Contrastive Learning based Semantic Communications

Shunpu Tang, Qianqian Yang, Lisheng Fan, Xianfu Lei, Arumugam Nallanathan, *Fellow, IEEE*, and George K. Karagiannidis, *Fellow, IEEE*

***Abstract*—Recently, there has been a growing interest in learning-based semantic communication because it can prioritize the preservation of meaningful semantic information over the accuracy of the transmitted symbols, resulting in improved communication efficiency. However, existing learning-based approaches still face limitations in defining semantic level loss and often struggle to find a good trade-off between preserving semantic information and preserving intricate details. In addition, the existing semantic communication approaches cannot effectively train semantic encoders and decoders without the support of downstream models. To address these limitations, this paper proposes a contrastive learning (CL)-based semantic communication system. First, inspired by practical observations, we introduce the concept of semantic contrastive loss and propose a semantic contrastive coding (SemCC) approach that treats data corruption during transmission as a form of data augmentation within the CL framework. Moreover, we propose a semantic re-encoding (SemRE) operation, which uses a duplicate of the semantic encoder deployed at the receiver to guide the entire training process when the downstream model is inaccessible. Further, we design the training procedure for SemCC and SemRE approaches, respectively, to balance the semantic information and intricate details. Finally, simulations are performed to demonstrate the superiority of the proposed approaches over competing approaches. In particular, our approaches achieve a significant accuracy improvement of up to 53% on the CIFAR-10 dataset with a bandwidth compression ratio of 1/24, and also obtain comparable image reconstruction quality as the bandwidth compression ratio is improved.**

***Index Terms*—Semantic communication, contrastive learning, joint source-channel coding, image transmission.**

I. Introduction

A. Backgrounds

The primary goal of digital communication system has been to reliably transmit bits through noisy channels, which is typically categorized as the technical level of communication. The classical information theory proposed by Shannon [2], provided the fundamental principle for achieving this goal, which introduced the concept of channel capacity to provide a theoretical upper bound on the data rate that ensures error-free transmission. Researchers have made considerable efforts to approximate the channel capacity by developing the advanced channel coding techniques such as *low-density parity check* (LDPC) [3] and *polar* code [4] in the 5G New Radio (NR). While modern digital communication systems based on these approaches have achieved remarkable progress, they do not explicitly consider the underlying meaning of the transmitted data and therefore treat all bits as equal, which could potentially lead to challenges in future applications.

In the upcoming Beyond 5G (B5G) and 6G networks, a large number of Internet of Things (IoT) devices will be deployed and various types of multimedia data will be transmitted to support novel applications such as smart cities, automated driving, virtual reality (VR), and augmented reality (AR) [5]–[7]. However, the large number of connections, data transfer requirements, and ultra-low latency demands will place a significant burden on the network infrastructure. To address these challenges, researchers are shifting their focus to improving communication efficiency within the constraints of available channel capacity. In this direction, the importance and meaning of the transmitted data are taken into account in the system design, and the concept of *semantic communication* has attracted increasing attention [8]–[11]. Semantic communication works under the semantic and effectiveness level of communication, which aims to prioritize the preservation of meaningful semantic information over the accuracy of transmitted symbols, leading to improved communication efficiency by transmitting only necessary information relevant to the specific task at the receiver. These characteristics of semantic communication can also better meet the requirements of the aforementioned applications in the context of B5G and 6G networks.

However, a major challenge in semantic communication is how to effectively extract semantic information at the transmitter while accurately reconstructing it at the receiver under constrained communication conditions. While recent efforts have leveraged advanced deep learning technologies in semantic communication systems, there are still some issues that need to be addressed, which are discussed below.

1) How to evaluate the loss of semantic level during the training process: While the performance of a semantic communication system can be effectively evaluated by the downstream task, it is crucial to note that the direct use of the loss functions of the downstream task, such as the

This paper was presented in part at the IEEE 98th Vehicular Technology Conference (VTC-Fall), Hong Kong, Oct. 2023 [1].

S. Tang is with the School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou, China, and also with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China (e-mail: tangshunpu@e.gzhu.edu.cn).

Q. Yang is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China, and with the Key Laboratory of Collaborative Sensing and Autonomous Unmanned Systems of Zhejiang Province, Zhejiang University, Hangzhou, China (e-mail: qianqianyang20@zju.edu.cn).

L. Fan is with the School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou, China (e-mail: lsfan@gzhu.edu.cn).

X. Lei is with the Institute of Mobile Communications, Southwest Jiaotong University, Chengdu, China (e-mail: xflei@home.swjtu.edu.cn).

A. Nallanathan is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London, U.K (e-mail: a.nallanathan@qmul.ac.uk).

G. K. Karagiannidis is with Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Greece and also with Artificial Intelligence and Cyber Systems Research Center, Lebanese American University (LAU), Lebanon (geokarag@auth.gr).

The corresponding author of this paper is L. Fan.

cross-entropy loss [12], may not fully match the intrinsic characteristics of semantic information, and may not guide the training of semantic encoder and decoder well, which could potentially lead to the degradation of effectiveness and robustness. Since semantic information is closely related to the meaningful content and contextual aspects of the data, these may be overlooked by the specific labels predicted by the downstream network and the corresponding use of cross-entropy loss. Therefore, it is necessary to integrate the inherent properties of semantic information into the semantic level loss.

2) How to train semantic encoders and decoders without the help of a pre-trained downstream network: The semantic communication system faces significant challenges in scenarios where the receiver is prohibited from accessing not only the weights, but also the architecture of the downstream model (i.e. the pretrained downstream is black-box). This limitation is primarily due to encryption and other security measures, and it severely hampers the simultaneous training of the semantic encoder and decoder in [12]. One possible solution to overcome this limitation is to retrain a deep neural network (DNN) that can act as a guide for training the semantic encoder and decoder. However, it is important to note that this solution may introduce additional system cost and complexity, which should be carefully considered in a practical system. Moreover, if the chosen architecture of the DNN differs from that of the pretrained downstream model, performance degradation may occur, which poses a challenge to this solution.

3) How to strike a good balance between preserving semantic information and preserving intricate details: For a well-designed semantic communication system, striking a good balance between preserving important semantic information and retaining intricate fine-grained details is essential. When the bandwidth resource is limited, the system may prefer to transmit the semantic information over intricate fine-grained details, and should gradually increase the level of detail as the bandwidth availability improves. However, existing approaches overemphasize semantic information, resulting in a loss of detail when the bandwidth resource is not limited.

B. Contributions

To address the above questions, we introduce the contrastive learning (CL)-based semantic communication system. We start with the first question by taking into account the inherent properties of semantic information. In particular, when we compare two unrelated entities, it becomes clear that their semantic information has significant differences. In contrast, the semantic information may not change much when a data augmentation operation is performed on the original entity. These observations can motivate us to incorporate the popular CL approach into semantic communication, since the principles of CL align closely with those of the semantic communication system and CL has demonstrated significant achievements across various domains, including computer vision [13]–[17], natural language processing (NLP) [18], and multimodal applications [19]–[21]. The works in [22], [23] first introduced the concept into semantic communication systems to help extract useful semantic information. However, these works still face limitations in evaluating the semantic level loss when these systems work in a noisy channel since they ignore the fact that in an ideal semantic communication system, the semantic information at the receiver should remain basically unchanged from its state before transmission, while still being distinguishable from those unrelated entities.

To this end, we propose the *semantic contrastive coding* (SemCC) approach, which deeply explores the introduction of CL process to the semantic communication system. Specifically, we replace the conventional data augmentation procedure with a wireless transmission process. This change is based on the idea that the distortion caused by the noise and fading characteristics of the wireless channel during transmission can be considered as a form of data augmentation. Therefore, we design the semantic-level loss for SemCC to ensure that the semantic distance between the original and reconstructed images is small enough while maintaining a considerable semantic distance between the reconstructed and irrelevant images for better discrimination in the downstream task.

To address the second question, we introduce the concept of *semantic re-encoding* (SemRE), which is inspired by the information bottleneck theory that only useful semantic information is allowed to pass through the semantic encoder. When the semantic information is initially used for data reconstruction, the reconstructed data, when passed through the semantic encoder again, should ideally acquire the same semantic information as the initial one. Therefore, the key design in this SemRE is that we deploy a semantic encoder at the receiver, which is copied from the one in the transmitter, and use it to guide the training of the semantic encoder and decoder.

Furthermore, we introduce training strategies to address the third question in the context of our semantic communication system. In particular, inspired by the approach presented in [12], we introduce a loss function that includes both observation loss and semantic level loss, with a hyper-parameter that controls the trade-off between these two components. We also design a fine-tuning approach for situations with an available downstream model, which aims to improve the inference performance of the downstream task.

Finally, simulations are performed to demonstrate the superiority of the proposed approaches over competing approaches. Without losing generality, we follow the concept of semantic communication in [9], [10] and focus on the specific tasks of image reconstruction and image classification at the receiver like [12]. In this context, we no longer pay attention to the typical metric of the technical level of communication such as bit error ratio (BER) and symbol error rate (SER). Instead, we evaluate the system performance based on the effectiveness of the received semantic information, using intrinsic task-related metrics such as image quality and inference accuracy. We compare the proposed approaches with the advanced semantic communication system in [12], [24], as well as the classical digital communication system. Simulation results show that the proposed approaches can achieve leading accuracy performance in the downstream task under a range of bandwidth compression ratios, and demonstrate remarkable adaptability

to both AWGN and Rayleigh fading channels with different noise levels, and also make a good trade-off between the image reconstruction quality and inference performance.

The main contributions of this paper are summarized as follows,

- We propose the SemCC approach, which integrates the concept of CL into semantic communication. By utilizing wireless transmission as a form of data augmentation in CL, SemCC ensures minimal semantic distance between original and reconstructed images while maintaining discrimination against irrelevant images.
- We introduce the SemRE operation, which uses a duplicate of the semantic encoder deployed at the receiver to guide the entire training process when the downstream model is inaccessible.
- We design the training procedure for SemCC and SemRE approaches, respectively, to balance the semantic information and intricate details.
- We conduct simulations to demonstrate the superiority of our approaches over existing methods in terms of inference accuracy, across various bandwidth compression ratios and channel conditions, and also obtain comparable image reconstruction quality as the bandwidth compression ratio is improved.

C. Structure

The rest of this paper is organized as follows. In Section II, we provide an overview of related work on semantic communication, covering its theoretical foundations and practical applications. Section III introduces the system model. In Section IV, we present the implementation details of the proposed CL-based semantic communication system. Section V describes the proposed SemRE operation and its training strategy. Simulation results are provided in Section VI. Finally, we conclude this work in Section VII.

