Multi-Task Semantic Communication: A Mutual Information-Aided Semi-Supervised Approach

Yining Wang, Wenqiang Yi, *Member, IEEE,* Shujun Han, *Member, IEEE,* Xiaodong Xu, *Senior Member, IEEE,* Ping Zhang, *Fellow, IEEE* and Arumugam Nallanathan, *Fellow, IEEE*

***Abstract*—In this paper, we design an end-to-end digital semantic communication system to transmit semantic symbols that simultaneously facilitate image classification tasks and reconstruction tasks. By training a mutual information-assisted joint source-channel coding (MIJSCC) framework, the learned semantic representation can incorporate both pixel-level generative information for reconstruction and structural discriminative information for classification, which are obtained label-free via global and local mutual information estimation and maximization, as well as mean square error (MSE) minimization. Then, the high-resolution semantic representation is quantized into finite constellation symbols to satisfy the hardware constraint on discrete control in practical radio frequency systems. Considering dynamic channel conditions in practical communication systems, we further design an adaptive MIJSCC framework with attention-based semantic enhancement (A-MIJSCC), which allows for the sequential activation of varying dimensions of the semantic representation according to channel signal-to-noise ratio. Compared to existing semantic communication frameworks that are dominated by end target and labels, the MIJSCC addresses the semi-supervised learning of intermediate semantics. Simulation results show that the proposed MIJSCC supports both image classification and reconstruction via task-agnostic semantic extraction, whose performance surpasses the benchmark frameworks. It is also demonstrated that the A-MIJSCC method facilitates the adaptive semantic transmission under varying channel conditions, which effectively reduces the transmission overhead while preserving task performance.**

***Index Terms*—Semantic communication, mutual information, task-oriented communications, joint source-channel coding (JSCC).**

I. Introduction

ARTIFICIAL intelligence (AI) and machine learning (ML) have demonstrated great potential in transforming wireless communications, significantly improving system performance and enabling semantic communication systems to optimize data transmission [1]. The deep learning (DL) enabled semantic communication [2] framework proposed in recent years has provided the future communication systems with the ability of knowledge perception and task comprehension [3], which can achieve prompt system reactions and dependable, efficient information exchange in Internet of Things (IoT) scenarios [4].

However, the preliminary semantic communication systems are independently designed according to a single target including reconstruction [5]–[7] and task execution [8]–[12], which fail to support the IoT scenarios with multiple tasks occurring simultaneously based on the same received content. Specifically, the transmitter extracts semantics based on the receiver's task objectives. When these objectives change, the semantic encoder should be retrained to match the new task, making it inflexible and difficult to adapt to the diverse and evolving requirements of the receiver.

Moreover, the existing task-driven semantic communication frameworks are fully-supervised, while the acquisition of massive data labels is expensive and challenging [13], indicating a limited scalability in practical applications. In this context, introducing semi-supervised learning helps alleviate the dependence on labeled data and improves the flexibility of the system in adapting to diverse downstream tasks using shared semantic representations. However, achieving this requires overcoming significant interdisciplinary challenges, making it essential to integrate semantic communication with the sixth-generation technologies by enabling more efficient, context-aware communication systems [14].

Fortunately, mutual information neural estimation (MINE) [15] has been proposed to explore semi-supervised or unsupervised learning of informative representations via neural networks, which overcomes the difficulty of calculating mutual information (MI) for high-dimensional features [16]. By incorporating MI regularization into multi-task semantic framework design, semantically informative representations can be extracted proactively, rather than relying solely on end-to-end loss-driven optimization such as label-matching or pixel recovery, thereby facilitating both generative and discriminative tasks.

A. Related Work and Motivations

Some novel work on semantic communications surpasses the restrictions of conventional frameworks with single end-to-end target. The authors in [17] designed a unified multi-modal semantic communication system to serve different tasks by activating different neural network layers via a multi-exit architecture. Tian *et al.* [18] proposed an asynchronous multi-task semantic communication framework with contrastive-

Yining Wang, Shujun Han, Xiaodong Xu and Ping Zhang are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: joanna_wyn@bupt.edu.cn; hanshujun@bupt.edu.cn; xuxiaodong@bupt.edu.cn; pzhang@bupt.edu.cn).

Wenqiang Yi is with the School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, U.K. (e-mail: w.yi@essex.ac.uk).

Arumugam Nallanathan is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K., and also with the Department of Electronic Engineering, Kyung Hee University, Yongin-si, Gyeonggi-do 17104, Korea (e-mail: a.nallanathan@qmul.ac.uk).

based encoder and task-related decoders, which can accomplish multiple tasks in a single transmission. In [19], a multi-task deep JSCC framework for image recovery and classification was derived based on coding rate reduction maximization, which can directly perform classification in the discriminate feature space and achieve data recovery simultaneously. [20] proposes a generative semantic communication system that supports both image reconstruction and segmentation by employing three Swin Transformers as the source semantic knowledge base at the transmitter to extract the multi-level features from the original image. At the receiver side, task-specific knowledge is generated based on hierarchical residual blocks. The authors in [21] proposed a data adaptation method for semantic communications with task-unaware transmitter, where dynamic data are transformed into a comparable form with the empirical data, thereby supporting arbitrary tasks predefined at the receiver. These studies promoted the practicality of semantic communication systems to facilitate multi-functional JSCC frameworks with the assistance of cutting-edge learning methods including contrastive learning [22], domain adaptation [23], coding rate reduction [24], and Transformer structures [25], which steer the field away from merely optimizing with regard to end-to-end targets but toward the learning of intermediate representations for universal tasks.

To achieve useful intermediate representations, recent studies analyzed from the view of information theory and attempted to improve the representation's suitability for downstream tasks. The authors in [26] employ an information-theoretic approach, infomax, where the end-to-end semantic coding procedure can exploit the statistical relations between different semantic interpretations from a single observation of the cooperative multi-task processing. Xie *et al.* [27] examined the intrinsic trade-off between the informativeness and the resilience to information distortion by training the semantic representations based on information bottleneck. The authors in [28] developed an extended rate-distortion problem for compact semantics extraction enabling multiple tasks with performance-transmission trade-off. In [29], the authors analyzed the relationship between the semantic signal length and the channel noise and proposed a packing sphere theory-based method to dynamically map the semantic signal into latent semantic codewords without noise overlap. The above mentioned work explored the compression and informative abilities of semantic representations from the perspective of information theory, which aims to preserve important knowledge as semantics instead of indiscriminately compressing the source input via neural network closed box.

From the above-mentioned works on multi-functional semantic communication framework design and information theory-based semantic representation learning, it can be inferred that the mutual information-based learning method can effectively reduce the reliance on task types and labels, thereby enhancing the adaptability and flexibility of the semantic communication framework. Inspired by this, we aim to design a multi-task semantic communication system, where informative semantic representations are obtained by unsupervised training and then transmitted over wireless channel to simultaneously facilitate various types of tasks at the receiver. The semantic representations should be competent enough to preserve the important information for all tasks, which can be achieved by evaluating and maximizing the MI between semantic representations and the source input. Moreover, the semantic representations should be transformed into feasible symbols before transmission, where further compression should be considered according to various channel conditions.

B. Contributions

In this paper, we consider a semi-supervised multi-task semantic communication system enabled by a mutual information-assisted joint source-channel coding (MIJSCC) framework, where both image classification and reconstruction tasks are conducted using the same semantic representation transmitted over the wireless channel. The main contributions are summarized as follows:

- We design a novel MIJSCC framework enabling the learning of both semantic representations and the end-to-end target to facilitate multi-task semantic communications, which aims to complete image classification and reconstruction tasks with the same received semantic representation. With global and local MI maximization enabled by Jensen-Shannon (JS)-based adversarial training, the semantic representation can not only learn the pixel-level information for reconstruction but also the discriminative feature for classification. Furthermore, with the assistance of mean square error (MSE) minimization, the MIJSCC encoder and decoder can be jointly trained to accomplish reconstruction-style objectives, thereby supporting image recovery at the receiver side.
- To implement the proposed MIJSCC framework in practical communication systems with limited RF capabilities, we integrate a standard asymmetric quantizer, which adapts the learned semantic representation for practical digital transmission. Specifically, we quantize the 32-bit float-number semantic representation into integers with fewer bits. Then the reshaped symbols can be mapped to discrete constellation points exhibiting larger point-distances, identifiable amplitudes and phases, thus can be seamlessly applied to existing communication systems.
- To further reduce the semantic transmission overhead, we design an adaptive MIJSCC framework with attention-based semantic enhancement (A-MIJSCC), which consecutively deactivates different numbers of dimensions in the semantic representation according to changing channel conditions. Moreover, to compensate for the potential performance loss caused by deactivation, we introduce a semantic enhancement module to reinforce the important dimensions in the masked semantic representations, thereby maintaining the task performance while reducing transmission overhead.
- To confirm the viability and superiority of our proposed MIJSCC framework, we conduct comprehensive experiments on CIFAR10 dataset. Simulation results demonstrate that compared with uniquely designed single-task frameworks and the conventional separate source-channel coding (SSCC) scheme, our proposed MIJSCC framework can leverage the received semantic representation

to accomplish multiple tasks simultaneously without suffering from the cliff effect. Furthermore, the effectiveness of the A-MIJSCC method is proved, which addresses its implementation under dynamic and resource-restricted environments.

The rest of this paper is organized as follows. Section II introduces the structure of the semi-supervised multi-task semantic communication system with semantic quantization. Section III describes the pipeline of the proposed MIJSCC framework and derives the principle of MI maximization. Section IV explains the implementation of the A-MIJSCC method for adaptive semantic transmission under changing channel conditions. Section V provides the numerical results. Finally, Section VI concludes this work.

II. Simultaneous Multi-Task Semantic Communication System

As shown in Fig. 1, this work considers end-to-end semantic communication at a wireless edge, where the blocklength of transmitted symbols is finite. Since the source-channel separation theorem is not applicable in this scenario, JSCC is utilized at both ends to explore the full potential of this communication [30].

A. Semantic Communication Framework

Two major tasks for semantic communications are classification and reconstruction, which are conventionally trained separately. As these two tasks share a similar latent space to describe different levels of visual information for the same object, we proposed a unified framework to train them simultaneously.

1) Transmitter Model: The source information is an image $\boldsymbol{s} \in \mathbb{R}^{L\times H\times C}$, where L, H, C denote the length, height, and color dimensions, respectively. With the aid of the JSCC encoder, the $\boldsymbol{s}$ is firstly convert to a high-level semantic representation $\boldsymbol{z}$, which is given by

$$\boldsymbol{z} = E_{\boldsymbol{\alpha}}(\boldsymbol{s}) \in \mathbb{R}^{2n}, \tag{1}$$

where $2n$ is the size of the semantic representation vector and $E_{\boldsymbol{\alpha}}(\cdot)$ represents the JSCC encoder model with parameters $\boldsymbol{\alpha}$. Then, the semantic representation $\boldsymbol{z}$ is reshaped into n complex value symbols as signals for modulating on the high-frequency carriers[1]. We normalize the power of the transmitted signal $\boldsymbol{x} \in \mathbb{C}^n$ as follows

$$\frac{1}{n}\mathbb{E}(\|\boldsymbol{x}\|^2) \leq 1. \tag{2}$$

Unlike pure source information compression, it is worth noting that the semantic representation contains not only the abstraction of the original image but also the compensation strategies for combating the randomness of wireless channels. The detailed reason is discussed in Section III.

2) Channel Model: This part presents the considered channel model with the necessary assumptions.

Assumption 1 (Noise-limited Channels): We assume each receiver has orthogonal time-frequency resource block, so the mutual-interference is ignored in this work.

Under Assumption 1, the received signal $\boldsymbol{y}$ can be expressed as follows

$$\boldsymbol{y} = h\boldsymbol{x} + \boldsymbol{n}, \tag{3}$$

where h denotes the channel coefficient and $\boldsymbol{n}$ represents the independent identically distributed (IID) channel noise vector, which follows the symmetric complex Gaussian distribution $\mathcal{CN}(0, \delta^2)$ with zero mean and variance δ^2.

As this work considers a single-antenna scenario, the existing channel estimation methods, e.g., the least squares (LS) estimation, are capable of providing sufficient estimation accuracy. Therefore, we performs channel equalization by multiplying $1/h$ on both sides of (3) to obtain the signal used for decoding as

$$\hat{\boldsymbol{y}} = \frac{\boldsymbol{y}}{h} = \boldsymbol{x} + \frac{\boldsymbol{n}}{h}, \tag{4}$$

and the signal-to-noise ratio (SNR) is (h/δ^2). Note that due to the existing mature channel estimation techniques, the performance in this work can be straightforwardly extended to different channel fading models, e.g., Rayleigh and Rician channels.

