The Model Inversion Eavesdropping Attack in Semantic Communication Systems

Yuhao Chen, Qianqian Yang[†], Zhiguo Shi, Jiming Chen
College of Control Science and Engineering, Zhejiang University, Hangzhou 310007, China
College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310007, China
The State Key Laboratory of Industrial Control Technology, Hangzhou 310007, China
{eeechenyh, qianqianyang20[†], shizg, cjm}@zju.edu.cn

Abstract—In recent years, semantic communication has been a popular research topic for its superiority in communication efficiency. As semantic communication relies on deep learning to extract meaning from raw messages, it is vulnerable to attacks targeting deep learning models. In this paper, we introduce the model inversion eavesdropping attack (MIEA) to reveal the risk of privacy leaks in the semantic communication system. In MIEA, the attacker first eavesdrops the signal being transmitted by the semantic communication system and then performs model inversion attack to reconstruct the raw message, where both the white-box and black-box settings are considered. Evaluation results show that MIEA can successfully reconstruct the raw message with good quality under different channel conditions. We then propose a defense method based on random permutation and substitution to defend against MIEA in order to achieve secure semantic communication. Our experimental results demonstrate the effectiveness of the proposed defense method in preventing MIEA.

I. Introduction

[illegible]

This paper is organized as follows. Section ?? introduces the basic ideas of semantic communications. In section ??, we present the proposed MIEA under both the white-box and black-box setting, and propose our defense method. In section ??, we evaluate the effectiveness of the proposed MIEA and the proposed defense method. Section ?? concludes our work.

II. Fundamentals

In this section, we provide the fundamentals of semantic communication and the eavesdropping performed by the attacker. We consider a semantic communication system which transmits images over wireless channels. As shown in Fig. ??, the transmitter of the semantic communication system consists of a semantic encoder and a channel encoder. The semantic encoder extracts the semantic features z from the raw image x, while the channel encoder maps z into the transmitted features $y_f \in \mathbb{R}^{h\times w\times c}$, where h, w and c denote the height, the width and the channel of the transmitted features respectively. Before transmission, y_f is reshaped into the transmitted symbols $y \in \mathbb{R}^{N\times 2}$, where $N = \frac{h\times w\times c}{2}$ and the two channels are the real parts and imaginary parts of the signal to be transmitted, respectively. y is then transmitted over a wireless channel, which we denote as the main channel to distinguish from the channel used by the attacker. The received signal $\hat{y}$ at the receiver side can be characterized by

$$\hat{y} = H_m y + n_m, \tag{1}$$

where H_m is a matrix which reflects the main channel effect such as multi-path propagation, fading and interference, while n_m is a zero-mean additive white Gaussian noise. The receiver of the semantic communication system consists of a channel decoder and a semantic decoder. The receiver first reshapes $\hat{y}$ back to the transmitted features $\hat{y}_f$. Then the channel decoder maps $\hat{y}_f$ back to the semantic features $\hat{z}$. The semantic decoder then reconstructs the image $\hat{x}$ from $\hat{z}$. We jointly train the semantic encoder, channel encoder, semantic decoder and channel decoder using the following loss function:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \|x - \hat{x}\|^2 + \lambda T(\hat{x}), \tag{2}$$

where N is the number of the training data batch and

$$T(x) = \sum_{i,j} \left(|\hat{x}_{i+1,j} - \hat{x}_{i,j}|^2 + |\hat{x}_{i,j+1} - \hat{x}_{i,j}|^2\right)^{\beta/2}. \tag{3}$$

The first term in (??) computes the mean square error (MSE) between x and $\hat{x}$. The second term $T(\hat{x})$ is the total variation [?] that measures the smoothness of the reconstructed image $\hat{x}$, where $\hat{x}_{i,j}$ denotes the pixel value at the position (i, j) and β controls the smoothness of the image, with larger β being more piecewise-smooth. The hyper-parameter λ balances the two terms. In our work, we choose $\beta = 1$ and $\lambda = 1$.

