Trustworthy Semantic-Enabled 6G Communication: A Task-oriented and Privacy-preserving Perspective

Shuaishuai Guo, *Senior Member, IEEE*, Anbang Zhang, Yanhu Wang, Chenyuan Feng, *Member, IEEE*, and Tony Q. S. Quek, *Fellow, IEEE*,

***Abstract*—Trustworthy task-oriented semantic communication (ToSC) emerges as an innovative approach in the 6G landscape, characterized by the transmission of only vital information that is directly pertinent to a specific task. While ToSC offers an efficient mode of communication, it concurrently raises concerns regarding privacy, as sophisticated adversaries might possess the capability to reconstruct the original data from the transmitted features. This article provides an in-depth analysis of privacy-preserving strategies specifically designed for ToSC relying on deep neural network-based joint source and channel coding (DeepJSCC). The study encompasses a detailed comparative assessment of trustworthy feature perturbation methods such as differential privacy and encryption, alongside intrinsic security incorporation approaches like adversarial learning to train the JSCC and learning-based vector quantization (LBVQ). This comparative analysis underscores the integration of advanced explainable learning algorithms into communication systems, positing a new benchmark for privacy standards in the forthcoming 6G era.**

I. Introduction

The sixth generation (6G) communication represents the next frontier in communication technology, promising significant advancements over the current 5th generation (5G) networks. 6G networks are projected to achieve data speeds in the range of terabits per second, a significant leap from the gigabit speeds of 5G [1]. The integration of artificial intelligence (AI) in 6G networks is anticipated to be more profound, with AI algorithms playing a crucial role in network management and user-centric services. Additionally, 6G is set to revolutionize industries by enabling new business models and services. The network's reliability, trustworthiness and timeliness will be critical in the scenario of massive mobile users with real-time response requirements [2]. Despite the promising prospects, the practical implementation of 6G continues to encounter numerous unprecedented hurdles, particularly for burst communications [3]. ToSC, emerging as a promising paradigm, is primarily characterized by its selective transmission of information [4]. This approach has garnered considerable attention [5], chiefly due to its proficiency in enhancing efficiency and reducing latency through the minimization of data transmission volume [6]. Furthermore, ToSC exhibits the capacity to offer a more customized and efficient user experience. The targeted and efficient nature of ToSC, therefore, not only aligns with the technological advancements envisaged in 6G communications but also caters to the nuanced demands of modern digital applications, ensuring a seamless and user-oriented interaction.

S. Guo, A. Zhang, and Y. Wang are with Shandong University, China; C. Feng is with EURECOM, France; T. Q. S. Quek is with Singapore University of Technology and Design, Singapore

To alleviate the misunderstandings and incorrect interpretations, ToSC can increase information clarity, relevance, transparency, credibility, and verifiability by concentrating on task-relevant information, implementing well-designed mechanisms of channel coding and feedback. Although this strategy inherently provides a certain level of privacy since selective data transmission eliminates the unnecessary data sharing, the information conveyed may still be vulnerable [7]. If intercepted, even these task-relevant data bits could reveal personal or sensitive information [8]. This risk is heightened by the advanced capabilities in machine learning and data analysis, which might enable adversaries to extract significant insights from minimal data. This work discusses ongoing research aimed at developing privacy-preserving methods specifically for ToSC relying on deep neural network-based joint source and channel coding (DeepJSCC). The task at hand is devising techniques that can be seamlessly incorporated into semantic communication without hindering their efficacy and efficiency. This work meticulously evaluates and contrasts feature perturbation methodologies, such as differential privacy and encryption techniques, with intrinsic security incorporation approaches like adversarial learning to the JSCC and learning-based vector quantization (LBVQ). This comprehensive comparative analysis highlights the potential for integrating sophisticated learning algorithms into contemporary communication systems, thereby establishing a new paradigm for privacy standards in the forthcoming 6G era.

II. Shift From Task-Agnostic Communications to ToSC and Privacy Challenges

Within the traditional framework of source-channel separation, the identification, representation, and transmission of information are rigorously addressed by rate-distortion theory and channel coding theory respectively. This paradigm, which prioritizes reconstruction-oriented compression and task-agnostic communication, has underpinned several iterations of digital communication systems. However, the advent of machine-to-machine communications and human-machine interactions necessitates a reassessment of this paradigm, considering that exact reconstructions are often of secondary importance from a machine's perspective. Notably, task-specific descriptors, derived via machine learning algorithms from latent feature spaces, are substantially more concise than their counterparts used for reconstruction purposes. Furthermore,