Assumption 2 (Continuous Constellation Symbols): Since the parameters of semantic representation are all decimals, we assume that the amplitude and phase of the modulated signal can vary continuously. In other words, the positions of the constellation symbols can be the entire constellation space.

Assumption 2 is important for semantic communication. In conventional communication systems, the spacing between constellation symbols is designed to counteract the noise effect on the received message. Consequently, employing a fixed modulation pattern with discrete constellation positions enhances robustness and reduces design complexity. However, in DL-enabled semantic communication, this spacing doesn't need to be excessively large to maintain the distinction among symbols. DL has the capability to autonomously determine the density of constellation symbols through training.

3) Receiver Model: Based on the applied reshaping scheme, the noised signal $\hat{\boldsymbol{y}}$ is recovered into the noised semantic representation vector $\hat{\boldsymbol{z}} \in \mathbb{R}^{2n}$, which is the input of the JSCC decoder for both the image classification and reconstruction tasks.

For the image reconstruction task, this work proposes an unsupervised JSCC decoder with the aid of MI, named MIJSCC decoder. The noised semantic representation $\hat{\boldsymbol{z}}$ is transmitted into the MIJSCC decoder $D_{\boldsymbol{\beta}}(\cdot)$ to generate the semantic reconstruction $\hat{\boldsymbol{s}} \in \mathbb{R}^{L\times H\times C}$ as

$$\hat{\boldsymbol{s}} = D_{\boldsymbol{\beta}}(\hat{\boldsymbol{z}}), \tag{5}$$

where $\boldsymbol{\beta}$ denotes the model parameters of the MIJSCC decoder. Note that the MIJSCC decoder model is jointly trained in an end-to-end manner with the JSCC encoder. The performance of image reconstruction tasks is evaluated by the

[1] The reshaping scheme can be any patterns, which will not affect the performance.

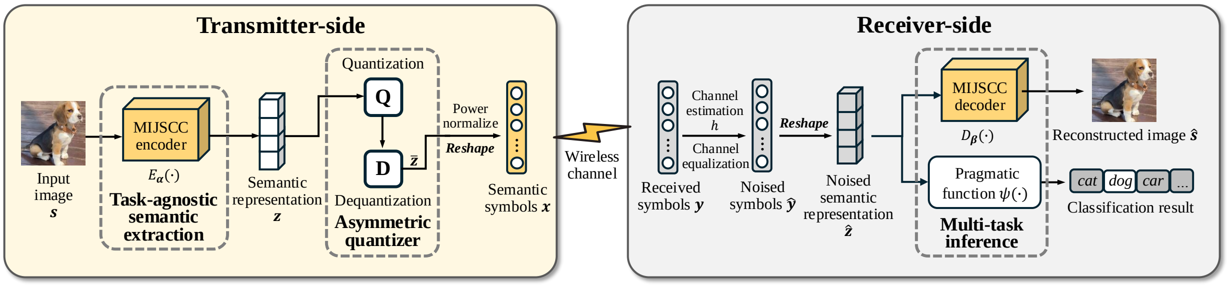


Fig. 1. The proposed simultaneous multi-task semantic communication system. The input image is encoded into semantic representation $\boldsymbol{z}$, which is quantized and dequantized via the asymmetric quantizer. Then, the dequantized semantic representation vector is reshaped and mapped into finite constellation points, which is transmitted over the wireless channel as semantic symbols $\boldsymbol{x}$. At the receiver side, channel equalization is performed, where the received symbols $\boldsymbol{y}$ is divided by the channel gain h from perfect channel estimation and the obtained noised symbols $\hat{\boldsymbol{y}}$ is reshaped into noised semantic representation vector $\hat{\boldsymbol{z}}$, which serves as the input of both image reconstruction and classification tasks.

peak signal-to-noise ratio (PSNR), which measures the visual difference between two images:

$$\begin{aligned}\mathrm{PSNR}(\boldsymbol{s},\hat{\boldsymbol{s}}) &= 10\log_{10}\frac{\mathrm{MAX}^2}{\mathrm{MSE}(\boldsymbol{s},\hat{\boldsymbol{s}})}\\ &= 10\log_{10}\frac{\mathrm{MAX}^2}{\frac{1}{LHC}\|\boldsymbol{s}-\hat{\boldsymbol{s}}\|^2},\end{aligned} \tag{6}$$

where MAX is the maximum value of the image pixels and $\mathrm{MSE}(.,.)$ denotes the function of calculating the mean squared error.

For the image classification task, the noised semantic representation $\hat{\boldsymbol{z}}$ is fed into a classifier, which is a given pragmatic function $\psi(\cdot)$ pre-designed according to the classification task [31] on the receiver-side. The classifier can be designed as a deep neural network (DNN) model, a K-Nearest Neighbor classifier or a support vector machine. The classifier is **not** jointly trained with the MIJSCC encoder since the extracted semantic representation is discriminative, which can be used for classification tasks directly. We provide additional discussion about this feature in Sections III-B and III-C. The accuracy of image classification tasks ACC is measured by the proportion of correctly classified image samples $N_{correct}$ to all test samples N_{test}:

$$\mathrm{ACC} = \frac{N_{correct}}{N_{test}} \times 100\%. \tag{7}$$

In conventional single-task semantic communication frameworks, the above two tasks can not be completed simultaneously. This is because the task-specific JSCC models are trained in an end-to-end manner with loss functions tailored to each task, which aim to bring the output closer to the task-intended value by adjusting model parameters. Since the objectives of different tasks require distinct optimization paths, it becomes challenging to achieve multiple goals using a combined target. Additionally, in the traditional end-to-end training paradigm, only the output layer is actively optimized based on the target, while the semantic representation derived from the intermediate layer is passively learned without supervision. As a result, the transmitted semantic representation is merely the intermediate outcome specific to a single task, which is not discriminative or interpretable.

To this end, the proposed MIJSCC framework aims to train both the semantic representation and the generated output, thereby facilitating the discriminative information extraction while preserving the generative ability of the MIJSCC decoder.

B. Quantization Transmission Model

To implement the proposed framework in a practical wireless system, we need a quantization transmission model that is compatible with the aforementioned MIJSCC. As stated in Assumption 2, the JSCC model jointly learns the constellation mapping of symbols from the source image and channel characteristics, while the obtained constellation points, such as clustered constellation [32], are disorganized in a large range due to the fine-grained 32-bit float-point symbols derived from the neural network. This nearly-continuous constellation design requires the ability of resolving high-resolution amplitude shifts and phase shifts, which violates the hardware constraints in current radio frequency systems. Therefore, the quantization scheme is vital for practical semantic communications.

1) Asymmetric Quantization: First, each float-point number $z_j \in \boldsymbol{z}$ is quantized into a q-bit integer z_j^q using the following quantization function:

$$z_j^q = \mathrm{clamp}(\mathrm{round}\left(\rho_s(z_j - z_{\mathrm{min}})\right); 0, 2^q - 1), \tag{8}$$

where ρ_s denotes the scale factor which maps the original range of 32-bit float point numbers into a smaller range of $[0, 2^q - 1]$. The scale factor can be calculated as

$$\rho_s = \frac{2^q - 1}{z_{\mathrm{max}} - z_{\mathrm{min}}}, \tag{9}$$

and the clamp function is exploited to remove the quantization outliers whose value exceeds $[0, 2^q - 1]$, which is defined as

$$\mathrm{clamp}(v; 0, 2^q - 1) = \begin{cases} 0, & v < 0 \\ v, & 0 \le v \le 2^q - 1 \\ 2^q - 1, & v > 2^q - 1. \end{cases} \tag{10}$$

In this way, each element of the semantic representation vector is quantized into q-bit integer z_j^q, thereby narrowing the size of constellation into $(2^q - 1)$ with identifiable amplitudes and phases [33].

2) Dequantization: Directly using the quantized values cause significant performance degradation in the MIJSCC framework, since the quantization leads to parameter mutations in the subsequent MIJSCC decoder layers. Therefore, dequantization operation is required to restore the pre-quantized value, which is formulated as

$$\hat{z}_j = \frac{z_j^q}{\rho_s} + z_{\min}. \tag{11}$$

The dequantized semantic representation vector can be expressed as $\hat{\boldsymbol{z}} = \{\hat{z}_j | 0 \leq j \leq |\hat{z}|\}$. Note that $\hat{\boldsymbol{z}}$ is still mapped into a finite constellation positions with size $2^q - 1$ but has a similar distribution as the original semantic representation vector $\boldsymbol{z}$. Moreover, the number of quantization bits q can be controlled for balancing the transmission overhead and task performance.

III. The Proposed MIJSCC framework

As in Fig. 2, we provide an in-depth description of the MIJSCC framework, which is the key enabler of the simultaneous multi-task semantic communication system. First, we introduce the principle of MI maximization which enables the unsupervised semantic representation learning. Then, we derive the global and local MI maximization method which assists the loss function design. Finally, we elaborate the training process of the proposed MIJSCC framework.

A. Mutual Information Maximization for Semantic Representation Learning

Unlike the conventional JSCC framework which passively learns the intermediate semantic representations, the proposed MIJSCC framework focuses on extracting key information from the original input and encoding it into learned semantics through initiative representation learning, rather than relying solely on a basic end-to-end approach.

One way to measure the quality of the semantic representation is to calculate the MI between the extracted semantic representation vector and the original input. Let S and Z denote the random variables sampled from the source image space $\mathcal{S}$ and the semantic representation space $\mathcal{Z}$, respectively, where the input image $\boldsymbol{s}$ and the extracted semantic representation $\boldsymbol{z}$ are viewed as different realizations of a random variable pair (S, Z). Then, the MI between S and Z can be formulated as

$$\begin{aligned}\mathcal{I}(S;Z) &= \mathbb{E}_{p(\boldsymbol{s},\boldsymbol{z})}\left[\log\frac{p(\boldsymbol{s},\boldsymbol{z})}{p(\boldsymbol{s})p(\boldsymbol{z})}\right] \\ &= D_{KL}\left(p(\boldsymbol{s},\boldsymbol{z}) \| p(\boldsymbol{s})p(\boldsymbol{z})\right),\end{aligned} \tag{12}$$

where $p(\boldsymbol{s},\boldsymbol{z})$ represents the joint probability density function of S and Z, and the associated marginal probability density functions are indicated by $p(\boldsymbol{s})$ and $p(\boldsymbol{z})$. $D_{KL}\left(p(\boldsymbol{s},\boldsymbol{z}) \| p(\boldsymbol{s})p(\boldsymbol{z})\right)$ denotes the Kullback-Leibler (KL) divergence between the joint and marginal probability distributions, which is equivalent to $\mathcal{I}(S;Z)$.

However, our goal is to maximize the MI, in this pursuit, we do not necessarily require its exact value. Moreover, the KL divergence is theoretically unbounded, which leads to infinite results during maximization. Therefore, following the method in DIM, we maximize the MI through maximizing the estimated Jensen-Shannon (JS) divergence, which is a bounded measurement defined as follows [34]

$$\begin{aligned}\hat{\mathcal{I}}^{(JS)}(S;Z) := \mathbb{E}_{p(\boldsymbol{s},\boldsymbol{z})}&\left[-\log\left(1+e^{-T_{\boldsymbol{\theta}}(\boldsymbol{s},\boldsymbol{z})}\right)\right] \\ -\mathbb{E}_{p(\boldsymbol{s})p(\boldsymbol{z})}&\left[\log\left(1+e^{T_{\boldsymbol{\theta}}(\boldsymbol{s}',\boldsymbol{z})}\right)\right],\end{aligned} \tag{13}$$

where $T_{\boldsymbol{\theta}}(\cdot)$ is a discriminator modeled by a neural network with parameters $\boldsymbol{\theta}$ and $\boldsymbol{s}'$ denotes a fake image sample. Specifically, the JS divergence is maximized in an adversarial manner, where the discriminator aims to maximize the score $T_{\boldsymbol{\theta}}(\boldsymbol{s},\boldsymbol{z})$ while minimizing the score $T_{\boldsymbol{\theta}}(\boldsymbol{s}',\boldsymbol{z})$. This process enables the discriminator to effectively distinguish whether the semantic representation originates from true source image, thereby maximizing the MI between the semantic representation and the source image (via the joint distribution term), while simultaneously reducing the dependency between the semantic representation and irrelevant samples (via the marginal distribution term).