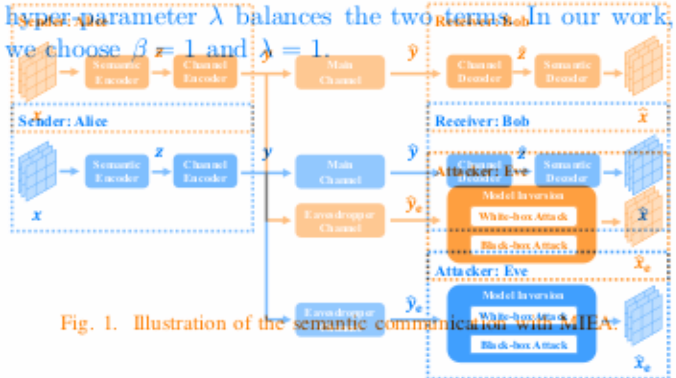


Fig. 1. Illustration of the semantic communication with MIEA.

Next, we introduce how an attacker eavesdrops the transmitted signal under the semantic communication system. We follow the naming convention in the security research, with Alice, Bob and Eve representing the sender, receiver and attacker respectively. Suppose Alice wants to send an image to Bob. As shown in the lower part of Fig. ??, since the transmitted symbols y is transmitted over the wireless channel, it can easily be captured by any unauthorized receiver. Assume that there exists an attacker Eve who intercepts y and attempts to reconstruct the raw image from it. The wireless channel between Alice and Eve is referred to as the eavesdropper channel [?]. The received signal at Eve is given by

$$\hat{y}_e = H_e y + n_e, \tag{4}$$

Similarly, H_e represents the eavesdropper channel matrix and n_e is a zero-mean additive white Gaussian noise. After eavesdropping $\hat{y}_e$, Eve is able to reconstruct the image, denoted as $\hat{x}_e$, which will be detailed section ??. Note that to avoid confusion, we use the received image to denote $\hat{x}$ received by Bob and the eavesdropped image to denote $\hat{x}_e$ eavesdropped by Eve.

III. The Proposed MIEA and its Defense

In this section, we first elaborate the idea of MIEA. To reconstruct $\hat{x}_e$, Eve performs MIA [?] using either the white-box attack or the black-box attack, which depends on the knowledge of the semantic encoder and channel encoder that Eve has. Then we propose an effective defense method that defends against both types of attack.

A. White-box Attack

In the white-box attack, Eve knows the parameters and structure of the semantic encoder and channel encoder. For example, the semantic communication system is publicly available or available through purchase, such as JPEG. In this case, Eve can directly use the semantic encoder and channel encoder to reconstruct the image. We denote the two encoders as a single function $f(\cdot)$ which maps a given image x to the transmitted symbol y, that is, $y = f(x)$.

The reconstructed image $\hat{\boldsymbol{x}}_e$ can be obtained by solving the following optimization problem:

$$\hat{\boldsymbol{x}}_e = \arg\min_{\boldsymbol{x}} \|\hat{\boldsymbol{y}}_e - f(\boldsymbol{x})\|^2 + \lambda T(\boldsymbol{x}). \tag{5}$$

The first term in (??) is the MSE between $\hat{\boldsymbol{y}}_e$ and $f(\boldsymbol{x})$, while the second term $T(\boldsymbol{x})$ is the total variation defined in (??) that guarantees the smoothness of $\boldsymbol{x}$. Similar to (??), we set $\beta = 1$ and $\lambda = 1$ here. The optimization problem (??) can be solved by performing the gradient descent, which iteratively updates the input $\boldsymbol{x}$ so that (??) can be minimized.

