSIM values, but also from the visual details (e.g., in the first image) and the color combinations (e.g., in the second image) of the reconstructed images. Comparing the PSNR/SSIM values of benchmarks 1 to 4 highlights the importance of implementing well-designed DNN architecture for the neural encoder-decoder pairs, which seems to have more impact on the reconstruction performance compared to the employed loss function.

ACKNOWLEDGMENT

An earlier version of this paper was presented in part at the IEEE Global Communications Conference, Kuala Lumpur, Malaysia, December 2023 [DOI: 10.1109/GLOBECOM54140.2023.10436928]. *(Mehdi Letafati and Seyyed Amirhossein Ameli Kalkhoran contributed equally to this work.)*

REFERENCES

[1] S. A. A. Kalkhoran, M. Letafati, E. Erdemir, B. H. Khalaj, H. Behroozi, and D. Gündüz, "Secure deep-JSCC against multiple eavesdroppers," in *Proc. IEEE Global Commun. Conf.*, Kuala Lumpur, Malaysia, Dec. 2023, pp. 3433–3438.

[2] F. Hu et al., "Cellular-connected wireless virtual reality: Requirements, challenges, and solutions," *IEEE Commun. Mag.*, vol. 58, no. 5, pp. 105–111, May 2020.

[3] M. Jankowski, D. Gündüz, and K. Mikolajczyk, "Wireless image retrieval at the edge," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 89–100, Jan. 2021.

[4] M. Letafati, H. Behroozi, B. H. Khalaj, and E. A. Jorswieck, "Deep learning for hardware-impaired wireless secret key generation with man-in-the-middle attacks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Madrid, Spain, Dec. 2021, pp. 1–6.

[5] M. Letafati, H. Behroozi, B. H. Khalaj, and E. A. Jorswieck, "Learning-based secret key generation in relay channels under adversarial attacks," *IEEE Open J. Veh. Technol.*, vol. 4, pp. 749–764, 2023.

[6] M. Letafati, A. Kuhestani, H. Behroozi, and D. W. K. Ng, "Jamming-resilient frequency hopping-aided secure communication for Internet-of-Things in the presence of an untrusted relay," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6771–6785, Oct. 2020.

[7] M. Letafati, H. Behroozi, B. H. Khalaj, and E. A. Jorswieck, "Hardware-impaired PHY secret key generation with man-in-the-middle adversaries," *IEEE Wireless Commun. Lett.*, vol. 11, no. 4, pp. 856–860, Apr. 2022.

[8] M. Letafati, H. Behroozi, B. H. Khalaj, and E. A. Jorswieck, "Wireless-powered cooperative key generation for e-health: A reservoir learning approach," in *Proc. IEEE 95th Veh. Technol. Conf. (VTC-Spring)*, Helsinki, Finland, Jun. 2022, pp. 1–7.

[9] M. Letafati, A. Kuhestani, K.-K. Wong, and M. J. Piran, "A lightweight secure and resilient transmission scheme for the Internet of Things in the presence of a hostile jammer," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4373–4388, Mar. 2021.

[10] A. Chorti et al., "Context-aware security for 6G wireless: The role of physical layer security," *IEEE Commun. Standards Mag.*, vol. 6, no. 1, pp. 102–108, Mar. 2022.

[11] M. Letafati, H. Behroozi, B. H. Khalaj, and E. A. Jorswieck, "On learning-assisted content-based secure image transmission for delay-aware systems with randomly-distributed eavesdroppers," *IEEE Trans. Commun.*, vol. 70, no. 2, pp. 1125–1139, Feb. 2022.

[12] M. Letafati, H. Behroozi, B. H. Khalaj, and E. A. Jorswieck, "Content-based medical image transmission against randomly-distributed passive eavesdroppers," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Montreal, QC, Canada, Jun. 2021, pp. 1–7.

[13] D. Gündüz, M. A. Wigger, T.-Y. Tung, P. Zhang, and Y. Xiao, "Joint source–channel coding: Fundamentals and recent progress in practical designs," *Proc. IEEE*, pp. 1–32, Nov. 2024, doi: 10.1109/JPROC.2024.3477331.

[14] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 3, pp. 567–579, Sep. 2019.

[15] D. B. Kurka and D. Gündüz, "DeepJSCC-f: Deep joint source-channel coding of images with feedback," *IEEE J. Sel. Areas Commun.*, vol. 1, no. 1, pp. 178–193, May 2020.

[16] T. Marchioro, N. Laurenti, and D. Gündüz, "Adversarial networks for secure wireless communications," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 8748–8752.