II. Related works

A. Basics on Theory of Semantic Communication

The authors in [25] defined the semantic information carried by a sentence in terms of logical probability during transmission. Building on Shannon and Weaver's theory, the authors in [8] introduced the concept of a semantic channel and proposed a model-theoretic approach to reliable semantic communication. In [26], the communication scenario between two intelligent beings was discussed, and the theoretical formulation of the goals of semantic communication was presented to demonstrate the necessity of semantic communication. Subsequently, G. Guler *et al.* investigated a semantic communication framework by considering the meanings of transmitted codewords over a noisy channel, and optimized the end-to-end average semantic error using a Bayesian approach [27]. Based on the aforementioned works, the concept of semantic information theory [28]–[32] has attracted increasing research interest in recent years, providing the theoretical foundation for the development of semantic communication in various directions.

B. Transmission Strategy of Semantic Communication

With the rapid growth of deep learning technology, researchers have started to explore the deployment of semantic communication system with the help of powerful semantic extraction provided by deep learning. In this direction, the authors in [24], [33], [34] proposed a deep learning based joint source-channel coding (DeepJSCC) for image data, where the encoder and decoder were designed based on autoencoder and jointly optimized for semantic information transmission to achieve a good image reconstruction quality. Then, the works in [35]–[37] extended DeepJSCC to different channel conditions and improved the image reconstruction quality under noisy channels. In addition, motivated by generative models, some works incorporated generative adversarial networks (GANs) to further reduce bandwidth consumption. For example, the authors in [38], [39] applied the GAN inversion methods [40] to regenerate the image at the receiver, which leads to significant improvements in communication efficiency. In [41], a joint semantic encoding-modulation system has been explored to facilitate the deployment of semantic communication in practical networks.

For text and speech data, the work in [42] extended DeepJSCC to reduce the BER while preserving the semantic information in sentences. Leveraging the transformer architecture, the authors in [43], [44] proposed a semantic communication approach for text, achieving a high semantic similarity between transmitted and received sentences. Guo *et al.* [45] explored the ability of pre-trained large language model (LLM) such as ChatGPT to extract semantic information by introducing a cross-layer manager, thus achieving lower semantic loss under limited bandwidth. In addition, the work in [46], [47] explored the semantic communication system for speech signals to reduce perceptual distortion.

C. Application of Semantic Communication

Not limited to data reconstruction at the receiver, semantic communication has been applied to various application scenarios to support the downstream task. In the work of [48]–[50], semantic communication is used to transmit the output of the mid-layer of a neural network (NN) to reduce the inference latency with the help of an edge server. The authors in [12] proposed a collaborative training framework for semantic communication, where users could train their semantic encoder to improve the performance of downstream vision inference tasks under limited bandwidth. Moreover, in [51], [52], the authors used semantic communication to support the Visual Question Answering (VQA) task by extracting and transferring the semantic information from the correlated multimodal data. The authors in [53] applied semantic communication in the UAV network, which enables efficient on-the-fly scene classification. Yang *et al.* [54] also introduced semantic communication into the complex vehicular networks, and jointly optimized the energy efficiency and semantic transmission reliability to support green V2V communication. The authos in [55] integrated semantic communication in mobie edge computing (MEC) network to support the efficient communication between the

edge server and the user equipment (UE), which helps reduce the energy consumption.

III. System model

This paper investigates a semantic communication system, where an NN-based semantic encoder and decoder are deployed in the transmitter and receiver, respectively. More specifically, we focus on the wireless image transmission in this paper, and use $\boldsymbol{x} \in \mathbb{R}^{n_c \times n_h \times n_w}$ to denote the transmitted source image, where n_c, n_h, and n_w correspond to the number of channels, height, and width of the image, respectively. To simplify, let $n = n_c \times n_h \times n_w$ stand for the input dimension of $\boldsymbol{x}$.

The transmission process begins with the semantic encoding, which is used to extract the semantic information of $\boldsymbol{x}$ and directly realize the non-linear mapping from semantic information into the k-dim complex-valued vector $\tilde{\boldsymbol{s}} \in \mathbb{C}^k$, given by

$$\tilde{\boldsymbol{s}} = \mathcal{E}_{\theta_1}(\boldsymbol{x}), \tag{1}$$

where $\mathcal{E}_{\theta_1}(\cdot)$ represents the semantic encoding operation with parameter θ_1. At this stage, it is important to consider the relationship between the output dimension k and the input dimension n in the context of the bandwidth constraint. Typically, $k < n$ should be satisfied to the bandwidth constraint, where k/n is referred to as the bandwidth compression ratio. In particular, a large bandwidth compression ratio indicates a favorable communication condition, while a small one indicates a limited use of bandwidth. In addition, a power normalization layer [24] is used at the end of the semantic coding network to satisfy the average power constraint of P at the transmitter, given by

$$\boldsymbol{s} = \sqrt{kP}\frac{\tilde{\boldsymbol{s}}}{\sqrt{\tilde{\boldsymbol{s}}^* \tilde{\boldsymbol{s}}}}, \tag{2}$$

where $\boldsymbol{s}$ is the channel input signals that meets the power constraint, and * denotes the conjugate transpose. Next, $\boldsymbol{s}$ is transmitted over the noisy channel, where both the additive Gaussian white noise (AWGN) channel and Rayleigh fading channels are considered in this paper. Specifically, for the AWGN channel, the received signals can be expressed as

$$\hat{\boldsymbol{s}} = \boldsymbol{s} + \boldsymbol{\epsilon}, \tag{3}$$

where $\hat{\boldsymbol{s}}$ is the received signals, and $\boldsymbol{\epsilon} \in \mathcal{CN}(0, \sigma^2 \boldsymbol{I})$ denotes the additional noise sample. In the case of Rayleigh fading channels, the received signals $\hat{\boldsymbol{s}}$ is given by

$$\hat{\boldsymbol{s}} = \boldsymbol{H} \cdot \boldsymbol{s} + \boldsymbol{\epsilon}, \tag{4}$$

where $\boldsymbol{H}$ is the channel parameter and we assume that $\boldsymbol{H}$ can be perfectly estimated through some pilot signals.

At the receiver, the semantic decoder will be used to reconstruct the original image $\hat{\boldsymbol{x}} \in \mathbb{R}^{n_c \times n_h \times n_w}$ from the corrupted $\hat{\boldsymbol{s}}$ according to

$$\hat{\boldsymbol{x}} = \mathcal{D}_{\theta_2}(\hat{\boldsymbol{s}}), \tag{5}$$

where $\mathcal{D}_{\theta_2}(\cdot)$ denotes the semantic decoding operation parameterized by θ_2. It is important to highlight that this semantic decoding operation aims to recapture the semantic information from the noisy signals and reconstruct the image $\hat{\boldsymbol{x}}$.

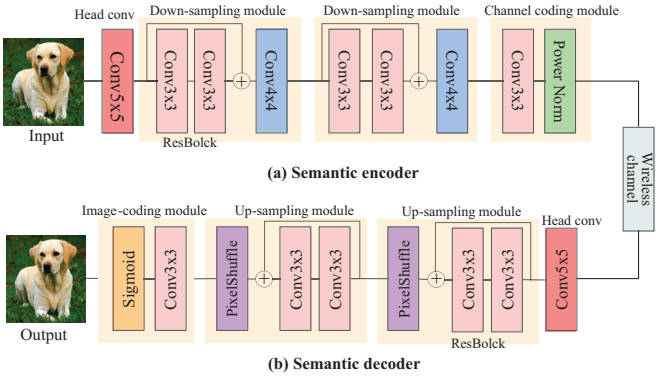

Fig. 1: Network architecture of the semantic encoder and decoder.

Subsequently, $\hat{\boldsymbol{x}}$ will be used to exert the downstream task and obtain the inference results through the following process

$$\boldsymbol{f}_{\boldsymbol{x}} = \mathcal{F}^b_{\phi_1}(\hat{\boldsymbol{x}}), \tag{6}$$

where $\mathcal{F}^b_{\phi_1}(\cdot)$ characterized by parameter ϕ_1 denotes the feature extraction operation performed by the backbone of the pretrained downstream model, and $\boldsymbol{f}_{\boldsymbol{x}} = \{\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)}, \cdots \boldsymbol{f}^{(C)}\}$ is the output feature map with C channels. The inference result $\hat{\boldsymbol{y}}$ can be obtained by passing $\boldsymbol{f}_{\boldsymbol{x}}$ to the classifier $\mathcal{F}^{cls}_{\phi_2}(\cdot)$ with parameter ϕ_2, which can be expressed as

$$\hat{\boldsymbol{y}} = \mathcal{F}^{cls}_{\phi_2}(\boldsymbol{f}_x). \tag{7}$$

From the above description, we can see that the semantic encoder and decoder play a key role in semantic communication. Moreover, preserving the semantic information in the reconstructed image is crucial for the inference performance, especially when the channel bandwidth is limited. Therefore, the architecture and training procedure of the semantic encoder and decoder require careful design.

Next, we will introduce the proposed CL-based semantic communication framework. Specifically, we will first present the architecture of the semantic encoder and decoder, and then provide the details of SemCC and the associated training procedure.

IV. CL based semantic communication

A. Architecture of Semantic Encoder and Decoder

The backbone of the pre-trained downstream model may not provide sufficient information to aid in reconstructing the image and can not mitigate the effect of channel noise. As a result, it cannot be utilized as a semantic encoder and decoder. Therefore, we train an extra semantic encoder and semantic decoder in this work. The architecture of the proposed semantic encoder and decoder is presented in Fig. 1. The semantic encoder consists of a 5×5 head convolution, two downsampling modules, and a channel coding module. Each downsampling module contains a basic block in ResNet [56] (we call it ResBolck) to capture the spatial feature of the image, and a 4×4 convolution with stride 2 to downsample the

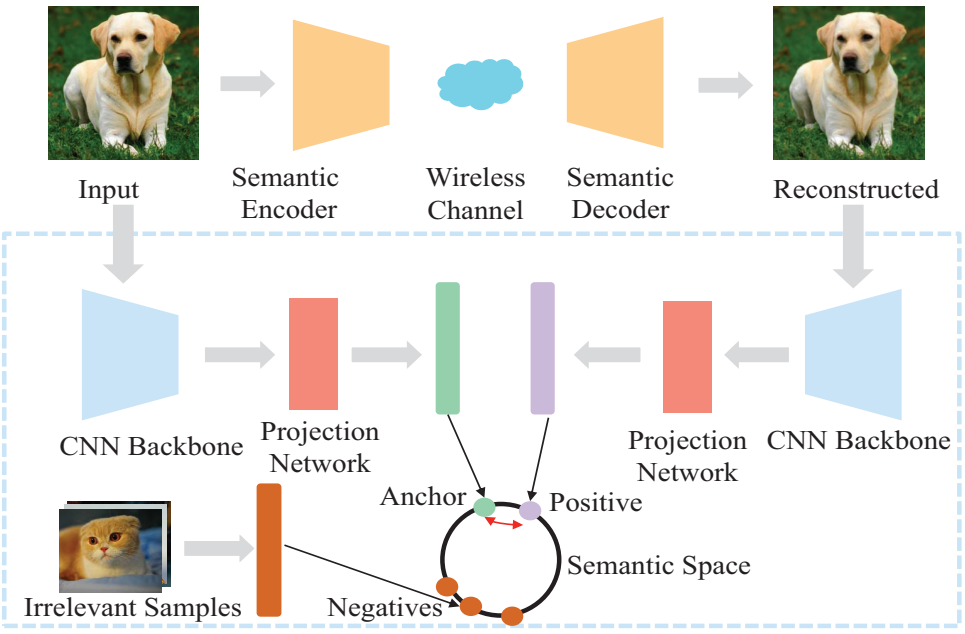

Fig. 2: Illustration of the proposed SemCC.

image. The channel coding module is used to mitigate channel corruption and output the k-dim complex-valued channel input that satisfies the bandwidth and power constraints.