Note that although the JS divergence cannot estimate the precise value of the MI, maximizing the JS divergence is equivalent to maximizing the MI (see Appendix A). Therefore, the MI maximization problem can be formulated as

$$\max_{\boldsymbol{\alpha},\boldsymbol{\theta}} \hat{\mathcal{I}}^{(JS)}(S;Z). \tag{14}$$

Hence, by jointly optimizing the MIJSCC encoder $E_{\boldsymbol{\alpha}}(\cdot)$ and the discriminator $T_{\boldsymbol{\theta}}(\cdot)$, the learned semantic representation summarizes the important information from the input image.

B. Global and Local Mutual Information Maximization

By focusing on different parts of the image (global or local), the learned semantic representations can be adjusted to contain information specific to different tasks. In Section III-A, we derive the MI between the original input and the semantic representation, which summarizes the feature of the whole image. Thus, (13) is also defined as the global MI

$$\Omega^{global}_{\boldsymbol{\alpha},\boldsymbol{\theta}_g} = \hat{\mathcal{I}}^{(JS)}(S;E_{\boldsymbol{\alpha}}(S)), \tag{15}$$

where $\boldsymbol{\theta}_g$ denotes the global discriminator model parameters. The global MI involves the entire image as the receptive field, thus facilitating the reconstruction tasks.

In contrast, for classification tasks, patches rather than the whole image can enhance the performance since they contain the spatial information of the local structure. Thus, we maximize the MI between each local patch of the original image and the semantic representation to learn the structural information for classification. As shown in Fig. 2, at the proposed MIJSCC encoder, the input image is first encoded into a feature map $\boldsymbol{f} \in \mathbb{R}^{N \times N \times C'}$ of $N \times N$ feature vectors, where C' denotes the depth of each feature vector,

$$\boldsymbol{f} = E_{\boldsymbol{\omega}}(\boldsymbol{s}), \tag{16}$$

where $\boldsymbol{\omega}$ represents the parameters of the first part of the MIJSCC encoder. As shown in Fig. 2, the MIJSCC encoder consists of two parts. The first part $E_{\boldsymbol{\omega}}(\cdot)$ extracts the feature map, where the $N \times N$ feature vectors correspondingly describe the

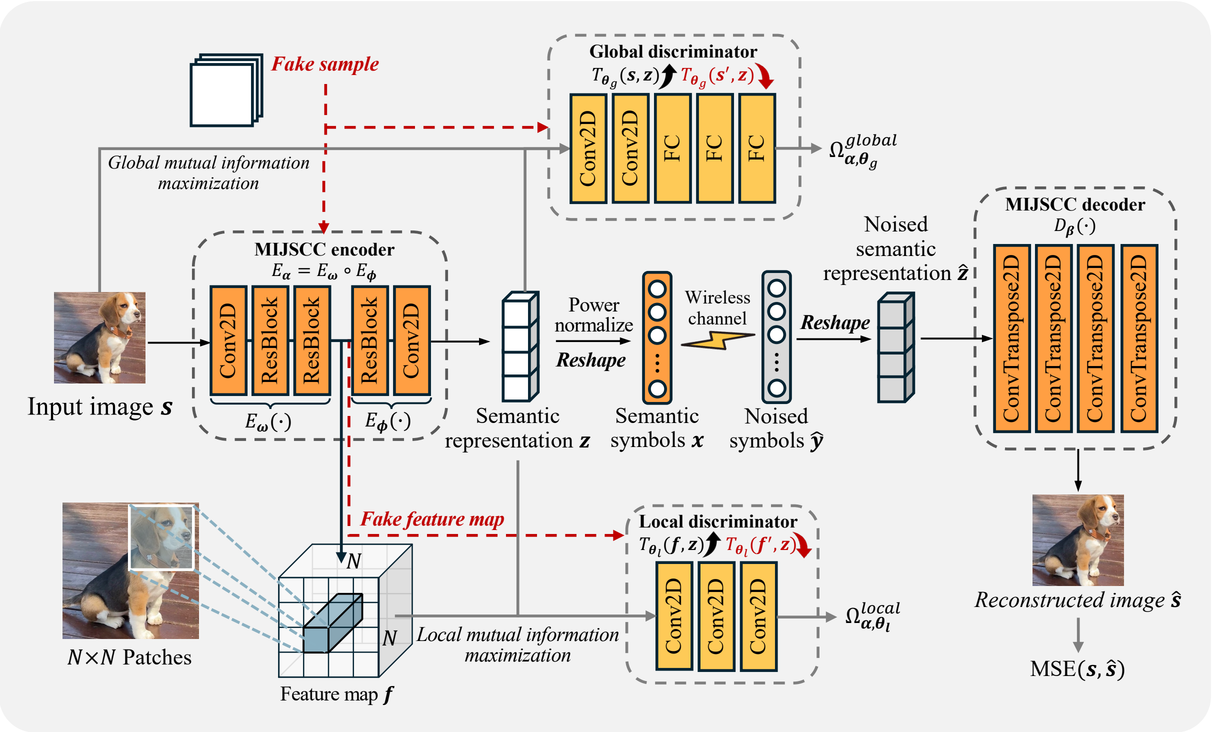


Fig. 2. The proposed MIJSCC framework. During the training process, the source image $\boldsymbol{s}$ is first encoded into feature map $\boldsymbol{f}$ and then encoded into semantic representation $\boldsymbol{z}$. At the same time, a fake sample is generated by permutation and encoded into a fake feature map. On one hand, both the real and fake samples are input into the global discriminator with the semantic representation to achieve global MI estimation $\Omega^{global}_{\boldsymbol{\alpha},\boldsymbol{\theta}_g}$. On the other hand, both the real and fake feature maps are input into the local discriminator with the semantic representation to obtain local MI estimation $\Omega^{local}_{\boldsymbol{\alpha},\boldsymbol{\theta}_l}$. At the receiver-side, the noised symbols $\hat{\boldsymbol{y}}$ are input into the MIJSCC decoder to reconstruct image $\hat{\boldsymbol{s}}$, where the reconstruction loss MSE$(\boldsymbol{s},\hat{\boldsymbol{s}})$ is derived to jointly optimize the MIJSCC codec together with the estimated MI.

information of the $N\times N$ patches from the original image [35]. The second part $E_{\boldsymbol{\phi}}(\cdot)$ summarizes the feature vectors into the semantic representation, so that $E_{\boldsymbol{\alpha}} = E_{\boldsymbol{\omega}} \circ E_{\boldsymbol{\phi}}$. Therefore, we derive the MI between each feature vector and the semantic representation $\boldsymbol{z}$ following the same method in Section III-A, then calculate the average MI of all patches as the local MI,

$$\begin{aligned}\Omega^{local}_{\boldsymbol{\alpha},\boldsymbol{\theta}_l} &= \frac{1}{N^2}\sum_{i=1}^{N^2}\hat{\mathcal{I}}^{(JS)}(F^{(i)};Z) \\ &= \frac{1}{N^2}\sum_{i=1}^{N^2}\hat{\mathcal{I}}^{(JS)}(E^{(i)}_{\boldsymbol{\omega}}(S);E_{\boldsymbol{\alpha}}(S)),\end{aligned} \tag{17}$$

where F denotes the random variable sampled from the feature map space $\mathcal{F}$ and $F^{(i)}$ denotes the i-th feature vector. $\boldsymbol{\theta}_l$ represents the parameters of the local discriminator model.

Different from conventional classification methods which rely on labeled data for supervised learning, the proposed semantic representation trained by local MI maximization captures the structural knowledge of the margins detected in each local patch, which is discriminative across image samples of different categories. This unsupervised classification approach focuses on identifying the similarities and differences of the visual characteristics among image samples, which can be applied label-free, rather than depending on the matching between each sample and its assigned label. By jointly training the MIJSCC encoder, global discriminator and local discriminator, both the global MI and the local MI can be optimized, thereby learning the semantic representation for both classification tasks and reconstruction tasks. Furthermore, the performance of reconstruction tasks is not only affected by the semantic representation quality, but also by the generator model. Therefore, in addition to maximizing the global MI that captures the global image feature, we also minimize the MSE between the original image $\boldsymbol{s}$ and the image $\hat{\boldsymbol{s}}$ generated by the MIJSCC decoder $D_{\boldsymbol{\beta}}(\cdot)$, which aims to decrease the pixel-level difference. Thus, the overall loss function of training the MIJSCC model can be expressed as

$$\mathcal{L}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta}_g,\boldsymbol{\theta}_l) = \text{MSE}(\boldsymbol{s},\hat{\boldsymbol{s}}) - \lambda(\mu_1\Omega^{global}_{\boldsymbol{\alpha},\boldsymbol{\theta}_g} + \mu_2\Omega^{local}_{\boldsymbol{\alpha},\boldsymbol{\theta}_l}), \tag{18}$$

where λ is the coefficient controlling the trade-off between reconstruction performance and semantic representation learning. μ_1 and μ_2 controls the focus on global or local MI maximization. By adjusting these hyperparameters, the proposed MIJSCC framework can either focus on classification/reconstruction tasks or balance the performance of both tasks and complete them simultaneously.

Remark 1: Compared to conventional single-task loss functions which either focus on minimizing the pixel-level reconstruction loss, including MSE and structural similarity index measure [36], or the cross-entropy loss for classification, the proposed loss function aims to minimize the generative loss

Algorithm 1 Training process of the proposed MIJSCC framework

```
Initialize: Parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}_g, \boldsymbol{\theta}_l$.
for epoch $= 1 \rightarrow 200$ do
    for each sample $\boldsymbol{s}$ in an image batch $\boldsymbol{S}$ do
        Input: Image $\boldsymbol{s}$.
        Shuffle image $\boldsymbol{s}$ to create a fake image $\boldsymbol{s'}$ for comparison.
        Extract feature map $\boldsymbol{f}$ and $\boldsymbol{f'}$ from the original and fake image via $E_{\boldsymbol{\omega}}(\boldsymbol{s})$ and $E_{\boldsymbol{\omega}}(\boldsymbol{s'})$, respectively.
        Derive the semantic representation $\boldsymbol{z}$ from the feature map $\boldsymbol{f}$ via $E_{\boldsymbol{\phi}}(\boldsymbol{f})$.
        Estimate global MI $\Omega^{global}_{\boldsymbol{\alpha},\boldsymbol{\theta}_g}$ with $\boldsymbol{s}$, $\boldsymbol{s}'$, and $\boldsymbol{z}$ via (13).
        Estimate local MI $\Omega^{local}_{\boldsymbol{\alpha},\boldsymbol{\theta}_l}$ with $\boldsymbol{f}$, $\boldsymbol{f}'$, and $\boldsymbol{z}$ via (17).
        for each dimension $j$ of the semantic representation vector $\boldsymbol{z}$ do
            Quantize each float number $z_j$ into $q$-bit integer $z_j^q$ via (8).
            Dequantize $z_j^q$ using (11) and restore the pre-quantized value as $\hat{z}_j$.
        end for
        Reshape the dequantized semantic representation vector into symbols $\boldsymbol{x}$ and perform power normalization via (2).
        Receive symbols $\boldsymbol{y}$ over wireless channel via (3) and obtain noised symbols $\hat{\boldsymbol{y}}$ via (4).
        Reshape $\hat{\boldsymbol{y}}$ into semantic representation vector $\hat{\boldsymbol{z}}$.
        Reconstruct image $\hat{\boldsymbol{s}}$ via $D_{\boldsymbol{\beta}}(\hat{\boldsymbol{z}})$.
        Calculate the loss value using (18) and update model parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}_g, \boldsymbol{\theta}_l$.
    end for
end for
```

MSE$(\boldsymbol{s}, \hat{\boldsymbol{s}})$ as well as maximize the global information $\Omega^{global}_{\boldsymbol{\alpha},\boldsymbol{\theta}_g}$ and local information $\Omega^{local}_{\boldsymbol{\alpha},\boldsymbol{\theta}_l}$ in the semantic representation simultaneously. Therefore, both the end-to-end target and the intermediate semantics are actively learned, while in conventional loss function design, the training process is merely dominated by the target of a single task. Moreover, since the MI can be learned in an unsupervised way, the proposed loss function is optimized without the assistance of labels, while in conventional task-driven semantic communication frameworks, the labels are indispensable.

