KAP-DJSCC: Key-Assisted Protection for Deep Joint Source-Channel Coding Under Model Inversion Eavesdropping Attacks

Mahshid Narimani Kenari, Nicola Cordeschi, Luigi Alfredo Grieco
Dept. of Electrical and Information Engineering
Politecnico di Bari
Bari, Italy
m.narimanikenari@phd.poliba.it, nicola.cordeschi@poliba.it, alfredo.grieco@poliba.it

***Abstract*—Semantic communication, particularly Deep Joint Source-Channel Coding (DJSCC), has emerged as a promising solution for efficient image transmission in future communication systems. However, the broadcast nature of wireless communication and correlation between the compressed and original data poses significant security risks, particularly in the form of Model Inversion Eavesdropping Attacks (MIEAs). Existing defense mechanisms suffer from limitations, such as high computational overhead and information loss due to quantization and adversarial training. As a result, these methods are unsuitable for applications that demand fast, secure, and reliable information transmission. In this paper, we propose a secure, lightweight, and reversible framework named Key-Assisted Protection for Deep Joint Source-Channel Coding (KAP-DJSCC), which utilizes a Diffie-Hellman (DH)-based key exchange to construct a shared secret transformation matrix. This pluggable process is applied to obscure the latent representation of the input without altering its structure or requiring retraining. Additionally, we introduce a novel MIEA variant, Key-assisted Model Inversion Eavesdropping Attack (KMIEA), in which the attacker attempts to guess the key. Simulation results confirm that KAP-DJSCC significantly degrades the eavesdropper's reconstruction performance in terms of Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Perceptual Image Patch Similarity (LPIPS) while preserving high fidelity for the legitimate receiver across varying Signal-to-Noise Ratio (SNR) levels.**

***Index Terms*—Secure DJSCC, wireless eavesdropping, 6G, secure image transmission, model inversion attacks**

I. Introduction

The sixth generation (6G) wireless communication system is expected to be inherently Artificial Intelligence (AI)-driven, supporting a wide range of applications such as intelligent transportation, virtual and augmented reality, and the industrial internet [1], [2]. In these applications, visual data plays a dominant role, with image and video content accounting for approximately 75% of current IP traffic. To transmit this massive volume of data, an efficient compression scheme is essential to avoid network congestion and transmission delays. To solve this issue, semantic communication has emerged as a promising paradigm that leverages AI to extract and transmit the most relevant semantic information from images over wireless channels [3].

Previous studies on semantic image transmission proposed Deep Joint Source-Channel Coding (DJSCC) methods that performed effectively under harsh channel conditions, including low Signal-to-Noise Ratio (SNR) and limited bandwidth [4]. The DJSCC approach presented in [4] maps image pixel values directly to complex-valued channel input symbols and learns noise-tolerant compressed representations. This mechanism helps avoiding the sudden quality degradation at low SNRs (cliff-effect) which commonly occurs in conventional communication systems that use separate source and channel coding. By employing DJSCC, image reconstruction quality degrades gracefully in the presence of adverse channel conditions.

Since DJSCC integrates characteristics of both traditional wireless communication and AI, it is susceptible to emerging attacks: those targeting wireless channels—due to their open and broadcast nature—and those aimed at Deep Neural Network (DNN) models. For example, attacks like eavesdropping, spoofing, man-in-the-middle, and their adaptation to DNNs pose a growing threat to data privacy and reliability of these networks. Since the DJSCC encoder leverages input redundancies for compression, the resulting channel input signal remains highly correlated with the original image [5]. While this correlation enhances robustness in image reconstruction, especially in poor channel conditions, it also introduces potential privacy leakage risks. An eavesdropper can capture the wireless signal through its own channel, train a surrogate DNN decoder, and attempt to reconstruct or interpret the transmitted content [6], [7]. This class of eavesdropping attacks, known as Model Inversion Eavesdropping Attacks (MIEAs), has recently gained increasing attention in the context of semantic communication security [7], [8].