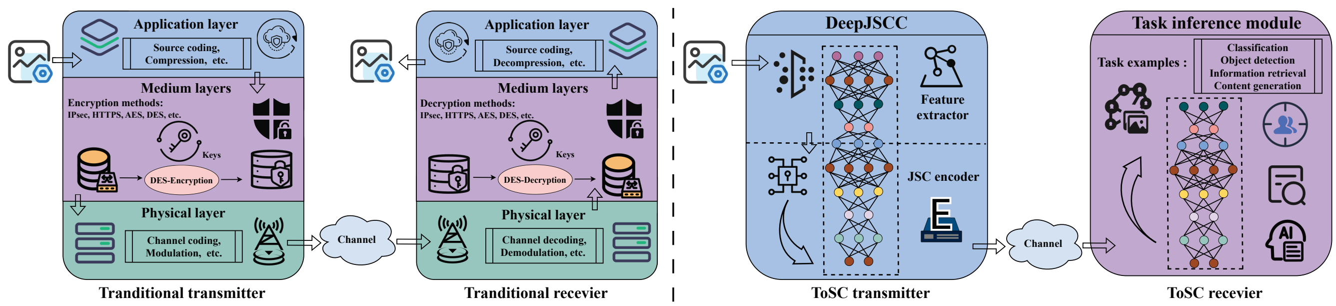


Fig. 1. Traditional transceivers versus ToSC transceivers.

communication systems trained end-to-end can significantly surpass those designed based on source-channel separation, across various performance metrics [4].

In ToSC systems, data transmission is meticulously tailored to align with the receiver's requirements. As delineated in Fig. 1, the ToSC transceivers differ from traditional transceivers. The ToSC system comprises two main components: the "DeepJSCC" and the "Task inference module". The DeepJSCC contains a JSCC encoder that plays the role of task-related information extraction, compression, and protection (against both channel noise and adverseries' attacks). The output of this passes through a channel and arrives at the task inference module. This module is outlined with examples of tasks such as classification, object detection, information retrieval, and content generation. Inside, it displays a multi-layered network topology that indicates intricate processing is occurring to perform the given tasks.

However, alongside these advantages, ToSC introduces significant privacy concerns. The main issue stems from the nature of the information being transmitted. Although ToSC systems transmit only task-specific information, this data can be sensitive and vulnerable to privacy infringements. Given the unpredictability of adversaries' objectives, it is imperative to devise a comprehensive and efficacious strategy for safeguarding a spectrum of private data. For instance, in the transmission of facial imagery, adversaries may undertake diverse strategies to extract personal attributes such as gender or skin tone, or alternatively, engage in facial recognition. In our research, we postulate the quality of image reconstruction—evaluated by metrics such as peak signal-to-noise ratio (PSNR)—as a quantifiable proxy for privacy preservation. Our rationale is predicated on the supposition that by amplifying the perturbation of data reconstructed by potential intruders, we indirectly shield various facets of personal information. This approach resonates with the principles of perturbative privacy preservation, notably exemplified by differential privacy paradigms. By escalating the degree of distortion, we can significantly diminish the likelihood or amplify the challenge for adversaries in gleaning sensitive information, thereby strengthening the robustness of privacy protection.

III. Privacy Preservation Methods for ToSC

Most ToSC systems are based on DeepJSCC architectures and employ an end-to-end training methodology to extract high-dimensional task-related channel-robust features for transmission. This approach ensures a coherent and automatically optimized process from data input to the final task output. Many traditional privacy protection techniques, such as k-anonymity, l-diversity, and t-closeness, are often designed to protect user privacy by modifying data and are not suitable for working with complex or high-dimensional data [8], [9], as the process of making records indistinguishable can lead to significant data loss or impracticality in datasets with numerous attributes. A thorough analysis of privacy-preserving strategies specifically designed for ToSC is provided, along with a detailed comparative assessment of feature perturbation methods and intrinsic security incorporation approaches.

A. Feature Perturbation Methods

Conventional approaches to AI inference privacy protection concentrate primarily on the safeguarding of original data, typically through direct preprocessing under the assumption that the entire AI model is in the possession of a potentially untrustworthy third party. However, in DeepJSCC-based ToSC, the DeepJSCC encoder (partial of the AI models) belongs to the transmitter, which plays the role of task-related information extraction, compression, and protection (against both channel noise and adverseries' attacks). The employment of DeepJSCC makes conventional privacy and security techniques applied in-between source and channel coding difficult in DeepJSCC. Applying these techniques in front of DeepJSCC will prevent the DeepJSCC to extract the task-related and channel-robust features, thus the only feasible is to process the output of DeepJSCC. Therefore, traditional data perturbations (e.g., noise addition, encryption) should be replaced by the feature perturbations on the DeepJSCC output. Next, we discuss two typical feature perturbation methods for privacy preservation.

- Differential privacy offers a mathematical framework for quantifying privacy loss. It provides strong privacy guarantees by adding noise to the DeepJSCC output. This approach is widely adopted due to its robustness and the theoretical guarantees. The addition of noise to protect individual data points can diminish the utility of data, especially in scenarios where precise data (i.e., semantic information) is used for task inference, as in ToSC. Too much noise can well protect privacy, however, it will inherently affect the task inference performance, that is the utility of the transmitted data.

