RepObE: Representation Learning-Enhanced Obfuscation Encryption Modular Semantic Task Framework

Limei Lin[1*], **Jinpeng Xu**[1], **Xiaoding Wang**[1], **Liang Chen**[1], **Sun-Yuan Hsieh**[2] and **Jie Wu**[3,4]

[1]College of Computer and Cyber Security, Fujian Provincial Key Laboratory of Network Security and Cryptology, Fujian Normal University, Fuzhou 350117, China
[2]Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan 701, Taiwan
[3]China Telecom Cloud Computing Research Institute, Beijing, 100088, China
[4]Department of Computer and Information Sciences, Temple University, PA 19122, USA
linlimei@fjnu.edu.cn, ubuj_175359@163.com, wangdin1982@fjnu.edu.cn, liangchen011208@gmail.com, hsiehsy@mail.ncku.edu.tw, jiewu@temple.edu

Abstract

Model inversion and adversarial attacks in semantic communication pose risks, such as content leaks, alterations, and prediction inaccuracies, which threaten security and reliability. This paper introduces, from an attacker's viewpoint, a novel framework called RepObE (Representation Learning-Enhanced Obfuscation Encryption Modular Semantic Task Framework) to secure semantic communication. This framework employs dynamic encryption during semantic extraction and feature transmission to hinder attackers from reconstructing data through eavesdropping, thus strengthening system privacy. To combat image communication task challenges, we propose a prototype adversarial collaborative alignment training approach enhanced by representation learning. This method extracts and encodes semantic features while using dynamic perturbation and robust optimization to improve system resilience against adversarial threats. The approach ensures reliable semantic communication in complex environments, maintaining performance while countering attacks using feature obfuscation, adversarial training, and representation learning. Experimental results demonstrate that our method surpasses existing techniques by more than 2% in resisting model inversion attacks on classification tasks. Visually, our method excels with minimal decipherable images for attackers. It also shows a 3% to 5% improvement in countering adversarial attacks on classification tasks.

1 Introduction

With 6G technology advancing swiftly, traditional techniques are increasingly unable to meet the requirements for high bandwidth, low latency, and intelligent processing. Semantic communication, which prioritizes sending the meaning of data over raw bits, enhances both communication efficiency and task accuracy [Lu *et al.*, 2024]. It also simplifies data transfer and boosts environmental awareness [Fu *et al.*, 2024]. This approach not only enhances efficiency but also introduces novel methods for intelligent information processing and resource optimization. Research has progressed from semantic extraction in single-mode data (e.g., text, images [Sagduyu *et al.*, 2024], audio [Chen *et al.*, 2024], video) to multimodal integration [Guo *et al.*, 2024], while addressing privacy and security issues, thus solidifying the theoretical and technical foundations of the field.

Recent advancements in semantic communication have transformed both practical applications and academia. In the Internet of Things (IoT), it boosts device interaction and resource management, enhancing system efficiency [Sang *et al.*, 2025]. For unmanned aerial vehicles (UAVs), it facilitates group coordination by extracting essential semantic data from missions, thereby reducing communication overhead and optimizing decision-making [Xu *et al.*, 2025]. In digital twin technology, which requires accurate, real-time synchronization between physical and virtual entities, semantic communication is crucial for industrial production monitoring [Du *et al.*, 2024]. Likewise, in vehicular networks, semantic communication tackles high latency and inefficiency issues in autonomous driving and V2V communication. By transmitting crucial semantic details about the environment and driving actions, it aids decision-making, improves safety, and optimizes traffic management [Xu *et al.*, 2023].

Recent research on privacy and security within semantic communication has gained traction due to the complexity of extracting and processing deep semantic information. To counter data breach risks, several privacy-preserving methods are suggested, such as adversarial learning [Wang *et al.*, 2025], location privacy protection [Qiu *et al.*, 2023], and federated learning frameworks [Wang *et al.*, 2024b]. Moreover, information bottleneck theory [Zhang *et al.*, 2024] and variational inference [Li *et al.*, 2024] have been examined to

*Corresponding author

improve feature extraction and model training, thereby indirectly enhancing privacy. However, these predominantly focus on data privacy, often overlooking crucial model security challenges. Existing approaches fall short against sophisticated threats, including model inversion attacks, which reconstruct sensitive data [Wang *et al.*, 2024a], and adversarial attacks, which disrupt systems and degrade performance using minimal perturbations [Zhang *et al.*, 2025]. This paper introduces an innovative solution, with the primary contributions summarized below.

Innovative Framework for Privacy Protection. We propose the RepObE framework, which dynamically encrypts data during the stages of semantic extraction and feature transmission. This framework effectively prevents attackers from reconstructing data through eavesdropping, significantly enhancing the privacy protection of the system.

Enhanced Resilience in Image Communication Tasks. To further strengthen the resilience of the system, we introduce a collaborative adversarial training approach of prototype alignment based on learning. This approach leverages representation learning to extract and encode semantic features while integrating dynamic perturbation and robust optimization techniques.

Fusion Approach for Reliable Semantic Communication. Our fusion approach ensures reliable and stable semantic communication in complex scenarios. It maintains high performance while defending against attacks through feature obfuscation, adversarial training, and representation learning.

Experimental Results Demonstrating Superior Performance. Experimental results show that our method outperforms existing methods by more than 2% in resisting model inversion attacks on the MNIST dataset's classification task. Visually, our approach performs the best, making it difficult for attackers to obtain meaningful images. Our method also improves by 3% to 5% in resisting adversarial attacks on the CIFAR10 dataset's classification task.

The rest of this paper is organized as follows. Section 2 is related work. Section 3 proposes the framework design. Section 4 presents the experimental results. Finally, Section 5 summarizes the paper.