In the context of protecting DJSCC systems against eavesdropping attacks, several studies have proposed data-driven solutions [5], cryptographic techniques [9], and information-hiding [10] mechanisms. For instance, [5] employs the concept of *privacy funnel* to balance the trade-off between maintaining high reconstruction quality and preventing the eavesdropper

© 2025 IFIP

from inferring sensitive information. The idea of privacy funnel optimization, as introduced in [11], seeks to minimize the mutual information between disclosed and private data. Alternatively, [9] employs encryption during training after extracting the latent space. The latent vector is quantized and subsequently handled as plaintext for encryption. Authors in [10] propose integrating lightweight adversarial modules into image transmission systems to mislead eavesdroppers. The method optimizes a weighted combination of privacy leakage, reconstruction error (Mean Squared Error (MSE)), and attack power, achieving improved security without significantly compromising image quality. These works assume a generic eavesdropper and do not address specific eavesdropping scenarios, particularly MIEAs.

To counter MIEAs in DJSCC, the early study in [7] utilizes permutation and substitution techniques applied directly to the latent representation. While this approach is resilient to both eavesdropping and channel noise, it requires twice the resources, as it sends two latent vectors instead of the original one. In addition, [12] applies steganography to create a semantically covert protection against MIEA. However, these mitigation methods mainly focus on data processing and architectural defense and do not integrate with telecommunication protocols. This limits their cross-layer adaptability and robustness in practical systems.

To the best of our knowledge, there are still some gaps in the literature that have not yet been addressed. Cryptographic methods [7], [9] can be resource-intensive, highlighting the need for simpler yet effective alternatives. Additionally, applying cryptography directly to the original image reduces its redundancy, which hinders compression using DNNs [9]. On the other hand, encrypting the latent space [9] also requires quantization, resulting in information loss. Moreover, data-driven models [5], [8] trade off full recovery and privacy by training the autoencoder with a combined loss function. Since these methods are application-specific and the defense is not pluggable, the model must be retrained when full information is needed at the receiver's side. In addition, existing approaches in the literature often neglect authentication methods and focus solely on DNN architecture design against MIEA. This discussion highlights the need for a fast and lightweight algorithm that provides security of DJSCC against MIEA by leveraging existing components in telecommunication protocols.

In this paper, we propose a novel secret key assisted protection method for DJSCC, named as Key-Assisted Protection for Deep Joint Source-Channel Coding (KAP-DJSCC) to protect the semantic communication systems against MIEA. Our proposed method leverages the Diffie-Hellman (DH) key exchange protocol to establish a shared secret key known only to the legitimate transmitter and receiver. Unlike conventional cryptographic approaches that rely on computationally intensive calculations and require quantization [9], the proposed scheme uses the shared key to generate a pseudo-random transformation matrix and applies a lightweight and reliable process to conceal the transmitted semantic information. Since the transform is fully reversible at the receiver side, the original data remains unaltered. This makes the method particularly suitable when accurate information recovery is required. Moreover, a novel MIEA, referred to as Key-assisted Model Inversion Eavesdropping Attack (KMIEA) has been proposed. Simulation results demonstrate that KAP-DJSCC effectively reduces the eavesdropper's ability to recover the original content through KMIEA and MIEA by lowering the similarity between the transmitted and source data.

The remainder of the paper is structured as follows: Section II describes the system model for DJSCC and details the proposed protection mechanism. Section III introduces a novel variant of the MIEA attack based on random key guessing. Section IV presents and analyzes the simulation results. Finally, Section V concludes the paper.

Notation: In this paper, we refer to the transmitter as *Alice*, the legitimate receiver as *Bob*, and the eavesdropper as *Eve*, following standard terminology in the literature.

II. System Model and Protection

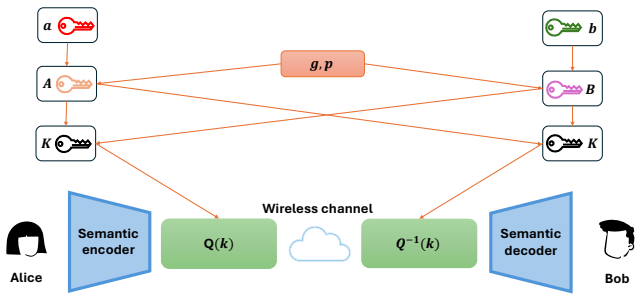

Fig. 1. Defense model

A. Encoding

In DJSCC, Alice employs an encoder $f : \mathbb{R}^N \to \mathbb{C}^c$ to extract semantic features from the input image $\mathbf{x}$ and compress them into a complex-valued latent vector $\mathbf{s}_{1\times c}$:

$$\mathbf{s} = f(\mathbf{x}; \phi), \tag{1}$$

where ϕ and c represent the trainable parameter set of the encoder and channel bandwidth according to [4], respectively.

