Task-Oriented Multi-User Semantic Communications

Huiqiang Xie, *Student Member, IEEE,* Zhijin Qin, *Senior Member, IEEE,* Xiaoming Tao, *Member, IEEE,*
and Khaled B. Letaief, *Fellow, IEEE*

Abstract—While semantic communications have shown the potential in the case of single-modal single-users, its applications to the multi-user scenario remain limited. In this paper, we investigate deep learning (DL) based multi-user semantic communication systems for transmitting single-modal data and multimodal data, respectively. We will adopt three intelligent tasks, including, image retrieval, machine translation, and visual question answering (VQA) as the transmission goal of semantic communication systems. We will then propose a Transformer based unique framework to unify the structure of transmitters for different tasks. For the single-modal multi-user system, we will propose two Transformer based models, named, DeepSC-IR and DeepSC-MT, to perform image retrieval and machine translation, respectively. In this case, DeepSC-IR is trained to optimize the distance in embedding space between images and DeepSC-MT is trained to minimize the semantic errors by recovering the semantic meaning of sentences. For the multimodal multi-user system, we develop a Transformer enabled model, named, DeepSC-VQA, for the VQA task by extracting text-image information at the transmitters and fusing it at the receiver. In particular, a novel layer-wise Transformer is designed to help fuse multimodal data by adding connection between each of the encoder and decoder layers. Numerical results will show that the proposed models are superior to traditional communications in terms of the robustness to channels, computational complexity, transmission delay, and the task-execution performance at various task-specific metrics.

Index Terms—Deep learning, semantic communications, multimodal fusion, multi-user communications, Transformer.

I. Introduction

Conventional communication systems are regarded as transmission pipes, in which the data are collected at the transmitters and reconstructed at the receivers. As we step into the era of connected intelligence [1], the widely deployed devices have been generating unprecedented amounts of multimodal data to serve various tasks, which makes conventional communications a new bottleneck and performance limit. There exist two ways to address this problem: 1) evolution of hardware to enlarge the system capacity and transmission rate, e.g., millimeter wave/terahertz communications [2], [3], massive antennas array [4], and reconfigurable intelligent surfaces [5];

Huiqiang Xie and Zhijin Qin are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK (e-mail: h.xie@qmul.ac.uk, z.qin@qmul.ac.uk).

Xiaoming Tao is with the Department of Electronic Engineering and also with the Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: taoxm@tsinghua.edu.cn).

Khaled B. Letaief is with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, and also with Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: eekhaled@ust.hk).

2) improvement of software to optimize the utilization of communication resources, e.g., data compression [6], resource multiplexing [7], and semantic communications [8]. In this work, we investigate the second approach. Moreover, we will focus on semantic communications, the new emerging communication paradigm, which has shown its superiority in handling the massive volume of data.

Semantic communications are content-aware, task-oriented, and semantic-related, in which only important, relevant, and useful information to the users/applications are extracted from a large amount of data and delivered to the destinations. The existing works on semantic communications can be categorized into two parts: full data reconstruction and task execution.

For the data reconstruction, semantic communications generally extract the global semantic information behind data and reconstruct the data based on the received semantic information. Farsad *et al.* [9] designed the initial deep joint source-channel coding for text transmission, in which the text sentences are encoded into fixed-length bit streams over simple channel environments. With the depth exploration in the semantic communications, Xie *et al.* [10] developed more powerful joint semantic-channel coding, named DeepSC, to encode text information into various length over complex channels. Moreover, Xie *et al.* [11] also proposed an environment-friendly semantic communication system, named L-DeepSC, for the capacity-limited devices. Besides, Bourtsoulatze *et al.* [12] investigated the initial deep image transmission semantic communication systems, in which the semantic and channel coding are optimized jointly. Kurka *et al.* [13] extended Bourtsoulatze's work with the channel feedback to improve the quality of image reconstruction. Weng *et al.* [14] developed an attention mechanism based semantic communication systems to reconstruct speech signals.

For the task-specific applications, only the semantic information useful for serving the task execution is extracted at the transmitter, which will be directly used for the decision making at the receiver. Lee *et al.* [15] developed an image classification-oriented semantic communications for improving the recognition accuracy rather than performing image reconstruction and classification separately. Jankowski *et al.* [16] considered image based re-identification for person or cars as the communication task, in which two schemes (digital and analog) are proposed to improve the retrieval accuracy. Except from image based tasks, Weng *et al.* [17] designed speech recognition-oriented semantic communications, named, DeepSC-SR, to directly recognize the speech signals into

texts. The prior works explore the possibility of semantic communications for transmitting signals in a single-modal single-user system. However, in practice, we must gather multimodal data from different users/devices, transmit over the air, and process/fuse multimodal data at the receiver. This motivates us to develop a multi-user semantic communication system to support multimodal data transmission. Our initial design of the MU-DeepSC is for serving the visual question answering (VQA) task to improve the answer accuracy [18], which adopts Long Short Term Memory (LSTM) for the text transmitter and Convolutional Neural Network (CNN) for the image transmitter. However, a unified framework to support various tasks with multimodal data is still missing in multi-user semantic communications.

Particularly, single-modal multi-user semantic communications represent the extension of single-modal single-user semantic communications, in which multiple single-modal intelligent tasks can be performed simultaneously but each user is only associated with one intelligent task. Multimodal multi-user semantic communications employ more than one user to serve one multimodal intelligent task, which is suitable for the emerging autonomous scenarios in daily life [19] and industry [20], i.e., autonomous checkout at retail stores [21], intelligent control at smart home [22], and human activity recognition in smart healthcare [23]. Such scenarios are achieved by collecting multimodal data from the various sensors so as to provide the information in a complementary manner and fuse them at the server/cloud. For the design of multi-user semantic communications, we face the following challenges:

Q1: *How to extract semantic information at the transmitter for both single-modal and multimodal multi-user semantic communications?*

Q2: *How to reduce the interference from other users for both single-modal and multimodal multi-user semantic communications?*

Q3: *How to process/fuse the received semantic information at the receiver for multi-user semantic communications to transmit multimodal data?*

In this paper, we investigate task-oriented multi-user semantic communications for transmitting data with single modality and multiple modalities by considering two types of sources: image and text. We choose image retrieval and machine translation for transmission data with single-modality, and one of the most challenging tasks, namely, the visual question answering (VQA) task, for illustrating transmission with multimodal data. The main contributions of this paper are summarized as follows:

- We propose a Transformer [24] based transmitter structure, which is applicable for both text and image transmission by effectively extracting semantic information for different tasks. This addresses the aforementioned *Q1*.
- We demonstrate the efficient methods for training the proposed structure. In particular, the transmitters and receiver in the proposed frameworks are trained jointly to eliminate distortion from the channels and interference from other users. This addresses the aforementioned *Q2*.
- Based on the proposed structure, we propose three different deep learning (DL) enabled multiuser semantic communication frameworks, named DeepSC-IR for image retrieval, DeepSC-MT for machine translation, and DeepSC-VQA for VQA. Specially, we propose a novel layer-wise Transformer, which can exploit more text information to guide image information, to fuse the text and image information. This addresses the aforementioned *Q3*.
- Based on extensive simulation results, the proposed frameworks outperform the traditional communication systems with lower requirements on the communication resources and improved system robustness at the low SNR regimes.

The rest of this paper is organized as follows. The related works of selected tasks and preliminaries are briefly reviewed in Section II. The system model is introduced in Section III. The proposed single-modal multi-user semantic communications are proposed in Section IV. Section V details the proposed multimodal multi-user semantic communications. Numerical results are presented in Section VI to show the performance of the proposed frameworks. Finally, Section VII concludes this paper.

