# Secure Deep-JSCC Against Multiple Eavesdroppers

Seyyed Amirhossein Ameli Kalkhoran*†, Mehdi Letafati*‡, Ecenaz Erdemir§, Babak Hossein Khalaj‡, Hamid Behroozi‡, and Deniz Gündüz§

‡ Electrical Engineering Department, Sharif University of Technology, Tehran, Iran
† Computer Engineering Department, Sharif University of Technology, Tehran, Iran
§ Department of Electrical and Electronic Engineering, Imperial College London, UK
Emails: ‡{mletafati@ee., behroozi@, khalaj@}sharif.edu; †ameli@ce.sharif.edu; §{e.erdemir17, d.gunduz}@imperial.ac.uk

*Abstract*—In this paper, a generalization of deep learning-aided joint source channel coding (Deep-JSCC) approach to secure communications is studied. We propose an end-to-end (E2E) learning-based approach for secure communication against multiple eavesdroppers over complex-valued fading channels. Both scenarios of colluding and non-colluding eavesdroppers are studied. For the colluding strategy, eavesdroppers share their logits to collaboratively infer private attributes based on ensemble learning method, while for the non-colluding setup they act alone. The goal is to prevent eavesdroppers from inferring private (sensitive) information about the transmitted images, while delivering the images to a legitimate receiver with minimum distortion. By generalizing the ideas of privacy funnel and wiretap channel coding, the trade-off between the image recovery at the legitimate node and the information leakage to the eavesdroppers is characterized. To solve this secrecy funnel framework, we implement deep neural networks (DNNs) to realize a data-driven secure communication scheme, without relying on a specific data distribution. Simulations over CIFAR-10 dataset verifies the secrecy-utility trade-off. Adversarial accuracy of eavesdroppers are also studied over Rayleigh fading, Nakagami-m, and AWGN channels to verify the generalization of the proposed scheme. Our experiments show that employing the proposed secure neural encoding can decrease the adversarial accuracy by 28%.

*Index Terms*—Secure Deep-JSCC, data-driven security, secrecy-utility trade-off, secure image transmission.

## I. Introduction

Driven by the growing interest in semantic communication systems [?], intelligent transmission of multimedia content has received much attention because of its various applications in augmented/virtual reality (AR/VR), Metaverse [?], and surveillance systems [?], [?]. The adoption and success of such services rely highly on the security of the delivered contents—communication systems should understand the desired "level of security" and intelligently adapt the transmission scheme accordingly [?], [?].

Connected intelligence is foreseen as the most significant driving force in the sixth generation (6G) of wireless communications. To this end, artificial intelligence and machine learning (AI/ML) algorithms are envisioned to be widely incorporated into 6G networks, realizing an "AI-native" air interface. Nevertheless, security issues at the *wireless edge* of 6G networks are still identified as open challenges [?]. The air interface of 6G systems encounters ever-rising attacks, such as eavesdropping, spoofing [?], and man-in-the-middle [?].

Recently, a considerable number of research has been dedicated to the utilization of deep learning (DL) techniques to optimize the performance of wireless systems, thanks to their outstanding performance and generalization capabilities [?], [?]. In the context of wireless security, autoencoders (composed of linear layers) are exploited to trade-off between data rate and security in additive white Gaussian noise (AWGN) wiretap channels [?]. To tackle the trade-off between the data rate and information leakage, a weighted sum of block error rate and information leakage is used as the loss function (LF) for the wiretap code design. The data fed into the autoencoder is combined with additional random bits to confuse the eavesdropper, while this also reduces the communication rate. Notably, most of the previous works, e.g., [?], [?], [?], focus on learning-aided secure channel coding, rather than taking into account the end-to-end (E2E) performance of secure communications. The content of the transmitted data is not addressed in these works and the entire bit-stream is equally treated as the secret information to be protected against an eavesdropper.

The E2E communication of images from a source node to a legitimate destination can be considered as a joint source channel coding (JSCC) problem. DL-aided JSCC design, a.k.a Deep-JSCC, has received significant attention thanks to its superior performance, particularly its lack of reliance on accurate channel state information [?]. However, [illegible] eavesdroppers, despite providing robustness against channel noise. Inspired by [?] and [?], we provide a generalization of the Deep-JSCC approach to secure communications against multiple eavesdroppers. [?] proposes a generative adversarial network (GAN)-inspired secure variational encoder-decoder (VAE)-based approach for AWGN wiretap channel against one eavesdropper.

In this paper, we consider E2E learning-based secure

*Equal contribution

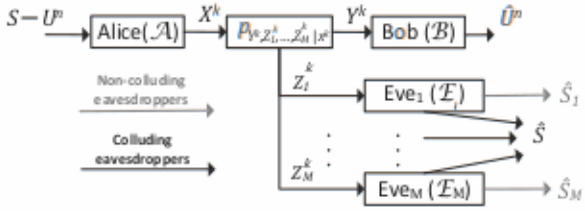


Fig. 1: Proposed system model for secure transmission against multiple eavesdroppers.