B. Key Agreement via Diffie–Hellman Protocol

Initially, Alice and Bob negotiate a shared secret key K using the DH protocol. They publicly agree on a large prime p and a generator g. Then, Alice selects a private key $a \in \mathbb{Z}_p$ and computes $A = g^a \bmod p$, while Bob selects a private key $b \in \mathbb{Z}_p$ and computes $B = g^b \bmod p$. After exchanging A and B, both compute the identical shared key:

$$K = B^a \bmod p = A^b \bmod p. \tag{2}$$

C. Key-based Transformation

The shared key K is used as a seed for a PCG64 pseudorandom generator [13]. Then, a random $c \times c$ matrix is generated and orthogonalized via QR decomposition. The orthogonal matrix $\mathbf{Q}$ which is suitable for geometric transformations, serves as a lightweight encryption tool for the latent vector. We refer to this process as $\Omega(k)$, which takes K and generates the invertible transformation matrix $\mathbf{Q} \in \mathbb{R}^{c \times c}$:

$$\mathbf{Q} = \Omega(K), \tag{3}$$

where $\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}$. This matrix is then applied to the latent vector $\mathbf{s}$:

$$\mathbf{s}' = \mathbf{s}\mathbf{Q}^\top. \tag{4}$$

This reversible procedure preserves overall structure of the latent space but masks the latent representation from unauthorized reconstruction.

D. Transmission and Recovery

After the security module, the output $\mathbf{s}'$ is normalized to satisfy a given average power budget [4]. The normalized transformed latent vector is transmitted over an Additive White Gaussian Noise (AWGN) channel. The received signal at Bob's side is modeled as:

$$\hat{\mathbf{s}} = \mathbf{h} \circ \mathbf{s}'_n + \mathbf{n}, \tag{5}$$

where $\mathbf{h}$ is the channel gain vector, $\circ$ denotes element-wise multiplication, $\mathbf{s}'_n$ shows the latent vector after transformation and power normalization, and $\mathbf{n} \sim \mathcal{CN}(0, \sigma^2 \mathbf{I})$ is complex Gaussian noise.

Since Bob also possesses the shared key K and is aware of $\Omega(k)$, he reconstructs the orthogonal matrix $\mathbf{Q}$ and inverts the transformation to recover the original latent vector:

$$\tilde{\mathbf{s}} = \hat{\mathbf{s}}\mathbf{Q}. \tag{6}$$

E. Decoding

Bob then applies the decoder $f^{-1} : \mathbb{C}^c \to \mathbb{R}^N$ to reconstruct the original image from the latent representation:

$$\hat{\mathbf{x}} = f^{-1}(\tilde{\mathbf{s}}; \psi), \tag{7}$$

where ψ denotes the decoder's trainable parameters. The orthogonal nature of $\mathbf{Q}$ ensures that this transformation introduces no distortion, thereby preserving reconstruction quality.

In contrast, Eve without knowledge of the shared key cannot guess $\mathbf{Q}$ and hence cannot invert the transformation. This incapability significantly degrades Eve's ability to recover the semantic representation $\mathbf{s}$, providing enhanced confidentiality without altering the autoencoder's architecture or training procedure.

F. Training

In DJSCC, the encoder and decoder are jointly trained to minimize the expected reconstruction loss over a training dataset $\mathcal{D}_{\text{train}} = \{\mathbf{x}^{(i)}\}_{i=1}^{M}$:

$$\mathcal{R}_M(\phi, \psi) = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{\text{train}}} \left[\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}) \right], \tag{8}$$

where $\mathcal{L}(\cdot)$ is a chosen loss function. In this paper we use MSE loss. Note that the encoder and decoder are trained jointly, whereas the key exchange and transformation steps are applied separately to the latent space as non-trainable operations. In essence, the protection module is designed to be pluggable. Figure 1 illustrates the DH-aided defense integrated with the DJSCC.