Notation: $\mathbb{C}^{n \times m}$ and $\mathbb{R}^{n \times m}$ represent sets of complex and real matrices of size $n \times m$, respectively. Bold-font variables denote matrices or vectors. $x \sim \mathcal{CN}(\mu, \sigma^2)$ means variable x follows a circularly-symmetric complex Gaussian distribution with mean μ and covariance σ^2. $(\cdot)^{\mathrm{T}}$ and $(\cdot)^{\mathrm{H}}$ denote the transpose and Hermitian, respectively. $\Re\{\cdot\}$ and $\Im\{\cdot\}$ refer to the real and imaginary parts of a complex number.

II. RELATED WORKS AND PRELIMINARIES

In this section, we will first introduce the definitions of the three intelligent tasks, including image retrieval, machine translation, and VQA. We then briefly review the related works on the three tasks. Because the designed models in the next sections mainly consist of the Transformer network, we will briefly introduce the preprocessing for image and text, and the main components for the Transformer network.

A. Image Retrieval

The image retrieval task aims to identify the top-k similar images by matching the sent image with those stored in a large server, and returns the similar ones to users. For example, the user uploads a dress image to Amazon app and wishes to find similar dress products. Such image retrieval tasks cannot be performed locally due to the centralized database.

Modern methods for image retrieval typically rely on DL based models by extracting compact image-level features [25] for image match or classification. Recent techniques mainly focus on two parts: deep network architectures and training algorithms. The deep network architectures include single feedforward pass models [26], multiple feedforward pass models [27], attention based models [28], and deep hashing embedding based models [29]. While the training algorithms focus on classification based learning [30], metric based learning [31], and unsupervised-based learning [32].

B. Machine Translation

One core of communications is to transmit the meanings behind the text, however, one of the major obstructs for communications is the different grammars and presentations for different languages. Therefore, for the machine translation task, the intention is that the transmitter sends one language, i.e., Chinese, and the receiver directly receives the desired language, i.e., English, which aims to broken the obstruct of communications and improve communication efficiency.

The recent successful approaches for machine translation problems are mostly based on the classic encoder-decoder structure [33], in which the encoder extracts the sentence-level intermediate features at source language and the decoder provides the entire sentence at target language based on the intermediate features. The representative models include CNN based models [34], Transformer based models [24], and RNN based models, i.e., LSTM networks [35] and Gated Recurrent Units (GRU) networks [36].

C. Visual Question Answering

In the VQA task, the semantic information from different users is correlated. One user may transmit the vision information collected by a camera while the other user sends the text information collected by a sensor. Then, the transmitted vision and text information from different users is employed to carry out the answers at the receiver.

The core of VQA tasks is multimodal data fusion techniques [37], in which the image and questions in text are first represented as global features and then fused by a multimodal fusion model to predict the answer. Recent approaches adopt the visual attention mechanism by attending image features with given question features, which include multimodal bilinear pooling methods [38], stacked attention network [39], bottom-up and top-down attention mechanism [40], and co-attention network [41].

D. Preliminaries

The *text preprocessing* includes two parts: tokenize and embedding. The input sentence is first splitted into scalar-wise tokens, each representing one word or one sub-word. These scalar-wise tokens are then mapped into vector-shaped tokens with learnable word vectors and used as the input to the Transformer. The *image preprocessing* also includes two parts: patchify and project. The input image is first decomposed into fixed-sized patches, e.g. 16x16. Each patch is linearly projected into vector-shaped tokens and used as an input to the Transformer. An extra learnable <CLS> token is added to the input sequence such that its corresponding output token serves as a global representation for the input sequence. The location prior is incorporated by adding a learnable one-dimension (1-D) positional encoding vector to the input tokens.

Transformer network consists of the encoder layers and decoder layers. Each encoder layer includes two main blocks: 1) a Multi-Headed Self Attention layer, which applies a self-attention operation to different projections of input tokens; and 2) a Feed-Forward layer. The decoder layer includes three main blocks: 1) a Multi-Headed Self Attention layer;

2) a Multi-Headed Guided Attention layer, which applies a attention operation to the projections of input tokens and the output tokens of encoder; and 3) a Feed-Forward layer. All blocks are preceded by layer normalization and followed by a skip connection.

III. SYSTEM MODEL

As shown in Fig. 1, we consider the multi-user semantic communication system, which consists of one receiver equipped with M antennas and K single-antenna transmitters. We will focus on the multi-user semantic communication system with single-modal data and multimodal data to transmit, respectively. The single-modal multi-user scenario means that each user transmits independent semantic information to perform its own task. The multimodal multi-user scenario indicates that the data from different users are semantically complementary.

A. Semantic Transmitter

As shown in Fig. 1, we denote the source data of the k-th user as $s_k^{\mathcal{Q}}$ with modality $\mathcal{Q} \subseteq \{\mathcal{I} : image, \mathcal{T} : text, \mathcal{V} : video, \mathcal{S} : speech\}$, where each source contains the semantic information. The semantic information is extracted first by

$$z_k^{\mathcal{Q}} = S\left(s_k^{\mathcal{Q}}; \boldsymbol{\alpha}_k^{\mathcal{Q}}\right), \tag{1}$$

where $z_k^{\mathcal{Q}} \in \mathbb{R}^{L_S \times 1}$ is the semantic information with length L_S and $S\left(; \boldsymbol{\alpha}_k^{\mathcal{Q}}\right)$ is the modality $\mathcal{Q}$ semantic encoder for the k-th user with learnable parameters $\boldsymbol{\alpha}_k^{\mathcal{Q}}$. Due to the limited communication resource and complex communication environment for wireless communications, the semantic information of the k-th user is compressed by

$$x_k^{\mathcal{Q}} = C\left(z_k^{\mathcal{Q}}; \boldsymbol{\beta}_k^{\mathcal{Q}}\right), \tag{2}$$

where $x_k^{\mathcal{Q}} \in \mathbb{C}^{L_C \times 1}$ is the transmitted complex signal with length $L_C < L_S$ and $C\left(; \boldsymbol{\beta}_k\right)$ is the k-th user joint source-channel (JSC) encoder for modality $\mathcal{Q}$ with learnable parameters, $\boldsymbol{\beta}_k$. The neural JSC encoder in semantic communications compresses semantic information to reduce the number of transmitted symbols, as well as improve the robustness to channel variations.

B. Semantic Receiver

When the transmitted signal passes a multiple-input multiple-output (MIMO) physical channel, the received signal, $\mathbf{Y} \in \mathbb{C}^{M \times L_C}$, at the receiver can be expressed as

$$\mathbf{Y} = \mathbf{H}\mathbf{X} + \mathbf{N}, \tag{3}$$

where $\mathbf{X}^T = \left[x_1^{\mathcal{Q}}, x_2^{\mathcal{Q}}, \cdots, x_K^{\mathcal{Q}}\right] \in \mathbb{C}^{L_C \times K}$ denotes transmit symbols from all K users, $\mathbf{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_K] \in \mathbb{C}^{M \times K}$ is the channel matrix between the BS and users. For the Rayleigh fading channel, the channel coefficient follows $\mathcal{CN}(0, 1)$; for the Rician fading channel, it follows $\mathcal{CN}(\mu, \sigma^2)$ with $\mu = \sqrt{r/(r+1)}$ and $\sigma = \sqrt{1/(r+1)}$, where r is the Rician coefficient. $\mathbf{N} \in \mathbb{C}^{M \times L_C}$ denotes the circular symmetric

Fig. 1. The framework of multi-user semantic communication systems

Gaussian noise. The elements of $\mathbf{N}$ are i.i.d with zero mean and variance σ_n^2, and SNR is defined by $\sum_k \left\| \boldsymbol{h}_k \boldsymbol{x}_k^{\mathcal{Q}} \right\|^2 / \sigma_n^2$.

