Contrastive Learning based Semantic Communication for Wireless Image Transmission

Shunpu Tang, Qianqian Yang, Lisheng Fan, Xianfu Lei, Yansha Deng, and Arumugam Nallanathan, *Fellow, IEEE*

Abstract—Recently, semantic communication has been widely applied in wireless image transmission systems as it can prioritize the preservation of meaningful semantic information in images over the accuracy of transmitted symbols, leading to improved communication efficiency. However, existing semantic communication approaches still face limitations in achieving considerable inference performance in downstream AI tasks like image recognition, or balancing the inference performance with the quality of the reconstructed image at the receiver. Therefore, this paper proposes a contrastive learning (CL)-based semantic communication approach to overcome these limitations. Specifically, we regard the image corruption during transmission as a form of data augmentation in CL and leverage CL to reduce the semantic distance between the original and the corrupted reconstruction while maintaining the semantic distance among irrelevant images for better discrimination in downstream tasks. Moreover, we design a two-stage training procedure and the corresponding loss functions for jointly optimizing the semantic encoder and decoder to achieve a good trade-off between the performance of image recognition in the downstream task and reconstructed quality. Simulations are finally conducted to demonstrate the superiority of the proposed method over the competitive approaches. In particular, the proposed method can achieve up to 56% accuracy gain on the CIFAR10 dataset when the bandwidth compression ratio is 1/48.

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

I. INTRODUCTION

Recently, semantic communication has emerged as a promising approach for efficient image transmission in wireless network, and attracted increasing research interests from academic. Compared with the conventional communication paradigm based on Shannon's theory, semantic communication aims to prioritize to preserve meaningful semantic information over focusing on the accuracy of transmitted symbols, which can significantly reduce the amount of data to be transmitted and improve the communication efficiency [1].

A major challenge in semantic communication for image transmission is to effectively extract semantic information at the transmitter while accurately reconstructing it at the receiver under limited communication conditions. To overcome

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

Q. Yang is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China (e-mail: qianqianyang20@zju.edu.cn).

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

Y. Deng is with the Department of Engineering, Kings College London, London, U.K (e-mail: yansha.deng@kcl.ac.uk).

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).
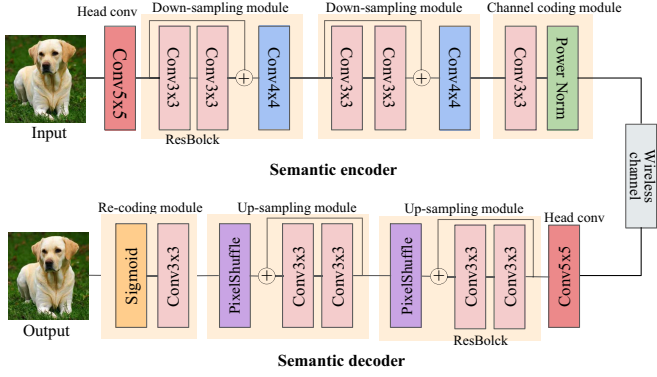
this issue, deep learning (DL) has been applied in semantic communication, due to its growing success in semantics extraction and reconstruction. In this direction, the authors in [2], [3] proposed a DL base joint source-channel coding (DeepJSCC), where the encoder and decoder were designed based on autoencoder and jointly optimized for semantic information transmission to achieve a good image reconstruction quality. Moreover, a collaborative training framework for semantic communication was proposed in [4], where users could train their semantic encoder to improve the performance of unknown downstream inference tasks. However, these approaches still suffer from limitations in achieving considerable inference performance and making a good trade-off between it and image reconstruction quality according to communication conditions.

In essence, the semantic distance between two irrelevant images is large due to different main semantic information, while the distance is small enough between two nearly identical images sharing the same semantic information. Motivated by this fact, we propose to integrate contrastive learning (CL) with semantic communication and design a two-stage training procedure. The key design in this framework is that we regard the image corruption that occurs during transmission over a limited channel as a form of data augmentation operation in [5] and propose the semantic contrastive coding to reduce the semantic distance (a.k.a semantic similarity) between the original and reconstructed image while maintain the semantic distance among irrelevant images for better discrimination. Based on contrastive loss, we design the two-stage training procedure and loss functions for jointly optimizing the encoder and decoder, thereby achieving a good balance between the inference performance in the downstream task and reconstruction quality. Simulations are finally conducted to demonstrate the superiority of the proposed method over the competitive approaches.