communication against multiple eavesdroppers for both colluding and non-colluding eavesdroppers, over AWGN as well as fading channels. For the scenario of colluding eavesdroppers, the adversaries share their logits to collaboratively infer private attributes based on the ensemble learning approach, while for the non-colluding setup they act alone. Please see Fig. ?? for an illustration of the communication scenario studied in this paper. Applications of our proposed framework include (but are not limited to) digital healthcare services, in which medical images are sent to or received from access points, while some private attributes, e.g., the identity of patients should be kept secret from potential eavesdroppers. Notably, previous works [?] and [?] only considered a single eavesdropper with a single-antenna transmitter and multiple parallel channels, respectively. In addition, both [?] and [?] are limited to static channels. Moreover, different from [?], no additional redundant bits are required to be added to the source image. In this paper, Rayleigh fading channel model is used to represent time-varying channel realizations during training, while inference is performed over Nakagami-$m$ and AWGN channels, in addition to Rayleigh fading. Note that the encoder and the decoder do not require any knowledge of the instantaneous channel gains in the proposed scheme.[1]

## II. System Model and Problem Statement

Consider the communication scenario depicted in Fig. ??, where a multi-antenna source node, Alice ($\mathcal{A}$) with $n_T$ antennas, aims to deliver an image $U^n \in \mathcal{U}^n$ to a destination node, Bob ($\mathcal{B}$), over $k$ uses of the communication channel, where $\mathcal{U}$ denotes the alphabet of source images. According to the JSCC literature [?], we refer to the image dimension, $n$, as the source bandwidth. The channel dimension $k$ characterizes the channel bandwidth, where we usually have $k < n$ to reflect the concept of bandwidth compression [?]. Image delivery should be kept secret from multiple eavesdroppers, denoted by $\text{Eve}_1(\mathcal{E}_1), \cdots, \text{Eve}_M(\mathcal{E}_M)$, which overhear the communication through their own channels, and want to infer a private (sensitive) attribute, e.g., diagnostic information regarding the source image, denoted by $S \in \mathcal{S}$ with a discrete alphabet $\mathcal{S}$. The eavesdroppers in the non-colluding setup act alone to infer the secret $S$, while for the colluding setup, "knowledge sharing" is also performed as they share their logits based on the concept of ensemble learning [?].

Alice maps the source information $U^n$ into a channel input codeword $X^k \in \mathbb{C}^{n_T \times k}$ via an encoding 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, $\frac{1}{k}\mathbb{E}[(X^k)^\dagger X^k] \leq P$, which will be satisfied for the realizations of the channel codeword in our data-driven approach. Channel outputs at $\mathcal{B}$ and $\mathcal{E}_m$ are denoted, respectively, by $Y^k$ and $Z^k_m \in \mathbb{C}^k$, with $m \in \{1, \ldots, M\}$. Transmission of data-streams over the air experiences independent realizations of the conditional channel distribution $p_{Y,Z_1,\ldots,Z_M|X}$. We consider both the AWGN and slow fading channels, where for the slow fading, we adopt two widely-used models of Rayleigh fading and Nakagami-$m$ channels, and assume the channel realization to remain constant for the duration of the transmission of a single image, i.e., for $k$ channel uses. Bob then applies a decoding function $f_{\mathcal{B}}$ to obtain $\hat{U}^n = f_{\mathcal{B}}(Y^k)$. Meanwhile, each Eve tries to extract the sensitive attribute $S$, from her observations $Z^k_m$, or by collaborating with other Eves, i.e., sharing their logits. 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 $I(S;Z^k_m)$. Inspired by the key idea of privacy funnel [?], we first formulate an optimization framework to characterize this trade-off, which we call secrecy funnel. This funnel-like framework is then solved in a data-driven manner by implementing DNNs. Mathematically, we aim to simultaneously minimize both the distortion $d(U^n, \hat{U}^n)$ at Bob and the information leakage, $I(S;Z^k_m)$, $m \in [M]$ to adversaries. This corresponds to the following problem

$$\underset{f_{\mathcal{A}}, f_{\mathcal{B}}}{\text{minimize}} \quad \mathbb{E}\left[d(U^n, \hat{U}^n)\right] + \frac{1}{M}\sum_{m\in[M]} w_m I(S;Z^k_m), \quad (1)$$

where $w_m$ shapes our secrecy funnel and adjusts the trade-off between the information leakage and the distortion. Numerical solution of optimization (??) is intractable due to the mutual information term. Hence, we apply a variational approximation of mutual information, which was proposed in [?]. Then, our optimization can be rewritten as:

$$\underset{f_{\Omega_{\mathcal{A}}}, f_{\Omega_{\mathcal{B}}}}{\text{minimize}} \quad \mathbb{E}\left[d(U^n, \hat{U}^n)\right] + \frac{1}{M}\sum_{m\in[M]} w_m \max_{q_{S|Z^k_m}} \mathbb{E}\left[\log q_{S|Z^k_m}(s|\boldsymbol{z}_m)\right], \quad (2)$$

where $q_{S|Z^k_m}(s|\boldsymbol{z}_m)$ characterizes the $\text{Eve}_m$'s posterior estimation corresponding to the correct distribution of $S$, given the observation $Z^k_m = \boldsymbol{z}_m$. To realize a data-driven approach, $f_{\mathcal{A}}$ and $f_{\mathcal{B}}$ are parameterized by DNNs with

[1]Notations: We denote the transpose, the conjugate transpose, and $\ell^2$-norm of a vector by $(\cdot)^T$, $(\cdot)^\dagger$, and $\|\cdot\|$, respectively. The expected value and the probability density function (pdf) of random variable (RV) $X$ are denoted by $\mathbb{E}[X]$ and $p_X(x)$, respectively. Realization vectors of RVs are represented by bold lowercase letters. The mutual information of RVs $X$ and $Y$ and the cross-entropy of two distributions $p$ and $q$ are shown, respectively, by $I(X;Y)$ and $H(p,q)$.