C. Training Method for MIJSCC

The training of MIJSCC is divided into two stages as shown in Fig. 3, where stage 1 is unsupervised focusing on reconstruction task and MI-based learning. Stage 2 is supervised to achieve the lightweight classifier according to personalized local tasks. During the training process, we update the model parameters of the MIJSCC encoder, MIJSCC decoder, global discriminator, and the local discriminator to realize both semantic representation learning and full-image generating. In order to achieve a MIJSCC codec that is robust under various channel conditions, we train the above models under SNR $\in \{6, 21\}$ dB, which is chosen randomly at each update. At the inference stage, the global and local discriminator are not used. For both image reconstruction tasks and classification tasks, the MIJSCC encoder is implemented for semantic representation extracting. The obtained semantic representation is directly input into a given classifier for image classification and the trained MIJSCC decoder for image reconstruction. Here we achieve the classifier by freezing the parameters of the MIJSCC encoder and training a 3-layer fully-connected network using the semantic representation as input. In practical use, this process can be completed on target devices with a small amount of local task labels, which fine-tunes the classifier based on the unsupervised, discriminative semantic representations instead of training from scratch under large-scale labeled-data as in other fully-supervised methods. The training and inference method of the proposed MIJSCC framework is summarized in Algorithm 1 and 2, respectively.

Algorithm 2 Inference process of the proposed MIJSCC framework

```
Input: Image $\boldsymbol{s}$.
Extract the semantic representation $\boldsymbol{z}$ from the original image via $E_{\boldsymbol{\alpha}}(\boldsymbol{s})$.
for each dimension $j$ of the semantic representation $\boldsymbol{z}$ do
    Quantize float number $z_j$ into $q$-bit integer $z_j^q$ via (8).
    Dequantize $z_j^q$ using (11) and restore the pre-quantized value as $\hat{z}_j$.
end for
Reshape the dequantized semantic representation vector into symbols $\boldsymbol{x}$ and perform power normalization via (2).
Receive symbols $\boldsymbol{y}$ over wireless channel via (3) and obtain noised symbols $\hat{\boldsymbol{y}}$ via (4).
Reshape $\hat{\boldsymbol{y}}$ into semantic representation vector $\hat{\boldsymbol{z}}$.
Perform image classification via $\psi(\hat{\boldsymbol{z}})$.
Reconstruct image $\hat{\boldsymbol{s}}$ via $D_{\boldsymbol{\beta}}(\hat{\boldsymbol{z}})$.
Output: Reconstructed image $\hat{\boldsymbol{s}}$ and classification result.
```

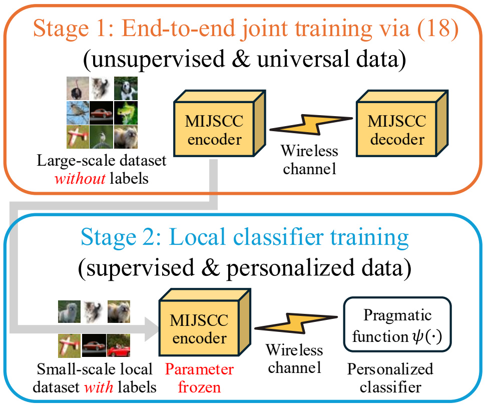

Stage 1: End-to-end joint training via (18)
(unsupervised & universal data)
Large-scale dataset
without labels
MIJSCC encoder
Wireless channel
MIJSCC decoder
Stage 2: Local classifier training
(supervised & personalized data)
Small-scale local dataset with labels
MIJSCC encoder
Parameter frozen
Wireless channel
Pragmatic function $\psi(\cdot)$
Personalized classifier

Fig. 3. The training process of the proposed MIJSCC framework.

IV. Adaptive MIJSCC with Attention-Based Semantic Enhancement

To enhance the robustness of the proposed MIJSCC framework against varying channel noise and further reduce the semantic transmission overhead, this section presents an adaptive MIJSCC framework, referred to as A-MIJSCC, which is shown in Fig. 4. By introducing adaptive channel masking and attention-based semantic enhancement, A-MIJSCC aims to dynamically adjust the number of activated semantic feature dimensions under different channel conditions, thereby improving semantic transmission efficiency while maintaining task performance.

In traditional communication, adaptive modulation and coding (AMC) can dynamically adjust the modulation scheme and coding rate based on channel conditions, adding redundancy to combat noise effects and enhancing data transmission reliability. Inspired by this, our study integrates the concept of dynamic neural networks [37] into the proposed MIJSCC framework to achieve adaptive channel coding. When the channel SNR is low, the transmitter activates more dimensions to mitigate the noise impact on the transmission of semantic representations. Otherwise, fewer dimensions will be activated to reduce the transmission overhead.

In practical application scenarios, devices can receive the knowledge of channel conditions through the feedback channel, including SNR, fading, etc. Thus, the channel condition can be input into a neural network as auxiliary knowledge, which is trained to generate the noise-adaptive channel mask. As illustrated in Fig. 4, a learnable noise encoder $N_{enc}(\cdot)$, designed as a 4-layer multi-layer perceptron (MLP), is used to map the scalar noise variance δ^2 into a noise-aware feature vector $\boldsymbol{g} \in \mathbb{R}^d$ with the same dimension as the semantic representation:

$$\boldsymbol{g} = N_{enc}(\delta^2), \tag{19}$$

where each component g_i is a non-negative, monotonically increasing function of δ^2 (as proven in Appendix B). This ensures the noise intensity meaningfully influence the activation of semantic representation dimensions, emphasizing that the channel mask in A-MIJSCC is not a set of random numbers but a deterministic, structured function learned from the noise variance δ^2.

Each component g_i of the noise feature vector $\boldsymbol{g}$ represents the activation potential of the corresponding dimension z_j of the semantic representation $\boldsymbol{z}$. However, the value distribution of $\boldsymbol{g}$ is irregular. Thus, directly determining the activation state of the semantic representation $\boldsymbol{z}$ according to $\boldsymbol{g}$ results in discontinuous activation, which requires additional indicator signals to be transmitted to the receiver. Therefore, to eliminate unnecessary communication overhead, the A-MIJSCC framework aims to generate a consecutive channel mask by multiplying $\boldsymbol{g}$ with an upper triangular matrix $\boldsymbol{U} \in \mathbb{R}^{d\times d}$, where d represents the dimension of the noise feature vector. Then, the updated noise feature vector $\tilde{\boldsymbol{g}}$ follows the decreasing trend:

$$\tilde{\boldsymbol{g}} = \boldsymbol{g} \cdot \boldsymbol{U}, \quad U_{ij} = \begin{cases} 1, & \text{if } i \le j, \\ 0, & \text{if } i > j. \end{cases} \tag{20}$$

As shown in (20), the updated noise feature $\tilde{\boldsymbol{g}}$ satisfies $\tilde{g}_j \ge \tilde{g}_{j+1}, \forall j \in [0, d-1]$. Unlike other feature pruning techniques [38], [39], which remove unnecessary dimensions without considering the pruned index, the proposed A-MIJSCC framework generates a consecutive, deterministic binary channel mask. Therefore, the transmitted dimensions are successively distributed and only one index indicating the ending position is required to be transmitted, significantly reducing the transmission and decoding overhead.

Next, a pruning threshold η is introduced to prune $\tilde{\boldsymbol{g}}$ and obtain the channel mask $\boldsymbol{m}$, where each component m_j satisfies:

$$m_j = \begin{cases} 1, & \text{if } \tilde{g}_j > \eta, \\ 0, & \text{if } \tilde{g}_j \le \eta. \end{cases} \tag{21}$$

By performing the Hadamard product between the semantic representation $\boldsymbol{z}$ and the channel mask $\boldsymbol{m}$, we obtain the masked semantic representation $\boldsymbol{z}'$ as:

$$\boldsymbol{z}' = \boldsymbol{m} \odot \boldsymbol{z}. \tag{22}$$

Noted that the pruning threshold η is a hyperparameter related to task performance and loss design. By tuning the value of η, unimportant dimensions will be temporarily deactivated without affecting the training of the MIJSCC framework, where the importance is automatically learned during the joint training of the channel mask generator and the MIJSCC model. Additionally, in each training batch, the value of δ^2 is randomly selected from a pre-defined noise variance range to serve as input of the noise encoder network $N_{enc}(\cdot)$, which can train $N_{enc}(\cdot)$ to generate channel masks that adapt to varying channel conditions.

To compensate for the potential semantic information loss introduced by channel mask operation, the masked semantic representation $\boldsymbol{z}'$ is further input into an attention-based semantic enhancement module $N_{attn}(\cdot)$, which consists of two fully connected layers and nonlinear activation functions. This module generates and dynamically adjusts the weight $\mathbf{w}$ for each dimension of $\boldsymbol{z}'$, re-weighting the activated dimensions to reinforce important semantic features while adapting to varying channel SNRs. The enhanced semantic representation is represented as:

$$\boldsymbol{z}^* = N_{attn}(\boldsymbol{z}') = \mathbf{w} \odot \boldsymbol{z}'. \tag{23}$$

Similar to the MIJSCC framework, the enhanced semantic representation $\boldsymbol{z}^*$ is quantized and then mapped into semantic symbols $\boldsymbol{x}^*$. In contrast, the A-MIJSCC framework activates and transmits only a subset of the original semantic representation dimensions, depending on the dynamic channel conditions. The ratio of activated dimensions is denoted by γ, and the receiver fills the inactive parts with zeros.

By jointly training the MIJSCC model, the noise encoder $N_{enc}(\cdot)$, and the semantic enhancement module $N_{attn}(\cdot)$, the proposed A-MIJSCC framework successively activates the important dimensions of the original semantic representation according to dynamic channel conditions. Afterwards, the attention-based module will enhance the masked semantic representation to mitigate potential performance degradation. Algorithm 3 provides the detailed training procedure of the A-MIJSCC framework.

Remark 2: The setting of pruning threshold η depends on datasets, tasks, as well as the required trade-off level between task performance and the semantic communication overhead.

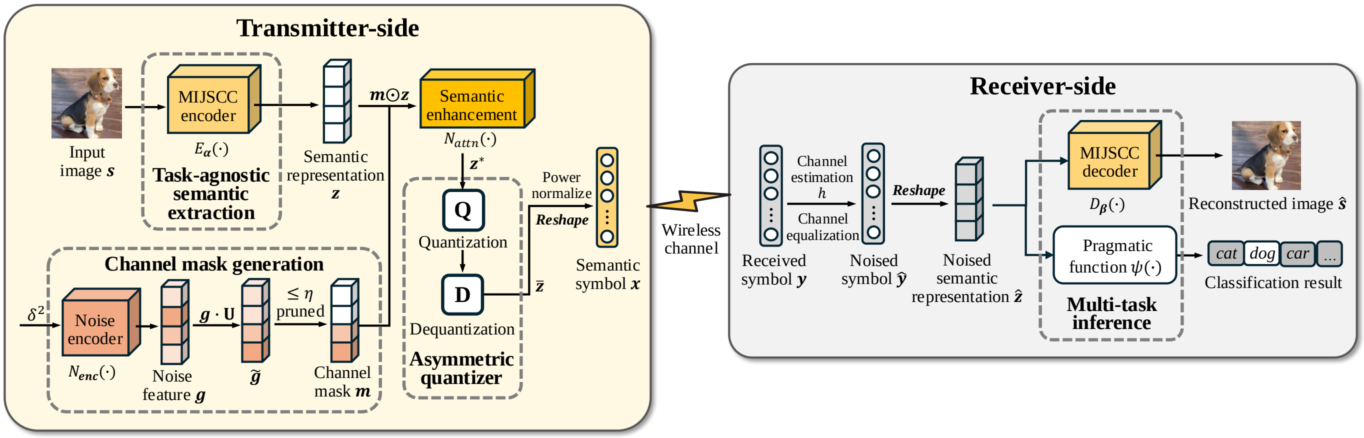

Transmitter-side
Input image $\boldsymbol{s}$
MIJSCC encoder
$E_{\boldsymbol{\alpha}}(\cdot)$
Task-agnostic semantic extraction
Semantic representation $\boldsymbol{z}$
$\boldsymbol{m}\odot\boldsymbol{z}$
Semantic enhancement
$N_{attn}(\cdot)$
$\boldsymbol{z}^*$
Q
Quantization
D
Dequantization
Asymmetric quantizer
Power normalize
Reshape
$\bar{\boldsymbol{z}}$
Semantic symbol $\boldsymbol{x}$
Channel mask generation
δ^2
Noise encoder
$N_{enc}(\cdot)$
Noise feature $\boldsymbol{g}$
$\boldsymbol{g}\cdot\mathbf{U}$
$\tilde{\boldsymbol{g}}$
$\leq\eta$ pruned
Channel mask $\boldsymbol{m}$
Wireless channel
Receiver-side
Received symbol $\boldsymbol{y}$
Channel estimation h
Channel equalization
Noised symbol $\hat{\boldsymbol{y}}$
Reshape
Noised semantic representation $\hat{\boldsymbol{z}}$
MIJSCC decoder
$D_{\boldsymbol{\beta}}(\cdot)$
Reconstructed image $\hat{\boldsymbol{s}}$
Pragmatic function $\psi(\cdot)$
Multi-task inference
cat
dog
car
...
Classification result

Fig. 4. The structure of the Adaptive MIJSCC with attention-based semantic enhancement. According to the channel mask generation module, the noise variance is first input into the noise encoder $N_{enc}(\cdot)$ to achieve the noise feature $\boldsymbol{g}$, which is multiplied with an unit upper triangular matrix. Then, the updated noise feature $\tilde{\boldsymbol{g}}$ is pruned with respected to the pruning threshold η to produce the channel mask $\boldsymbol{m}$ that performs Hadamard product with the semantic representation $\boldsymbol{z}$. Finally, the masked semantic representation is enhanced by the attention-based module $N_{attn}(\cdot)$, which aims to compensate for the semantic information loss that may be caused by channel masking operations.