Subsequently, the transmission signals are recovered by the linear minimum mean-squared error (L-MMSE) detector with the estimated channel state information (CSI),

$$\widehat{\mathbf{X}} = \widehat{\mathbf{H}}^H \left(\widehat{\mathbf{H}} \widehat{\mathbf{H}}^H + \sigma_n^2 \mathbf{I} \right)^{-1} \mathbf{Y}, \tag{4}$$

where $\widehat{\mathbf{X}}^T = \left[\hat{\boldsymbol{x}}_1^{\mathcal{Q}}; \hat{\boldsymbol{x}}_2^{\mathcal{Q}}; \cdots; \hat{\boldsymbol{x}}_K^{\mathcal{Q}} \right] \in \mathbb{C}^{L_C \times K}$ is the recovered transmission signals, $\widehat{\mathbf{H}} = \mathbf{H} + \Delta\mathbf{H}$ is the estimated CSI, in which $\Delta\mathbf{H}$ is the estimation error with $\Delta\mathbf{H} \in \mathcal{CN}(0, \sigma_e^2)$.

The semantic information from the k-th user, $\hat{\boldsymbol{z}}_k^{\mathcal{Q}} \in \mathbb{R}^{L_S \times 1}$, is recovered by the JSC decoder as

$$\hat{\boldsymbol{z}}_k^{\mathcal{Q}} = C^{-1} \left(\hat{\boldsymbol{x}}_k^{\mathcal{Q}}; \gamma_k^{\mathcal{Q}} \right), \tag{5}$$

where $C^{-1} \left(\hat{\boldsymbol{x}}_k^{\mathcal{Q}}; \gamma_k^{\mathcal{Q}} \right)$[a] is JSC decoder for the k-th user with the modality $\mathcal{Q}$ and the learned parameters $\gamma_k^{\mathcal{Q}}$. The JSC decoder aims to decompress the semantic information while mitigating the effects of channel distortion and inter-user interference. According to the independence of transmission semantic information, we will have the single-modal semantic receiver and the multimodal semantic receiver.

1) Single-Modal Semantic Receiver: For single-modal semantic transmission, the semantic information from each user is exploited to perform different tasks independently. The recovered semantic information is employed for the task of the k-th user by

$$\boldsymbol{p}_k^{\mathcal{Q}} = S^{-1} \left(\hat{\boldsymbol{z}}_k^{\mathcal{Q}}; \boldsymbol{\varphi}_k^{\mathcal{Q}} \right), \tag{6}$$

where $\boldsymbol{p}_k^{\mathcal{Q}}$ is the result of the task, i.e., the translated sentence for the machine learning task, and retrieval results for the image retrieval task. $S^{-1}(; \boldsymbol{\varphi}_k^{\mathcal{Q}})$ is the modality $\mathcal{Q}$ semantic decoder for the k-th user with learning parameters $\boldsymbol{\varphi}_k^{\mathcal{Q}}$.

2) Multimodal Semantic Receiver: With the multimodal semantic information, the final task is performed directly by

[a]In order to reduce the number of representation symbols, we use $\cdot^{-1}$ here to represent the decoder.

merging the semantic information from different users. This is expressed by

$$\boldsymbol{p} = S^{-1} \left(\hat{\boldsymbol{z}}_1^{\mathcal{Q}}, \hat{\boldsymbol{z}}_2^{\mathcal{Q}}, \cdots, \hat{\boldsymbol{z}}_K^{\mathcal{Q}}; \boldsymbol{\varphi}_{(1,2,\cdots,K)} \right), \tag{7}$$

where $\boldsymbol{p}$ is the results of the multimodal task and $S^{-1} \left(; \boldsymbol{\varphi}_{(1,2,\cdots,K)} \right)$ is the multimodal semantic decoder with learnable parameters $\boldsymbol{\varphi}_{(1,2,\cdots,K)}$.

IV. SINGLE-MODAL MULTI-USER SEMANTIC COMMUNICATIONS

In this section, we focus on the multi-user semantic communication system to transmit single-modal data from multiple users. We propose semantic communication systems for the image retrieval task (i.e., DeepSC-IR), and the machine translation task (i.e., DeepSC-MT). Particularly, we adopt the vision Transformer for image understanding and text Transformer for text understanding, in which the vision Transformer and text Transformer are assumed to have the same network structure.

A. Image Retrieval Task

Assume that $\mathcal{D}_k^{\mathcal{I}} = \left\{ \left(\boldsymbol{s}_{k,j}^{\mathcal{I}}, l_{k,j}^{\mathcal{I}} \right) \right\}_{j=1}^{D}$ with size D is the training image dataset for the k-th user, where $\boldsymbol{s}_{k,j}^{\mathcal{I}}$ and $l_{k,j}^{\mathcal{I}}$ are the j-th image and its corresponding label in $\mathcal{D}_k^{\mathcal{I}}$, respectively. $S_{\mathrm{IR}} \left(; \boldsymbol{\alpha}_k^{\mathcal{I}} \right)$, $C_{\mathrm{IR}} \left(; \boldsymbol{\beta}_k^{\mathcal{I}} \right)$, and $C_{\mathrm{IR}}^{-1} \left(; \boldsymbol{\gamma}_k^{\mathcal{I}} \right)$ represent the semantic encoder, JSC encoder, and JSC decoder of the i-th user for the image retrieval task, respectively.

1) Model Description: The proposed image retrieval network is shown in Fig. 2. Specifically, the DeepSC-IR transmitter consists of an image semantic encoder to extract image semantic information to be transmitted and a JSC encoder to compress the semantic information, where the semantic encoder includes multiple vision Transformer layers and the JSC encoder uses dense layers with different units. Especially, we choose only the <CLS> vector-token to be transmitted as it represents the global image information. After transmitting and performing signal detection, the DeepSC-IR receiver employs the JSC decoder with different units to recover the transmitted image semantic information.

Fig. 2. The network structure of single-modal multi-user semantic communications, which contains the DeepSC-IR transceiver and DeepSC-MT transceiver.

The recovered semantic information after the JSC decoder at the receiver can be used to match the other image semantic information in the database by computing the euclidean distance to find similar images as

$$d(\boldsymbol{z}_{k,j}^{\mathcal{I}}, \boldsymbol{z}_{k,i}^{\mathcal{I}}) = \left\| \boldsymbol{z}_{k,j}^{\mathcal{I}} - \boldsymbol{z}_{k,i}^{\mathcal{I}} \right\|_2. \tag{8}$$

The euclidean distance becomes the cosine similarity when $\boldsymbol{z}_{k,j}^{\mathcal{I}}$ and $\boldsymbol{z}_{k,i}^{\mathcal{I}}$ are l^2 normalized.

2) Training Algorithm: As shown in Algorithm 1, the training process of the DeepSC-IR consists of two phases due to different loss functions. The first phase is to train the semantic encoder, and the second phase is to train the JSC codec.

In the first phase, the semantic encoder will be trained by the function, `Train Semantic Encoder`. Different from other tasks, image retrieval is performed by computing the distance between images to return similar images. Therefore, we choose metric learning, as one type of self-supervised learning, as the learning paradigm. Such paradigm aims at minimizing the distance between images belonging to the same category and maximizing the distance between images belonging to different categories. The loss function is expressed by

$$\begin{aligned}
\mathcal{L}_{\mathrm{IR}} = & \mathbb{E}\left[\sum_{l_{k,j}^{\mathcal{I}}=l_{k,i}^{\mathcal{I}}} \left(1 - (\boldsymbol{z}_{k,j}^{\mathcal{I}})^{\mathrm{T}} \boldsymbol{z}_{k,i}^{\mathcal{I}}\right) \right] \\
& + \mathbb{E}\left[\sum_{l_{k,j}^{\mathcal{I}} \neq l_{k,i}^{\mathcal{I}}} \left((\boldsymbol{z}_{k,j}^{\mathcal{I}})^{\mathrm{T}} \boldsymbol{z}_{k,i}^{\mathcal{I}} - \xi\right)_+ \right],
\end{aligned} \tag{9}$$

where the operator $(x)_+$ returns $\max(x, 0)$, $\boldsymbol{z}_{i,j}$ is the image semantic information, ξ is a constant margin to prevent the training signal from being overwhelmed by easy negatives. After training the semantic encoder with (9), the semantic encoder becomes capable of extracting semantic image information, which returns a smaller euclidean distance if they are from images within the same category.