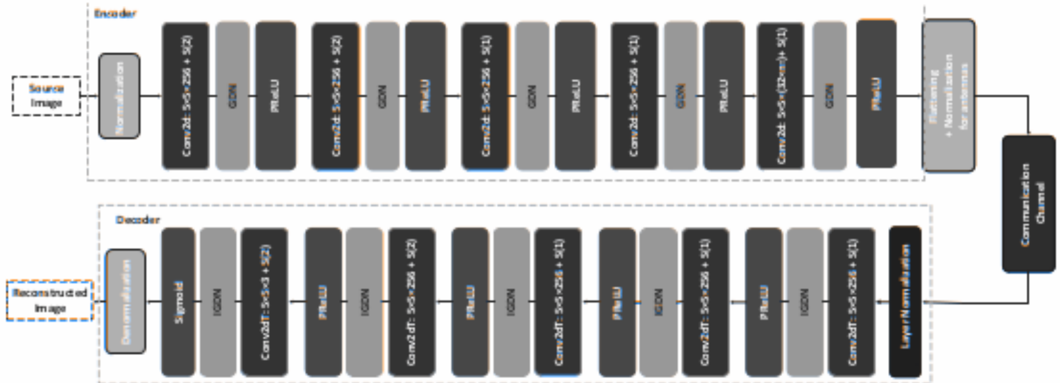


Fig. 2: Proposed DNNs at Alice (encoder) and Bob (decoder). The notation $f \times a \times a$ denotes a convolutional layer with $f$ filters of spatial extent $a$. Moreover, $S(\cdot)$ denotes the stride, which can be downsampling (at the encoder) or upsampling (at the decoder) [?]. At the output of the last PReLU layer, which consists of $2k$ layer elements, we employ a flattening layer for each of the $n_T$ antennas, to reshape the encoded tensor to a data-stream. The encoded latent sequence is further normalized, such that the channel input satisfies the average transmit power constraint.

parameters $\Omega_{\mathcal{A}}$ and $\Omega_{\mathcal{B}}$, respectively. Details of the DNN architectures and the training strategies are given in the next section. The approximation in (??) can be interpreted as the sample-wise negative cross-entropy (CE) between the distribution over adversaries' predictions and the true distribution of sensitive attributes. To solve (??) in a data-driven manner, we assume that the Eves also employ adversarially-trained DNNs and try to infer the sensitive attributes of the transmitted images, where $\Theta_{E,m}$ parameterizes the adversarial network of Eve$_m$. Accordingly, we can formulate the following LF.

Fig. 3: Implemented DNN at adversaries for extracting sensitive information.

$$\mathcal{L}(\Omega_{\mathcal{A}},\Omega_{\mathcal{B}},\boldsymbol{\Theta}_E) = \mathbb{E}\left[d(U^n, f_{\Omega_{\mathcal{B}}}(Y^k))\right] + \frac{1}{M}\sum_{i\in[M]} w_m \max_{\Theta_{E,m}} \mathbb{E}\left[\log q_{\Theta_{E,m}}(s|\boldsymbol{z}_m)\right], \quad (3)$$

where $\boldsymbol{\Theta}_E \triangleq (\Theta_{E,1},\cdots,\Theta_{E,M})$, and $q_{\Theta_{E,m}}(s|\boldsymbol{z}_m)$ formulates the approximated adversarial likelihood regarding the correct value $S = s$, estimated by the DNN of the $m$'th adversary. Details of the proposed DNN architectures and the training strategies are given in the following section.

## III. DNN Architectures

According to the DeepJSCC concept [?], [?], we employ autoencoder DNNs to directly map the image pixels to channel input symbols. In this regard, $\mathcal{A}$ 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 $X^k$. The block diagram of the DNNs employed for the encoder and decoder components of legitimate parties is illustrated in Fig. ??. In addition, the DNN structure employed by each of the adversaries is demonstrated in Fig. ??.

According to the proposed system model, adversaries utilize DNNs to facilitate the inference of sensitive information $S$ from their received signals. $S$ can be the class labels of the images [?], [?]. For example, the identity of patients within medical imaging in e-health applications. To infer sensitive attributes from images, each adversary employs the DNN architecture illustrated in Fig. ??, where the dimension of the output neurons, $L$, equals the cardinality of the secrets $|S|$.[2] The output of the softmax layer produces an adversarial likelihood estimation regarding the posterior distribution $q_{\Theta_{E,m}}(s_m|\boldsymbol{z}_m)$. Invoking (??), one can say that each Eve tries to minimize its CE between the adversarially-estimated posterior distribution $q_{\Theta_{E,m}}(s_m|\boldsymbol{z}_m)$ and the ground-truth, which is represented by the one-hot encoded vector of $S$, denoted by $\boldsymbol{\varepsilon}_s \in \{0,1\}^L$. Notably, having lower CE values results in higher similarity between the adversarial posterior distribution and the ground-truth, which increases the information leakage in terms of CE. Meanwhile, $\mathcal{A}$ and $\mathcal{B}$ try to jointly minimize the reconstruction distortion and the information leakage measured by the negative CE metric. Hence, the sample-wise communication framework can be reformulated as:

$$\underset{}{\text{minimize}}\ \mathcal{L}(\Omega_{\mathcal{A}},\Omega_{\mathcal{B}},\boldsymbol{\Theta}_E) = \mathbb{E}_{p(\boldsymbol{u},\hat{\boldsymbol{u}})}\left[d(\boldsymbol{u},\hat{\boldsymbol{u}})\right] + \frac{1}{M}\sum_{m\in[M]} w_m \max_{\Theta_{E,m}} \left(-H(q_{\Theta_{E,m}}(s|\boldsymbol{z}_m),\boldsymbol{\varepsilon}_s)\right), \quad (4)$$

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, taking into account the randomness in the input image and the channel. Since the true distribution $p(\boldsymbol{u})$ is often unknown, we estimate the expected distortion measure using samples $\boldsymbol{u}_i$ from an available dataset $\mathcal{D}_u$ by computing $\mathbb{E}_{p(\boldsymbol{u},\hat{\boldsymbol{u}})}[d(\boldsymbol{u}, f_{\Omega_{\mathcal{B}}}(\boldsymbol{y}))] \approx \frac{1}{N_u}\sum_{i\in[N_u]} d(\boldsymbol{u},\hat{\boldsymbol{u}})$, where $N_u = |\mathcal{D}_u|$. It is assumed that we know the sensitive attribute in which the eavesdroppers are interested, as well as their channel models. Both of these assumptions are common in the privacy [?], [?] and wiretap channel [?], [?] literature. We do not need to know the instantaneous channel gains, but use their distributions to sample channel realizations during training.

*Training Procedure:* In order to train our system based on (??), we follow an iterative procedure. Intuitively, the network is faced with a minimax game, i.e., the competition between legitimate autoencoder and the adversarial DNNs.

[2] It would be interesting to study the performance when the eavesdroppers employ more complicated architectures, such as VGG and ResNet variations, to infer the secret. This will be addressed in our future works.

ability distribution of the original and the reconstructed image, taking into account the functions of Alice and Bob and the channel. Since the true distribution $p(\boldsymbol{u})$ is often unknown, we estimate the expected distortion measure using samples $\boldsymbol{u}_j$ from an available dataset $\mathcal{D}_u$ by computing $\mathbb{E}_{p(\boldsymbol{u},\hat{\boldsymbol{u}})}[d(\boldsymbol{u}, f_{\Omega_B}(\boldsymbol{y}))] \approx \frac{1}{N_u}\sum_{\boldsymbol{u}\in\mathcal{D}_u} d(\boldsymbol{u},\hat{\boldsymbol{u}})$, where $N_u = |\mathcal{D}_u|$. It is assumed that we know the sensitive attribute in which the eavesdroppers are interested, as well as their channel models. Both of these assumptions are common in the privacy [?], [?] and wiretap channel [?], [?], [?] literature. We do not need to know the instantaneous channel gains, but use their distributions to sample channel realizations during training.

**Training Procedure:** In order to train the system based on (??), we follow an iterative procedure. Intuitively, the network nodes are faced with a minimax game, i.e., the competition between legitimate autoencoder and the adversarial DNNs. Hence, the following strategy is run through our proposed E2E system. The encoder and decoder function of Alice and Bob should jointly minimize their LF, denoted by $\mathcal{L}_{AB}$:

$$\mathcal{L}_{AB} = \frac{1}{N_u}\sum_{\boldsymbol{u}\in\mathcal{D}_u}\left(d(\boldsymbol{u},\hat{\boldsymbol{u}}) - \sum_{m\in[M]} w_m H\big(q_{\Theta_{E,m}}(s(\boldsymbol{u})|\boldsymbol{z}_m(\boldsymbol{u})), \varepsilon_s(\boldsymbol{u})\big)\right) \quad (5)$$

The distortion measure we consider for our legitimate loss function $\mathcal{L}_{AB}$ is a mixture of the average mean squared error (MSE), denoted by $\Delta^{\mathsf{MSE}}$, and the structural similarity index (SSIM), $\Delta^{\mathsf{SSIM}}$, between the input image $\boldsymbol{u}$ and the recovered version $\hat{\boldsymbol{u}}$ at the output of Bob's DNN. Therefore, we assume $d(\cdot,\cdot)$ to be measured as follows

$$d(\boldsymbol{u},\hat{\boldsymbol{u}}) = \Delta^{\mathsf{MSE}}(\boldsymbol{u},\hat{\boldsymbol{u}}) + \alpha\Delta^{\mathsf{SSIM}}(\boldsymbol{u},\hat{\boldsymbol{u}}), \quad (7)$$