Algorithm 3 Training process of the A-MIJSCC framework

```
Initialize: Parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}_g, \boldsymbol{\theta}_l$, noise encoder network $N_{enc}(\cdot)$, semantic enhancement module $N_{attn}(\cdot)$ and pruning threshold $\eta$.
for epoch $= 1 \to 200$ do
    for each sample $\boldsymbol{s}$ in an image batch $\boldsymbol{S}$ do
        Input: Image $\boldsymbol{s}$, current noise variance $\delta^2$.
        Perform step 5-9 in Algorithm 1 and obtain $\boldsymbol{z}$
        Obtain the noise feature $\boldsymbol{g}$ via $N_{enc}(\delta^2)$.
        Obtain the decreasing noise feature $\tilde{\boldsymbol{g}}$ using (20).
        Prune $\tilde{\boldsymbol{g}}$ according to $\eta$ and obtain $\boldsymbol{m}$ via (21).
        Obtain masked semantic representation $\boldsymbol{z}'$ via (22).
        Generate attention weight $\boldsymbol{w}$ and obtain enhanced semantic representation $\boldsymbol{z}^*$ via (23).
        Quantize $\boldsymbol{z}^*$ and transmit to the receiver via step 10 to 15 in Algorithm 1 to obtain $\hat{\boldsymbol{y}}$.
        Reshape $\hat{\boldsymbol{y}}$ into vector and fill the masked dimensions with zero to obtain $\hat{\boldsymbol{z}}$.
        Reconstruct image $\hat{\boldsymbol{s}}$ via $D_{\boldsymbol{\beta}}(\hat{\boldsymbol{z}})$.
        Compute the loss function (18) to update $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}_g, \boldsymbol{\theta}_l$, $N_{enc}(\cdot)$ and $N_{attn}(\cdot)$.
    end for
end for
```

V. Simulation Results and Discussions

In this section, we evaluate the performance of the simultaneous multi-task semantic communication system enabled by the MIJSCC framework. The proposed MIJSCC framework can be jointly trained on a server and moved to real network devices for inference. First, we describe the experiment settings including the dataset, hyperparameters, and model structures. Then, we explain the comparison benchmarks. Finally, we provide a detailed discussion on simulation results.

TABLE I
The DNN Structures of MIJSCC Encoder and Decoder

	Layer	Output Size	Activation
MIJSCC Encoder	Conv2D+BatchNorm	$64 \times 32 \times 32$	ReLU
	ResidualBlock	$64 \times 16 \times 16$	ReLU
	ResidualBlock	$128 \times 8 \times 8$	ReLU
	ResidualBlock	$256 \times 4 \times 4$	ReLU
	Conv2D	$512 \times 1 \times 1$	None
MIJSCC Decoder	ConvTranspose2D+BatchNorm	$256 \times 4 \times 4$	ReLU
	ConvTranspose2D+BatchNorm	$128 \times 8 \times 8$	ReLU
	ConvTranspose2D+BatchNorm	$64 \times 16 \times 16$	ReLU
	ConvTranspose2D+BatchNorm	$3 \times 32 \times 32$	Sigmoid

A. Experiment Settings

We evaluate the proposed simultaneous multi-task semantic communication system on CIFAR10 dataset to complete classification tasks and reconstruction tasks at the same time. We set the trade-off coefficient λ between the MSE loss and the MI in a collection of $\{0.1, 0.05, 0.01, 0.005, 0.001, 0.0005\}$, then test the system under each value to find a point which has satisfying performance for both tasks (see Section V-B). Afterwards, we fix the value of λ as 0.05 and 0.005 during the tests. The hyperparameters μ_1 and μ_2 for global and local MI maximization are fixed to 0.01 and 1, respectively. The MIJSCC encoder, decoder, global discriminator and local discriminator are jointly trained for 200 epochs using the Adam optimizer with learning rate=0.001 and batch size=32.

The model structures and output sizes of the MIJSCC encoder and decoder are demonstrated in Table I. The MIJSCC encoder mainly consists of three downsampling residual blocks followed by a convolutional layer which controls the dimension of the semantic representation vector. The MIJSCC decoder is composed of four deconvolutional layers, each is followed by a batch normalization layer and an activation layer. The model structures of the global and local discriminators are shown in Table II. In the global discriminator, the image is flattened into vector and then concatenated with the semantic representation vector to composite the input of

TABLE II
The Model Structures of Global and Local Discriminator

	Layer	Activation
Global Discriminator	Conv2D	ReLU
	Conv2D	None
	Flatten+Concatenate	None
	Fully-connected Layer	ReLU
	Fully-connected Layer	ReLU
	Fully-connected Layer	None
Local Discriminator	Expand+Concatenate	None
	Conv2D	ReLU
	Conv2D	ReLU
	Conv2D	None

fully-connected layers which derive the score. In the local discriminator, the semantic representation is first expanded into the same size of the feature map and then concatenated with the feature map to composite the input of convolutional layers.

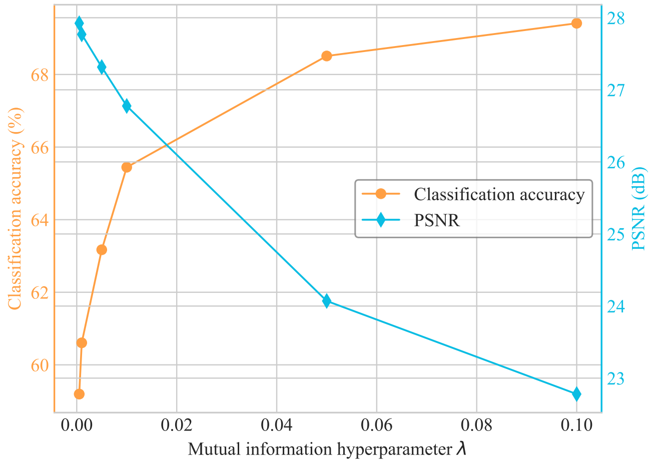


Fig. 5. The classification accuracy and the PSNR of the proposed MIJSCC framework w.r.t. MI hyperparameter λ.

B. Performance Evaluation

We compare our proposed simultaneous multi-task semantic communication system with the following benchmarks:

- **Deep JSCC with MSE loss [5]:** We compare with the single-task semantic communication system for image reconstruction using MSE as loss function, which is a representative and widely-recognized baseline in the field of image semantic communication. The Deep JSCC framework has the same DNN structure with our MIJSCC framework for fair comparison.
- **Separate source-channel coding (SSCC):** We compare with the conventional communication system with better portable graphics (BPG) as the source coding method and consider error-free transmission within the channel capacity $\Delta = \log_2(1 + \text{SNR})$. Therefore, we can obtain the maximum required compression ratio as

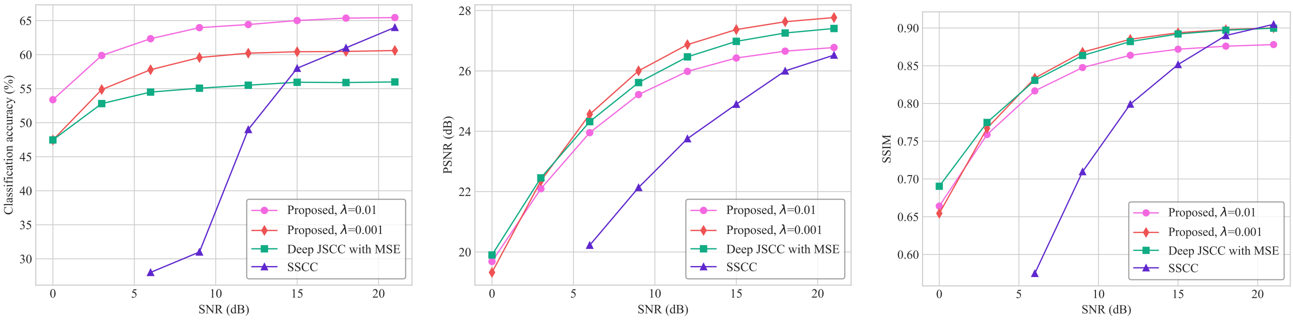
$$R_{\max} = \frac{bLHC}{n\Delta}, \tag{24}$$

where $b = 8$ denotes the number of bits allocated to each color channel. Note that BPG offers significantly better compression performance than JPEG, which serves as a strong source coding benchmark. Additionally, for BPG image compression, there exists a maximum achievable compression ratio $R_{\max}^{\text{BPG}}$. Therefore, $R_{\max} > R_{\max}^{\text{BPG}}$ may occurs when facing bad channel condition (low SNR) or less transmitted symbols. In this circumstance, the SSCC scheme is infeasible since the required compression ratio exceeds the ability of the BPG compression algorithm. To guarantee fair comparisons, we input the BPG-compressed image into a pretrained classification model which has the same structure as the MIJSCC encoder plus the pragmatic classifier.

In Fig. 5, we demonstrate the impact of the hyperparameter λ on the performance of both the classification task and the reconstruction task. As defined in Section III-B, λ controls the trade-off between the MI and the MSE loss. A larger value of λ leads to greater focus on the maximization of the MI, which promotes the learning of semantic representation. Otherwise, a smaller value of λ reflects a stronger emphasis on image recovery, with $\lambda = 0$ equivalent to conventional single-task semantic communication framework for only image reconstruction. It can be observed that λ obtains the trade-off between task performance of classification and reconstruction. To facilitate the simultaneous implement of both tasks, a proper λ value should be chosen to support the training of both the semantic representation and the end-to-end target.

Then, we separately analyze the performance of the proposed MIJSCC framework on classification tasks and reconstruction tasks. As shown in Fig. 6(a), we compare the task performance of classification among different schemes, including the MIJSCC-based method under $\lambda = 0.01$, $\lambda = 0.001$ and two baseline methods. It can be observed that larger λ leads to higher classification accuracy, reflecting that discriminative features are learned through MI maximization, thereby supporting image classification without the supervision of labels. However, the Deep JSCC with MSE loss only focuses on minimizing the pixel-level differences between the original and reconstructed images while neglecting the learning of useful semantics in the representation. Therefore, the semantic representation extracted from the MSE-based JSCC cannot be directly used for classification tasks. Secondly, the conventional SSCC scheme based on BPG compression fails to perform the classification task under bad channel conditions since the required compression ratio too high even under full channel capacity, which exceeds the ability of the BPG algorithm. Additionally, the SSCC suffers from the cliff effect compared with our proposed MIJSCC framework, although it achieves a comparable performance at 21 dB. Note that although the performance of advanced task-oriented semantic communication frameworks has surpassed 85%, these frameworks are uniquely designed for classification task and cannot be used for reconstruction.

The performance of image reconstruction tasks under different schemes is shown in Fig. 6(b) and 6(c) in terms of PSNR and Structural Similarity Index Measure (SSIM), respectively. It can be seen that the proposed MIJSCC framework with $\lambda = 0.001$ has only negligible performance loss in low-SNR



(a) The classification accuracy w.r.t. SNR.