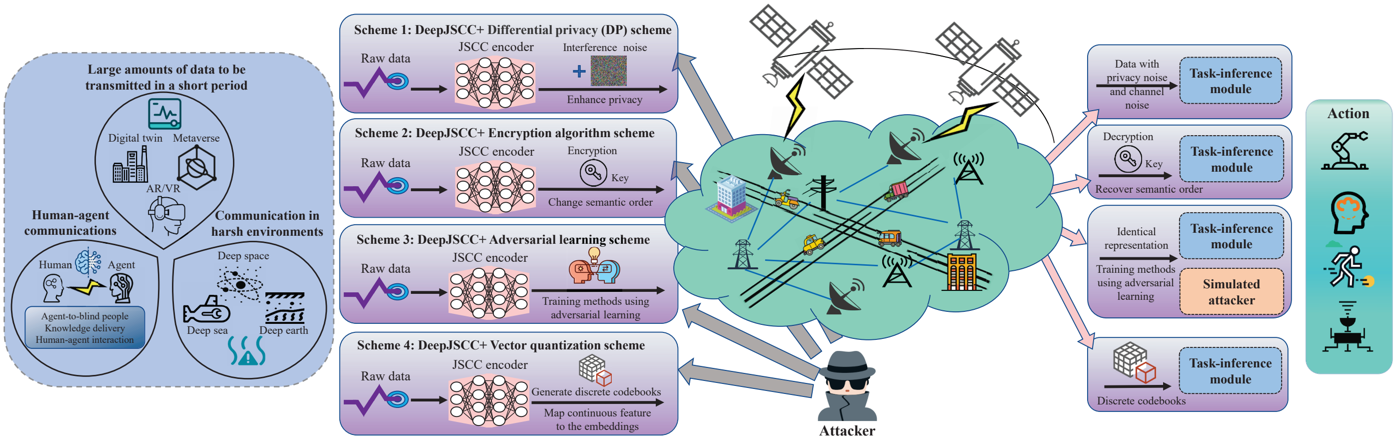


Fig. 2. Transceiver structure of four privacy protection methods: differential privacy, encryption, adversarial learning, and vector quantization.

- Encryption serves as a fundamental method for securing the output of DeepJSCC, by converting it into a format that is not readily intelligible or accessible without the appropriate decryption key. However, implementing encryption and decryption processes can significantly decelerate both the training and inference phases of ToSC. Additionally, in the absence of channel coding, encryption becomes vulnerable and may compromise the integrity of decryption. To mitigate these challenges, only methods such as simple encryption or error-robust encryption techniques are recommended for ToSC. An example of such a technique is the key-guided feature shuffling, as discussed in the recent literature [10]. Besides, encryption-based privacy-preserving methods are facing an additional key-sharing burden and the challenge of key leakage.

B. Intrinsic Security Incorporation Strategies

It is key to design and train DeepJSCC with intrinsic security. Adversarial learning, as a commonly-used technique, involves training a model in a competitive setting where two models, usually called the generator and the discriminator, are pitted against each other. The generator aims to produce data that is indistinguishable from real data, while the discriminator's goal is to accurately distinguish between the generator's fake data and real data. This process helps to improve the performance in both models. Adversarial learning plays a crucial role in privacy protecting of ToSC. By training models in an adversarial setting, it is possible to generate feature representations that are informative for intended tasks but are simultaneously difficult for adversarial models to exploit. This is particularly useful in scenarios where sensitive data needs to be protected from potential privacy breaches. The adversarial process ensures that while the essential characteristics of the data are preserved for the task at hand, the ability of an adversary to extract sensitive information from the data is minimized. This approach is instrumental in developing models that balance the need for utility and privacy.

It is noteworthy that the transmitter does not know the architectures of the adversary neural networks in the design. In practice, the adversary is not fixed, and it is impractical to obtain the adversary's network architecture. To defend against model inversion attacks, one can construct a simulated adversary and utilize adversarial learning to assist in ToSC training [11]. Even if the adversary's inversion network architecture is unknown, the approach still provides effective privacy protection via a simulated adversary.

Besides adversary learning, LBVQ uses discrete latent representations which are more robust against inversion. This means that it's harder to reconstruct the original input data accurately from the latent representations, thereby providing a layer of privacy. The quantization step in LBVQ, which converts continuous latent variables into discrete ones, acts as a bottleneck, reducing the amount of detailed information that can be decoded from the latent space. This inherent characteristic of LBVQ makes it a suitable choice for tasks where preserving privacy is crucial, as it inherently limits the detailed information that can be extracted from the encoded representations.

The integration of LBVQ within ToSC systems presents an additional benefit regarding compatibility with existing digital communication frameworks. In current ToSC models, features extracted by neural networks are typically continuous, aligning well with analog communications but not with prevalent digital communication systems. LBVQ, on the other hand, produces discrete representations that can be directly mapped to digital modulation symbols. This compatibility is instrumental in facilitating a seamless transition from well-established digital communication systems to the emerging ToSC paradigms, thereby bridging the gap between current technologies and future communication methodologies.