II. SYSTEM MODEL

This paper investigates a semantic communication system for wireless image transmission, where a convoloutional neural network (CNN) based semantic encoder and decoder are deployed in the transmitter and receiver, respectively. The semantic encoder is used to extract the semantic information of input image $\boldsymbol{x} \in \mathbb{R}^{c \times h \times w}$ 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, and c, h, and w denote the number of channels,

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

Fig. 2: Illustration of the proposed semantic contrastive coding.

height, and width of the image, respectively. To simplify, we use $n = c \times h \times w$ to stand for the dimension of $\boldsymbol{x}$. Typically, $k < n$ should be satisfied to meet the bandwidth constraint, and k/n is referred to as the bandwidth compression ratio. In particular, a large bandwidth compression ratio indicates a favorable communication condition, whereas a small one denotes a limited usage of bandwidth. In addition, a power normalization layer [2] is used at the end of the semantic encoding network to satisfy the average power constraint $\frac{1}{k}\mathbb{E}[\boldsymbol{s}^*\boldsymbol{s}] \leq P$ at the transmitter, which can be written as

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

where $\boldsymbol{s}$ is the channel input signal that meets the power constraint, and * denotes the conjugate transpose. Next, $\boldsymbol{s}$ will be transmitted over the additive white Gaussian (AWGN) channel, given by

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

where $\hat{\boldsymbol{s}}$ is the received signal, and $\boldsymbol{\epsilon} \in \mathbb{C}^k$ denotes the independent and identically distributed (IID) channel noise sample, which follows symmetric complex Gaussian distribution $\mathcal{CN}(0, \sigma^2\boldsymbol{I})$ with zero mean and variance σ^2.

The semantic decoder deployed at the receiver will reconstruct the original image $\hat{\boldsymbol{x}} \in \mathbb{R}^{c \times h \times w}$ from $\hat{\boldsymbol{s}}$ according to

$$\hat{\boldsymbol{x}} = \mathcal{D}_{\theta_2}(\hat{\boldsymbol{s}}), \tag{4}$$

where $\mathcal{D}_{\theta_2}(\cdot)$ is the semantic decoding operation parameterized by θ_2. Subsequently, $\hat{\boldsymbol{x}}$ will be used to exert downstream task and obtain the inference results through the following process

$$\boldsymbol{f_x} = \mathcal{F}^b_{\phi_1}(\hat{\boldsymbol{x}}), \tag{5}$$

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

$$\hat{\boldsymbol{y}} = \mathcal{F}^{cls}_{\phi_2}(\boldsymbol{f_x}). \tag{6}$$

Since maintaining semantic information in the reconstructed image is crucial for the inference performance, especially when the channel bandwidth is limited. Therefore, it is of vital importance to design the semantic encoder and decoder, as well as the training procedure.

III. PROPOSED FRAMEWORK

In this section, we will introduce the proposed CL based semantic communication framework. Specifically, we first present the architecture of the semantic encoder and decoder, and then we will provide the details of semantic contrastive coding and the training procedure.

A. Architecture of Semantic Encoder and Decoder

The network architecture of the semantic encoder and decoder plays a critical role in the extraction of semantic information. Therefore, we do not utilize the straightforward approach of stacked convolutional layers in [2], as this simple architecture lacks this ability. The architecture of the proposed semantic encoder and decoder are presented in Fig. 1. The semantic encoder comprises a 5×5 head convolution, two down-sampling modules, and a channel coding module. Each down-sampling module includes a basic block in ResNet [6] (we refer to as ResBolck) for capturing the spatial feature of the image, and a 4×4 convolution with stride 2 for down-sampling the 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 constraint.