In order to compress semantic redundancy while overcoming the distortion from the channels, the JSC codec is trained in the second phase. The mean-squared error (MSE) is employed as the loss function to minimize the difference between the

transmitted and recovered semantic image information, which is represented as

$$\mathcal{L}_{\mathrm{MSE}} = \mathbb{E}\left[\left\| \hat{\boldsymbol{z}}_{k,j}^{\mathcal{I}} - \boldsymbol{z}_{k,j}^{\mathcal{I}} \right\|_2^2 \right], \tag{10}$$

where $\hat{\boldsymbol{z}}_{k,j}^{\mathcal{I}}$ is the semantic image information recovered at receiver and $\boldsymbol{z}_{k,j}^{\mathcal{I}}$ is the transmitted semantic image information. By minimizing the $\mathcal{L}_{\mathrm{MSE}}$, the JSC codec will learn to compress and decompress semantic image information for fewer transmitted symbols while keeping the semantic recovery accurately by dealing with the distortion and interference jointly from the channels and inter-users.

B. Machine Translation Task

Assume $\mathcal{D}_k^{\mathcal{T}} = \left\{ (\boldsymbol{s}_{k,j}^{\mathcal{T}}, \boldsymbol{p}_{k,j}^{\mathcal{T}}) \right\}_{j=1}^D$ with size D as the training text dataset for the k-th user, where $\boldsymbol{s}_{k,j}^{\mathcal{T}}$ and $\boldsymbol{p}_{k,j}^{\mathcal{T}}$ are the j-th sentence in the source language and the translated sentence in the target language, respectively. $\boldsymbol{s}_{k,j}^{\mathcal{T}}[n]$ and $\boldsymbol{p}_{k,j}^{\mathcal{T}}[n]$ represent the n-th word in sentence $\boldsymbol{s}_{k,j}^{\mathcal{T}}$ and $\boldsymbol{p}_{k,j}^{\mathcal{T}}$, respectively. $S_{\mathrm{MT}}\left(; \boldsymbol{\alpha}_k^{\mathcal{T}}\right)$, $C_{\mathrm{MT}}\left(; \boldsymbol{\beta}_k^{\mathcal{T}}\right)$, $C_{\mathrm{MT}}^{-1}\left(; \boldsymbol{\gamma}_k^{\mathcal{T}}\right)$, and $S_{\mathrm{MT}}^{-1}\left(; \boldsymbol{\varphi}_k^{\mathcal{T}}\right)$ represent the semantic encoder, JSC encoder, JSC decoder, and semantic decoder of the k-th user for the machine translation task, respectively.

1) Model Description: The proposed machine translation network is shown in Fig. 2. The transmitter includes a text semantic encoder and a text JSC encoder to extract and compress the semantic text information, respectively, where the text semantic encoder adopts multiple Transformer encoder layers and the designed text JSC encoder in Fig. 2 is with multiple dense layers. At the receiver, the designed text JSC decoder recovers the semantic text information from distorted signals. Subsequently, the semantic decoder consists of multiple Transformer decoder layers to derive the translated sentence based on the recovered semantic text information.

2) Training Algorithm: As shown in Algorithm 2, the training process of DeepSC-MT consists of three phases: `Train Semantic Codec`, `Train JSC Codec`, and `Train Whole Network`.

The first phase is `Train Semantic Codec`. The semantic codec, $S_{\mathrm{MT}}\left(; \boldsymbol{\alpha}_k^{\mathcal{T}}\right)$ and $S_{\mathrm{MT}}^{-1}\left(; \boldsymbol{\varphi}_k^{\mathcal{T}}\right)$, will be trained firstly by the cross-entropy (CE) loss function, which enables the

---

Algorithm 1: DeepSC-IR Training Algorithm.

Initialization: The training dataset $\mathcal{D}_k^{\mathcal{I}}$ and the batch size B.

1 Function `Train Semantic Encoder()`:

    Input: Choose mini-batch data $\left\{\left(\boldsymbol{s}_{k,j}^{\mathcal{I}}, l_{k,j}^{\mathcal{I}}\right)\right\}_{j=n}^{n+B}$ from $\mathcal{D}_k^{\mathcal{I}}$.

2 $\left\{S_{\text{IR}}\left(\boldsymbol{s}_{k,j}^{\mathcal{I}}; \boldsymbol{\alpha}_k^{\mathcal{I}}\right)\right\}_{j=n}^{n+B} \rightarrow \left\{\boldsymbol{z}_{k,j}^{\mathcal{I}}\right\}_{j=n}^{n+B}$,

3   Compute the $\mathcal{L}_{\text{IR}}$ by (9) with $\left\{\boldsymbol{z}_{k,j}^{\mathcal{I}}\right\}_{j=n}^{n+B}$,

4   Train $\boldsymbol{\alpha}_k^{\mathcal{I}} \rightarrow$ Gradient descent with $\mathcal{L}_{\text{IR}}$,

    Return: $S_{\text{IR}}\left(;\boldsymbol{\alpha}_k^{\mathcal{I}}\right)$.

5 Function `Train JSC Codec()`:

    Input: The semantic image information $\left\{\boldsymbol{z}_{k,j}^{\mathcal{I}}\right\}_{j=n}^{n+B}$.

6   **for** $j = n \rightarrow n + B$ **do**

7   **Transmitter**:

8 $C_{\text{IR}}\left(\boldsymbol{z}_{k,j}^{\mathcal{I}}; \boldsymbol{\beta}_k^{\mathcal{I}}\right) \rightarrow \boldsymbol{x}_{k,j}^{\mathcal{I}}$,

9   Transmit $\boldsymbol{x}_{k,j}^{\mathcal{I}}$ over the channel,

10   **Receiver**:

11   Receive $\mathbf{Y}$,

12   MIMO detection by (4) to get $\hat{\boldsymbol{x}}_{k,j}^{\mathcal{I}}$,

13 $C_{\text{IR}}^{-1}\left(\hat{\boldsymbol{x}}_{k,j}^{\mathcal{I}}; \boldsymbol{\gamma}_k^{\mathcal{I}}\right) \rightarrow \hat{\boldsymbol{z}}_{k,j}^{\mathcal{I}}$,

14   Compute the $\mathcal{L}_{\text{MSE}}$ by (10) with $\boldsymbol{z}_{k,j}^{\mathcal{I}}$, $\hat{\boldsymbol{z}}_{k,j}^{\mathcal{I}}$,

15   Train $\boldsymbol{\beta}_k^{\mathcal{I}}, \boldsymbol{\gamma}_k^{\mathcal{I}} \rightarrow$ Gradient descent with $\mathcal{L}_{\text{MSE}}$,

    Return: $C_{\text{IR}}\left(;\boldsymbol{\beta}_k^{\mathcal{I}}\right), C_{\text{IR}}^{-1}\left(;\boldsymbol{\gamma}_k^{\mathcal{I}}\right)$.

Algorithm 2: DeepSC-MT Training Algorithm.

Initialization: The training dataset $\mathcal{D}_k^{\mathcal{T}}$ and the batch size B.

1 Function `Train Semantic Codec()`:

    Input: Choose mini-batch data $\left\{\left(\boldsymbol{s}_{k,j}^{\mathcal{T}}, \boldsymbol{p}_{k,j}^{\mathcal{T}}\right)\right\}_{j=n}^{n+B}$ from $\mathcal{D}_k^{\mathcal{T}}$.