where $\Delta^{\mathsf{MSE}}(\boldsymbol{u},\hat{\boldsymbol{u}}) = \frac{1}{n}\|\boldsymbol{u}-\hat{\boldsymbol{u}}\|^2$, $\Delta^{\mathsf{SSIM}}(\boldsymbol{u},\hat{\boldsymbol{u}}) = 1 - \mathsf{SSIM}(\boldsymbol{u},\hat{\boldsymbol{u}})$, and $\alpha$ is a tuning parameter representing the contribution of the SSIM metric. The SSIM measure between two images $I$ and $K$ is defined as

$$\mathsf{SSIM}(I,K) \triangleq \frac{(2\mu_I\mu_K + c_1)(2\sigma_{IK} + c_2)}{(\mu_I^2+\mu_K^2+c_1)(\sigma_I^2+\sigma_K^2+c_2)}, \quad (8)$$

where $\mu_I$, $\mu_K$, $\sigma_I$, $\sigma_K$, and $\sigma_{IK}$ are the local means, standard deviations, and cross-covariance for images $I$ and $K$, while $c_1$ and $c_2$ are two adjustable constants [?]. The rationale behind the proposed distortion metric is that we not only aim to recover every pixel of images with minimum error (captured via the MSE measure), but also want to obtain a good-quality reconstruction from the *human perception* point-of-view.

Each step of the joint training of $\mathcal{A}$-to-$\mathcal{B}$ autoencoder is followed by a training step for the adversarial DNNs. Eavesdroppers aim to minimize the CE between their estimated likelihood $q_{\Theta_{E,m}}(s|\boldsymbol{z}_m)$ and the ground-truth vector $\varepsilon_s$ corresponding to $S$. Hence, the following LF is employed for training the DNN of Eve$_m$ for $m \in [M]$:

$$\mathcal{L}_{E_m} = \frac{1}{N_u}\sum_{\boldsymbol{u}\in\mathcal{D}_u} H\big(q_{\Theta_{E,m}}(s(\boldsymbol{u})|\boldsymbol{z}_m(\boldsymbol{u})), \varepsilon_s(\boldsymbol{u})\big) \quad (9)$$

Note that the adversarial DNNs can be trained in parallel. Then for the case of colluding eavesdroppers, an additional step of "knowledge sharing" is performed. In this case, the adversaries share their individually-extracted logits, and a weighted sum of these logits is exploited for the inference of private attributes, where the logit weights are trained in the colluding framework.

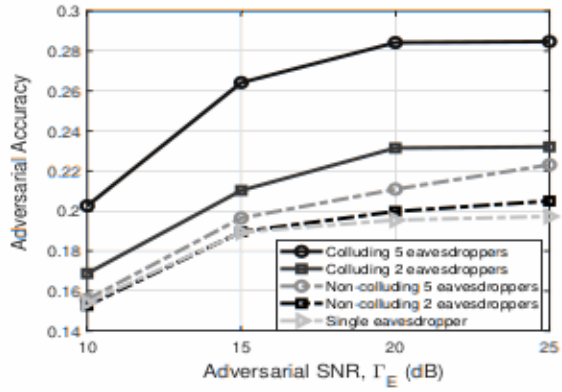


Fig. 4: Total adversarial accuracy over Rayleigh channels.

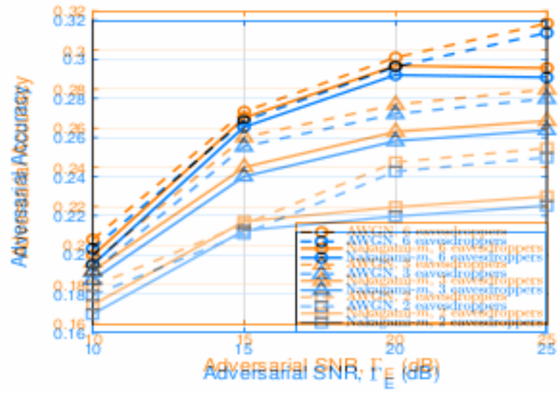


Fig. 5: Colluding adversarial accuracy over AWGN and Nakagami.

## IV. Evaluations

In this section, we evaluate the performance of the proposed scheme over both AWGN and complex fading (Rayleigh and Nakagami-$m$) communication channels. We address the generalization capability of our proposed scheme for different communication scenarios and over a wide range of signal-to-noise ratio (SNR) values, to highlight the *data efficiency* of our proposed learning-based security solution. We also address the *secrecy-utility trade-off* for the proposed learning-based approach. We evaluate our proposed secure framework using images with dimension $32 \times 32 \times 3$ (height, width, channels) from CIFAR-10 dataset [?]. The dataset consists of 60000 colored images of size $32 \times 32$ 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 $S$, either individually (the non-colluding case), or by learning from the combination of the adversarial logits (the colluding setup). The common secret $S$ here is considered as the class of CIFAR-10 images with $|S| = L = 10$. For simplicity, we consider a single weight in the LF, that is, $w_m = w = 5, \forall m$. We also set $\alpha = 0.1$ and $N_T = 4$ antennas. These parameters are set after conducting extensive experiments and training the DNNs with a wide range of values for $w_m$ and $\alpha$, where we have omitted the results of fine-tuning step due to space limitations. Transmit SNRs of communication links are defined as $\Gamma_B = 10\log_{10}\frac{P}{\sigma_L^2}$ dB and $\Gamma_E = 10\log_{10}\frac{P}{\sigma_E^2}$ dB, representing the ratio of the average power of the channel input to the average noise power of legitimate $\sigma_L^2$ and adversarial nodes $\sigma_E^2$, respectively. During training, we set $\Gamma_B = 20$dB and $\Gamma_E = 15$dB, respectively, while we test the performance over different values of channel SNRs during the inference. In addition, the bandwidth compression ratio is set to $\frac{k}{n} = \frac{1}{3}$. For the training, we sample channel realizations from the general case of complex Rayleigh fading model with average $\Gamma_B$ and $\Gamma_E$ values stated above. Nevertheless, during the inference phase we study the performance in different scenarios of AWGN and Nakagami-$m$ channels (for $m = 3$). While we do assume known channel models in our simulations, which we use 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. DNN architectures are implemented using Python3 with Tensorflow.[3] The codes were run on Intel(R) Xeon(R) Silver 4114 CPU running at 2.20GHz with GeForce RTX 2080 Ti GPU. To minimize the LFs, the widely-adopted Adam optimizer is chosen [?] with a learning rate of $10^{-4}$. We fix the number of training episodes to $N_{\text{episode}} = 200$, and the batch size to $m = 128$.