C. Transceiver Structure Comparison

Comparing feature perturbation methods and encryption with intrinsic security incorporation strategies reveals distinct strengths and weaknesses. Their differences in transceiver structures are demonstrated in Fig. 2. Feature perturbation methods like differential privacy and encryption offer robust theoretical guarantees for privacy. Differential privacy provides a quantifiable measure of privacy by adding noise to the

TABLE I
COMPLEXITY, COST AND LATENCY OF FOUR PRIVACY-PRESERVING MECHANISMS EVALUATED IN THE CIFAR-10 DATASET.

	DeepJSCC	DeepJSCC-DP	DeepJSCC-Encryption	IBAL	DeepJSCC-LBVQ
FLOPs	0.085 G	0.085 G	0.085 G	0.382 G	0.477 G
Params	3.19 M	3.19 M	3.19 M	10.91 M	12.61 M
Train Time for 1 Epoch	8.23 s	8.93 s	8.79 s	32.5 s	37.30 s
Test Time for 1 Instance	0.002 s	0.002 s	0.002 s	0.002 s	0.004 s

features. Encryption provides strong security for DeepJSCC output in transit but does not address the unique challenges of vulnerability to channel errors and real-time processing. On the other hand, adversarial learning offers an intrinsic security incorporation approach in ToSC. By introducing the adversary loss in the training phase of DeepJSCC, attacking-robust communication models can be obtained. Adversarial learning maintains strong performance even in the face of unknown or intentional attacks. However, the complexity and computational demands of adversarial models are notable challenges in the training phase. Additionally, continuous updates and model improvements may be needed to cope with new attack methods. LBVQ offers an effective approach for safeguarding feature privacy through the map of features into smaller vector spaces. This method significantly reduces feature dimensions while upholding the integrity of information. Nevertheless, the process of dimension reduction might result in potential information loss. And another challenge is the need for appropriate vector quantization based on the specific tasks and data types involved.

D. Cost and Delay Comparison

Tab. III-B shows the computational cost, complexity, and communication latency of four privacy schemes running on an 11th Gen Intel(R) processor at 2.50 GHz and a single 3060 CPU core, including the specific number of floating-point operations (FLOPs) in the whole training process, the number of parameters in all the neural network model, the time for a single training epoch (batch-size: 512) and the task-inference time of a single image input.

The mechanisms evaluated are DeepJSCC, DeepJSCC-DP, DeepJSCC-Encryption, IBAL, and DeepJSCC-LBVQ. The results indicate that DeepJSCC, DeepJSCC-DP, and DeepJSCC-Encryption demonstrate similar computational requirements and maintain a lower complexity profile, as evidenced by their FLOPs and parameter counts. In contrast, IBAL and DeepJSCC-LBVQ necessitate substantially greater resources, manifesting in increased FLOPs, a higher number of parameters, and extended training durations. Despite this, the inference times for all configurations, with the exception of DeepJSCC-LBVQ, remain comparably low. This suggests that all configurations, barring the LBVQ variant, achieve high efficiency during the inference phase. The LBVQ variant, however, incurs additional delays due to its discrete codebook mapping and remapping processes. Importantly, when compared to the baseline DeepJSCC model without privacy enhancements, the DeepJSCC-DP, DeepJSCC-Encryption, and IBAL models exhibit negligible increases in task-inference time, rendering them particularly suitable for applications requiring low-latency and privacy-sensitive remote inference.

IV. EXPERIMENTS AND DISCUSSIONS

Two experimental evaluations (i.e. image classification and face recognition) are provided to assess both the task performance and the degree of privacy preservation within end-to-end ToSC systems, along with the attacker settings to improve the overall privacy-preserving process.

A. Experimental Settings

1) Dataset and Attack Setups: CIFAR-10 dataset and CelebA dataset are adopted for image classification task and face recognition task, respectively. The former comprises $60,000$ color images, each measuring 32×32 pixels, and categorized into 10 distinct classes. The latter contains over 200,000 celebrity images with 40 attribute annotations per image, whose images cover a rich range of human pose variations and complex and diverse background information. To ensure comparability, all algorithmic networks under consideration are configured to produce outputs of identical dimensions. Furthermore, we incorporate a hypothetical scenario involving an adversarial attack network. This network is designed to execute model inversion via a black-box attack approach. It is posited to have continuous access to the network model on the target device, thereby enabling it to attempt image reconstruction. This scenario is pivotal in assessing the robustness of our framework against potential security breaches.

2) Performance Metrics: For the image classification task, classification accuracy and PSNR of the attacker are employed. The former serves as an indicator of inference performance, with higher classification accuracy signifying more effective inference capabilities. The latter, the PSNR value, is utilized to gauge the level of privacy protection. In this context, a lower PSNR value in the reconstructed images indicates that its attacker steals the transmitted data, having a worse distortion of the reconstructed image, which also reflects the stronger privacy preservation, as it indicates that the reconstructed image has a reduced clarity or fidelity, which prevents unauthorised interpretation. For the face recognition task, the top-1 accuracy and the reconstructed image of the attacker are used. The former is a metric of recognition performance and is used to judge the task performance. The latter, i.e., the direct reconstruction effect of the image, is used to measure the level of privacy protection.