2   **for** $j = n \rightarrow n + B$ **do**

3 $S_{\text{MT}}\left(\boldsymbol{s}_{k,j}^{\mathcal{T}}; \boldsymbol{\alpha}_k^{\mathcal{T}}\right) \rightarrow \boldsymbol{z}_{k,j}^{\mathcal{T}}$,

4 $S_{\text{MT}}^{-1}\left(\boldsymbol{z}_{k,j}^{\mathcal{T}}; \boldsymbol{\varphi}_k^{\mathcal{T}}\right) \rightarrow \hat{\boldsymbol{p}}_{k,j}^{\mathcal{T}}$,

5   Compute $\mathcal{L}_{\text{MT}}$ by (11) with $\boldsymbol{p}_{k,j}^{\mathcal{T}}$ and $\hat{\boldsymbol{p}}_{k,j}^{\mathcal{T}}$.

6   Train $\boldsymbol{\alpha}_k^{\mathcal{T}}, \boldsymbol{\varphi}_k^{\mathcal{T}} \rightarrow$ Gradient descent with $\mathcal{L}_{\text{MT}}$.

    Return: $S_{\text{MT}}\left(;\boldsymbol{\alpha}_k^{\mathcal{T}}\right)$ and $S_{\text{MT}}^{-1}\left(;\boldsymbol{\varphi}_k^{\mathcal{T}}\right)$.

7 Function `Train JSC Codec()`:

    Input: The semantic text features $\left\{\boldsymbol{z}_{k,j}^{\mathcal{T}}\right\}_{j=n}^{n+B}$.

8   **for** $j = n \rightarrow n + B$ **do**

9   **Transmitter**:

10 $C_{\text{MT}}\left(\boldsymbol{z}_{k,j}^{\mathcal{T}}; \boldsymbol{\beta}_k^{\mathcal{T}}\right) \rightarrow \boldsymbol{x}_{k,j}^{\mathcal{T}}$,

11   Transmit $\boldsymbol{x}_{k,j}^{\mathcal{T}}$ over the channel.

12   **Receiver**:

13   Receive $\mathbf{Y}$,

14   MIMO detection by (4) to get $\hat{\boldsymbol{x}}_{k,j}^{\mathcal{T}}$,

15 $C_{\text{MT}}^{-1}\left(\hat{\boldsymbol{x}}_{k,j}^{\mathcal{T}}; \boldsymbol{\gamma}_k^{\mathcal{T}}\right) \rightarrow \hat{\boldsymbol{z}}_{k,j}^{\mathcal{T}}$,

16   Compute $\mathcal{L}_{\text{MSE}}$ with (12).

17   Train $\boldsymbol{\beta}_k^{\mathcal{T}}, \boldsymbol{\gamma}_k^{\mathcal{T}} \rightarrow$ Gradient descent with $\mathcal{L}_{\text{MSE}}$.

    Return: $C_{\text{MT}}\left(;\boldsymbol{\beta}_k^{\mathcal{T}}\right)$ and $C_{\text{MT}}^{-1}\left(;\boldsymbol{\gamma}_k^{\mathcal{T}}\right)$.

18 Function `Train Whole Network()`:

    Input: Choose mini-batch data $\left\{\left(\boldsymbol{s}_{k,j}^{\mathcal{T}}, \boldsymbol{p}_{k,j}^{\mathcal{T}}\right)\right\}_{j=n}^{n+B}$ from $\mathcal{D}_k^{\mathcal{T}}$.

19   **for** $j = n \rightarrow n + B$ **do**

20   Repeat line 3-4, 11-16, and 4 to get $\hat{\boldsymbol{p}}_{k,j}^{\mathcal{T}}$,

21   Compute $\mathcal{L}_{\text{MT}}$ by (11) with $\boldsymbol{p}_{k,j}^{\mathcal{T}}$ and $\hat{\boldsymbol{p}}_{k,j}^{\mathcal{T}}$.

22   Train $\boldsymbol{\alpha}_k^{\mathcal{T}}, \boldsymbol{\beta}_k^{\mathcal{T}}, \boldsymbol{\gamma}_k^{\mathcal{T}}, \boldsymbol{\varphi}_k^{\mathcal{T}} \rightarrow$ Gradient descent with $\mathcal{L}_{\text{MT}}$.

    Return: $S_{\text{MT}}\left(;\boldsymbol{\alpha}_k^{\mathcal{T}}\right)$, $S_{\text{MT}}^{-1}\left(;\boldsymbol{\varphi}_k^{\mathcal{T}}\right)$, $C_{\text{MT}}\left(;\boldsymbol{\beta}_k^{\mathcal{T}}\right)$, and $C_{\text{MT}}^{-1}\left(;\boldsymbol{\gamma}_k^{\mathcal{T}}\right)$.

---

model to convert the meaning to the target sentence by learning the target language word distribution. The CE loss function is represented by

$$\mathcal{L}_{\text{MT}} = \mathbb{E}\left[-\sum_n P(\boldsymbol{p}_{k,j}^{\mathcal{T}}[n])\log\left(P(\hat{\boldsymbol{p}}_{k,j}^{\mathcal{T}}[n])\right)\right], \qquad (11)$$

where $P(\hat{\boldsymbol{p}}_{k,j}^{\mathcal{T}}[n])$ is the predicted probability that the n-th word appears in sentence $\hat{\boldsymbol{p}}_{k,j}^{\mathcal{T}}$, and $P(\boldsymbol{p}_{k,j}^{\mathcal{T}}[n])$ is the real probability that the n-th word appears in the sentence $\boldsymbol{p}_{k,j}^{\mathcal{T}}$. After convergence, the model learns the syntax, phrase, the meaning of words in the target language.

In the second training phase that is listed as `Train JSC Codec` of Algorithm 2, the JSC codec, $C_{\text{MT}}(;\boldsymbol{\beta}_k^{\mathcal{T}})$ and $C_{\text{MT}}^{-1}(;\boldsymbol{\gamma}_k^{\mathcal{T}})$, are also trained to learn the compress and decompress semantic text information, as well as deal with the channel distortion and multi-user interference with the MSE loss function given by

$$\mathcal{L}_{\text{MSE}} = \mathbb{E}\left[\left\|\hat{\boldsymbol{z}}_{k,j}^{\mathcal{T}} - \boldsymbol{z}_{k,j}^{\mathcal{T}}\right\|_2^2\right], \qquad (12)$$

where $\hat{\boldsymbol{z}}_{k,j}^{\mathcal{T}}$ is the recovered semantic text information at the receiver and $\boldsymbol{z}_{k,j}^{\mathcal{T}}$ is the transmitted semantic text information.

Different from the DeepSC-IR training algorithm, there exists a semantic decoder at the DeepSC-MT receiver. This means that semantic errors between $\hat{\boldsymbol{z}}_{k,j}^{\mathcal{T}}$ and $\boldsymbol{z}_{k,j}^{\mathcal{T}}$ can be mit-

igated by jointly training the whole system shown as `Train Whole Network` in Algorithm 2 with the loss function (11).

V. MULTIMODAL MULTI-USER SEMANTIC COMMUNICATIONS

In this section, the multimodal multi-user semantic communications are investigated for serving the VQA task, namely DeepSC-VQA, in which the transmitters adopt the same structures as that of DeepSC-IR for images and DeepSC-MT for texts. They also share the same JSC decoder design. Particularly, a novel semantic decoder is proposed to merge the image-text semantic information.

Fig. 3. The proposed network structure of multimodal multi-user semantic communication system with DeepSC-VQA transceiver.