[17] E. Erdemir, P. L. Dragotti, and D. Gündüz, "Privacy-aware communication over a wiretap channel with generative networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Singapore, May 2022, pp. 2989–2993.

[18] J. Xu, T.-Y. Tung, B. Ai, W. Chen, Y. Sun, and D. D. Gündüz, "Deep joint source-channel coding for semantic communications," *IEEE Commun. Mag.*, vol. 61, no. 11, pp. 42–48, Nov. 2023.

[19] K.-L. Besser, C. R. Janda, P.-H. Lin, and E. A. Jorswieck, "Flexible design of finite blocklength wiretap codes by autoencoders," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 2512–2516.

[20] K.-L. Besser, P.-H. Lin, C. R. Janda, and E. A. Jorswieck, "Wiretap code design by neural network autoencoders," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3374–3386, 2019.

[21] R. Fritschek, R. F. Schaefer, and G. Wunder, "Deep learning based wiretap coding via mutual information estimation," in *Proc. 2nd ACM Workshop Wireless Secur. Mach. Learn.*, New York, NY, USA, Jul. 2020, pp. 74–79.

[22] T.-Y. Tung and D. Gündüz, "Deep joint source-channel and encryption coding: Secure semantic communications," in *Proc. IEEE Int. Conf. Commun.*, Rome, Italy, May 2023, pp. 5620–5625.

[23] J. Xu, B. Ai, W. Chen, N. Wang, and M. Rodrigues, "Deep joint source-channel coding for image transmission with visual protection," *IEEE Trans. Cognit. Commun. Netw.*, vol. 9, no. 6, pp. 1399–1411, Dec. 2023.

[24] M. Zhang, Y. Li, Z. Zhang, G. Zhu, and C. Zhong, "Wireless image transmission with semantic and security awareness," *IEEE Wireless Commun. Lett.*, vol. 12, no. 8, pp. 1389–1393, Aug. 2023.

[25] S. Tang, C. Liu, Q. Yang, S. He, and D. Niyato, "Secure semantic communication for image transmission in the presence of eavesdroppers," 2024, *arXiv:2404.12170*.

[26] Y. E. Sagduyu, T. Erpek, S. Ulukus, and A. Yener, "Is semantic communication secure? A tale of multi-domain adversarial attacks," *IEEE Commun. Mag.*, vol. 61, no. 11, pp. 50–55, Nov. 2023.

[27] J. Ballé, V. Laparra, and E. P. Simoncelli, "Density modeling of images using a generalized normalization transformation," Nov. 2025, *arXiv:1511.06281*.

[28] N. Merhav, "On joint coding for watermarking and encryption," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 190–205, Jan. 2006.

[29] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Médard, "From the information bottleneck to the privacy funnel," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Nov. 2014, pp. 501–505.

[30] E. Bjornson, E. A. Jorswieck, M. Debbah, and B. Ottersten, "Multiobjective signal processing optimization: The way to balance conflicting metrics in 5G systems," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 14–23, Nov. 2014.

[31] B. Razeghi, F. P. Calmon, D. Gunduz, and S. Voloshynovskiy, "Bottlenecks CLUB: Unifying information-theoretic trade-offs among complexity, leakage, and utility," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 2060–2075, 2023.

[32] A. Tripathy, Y. Wang, and P. Ishwar, "Privacy-preserving adversarial networks," in *Proc. 57th Annu. Allerton Conf. Commun. Control Comput. (Allerton)*, 2019, pp. 495–505.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[34] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[35] J. Snell, K. Ridgeway, R. Liao, B. D. Roads, M. C. Mozer, and R. S. Zemel, "Learning to generate images with perceptual similarity metrics," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 4277–4281.

[36] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Mar. 2017.

[37] M. Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Accessed: Mar. 25, 2025. [Online]. Available: https://www.tensorflow.org/

[38] A. Krizhevsky, "Learning multiple layers of features from tiny images," Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 2009.

[39] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–13.

[41] M. Letafati, S. Ali, and M. Latva-Aho, "Conditional denoising diffusion probabilistic models for data reconstruction enhancement in wireless communications," *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 3, pp. 133–146, 2025.

[42] M. Letafati, S. Ali, and M. Latva-Aho, "Diffusion model-aided data reconstruction in cell-free massive MIMO downlink: A computation-aware approach," *IEEE Wireless Commun. Lett.*, vol. 13, no. 11, pp. 3162–3166, Nov. 2024.

[43] M. Letafati, S. Ali, and M. Latva-Aho, "Denoising diffusion probabilistic models for hardware-impaired communications," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Dubai, United Arab Emirates, Apr. 2024, pp. 1–6.