Figs. ?? and ?? show the adversarial accuracy of the proposed scheme vs. the SNR of adversarial links ($\Gamma_E = \Gamma_B = 5$ dB) for both scenarios of colluding and non-colluding eavesdroppers. For the colluding setup, the total adversarial accuracy refers to the overall accuracy of adversaries in correctly finding the ground-truth $s_i$ from their aggregated logits, while for the non-colluding benchmarks, the mean accuracy across eavesdroppers is plotted for the sake of comparison. One can observe that increasing the number of eavesdroppers leads to 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 sharing" among eavesdroppers through the ensemble learning process [?]. This actually helps them learn the secret more accurately. The figures also indicate that by increasing the quality of adversarial links, i.e., increasing $\Gamma_E$, the accuracy of adversaries increases by at most 10%. This is because higher SNR values result in having less distorted (less noisy) observations at the eavesdroppers, resulting in more accurate estimations about the posterior adversarial distribution $q_{\theta_{E,m}}(s|z_m)$. The amount of increase in the adversarial accuracy reduces with the increase in $\Gamma_E$, which highlights the limitation of eavesdroppers in the proposed secure scheme. Fig. ?? also highlights the generalization capability of our learning-based framework extended to the AWGN and Nakagami-$m$ (for $m = 3$) channel scenario. It shows that we can achieve almost similar trends in channel scenarios other than Rayleigh fading, despite training the networks with a Rayleigh channel model.

[3] https://www.tensorflow.org/

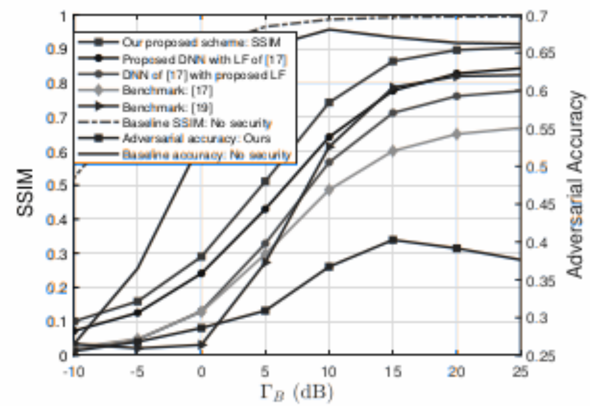


Fig. 4: Ablation study for $M = 3$ non-colluding eavesdroppers.

Fig. ?? illustrates the data reconstruction performance at Bob and the total adversarial accuracy, when having $M = 3$ non-colluding eavesdroppers. One can infer from the figure that our proposed system outperforms the benchmarks in terms of the reconstruction performance. Accordingly, 20% and 10% performance gain is achieved by our proposed scheme compared with [?] and [?], respectively. The ablation examinations conducted in this figure show that both the implemented DNNs and the proposed LFs for optimizing the framework contribute to the system's performance compared with other benchmarks. The figure also implies that increasing $\Gamma_B$ results in having higher SSIM values. This is because increasing $\Gamma_B$ can result in less disturbed observations at Bob, which facilitates the image reconstruction performance. *Data efficiency and generalizability* of our proposed scheme are also validated, since we have trained our DNNs with a fixed SNR $\Gamma_B = 20$ dB, while the performance gain of our approach during inference holds for varying SNRs. Further, Fig. ?? highlights that if we ignore the eavesdroppers during the training of $A$-$B$ pair, and set the $\alpha_m = 0, \forall m \in [M]$, our proposed scheme can achieve almost perfect (SSIM ≈ 1) data recovery. The impact of the eavesdroppers on Bob's performance can be studied in this figure as well, where having 3 eavesdroppers can impose 10% decrement in the reconstruction performance of Bob. Finally, to indicate the importance of our proposed adversary-aware scheme in terms of preventing leakage, we can observe from the figure that if we do not employ secure neural encoding (i.e., ignoring eavesdroppers during the training of $A$-$B$ pair), the adversarial accuracy is increased by about 28%. This clearly highlights the importance of employing our proposed learning-based secure encoding scheme.