A. Model Description

Assume that the k-th user for image transmission and the i-th user for text transmission, $\mathcal{D}_{k,i}^{\mathcal{I},\mathcal{T}} = \left\{ \left(s_{k,j}^{\mathcal{I}}, s_{i,j}^{\mathcal{T}}, l_{(k,i),j} \right) \right\}_{j=1}^{D}$ with size D is the training dataset, where $s_{k,j}^{\mathcal{I}}$ is the j-th image from the k-th user, $s_{i,j}^{\mathcal{T}}$ is the j-th text from the i-th user, and $l_{(k,i),j}$ is the answer label for $s_{k,j}^{\mathcal{I}}$ and $s_{i,j}^{\mathcal{T}}$. $S_{\text{VQA}}\left(; \boldsymbol{\alpha}_k^{\mathcal{I}} \right)$, $C_{\text{VQA}}\left(; \boldsymbol{\beta}_k^{\mathcal{I}} \right)$, $C_{\text{VQA}}^{-1}\left(; \boldsymbol{\gamma}_k^{\mathcal{I}} \right)$ are the image semantic encoder, image JSC encoder, and image JSC decoder of the i-th user, respectively. $S_{\text{VQA}}\left(; \boldsymbol{\alpha}_i^{\mathcal{T}} \right)$, $C_{\text{VQA}}\left(; \boldsymbol{\beta}_i^{\mathcal{T}} \right)$, $C_{\text{VQA}}^{-1}\left(; \boldsymbol{\gamma}_i^{\mathcal{T}} \right)$ are the text semantic encoder, text JSC encoder, and text JSC decoder of the k-th user, respectively. $S_{\text{VQA}}^{-1}\left(; \boldsymbol{\varphi}_{(k,i)} \right)$ represents joint semantic decoder of the i-th and the t-th user for the VQA task.

As shown in Fig. 3, the proposed DeepSC-VQA network consists of one image transmitter, one text transmitter, and one receiver for simplicity. For the DeepSC-VQA transmitters and receivers, we adopt the same structures as the image transmitter of DeepSC-IR and text transmitter of DeepSC-MT to unify the transmitter paradigm. At the receiver, the structures of the image JSC decoder and text JSC decoder are also the same as that of the image JSC decoder in DeepSC-IR and that of the text JSC decoder in DeepSC-MT. Besides, we develop a new semantic decoder network for image-text information fusion, which includes two modules: information query module and information fusion module.

1) Information Query: The layer-wise Transformer is adopted. Different from the classic Transformer, where the decoder layers exploit the output tokens of the last layer of encoder as the input, the layer-wise Transformer employs the output tokens of each encoder layer as the input of each decoder layer. Such a design can leak more text information than classic Transformer and guide the image information query in the decoder more efficiently, which does not introduce any costs.

2) Information Fusion: After the information query, the layer-wise Transformer has already captured keywords in the text information and the corresponding regions in image information, which has reflected in the output tokens. We will then need to fuse keywords and the corresponding image regions to get the answer. As mentioned in Section II, the <CLS> token represents the global descriptor. Therefore, the <CLS> tokens in the output tokens of the Transformer encoder and Transformer decoder represent the global text information and global image information, respectively. Using the text <CLS>

and image <CLS>, we design the information fusion module as shown in Fig. 3, where dropout layers are used here to avoid over-fitting.

B. Training Algorithm

Similar to the DeepSC-MT training algorithm, the DeepSC-VQA is trained jointly by three phases but with different loss functions.

The first phase is `Train Semantic Codec`, the semantic codec of DeepSC-VQA, $S_{\text{VQA}}\left(; \boldsymbol{\alpha}_k^{\mathcal{I}} \right)$, $S_{\text{VQA}}\left(; \boldsymbol{\alpha}_i^{\mathcal{T}} \right)$, $S_{\text{VQA}}^{-1}\left(; \boldsymbol{\varphi}_{(k,i)} \right)$, is trained jointly by the CE loss function,

$$\mathcal{L}_{\text{VQA}} = \mathbb{E}\left[-P\left(l_{(k,i),j} \right) \log \left(P\left(\hat{l}_{(k,i),j} \right) \right) \right], \quad (13)$$

where $P(l_{(k,i),j})$ and $P(\hat{l}_{(k,i),j})$ are the real and predicted probability of answer, respectively. By reducing the loss value of CE, the network learns to predict the answer with the highest probability of accuracy.

After training the semantic codec, JSC codecs are trained to compress by JSC encoder to reduce the number of transmitted symbols, and then decompress by the JSC decoder to recover semantic information accurately over multiple user physical channels. The image and text JSC codec are trained jointly by the function `Train JSC Codec`, in which the loss function is designed as

$$\mathcal{L}_{\text{MSE}}^{(\text{VQA})} = \mathbb{E}\left[\left\| \hat{z}_{k,j}^{\mathcal{I}} - z_{k,j}^{\mathcal{I}} \right\|_2^2 + \left\| \hat{z}_{i,j}^{\mathcal{T}} - z_{i,j}^{\mathcal{T}} \right\|_2^2 \right], \quad (14)$$

where $z_{k,j}^{\mathcal{I}}$ and $z_{i,j}^{\mathcal{T}}$ are the transmitted semantic image and text information, respectively. $\hat{z}_{k,j}^{\mathcal{I}}$ and $\hat{z}_{i,j}^{\mathcal{T}}$ are the recovered semantic image and text information at the receiver, respectively.

There exists error propagation from the JSC decoders to the semantic receiver because of the imperfect semantic information recovery in the low SNR regimes. Therefore, the whole DeepSC-VQA network is trained jointly with loss function (13) to reduce the error propagation, which is the function `Train Whole Network`.

VI. SIMULATION RESULTS

In this section, we compare the proposed multi-user semantic communication systems with traditional source coding and channel coding methods over various channels, in which both the perfect and imperfect CSI are considered.

A. Implementation Details

1) The Datasets: We choose four popular datasets commonly used for the image retrieval task. *Stanford Online Products* [42] consists of 120,053 online products images representing 22,634 categories, in which 11,318 categories are used for training and the remaining 11,316 categories are used for testing. *CUB-200-2011* [43] has 200 bird categories with 11,789 images. We split the first 100 classes for training and the rest of 100 classes for testing. *Cars196* [44] contains 16,185 images corresponding to 196 car categories with the first 98 categories to be used for training. The remaining 98 categories are used for testing. *In-Shop Clothes* [45] contains 72,172 cloth images belonging to 7,986 categories, in which 3997 categories are used for training and the other 3985 categories will be used for testing.

For the machine translation task, we adopt the *WMT 2018 Chinese-English news track*, which contains 202,221 pairs for training and 50,556 pairs for testing. The dataset is filtered into the length of English sentences with 5 to 75 words.

For the VQA task, we adopt the popular dataset: *CLEVR* [46], which consists of a training set of 70,000 images and 699,989 questions and a test set of 15,000 images and 149,991 questions.

2) Training Settings: The image semantic encoder of DeepSC-IR is based on the public implementation of DeiT-small model[b] with 6 Transformer encoder layers. The setting of the `Train Semantic Encoder` of DeepSC-IR is the Adam optimizer with learning rate 3×10^{-5}, weight decay 5×10^{-4}, batch size of 64, and epoch of 40. The setting of the `Train JSC Encoder` of DeepSC-IR is the Adam optimizer with learning rate 1×10^{-3}, batch size of 64, and epoch of 100. During the training phase, the data augmentation is used to resize the image to 256×256 and then take a random crop of size 224×224 combined with random horizontal flipping. In the test phase, the images are resized to 256×256 first and centrally cropped to 224×224.

The text semantic codec of DeepSC-MT is based on the public implementation of the Transformer model[c] with 6 Transformer encoder layers and decoder layers. The setting of the `Train Semantic Codec` of DeepSC-MT is the Adam optimizer with learning rate 1×10^{-5}, betas of 0.9 and 0.98, batch size of 64, and epoch of 10. The setting of the `Train JSC Codec` of DeepSC-MT is the Adam optimizer with learning rate 1×10^{-3}, batch size of 64, and epoch of 20. The setting of the `Train Whole Network` of DeepSC-MT is the same as that of `Train Semantic Codec` but with epoch of 20.