Furthermore, we adopt a symmetric architecture in the decoder, which consists of a 5×5 head convolution, two upsampling modules, and a recoding module. In the upsampling module, ResBlocks are also used as in the encoder, and we adopt the pixel shuffle technology [57] to upsample the image, as it can provide a more efficient computing paradigm and better reconstruction performance compared to the transposed convolution used in [24]. The image coding module consists of a 3×3 convolution followed by the sigmoid activated function to produce the reconstructed image. Notably, the batch normalization and parametric rectified linear unit (PReLU) activated function are followed with all convolutions if not specified.

B. Semantic Contrastive Coding

The details of the proposed SemCC are shown in Fig. 2. The process begins with the semantic encoding and decoding for a typical image $\boldsymbol{x}$ in a training batch $\mathcal{B}$, where we can obtain the reconstructed $\hat{\boldsymbol{x}}$. We use this process to replace the conventional data augmentation procedure, and we regard $\hat{\boldsymbol{x}}$ as an augmented sample of $\boldsymbol{x}$. The backbone of pretrained downstream mode $\mathcal{F}^b_{\phi_1}(\cdot)$ is applied to $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$, which generates the feature maps $\boldsymbol{f}_{\boldsymbol{x}} = \mathcal{F}^b_{\phi_1}(\boldsymbol{x})$ and $\boldsymbol{f}_{\hat{\boldsymbol{x}}} = \mathcal{F}^b_{\phi_1}(\hat{\boldsymbol{x}})$, respectively. Next, a fully connected projection network $\mathcal{P}_\psi(\cdot)$ with learnable parameter ψ followed by a normalization operation maps the features into the semantic space defined as a hypersphere[1], where samples are represented as tensors based on their semantic information in this sapce. As mentioned earlier, samples with similar semantic information are close together, while those with different semantic information are farther apart in this space.

During the training stage, $\mathcal{P}_\psi(\cdot)$ can be updated to enhance the understanding of features, thereby learning the mapping from features to semantics. Specifically, the projected results of $\boldsymbol{f}_{\boldsymbol{x}}$ and $\boldsymbol{f}_{\hat{\boldsymbol{x}}}$ can be represented as $\boldsymbol{q}_{\boldsymbol{x}} = \mathcal{P}_\psi(\boldsymbol{f}_{\boldsymbol{x}})$ and $\boldsymbol{q}_{\hat{\boldsymbol{x}}} = \mathcal{P}_\psi(\boldsymbol{f}_{\hat{\boldsymbol{x}}})$, respectively, where $\boldsymbol{q}_{\boldsymbol{x}}$ is referred to as the *anchor*, and $\boldsymbol{q}_{\hat{\boldsymbol{x}}}$ is called the *positive*. We can apply the widely used cosine similarity between *anchor* and *positive* to define the semantic distance between $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ since it is suitable for comparing the similarity between points in such high-dimensional space. It is notable that when we focus on $\hat{\boldsymbol{x}}$, we can regard it as the *anchor*, and $\boldsymbol{x}$ as the *positive* instead.

For the remaining samples $\boldsymbol{m} \in \mathcal{B}/\{\boldsymbol{x}\}$ within training batch $\mathcal{B}$, the same procedure will be followed. Specifically, we can obtain the feature map $\boldsymbol{f}_{\boldsymbol{m}} = \mathcal{F}_b(\boldsymbol{m})$ and $\boldsymbol{f}_{\hat{\boldsymbol{m}}} = \mathcal{F}_b(\hat{\boldsymbol{m}})$ by feeding $\boldsymbol{m}$ and $\hat{\boldsymbol{m}}$ into the backbone of pretrained downstream model respectively. Then, we project them into the semantic space using $\mathcal{P}_\psi(\cdot)$, where $\boldsymbol{q}_{\boldsymbol{m}}$ and $\boldsymbol{q}_{\hat{\boldsymbol{m}}}$ are referred to as the *negative* of $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$, respectively. Similarly, the semantic distance among them can be defined as the cosine similarity between *anchor* and *negative*.

To simply the expression, we define $\mathcal{B}^*$ as the augmented version of $\mathcal{B}$, which comprises both of the original samples from $\mathcal{B}$ and the reconstructed ones, and $|\mathcal{B}^*| = 2|B|$ is satisfied. We also define $\boldsymbol{x}^*$ as the *positive* of $\boldsymbol{x} \in \mathcal{B}^*$. The objective of SemCC is to minimize the semantic distance between the original and reconstructed images while maximizing the semantic distance among the original image and the irreverent images. Therefore, we can use the InfoNCE function [13] to define the semantic contrastive loss, which can be expressed as

$$\mathcal{L}_{sem} = \mathbb{E}_{\boldsymbol{x} \in \mathcal{B}^*} \left\{ -\log \frac{\exp(\boldsymbol{q}_{\boldsymbol{x}} \cdot \boldsymbol{q}_{\boldsymbol{x}^*}/\tau)}{\sum_{\boldsymbol{m} \in \mathcal{B}^*/\{\boldsymbol{x}\}} \exp(\boldsymbol{q}_{\boldsymbol{x}} \cdot \boldsymbol{q}_{\boldsymbol{m}}/\tau)} \right\}, \quad (8)$$

[1]The output of the projection network is typically represented as tensors, which can be straightforwardly normalized into a unit hypersphere. This approach is widely used in the domain of representation learning, as it can help improve training stability. More details about this can be found in [58].

where $\tau > 0$ is the temperature coefficient used to smooth the probability distribution. Next, we will introduce how to take into account the SemCC and semantic contrastive loss to design the loss function and training procedure.

C. Loss Function and Training Procedure

Based on the SemCC, we design a two-stage training strategy to optimize the semantic encoder and decoder. The first stage is pre-training, where we use the SemCC approach to train the weights of the encoder θ_1, the decoder θ_2, and the projection network ψ simultaneously. Since it is difficult to achieve a fast convergence speed when we only optimize the semantic contrastive loss, we combine the semantic contrastive loss with the reconstructed loss between $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$, since reducing the reconstructed loss can help improve the convergence speed in the early training rounds. Specifically, we use the Mean Square Error (MSE) function to evaluate the reconstruction loss for the training batch $\mathcal{B}$, which can be expressed as

$$\mathcal{L}_{rec} = \mathbb{E}_{\boldsymbol{x}\in\mathcal{B}}\left\{\frac{1}{n}||\boldsymbol{x}-\hat{\boldsymbol{x}}||_2^2\right\}. \tag{9}$$

Then, the loss function in the first training stage can be summarized as the linear combination, given by

$$L_1 = \alpha_1\mathcal{L}_{rec} + (1-\alpha_1)\mathcal{L}_{sem}, \tag{10}$$

where $\alpha_1 \in [0,1]$ is a hyperparameter that controls the tradeoff between the two parts of the loss function. For example, we can set $\alpha_1 = k/n$ in the practical semantic communication system. In this context, the system prioritizes the preservation of semantic information over the reconstructed quality when the bandwidth compression ratio is small. In contrast, as the bandwidth compression ratio increases, the system shifts its focus to preserving the reconstructed quality.

In the second training stage, we aim to further optimize the performance of the semantic communication system by jointly fine-tuning the encoder and decoder with a small learning rate to achieve significant inference performance and reconstructed image quality, especially when the bandwidth compression ratio is low. The loss function of this stage can be expressed as

$$L_2 = \alpha_2\mathcal{L}_{rec} + (1-\alpha_2)\mathcal{L}_{Task}, \tag{11}$$

where $\alpha_2 \in [0,1]$ is a hyper-parameter like α_1 and $\mathcal{L}_{Task}$ is the loss function of the downstream task. Specifically, when the downstream task is a classification problem, the cross-entropy function can be employed to model the loss, given by

$$\mathcal{L}_{Task} = \mathbb{E}_{\boldsymbol{x}\in\mathcal{B}}\left\{-\frac{1}{N_{cls}}\sum_{i=1}^{N_{cls}} y_{\boldsymbol{x},i}\log(\hat{y}_{\boldsymbol{x},i})\right\}, \tag{12}$$

where $y_{\boldsymbol{x},i}$ and $\hat{y}_{\boldsymbol{x},i}$ represent the ground-truth and the predicted probability of the i-th class, respectively. Notation N_{cls} denotes the number of classes in the dataset. The whole training procedure is summarized in Algorithm 1, where N_{epochs} and $N_{\text{Fine-tuning}}$ represent the number of training and fine-tuning epochs, respectively.

Algorithm 1: SemCC Training Procedure

```
// Training Stage 1: Pre-training
for epoch ← 1 to N_epochs do
    Sample a batch B from the knowledge base;
    The transmitter encodes each x ∈ B with E_θ1(·);
    The receiver decodes and obtains x̂ with D_θ2(·);
    Extract feature maps f_x and f_x̂ using F^b_φ1(·) for x ∈ B;
    Project feature maps to semantic space using p_ψ(·);
    Calculate reconstruction loss L_rec using (9);
    Calculate semantic contrastive loss L_sem based on (8);
    Calculate combined loss L_1 based on (10);
    Update θ1, θ2, and ψ using SGD.
end
// Training Stage 2: Fine-tuning
for epoch ← 1 to N_Fine-tuning do
    Sample a batch B from the knowledge base;
    The transmitter encodes each x ∈ B with E_θ1(·);
    The receiver decodes and obtains x̂ with D_θ2(·);
    Extract feature maps f_x and f_x̂ using F^b_φ1(·) for x ∈ B;
    Send the feature map to the classifier F^cls_φ2(·);
    Calculate reconstruction loss L_rec using (9);
    Calculate loss of the downstream task L_Task based on (12);
    Calculate combined loss L_2 based on (11);
    Update θ1 and θ2 using SGD.
end
```

V. Semantic Re-encoding with Inaccessible Downstream Model

In this section, we will discuss a more general scenario where the architecture and weights of the downstream network are not accessible. Specifically, we will introduce an alternative approach, namely SemRE, to address these issues, and then present a soft update paradigm for the semantic encoder. After that, we will provide the updated loss function and the training procedure.