2 Related Works

2.1 Image Semantic Communication

Advancements in AI and communication technologies have prioritized efficient image data transmission while retaining critical information. In image classification, Liu *et al.* employed semantic compression to reduce transmission load and latency while improving performance [Liu *et al.*, 2021], and Hu *et al.* integrated masked VQ-VAE with adversarial training to enhance robustness and multitask optimization [Hu *et al.*, 2023]. For UAV scene classification, Kang *et al.* used deep reinforcement learning to select key semantic blocks, optimizing efficiency [Kang *et al.*, 2022]. In segmentation and autonomous driving, Pan *et al.* designed a system that significantly improved segmentation performance [Pan *et al.*, 2023]. For image retrieval, Jankowski *et al.* [Jankowski *et al.*, 2020] and Xie *et al.* [Xie *et al.*, 2022] optimized single-user and multi-user scenarios, enhancing retrieval and transmission. In reconstruction tasks, Yang *et al.* leveraged the Swin Transformer to improve representation and efficiency [Yang *et al.*, 2023]. For multitask scenarios, Zhang *et al.* utilized semantic encoding with a data adaptation network [Zhang *et al.*, 2023], and Wu *et al.* introduced cross-task transfer to reduce storage needs and enhance performance [Wu *et al.*, 2022]. Additionally, Liu *et al.* demonstrated a task-oriented framework that effectively reduced data volume and latency in classification and reconstruction tasks [Liu *et al.*, 2022].

2.2 Privacy and Security of Neural Networks

Semantic communication enhances transmission efficiency but faces significant challenges in data security and privacy protection [Nan *et al.*, 2023]. Neural network-based encoders and decoders are vulnerable to threats like model inversion, inference, and adversarial attacks, leading to potential privacy breaches and compromised security [Liu *et al.*, 2024a]. To mitigate these threats, researchers have proposed defense strategies such as combining information bottleneck theory with adversarial learning [Wang *et al.*, 2024a], employing adversarial training for robustness, and implementing hierarchical defenses for vehicular metaverse applications [Kang *et al.*, 2023]. Random feature space perturbations [Xu *et al.*, 2021] and encryption methods like homomorphic encryption and secure multi-party computation [Liu *et al.*, 2024b; Patra *et al.*, 2021] further protect sensitive data and model parameters. However, existing approaches still fall short in addressing privacy protection and security for image-focused semantic communication systems.

Existing literature on semantic communication for image data has improved efficiency but lacks adequate solutions for data security and privacy. Despite proposed defense strategies, neural network-based encoders and decoders remain vulnerable to threats. This paper introduces the Representation Learning-Enhanced Obfuscation Encryption Modular Semantic Task Framework to enhance both efficiency and security of image-focused semantic communication systems.

3 RepObE Framework Design

The proposed system model for this paper can be found in supplementary materials. In this section, we propose the RepObE framework for secure semantic communication, as shown in Figure 1. We dynamically encrypt data during processing to prevent eavesdropping and enhance privacy. For image communication, we introduce a representation learning-enhanced adversarial training approach. This approach encodes features and integrates techniques to strengthen robustness against attacks. It ensures reliable communication in complex scenarios while maintaining high performance. The training algorithm is shown in Algorithm 1.

3.1 Design of Semantic Encoder-Decoder

Design of Semantic Encoder. The input 224×224 image is processed through a 7×7 convolutional layer, followed by max-pooling, reducing the feature map to 56×56 with 64 channels. Two ResNet blocks extract deep features, maintaining the size and channel count. Subsequent residual blocks

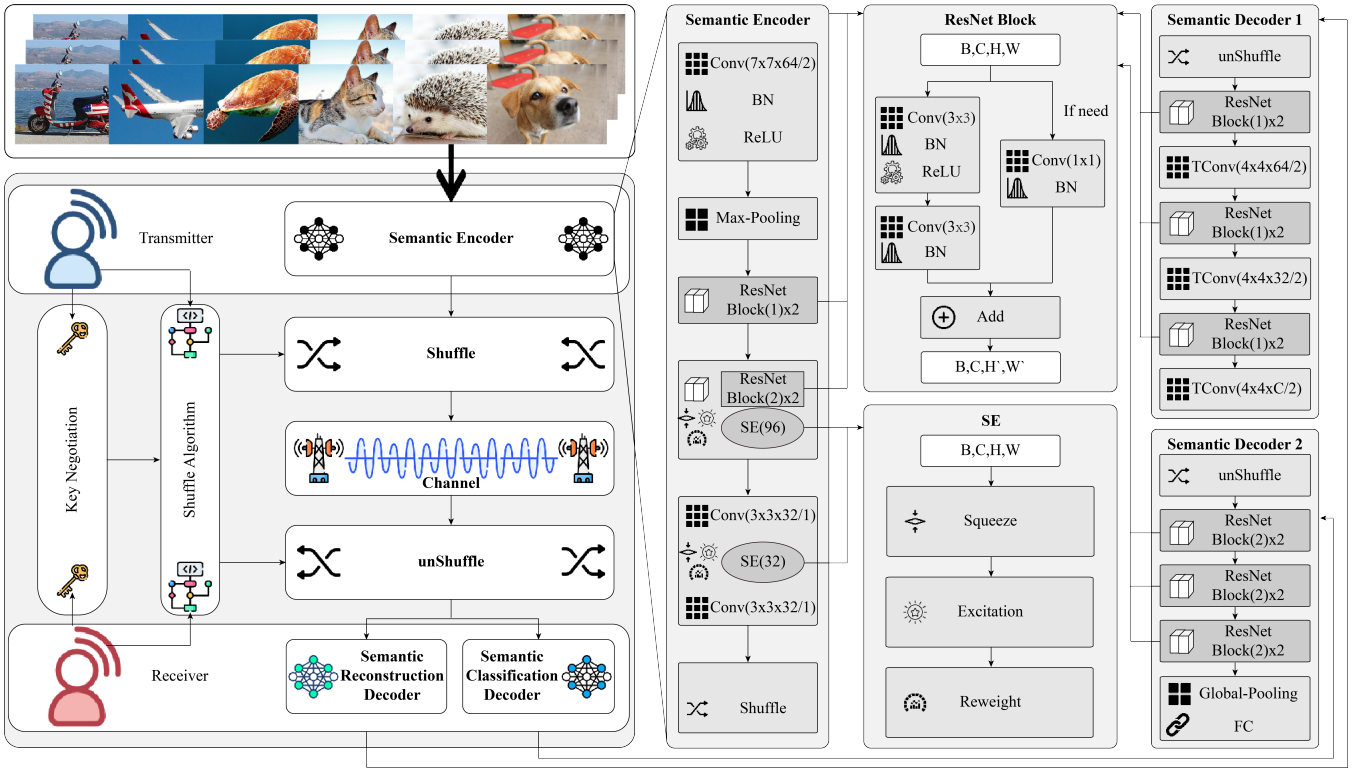


Figure 1: RepObE: Representation learning-enhanced obfuscation encryption modular semantic task framework.

with an SE module increase channels to 96 and reduce the size to 28×28. A convolutional layer reduces the channels to 4, followed by 3×3 convolutions with an SE module, adjusting the channels between 32 and 4. A final channel obfuscation operation produces the output.