The image semantic encoder of DeepSC-VQA is also based on the pre-trained DeiT-small model but the other parts are trained from scratch, where the text semantic encoder is with 6 Transformer encoder layers and the semantic decoder is with 4 Transformer encoder layers and decoder layers. We freeze the image semantic encoder to speed up training. The output dimension for the vision Transformer and text Transformer are set differently, which requires the dimension increasing oper-

[b]https://github.com/facebookresearch/deit.
[c]https://huggingface.co/Helsinki-NLP.

ations after the image JSC decoder. The dimension-increasing operations successively include the dropout layer, dense layer from 384 to 512, ELU activation layer, dropout layer, and dense layer from 512 to 512, and ELU activation layer. The setting of the `Train Semantic Codec` of DeepSC-VQA is the Adam optimizer with learning rate 1×10^{-4}, betas of 0.9 and 0.98, batch size of 64, and epoch of 80. The setting of the `Train JSC Codec` of DeepSC-VQA is the Adam optimizer with learning rate 1×10^{-3}, batch size of 128, and epoch of 30. The setting of `Train Whole Network` of DeepSC-MT is the same as that of the `Train Semantic Codec` but with epoch of 10. The data augmentation is used to resize images to 224×224 with BICUBIC interpolation for both training and testing.

3) Benchmarks and Performance Metrics: Our benchmark will adopt several typical source and channel coding methods.

- Error-free Transmission: The full, noiseless images and texts are delivered to the receiver, which will serve as the upper bound.
- Traditional Methods: To perform the source and channel coding separately, we use the following technologies, respectively:
  - 8-bit unicode transformation format (UTF-8) encoding for text source coding, a commonly used method in text compression;
  - Joint photographic experts group (JEPG) for image source coding, a widely used method in image compression;
  - Turbo coding for text channel coding, popular channel coding for a small size file;
  - Low-density parity-check code (LDPC) for image channel coding, and classic channel coding for big size files.

In the simulation, all coding rates of channel codings are 1/3. Perfect and imperfect CSI are set with $\sigma_e^2 = 0$ and $\sigma_e^2 = 0.025$, respectively. We set $r = 2$ for Rician channels and $\mathbf{H} = \mathbf{I}$ for AWGN channels.

The Recall@1 evaluation metric is adopted as performance metric for the image retrieval task. Bi-lingual evaluation understudy (BLEU) score is adopted for the machine translation task. Answer accuracy is used for VQA task.

B. Single-Modal Multi-User Semantic Communication

The Recall@1 performance comparison for different channels on CUB-200-2011 and for different datasets over Rician channels are shown in Fig. 4 and Fig. 5, respectively. From Fig. 4, for different channels on CUB-200-2011, the proposed DeepSC-IR provides a significant gain at the low SNR regimes and approaches to the upper bound at the high SNR regimes among the reported methods, outperforming the JPEG-LDPC with 8-QAM by a margin of more than 24dB gain for 0.4 Recall@1 over fading channels. Even when using imperfect CSI, the DeepSC-IR still outperforms the benchmarks with slight performance degradation at Recall@1. From Fig. 5, for different datasets over Rician channels, the DeepSC-IR also outperforms the JPEG-LDPC with 8-QAM in the three popular datasets at Recall@1 with more than 24 dB gain, respectively.

Fig. 4. Recall@1 comparison between DeepSC-IR and JPEG-LDPC with 8-QAM over different channels, in which the dataset is CUB-200-2011.

(a) Stanford Online Products.　　(b) Cars196.　　(c) In-shop Clothes.

Fig. 5. Recall@1 comparison between DeepSC-IR and JPEG-LDPC with 8-QAM for different datasets under Rician channels.

(a) AWGN Channels.　　(b) Rayleigh Channels.　　(c) Rician Channels.

Fig. 6. BLEU score comparison between DeepSC-MT and UTF-8-Turbo with QPSK for English-to-Chinese under AWGN channels, Rayleigh channels, and Rician channels.

Besides, exploiting imperfect CSI considerably decreases the performance at Recall@1 for the traditional method, especially in In-Shop Clothes but is only with a slightly performance degradation for DeepSC-IR.

The BLEU score performance comparison for different channels on English-to-Chinese is reported in Fig. 6 and on Chinese-to-English is shown in Fig. 7. From Fig. 6, on English-to-Chinese over different channels, the DeepSC-MT outperforms the UTF-8-Turbo with QPSK at the low SNR regimes over AWGN, as well as at all SNR regimes over fading channels. More inaccurate CSI decreases BLEU score for both systems, in which the DeepSC-MT outperforms the benchmark and retains its high robustness to imperfect CSI.

On Chinese-to-English over fading channels in Fig. 7, the DeepSC-MT performs well except at the high SNR regimes. Although the UTF-8-Turbo in BSPK has a higher BLEU score than DeepSC-MT as SNR increases, it performs worse than DeepSC-MT at all SNR regimes w.r.t. imperfect CSI.

C. Multimodal Multi-User Semantic Communication

The answer accuracy performance comparison for VQA task over different channels is presented in Fig. 8, in which the benchmark consists of UTF-8-Turbo with BPSK for text and JPEG-LDPC with 8-QAM for image. The DeepSC-VQA outperforms the benchmark at the low SNR regimes over the AWGN channels and at all SNR regimes over fading

(a) AWGN Channels. (b) Rayleigh Channels. (c) Rician Channels.

Fig. 7. BLEU score comparison between DeepSC-MT and UTF-8-Turbo with BPSK for Chinese-to-English under AWGN channels, Rayleigh channels, and Rician channels.

(a) AWGN Channels. (b) Rayleigh Channels. (c) Rician Channels.

Fig. 8. Answer accuracy comparison between DeepSC-VQA and traditional methods, including UTF-8-Turbo with BPSK for text and JPEG-LDPC with 8-QAM for image, in which different channels are considered.

(a) Image Retrieval. (b) Machine Translation. (c) VQA.

Fig. 9. Recall@1, BLEU score, and answer accuracy comparisons versus the number of users over Rician channel with SNR=18dB.

channels. In particular, the DeepSC-VQA achieves the upper bound at approximate SNR=9dB over fading channels. The answer accuracy of benchmark considerably decreases from the AWGN to fading channels for benchmarks but experiences only little performance degradation at the low SNR regimes and no performance loss at the high SNR regimes for DeepSC-VQA. Similarly, for imperfect CSI, the robustness of DeepSC-VQA is also better than that of benchmark with more than 24dB gain at 0.7 answer accuracy. This also verifies the effectiveness of the design of semantic decoder of DeepSC-VQA.

D. Different Number of Users

In Fig. 9, different tasks versus the different number of users are compared. All proposed methods perform steadily as the number of users increases but the benchmarks experience performance improvement or degradation. The difference in performance trends between benchmarks are because of the gains from channel coding and low-order modulation methods. Both for image retrieval task and VQA task, the DeepSC-IR and DeepSC-VQA outperform their benchmarks at Recall@1 and at answer accuracy, respectively, in which the performance at Recall@1 and answer accuracy of benchmarks decrease first and achieve floor as the number of users increases. For the

TABLE I
THE NUMBER OF TRANSMITTED SYMBOLS COMPARISON BETWEEN MULTI-USER SEMANTIC COMMUNICATION SYSTEMS AND TRADITIONAL
SOURCE-CHANNEL COMMUNICATION SYSTEMS.