Moreover, we adopt a symmetrical architecture in the decoder, which consists of a 5×5 head convolution, two up-sampling modules, and a re-coding module. In the up-sampling module, ResBolcks are also used as in the encoder and we adopt the Pixel-Shuffle technology [7] to up-sample the image, as it can provide a more efficient computing paradigm and better reconstruction performance compared to transposed convolution used in [2]. The re-coding module consists of a 3×3 convolution followed by the Sigmoid activated function to generate 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 key design of semantic contrastive coding is inspired by the success of CL which employs data augmentation to generate samples with similar vision representation and minimizes the distance among them to pretrain the backbone. We modify the CL process to adapt it for the semantic communication system. Specifically, we replace the data augmentation with the process of wireless transmission, as the image corruption that occurs during the transmission can be viewed as a form of data augmentation, and the original image and reconstructed one should keep a small semantic distance for an efficient semantic communication system. Moreover, we utilize a pretrained backbone to extract features and a learnable projection network to map these features into the semantic space. In further, by incorporating the contrastive loss in semantic space, we jointly optimize the semantic encoder and decoder rather than pretraining the backbone in CL.

The details of the proposed semantic contrastive coding 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 $\hat{\boldsymbol{x}}$. The backbone network $\mathcal{F}^b_{\phi_1}(\cdot)$ is applied to $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$, which generates the feature maps $\boldsymbol{f_x} = \mathcal{F}^b_{\phi_1}(\boldsymbol{x})$ and $\boldsymbol{f_{\hat{x}}} = \mathcal{F}^b_{\phi_1}(\hat{\boldsymbol{x}})$, respectively. Next, a fully connected projection network $\mathcal{P}_\psi(\cdot)$ with learnable parameter ψ followed by a normalization operation maps the features into a semantic space defined as a hypersphere. 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 $\boldsymbol{f_x}$ and $\boldsymbol{f_{\hat{x}}}$ can be represented as $\boldsymbol{q_x} = \mathcal{P}_\psi(\boldsymbol{f_x})$ and $\boldsymbol{v_+} = \mathcal{P}_\psi(\boldsymbol{f_{\hat{x}}})$, respectively, where $\boldsymbol{q_x}$ is referred to as the *anchor*, and $\boldsymbol{v_+}$ is called the *positive*. We can use the cosine similarity between *anchor* and *positive* to define the semantic distance between $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$.

For the remaining samples $\boldsymbol{m} \in \mathcal{B}/\{\boldsymbol{x}\}$ within training batch $\mathcal{B}$, the same procedures will be followed. Specifically, we can obtain the feature map $\boldsymbol{f_m} = \mathcal{F}_b(\boldsymbol{m})$ by feding $\boldsymbol{m}$ into the backbone network, and then project $\boldsymbol{f_m}$ into the semantic space using $\boldsymbol{v_m} = \mathcal{P}_\psi(\boldsymbol{f_m})$, where we refer $\boldsymbol{v_m}$ to as the *negative*. Similar, the semantic distance between $\boldsymbol{x}$ and $\boldsymbol{m}$ can be defined as the cosine similarity between *anchor* and *negative*. The objective of semantic contrastive coding is to minimize the semantic distance between the original and reconstructed image while maximizing the semantic distance among the original image and the irrelevent images. Therefore, we can use the InfoNCE function [5] to define the semantic contrastive loss for the training batch $\mathcal{B}$, which can be expressed as

$$\mathcal{L}_{sem} = \mathbb{E}_{\boldsymbol{x}\in\mathcal{B}}\left\{ -\log \frac{\exp(\boldsymbol{q_x} \cdot \boldsymbol{v_+}/\tau)}{\sum_{\boldsymbol{m}\in\mathcal{B}/\{\boldsymbol{x}\}} \exp(\boldsymbol{q_x} \cdot \boldsymbol{v_m}/\tau)} \right\}, \quad (7)$$