3) Approaches for Evaluation: To comprehensively demonstrate the efficacy of diverse privacy-preserving approaches, we have meticulously selected four state-of-the-art ToSC schemes, including:

- **DeepJSCC-DP:** DeepJSCC, originally designed for data-oriented communication systems, utilizes deep neural network-based encoders to map data directly to channel input symbols. To facilitate a comparative analysis of both performance and security aspects, we propose the integration of differential privacy mechanism [12], specifically through the injection of Laplacian noise into the transmission characteristics. The differential privacy mechanism allows for precise control over the level of privacy by adjusting the privacy budget, which is set at 0.05, 0.1, and 0.9 for our experiments. A pivotal aspect to note is that a smaller privacy budget correlates with a higher volume of noise injected into the transmitted features. This increased noise level consequently leads to stronger privacy protection, as it more effectively obscures the original data features, thereby enhancing the security against potential data breaches or unauthorized data reconstruction efforts.
- **DeepJSCC-Encryption:** In DeepJSCC-Encryption, the encoder not only processes the data to extract features for the JSCC, but also integrates an encryption operation into its output [10]. At the encoder, encryption operations are performed on the encoded semantic features. This dual-functionality approach effectively combines feature extraction and encoding with a layer of cryptographic security. At the receiver, the process is reversed. The encoded and encrypted features are subjected to a decryption operation, a critical step for regaining the original data characteristics. Post-decryption, these features are then utilized for the intended classification and reconstruction tasks. This mechanism ensures that the data remains secure during transmission, only becoming accessible and usable upon successful decryption at the intended destination.
- **IBAL:** IBAL [11] represents a novel privacy-preserving scheme within the context of ToSC, leveraging the principles of adversarial learning. This method uniquely trains the encoder to effectively deceive the potential adversaries. It does so by optimizing the encoder to maximize the distortion in the data reconstruction process. Such a strategy is designed to thwart unauthorized attempts at data reconstruction, thereby enhancing the privacy and security of the transmitted information.
- **DeepJSCC-LBVQ:** DeepJSCC-LBVQ [13] represents a sophisticated ToSC scheme that incorporates digital modulation. Central to this approach is the implementation of a robust encoding framework, which is underpinned by a learned codebook. Its primary objective is to enhance communication robustness in response to channel variations. The essence of DeepJSCC-LBVQ lies in its ability to effectively balance the trade-off between informativeness and robustness. By employing a learned codebook, the scheme adapts to varying channel conditions, ensuring that the integrity and reliability of the transmitted data are maintained, even in challenging communication environments. This adaptability makes it a significant contender in scenarios where channel variability is a critical factor.

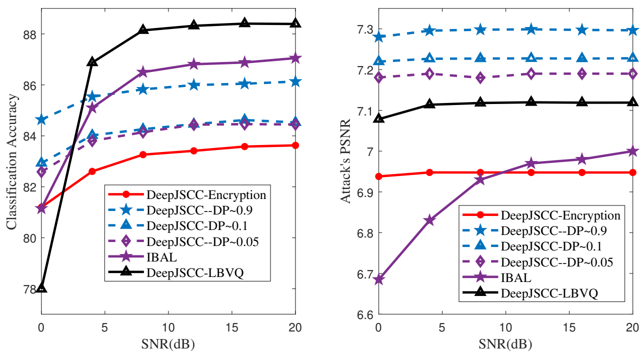
The inference performance and the quality of image reconstruction are critically influenced by the dimensionality of the encoded representation. To facilitate fair and equitable comparisons across the different methods, we have standardized the dimensionality of the representations encoded by all methods under consideration. For the continuous representation methods such as DeepJSCC-DP, DeepJSCC-Encryption, and IBAL, we employ a full-resolution constellation modulation technique. This approach is instrumental in maintaining the integrity and resolution of the encoded data during the modulation process. In contrast, for DeepJSCC-LBVQ, which is a discrete method, we utilize the M-ary quadrature amplitude modulation (QAM) scheme. This choice is tailored to suit the discrete nature of the representations encoded by DeepJSCC-LBVQ, ensuring that the modulation process is compatible with the encoding method. Additionally, to further ensure the impartiality of our evaluation, we have standardized the settings across all adversary attack networks. This uniformity in settings is vital for ensuring that each method is subjected to non-discriminatory attacks, thereby providing a more accurate and fair assessment of private level. By adopting this approach, we aim to offer a comprehensive and unbiased comparison of each method's ability to protect privacy under equivalent adversarial conditions.