Design of Semantic Reconstruction Decoder. The feature map of size $4 \times 28 \times 28$ first undergoes an unShuffle operation for channel rearrangement. It then passes through two ResNet blocks, producing a feature map of size $64 \times 28 \times 28$. A transposed convolution increases the size to $64 \times 56 \times 56$, followed by two more ResNet blocks that reduce the channels to 32 while maintaining the size. Another transposed convolution expands the size to $32 \times 112 \times 112$. Two additional ResNet blocks further reduce the channels to 16, keeping the size unchanged. Finally, a transposed convolution adjusts the channels to C, outputting an image of size $C \times 224 \times 224$.

Design of Semantic Classification Decoder. The $4 \times 28 \times 28$ input is first unshuffled to rearrange channels. It then passes through two ResNet blocks, producing an output with 64 channels and a feature map size of $64 \times 14 \times 14$. Next, two additional ResNet blocks increase the channels to 128, reducing the size to $128 \times 7 \times 7$. Another two ResNet blocks further increase the channels to 256 and reduce the size to $256 \times 4 \times 4$. Finally, global pooling is applied, followed by a fully connected layer, generating the final classification output with a dimension of 10.

Optimization of Semantic Encoder-Decoder. To ensure secure and practical transmission, the input image size $I \in R^{C \times H \times W}$ is set to at least 224×224. Both the encoder and decoder leverage convolutional neural networks, incorporating ResNet blocks and SE attention mechanisms for effective feature extraction. For an input image of resolution 224×224, the feature map resolution is progressively reduced to 28×28 through convolution, pooling, and downsampling, while the number of channels increases from 3 to 4. ResNet blocks enhance feature extraction, while SE blocks capture rich semantic features, improving the model's representational capacity. Multi-step downsampling minimizes computational complexity. A confusion layer, utilizing a shift algorithm for image confusion and deconfusion, is integrated at the sender's input and receiver's output. Each ResNet block comprises two 3×3 convolution layers and a residual connection. The formula for a ResNet block is

$$F' = ReLU(Rn(F) + F), \tag{1}$$

where the input feature map is F, $Rn(F)$ represents the result of two convolution operations, and the output feature map is F'. The SE Block includes global average pooling, which compresses the spatial dimensions $H \times W$ to the channel dimension, extracting global features for each channel. Then, two fully connected layers are used to generate the attention weights for the channels, and finally, each channel is weighted to enhance the representation of important features. The formula is

$$w = \sigma(w_2 \cdot ReLU(w_1 \cdot GAP(F))), F' = w \cdot F, \tag{2}$$

where w_1 and w_2 are model weights, GAP denotes global average pooling, and w represents the computed weights. The

Algorithm 1 RepObE Training Algorithm

Require: A dataset D_i, epochs e, batch size B, initialize semantic encoder network $S_\theta(\cdot)$, semantic reconstruction/classification decoder network $G_{\theta_1}(\cdot)$/$G_{\theta_2}(\cdot)$.
Ensure: The trained $S_\theta(\cdot)$, $G_{\theta_1}(\cdot)$ or $G_{\theta_2}(\cdot)$.
1: The transmitter and receiver negotiate the key $K = (K_r, K_c, r, c)$;
2: Set epoch counter $e = 1$;
3: **while** the training stop condition is not met **do**
4: The transmitter generates adversarial sample B' based on PGD algorithm and dataset B;
5: The transmitter encrypts $\chi = S_\theta(B' + B)$ based on the key K and formulas (4), (5), (6), (7);
6: Transmit χ over the channel, receiving $\widetilde{\chi}$ at the destination;
7: The receiver decrypts $\widetilde{\chi}$ based on the key K and formulas (8), (9), (10), (11);
8: The receiver calculates the loss function L_{rec} (for the reconstruction task) or L_{class} (for the classification task);
9: Update parameters of semantic encoder network $S_\theta(\cdot)$ and decoder network $G_{\theta_1}(\cdot)$ or $G_{\theta_2}(\cdot)$;
10: $e = e + 1$;
11: **end while**

feature transmitted over the channel is a k-dimensional vector z. Considering the output power p from the sender, we impose an average power constraint on the transmitted signal vector, expressed as

$$\hat{z} = \sqrt{kp} \cdot z / \sqrt{z \cdot z}. \tag{3}$$

The encoder and decoder are jointly trained, with the channel modeled as a non-trainable layer.

3.2 Confusion Offset

Before entering the input channel, assume the semantic encoder at the sender needs to send a feature map of size $F_s \in R^{a \times H' \times W'}$, where a represents the number of channels in the feature map, H' is the height, and W' is the width of the feature map. The sender and receiver negotiate a key $K = (K_r, K_c, r, c)$, when K_r is the row permutation key of size H', K_c is the column permutation key of size W', r is the row offset base, and c is the column offset base.

The encryption process consists of the following four steps.

Row Confusion. Rearrange the rows of each channel a of the feature map F according to the order specified by the key K_r. Let the i-th row of F be denoted as $F[i, :]$, and the confusion feature map is denoted as F_1 as follows.

$$F_1[a, i, :] = F_s[a, K_r[i], :], \tag{4}$$

where $K_r[i]$ is the i-th element in K_r, indicating that the $K_r[i]$-th row of the original feature map is mapped to the i-th row in the confused feature map.