B. Black-box Attack

In the black-box attack, Eve lacks knowledge of the parameters and structures of both encoders. In this case, Eve uses an inverse network of the two encoders, denoted as $f^{-1}(\cdot)$, to inverse $\hat{\boldsymbol{y}}_e$ back to $\hat{\boldsymbol{x}}_e$. Specifically, f^{-1} takes $\hat{\boldsymbol{y}}_e$ as input and outputs $\boldsymbol{x}$, i.e., $f^{-1}(\hat{\boldsymbol{y}}_e) = \boldsymbol{x}$. To train f^{-1}, we assume that Eve can feed a batch of samples $\mathbb{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_m\}$ into the encoder and capture the corresponding transmitted symbols $\mathbb{Y}_e = \{\hat{\boldsymbol{y}}_1, \hat{\boldsymbol{y}}_2, ..., \hat{\boldsymbol{y}}_m\}$, where m is the number of the samples. Eve then trains f^{-1} using $\mathbb{Y}_e$ as the input and $\mathbb{X}$ as the ground truth output. We use the l_2 norm as the loss function and employ stochastic gradient descent to train the inverse network:

$$f^{-1} = \arg\min_{g} \frac{1}{m}\sum_{i=1}^{m} \|g(\hat{\boldsymbol{y}}_i) - \boldsymbol{x}_i\|^2, \tag{6}$$

where we use g to represent the inverse network being optimized. Once the inverse network is trained, Eve is able to reconstruct the image from any newly eavesdropped signal.

C. Defense Method against MIA

To defend against MIA in deep learning, researchers have proposed various techniques such as differential privacy (DP) [?] or attacker-aware training [?]. The DP technique involves adding an additive Laplacian noise to the raw image, while the attacker-aware training adds a regularization term to the loss function during training, which maximizes the MSE between the reconstructed image and the raw image. Both methods prevent Eve from reconstructing high-quality images while maintaining the performance of downstream tasks. However, in the transmission task considered in this paper, both Bob and Eve attempt to reconstruct the image from $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{y}}_e$ respectively. Furthermore, we assume that both the main channel and eavesdropper channel are AWGN channel, i.e., $H_m = H_e$. Thus the difference between $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{y}}_e$, which is $n_m - n_e$, is relatively small. Therefore, if Eve fails to reconstruct high-quality images under the defense methods above, Bob will also fail, which contradicts the goal of the transmission task. To prevent Eve from reconstructing images as well as maintaining the image quality received by Bob, an intuitive solution is to encrypt $\boldsymbol{y}_f$ using cryptography algorithms, but this will incur a large computation overhead. To reduce the computation overhead, we propose a defense method based on random permutation and substitution of the transmitted features $\boldsymbol{y}_f$, which can simultaneously defend against both types of attacks.

Random Permutation and Substitution. We first introduce the random permutation operation. For the transmitted features $\boldsymbol{y}_f$, we randomly permute the features along the first dimension h. We define the permutation scheme P as a random permutation of the array $[0, 1, ..., h-1]$, where each element represents the index (0-indexed) of $\boldsymbol{y}_f$. After applying P, we obtain the permuted transmitted features $\boldsymbol{y}_f^P$.

Then we perform the substitution operation on $\boldsymbol{y}_f^P$ by swapping some of the y_i with y'_i from another transmitted features $\boldsymbol{y}'_f$, where i in y_i and y'_i indicates that substitution is performed in the same position of both transmitted features. Note that we remove the subscript f in y_i to avoid complex notations. If Alice sends several images to Bob, $\boldsymbol{y}'_f$ can be the transmitted features of the next image. If Alice only sends one image to Bob, $\boldsymbol{y}'_f$ can be a random-noise tensor, which will also be sent to Bob. Similarly, we define the substitution scheme S as a sub-array of the array $[0, 1, ..., h-1]$, where every index i indicates the y_i that should be substituted. After substitution, $\boldsymbol{y}_f^P$ becomes $\boldsymbol{y}_f^S$.