A. Semantic Re-encoding

When the weights of the pretrained downstream model are not accessible (i.e. the pretrained downstream is black-box), we cannot use its pre-trained backbone to extract features and subsequently map them to the semantic space, as back-propagation cannot be performed. A simple straightforward solution is to initialize a DNN model randomly and pre-train it using the label information. This pre-trained random model can then guide the training of the proposed SemCC or DeepSC. However, this approach introduces additional system overhead and training latency. Moreover, if the chosen architecture of the random DNN differs from that of the pretrained downstream model, it may result in performance degradation. Therefore, we propose to use only the label information to train the semantic encoder and decoder, which

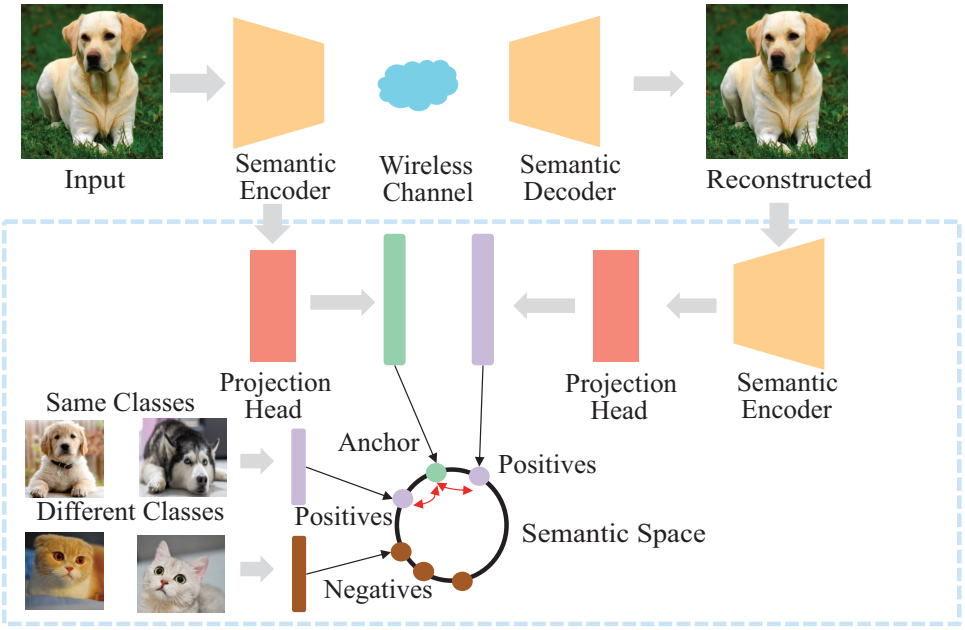

Fig. 3: Illustration of the proposed SemRE.

$$\mathcal{L}_{sem,R} = \mathbb{E}_{\boldsymbol{x}\in\mathcal{B}^*}\left\{ -\frac{1}{|\mathcal{S}_x|}\sum_{\boldsymbol{x}^*\in\mathcal{S}_x} \log \frac{\exp(\boldsymbol{q}_{\boldsymbol{x}}\cdot\boldsymbol{q}_{\boldsymbol{x}^*}/\tau)}{\sum_{\boldsymbol{m}\in\mathcal{B}^*/\{\mathcal{S}_{\boldsymbol{x}}\}}\exp(\boldsymbol{q}_{\boldsymbol{x}}\cdot\boldsymbol{q}_m/\tau)}\right\}. \tag{13}$$

provides a complementary technique between DeepSC and the proposed SemCC. Specifically, we propose to use the semantic encoder to re-encode the reconstructed image instead of the feature extraction operation. This is motivated by the concept of information bottleneck [59] in deep learning, where the network acts as a bottleneck and the only useful information from the input is agreed to pass through itself. In other words, the unimportant information is filtered out in the process. In essence, the ideal semantic encoder should play such a role, i.e., only the semantic information can be retained after semantic encoding, and the task-irrelevant information is removed. It is worth noting that, according to the principles of *data processing inequality* (DPI) [60], no additional semantic information is generated during encoding, transmission, and decoding processing. Therefore, when the reconstructed sample at the receiver is fed back to the semantic encoder, the output should resemble the previous encoding results for an ideal semantic communication system.

We then provide a detailed description of the SemRE and the modification of the semantic contrastive loss. As shown in Fig. 3, in contrast to SemCC, where both the original $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ are fed into the backbone of the pre-trained downstream model, the proposed SemRE approach only needs to perform re-encoding at the receiver, since we have already obtained the encoding results at the sender. Let $\tilde{\boldsymbol{s}}^r = \mathcal{E}_{\theta_1^r}(\hat{\boldsymbol{x}})$ denote the re-encoding operation at the receiver, where θ_1^r represents the parameters of the re-encoder. In particular, θ_1^r is updated based on θ_1 and we will introduce how this update is achieved. After that, the power normalization in (2) is used to obtain the results $\boldsymbol{s}^r$.

Next, we can use $\boldsymbol{s}$ and $\boldsymbol{s}^r$ to perform CL. Specifically, similar to SemCC, SemRE still uses projection meshes. For simplicity, we use $\boldsymbol{q}_{\boldsymbol{x}}$ and $\boldsymbol{q}_{\boldsymbol{x}^*}$ to denote the anchor and positive projection results. To obtain the *negative*, we select samples from the same batch and use cosine similarity to evaluate the semantic distance. It is important to note that we do not consider all remaining samples within the same batch as *negative*, because the ability of semantic extraction in this scenario is limited and we cannot obtain rich and fine-grained semantic information due to the lack of the pre-trained downstream model. Blindly pulling samples away from each other within the same class would degrade the system performance. In other words, it is advisable to consider the semantic similarities between different samples belonging to the same class. For these reasons, we adopt a supervised method in [17]. Specifically, the label of each sample is used to facilitate CL, and we define $\mathcal{S}_{\boldsymbol{x}}$ as the *positive set* containing samples belonging to the same class with $\boldsymbol{x}$. Thus, we can derive the semantic contrastive loss of the SemRE approach as (13). Then we can use (13) to replace (10) to perform the gradient descent.

B. Soft Update Approach

In the training process of the proposed SemRE approach, the semantic encoder plays a key role. It first extracts the semantic information and encodes it before transmission. Then, at the receiver, it extracts the semantic information again to evaluate the quality of the received semantic information. However, training such a semantic encoder is challenging in practice. Because, it is a self-guided process, i.e., the semantic encoder evaluates its own performance and the weights of the

Algorithm 2: Semantic Re-Coding Training Procedure

for *epoch* $\leftarrow 1$ *to* N_{epochs} **do**
  Sample a batch $\mathcal{B}$ from the knowledge base;
  The transmitter encodes each $\boldsymbol{x} \in \mathcal{B}$ with $\mathcal{E}_{\theta_1}(\cdot)$;
  The receiver decodes and obtains $\hat{\boldsymbol{x}}$ with $\mathcal{D}_{\theta_2}(\cdot)$;
  The receiver re-encodes $\hat{\boldsymbol{x}}$ with $\mathcal{E}_{\theta_1^r}(\cdot)$;
  Project the encoded results from both of the transmitter and receiver to semantic space using $p_\psi(\cdot)$;
  Calculate reconstruction loss $\mathcal{L}_{rec}$ based on (9);
  Calculate semantic contrastive loss $\mathcal{L}_{sem,R}$ based on (13);
  Calculate combined loss L_1 based on (10);
  Update θ_1, θ_2 and ψ using SGD;
  if *epoch* $\mod N_{update} = 0$ **then**
    Update $\theta_1^r \leftarrow \beta\theta_1^r + (1-\beta)\theta_1$
  end
end

semantic encoder are also updated dynamically. This makes the semantic encoder at the receiver fail to provide stable evaluation, which complicates the optimization of the entire process.

Inspired by weight update strategies in deep reinforcement learning, as exemplified by *deep Q-Networks* (DQN) [61] and *deep deterministic policy gradients* (DDPG) [62], we propose a soft update approach for the semantic encoder to address these challenges. This approach decouples the evaluation and update steps in the training process. Specifically, the semantic encoder at the receiver does not update its weights after each training batch, as it does at the transmitter. Instead, its weights are updated periodically to achieve better training stability. The detailed soft-update approach can be expressed as

$$\theta_1^r \leftarrow \beta\theta_1^r + (1-\beta)\theta_1, \tag{14}$$

where $\beta \in [0,1]$ is a hyper-parameter which controls the update magnitude. We finally summarize the whole training process of SemRE as shown in Algorithm 2, where N_{update} is the update interval.

VI. Simulations

A. Simulations Settings

To verify the effectiveness of the proposed framework, we conduct experiments on CIFAR-10, which contains 60,000 32×32 color images divided into 10 classes. The training set contains 50,000 images, while the test set contains 10,000 images. A pre-trained ResNet-20 [56] is used as the backbone and classifier of the downstream model for inference[2].

The projection network adopts a two-layer fully connected structure with an output dimension of 32. The number of training epochs for the pre-training and fine-tuning is set to 200 and 50, respectively, with a batch size of 128. We also use the Adam optimizer with a learning rate of 0.001 for the first pre-training stage and 0.0001 for the second fine-tuning stage. These learning rates are adjusted every 50 epochs with a decay factor of 0.5.

For the network environment, we set the transmit power to unity and the transmit SNR to 20dB and 5dB for normal and noisy environments, respectively. In addition, we assume that the receiver can estimate the channel parameters perfectly in the case of Rayleigh channel. We compared the proposed approaches with the advanced DL-based semantic communication approaches, which are listed as follows,

- **SemCC**: The proposed CL based semantic communication approach, where the pretrained backbone of the downstream task is used in the training process.
- **SemRE**: The proposed SemRE strategy and no pre-trained backbone is adopted in this case.
- **DeepJSCC** [24]: Deep learning based source-channel joint coding that maps the original input to the channel input through the structure of an autoencoder.
- **DeepSC** [12]: The SOTA deep learning based semantic communication framework to support downstream inference task. DeepSC trains the semantic encoder and decoder with both semantic loss provided by the whole pre-trained ResNet-20 and observation loss in (11) to achieve efficient semantic information transmission. Note that the hyper-parameter α_2 is set to the same as it in the fine-tuning stage of the proposed approaches.

For fair comparison, the architectures of the encoders and decoders in these approaches are set to be the same, and the network environment settings are kept consistent across all experiments, if not specified.

Moreover, we also compare the performance of the proposed approaches with conventional digital communication using separate source and channel coding under the same bandwidth compression ratio. For the source coding, we leverage the SOTA image compression algorithm named better portable graphics (BPG)[3], which is based on the intra-frame encoding approach of the high-efficiency video coding (HEVC, aka H.265) standard. As for the channel coding, we integrate LDPC code configured according to the IEEE 802.16E standard (Mobile WIMAX), where the block length of 2304 and rates of 1/2, 2/3 and 3/4 are adopted in our simulations. In addition, we use the quadrature amplitude modulation (QAM) with order of 4,16 and 64. Notably, we only report the results of the optimal combination of LDPC rates and modulation schemes for simplicity.

In further, we present the upper bound performance of the digital communication approach, denoted as BPG+Capacity, which realizes capacity-achieving transmission based on Shannon theorem for a given transmit SNR, with the assumption of error-free transmission. Hence, practical digital transmission schemes incorporating channel coding and modulation can not outperform this upper bound.