Column Offset. After the row confusion is completed, the feature map F_1 undergoes a column shift for each row based on the column offset base c and the current row number i. The offset feature map is denoted by F_2 as follows.

$$F_2[a, i, j] = F_1[a, i, (j + i + c) \bmod W']. \tag{5}$$

The column offset of each row increases incrementally. The 0-th row has an offset of $0 + c$, the 1st row has an offset of $1 + c$, and so on, until the $(H' - 1)$-th row, where the offset is $H' - 1 + c$.

Column Confusion. For each channel c, the columns of F_2 are permuted according to the order specified by the key K_c. The permuted feature map is denoted by F_3, and the operation is defined as

$$F_3[a, :, j] = F_2[a, :, K_c[j]], \tag{6}$$

where $K_c[j]$ is the j-th element in K_c, indicating that the $K_c[j]$-th column of the original feature map is mapped to the j-th column in the confused feature map.

Row Offset. After the column confusion is completed, the feature map F_3 undergoes a row shift for each column based on the row offset base r and the current column index j. The offset feature map is denoted by F_{en} as follows.

$$F_{en} = F_3[a, (i + j + r) \bmod H', j]. \tag{7}$$

The row offset for each column increases incrementally. The 0-th column has an offset of $0+r$, the 1st column has an offset of $1 + r$, and so on, until the $(W' - 1)$-th column, where the offset is $(W' - 1) + r$.

The decryption process is composed of the following four steps.

Inverse Row Offset.

$$F_1' = F_{en}[a, (i - j - r) \bmod H', j]. \tag{8}$$

Inverse Column Confusion.

$$F_2'[a, :, K_c[j]] = F_1'[a, :, j]. \tag{9}$$

Inverse Column Offset.

$$F_3'[a, i, j] = F_2'[a, i, (j - i - c) \bmod W']. \tag{10}$$

Inverse Row Confusion.

$$F_{de}[a, K_r, :] = F_4[a, i, :]. \tag{11}$$

Theorem 1. *The encryption process consists of row confusion, column offset, column confusion, and row offset. These steps introduce significant complexity for attackers attempting to reconstruct the original feature map. The total key space size of the encryption process is given as follows.*

$$|K| = H'! \cdot W'! \cdot H \cdot W. \tag{12}$$

Proof. The row confusion step uses a key K_r, which is a permutation of length H'. The total number of possible row permutations is $H'!$.

The column confusion step uses a key K_c, which is a permutation of length W'. The total number of possible column permutations is $W'!$.

The row offset r and column offset c are values within the range from 0 to $H'-1$ and 0 to $W'-1$, respectively, providing H and W possible offset values for rows and columns.

Combining all the above steps, the total size of the key space is the product of all possible combinations, given by formula (12). □

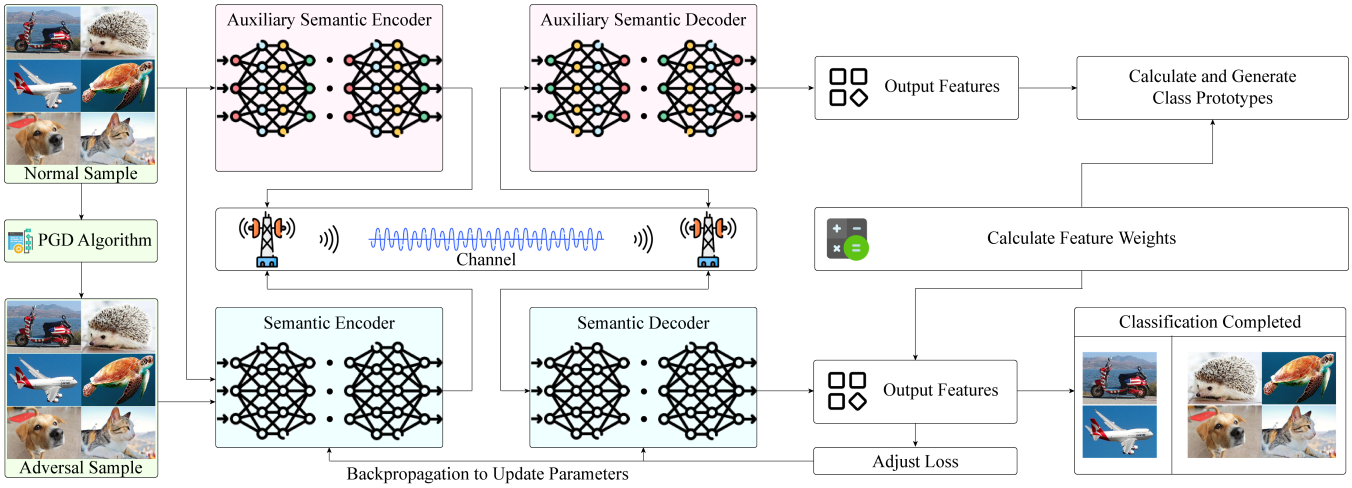


Figure 2: Representation learning-enhanced prototype alignment collaborative adversarial training.

3.3 Representation Learning-Enhanced Prototype Alignment Collaborative Adversarial Training

To defend against adversarial attacks, we adopt an adversarial training strategy to improve the robustness of the model, as shown in Figure 2 Adversarial training aims to enhance the model's ability to correctly classify natural samples while resisting the interference caused by adversarial samples by incorporating adversarial examples into the training process. The goal of generating adversarial samples is to maximize the model's loss function by adding perturbations to the input sample I. We use the PGD algorithm to generate adversarial samples. PGD is an iterative optimization method that incrementally finds the most disruptive adversarial samples through multiple updates. In each iteration, the loss function is denoted as $L_{\text{CE}}(M_\theta(I_{\text{adv}}), y)$, and the perturbation that increases the loss is denoted as formula (14).