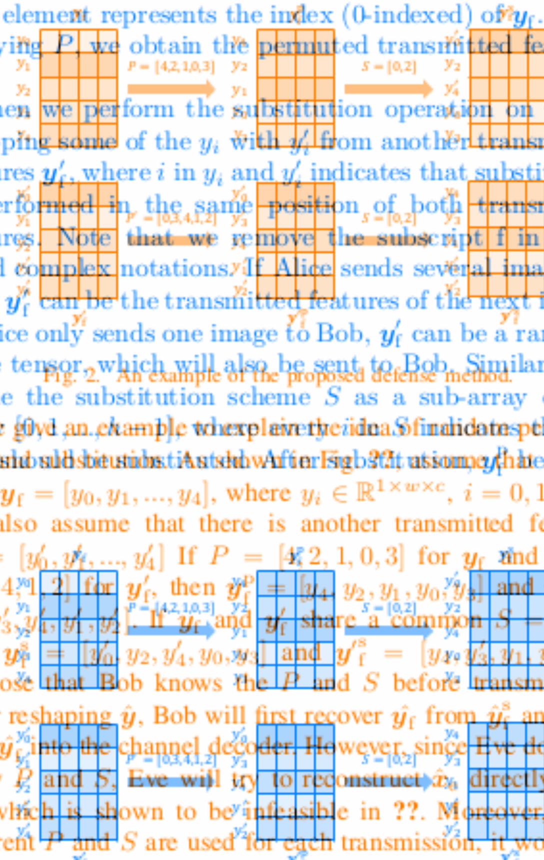


Fig. 2. An example of the proposed defense method.

We give an example to explain the idea. Suppose $\boldsymbol{y}_f = [y_0, y_1, ..., y_4]$, where $y_i \in \mathbb{R}^{1\times w\times c}$, $i = 0, 1, ..., 4$. We also assume that there is another transmitted features $\boldsymbol{y}'_f = [y'_0, y'_1, ..., y'_4]$. If $P = [4, 2, 1, 0, 3]$ for $\boldsymbol{y}_f$ and $P' = [0, 3, 4, 1, 2]$ for $\boldsymbol{y}'_f$, then $\boldsymbol{y}_f^P = [y_4, y_2, y_1, y_0, y_3]$ and $\boldsymbol{y}'^P_f = [y'_0, y'_3, y'_4, y'_1, y'_2]$. If $\boldsymbol{y}_f$ and $\boldsymbol{y}'_f$ share a common $S = [0, 2]$, then $\boldsymbol{y}_f^S = [y'_0, y_2, y'_4, y_0, y_3]$ and $\boldsymbol{y}'^S_f = [y_4, y'_3, y_1, y'_1, y'_2]$. Suppose that Bob knows the P and S before transmission. After reshaping $\hat{\boldsymbol{y}}$, Bob will first recover $\hat{\boldsymbol{y}}_f$ from $\hat{\boldsymbol{y}}_f^S$ and then feed $\hat{\boldsymbol{y}}_f$ into the channel decoder. However, since Eve does not know P and S, Eve will try to reconstruct $\hat{\boldsymbol{x}}_e$ directly from $\hat{\boldsymbol{y}}_f^S$, which is shown to be infeasible in ??. Moreover, since different P and S are used for each transmission, it would be difficult for Eve to determine the correct P and S for each eavesdropped signal.

Scheme Selection. The proposed method is dependent on Bob having knowledge of P and S. ... it would be difficult for Eve to determine the correct P and S for each eavesdropped signal.