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

C. Loss Function and Training Procedure

Based on the semantic contrastive coding, we design a two-stage training strategy to optimize the semantic encoder and decoder. The first stage is pre-training, where we employ the semantic contrastive coding approach to train the weights of encoder θ_1, decoder θ_2 and project network ψ simultaneously. However, it is difficult to achieve a fast convergence speed when we only optimize semantic contrastive loss. To tackle this issue, we combine it 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 apply the mean square error (MSE) function to evaluate the reconstruction loss for training batch $\mathcal{B}$, which can be expressed as

$$\mathcal{L}_{rec} = \mathbb{E}_{\boldsymbol{x}\in\mathcal{B}}\left\{ \frac{1}{n}||\boldsymbol{x} - \hat{\boldsymbol{x}}||_2^2 \right\}. \quad (8)$$

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

$$L_1 = \alpha_1 \mathcal{L}_{rec} + (1 - \alpha_1)\mathcal{L}_{sem}, \quad (9)$$

where $\alpha_1 \in [0, 1]$ is a hyper-parameter that controls the trade-off between the two part loss functions. For instance, we can set to $\alpha = k/n$ 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 is small. In contrast, as the bandwidth compression increases, the system shifts its focus towards maintaining 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, decoder, and classifier with a small learning rate to achieve considerable inference performance and reconstructed image quality. One reason of fine-tuning the classifier is that the weights of the backbone and classifier are typically trained without considering channel corruption, which causes that the outputs of the backbone network may undergo substantial changes when the reconstructed images are inputted instead of the original images. This can result in a performance degradation. Therefore, fine-tuning the classifier with the semantic encoder and decoder can mitigate this issue and help enhance the semantics transmission. The loss function of this stage can be expressed as

$$L_2 = \alpha_2 \mathcal{L}_{rec} + (1 - \alpha_2)\mathcal{L}_{Task}, \quad (10)$$

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

$$\mathcal{L}_{Task} = \mathbb{E}_{\boldsymbol{x}\in\mathcal{B}}\left\{ -\frac{1}{N_{cls}}\sum_{i=1}^{N_{cls}} y_i \log(\hat{y}_i) \right\}, \quad (11)$$

where y_i and $\hat{y}_i$ 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.
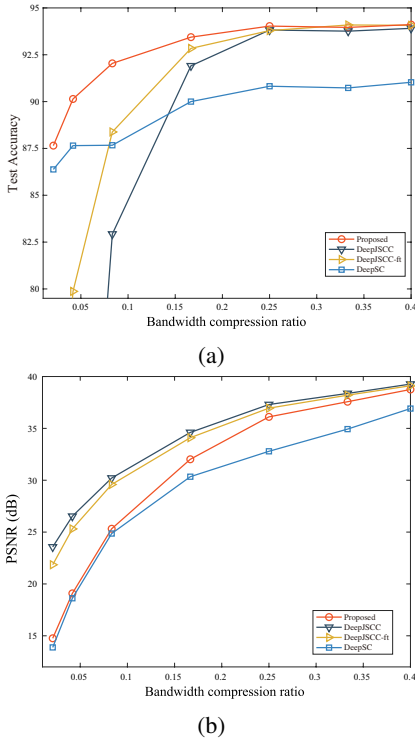
(a)

(b)

Fig. 3: Performance on CIFAR-10 versus the bandwidth compression ratio with SNR = 20dB.

(a)

(b)

Fig. 4: Performance on CIFAR-10 versus the bandwidth compression ratio with SNR = 5dB.

IV. SIMULATIONS

In order 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 comprises 50,000 images, while the test set contains 10,000 images. A pre-trained ResNet-56 [6] is used as the backbone network and classifier for downstream inference, while the structure of the semantic encoder and decoder is shown in Fig. 1. The projection network adopts a two-layer fully connected structure with an output dimension of 32. The number of training epochs for the two stages is set to 200 and 100, respectively, with a batch size of 128. Besides, we use the Adam optimizer with a learning rate of 0.01 for the first pre-training stage and 0.0001 for the second fine-tuning stage. These learning rates will be adjusted every 50 epochs with a decay factor of 0.5. We compare the proposed method with the following DL-based semantic communication methods,