Additionally, to enhance the representation learning capability of the model, we incorporate a contrastive learning framework during adversarial training. Contrastive learning leverages the relationships between adversarial and natural samples to enforce robust feature representations. Specifically, we treat adversarial samples I_{adv} and their corresponding natural counterparts I_{nat} as positive pairs while ensuring that different samples form negative pairs. The contrastive loss is expressed as

$$L_{\text{contrast}} = -\log \frac{\exp(\text{sim}(z_{\text{adv}}, z_{\text{nat}})/\lambda)}{\sum_i \exp(\text{sim}(z_{\text{adv}}, z_i)/\lambda)}, \tag{13}$$

where z_{adv} and z_{nat} are the embeddings of adversarial and natural samples, $sim(\cdot)$ represents the cosine similarity, and λ is the temperature parameter. By minimizing L_{contrast}, the model learns robust embeddings that are invariant to adversarial perturbations, improving both robustness and representation quality.

The goal of this paper is to maximize the model's adversarial robustness through adversarial training, while maintaining the classification performance on natural samples as much as possible. First, the adversarial loss is maximized to generate adversarial examples that can deceive the model, causing its predictions to deviate from the correct labels as follows.

$$\max_{\|I_{adv}-I_{nat}\|} D_{KL}\left(M_\theta(I_{nat}) \parallel M_\theta(I_{adv})\right), \tag{14}$$

where I_{nat} represents the adversarial sample generated by maximizing the adversarial robustness loss. The next step is to input both adversarial and natural samples into the model, and optimize it to make the model's prediction on adversarial samples as close as possible to that of natural samples. The adversarial training loss function introduced is

$$\min_\theta E_{(I,y)\sim D}\left[L_{CE}\left(M_\theta(I_{nav}), y\right) + formula\ (14)\right]. \tag{15}$$

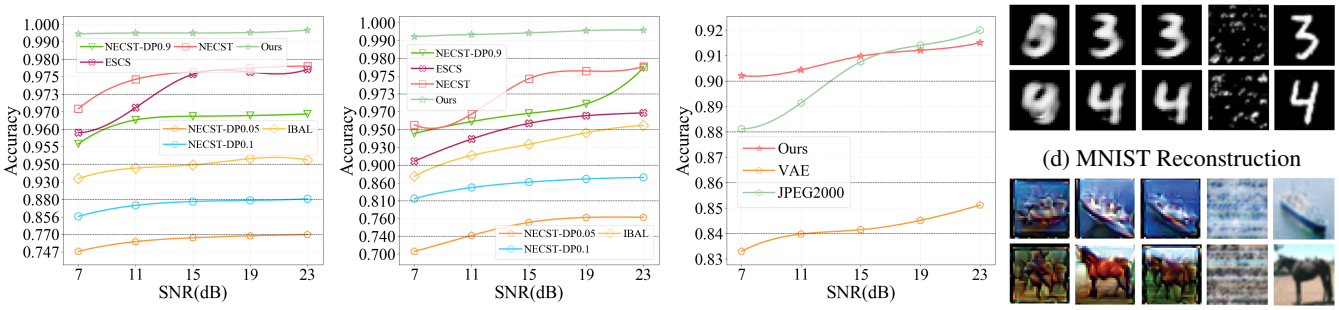
In each training iteration, the sender inputs natural samples and generates corresponding adversarial samples based on the current state of the model. Then, both natural and adversarial samples are fed into the model for training. To minimize the loss, we have designed a new loss function as follows.

$$\begin{aligned} L_{\text{class}} = &-\log \frac{\exp(\text{sim}(z_{\text{adv}}, z_{\text{nat}})/\lambda)}{\sum_i \exp(\text{sim}(z_{\text{adv}}, z_i)/\lambda)} \\ &+ \frac{1}{N}\sum_{n=1}^{N} L_{\text{CE}}\left(M_\theta(I_{\text{nat}}), y\right) \\ &+ \frac{\epsilon}{N}\sum_{n=1}^{N} \hat{w}(I_{\text{adv}}) D_{\text{KL}}\left(M_\theta(I_{\text{nat}}) \parallel M_\theta(I_{\text{adv}})\right). \end{aligned} \tag{16}$$

We introduce $\hat{w}(I_{\text{adv}})D_{KL}\left(M_\theta(I_{nat}) \parallel M_\theta(I_{adv})\right)$ as the loss for adversarial samples, which aims to reduce the prediction discrepancy between natural samples and adversarial samples. ϵ is a hyperparameter used to adjust the magnitude of KL divergence loss. The KL divergence is used to quantify the difference in output distributions between natural and adversarial samples, and is given by the following formula.

$$D_{KL}(P \parallel Q) = \sum P(x) \log P(x)/Q(x), \tag{17}$$

where P, Q represent the output distributions of natural and adversarial samples, respectively. By minimizing KL divergence, the model's prediction distributions for natural and adversarial samples are consistent, thereby enhancing robustness.



(a) MNIST under AGWN (b) MNIST under Rayleigh (c) CIFAR-10 under Rayleigh (d) MNIST Reconstruction (e) CIFAR-10 Reconstruction

Figure 3: Experimental study on classification performance of MNIST and CIFAR-10 datasets under different SNRs and channels.

Inspired by prototype alignment, suppose the dataset contains K classes. We first train an auxiliary model, and after training, the K-dimensional logits output for natural samples are denoted as the feature $p(I) \in R^K$, where

$$p(I) = [p_1(I), p_2(I), \ldots, p_K(I)], \tag{18}$$

and $p(x)$ refers to the high-dimensional feature before the model's output layer (or the logits output), which can be used as a high-level semantic representation of samples during adversarial training and model optimization.

For each class $y \in K$, the prototype $p_y \in R^K$ is the mean of each dimension of the feature of all samples in that class. The calculation is listed as follows.

$$p_y = (1/N_y) \sum_{i=1}^{N_y} p(I). \tag{19}$$

The goal is to bring the model's logits output for adversarial samples closer to the prototype of the corresponding class in natural samples, reducing the prediction discrepancy between natural and adversarial samples. The adversarial samples are assigned different weights $w(I_{adv})$ based on their deviation from the prototype of natural samples, where the weight is inversely proportional to the distance from the natural sample. A greater weight is assigned to adversarial samples that deviate further from the natural sample prototype, and a smaller weight is assigned to those closer to the prototype. The weight calculation formula is listed as follows.

$$w(I_{adv}) = 1 - \exp\left(-\|M_\theta(I_{adv}) - p_y\|^2/\tau\right), \tag{20}$$

where $\|M_\theta(I_{adv}) - p_y\|^2$ is the squared Euclidean distance between adversarial sample $M_\theta(I_{adv})$ and class prototype p_y.