(b) PSNR w.r.t. SNR.

(c) SSIM w.r.t. SNR.

Fig. 6. The performance of classification and reconstruction tasks w.r.t. SNR under different communication schemes.

region compared with the MSE-based deep JSCC framework, which is specifically designed for image reconstruction targets. Moreover, the MIJSCC framework outperforms the deep JSCC under good channel conditions in terms of PSNR, indicating that MI maximization can somehow assist the learning of end-to-end goals. Furthermore, the MIJSCC framework with $\lambda = 0.01$ undergoes marginal performance loss in terms of both PSNR and SSIM due to the performance trade-off between image reconstruction tasks and classification tasks, however, it significantly outperforms the conventional SSCC framework in low-SNR region. Thus, it is proved that employing both representation learning and end-to-end target training with the proposed MIJSCC framework can effectively support discriminative feature learning while maintaining image structure consistency. Fig. 7 presents the visualization results of the image reconstruction task under different semantic communication frameworks when SNR = 9 dB and SNR = 12 dB. It is observed that the proposed MIJSCC framework achieves comparable performance to Deep JSCC with MSE when $\lambda = 0.001$.

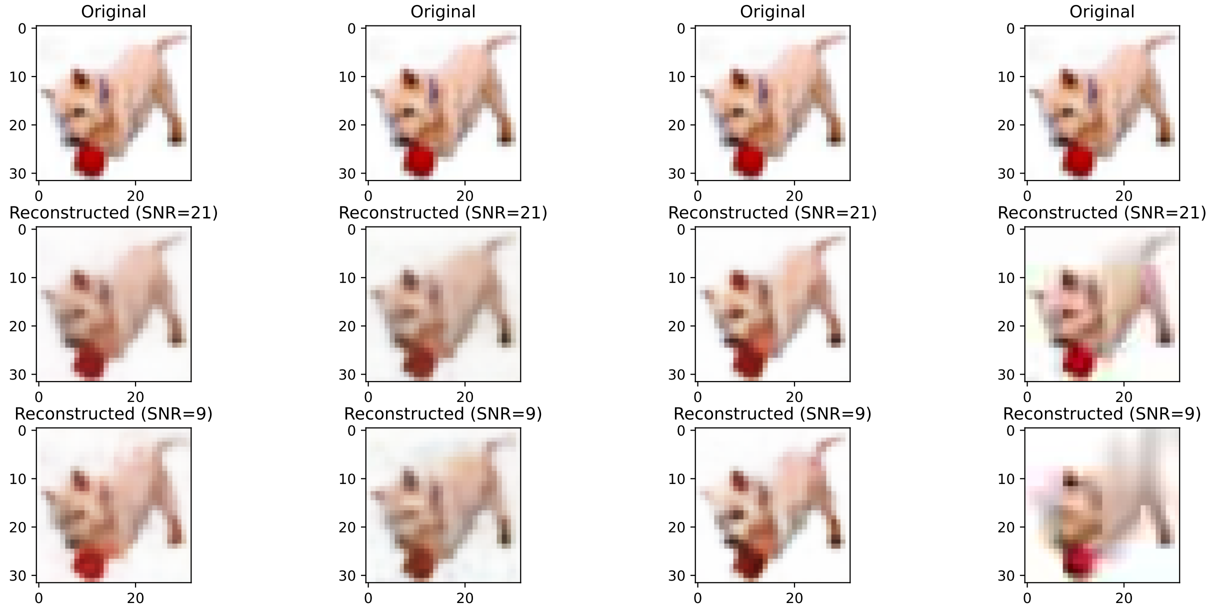
To further verify the robustness of the MIJSCC framework, we conducted additional experiments under Rayleigh fading channel incorporating LS channel estimation. We trained and evaluated the MIJSCC using the estimated channel state information (CSI). The results show that imperfect CSI has a negligible impact on the final performance, as demonstrated in Fig. 8, which verifies the scalability of the MIJSCC method under practical channel conditions.

Fig. 9 verifies the impact of global and local MI in the MIJSCC loss function on the performance of image classification tasks under different MI hyperparameter λ. According to (18) in Section III-B, a larger λ refers to greater effect of MI optimization. It can be observed that when λ increases, the performance degradation of image classification tasks becomes obvious when local MI is removed. This proves that the optimization of local MI is directly related to the discriminability of semantic representations, thereby determining the performance of image classification tasks.

Furthermore, to evaluate the effectiveness of semantic quantization for transmitted symbols, we compare the performance of the MIJSCC framework with analog transmission (without symbol quantization), 8-bit quantization, and 4-bit quantization. For example, Fig. 10(a) represents the constellation points before symbol quantization, which exhibits full-resolution constellation mapping and contains the full information extracted from the neural network represented by single-precision floating-point format (Float32) in 32 bits. Although this analog method preserves more information, it maps the transmitted symbols into nearly-continuous constellation points with indistinguishable amplitudes and phases, which is unachievable for the implementation of RF modules in current communication systems. Through quantizing the semantic representation vector into 4-bit integers, the reshaped symbols will be mapped into less constellation points with a more discrete distribution, as shown in Fig. 10(b). It can be observed that 4-bit constellation points exhibit a similar distribution compared with the full-resolution constellation points of the same image, but maintains a larger point-distance, thereby providing distinguishable amplitude-shift and phase-shift for further implementations in practical communication systems with limited RF capabilities.

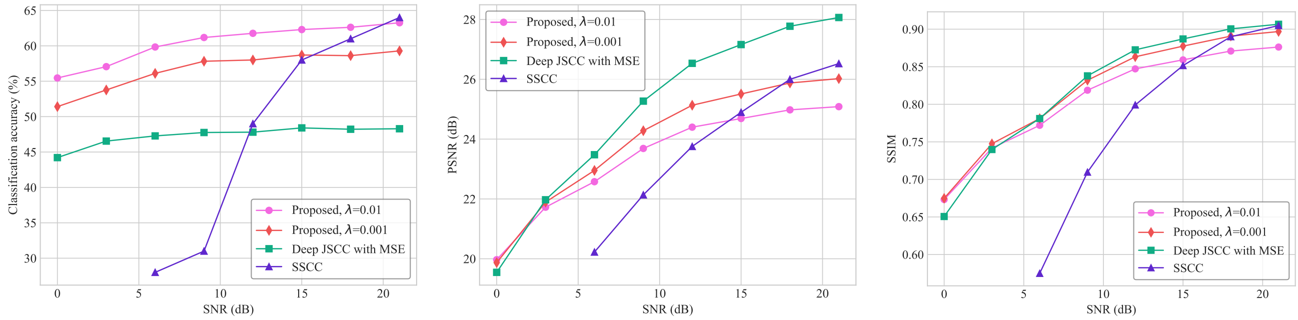
Then, we evaluate the performance of both classification tasks and image reconstruction tasks on the quantized MIJSCC framework with different quantization precision. We demonstrate the classification accuracy and the PSNR among MIJSCC with analog transmission, 8-bit quantization, and 4-bit quantization in Fig. 11(a) and Fig. 11(b), respectively. It can be observed that, compared to the continuous constellation points in analog signal transmission, the image classification accuracy after 8-bit quantization experiences only a slight reduction, while the PSNR remains almost unchanged. This suggests that there is inherent redundancy in the semantic symbols represented by 32-bit floating-point numbers. The optimized 8-bit quantized symbols are able to efficiently utilize bandwidth resources without significantly compromising task performance. However, when the quantization is reduced to 4 bits, the performance of the MIJSCC framework for both image classification and reconstruction tasks deteriorates. This is primarily due to the low-precision quantization, where fine-grained semantic information is mapped to a limited number of constellation points. As a result, critical information required for semantic decoding and noise mitigation is lost. Consequently, this reduces the discriminative and generative capabilities of the semantic representation.

We further evaluate the effect of the channel mask by comparing the fixed-length MIJSCC framework (without channel



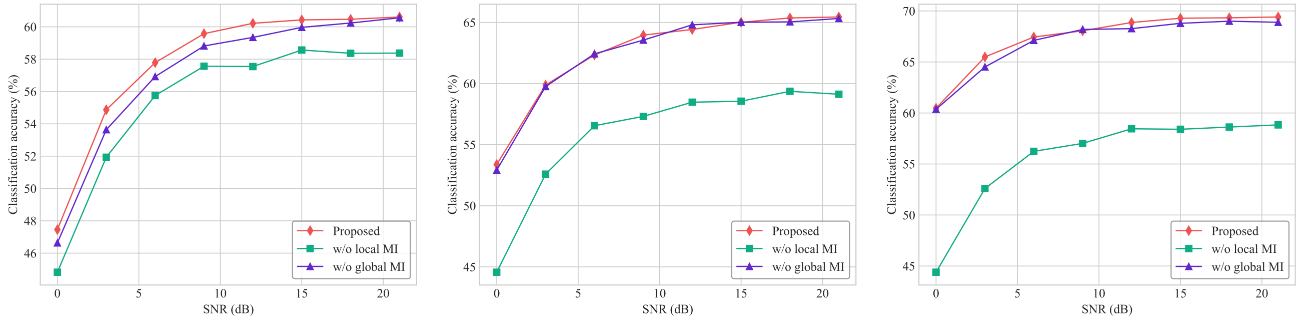
(a) Proposed, $\lambda = 0.001$ (b) Proposed, $\lambda = 0.01$ (c) Deep JSCC with MSE (d) SSCC

Fig. 7. Visualization examples of image reconstruction tasks under different semantic communication frameworks.



(a) Classification accuracy (b) PSNR (c) SSIM

Fig. 8. The performance of classification and reconstruction tasks w.r.t. SNR under Rayleigh fading channel with LS channel estimation.

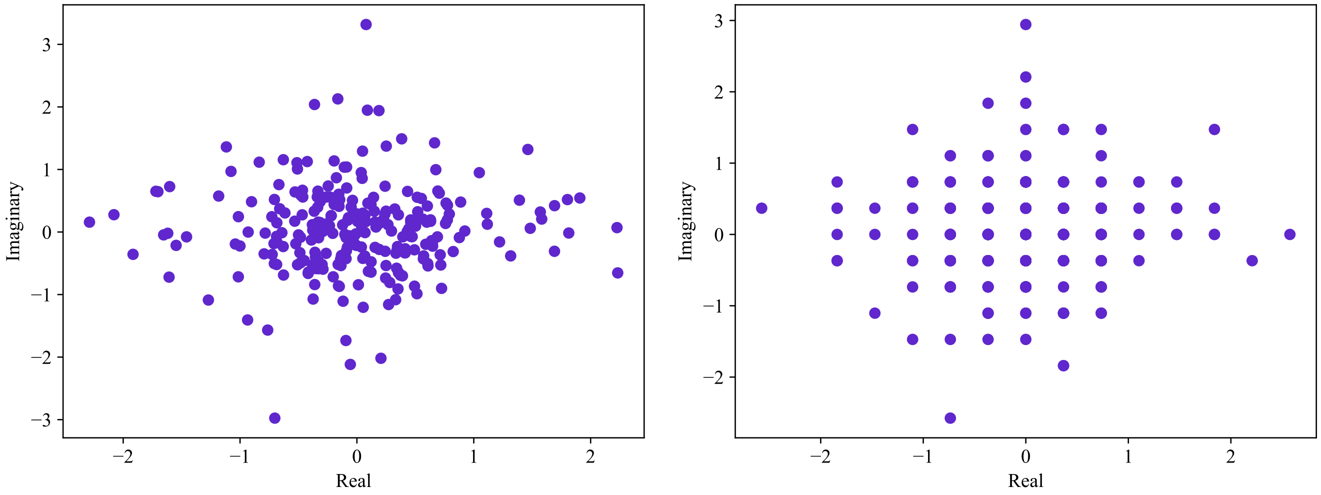


(a) λ=0.001 (b) λ=0.01 (c) λ=0.1

Fig. 9. The impact of removing global or local MI in the MIJSCC loss function on image classification tasks.

masking) and the proposed A-MIJSCC framework (with adaptive channel masking) under different pruning thresholds and channel conditions. The 8-bit quantized MIJSCC framework is used as the pre-trained model, and we fine-tune the A-MIJSCC framework for 20 epochs. As shown in Fig. 12(a), the ratio of activated dimensions γ decreases as the SNR increases. It indicates that under improved channel conditions, the A-MIJSCC framework tends to prune more redundant dimensions, thereby reducing the semantic transmission overhead. Moreover, as the pruning threshold η increases, the A-MIJSCC framework becomes aggressive in pruning unimportant dimensions.