B. Effectiveness

Fig. 4 compares the accuracy performance of DeepJSCC, DeepSC, the conventional digital communication and the

[2]The pre-trained weights can be found at https://github.com/chenyaofo/pytorch-cifar-models.

[3]https://bellard.org/bpg/.

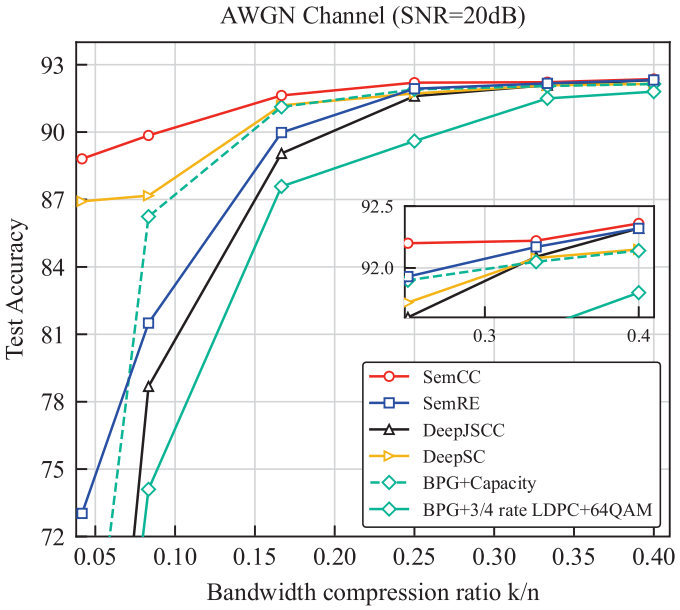

Fig. 4: Test Accuracy versus the bandwidth compression ratio under AWGN channel, where SNR is 20dB.

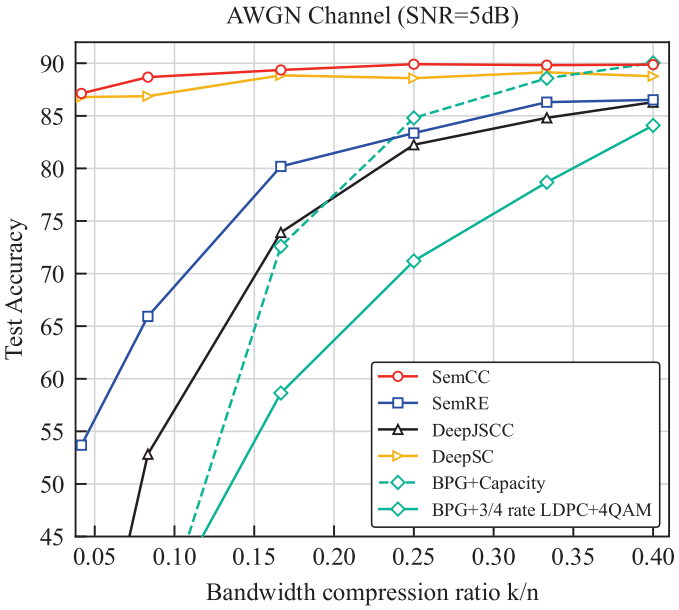

Fig. 6: Test Accuracy versus the bandwidth compression ratio under AWGN channel, where SNR is 5dB.

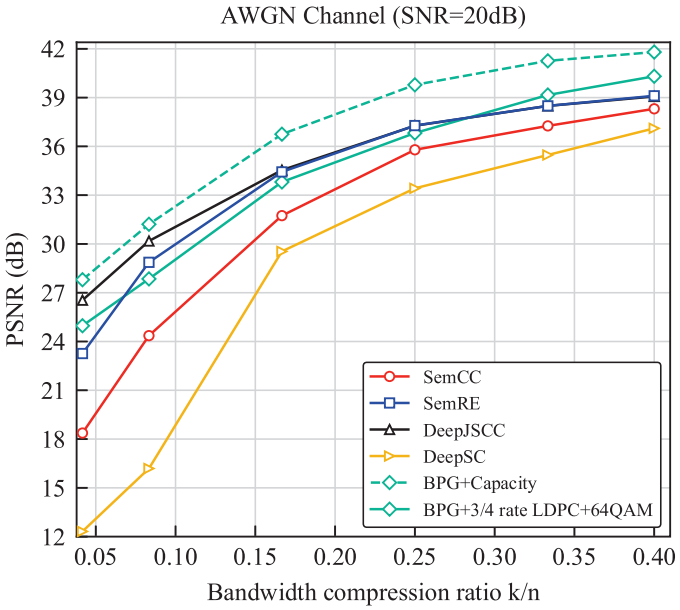

Fig. 5: PSNR versus the bandwidth compression ratio under AWGN channel, where SNR is 20dB.

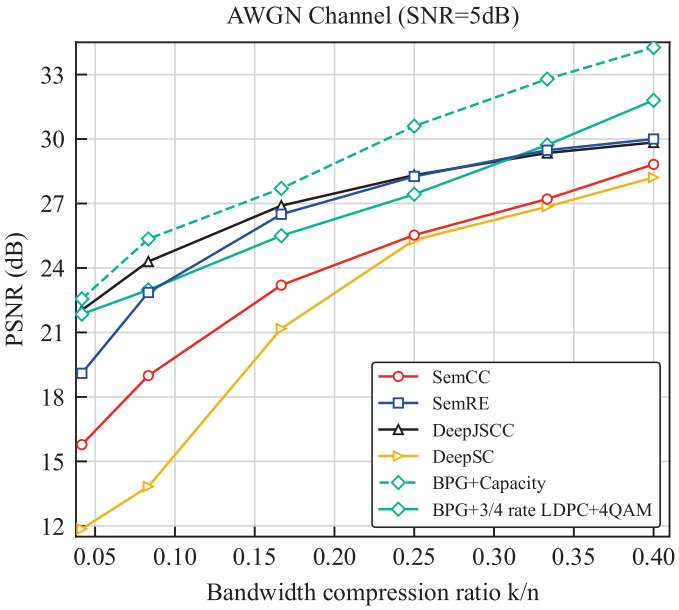

Fig. 7: PSNR versus the bandwidth compression ratio under AWGN channel, where SNR is 5dB.

proposed approaches including SemCC and SemRE under AWGN channel where the SNR is set to 20dB and the bandwidth compression ratio k/n varies from 1/24 to 1/2.5. For the digital communication system, the combination of 3/4 rate LDPC and 64QAM is used. This figure clearly shows that the proposed SemCC consistently outperforms the compared ones in terms of accuracy. In particular, when the compression ratio is 1/2.5, all approaches can transmit rich semantic information to support the downstream task, resulting in high accuracy levels of about 92.3%. As the bandwidth compression ratio decreases, the proposed SemCC still maintains a comparable accuracy performance. For example, the proposed SemCC can achieve accuracy levels of 89.85% and 88.81% at bandwidth compression ratios of 1/12 and 1/24, respectively, which outperforms DeepSC by about 2% at the corresponding bandwidth compression ratios and also shows an accuracy gain of up to 40% and 26% over DeepJSCC and BPG+Capcacity, respectively. In addition, the SemRE approach is also superior to DeepJSCC, achieving an accuracy gain of up to 25% at a bandwidth compression ratio of 1/24. It is important to note that both approaches do not use a pre-trained backbone during the training process. These results suggest that the proposed approaches can effectively extract semantic information to meet the requirements of the

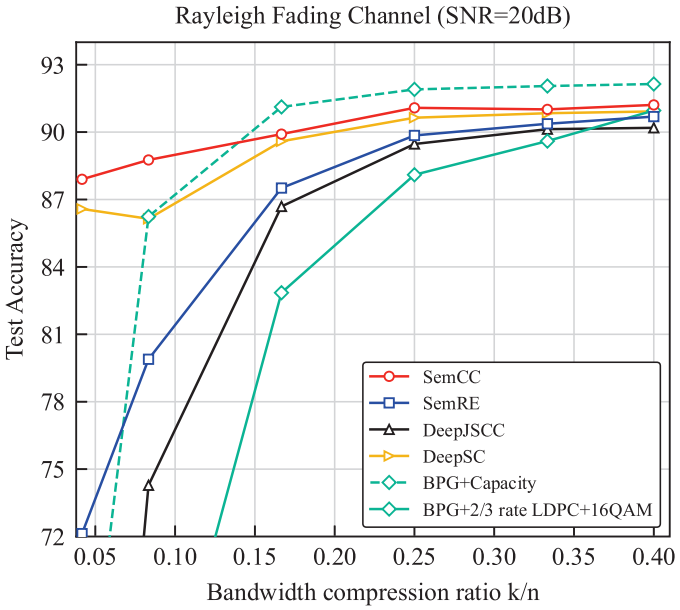

Fig. 8: Test accuracy versus the bandwidth compression ratio under Rayleigh fading channels, where SNR is 20dB.

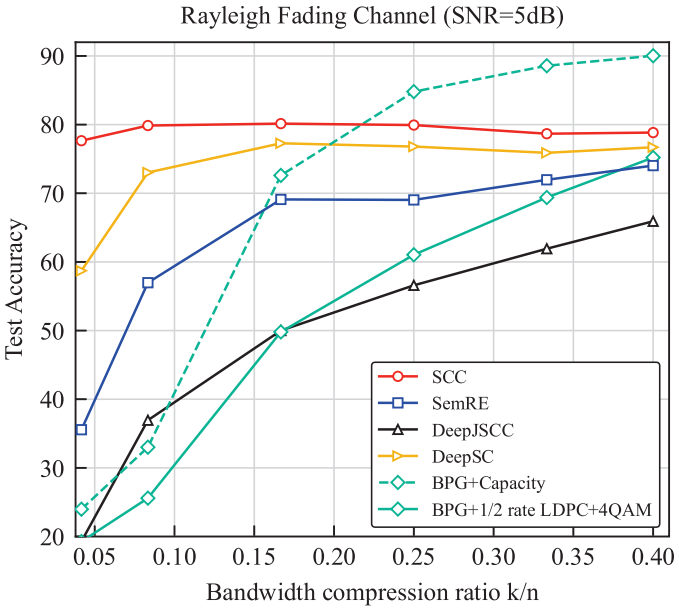

Fig. 10: Test Accuracy versus the bandwidth compression ratio under Rayleigh fading channels, where SNR is 5dB.