Fig. ?? studies the impact of [illegible] parameters [illegible] hyper-parameters [illegible] coefficients [illegible] LFs [illegible] CIFAR-10 dataset. The figure indicates that [illegible] higher [illegible] SSIM [illegible]

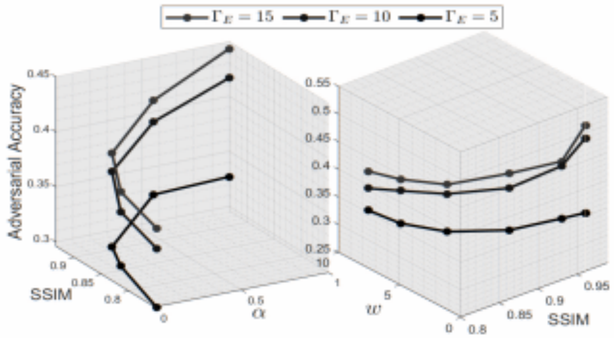


Fig. 7: Secrecy-utility trade-off.

in having higher SSIM values. This is because increasing $\Gamma_B$ can result in less distorted observations at Bob, which facilitates the image reconstruction performance. Data efficiency and generalizability of our proposed scheme are also validated, since we have trained our DNNs with a fixed SNR $\Gamma_B = 20$ dB, while the performance gain of our approach during inference holds for various SNRs. Furthermore, Fig. ?? highlights that, if we ignore the eavesdroppers during the training of $\mathcal{A}$-$\mathcal{B}$ pair, and set $w_m = 0, \forall m \in [M]$, our proposed scheme can achieve almost perfect (SSIM = 1) data recovery. The impact of the eavesdroppers on Bob's performance can be studied in this figure as well, where having 3 eavesdroppers can impose 10% decrease in the reconstruction performance of Bob. Finally, to indicate the importance of our proposed adversary-aware scheme in terms of preventing leakage, we can observe from the figure that if we do not employ secure neural encoding (i.e., ignoring the eavesdroppers during the training of $\mathcal{A}$-$\mathcal{B}$ pair), the adversarial accuracy is increased by about 28%. This clearly highlights the importance of employing our proposed learning-based secure encoding scheme.

Fig. ?? studies the impact of tuning parameters $\alpha$ and $w$ on the adversarial performance of our proposed system. These hyper-parameters are the coefficients associated with utility and secrecy adjustment terms within our training LFs in (??) and (??), respectively. For this experiment, the adversarial performance is captured by investigating the accuracy of adversaries in correctly finding the ground-truth label $\varepsilon_s$ (representing the sensitive information $S$) among the labels of CIFAR-10 dataset. The figure indicates that by increasing $\alpha$, higher values of SSIM can be achieved, since more emphasis on the SSIM error $\Delta^{\mathsf{SSIM}}$ is put based on (??). However, the accuracy of adversaries in extracting the sensitive information also increases, which verifies the secrecy-utility trade-off. Notably, by increasing $\alpha$ to values more than 0.1, a jump in the adversarial accuracy can be observed, which leads us choosing $\alpha = 0.1$ for our network. Similarly, by increasing $w$, the emphasis goes toward the secrecy criteria introduced in (??), (??), and (??), which leads to the reduction in adversarial accuracy and achieved SSIM, verifying the secrecy-utility trade-off as well. Fig. ?? also shows that increasing $\Gamma_E$ can improve the adversarial accuracy in finding the sensitive data $\varepsilon_s$. Notably, the amount of increase in the adversarial accuracy reduces with the increase in $\Gamma_E$ which highlights the limitation of adversarial nodes based on our proposed secure scheme.

## V. Conclusions and Future Directions

We proposed a learning-based approach for E2E secure image delivery against multiple eavesdroppers over AWGN and complex-valued fading channels. We considered both scenarios of colluding and non-colluding eavesdroppers over CIFAR-10 dataset. For the colluding strategy, eavesdroppers collaborate to infer private data from their observations (channel outputs), using ensemble learning, while for the non-colluding setup they act alone. Meanwhile, the legitimate parties aim to have a secure communication with minimum average distortion. Employing autoencoders, we proposed a secrecy funnel framework to achieve both secrecy and utility, where we also take into account the perceptual quality of image transmission within our LF. Evaluations validate the performance of our proposed scheme compared with existing benchmarks, while addressing the secrecy-utility trade-off. Our proposed system is also shown to be generalizable to a wide range of SNRs and different communication scenarios.

*Future Directions:* A problem to be studied is training the systems over real-time wireless channels. The challenge here is the relatively long coherence time compared to the rate at which data samples can be processed for training. Accordingly, only a few channel realizations are observed over every minibatch, which will be an important issue for training communication systems that are supposed to generalize well to a wide range of channel realizations.

## References


[1] D. Gündüz, Z. Qin, I. Estella Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Beyond transmitting bits: Context, semantics, and task-oriented communications," arXiv:2207.09353, 2022.

[2] M. Letafati and S. Otoum, "On the privacy and security for e-health services in the metaverse: An overview," Ad Hoc Networks, Aug. 2023, https://doi.org/10.1016/j.adhoc.2023.103262.

[3] Q. Yang and D. Gündüz, "Coded caching and content delivery with heterogeneous distortion requirements," IEEE Trans. Inf. Theory, vol. 64, no. 6, pp. 4347–4364, Jun. 2018.