III. KMIEA

In MIEA [6], [7], Eve queries the encoder illegally, receives the encoded data via its channel, and trains its *surrogate decoder*. We propose KMIEA, a *surrogate model training* [6] attack which trains its model after guessing K. In this attack, Eve leverages auxiliary information and illicitly accesses Alice's model predictions to train a surrogate model that approximates Bob's behavior. Since Eve does not know the true K, it uses a randomly chosen key K_e to simulate the defense mechanism. It is important to note that while Eve has no knowledge of Bob's decoder, it has full access to the encoder and knows the applied defense process in the transmitter's side. Since Eve does not know Bob's decoder or the reverse transform, it uses K_e to train its surrogate decoder on a transformed latent vector, enabling it to handle both the transformation and AWGN noise simultaneously. In fact, the model learns the mapping introduced by K_e. More specifically, KMIEA consists of the following steps:

- Query the encoder using an auxiliary training set D_{aux};
- Apply K_e to the queries using eq. (3) and (4).
- Receive the queries via the eavesdropping AWGN channel with its specific training noise $\mathbf{n}_e$;
- Use transformed noisy queries as inputs, and matching D_{aux} samples as targets, to form D_{adv} for inverse network training;
- Train the surrogate model that utilizes D_{adv} to approximate the inverse mapping.

For an image reconstruction attack in KMIEA, both the quality of D_{aux} and the design of the surrogate model are critical. Additionally, Eve's level of access to information about the encoder and the defense method significantly impacts the strength of the attack. All the steps employed in KMIEA and the training procedure are illustrated in Figure 2.

IV. Simulation results

In our simulations, we evaluated performance using widely adopted image quality metrics, including pixel-wise measures such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM), as well as the perceptual similarity metric, Perceptual Image Patch Similarity (LPIPS). The legitimate DJSCC system is implemented based on the baseline

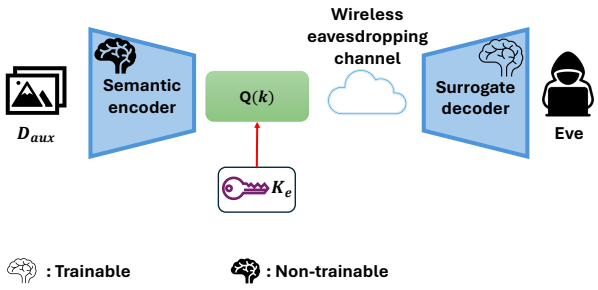

Fig. 2. KMIEA training model

architecture proposed in [4], while the architecture of Eve's surrogate decoder is presented in Table I. All models are implemented in Python using the TensorFlow framework, and trained with a learning rate of 0.0001 for both Bob and Eve. The CIFAR-10 dataset is used, with 40000 images allocated for training Bob, 10000 images for training Eve's surrogate decoder, and 10000 images for testing. Eve and Bob are both trained at SNR of 20 dB.

TABLE I
EVE'S DECODER ARCHITECTURE SUMMARY

No.	Layer
1	Conv2DTranspose, 5×5, stride 1
2	PReLU
3	Conv2DTranspose, 5×5, stride 2
4	PReLU
5	Conv2DTranspose, 5×5, stride 2
6	Sigmoid

Figure 3 illustrates the reconstructed image quality evaluated using the aforementioned metrics across different channel SNRs. The curve labeled Bob, KAP-DJSCC represents the reconstruction quality at Bob's side when the proposed defense mechanism is applied, while Bob, DJSCC shows the performance of the standard DJSCC system without any protection. The curves Eve,KMIEA and Eve,MIEA depict Eve's reconstruction performance when KAP-DJSCC is applied under KMIEA and MIEA [7], respectively. Furthermore, Eve,MIEA (Surrogate) curve provides the comparison with the benchmark DJSCC under MIEA. The results demonstrate that the proposed KAP-DJSCC method achieves image reconstruction quality nearly identical to the baseline DJSCC in all subfigures. This confirms that the defense mechanism introduces no recognizable information loss or decoding failure, thanks to utilizing reversible transformation applied to the transmitted latent space. Moreover, Eve with MIEA can achieve results closely converging to Bob when KAP-DJSCC is not implemented. In contrast, the defense substantially degrades Eve's reconstruction quality. While key guessing (K_e) and training a surrogate decoder based on it in KMIEA improves Eve's ability, her performance remains significantly inferior to Bob. Since the secret key is exchanged via the DH protocol, successful guessing is highly improbable, ensuring robust protection against MIEAs.