- DeepJSCC [2]: DL-based source-channel joint coding that maps the original input to the channel input through the structure of an autoencoder.
- DeepJSCC-ft: an extension of DeepJSCC with the second stage fine-tuning strategy of our proposed method, in which the encoder, decoder and classifier are updated with a small learning rate.
- DeepSC [4]: a DL-based semantic coding framework that trains the semantic encoder and decoder with both semantic and observation losses to achieve efficient semantic information transmission. For a fair comparison, we freeze its encoder and decoder after training and retrained the classifier with a small learning rate.
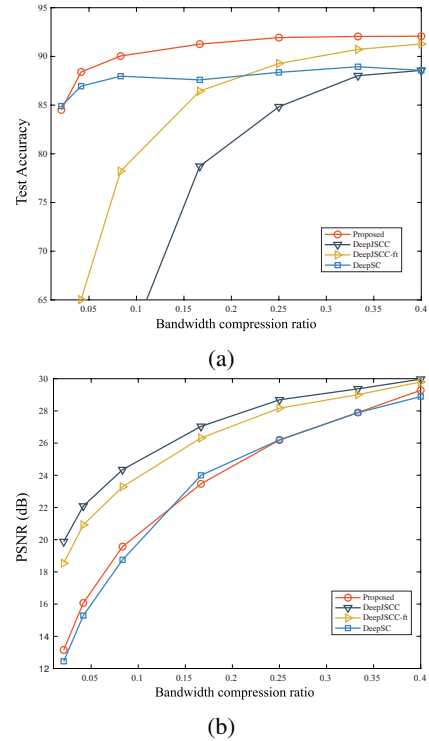
Fig. 3a compares the accuracy performance of the proposed method, DeepJSCC, DeepJSCC-ft, and DeepSC, where SNR is set to 20dB, and the bandwidth compression ratio k/n varies from 1/48 to 1/2.5. From this figure, we can find that the proposed method consistently outperforms or matches the competitive methods in accuracy. Specifically, when the compression ratio is 1/2.5, the proposed method, DeepJSCC, and DeepJSCC-ft achieve the test accuracy of about 94%, whereas DeepSC performs poorly with the accuracy of only 91%. As the bandwidth compression ratio decreases, the proposed method still maintains a comparable accuracy performance. For instance, the proposed method can achieve accuracy of 90.14% and 87.65% at bandwidth compression ratios of 1/24 and 1/48, respectively, which outperforms DeepSC by about 2.5% and 1% at the corresponding bandwidth compression ratios and also shows an accuracy gain of up to 56% over DeepJSCC. These results indicate that the proposed framework can effectively extract semantic information to meet the requirements of downstream task and removes irrelevant redundant information to ensure the semantic information can be successfully transmitted, especially when channel bandwidth is limited.

Fig. 3b presents the peak signal-to-ration (PSNR) of the proposed and the three completing methods , where SNR is set to 20dB and the bandwidth compression ratio varies from 1/48 to 1/2.5. As shown in the figure, we can find that as the bandwidth compression ratio increases, the PSNRs of all methods get improved. Although the proposed method sacrifices some image quality to prioritize semantic information when the bandwidth compression ratio is low, it can quickly catch up

Original	DeepJSCC	DeepJSCC-ft	DeepSC	Proposed

| PSNR\|MS-SSIM | 20.23dB\|0.83 | 20.93dB\|0.82 | 16.85dB\|0.68 | 19.83dB\|0.82 |

(a)