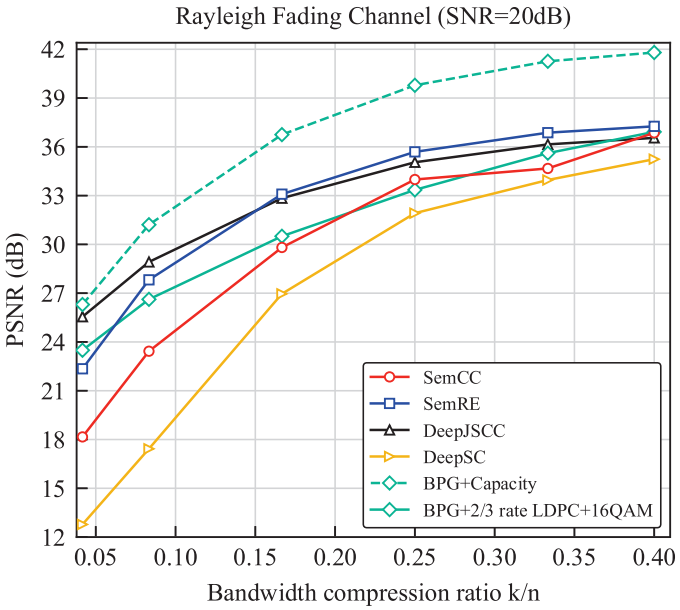

Fig. 9: PSNR versus the bandwidth compression ratio under Rayleigh fading channels, where SNR is 20dB.

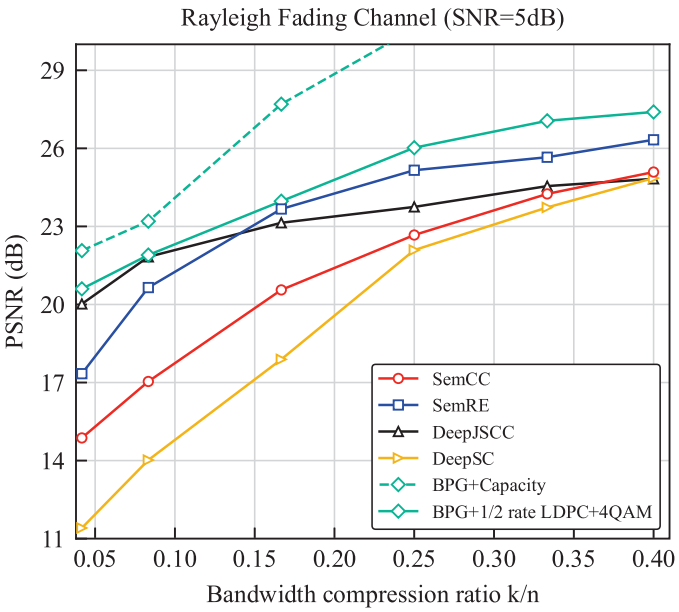

Fig. 11: PSNR versus the bandwidth compression ratio under Rayleigh fading channels, where SNR is 5dB.

downstream task and remove irrelevant redundant information to ensure that the semantic information can be successfully transmitted. This is particularly beneficial in scenarios where the channel bandwidth is limited.

Fig. 5 presents the peak signal-to-noise ratio (PSNR) comparison of the proposed SemCC and SemRE and the four complementary approaches, where the SNR is set to 20dB and the bandwidth compression ratio varies from 1/24 to 1/2.5. For the digital communication system, the combination of 3/4 rate LDPC and 64QAM is used. As shown in the figure, we can see that as the bandwidth compression ratio increases, the PSNRs of all the approaches improve and the conventional approach performs best when the bandwidth compression ratio is larger than 1/3. Although SemCC and SemRE sacrifice some image quality to prioritize semantic information when the bandwidth compression ratio is low, they can quickly catch up with the PSNR of DeepJSCC at higher compression ratios. Specifically, the proposed SemCC achieves a PSNR of 38.31dB, which is close to the 39.07dB of DeepJSCC, and outperforms DeepSC with 37.11dB when the bandwidth compression ratio is 1/2.5. Moreover, the SemRE approach achieves the same PSNR performance as the DeepJSCC when the bandwidth compression ratio is greater than 1/6. These results indicate that the proposed SemCC and SemRE approaches

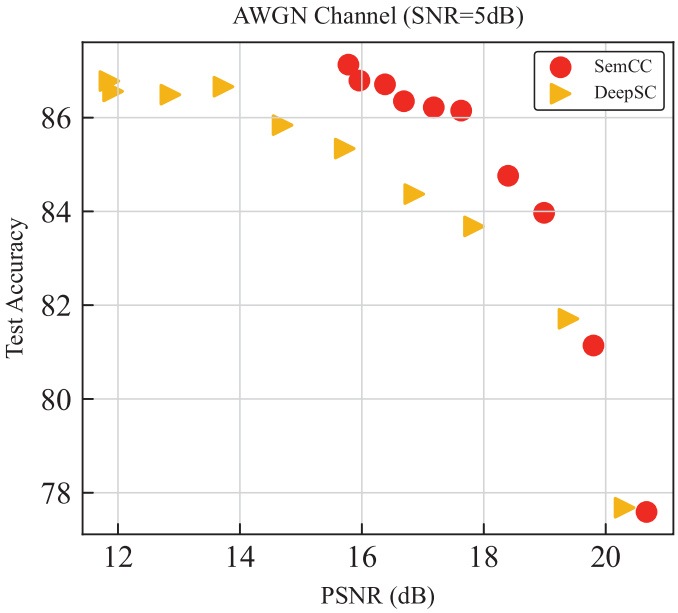

Fig. 12: Test Accuracy on CIFAR-10 versus PSNR under AWGN channel with SNR of 5dB, where the bandwidth compression ratio is set to 1/24.

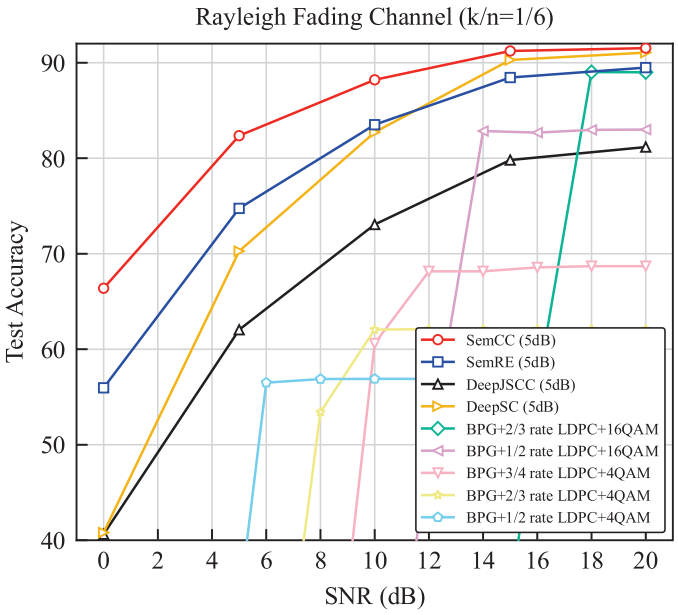

Fig. 13: Test Accuracy on CIFAR-10 versus test SNR under Rayleigh channel. The semantic encoder and decoder are trained at an SNR of 5dB, and the high-performance RepVGG16 is utilized as the downstream model.

can prioritize the transmission of semantic information over irrelevant background information to ensure the performance of the downstream task in bandwidth-limited scenarios, and meanwhile transmit enough background information to obtain good image quality when the bandwidth is not a bottleneck. These results further demonstrate the effectiveness of the proposed approaches.

Fig. 6 and Fig. 7 show the performance comparison of several approaches under low SNR in terms of accuracy

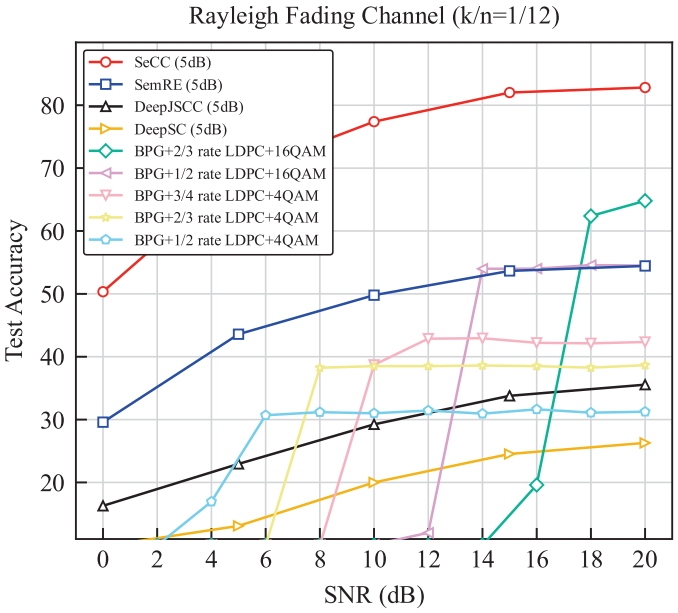

Fig. 14: Test Accuracy on CIFAR-10 versus test SNR under Rayleigh channel. The semantic encoder and decoder are trained at an SNR of 5dB, and the lightweight ShuffleNet-V2 is utilized as the downstream model.

and PSNR, respectively. Specifically, both figures consider a low SNR of 5dB, and the bandwidth compression ratio varies from 1/24 to 1/2.5. Moreover, 3/4 rate LDPC and 4QAM is used in this case. From Fig. 6, we can see that the proposed SemCC still shows the superiority in terms of accuracy compared to the competitive ones, indicating its robustness in low SNR scenarios. From Fig. 7, we can find that the proposed approaches can adaptively sacrifice the global information to obtain comparable semantic performance when the bandwidth compression ratio is low, and meanwhilebtain sufficient reconstructed quality in terms of PSNR as the bandwidth compression ratio increases. These results in both figures further verify the effectiveness and robustness of the proposed approaches in low SNR scenarios.

To further evaluate the performance of several approaches, we perform a comparison under Rayleigh fading channels. Fig. 8 shows the accuracy performance where the SNR is 20dB and the bandwidth compression ranges from 1/24 to 1/2.5. For the digital communication system, we utilize a combination of 2/3 rate LDPC and 16QAM. From this figure, we can see that all approaches experience performance degradation under Rayleigh fading channels compared to the AWGN channel. However, the proposed SemCC still achieves a leading level of accuracy. Specifically, at a bandwidth compression ratio of 1/2.5, the proposed SemCC achieves an accuracy of 91.21%, which is approximately 1% higher than DeepJSCC and 0.3% higher than DeepSC. As the bandwidth compression ratio decreases, the proposed SemCC demonstrates its adaptability to Rayleigh fading channels and still achieves an accuracy of about 90% when the bandwidth compression ratio ranges from 1/4 to 1/24, which achieves accuracy gains of up to 46.08% and 2.62% over DeepJSCC and DeepSC, respectively, and

TABLE I: Ablation study on AGWN channels (SNR=5dB)

	$k/n = 1/24$			$k/n = 1/12$			$k/n = 1/6$		
Metrics	Baseline	w/o. FT	w/o. CL	Baseline	w/o. FT	w/o. CL	Baseline	w/o. FT	w/o. CL
ACC	**87.13**	83.48 (-3.65)	**86.78 (-0.26)**	**88.68**	**88.16 (-0.52)**	86.87 (-1.81)	**89.32**	**89.18 (-0.14)**	88.85 (-0.47)
PSNR (dB)	**15.87**	**16.00 (+0.13)**	11.86 (-4.01)	**18.99**	**18.97 (-0.02)**	13.83 (-5.16)	**23.60**	**23.18 (-0.42)**	21.28 (-2.32)

TABLE II: Ablation study on Rayleigh fading channels (SNR=5dB)

	$k/n = 1/24$			$k/n = 1/12$			$k/n = 1/6$		
Metrics	Baseline	w/o. FT	w/o. CL	Baseline	w/o. FT	w/o. CL	Baseline	w/o. FT	w/o. CL
ACC	**77.65**	**71.00 (-6.65)**	65.22 (-12.43)	**79.86**	**76.99 (-2.87)**	73.02 (-6.84)	**80.14**	**78.79 (-1.35)**	77.28 (-2.86)
PSNR (dB)	**14.87**	**14.94 (+0.07)**	11.41 (-3.46)	**17.04**	**16.94 (-0.10)**	14.02 (-3.02)	**20.65**	**20.48 (-0.17)**	17.89 (-2.76)

even outperforms the upper bound performance of the digital communication approach. Moreover, the SemRE also shows its superiority in exploiting the Rayleigh fading characteristics compared to DeepJSCC and BPG+Capacity, achieving accuracy gains of up to 30.31% and 10.27%, respectively. These results further demonstrate the effectiveness of the proposed SemCC and SemRE approaches under Rayleigh fading channels.