Scheme Selection. The proposed method is dependent on Bob having knowledge of P and S before transmission. Hence it is necessary for Alice and Bob to share two common sets of schemes, namely the permutation scheme set $\mathbb{P}$ and the substitution set $\mathbb{S}$, which are kept secret from Eve. Both sets comprise multiple schemes that can be employed for permutation and substitution. Before each image transmission, Alice generates a value pair $V = \{p, s\}$, which is used to select the corresponding P and S from $\mathbb{P}$ and $\mathbb{S}$, respectively. V is first encrypted using a secret key K shared between Alice and Bob. Then the encrypted V is transmitted to Bob, which cannot be modified by the main channel. Hence error-free techniques such as error correction and retransmission are utilized to transmit V. After receiving the $\hat{\boldsymbol{y}}_{\mathrm{s}}$ and the encrypted V, Bob decrypts V using K and determines P and S, from which $\hat{\boldsymbol{y}}$ can be recovered.

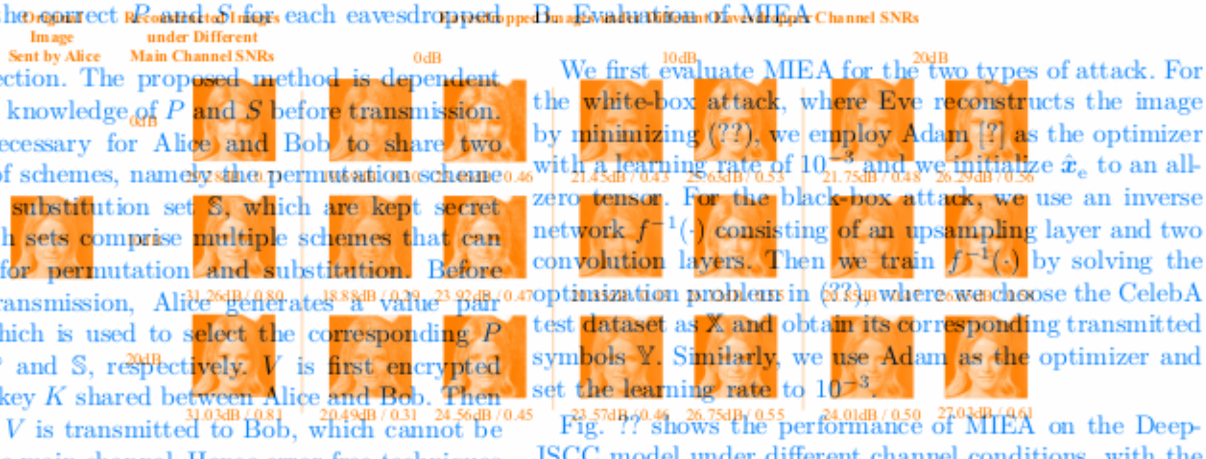


Fig. 3. Visualization of MIEA for the white-box attack and the black-box attack under different channel conditions. For each channel condition, the eavesdropped image by the white-box attack is displayed on the left and the one obtained by the black-box attack is on the right.

IV. Evaluations

In this section, we present our experiments to evaluate MIEA and the proposed defense method. We first evaluate [illegible]

different DeepJSCC models for each SNR value. For each evaluation, we consider the SNR of both channels to be [illegible]

A. Evaluation Setup

[illegible]

B. Evaluation of MIEA

We first evaluate MIEA for the two types of attack. For the white-box attack, where Eve reconstructs the image by minimizing (??), we employ Adam [?] as the optimizer with a learning rate of 10^{-3} and we initialize $\hat{\boldsymbol{x}}_{\mathrm{e}}$ to an all-zero tensor. For the black-box attack, we use an inverse network $f^{-1}(\cdot)$ consisting of an upsampling layer and two convolution layers. Then we train $f^{-1}(\cdot)$ by solving the optimization problem in (??), where we choose the CelebA test dataset as $\mathbb{X}$ and obtain its corresponding transmitted symbols $\mathbb{Y}$. Similarly, we use Adam as the optimizer and set the learning rate to 10^{-3}.