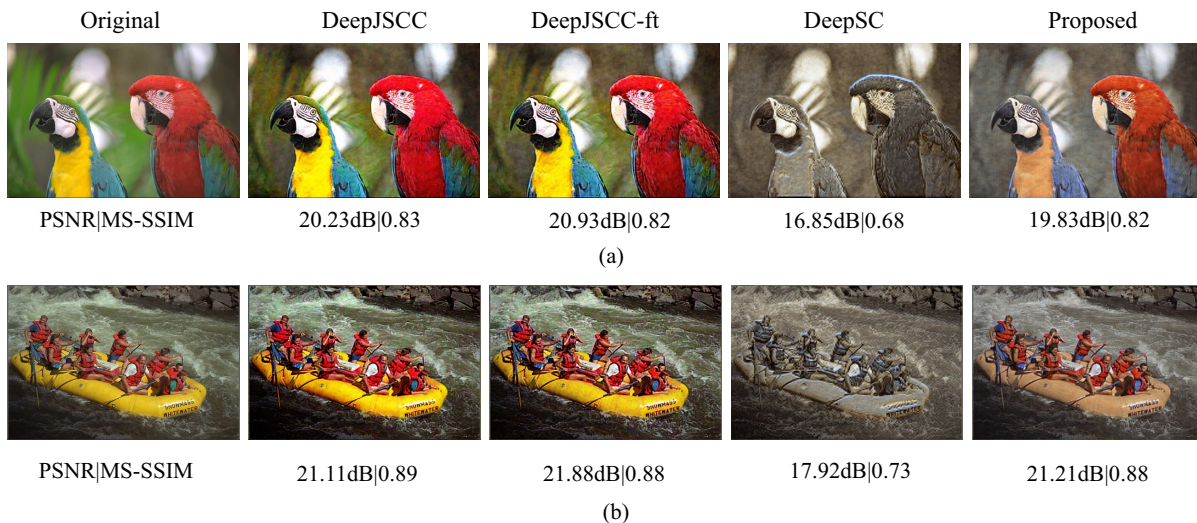| PSNR\|MS-SSIM | 21.11dB\|0.89 | 21.88dB\|0.88 | 17.92dB\|0.73 | 21.21dB\|0.88 |

(b)

Fig. 5: Visualization comparison of several methods on the Kodak dataset, where SNR = 20dB and bandwidth compression ratio is 1/48.

with DeepJSCC's PSNR at higher compression ratios. Specifically, the proposed method achieves a PSNR of 38.75dB, which is close to the 39.27dB of DeepJSCC, and outperforms DeepSC with 36.91dB when the bandwidth compression ratio is 1/2.5. These results indicate that the proposed method can prioritize to transmit semantic information over irrelevant background information to ensure the performance of downstream task in bandwidth-limited scenarios, and meanwhile transmit enough background information to obtain good image quality when the bandwidth is not a bottleneck. This further demonstrates the effectiveness of the proposed method.

Fig. 4a and Fig. 4b present the performance comparison of several methods under poor channel conditions in terms of accuracy and PSNR, respectively. Specifically, both figures consider a low SNR of 5dB, and the bandwidth compression ratio varies from 1/48 to 1/2.5. From Fig. 4a, we can observe that the proposed methods still shows the superiority in terms of accuracy compared to the three competitive methods, indicating its robustness in low SNR scenarios. From Fig. 4b, we can find the proposed framework can adaptively sacrifice the global information to obtain comparable semantic performance when the bandwidth compression ratio is low, and meanwhile obtain enough reconstructed quality in terms of PSNR as bandwidth compression ratio decreased. These results in both figures further verify the effectiveness and robustness of the proposed method in low SNR scenarios.

We also provide the visualization comparison of several methods on the Kodak dataset in Fig. 5, where the encoder and decoder are trained on the STL10 dataset, SNR is set to 20dB and the bandwidth compression ratio is 1/48. From this figure, we can observe that the proposed can effectively remove redundant background information and meanwhile preserve the main semantic information, resulting in less image corruption in semantic regions (e.g., macaws and rafters in this figures) compared to the competitive methods. Moreover, the proposed method achieves similar PSNR and multi-scale structural similarity (MS-SSIM) performance with DeepJSCC and DeepJSCC-ft, indicating the effectiveness of the proposed

method in reconstructing semantic information. These results further demonstrate the superiority of the proposed method in achieving leading accuracy in downstream tasks over the compared methods.

V. CONCLUSION

In this paper, we proposed a CL-based semantic communication framework for wireless image transmission. The framework incorporated semantic contrastive coding and a two-stage training procedure to enhance the extraction of semantic information, and in order toachieve a better trade-off between reconstruction quality and the performance of downstream task. We evaluated the effectiveness of the proposed methods through simulations on CIFAR-10 and Kodak datasets, which simulated results demonstrated the superiority of our approach over existing methods.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.