The distance reflects how much an adversarial sample deviates from the natural sample distribution, with larger distances indicating harder-to-classify samples. The function $1-\exp(\cdot)$, a variant of the Gaussian kernel, maps the distance to a weight range of $[0, 1)$. The smoothing parameter τ controls the weight's sensitivity to distance. A larger τ results in more uniform optimization, while a smaller τ increases the weight of samples farther from the prototype, emphasizing classification boundaries. Weights are normalized to ensure the average weight remains 1, dynamically enhancing optimization for difficult samples without diminishing the overall adversarial loss impact as follows.

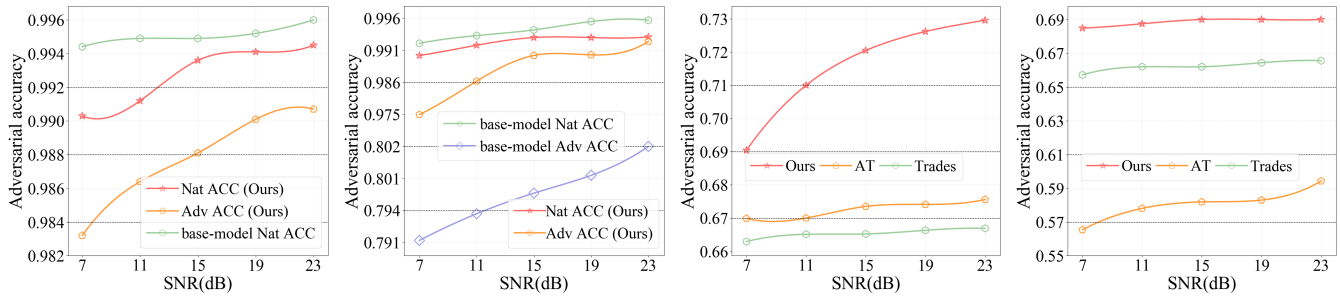
$$\hat{w}(I_{\text{adv}}) = N w(I_{\text{adv}}) / \sum_{n=1}^{N} w(I_{\text{adv}}). \tag{21}$$

4 Performance Evaluations

The experiments in this study utilized the MNIST and CIFAR-10 datasets, representing handwritten digit classification and more complex color image classification tasks, respectively. To simulate channel interference in a real wireless communication environment, the communication channel in the experiments was modeled as either an AWGN channel or a Rayleigh fading channel. The Signal-to-Noise Ratio (SNR) range was set from 7 to 23, covering typical communication conditions from low to high SNR. This experimental design aims to comprehensively evaluate the robustness of the proposed semantic communication framework across different datasets and channel environments, as well as its capability in privacy protection and resilience against adversarial attacks. Please refer to the supplementary materials for the experimental settings designed in this study.

4.1 Model Reversal Attack Experiment

As shown in Figures 3a and 3b, RepObE demonstrated superior performance on MNIST dataset under both AWGN and Rayleigh fading channels, with accuracy steadily improving as SNR increased (e.g., from 0.9943 to 0.9966 under AWGN and from 0.9920 to 0.9956 under Rayleigh). In contrast, NECST-DP [Choi *et al.*, 2019] exhibited significantly lower performance, particularly NECST-DP0.05. While IBAL [Wang *et al.*, 2024a] and ESCST performed well at low to medium SNRs, their improvements plateaued at higher SNRs, ultimately lagging behind ours. Additionally, NECST consistently outperformed NECST-DP across all SNR levels, reflecting the trade-off between privacy protection and performance. As shown in Figure 3c, RepObE exhibited robust performance on CIFAR-10 dataset under Rayleigh fading channel, improving from 0.9021 at SNR = 7 to 0.9151 at SNR = 23. While VAE [Mehrasa *et al.*, 2019] consistently underperformed with a maximum of 0.8512, JPEG2000 [Christopoulos *et al.*, 2000] slightly outperformed ours at high SNR (0.9203 at SNR = 23). Overall, ours demonstrated the best robustness under low to medium SNRs, while JPEG2000 showed a slight advantage at high SNRs.



(a) MNIST under AWGN (b) MNIST under Rayleigh (c) CIFAR-10 under AWGN (d) CIFAR-10 under Rayleigh

Figure 4: Adversarial attack experiments on MNIST and CIFAR-10 datasets under different SNRs and channels.

In Figure 3d and Figure 3e, the first four columns show images reconstructed by attackers using IBAL, NECST, NECST-DP0.9, and ours, respectively, under model inversion attacks. The fifth column displays images received by legitimate users with ours. Figure 3d presents results for MNIST digits "3" and "4", while Figure 3e shows CIFAR-10 categories "ship" and "horse". Visually, attacker-reconstructed images with ours are almost unrecognizable, containing no meaningful information, whereas legitimate users received clear and usable images.

These results confirm the effectiveness of our method in resisting model inversion attacks while maintaining high-quality image reconstruction and classification. Attackers failed to extract meaningful information, while legitimate users successfully received clear images, achieving a strong balance between privacy protection and task performance.

4.2 Adversarial Attack Experiment

As shown in Figure 4a the experimental results under the AWGN channel with MNIST dataset and adversarial training demonstrate that RepObE maintains high performance on both natural samples accuracy (Nat ACC) and adversarial samples accuracy (Adv ACC). Across all SNR conditions, the Nat ACC of our method remained stable between 0.9903 and 0.9945, closely matching the baseline model's Nat ACC range of 0.9944 to 0.9965, indicating that adversarial training has minimal impact on the classification performance of natural samples. Simultaneously, the Adv ACC of our method improved from 0.9832 at SNR = 7 to 0.9907 at SNR = 23, demonstrating strong robustness against adversarial attacks.