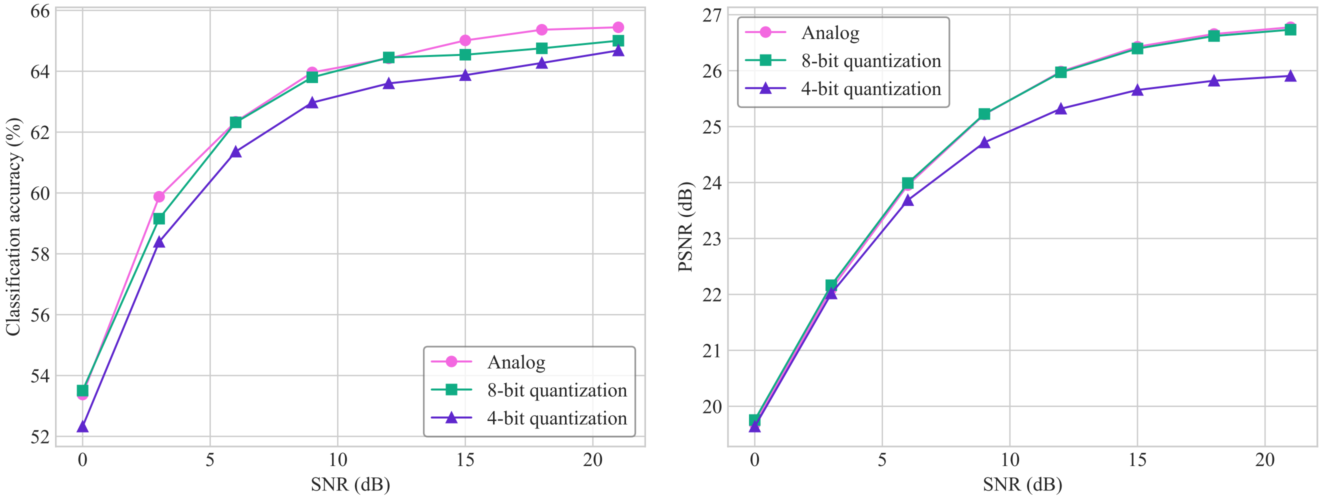
Fig. 12(b) and 12(c) respectively demonstrate the perfor-



(a) Constellation points before quantization.

(b) Constellation points after 4-bit quantization.

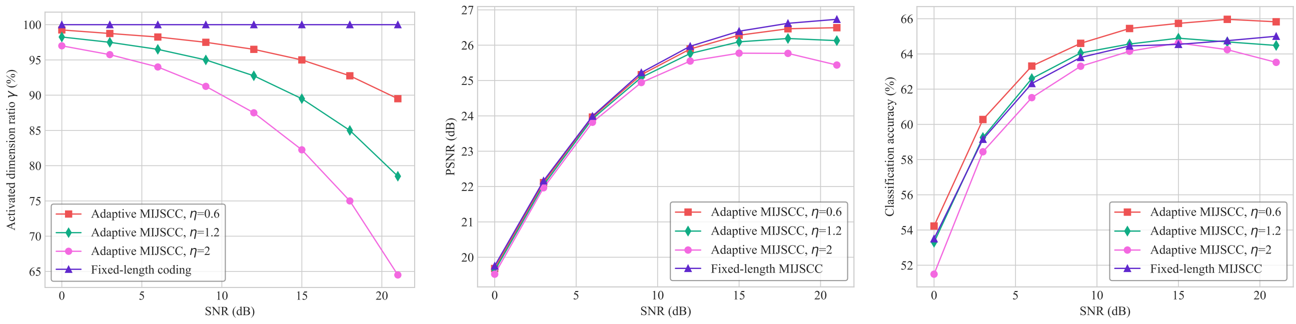
Fig. 10. The distribution of constellation points before and after quantization.



(a) The classification accuracy w.r.t. SNR.

(b) The PSNR w.r.t. SNR.

Fig. 11. The performance of classification and reconstruction tasks w.r.t. SNR under different quantization bits.



(a) Ratio of activated dimensions

(b) PSNR

(c) Classification accuracy

Fig. 12. Ratio of activated dimensions in the semantic representation and the performance of reconstruction and classification tasks w.r.t. SNR under different pruning thresholds η compared with fix-length MIJSCC.

mance of the A-MIJSCC framework in image reconstruction and classification tasks under different pruning thresholds and channel SNRs. To intuitively reflect the impact of adaptive channel mask on task performance, the semantic enhancement module is temporarily omitted. It can be observed that the proposed A-MIJSCC framework can effectively reduce the

TABLE III
PERFORMANCE COMPARISON OF A-MIJSCC ON RECONSTRUCTION TASKS

Methods		Performance of reconstruction tasks (PSNR/dB)							
		0dB	3dB	6dB	9dB	12dB	15dB	18dB	21dB
Adaptive MIJSCC	$\eta = 0.6$	19.68	22.11	23.97	25.17	25.89	26.28	26.46	26.50
	$\eta = 1.2$	19.62	22.04	23.92	25.09	25.76	26.10	26.19	26.13
	$\eta = 2$	19.52	21.97	23.82	24.94	25.55	25.77	25.77	25.44
Adaptive MIJSCC with attention	$\eta = 0.6$	20.17	22.20	23.87	25.09	25.90	26.37	26.61	26.72
	$\eta = 1.2$	20.13	22.13	23.81	25.00	25.83	26.26	26.48	26.54
	$\eta = 2$	20.07	22.06	23.72	24.91	25.70	26.06	26.18	26.16

TABLE IV
PERFORMANCE COMPARISON OF A-MIJSCC ON CLASSIFICATION TASKS

Methods		Performance of classification tasks (ACC/%)							
		0dB	3dB	6dB	9dB	12dB	15dB	18dB	21dB
Adaptive MIJSCC	$\eta = 0.6$	54.23	60.28	63.31	64.60	65.44	65.74	65.97	65.83
	$\eta = 1.2$	53.31	59.25	62.60	64.05	64.56	64.89	64.68	64.48
	$\eta = 2$	51.49	58.44	61.52	63.30	64.15	64.62	64.24	63.52
Adaptive MIJSCC with attention	$\eta = 0.6$	53.88	60.11	63.71	65.28	65.57	65.85	65.78	65.89
	$\eta = 1.2$	53.93	59.17	62.93	64.39	65.53	65.63	65.41	65.50
	$\eta = 2$	52.06	58.85	62.15	64.13	64.82	64.79	64.63	64.31

semantic transmission overhead according to the channel SNR while maintaining the stability of task performance. Specifically, when $\eta = 0.6$, the proposed A-MIJSCC tends to reserve more dimensions, without significantly affecting the robustness of multi-task semantic communication. For image classification tasks, pruning redundant dimensions helps reduce the interference from irrelevant information, thus improving the classification accuracy. When $\eta = 2$, the A-MIJSCC tends to prune less critical dimensions in exchange for lower semantic transmission overhead. However, when the SNR continues to increase (e.g., SNR = 21 dB), both task performance begins to degrade. This is because the A-MIJSCC framework tends to generate a stricter channel mask under better channel conditions, which filters out useful semantic information.

In order to compensate for the task performance degradation caused by the strict channel mask under high SNR, a semantic enhancement module based on the attention mechanism is introduced in the A-MIJSCC framework. The attention weights $\boldsymbol{w}$ of each dimension are gradually updated during the joint training process to enhance the masked semantic representation. Tables IV and III respectively show the performance of the A-MIJSCC framework on image reconstruction and classification tasks before and after semantic enhancement. Fig. 13(b) and 13(a) show the performance-overhead ratio of image classification tasks and restoration tasks under different semantic communication schemes and SNRs, respectively.

We can see that the A-MIJSCC with semantic enhancement module further improves the performance-overhead ratio of image classification and reconstruction tasks, which indicates that the attention mechanism can enhance the representation ability of semantics under consistent dimensions, thereby improving the transmission efficiency and robustness of the multi-task semantic communication system in scenarios with dynamic channel conditions and limited wireless resources.

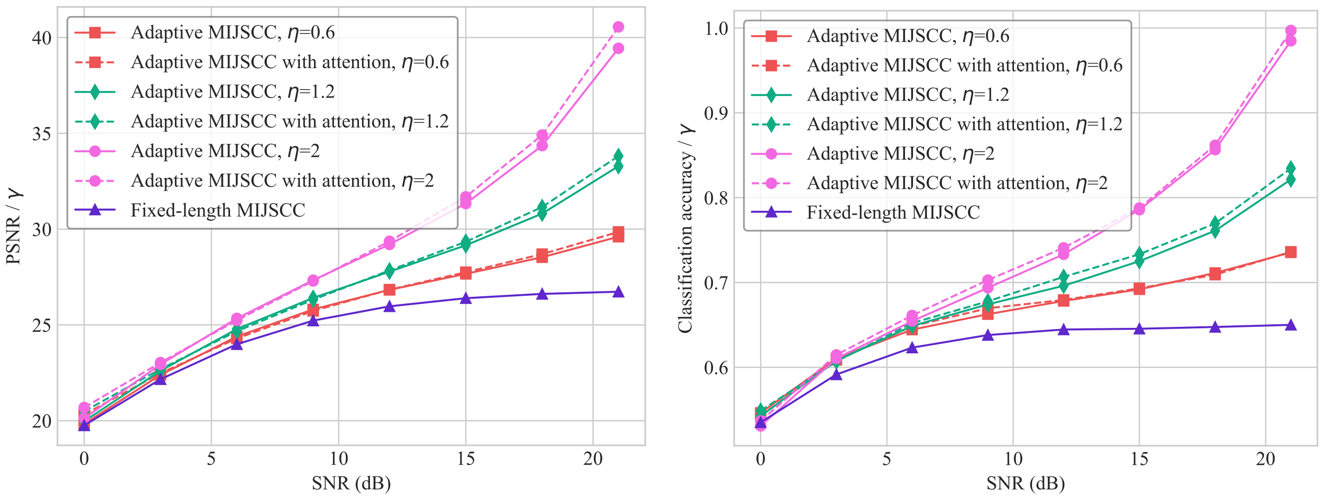
VI. CONCLUSION

In this paper, a semi-supervised multi-task semantic communication system based on MIJSCC framework is designed, where semantic representations are learned via global and local MI maximization to obtain discriminate features for direct classification, and the end-to-end target is learned via MSE minimization to derive pixel-level information for reconstruction. To seamlessly implement the proposed MIJSCC framework in practical communication systems with limited RF capabilities, the continuous semantic representations are quantized and mapped into discrete symbols with larger constellation point distance for amplitude and phase identification. To support adaptive transmission under changing channel conditions, an A-MIJSCC framework is introduced to consecutively activate important dimensions in semantic representations according to the channel noise. Simulation results demonstrate that compared with benchmark frameworks including single-task JSCC with MSE and conventional SSCC, the proposed MIJSCC supports multiple tasks with single transmission by extracting task-agnostic semantics. Moreover, the A-MIJSCC is verified to facilitate adaptive semantic transmission under varying channel environments with lower transmission overhead, while maintaining similar task performance with the MIJSCC.

APPENDIX A
RELATIONSHIP BETWEEN JENSEN-SHANNON DIVERGENCE AND MI

Denote $p(x)$ and $p(y)$ as two marginal distributions, $P = p(x, y)$ and $Q = p(x)p(y)$ stand for the joint and the product of marginals, respectively. The JS divergence between P and Q is the variant of their KL divergence, which is defined as

$$D_{JS}(P\|Q) = \frac{1}{2}D_{KL}\left(P\|\frac{P+Q}{2}\right) + \frac{1}{2}D_{KL}\left(Q\|\frac{P+Q}{2}\right). \tag{A.1}$$



(a) Performance-overhead ratio on reconstruction tasks

(b) Performance-overhead ratio on classification tasks

Fig. 13. The performance-overhead ratio comparison between the A-MIJSCC w/o attention and the A-MIJSCC with attention w.r.t. SNR under different pruning thresholds η.

Thus, we consider a mixture distribution $m(x,y)=\frac{P+Q}{2}=\frac{1}{2}(p(x,y)+p(x)p(y))$, which follows $m(x)=p(x)$, $m(y)=p(y)$, and $m(y|x)=\frac{1}{2}(p(y|x)+p(y))$. With constants discarded, (A.1) is rewritten in (A.2) at the top of the next page, where the term inside the expectation is a monotonically increasing convex function of $\frac{p(y|x)}{p(y)}$.