In subfigures (a) and (b), the trends of the curves are similar, as both PSNR and SSIM are pixel-level quality metrics. Consistent with typical DJSCC schemes [4], the proposed method and both attack models exhibit saturation at SNRs higher than the training SNR. In subfigure (a), the maximum degradation in Eve's reconstruction quality under KMIEA occurs at 25 dB, amounting to a 30.1 % reduction relative to the no-attack case, while under MIEA it amounts to a 51.7 % reduction. A similar trend is observed in subfigure (b), where decoding failure measured by SSIM peaks at 62.5% under KMIEA and 92.0% under MIEA. Unlike PSNR and SSIM, the lower values in LPIPS indicate better perceptual similarity. The defense method proves highly effective in this regard, substantially reducing perceptual reconstruction quality at Eve's side with maximum differences of 52.7% and 55.3% under KMIEA and MIEA, respectively. Although only the maximum differences are reported here, the performance gap between Bob and Eve remains consistently significant across all SNRs, demonstrating the robustness and effectiveness of the proposed KAP-DJSCC scheme.

Figure 4 illustrates qualitative reconstruction results at SNR levels of 20 dB and 0 dB for the proposed scheme and both attack models. At SNR = 20 dB, Bob with KAP-DJSCC recover the image with high fidelity compared to DJSCC. The results confirm that the proposed DH-aided latent space transformation preserves reconstruction quality for Bob, maintains image texture, and accurately recovers the semantic content. In contrast, Eve's reconstructions under both attack models exhibit severe quality degradation when KAP-DJSCC is applied. KMIEA achieves relatively better values and reveals scrambled textures and color regions compared to MIEA, but the image still remains difficult to interpret. This proves the advantage of using a secret key exchange. Under MIEA, only noisy and unrecognizable outputs are obtained. In very low SNR regime (0 dB), Bob's reconstructions degrade slightly when KAP-DJSCC is applied while the semantic content remains clearly visible. Notably, Eve under MIEA produces no meaningful reconstruction, confirming that KAP-DJSCC effectively obscures the latent representations and protects the semantic contents against eavesdropping.

V. CONCLUSION

This paper introduced KAP-DJSCC, a lightweight and reversible defense framework that safeguards semantic image transmission via DJSCC against MIEA. Using a DH-based key exchange protocol, the proposed method applies a key-assisted orthogonal transformation to the transmitted latent space, which enhances security without altering the original meaning or requiring retraining. To assess the efficiency, we introduced a new attack, KMIEA, which performs key guessing and surrogate model training. The simulations using PSNR, SSIM, and LPIPS metrics demonstrate that KAP-DJSCC preserves high reconstruction quality at the legitimate receiver while significantly degrading the eavesdropper's performance across all channel SNRs. Visual results also confirm the effectiveness of KAP-DJSCC in protecting semantic content. Overall, KAP-

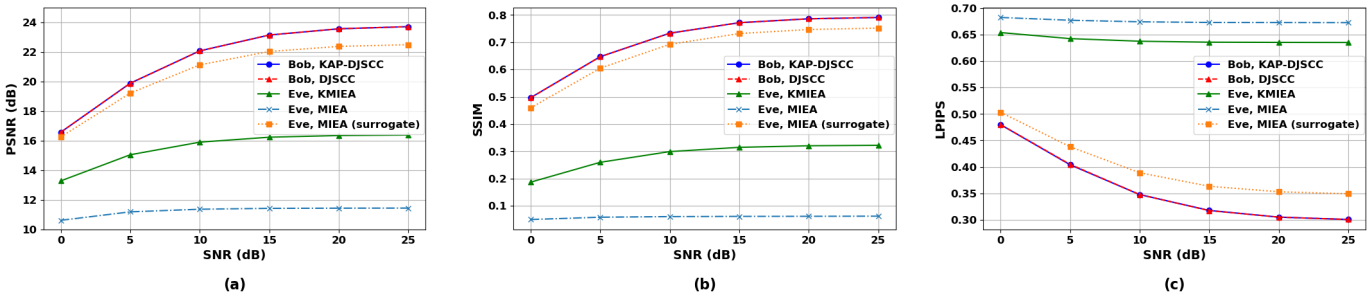

Fig. 3. PSNR, SSIM, and LPIPS vs. SNR for Bob and Eve under different schemes. The proposed KAP-DJSCC method preserves high reconstruction quality for Bob while effectively degrading Eve's performance under both KMIEA and MIEA attacks.