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

[5] M. Letafati, H. Behroozi, B. H. Khalaj, and E. A. Jorswieck, "Content-based medical image transmission against randomly-distributed passive eavesdroppers," IEEE Int. Conf. Comm. Workshop (ICCW), Montreal, Canada, Jun. 2021, pp. 1–7.

[6] ——, "On learning-assisted content-based secure image transmission for delay-aware systems with randomly-distributed eavesdroppers," IEEE Trans. Comm., vol. 70, no. 2, pp. 1125–1139, Feb. 2022.

[7] A. Chorti, A. N. Barreto, S. Köpsell, M. Zoli, M. Chafii, P. Sehier, G. Fettweis, and H. V. Poor, "Context-aware security for 6G wireless: The role of physical layer security," 2021, arXiv:2101.01536.

[8] 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 Comm. Lett., [illegible] pp. 856–860, 2022.

[9] 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," [illegible] (GLOBECOM), Madrid, Spain, Dec. 2021, pp. 1–6.

[10] E. Bourtsoulatze, D. Burth 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.

[11] D. B. Kurka and D. Gündüz, "Bandwidth-agile image transmission with deep joint source-channel coding," IEEE Trans. Wireless Comm., vol. 20, no. 12, pp. 8081–8095, Dec. 2021.

[12] M. Letafati, H. Behroozi, B. H. Khalaj, and E. A. Jorswieck, "Wireless-powered cooperative key generation for e-health: A reservoir learning approach," 2022 IEEE 95th Vehicular Technology Conference (VTC-Spring), Helsinki, Finland, Jun. 2022, pp. 1–7.

[13] [illegible] "Wiretap code design by neural network autoencoders," IEEE Trans. Inf. Forensics Secur., vol. 15, pp. 3374–3386, Oct. 2019.

[14] R. Fritschek, R. F. Schaefer, and G. Wunder, "Deep learning for channel coding via neural mutual information estimation," in Proc. 20th Int. Workshop Sig. Proc. Adv. Wireless Comm., pp. 1–5, July 2019.

[15] R. Fritschek, R. F. Schaefer, and G. Wunder, "Deep learning based wiretap coding via mutual information estimation," in Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning (WiseML '20), NY, USA, Jul. 2020, pp. 74–79.

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

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

[18] D. B. Kurka and D. Gündüz, "DeepJSCC-f: Deep joint source-channel coding of images with feedback," [illegible] vol. 1, no. 1, pp. 178–193, May 2020.

[19] R. Polikar, "Bootstrap-inspired techniques in computational intelligence," IEEE Signal Process. Mag., vol. 24, no. 4, pp. 59–72, 2007.

[20] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Médard, "From the information bottleneck to the privacy funnel," in [illegible] Workshop (ITW), [illegible] 2014.

[21] D. Barber and F. Agakov, "The IM algorithm: A variational approach to information maximization," [illegible] pp. 201–208.

[22] 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.

[23] A. Krizhevsky, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., 2009.

[8] 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 Comm. Lett., vol. 11, no. 4, pp. 856–860, Apr. 2022.

[9] 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," 2021 IEEE Global Communications Conference (GLOBECOM), Madrid, Spain, Dec. 2021, pp. 1–6.

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

[11] D. B. Kurka and D. Gündüz, "Bandwidth-agile image transmission with deep joint source-channel coding," IEEE Trans. Wireless Comm., vol. 20, no. 12, pp. 8081–8095, Dec. 2021.

[12] M. Letafati, H. Behroozi, B. H. Khalaj, and E. A. Jorswieck, "Wireless-powered cooperative key generation for e-health: A reservoir learning approach," 2022 IEEE 95th Vehicular Technology Conference (VTC-Spring), Helsinki, Finland, Jun. 2022, pp. 1–7.

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

[14] R. Fritschek, R. F. Schaefer, and G. Wunder, "Deep learning for channel coding via neural mutual information estimation," in Proc. 20th Int. Workshop Sig. Proc. Adv. Wireless Comm., pp. 1–5, July 2019.

[15] R. Fritschek, R. F. Schaefer, and G. Wunder, "Deep learning based wiretap coding via mutual information estimation," in Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning (WiseML'20), NY, USA, Jul. 2020, pp. 74–79.

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

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

[18] D. B. Kurka and D. Gündüz, "DeepJSCC-f: Deep joint source-channel coding of images with feedback," IEEE Journal on Selected Areas in Information Theory, vol. 1, no. 1, pp. 178–193, May 2020.

[19] R. Polikar, "Bootstrap-inspired techniques in computational intelligence," IEEE Signal Process. Mag., vol. 24, no. 4, pp. 59–72, 2007.

[20] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Médard, "From the information bottleneck to the privacy funnel," in Proc. of IEEE Information Theory Workshop (ITW), 2014.

[21] D. Barber and F. Agakov, "The IM algorithm: A variational approach to information maximization," Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference, vol. 16, pp. 201–208.

[22] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," IEEE Transactions on Computational Imaging, vol. 3, pp. 47–57, Mar. 2017.

[23] A. Krizhevsky, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., 2009.

[24] D.P. Kingma and L.J. Ba, "Adam: A method for stochastic optimization," 2015 International Conference on Learning Representations (ICLR), San Diego, May 2015, pp. 1–13.