Since the pointwise MI is formulated as $\ln\frac{p(x,y)}{p(x)p(y)}=\ln\frac{p(y|x)}{p(y)}$, it can be verified that maximizing the JS divergence between P and Q is equivalent to maximizing the MI between x and y.

Appendix B
Mathematical Properties of the MLP Function

The noise encoder network $N_{enc}(\cdot)$ is composed of M linear layers, each followed by a Tanh activation. Therefore, the noise feature $\boldsymbol{g}$ can be expressed as a composition of M non-linear function as

$$\boldsymbol{g}=N_{enc}(\delta^2)=\boldsymbol{u}_M\circ\boldsymbol{u}_{M-1}\cdots\boldsymbol{u}_1(\delta^2), \tag{B.1}$$

where $\boldsymbol{u}_m$ denotes the m-th layer of the MLP network and is expressed by

$$\boldsymbol{u}_m(\boldsymbol{x})=\tanh\left(\boldsymbol{H}^{(m)}\boldsymbol{x}+\boldsymbol{b}^{(m)}\right). \tag{B.2}$$

$\boldsymbol{H}^{(m)}$ and $\boldsymbol{b}^{(m)}$ represent the weight and bias of the m-th layer. As stated in Section IV, each output g_j should be a non-negative increasing function to ensure its monotonicity with the noise variance. Therefore, with a fixed pruning threshold η, more dimensions will be activated under low-SNR circumstances. Thus, g_j should meet the subsequent requirements:

$$g_j\geq 0;\quad g_j'=\frac{\partial g_j}{\partial\delta^2}\geq 0. \tag{B.3}$$

Since g_j can be expressed as

$$g_j=u_{M,j}\circ\boldsymbol{u}_{M-1}\cdots\boldsymbol{u}_1(\delta^2), \tag{B.4}$$

where $u_{M,j}$ denotes the j-th output dimension of layer $\boldsymbol{u}_M$. Thus, the derivative of g_j can be calculated according to the chain rule:

$$g_j'=\boldsymbol{u}_{M,j}'\circ\boldsymbol{u}_{M-1}'\cdots\boldsymbol{u}_1'(\delta^2), \tag{B.5}$$

where $\boldsymbol{u}_m'$ is obtained as the Jacobian matrix of $\boldsymbol{u}_m$, and $\boldsymbol{u}_{M,j}'$ is the j-th row of $\boldsymbol{u}_M'$. Therefore, the derivation of (B.2) is obtained as[2]

$$\boldsymbol{u}_m'(\boldsymbol{x})=\text{diag}\left(1-\tanh\left(\boldsymbol{H}^{(m)}\boldsymbol{x}+\boldsymbol{b}^{(m)}\right)\right)\cdot\boldsymbol{H}^{(m)}. \tag{B.6}$$

Under the setting of $\boldsymbol{H}^{(m)}=\text{abs}(\widetilde{\boldsymbol{H}}^{(m)})$ and small bias, both constraints in (B.3) are satisfied, indicating that g_j is a non-negative increasing function of the noise variance[3].

References

[1] Z. Qin, L. Liang, Z. Wang, S. Jin, X. Tao, W. Tong, and G. Y. Li, "AI empowered wireless communications: From bits to semantics," *Proceedings of the IEEE*, vol. 112, no. 7, pp. 621–652, Jul. 2024.

[2] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Trans. Signal Process.*, vol. 69, pp. 2663–2675, 2021.

[3] P. Zhang, W. Xu, H. Gao, K. Niu, X. Xu, X. Qin, C. Yuan, Z. Qin, H. Zhao, J. Wei *et al.*, "Toward wisdom-evolutionary and primitive-concise 6G: A new paradigm of semantic communication networks," *Engineering*, vol. 8, pp. 60–73, 2022.

[4] W. Yang, H. Du, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. Shen, and C. Miao, "Semantic communications for future internet: Fundamentals, applications, and challenges," *IEEE Commun. Surv. Tutor.*, vol. 25, no. 1, pp. 213–250, 2023.

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

[6] J. Dai, S. Wang, K. Tan, Z. Si, X. Qin, K. Niu, and P. Zhang, "Nonlinear transform source-channel coding for semantic communications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 8, pp. 2300–2316, Jun. 2022.

[7] B. Xu, R. Meng, Y. Chen, X. Xu, C. Dong, and H. Sun, "Latent semantic diffusion-sased channel adaptive de-noising SemCom for future 6G systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2023, pp. 1229–1234.

[2] $\tanh(x)=\frac{e^x-e^{-x}}{e^x+e^{-x}}$, thus $\tanh'(x)=1-\tanh(x)$.

[3] $\widetilde{\boldsymbol{H}}^{(m)}$ denotes the actual weights of the m-th MLP layer, and $\text{abs}(\cdot)$ represents the elementwise absolute value function.

$$
\begin{aligned}
&D_{JS}(p(x,y)\|p(x)p(y)) \propto D_{KL}(p(x,y)\|m(x,y)) + D_{KL}(p(x)p(y)\|m(x,y)) \\
&\propto \mathbb{E}_{p(x,y)}\left[\log\frac{p(x,y)}{m(x,y)}\right] + \mathbb{E}_{p(x)p(y)}\left[\log\frac{p(x)p(y)}{m(x,y)}\right] \\
&\propto \mathbb{E}_{p(x)}\left[\mathbb{E}_{p(y|x)}\left[\log\frac{p(y|x)p(x)}{m(y|x)m(x)}\right] + \mathbb{E}_{p(y)}\left[\log\frac{p(y)p(x)}{m(y|x)m(x)}\right]\right] \\
&\propto \mathbb{E}_{p(x)}\left[\mathbb{E}_{p(y|x)}\left[\log\frac{p(y|x)}{p(y)} - \log\left(1+\frac{p(y|x)}{p(y)}\right)\right] + \mathbb{E}_{p(y)}\left[-\log\left(1+\frac{p(y|x)}{p(y)}\right)\right]\right] \\
&\propto \mathbb{E}_{p(x)}\left[\mathbb{E}_{p(y|x)}\left[\log\frac{p(y|x)}{p(y)}\right] - 2\mathbb{E}_{m(y|x)}\left[\log\left(1+\frac{p(y|x)}{p(y)}\right)\right]\right] \\
&\propto \mathbb{E}_{p(x)}\left[\mathbb{E}_{p(y|x)}\left[\log\frac{p(y|x)}{p(y)} - 2\frac{m(y|x)}{p(y|x)}\log\left(1+\frac{p(y|x)}{p(y)}\right)\right]\right] \\
&\propto \mathbb{E}_{p(x)}\left[\mathbb{E}_{p(y|x)}\left[\log\frac{p(y|x)}{p(y)} - \left(1+\frac{p(y)}{p(y|x)}\right)\log\left(1+\frac{p(y|x)}{p(y)}\right)\right]\right],
\end{aligned} \tag{A.2}
$$

[8] Z. Weng, Z. Qin, X. Tao, C. Pan, G. Liu, and G. Y. Li, "Deep learning enabled semantic communications with speech recognition and synthesis," *IEEE Trans. Wirel. Commun.*, vol. 22, no. 9, pp. 6227–6240, Feb. 2023.

[9] 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, Jul. 2022.

[10] X. Kang, B. Song, J. Guo, Z. Qin, and F. R. Yu, "Task-oriented image transmission for scene classification in unmanned aerial systems," *IEEE Trans. Commun.*, vol. 70, no. 8, pp. 5181–5192, Aug. 2022.

[11] Y. Wang, S. Han, X. Xu, H. Liang, R. Meng, C. Dong, and P. Zhang, "Feature importance-aware task-oriented semantic transmission and optimization," *IEEE Trans. Cogn. Commun. Netw.*, vol. 10, no. 4, pp. 1175–1189, Aug. 2024.

[12] Y. Wang, W. Ni, W. Yi, X. Xu, P. Zhang, and A. Nallanathan, "Federated contrastive learning for personalized semantic communication," *IEEE Commun. Lett.*, vol. 28, no. 8, pp. 1875–1879, Aug. 2024.

[13] C. Wang, X. Cao, L. Guo, and Z. Shi, "DualMatch: Robust semi-supervised learning with dual-level interaction," in *Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Springer, 2023, pp. 102–119.

[14] T. M. Getu, G. Kaddoum, and M. Bennis, "Semantic communication: A survey on research landscape, challenges, and future directions," *Proceedings of the IEEE*, vol. 112, no. 11, pp. 1649–1685, Nov. 2024.

[15] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *Int. Conf. Mach. Learn.* PMLR, 2018, pp. 531–540.

[16] S. Kleinegesse and M. U. Gutmann, "Bayesian experimental design for implicit models by mutual information neural estimation," in *Int. Conf. Mach. Learn.* PMLR, 2020, pp. 5316–5326.

[17] G. Zhang, Q. Hu, Z. Qin, Y. Cai, G. Yu, and X. Tao, "A unified multi-task semantic communication system for multimodal data," *IEEE Trans. Commun.*, vol. 72, no. 7, pp. 4101–4116, Feb. 2024.

[18] Z. Tian, H. Vo, C. Zhang, G. Min, and S. Yu, "An asynchronous multi-task semantic communication method," *IEEE Netw.*, Oct. 2023.

[19] Z. Lyu, G. Zhu, J. Xu, B. Ai, and S. Cui, "Semantic communications for image recovery and classification via deep joint source and channel coding," *IEEE Trans. Wirel. Commun.*, vol. 23, no. 8, pp. 8388–8404, Jan. 2024.

[20] W. Yuan, J. Ren, C. Wang, R. Zhang, J. Wei, D. I. Kim, and S. Cui, "Generative semantic communication for joint image transmission and segmentation," *arXiv preprint arXiv:2411.18005*, 2024.

[21] 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 J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 170–185, Nov. 2022.

[22] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Int. Conf. Mach. Learn.* PMLR, 2020, pp. 1597–1607.

[23] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 3722–3731.

[24] Y. Yu, K. H. R. Chan, C. You, C. Song, and Y. Ma, "Learning diverse and discriminative representations via the principle of maximal coding rate reduction," *Advances Neural Inf. Process. Syst.*, vol. 33, pp. 9422–9434, 2020.

[25] A. Waswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[26] A. H. Razlighi, C. Bockelmann, and A. Dekorsy, "Semantic communication for cooperative multi-task processing over wireless networks," *IEEE Wireless Commun. Lett.*, vol. 13, no. 10, pp. 2867–2871, 2024.

[27] 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.*, Jun. 2023.

[28] F. Liu, Z. Sun, Y. Yang, C. Guo, and S. Zhao, "Rate adaptable multi-task-oriented semantic communication: An extended rate-distortion theory based scheme," *IEEE Internet Things J.*, vol. 11, no. 9, pp. 15 557–15 570, 2024.

[29] H. Xie, Z. Qin, and G. Y. Li, "Semantic communication with memory," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 8, pp. 2658–2669, Jun. 2023.

[30] S. Vembu, S. Verdu, and Y. Steinberg, "The source-channel separation theorem revisited," *IEEE Trans. Inf. Theory*, vol. 41, no. 1, pp. 44–54, 1995.

[31] J. Dai, P. Zhang, K. Niu, S. Wang, Z. Si, and X. Qin, "Semantic coded transmission: Architecture, methodology, and challenges," *arXiv preprint*, 2021.

[32] B. Zhu, J. Wang, L. He, and J. Song, "Joint transceiver optimization for wireless communication PHY using neural network," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1364–1373, Mar. 2019.

[33] H. Xie and Z. Qin, "A lite distributed semantic communication system for internet of things," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 142–153, 2020.

[34] S. Nowozin, B. Cseke, R. Tomioka, and GAN, "Training generative neural samplers using variational divergence minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 271–279.

[35] Q. Yang, H.-X. Yu, A. Wu, and W.-S. Zheng, "Patch-based discriminative feature learning for unsupervised person re-identification," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 3633–3642.

[36] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[37] Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang, "Dynamic neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7436–7456, Oct. 2021.

[38] Z. Wu, T. Nagarajan, A. Kumar, S. Rennie, L. S. Davis, K. Grauman, and R. Feris, "Blockdrop: Dynamic inference paths in residual networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2018, pp. 8817–8826.

[39] Z. Chen, Y. Li, S. Bengio, and S. Si, "You look twice: Gaternet for dynamic filter selection in CNNs," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Oct. 2019, pp. 9172–9180.