B. *Information Leakage Under Adversarial Attack*

Regarding the adversarial's description, we assume ToSC system is under black-box model inversion attacks [14], where the adversaries reconstruct the received features as raw input using DNNs, and obtain users' privacy. Specifically, the adversary network is a DNN designed by the adversary and deployed on the adversary's device. And the transmitter's coded features are illegally accessed by the adversary. The adversary then attempts to generate an approximate reconstruction of the user's data based on the stolen transmission data. To improve the system's capability to combat attacks, we make a weak assumption that attacker knows the codebook and can continuously access the trained encoder. As the intentions of attacker are not known, we consider a universal loss function for training the attacker neural network, which is to minimize the distortion of reconstructed data. For image transmission task, the loss function for training attacker can be expressed as

$$\mathcal{L}_{Attack} = \mathcal{L}_{MSE} + \mathcal{L}_{PER}, \tag{1}$$

where $\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i\|^2$ represents the MSE with N denoting the number of samples, $\mathbf{x}_i$ representing the i-th sample, and $\tilde{\mathbf{x}}_i$ standing for the reconstructed data samples by the attacker; $\mathcal{L}_{PER} = \frac{1}{N}\sum_{i=1}^{N}\|VGN(\boldsymbol{x}_i) - VGN(\tilde{\boldsymbol{x}}_i)\|^2$ represents the perception loss with $VGN(\cdot)$ being the first three layers of the pretrained visual geometry group's (VGG) neural network [15].



(a) Classification accuracy. (b) Attacker's PSNR.

Fig. 3. Classification Performance versus SNR on the CIFAR-10 dataset.

C. Experimental Results – Image Classification

Each scheme undergoes training at a specific signal-to-noise ratio (SNR), denoted as $\mathrm{SNR}_{\mathrm{train}}$, which is set at 12 dB. The testing phase involves varying SNR levels, specifically at $\mathrm{SNR}_{\mathrm{test}}$ values of 0 dB, 4 dB, 8 dB, 12 dB, 16 dB, and 20 dB. To maintain consistency and fairness, both discrete and continuous, the dimension of the encoded representation is uniformly set to 128. This uniformity ensures that any observed differences in performance are attributable to the scheme's inherent characteristics rather than discrepancies in encoded representation size. Specifically, for DeepJSCC-LBVQ method, which employs a discrete approach, we use a codebook of size 16. This size is selected as it offers a balance between complexity and performance.

The analysis of Figs. 3(a) and 3(b) reveals the significant insights into the image classification performance of various privacy-preserving schemes under an additive white Gassian noise (AWGN) channel using the CIFAR-10 dataset. First, the classification accuracies of IBAL and DeepJSCC-LBVQ are significantly higher under regular channel conditions (i.e. SNR $\geq$ 4dB) compared to DeepJSCC-DP and DeepJSCC-Encryption. Moreover, these two schemes also exhibit superior privacy protection capabilities than DeepJSCC-DP method. Under low SNR regimes (i.e. SNR $<$ 4dB), DeepJSCC-DP and DeepJSCC-Encryption show better robustness in terms of classification accuracy. At this point, DeepJSCC-LBVQ classification accuracy decays, due to the wide deflection of discrete features caused by the codebook indexing of the channel transmission, but still maintains good privacy preserving ability. And IBAL shows the best privacy-preserving capability, as it focuses on trade-off between both privacy and task, with a slight bias at low SNR.

Notably, DeepJSCC-DP, with its increased Laplacian noise injection, offers improved privacy protection at the cost of task performance. For instance, DeepJSCC-DP with a privacy budget of 0.05 achieves similar privacy protection levels as DeepJSCC-LBVQ, but its classification accuracy falls behind by approximately 3 to 4 when SNR is greater than 4dB. DeepJSCC-Encryption presents the best privacy protection among the compared methods. However, this comes at the expense of task performance, failing to strike an optimal balance between privacy and utility. This contrast highlights the superiority of the intrinsic security incorporation schemes (IBAL and DeepJSCC-LBVQ) over feature perturbation methods in achieving a well-balanced effect in both task performance and privacy preservation. The advanced IBAL scheme particularly stands out for its capability to improve privacy protection without significantly compromising task inference performance, achieving an optimal privacy-utility trade-off compared to the baseline methods. A key factor contributing to DeepJSCC-LBVQ's superior task performance is that its discrete representation, enhanced by the learned codebook, contains more informative messages, thereby leading to better performance. This underscores the effectiveness of discrete representation encoding in ToSC systems.

TABLE II
THE TOP-1 ACCURACY OF FACE RECOGNITION IN THE CELEBA DATASET.

	Mustache	Smiling	Wavy Hair
DP-0.05	73.3%	66.8%	69.6%
DP-0.1	75.5%	66.7%	70.7%
DP-0.9	76.5%	68.3%	70.8%
Encryption	77.8%	65.7%	72.6%
IBAL	76.7%	69.9 %	71.7%
LBVQ	85.0%	80.6%	79.4%

D. Experimental Results – Face Recognition

Regarding the face recognition task, we consider a few single attributes as the target of the retrieval, e.g., smiling, moustache, wavy hair. Furthermore, the network requires significantly less information to predict a single characteristic than it does for 40 attributes recognition, which should make it simpler to create privacy-protected features. As shown in Tab. IV-C, DeepJSC-LBVQ scheme achieves better task performance when performing face recognition with arbitrarily selected different attributes. This scheme, despite the fact that it undergoes dimensionality reduction and may lead to potential loss of information, adequately extracts the features needed for the face, maintains the integrity of the information and therefore ensures higher task performance. IBAL, on the other hand, a pre-trained model with adversarial learning, is also better able to extract the required features for the face recognition task. The performance of face recognition under all three attributes is better than the DeepJSCC-DP scheme, but under some attributes (e.g., mustache, wavy hair) the performance is similar to the DeepJSCC-Encryption scheme. The reason is that IBAL considers the overall trade-off between performance and privacy, and outperforms both schemes in terms of privacy preservation, shown in Fig. 4. Longitudinally, the common DeepJSCC-DP and DeepJSCC-Encryption schemes have a greater loss (about 5%-15% drop) in performance compared to the above DeepJSC-LBVQ scheme, i.e. on the mustache, on the smiling, and on the wavy hair attributes.