As shown in Figure 4b, the experimental results under Rayleigh fading channel with MNIST dataset and adversarial training ($\epsilon = 0.016$) indicate that RepObE outperforms the baseline model in both Nat ACC and Adv ACC. Across all SNR conditions, the Nat ACC achieved by RepObE ranged from 0.9902 to 0.9931, closely approaching the baseline model's Nat ACC range of 0.9921 to 0.9957, demonstrating stability in classifying natural samples. Moreover, the Adv ACC of RepObE was significantly higher than that of the baseline model. The Adv ACC of RepObE improved from 0.9745 at SNR = 7 to 0.9924 at SNR = 23, whereas the baseline model's Adv ACC only increased from 0.7912 to 0.8023.

The experimental results in Figures 4c and 4d demonstrate the superiority of the proposed method (ours) over comparative methods AT and Trades on the CIFAR-10 dataset with adversarial samples ($\epsilon = 0.016$, step size 0.0032, 5 iterations). Under the AWGN channel (Figure 4c), ours achieved steady performance gains, improving from 0.6904 at SNR = 7 to 0.7296 at SNR = 23, reflecting strong robustness and stability. In comparison, AT and Trades [Zhang *et al.*, 2019] showed limited improvements, with maximum accuracies of 0.6756 and 0.6669 at SNR = 23, respectively, consistently lagging behind ours. Similarly, under the Rayleigh fading channel (Figure 4d), ours outperformed across all SNR levels, with accuracy increasing from 0.6849 at SNR = 7 to 0.7099 at SNR = 23. In comparison, the AT method exhibited relatively low performance under low SNR conditions, with an accuracy of only 0.5654 at SNR = 7, improving to 0.5943 at SNR = 23, but remaining consistently lower than ours. The Trades method showed relatively stable performance across SNR conditions, but its maximum accuracy was only 0.6656 at SNR = 23, which was significantly inferior to ours.

These results highlight the notable advantage of the proposed method in robustness against adversarial sample interference and under complex channel conditions.

5 Conclusion

This paper introduces the RepObE framework for secure semantic communication, emphasizing defense against attackers. By encrypting data dynamically during semantic extraction and feature transmission, the framework deters data reconstruction via eavesdropping, boosting privacy protection. We improved system resilience in image communication tasks using a representation learning-enhanced prototype alignment collaborative adversarial training method. This integrated approach, with dynamic perturbation and robust optimization, provides reliable semantic communication in complex situations. Experimental results show our method surpasses existing ones, with over a 2% increase in resisting model inversion attacks and a 3% to 5% improvement in adversarial attack resistance in classification tasks. Visually, it excels, making it difficult for attackers to acquire useful images. Future work will explore the RepObE framework's application beyond image communication and additional techniques to enhance system robustness and efficiency against evolving adversarial threats.

Acknowledgments

This work is supported by National Natural Science Foundation of China under grants 62171132 and U1905211, and Natural Science Foundation of Fujian Province under grant 2024J09032.

References

[Chen *et al.*, 2024] X. Chen, J. Wang, L. Xu, J. Huang, and Z. Fei. A perceptually motivated approach for low-complexity speech semantic communication. *IEEE Internet of Things Journal*, 11(12):22054–22065, 2024.

[Choi *et al.*, 2019] K. Choi, K. Tatwawadi, T. Weissman, and S. Ermon. NECST: Neural joint source-channel coding. In *Proceedings of the International Conference on Machine Learning*, pages 1182–1192, 2019.

[Christopoulos *et al.*, 2000] C. Christopoulos, A. Skodras, and T. Ebrahimi. The JPEG2000 still image coding system: An overview. *IEEE Transactions on Consumer Electronics*, 46(4):1103–1127, 2000.

[Du *et al.*, 2024] B. Du, H. Du, H. Liu, D. Niyato, P. Xin, and J. Yu. YOLO-Based semantic communication with generative AI-aided resource allocation for digital twins construction. *IEEE Internet of Things Journal*, 11(5):7664–7678, 2024.

[Fu *et al.*, 2024] Y. Fu, W. Cheng, W. Zhang, and J. Wang. Scalable extraction based semantic communication for 6g wireless networks. *IEEE Communications Magazine*, 62(7):96–102, 2024.

[Guo *et al.*, 2024] J. Guo, H. Chen, B. Song, Y. Chi, C. Yuen, and F. R. Yu. Distributed task-oriented communication networks with multimodal semantic relay and edge intelligence. *IEEE Communications Magazine*, 62(6):82–89, 2024.

[Hu *et al.*, 2023] Q. Hu, G. Zhang, Z. Qin, Y. Cai, G. Yu, and G. Y. Li. Robust semantic communications with masked VQ-VAE enabled codebook. *IEEE Transactions on Wireless Communications*, 22(12):8707–8722, 2023.

[Jankowski *et al.*, 2020] M. Jankowski, D. Gündüz, and K. Mikolajczyk. Wireless image retrieval at the edge. *IEEE Journal on Selected Areas in Communications*, 39(1):89–100, 2020.

[Kang *et al.*, 2022] X. Kang, B. Song, J. Guo, Z. Qin, and F. R. Yu. Task-oriented image transmission for scene classification in unmanned aerial systems. *IEEE Transactions on Communications*, 70(8):5181–5192, 2022.

[Kang *et al.*, 2023] J. W. Kang, J. Y. He, H. Y. Du, Z. Xiong, Z. Yang, and X. Huang. Adversarial attacks and defenses for semantic communication in vehicular metaverses. *IEEE Wireless Communications*, 30(4):48–55, 2023.

[Li *et al.*, 2024] Y. Li, Z. Shi, H. Hu, Y. Fu, H. Wang, and H. Lei. Secure semantic communications: From perspective of physical layer security. *IEEE Communications Letters*, 28(10):2243–2247, 2024.

[Liu *et al.*, 2021] C. Liu, C. Guo, Y. Yang, F. Yan, and Q. Sun. Semantic communication methods for intelligent tasks in AI-powered internet of things. *Journal on Communications*, 42(11):97–108, 2021.