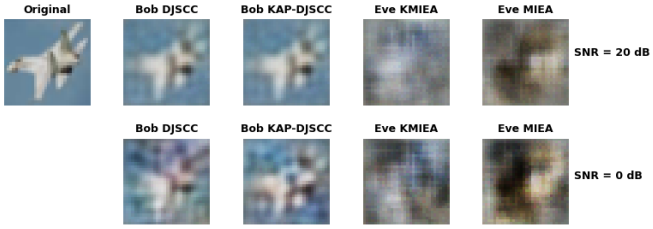

Fig. 4. Reconstructed images at SNR = 20 dB (top) and 0 dB (bottom). Shown are outputs from Bob (with/without KAP-DJSCC) and Eve (KMIEA and MIEA) under KAP-DJSCC.

DJSCC presents a practical and computationally efficient solution for security in future 6G semantic communication systems, bridging the gap between cryptographic key exchange and semantic security. For future work, a more advanced DH method such as Elliptic-Curve DH is recommended. The basic DH, although efficient, remains vulnerable to man-in-the-middle attacks, which could also serve as an entry point for multidomain attacks.

Acknowledgment

This work has received funding from the Smart Networks and Services Joint Undertaking (SNS JU) project 6G-GOALS under the European Union's Horizon Europe research and innovation program under Grant Agreement No 101139232.

References

[1] Z. Qin, X. Tao, J. Lu, W. Tong, and G. Y. Li, "Semantic communications: Principles and challenges," *arXiv preprint arXiv:2201.01389*, 2021.

[2] E. C. Strinati, P. Di Lorenzo, V. Sciancalepore, A. Aijaz, M. Kountouris, D. Gündüz, P. Popovski, M. Sana, P. A. Stavrou, B. Soret *et al.*, "Goal-oriented and semantic communication in 6g ai-native networks: The 6g-goals approach," in *2024 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*. IEEE, 2024, pp. 1–6.

[3] D. Huang, F. Gao, X. Tao, Q. Du, and J. Lu, "Toward semantic communications: Deep learning-based image semantic coding," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 55–71, 2022.

[4] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 567–579, 2019.

[5] M. Letafati, S. A. A. Kalkhoran, E. Erdemir, B. H. Khalaj, H. Behroozi, and D. Gündüz, "Deep joint source channel coding for privacy-aware end-to-end image transmission," *IEEE Transactions on Machine Learning in Communications and Networking*, 2025.

[6] S. V. Dibbo, "Sok: Model inversion attack landscape: Taxonomy, challenges, and future roadmap," in *2023 IEEE 36th Computer Security Foundations Symposium (CSF)*. IEEE, 2023, pp. 439–456.

[7] Y. Chen, Q. Yang, Z. Shi, and J. Chen, "The model inversion eavesdropping attack in semantic communication systems," in *GLOBECOM 2023-2023 IEEE Global Communications Conference*. IEEE, 2023, pp. 5171–5177.

[8] Y. Wang, S. Guo, Y. Deng, H. Zhang, and Y. Fang, "Privacy-preserving task-oriented semantic communications against model inversion attacks," *IEEE Transactions on Wireless Communications*, vol. 23, no. 8, pp. 10 150–10 165, 2024.

[9] T.-Y. Tung and D. Gündüz, "Deep joint source-channel and encryption coding: Secure semantic communications," in *ICC 2023-IEEE International Conference on Communications*. IEEE, 2023, pp. 5620–5625.

[10] B. He, F. Wang, and T. Q. Quek, "Secure semantic communication via paired adversarial residual networks," *IEEE Wireless Communications Letters*, 2024.

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

[12] S. Tang, Y. Chen, Q. Yang, R. Zhang, D. Niyato, and Z. Shi, "Towards secure semantic communications in the presence of intelligent eavesdroppers," *arXiv preprint arXiv:2503.23103*, 2025.

[13] M. E. O'neill, "Pcg: A family of simple fast space-efficient statistically good algorithms for random number generation," *ACM Transactions on Mathematical Software*, vol. 204, pp. 1–46, 2014.