Fig. ?? shows the performance of MIEA on the DeepJSCC model under different channel conditions, with the SSIM and PSNR given below each image. The first two columns in Fig. ?? are baselines for comparisons with the images eavesdropped by MIEA, where the first column [illegible] be observed that the SSIM and PSNR values in the black-box attack are generally larger than those in the white-box attack. This is because the black-box attack requires training $f^{-1}(\cdot)$ before reconstructing any image from the eavesdropped signal, which needs many samples from $\mathbb{X}$ and $\mathbb{Y}$. In contrast, the white-box attack directly reconstructs the image from the eavesdropped signal without any training in advance. Although the SSIM and the PSNR of the eavesdropped images in both attacks are lower than those of the images received by Bob, the eavesdropped images are visually recognizable and their privacy is compromised, which confirms the effectiveness of MIEA and reveals the risk of privacy leaks in current semantic communication.

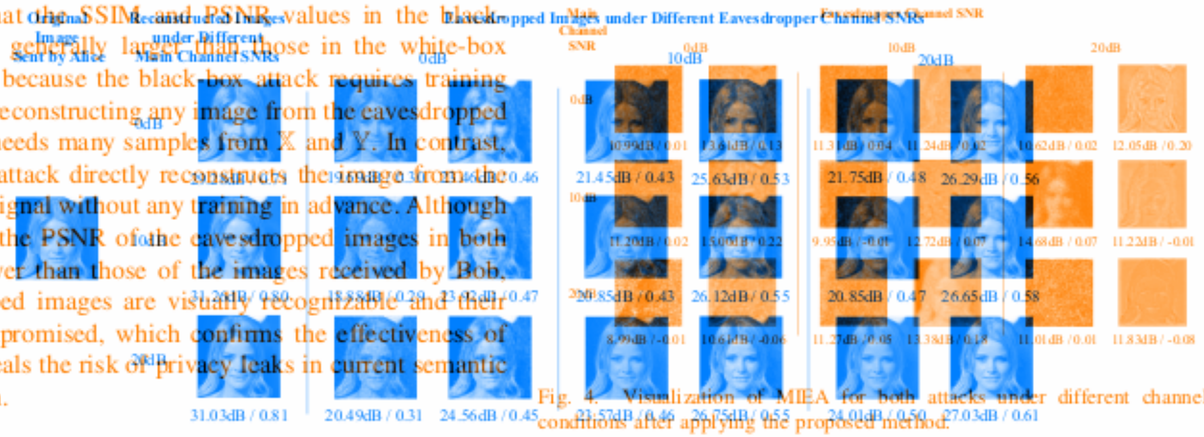


Fig. 3. Visualization of MIEA for the white-box attack and the black-box attack under different channel conditions. For each channel condition, the eavesdropped image by the white-box attack is displayed on the left and the one obtained by the black-box attack is on the right.

Fig. 4. Visualization of MIEA for both attacks under different channel conditions after applying the proposed method.

TABLE I
THE AVERAGE SSIM AND PSNR OF THE EAVESDROPPED IMAGES UNDER DIFFERENT CHANNEL CONDITIONS

	Main Channel SNR		
	0dB	10dB	20dB
Reconstructed Images by Bob	30.02dB / 0.70	32.28dB / 0.79	33.28dB / 0.80
EC[1] SNR 0dB	17.44dB / 0.31 21.65dB / 0.46	16.69dB / 0.30 22.24dB / 0.46	18.58dB / 0.32 22.85dB / 0.46
EC SNR 10dB	18.58dB / 0.40 23.50dB / 0.53	17.80dB / 0.40 23.76dB / 0.54	20.56dB / 0.43 24.33dB / 0.56
EC SNR 20dB	18.74dB / 0.43 23.88dB / 0.57	17.94dB / 0.42 23.98dB / 0.59	20.84dB / 0.46 24.41dB / 0.61

[1] EC refers to the eavesdropper channel.