[Liu *et al.*, 2022] C. Liu, C. Guo, Y. Yang, J. Chen, and M. Zhu. Semantic communication for intelligent tasks: Theories, techniques, and challenges. *Journal on Communications*, 43(6):41–57, 2022.

[Liu *et al.*, 2024a] C. Liu, B. X. Chen, W. Shao, W. C. Zhang, K. K. L. Wong, and Y. Zhang. Unraveling attacks to machine-learning-based IoT systems: A survey and the open libraries behind them. *IEEE Internet of Things Journal*, 11(11):19232–19255, 2024.

[Liu *et al.*, 2024b] F. R. Liu, X. Xie, and Y. Yu. Scalable multi-party computation protocols for machine learning in the honest-majority setting. In *Proceedings of the 33rd USENIX Security Symposium*, pages 1939–1956, 2024.

[Lu *et al.*, 2024] Z. Lu, R. Li, K. Lu, X. Chen, E. Hossain, and Z. Zhao. Semantics-empowered communications: A tutorial-cum-survey. *IEEE Communications Surveys & Tutorials*, 26(1):41–79, 2024.

[Mehrasa *et al.*, 2019] N. Mehrasa, A. A. Jyothi, T. Durand, J. He, L. Sigal, and G. Mori. A variational auto-encoder model for stochastic point processes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3165–3174, 2019.

[Nan *et al.*, 2023] G. S. Nan, Z. C. Li, J. L. Zhai, Q. Cui, G. Chen, X. Du, X. Zhang, X. Tao, Z. Han, and T. Q. S. Quek. Physical-layer adversarial robustness for deep learning-based semantic communications. *IEEE Journal on Selected Areas in Communications*, 41(8):2592–2608, 2023.

[Pan *et al.*, 2023] Q. Pan, H. Tong, J. Lv, T. Luo, Z. Zhang, C. Yin, and J. Li. Image segmentation semantic communication over internet of vehicles. In *Proceedings of the IEEE Wireless Communications and Networking Conference*, pages 1–6, 2023.

[Patra *et al.*, 2021] A. Patra, T. Schneider, A. Suresh, and H. Yalame. ABY2.0: Improved mixed-protocol secure two-party computation. In *Proceedings of the 30th USENIX Security Symposium*, pages 2165–2182, 2021.

[Qiu *et al.*, 2023] G. Qiu, G. Tang, C. Li, L. Luo, D. Guo, and Y. Shen. Differentiated location privacy protection in mobile communication services: A survey from the semantic perception perspective. *ACM Computing Surveys*, 56(3):60:1–60:36, 2023.

[Sagduyu *et al.*, 2024] Y. E. Sagduyu, T. Erpek, A. Yener, and S. Ulukus. Joint sensing and semantic communications with multi-task deep learning. *IEEE Communications Magazine*, 62(9):74–81, 2024.

[Sang *et al.*, 2025] N. H. Sang, N. D. Hai, N. D. D. Anh, N. C. Luong, V. Nguyen, and S. Gong. Wireless power transfer meets semantic communication for resource-constrained IoT networks: A joint transmission mode selection and resource management approach. *IEEE Internet of Things Journal*, 12(1):556–568, 2025.

[Wang *et al.*, 2024a] 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*, 22(8):10150–10165, 2024.

[Wang *et al.*, 2024b] Y. Wang, W. Ni, W. Yi, X. Xu, P. Zhang, and A. Nallanathan. Federated contrastive learning for personalized semantic communication. *IEEE Communications Letters*, 28(8):1875–1879, 2024.

[Wang *et al.*, 2025] W. Wang, Z. Tian, C. Zhang, and S. Yu. SCU: An efficient machine unlearning scheme for deep learning enabled semantic communications. *IEEE Transactions on Information Forensics and Security*, 20:547–558, 2025.

[Wu *et al.*, 2022] Q. Wu, F. Liu, H. Xia, and T. Zhang. Semantic transfer between different tasks in the semantic communication system. In *Proceedings of the 2022 IEEE Wireless Communications and Networking Conference*, pages 566–571, 2022.

[Xie *et al.*, 2022] H. Xie, Z. Qin, X. Tao, and K. B. Letaief. Task-oriented multi-user semantic communications. *IEEE Journal on Selected Areas in Communications*, 40(9):2584–2597, 2022.

[Xu *et al.*, 2021] Q. L. Xu, G. H. Tao, S. Y. Cheng, and X. Zhang. Towards feature space adversarial attack by style perturbation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10523–10531, 2021.

[Xu *et al.*, 2023] W. Xu, Y. Zhang, F. Wang, Z. Qin, C. Liu, and P. Zhang. Semantic communication for the internet of vehicles: A multiuser cooperative approach. *IEEE Vehicular Technology Magazine*, 18(1):100–109, 2023.

[Xu *et al.*, 2025] X. Xu, C. He, X. Li, and J. Xu. Joint optimization trajectory and resource allocation for UAV-assisted semantic communications. *Physical Communication*, 68:102555, 2025.

[Yang *et al.*, 2023] K. Yang, S. Wang, J. Dai, K. Tan, K. Niu, and P. Zhang. WITT: A wireless image transmission transformer for semantic communications. In *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2023.

[Zhang *et al.*, 2019] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. El Ghaoui, and M. I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7472–7482, 2019.

[Zhang *et al.*, 2023] H. Zhang, S. Shao, M. Tao, X. Bi, and K. B. Letaief. Deep learning-enabled semantic communication systems with task-unaware transmitter and dynamic data. *IEEE Journal on Selected Areas in Communications*, 41(1):170–185, 2023.

[Zhang *et al.*, 2024] A. Zhang, Y. Wang, and S. Guo. On the utility-informativeness-security trade-off in discrete task-oriented semantic communication. *IEEE Communications Letters*, 28(6):1298–1302, 2024.

[Zhang *et al.*, 2025] C. Zhang, M. Hu, W. Li, and L. Wang. Adversarial attacks and defenses on text-to-image diffusion models: A survey. *Information Fusion*, 114:102701, 2025.