Fig. 9 shows the PSNR comparison under Rayleigh fading channels with an SNR of 20dB. For the digital communication system, a combination of 2/3 rate LDPC and 16QAM is used. From this figure, we can see that both the proposed SemCC and SemRE approaches prioritize the transmission of semantic information at low bandwidth compression ratios while preserving enough detail to improve image quality as the bandwidth compression ratios increase. It is noteworthy that the SemRE approach achieves a superior PSNR performance compared to DeepJSCC when the bandwidth compression ratio is greater than 1/6, and the proposed SemCC also outperforms DeepJSCC and the conventional digital communication approach when the bandwidth compression ratio is 1/2.5. This can be attributed to the introduced CL and the approach of replacing the data augmentation with a practical wireless channel, which helps to mitigate the effect of Rayleigh fading channels. In addition, the presence of rich semantic information plays a crucial role in image reconstruction at the receiver.

Fig. 10 and Fig. 11 show the performance comparison under Rayleigh fading, and the SNR is set to 5dB. In this scenario, the characteristics of fading and channel noise pose even greater challenges to the semantic communication system. From Fig. 10, we can see that the test accuracy of all approaches deteriorates significantly. Compared to the competing approaches, the accuracy gains of the proposed SemCC increase with a smaller bandwidth compression ratio, and it outperforms DeepSC and DeepJSCC by up to 18.65% and 57%, respectively. Moving to Fig. 11, we find that the SemRE approach still achieves superior image quality in terms of PSNR compared to other approaches when the bandwidth compression ratio is larger than 1/6. It achieves a gain of up to 1.5 dB over DeepJSCC. Moreover, the proposed SemCC also outperforms DeepJSCC when the bandwidth compression ratio is 1/2.5. These results further demonstrate the effectiveness of the proposed approaches in overcoming the challenges of channel environments and successfully balancing the semantic information and image details according to the bandwidth compression ratio.

Next, we vary the trade-off parameters in (10) and (11) across a broad range to adjust the PSNR value and observe the corresponding performance in terms of accuracy under AWGN channels, where the BCR is set to 1/24 and SNR is set to 5dB. As shown in Fig. 12, the test accuracy of the proposed SemCC and DeepSC decreases as the PSNR increases, which indicates that both the proposed SemCC and DeepSC can balance the trade-off between the semantic information and image quality. However, the test accuracy of DeepSC is more sensitive to the trade-off parameter, while the proposed SemCC can maintain a higher accuracy level across a broad range of trade-off parameters and the corresponding PSNR values. Specifically, the proposed SemCC achieves the test accuracy of 87.13% and 86.71% when the PSNR is about 15.87dB and 16.38dB, respectively, while DeepSC achieves the same-level accuracy with a lower PSNR value of 12-13dB. These results further demonstrate the effectiveness of the proposed SemCC in the trade-off between the semantic information and image quality.

We also conduct ablation studies, as shown in Table I and Table II. We consider our proposed SemCC with two-stage training as the baseline. We then remove the first stage of CL pre-training (denoted as w/o. CL) and the second stage of fine-tuning (denoted as w/o. FT), respectively, to evaluate their individual impact. It is notable that for w/o. CL, we used a larger learning rate to train the semantic encoder and decoder from scratch than that used in the fine-tuning stage and in fact, SemCC degrades to DeepSC in this case. Specifically, Table I provides the performance comparison, where AWGN channel with SNR of 5dB is set and the bandwidth compression k/n is set to 1/24, 1/12 and 1/6, respectively. From this table, we can find that the baseline achieves the highest accuracy and PSNR, while the baseline without fine-tuning can still outperform the one without CL pre-training in most cases. These results indicate that the gains of the proposed SemCC mainly come from the CL pre-training, and the fine-tuning also contributes to improved accuracy performance, especially when the bandwidth compression ratio is $1/24$.

Similar results of ablation studies on Rayleigh fading channels are presented in Table II, where SNR is set to 5dB. From

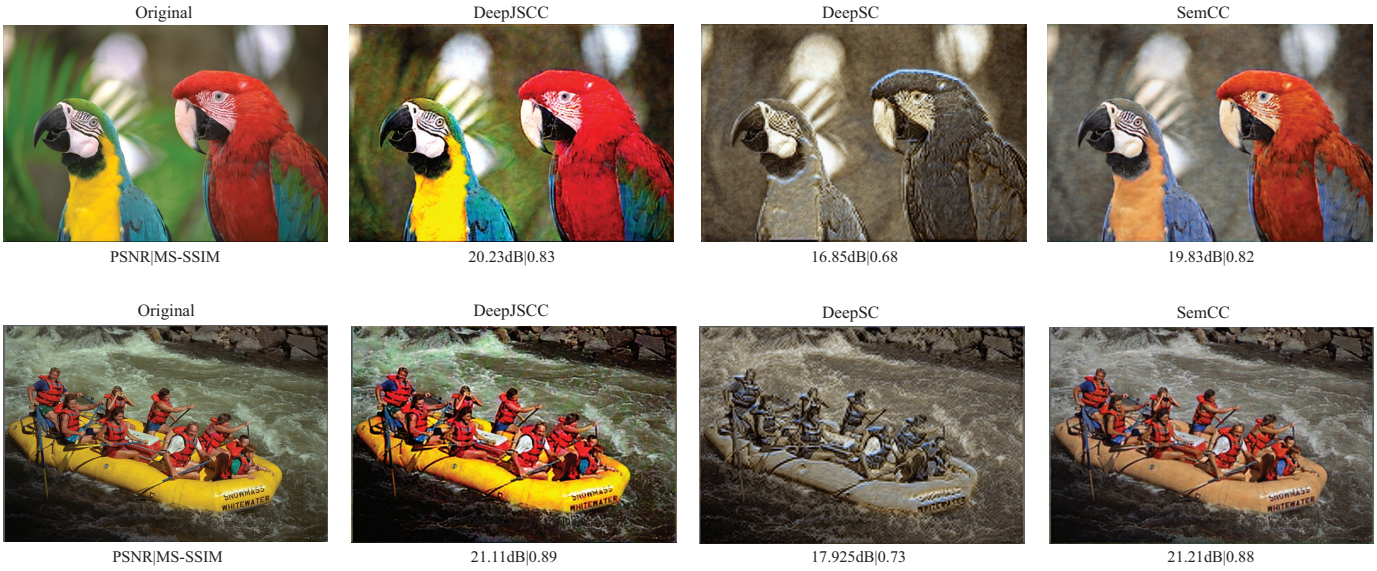

Fig. 15: Visual comparison of the reconstructed image under AWGN channel with an SNR of 20dB and a bandwidth compression ratio of 1/48. The proposed approaches effectively preserves the semantic information of the colorful macaws and removes the irrelevant background.

this table, we can observe that a larger gain is obtained by the CL pre-training over training from scratch compared to that under AWGN channel, which further demonstrates the effectiveness of the proposed CL-based pre-training, as well as the benefits of fine-tuning.

C. Robustness

To verify the robustness of the proposed SemCC and SemRE approaches, we consider a more challenging scenario where there is a mismatch between the training SNR and the test SNR. In addition, the model architecture for the downstream task is different from the one used in the training process. In particular, we consider the state-of-the-art (SoTA) RepVGG16 [63] and the lightweight ShuffleNet [64] as the downstream models. RepVGG16 and ShuffleNet are more powerful and less powerful, respectively, compared to the ResNet-20 used in the training phase. Therefore, we can evaluate the robustness of the proposed approaches by assessing whether the preserved semantic information is general enough to work properly with a more powerful pre-trained model downstream and whether it is sufficient and appropriate for the lightweight model. We also provide the performance of the conventional digital communication system under the same condition, where we use the serval combination of LDPC rates and modulation schemes to achieve the best performance. In Fig. 13, we present the test accuracy comparison under Rayleigh fading channels for different SNRs, where the semantic encoder and the semantic decoder are both trained at the SNR of 5dB, and the bandwidth ratio is set to 1/6. The model architecture for the downstream task is ResNet20 during the training phase, while in the test phase, we use RepVGG16 [63]. This change may indicate an upgrade in the GPU device of the receiver, which allows the use of a more powerful deep model. From this figure, we can observe that despite the mismatched SNR and the use of a more complex downstream model, the proposed approaches still demonstrate competitive test accuracy levels by providing enough semantic information to the downstream model and the ability to protect it from fading and noise. It is also important to note that the proposed SemRE, which utilizes only label information, demonstrates better performance compared to DeepSC when the downstream model is unknown. These results suggest that our semantic communication system maintains its robustness in the face of real-world variations.

In Fig. 14, we continue to evaluate the test accuracy under Rayleigh fading channels, where we employ a lightweight model architecture. Specifically, ShuffleNet [64] is employed for the downstream model, and the bandwidth compression ratio is set to 1/12 to simulate computational resource and bandwidth constraints. In this scenario, we observe a significant degradation in the accuracy performance of DeepSC, as it primarily emphasizes specific semantic information and shows its sensitivity to different model architectures of the downstream task. In contrast, our proposed approaches demonstrate that they can provide more general semantic communication by maintaining competitive test accuracy levels, reaching up to 80%, and showing gains of up to 40% over DeepSC. In further, the proposed approaches can effectively mitigate the cliff effect under various channel conditions. These results highlight the robustness of our semantic communication system in scenarios with limited computational and bandwidth resources.