TABLE II
THE AVERAGE SSIM AND PSNR OF THE EAVESDROPPED IMAGES BY MIEA AFTER DEFENSE

EC \ MC[1]	0dB	10dB	20dB
0dB	8.03dB / 0.02 11.36dB / 0.11	8.74dB / 0.02 11.41dB / 0.07	6.94dB / 0.00 12.51dB / 0.07
10dB	8.55dB / 0.04 11.72dB / 0.16	7.70dB / -0.01 11.34dB / 0.11	9.07dB / 0.05 13.22dB / 0.13
20dB	8.02dB / 0.03 12.59dB / 0.21	13.31dB / 0.09 11.55dB / 0.10	8.39dB / 0.02 11.54dB / 0.07

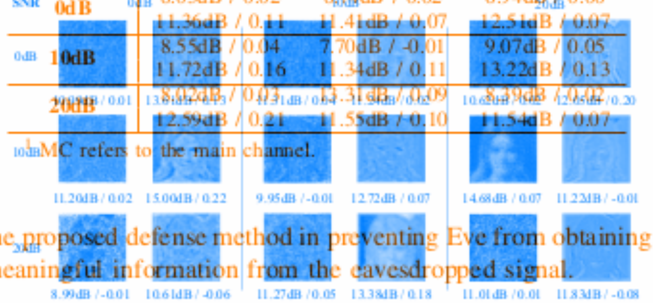
[1] MC refers to the main channel.

C. Evaluation of the Proposed Defense Method

Next, we evaluate the proposed defense method by repeating the evaluation of MIEA in section ?? with the defense method implemented on y_r.

Fig. ?? visualizes the eavesdropped images by the two attack types after applying the proposed defense method, using the same individual as in Fig. ??. It can be observed that the eavesdropped images are visually unrecognizable, demonstrating the effectiveness of the proposed defense method in preventing Eve from eavesdropping on raw images. We can also ... attack is superior to that against the black-box attack. This is because that Eve has no prior knowledge of ... $f^{-1}(x)$ ... the effectiveness of the proposed defense method in preventing Eve from obtaining meaningful information from the eavesdropped signal.

In addition, we provide the average SSIM and PSNR of the eavesdropped images ... in Table ?? ... PSNR are relatively small, which indicates the effectiveness of



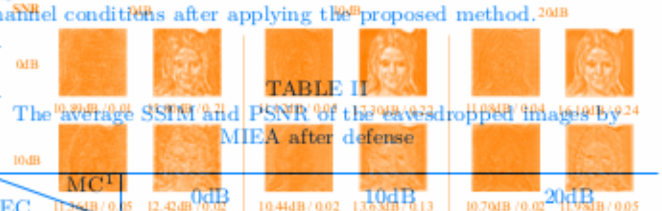


Fig. 5. Visualization of MIEA for both attacks under different channel conditions after applying random permutation.

Next, we conduct an ablation study to further validate our proposed ... random permutation ... the average SSIM and PSNR are relatively small, which indicates the effectiveness of the proposed defense method in preventing Eve from obtaining meaningful information from the eavesdropped signal.

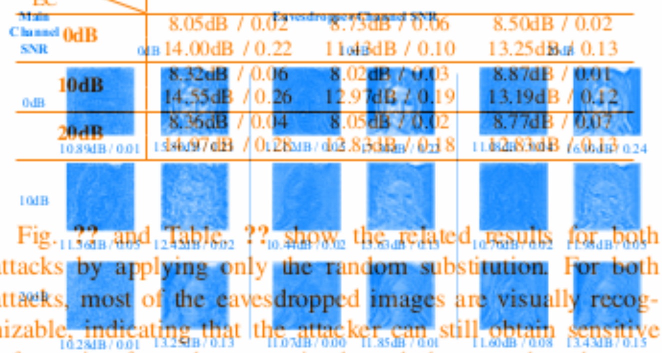


Fig. 5. Visualization of MIEA for both attacks under different channel conditions after applying only the random permutation.