Next, we investigate the images of the Celeba dataset obtained after reconstructing the transmitted signals of these

Fig. 4. CelebA: from left to right, original, DeepJSCC-DP-0.9, DeepJSCC-DP-0.1, DeepJSCC-DP-0.05, DeepJSCC-Encryption, IBAL, DeepJSCC-LBVQ

schemes using model inversion attacks. The results demonstrated in Fig. 4 can be seen that the attacker reconstructs the image worse with IBAL and DeepJSCC-LBVQ. IBAL even affects the attacker's judgement on the gender reconstruction of men and women, which brings great protection. Combined with the above analysis of performance it can be seen that IBAL and DeepJSCC-LBVQ improve privacy protection while ensuring better task inference performance. And as Laplacian noise increases, DeepJSCC-DP scheme improve their privacy protection, i.e., the attacker reconstructs a blurrier image. In addition, the DeepJSCC-encryption scheme also provides good privacy protection because it completely disrupts the order of features in the image transmission, while the attacker only steals the image data and cannot carry out the complete reconstruction process later. In a comprehensive analysis, IABL and DeepJSCC-LBVQ schemes achieve a better privacy-utility trade-off compared to the baseline scheme.

V. Future Research Directions

Regarding securing privacy in ToSC, several research directions merit further exploration.

A. Balancing Utility, Efficiency, and Privacy

In ToSC systems, achieving a harmonious balance between utility, efficiency, and privacy is a multifaceted challenge. Utility is paramount, as the primary goal is to transmit data that is semantically relevant to the task at hand. However, ensuring utility often requires processing and transmitting detailed information, which can inadvertently expose sensitive data, thereby conflicting with privacy goals. Efficiency is crucial in the high-volume, high-speed environment of 6G networks. Efficient data transmission not only conserves bandwidth but also reduces latency, enhancing the overall performance of the network. Privacy is perhaps the most challenging aspect to balance. While feature perturbation methods like encryption provide robust security, they do not address the unique challenges of real-time, semantically-rich communication. Advanced strategies, although more adaptable, come with their own set of risks, such as the inexplicability and uncertainty of black-box deep learning. To achieve this balance, a multi-layered approach [7] is often required. This involves combining feature perturbation privacy methods with intrinsic security incorporation techniques, continuously adapting to the changing dynamics of the network and data. Regular audits and updates to the privacy protocols are also essential to respond to new threats and technologies.

B. Exploring Generative AI for Privacy Preserving

Generative AI presents a novel approach to preserving privacy in ToSC. It focuses on creating data that is semantically similar to, but distinct from, the original dataset, thereby enabling the use of valuable data without exposing sensitive information. In ToSC, where the goal is to transmit semantically relevant information, generative AI can be used to produce high-quality synthetic data that maintains the statistical properties of the original dataset. This ensures that the utility of the data is not compromised, which is crucial for the effective functioning of ToSC systems. Additionally, since the synthetic data does not directly correspond to real user data, the risk of privacy breaches is significantly reduced. However, the use of generative AI in privacy preservation also poses challenges. One key issue is ensuring that the synthetic data does not retain any indirect identifiers that could lead to privacy breaches. This requires careful design and continuous evaluation of the generative models. Moreover, the computational complexity of training generative models can be a limiting factor.

C. Transfer Learning for Task, Data and Channel Adaption

In learning-based privacy-preserving ToSC, the system needs to be retrained as either the task, data or channel varies. Transfer learning can be instrumental for task/data/channel adaptation, allowing communication systems to efficiently adapt to new domains using pre-existing knowledge. Task adaptation focuses on applying learned models to new but related tasks. In ToSC, this could involve using a model trained for one semantic communication task (e.g., image recognition) and adapting it for another (e.g., image retrieval). Transfer learning enables ToSC systems to quickly adjust to new tasks without the need for extensive retraining, thereby saving time and computational resources. Data adaptation is essential due to the variability of data types and sources. Transfer learning allows for the utilization of pre-trained models on one type of data (like text) and adapts them for different types (such as images or sensors data). This flexibility is particularly beneficial in multi-modal communication scenarios where different types of data need to be processed and transmitted seamlessly. Transfer learning can be also employed to adapt ToSC systems to varying channel conditions such as interference, signal attenuation, and mobility. By learning from data transmitted under different channel conditions, a ToSC system can predict and adjust its parameters for optimal performance, even in less-than-ideal transmission environments.