D. Visualization

We also provide visual comparisons of different approaches using the Kodak dataset in Fig. 15, where the encoder and decoder are trained on the STL10 dataset, the SNR is 20dB, and the bandwidth compression ratio is 1/48. From this figure, we can find that the quality of the reconstructed image

for DeepJSCC, DeepSC, and the proposed SemCC all get deteriorated in this large compression ratio, but the proposed SemCC can effectively preserve the semantic information. For example, the proposed SemCC effectively preserves the semantic information of the colorful macaws, people, and rafters, and removes the irrelevant background. This is particularly beneficial in scenarios where the channel bandwidth is limited and can explain the reasons for the superior performance of the proposed SemCC in the downstream task. On the other hand, DeepJSCC treats all information as important and attempts to reconstruct the background, leading to the loss of semantic information about colorful macaws and rafters. Although DeepSC can extract textual information, it still fails to preserve the colorful macaws and rafters, significantly deteriorating image quality. These results further demonstrate the effectiveness of the proposed approaches in preserving semantic information and removing irrelevant background information.

VII. Conclusion

In this paper, we investigated a CL-based semantic communication system. Our contribution was to introduce the concept of semantic contrastive loss, which provides a more reasonable evaluation of semantic-level aspects during the training process. Moreover, we modified the CL procedure by replacing the traditional data augmentation with a practical wireless channel and proposed the SemCC approach, which allows us to comprehensively exploit the impact of the channel on the transmission of semantic information. We also proposed the SemRE approach, which uses a copy of the semantic encoder to guide the whole training process, to address the problem of an inaccessible downstream model. Further, we designed training procedures for SemCC and SemRE, respectively, which achieved a good trade-off between preserving semantic information and retaining intricate details. Finally, we conducted simulations under various conditions, including different bandwidth compression ratios, SNRs, and downstream model configurations, to demonstrate the effectiveness and robustness of the proposed approaches.

References

[1] S. Tang, Q. Yang, L. Fan, X.Lei, Y. Deng, and A. Nallanathan, "Contrastive learning based semantic communication for wireless image transmission," in *IEEE Veh. Technology Conf. (VTC-Fall)*, Oct. 2023, pp. 1–6.

[2] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.

[3] R. G. Gallager, "Low-density parity-check codes," *IEEE Trans. Inf. Theory*, vol. 8, no. 1, pp. 21–28, 1962.

[4] E. Arikan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3051–3073, 2009.

[5] K. B. Letaief, Y. Shi, J. Lu, and J. Lu, "Edge artificial intelligence for 6G: Vision, enabling technologies, and applications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 5–36, 2022.

[6] X. Duan, X. Wang, L. Lu, N. X. Shi, C. Liu, T. Zhang, and T. Sun, "6G architecture design: From overall, logical and networking perspective," *IEEE Commun. Mag.*, vol. 61, no. 7, pp. 158–164, 2023.

[7] C. Wang, X. You, X. Gao, X. Zhu, Z. Li, C. Zhang, H. Wang, Y. Huang, Y. Chen, H. Haas, J. S. Thompson, E. G. Larsson, M. D. Renzo, W. Tong, P. Zhu, X. Shen, H. V. Poor, and L. Hanzo, "On the road to 6G: Visions, requirements, key technologies, and testbeds," *IEEE Commun. Surv. Tutorials*, vol. 25, no. 2, pp. 905–974, 2023.

[8] J. Bao, P. Basu, M. Dean, C. Partridge, A. Swami, W. Leland, and J. A. Hendler, "Towards a theory of semantic communication," in *IEEE Netw. Sci. Workshop*, 2011, pp. 110–117.

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

[10] Z. Qin, X. Tao, J. Lu, and G. Y. Li, "Semantic communications: Principles and challenges," *arXiv:2201.01389*, 2022.

[11] Y. Shi, Y. Zhou, D. Wen, Y. Wu, C. Jiang, and K. B. Letaief, "Task-oriented communications for 6G: Vision, principles, and technologies," *IEEE Wirel. Commun.*, vol. 30, no. 3, pp. 78–85, 2023.

[12] 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, 2023.

[13] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 119, 2020, pp. 1597–1607.

[14] X. Chen, H. Fan, R. B. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv*.

[15] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020, pp. 9726–9735.

[16] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Adv. Neural Inf. Process. Syst. (Neural-IPS)*, 2020.

[17] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Adv. Neural Inf. Process. Syst. (Neural-IPS)*, vol. 33, pp. 18 661–18 673, 2020.

[18] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," in *Proc. Empir. Methods Nat. Lang. Process. (EMNLP)*, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds., 2021, pp. 6894–6910.

[19] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 139, 2021, pp. 8748–8763.

[20] Y. Zeng, C. Jiang, J. Mao, J. Han, C. Ye, Q. Huang, D. Yeung, Z. Yang, X. Liang, and H. Xu, "Clip^2: Contrastive language-image-point pretraining from real-world point cloud data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*. IEEE, 2023, pp. 15 244–15 253.

[21] Q. Zhou, G. Pang, Y. Tian, S. He, and J. Chen, "AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection," in *Int. Conf. Learn. Represent. (ICLR)*, 2024.

[22] C. Chaccour and W. Saad, "Disentangling learnable and memorizable data via contrastive learning for semantic communications," in *Asilomar Conf. Signals, Sys., Comput. (ACSSC)*. IEEE, 2022, pp. 1175–1179.

[23] Z. Tian, H. Vo, C. Zhang, G. Min, and S. Yu, "An asynchronous multi-task semantic communication method," *IEEE Netw.*, pp. 1–1, 2023.

[24] 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, 2019.

[25] R. Carnap, Y. Bar-Hillel *et al.*, "An outline of a theory of semantic information," *Res. Lab. Electron., MIT.*, 1952.

[26] B. Juba and M. Sudan, "Universal semantic communication i," in *Proc. of ACM Symp. Theory Comput.*, 2008, pp. 123–132.

[27] B. Güler, A. Yener, and A. Swami, "The semantic communication game," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 4, pp. 787–802, 2018.

[28] Y. Zhong, "A theory of semantic information," *China Commun.*, vol. 14, no. 1, pp. 1–17, 2017.

[29] Y. Shao, Q. Cao, and D. Gunduz, "A theory of semantic communication," *arXiv:2212.01485*, 2022.

[30] J. Tang, Q. Yang, and Z. Zhang, "Information-theoretic limits on compression of semantic information," *arXiv:2306.02305*, 2023.

[31] S. Kobus, T.-Y. Tung, and D. Gündüz, "Goal-oriented compression with a constrained decoder," in *IEEE Int. Symp. Inf. Theory (ISIT)*, 2023, pp. 868–873.

[32] F. Pase, S. Kobus, D. Gündüz, and M. Zorzi, "Semantic communication of learnable concepts," in *IEEE Int. Symp. Inf. Theory (ISIT)*, 2023, pp. 731–736.

[33] D. B. Kurka and D. Gündüz, "DeepJSCC-f: Deep joint source-channel

coding of images with feedback," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 178–193, 2020.

[34] T.-Y. Tung, D. B. Kurka, M. Jankowski, and D. Gündüz, "DeepJSCC-Q: Constellation constrained deep joint source-channel coding," *IEEE J. Sel. Areas Inf. Theory*, vol. 3, no. 4, pp. 720–731, 2022.

[35] M. Ding, J. Li, M. Ma, and X. Fan, "SNR-adaptive deep joint source-channel coding for wireless image transmission," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 1555–1559.

[36] M. Yang and H. Kim, "Deep joint source-channel coding for wireless image transmission with adaptive rate control," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2022, pp. 5193–5197.

[37] J. Xu, B. Ai, W. Chen, A. Yang, P. Sun, and M. Rodrigues, "Wireless image transmission using deep source channel coding with attention modules," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2315–2328, 2022.

[38] E. Erdemir, T.-Y. Tung, P. L. Dragotti, and D. Gündüz, "Generative joint source-channel coding for semantic image transmission," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 8, pp. 2645–2657, 2023.

[39] T. Han, J. Tang, Q. Yang, Y. Duan, Z. Zhang, and Z. Shi, "Generative model based highly efficient semantic communication approach for image transmission," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2023, pp. 1–5.

[40] W. Xia, Y. Zhang, Y. Yang, J.-H. Xue, B. Zhou, and M.-H. Yang, "Gan inversion: A survey," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 45, no. 3, pp. 3121–3138, 2022.

[41] Y. Bo, Y. Duan, S. Shao, and M. Tao, "Learning based joint coding-modulation for digital semantic communication systems," in *Int. Conf. Wirel. Commun. Signal Process. (WCSP)*, 2022, pp. 1–6.

[42] N. Farsad, M. Rao, and A. Goldsmith, "Deep learning for joint source-channel coding of text," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2018, pp. 2326–2330.

[43] 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.

[44] X. Peng, Z. Qin, D. Huang, X. Tao, J. Lu, G. Liu, and C. Pan, "A robust deep learning enabled semantic communication system for text," in *IEEE Glob. Commun. Conf. (GLOBECOM)*, 2022, pp. 2704–2709.

[45] S. Guo, Y. Wang, S. Li, and N. Saeed, "Semantic importance-aware communications using pre-trained language models," *IEEE Commun. Lett.*, vol. 27, no. 9, pp. 2328–2332, 2023.

[46] T. Han, Q. Yang, Z. Shi, S. He, and Z. Zhang, "Semantic-preserved communication system for highly efficient speech transmission," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 245–259, 2023.

[47] Z. Weng and Z. Qin, "Semantic communication systems for speech transmission," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2434–2444, 2021.

[48] J. Shao and J. Zhang, "Bottlenet++: An end-to-end approach for feature compression in device-edge co-inference systems," in *IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2020, pp. 1–6.

[49] 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, 2022.

[50] J. Shao, Y. Mao, and Jun Zhang, "Task-oriented communication for multidevice cooperative edge inference," *IEEE Trans. Wirel. Commun.*, vol. 22, no. 1, pp. 73–87, 2023.

[51] H. Xie, Z. Qin, and G. Y. Li, "Task-oriented multi-user semantic communications for VQA," *IEEE Wirel. Commun. Lett.*, vol. 11, no. 3, pp. 553–557, 2022.

[52] 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, 2022.

[53] 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, 2022.

[54] W. Yang, X. Chi, L. Zhao, Z. Xiong, and W. Jiang, "Task-driven semantic-aware green cooperative transmission strategy for vehicular networks," *IEEE Trans. Commun.*, vol. 71, no. 10, pp. 5783–5798, 2023.

[55] Y. Wu, S. Tang, L. Zhang, L. Fan, X. Lei, and X. Chen, "Resilient machine learning based semantic-aware mec networks for sustainable next-g consumer electronics," *IEEE Trans. Consum. Electron.*, pp. 1–1, 2023.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2016, pp. 770–778.

[57] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2016, pp. 1874–1883.

[58] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 119, 2020, pp. 9929–9939.

[59] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *IEEE Inf. Theory Workshop (ITW)*, 2015, pp. 1–5.

[60] M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information*. Cambridge Univ. Press., 2010.

[61] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. A. Riedmiller, A. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nat.*, vol. 518, no. 7540, pp. 529–533, 2015.

[62] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *Int. Conf. Learn. Represent. (ICLR)*, 2016.

[63] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "RepVGG: Making vgg-style convnets great again," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021, pp. 13 728–13 737.

[64] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient cnn architecture design," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018.