TABLE III
The average SSIM and PSNR of the eavesdropped images when applying only random permutation

EC \ MC	0dB	10dB	20dB
0dB	8.05dB / 0.02 14.00dB / 0.22	8.73dB / 0.06 11.43dB / 0.10	8.50dB / 0.02 13.25dB / 0.13
10dB	8.32dB / 0.06 14.55dB / 0.26	8.02dB / 0.03 12.97dB / 0.19	8.87dB / 0.01 13.19dB / 0.12
20dB	8.36dB / 0.04 14.97dB / 0.28	8.05dB / 0.02 12.83dB / 0.18	8.77dB / 0.07 12.83dB / 0.13

Next, we conduct an ablation study to further validate our proposed method. Fig. ?? and Table. ?? demonstrate the eavesdropped images and the average SSIM and PSNR for both attacks by applying only the random permutation. As shown in Fig. ??, when only the random substitution is applied, the white-box attack can be effectively defended, while the black-box attack can still reconstruct visually recognizable images for some P and S. Moreover, the average SSIM and PSNR for the black-box attack in Table. ?? are larger than those in Table. ??, which demonstrates that only random permutation is insufficient for defending against MIEA.

Fig. ?? and Table. ?? show the related results for both attacks by applying only the random substitution. For both attacks, most of the eavesdropped images are visually recognizable, indicating that the attacker can still obtain sensitive information from the transmitted symbols, even though some of the semantic features have been substituted. The average SSIM and PSNR in Table. ?? are larger than those in Table. ??, which means that the random permutation is more effective than random substitution in defending against MIEA. From the ablation study, we can observe that the proposed defense method outperforms both the random-permutation-based and random-substitution-based defense methods, demonstrating that both permutation and substitution are essential for the effectiveness of the proposed defense method.

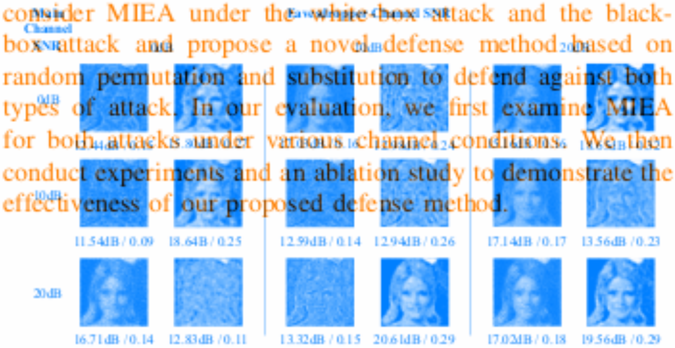


Fig. 6. Visualization of MIEA for both attacks under different channel conditions after applying only the random substitution.

TABLE IV
The average SSIM and PSNR of the eavesdropped images when applying only random substitution

EC \ MC	0dB	10dB	20dB
0dB	8.99dB / 0.14 16.06dB / 0.28	8.91dB / 0.10 15.00dB / 0.25	15.80dB / 0.16 14.70dB / 0.20
10dB	15.62dB / 0.19 14.68dB / 0.26	9.16dB / 0.14 14.49dB / 0.26	10.43dB / 0.16 15.80dB / 0.27
20dB	12.68dB / 0.18 15.78dB / 0.30	15.35dB / 0.18 14.53dB / 0.27	16.13dB / 0.21 17.29dB / 0.28

V. Conclusion

In this paper, we propose MIEA to expose privacy risks in semantic communication. MIEA enables an attacker to eavesdrop on the transmitted symbols through an eavesdropper channel and reconstruct the raw message by inverting the DL model employed in the semantic communication system. We consider MIEA under the white-box attack and the black-box attack and propose a novel defense method based on random permutation and substitution to defend against both types of attack. In our evaluation, we first examine MIEA for both attacks under various channel conditions. We then conduct experiments and an ablation study to demonstrate the effectiveness of our proposed defense method.