D. Integrating Physical Layer Security

In ToSC, integrating physical layer security is paramount for preserving privacy by safeguarding the confidentiality, integrity, and availability of transmitted semantic information, especially when conventional security measures at higher layers are ineffective. Physical layer security strengthens communication systems by providing an additional defense layer against eavesdropping and jamming attacks. The resilience of the communication channel against compromise by adversaries could be enhanced by leveraging inherent properties of the communication medium, such as signal attenuation and channel randomness. However, there are still multifaceted challenges. Implementation often introduces overhead in bandwidth utilization, and energy consumption, necessitating careful consideration of trade-offs between security and system performance. Moreover, ToSC systems operate in dynamic and unpredictable environments, characterized by rapid fluctuations in channel conditions. Consequently, it is imperative to devise mechanisms resilient to adversarial threats and adaptable to diverse operational conditions, including variations in channel quality and eavesdropper locations.

VI. Conclusion

This article highlights the significance of transmitting task-specific essential information efficiently while addressing the critical issue of privacy preservation. It includes a comprehensive analysis of privacy-preserving strategies for ToSC, comparing feature perturbation methods like differential privacy and encryption with intrinsic security incorporation approaches such as adversarial learning and LBVQ. The research also explores experimental evaluations of these methods, assessing their performance and privacy protection capabilities. Finally, potential avenues for future study on privacy security and information trustworthiness are prospected.

References

[1] W. Saad, M. Bennis, and M. Chen, “A vision of 6G wireless systems: Applications, trends, technologies, and open research problems,” *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, May 2020.

[2] C. Feng, H. H. Yang, D. Hu, Z. Zhao, T. Q. S. Quek, and G. Min, “Mobility-aware cluster federated learning in hierarchical wireless networks,” *IEEE Transactions on Wireless Communications*, vol. 21, no. 10, pp. 8441–8458, Oct. 2022.

[3] Q. Hu, G. Zhang, Z. Qin, Y. Cai, G. Yu, and G. Y. Li, “Robust semantic communications with masked vq-vae enabled codebook,” *IEEE Trans. Wireless Commun.*, vol. 22, no. 12, pp. 8707–8722, Apr. 2023.

[4] D. Gündüz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, “Beyond transmitting bits: Context, semantics, and task-oriented communications,” *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 5–41, Jan. 2023.

[5] H. Xie, Z. Qin, X. Tao, and K. B. Letaief, “Task-oriented multi-user semantic communications,” *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2584–2597, Sept. 2022.

[6] J. Shao, Y. Mao, and J. Zhang, “Learning task-oriented communication for edge inference: An information bottleneck approach,” *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 197–211, Jan. 2022.

[7] A. Zhang, Y. Wang, and S. Guo, “On the utility-informativeness-security trade-off in discrete task-oriented semantic communication,” *IEEE Commun. Lett.*, vol. 28, no. 6, pp. 1298–1302, Mar. 2024.

[8] T. Xiao, Y.-H. Tsai, K. Sohn, M. Chandraker, and M.-H. Yang, “Adversarial learning of privacy-preserving and task-oriented representations,” in *Proc. AAAI Conf. Artif. Intell.*, New York, USA, Feb. 2020, pp. 12 434–12 441.

[9] A. Li, Y. Duan, H. Yang, Y. Chen, and J. Yang, “TIPRDC: Task-independent privacy-respecting data crowdsourcing framework for deep learning with anonymized intermediate representations,” in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD)*, Virtual, Online, USA, Aug. 2020, pp. 824–832.

[10] X. Luo, Z. Chen, M. Tao, and F. Yang, “Encrypted semantic communication using adversarial training for privacy preserving,” *IEEE Commun. Lett.*, vol. 27, no. 6, pp. 1486–1490, June 2023.

[11] Y. Wang, S. Guo, Y. Deng, H. Zhang, and Y. Fang, “Privacy-preserving task-oriented semantic communications against model inversion attacks,” *IEEE Trans. Wireless Commun.*, pp. 1–1, Mar. 2024.

[12] A. Xiong, T. Wang, N. Li, and S. Jha, “Towards effective differential privacy communication for users’ data sharing decision and comprehension,” in *Proc. IEEE S&P*, San Francisco, USA, July 2020, pp. 392–410.

[13] S. Xie, S. Ma, M. Ding, Y. Shi, M. Tang, and Y. Wu, “Robust information bottleneck for task-oriented communication with digital modulation,” *IEEE J. Sel. Areas Commun.*, vol. 41, no. 8, pp. 2577–2591, Aug. 2023.

[14] M. Fredrikson, S. Jha, and T. Ristenpart, “Model inversion attacks that exploit confidence information and basic countermeasures,” in *Proc. ACM. Conf. Computer. Commun. Secur. (ACM), Denver, CO, USA*, Oct. 2015, pp. 1322–1333.

[15] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. Int. Conf. Learn. Represent. (ICLR), San Diego, CA, USA*, May. 2015.