Task	Dataset	Methods	Average Number of Transmitted Symbols for One Image or One Word	Ratio
Image Retrieval	Cars196 CUB-200-2011 In-Shop Clothes Stanford Online Products	DeepSC-IR / JPEG-LDPC with 8-QAM	$128/499,920$ $128/247,312$ $128/60,696$ $128/174,808$	0.02% 0.05% 0.21% 0.07%
Machine Translation	English-to-Chinese Chinese-to-English	DeepSC-MT / UTF-8-Turbo with QPSK DeepSC-MT / UTF-8-Turbo with BPSK	$77/76$ $77/68$	101.31% 113.23%
VQA	CLEVR: Text CLEVR: Image	DeepSC-VQA / UTF-8-Turbo with BPSK DeepSC-VQA / JPEG-LDPC with 8-QAM	$77/152$ $25,216/55,624$	50.66% 45.33%

machine translation task, the BLEU score of the benchmark increases with the number of users, making the benchmark outperform DeepSC-MT with respect to perfect CSI. Besides, for imperfect CSI, all proposed semantic communication systems outperform the corresponding benchmarks with relatively little performance degradation.

E. Number of Transmitted Symbols

The numbers of transmission symbols for different methods are compared in Table I. For image transmission, the proposed multi-user semantic communication systems significantly decrease the number of transmission symbols, especially for the image retrieval task with the DeepSC-IR only transmitting 0.02% symbols of the benchmarks for one image. For text transmission, although the proposed methods transmit a similar or slightly more number of symbols compared with the benchmark in machine translation task, they achieve approximately 50% saving in the numbers of symbols when the benchmark employs a lower order modulation in the VQA task. This suggests that the proposed multi-user semantic communications can decrease the transmission delay with a lower number of transmission symbols and hence are suitable for lower latency scenarios.

F. Computational Complexity

The computational complexity for different methods[d] is compared in Table II. For image transmission, all of the proposed methods have a lower computational complexity than traditional methods, in which the complexity of DeepSC-IR can decrease by more than one order of magnitude. For text transmission, the proposed DeepSC-MT shows a similar computational complexity in English transmission but has a slightly higher computational complexity in the Chinese transmission compared to the benchmarks. Such a slightly higher computational complexity can provide robustness to noise in low SNRs. This suggests that the proposed multi-user semantic communication systems achieve lower power consumption when transmitting a large size of data.

---

[d]We only analyze the complexity of channel coding for both methods because the other parts are shared in both methods and the complexity of source coding is low and can be omitted.

G. Visualization Results

The visualized results for the considered tasks including image retrieval, machine translation, and VQA are shown in Fig. 10, Table III, and Fig. 11, respectively. Fig. 10 shows *top-4* similar image retrieval results for DeepSC-IR and JPEG-LDPC at 18 dB over Rician channels. The proposed DeepSC-IR returns similar images successfully with the query image but the traditional method fails due to the destroyed received image. Table III provides the received translation results on Chinese-to-English. The proposed DeepSC-MT demonstrates reasonable translations results in both scenarios with perfect and imperfect CSI, but the traditional method fails to convey the sentence when CSI cannot be estimated exactly.

Fig. 11 shows the results of the VQA task and the attention visualizations for the layer-wise Transformer. The proposed DeepSC-VQA correctly answers the question. In the attention visualizations, the proposed DeepSC-VQA can effectively query the key regions in the image layer by layer with the received semantic image and text information. Specifically, the words "*shape, red tiny*" in the sentence has a higher magnitude in the first layer, finding the key red tiny object in the image. The second layer highlights the words "*are there other*" to find other objects in the image and neglect the red tiny object. The third and fourth layers double-check the other objects with the red tiny object to give the final answer.

VII. CONCLUSIONS

In this paper, we have explored task-oriented multi-user semantic communications to transmit data with single-modality and multiple modalities, respectively. We considered two single-modal tasks, image retrieval and machine translation, as well as one multimodal task, visual question answering (VQA). In this context, we have proposed three Transformer based transceivers, DeepSC-IR, DeepSC-MT, and DeepSC-VQA, which share the same transmitter structures but with different receiver structures. Each transceiver is trained jointly by the proposed training algorithm. In addition, all of the proposed multi-user semantic communication systems were found to outperform the traditional ones in the low SNR regimes and provide graceful performance degradation with imperfect CSI. For both image retrieval and VQA tasks, the proposed DeepSC-IR and DeepSC-VQA can provide more than 18 dB gain and reduce by more than 50% the number of

TABLE II

COMPUTATIONAL COMPLEXITY COMPARISON BETWEEN MULTI-USER SEMANTIC COMMUNICATION SYSTEMS AND TRADITIONAL SOURCE-CHANNEL
COMMUNICATION SYSTEMS.

Task	Dataset	Methods	Computational Complexity	
			Additions	Multiplications
Image Retrieval	Cars196 CUB-200-2011 In-Shop Clothes Stanford Online Products	DeepSC-IR / JPEG-LDPC with 8-QAM	$8.2 \times 10^5/9.0 \times 10^9$ $8.2 \times 10^5/4.4 \times 10^9$ $8.2 \times 10^5/1.0 \times 10^9$ $8.2 \times 10^5/3.1 \times 10^9$	$8.2 \times 10^5/1.7 \times 10^{10}$ $8.2 \times 10^5/8.4 \times 10^9$ $8.2 \times 10^5/2.1 \times 10^9$ $8.2 \times 10^5/6.0 \times 10^9$
Machine Translation	English-to-Chinese Chinese-to-English	DeepSC-MT / UTF-8-Turbo with QPSK DeepSC-MT / UTF-8-Turbo with BPSK	$5.9 \times 10^5/1.0 \times 10^5$ $5.9 \times 10^5/4.5 \times 10^4$	$5.9 \times 10^5/1.6 \times 10^5$ $5.9 \times 10^5/7.3 \times 10^4$
VQA	CLEVR: Text CLEVR: Image	DeepSC-VQA / UTF-8-Turbo with BPSK DeepSC-VQA / JPEG-LDPC with 8-QAM	$5.9 \times 10^5/1.0 \times 10^5$ $1.6 \times 10^8/1.0 \times 10^9$	$5.9 \times 10^5/1.6 \times 10^5$ $1.6 \times 10^8/1.9 \times 10^9$

Fig. 10. Part of the results for image retrieval at 18 dB over Rician Channels.

TABLE III

PART OF THE RESULTS FOR MACHINE TRANSLATION ON CHINESE TO ENGLISH AT 18 dB OVER RICIAN CHANNELS.

Transmission Sentence	与此同时,过去二十年来卡塔尔想尽办法扩张其影响力并已经收到成效,该国的吸引力已经蔚为可观。
Reference Translated Sentence	Meanwhile, Qatar's diligent efforts to expand its influence over the last two decades have paid off, with the country developing considerable power of attraction.
DeepSC-MT perfect CSI	Meanwhile, over the past two decades, Qatar has been able to expand its influence by doing everything it can, and has been successful, and the country's appeal is already strong.
DeepSC-MT imperfect CSI	Meanwhile, over the past two decades, Qatar has been able to reap the benefits of its efforts to expand its reach, and the country's appeal is already strong.
UTF-8-Turbo with BPSK perfect CSI	At the same time, over the past two decades Qatar had tried to expand its influence and had already borne fruit, and the country's appeal had become significant.
UTF-8-Turbo with BPSK imperfect CSI	Could not decode

transmission symbols and computational complexity compared to traditional communications. In particular, compared with traditional methods, DeepSC-IR only needs 1‰ transmission symbols on average and decreases the complexity by more than one order of magnitude. As we result, we can conclude that multi-user semantic communication systems are an attractive alternative to traditional communication systems for particular tasks.

REFERENCES

[1] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y. A. Zhang, "The roadmap to 6g: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, Aug. 2019.

[2] P. Wang, Y. Li, L. Song, and B. Vucetic, "Multi-gigabit millimeter wave wireless communications for 5G: From fixed access to cellular networks," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 168–178, Jan. 2015.

[3] C. Han and Y. Chen, "Propagation modeling for wireless communications in the Terahertz band," *IEEE Commun. Mag.*, vol. 56, no. 6, pp. 96–101, Jun. 2018.

[4] L. Lu, G. Y. Li, A. L. Swindlehurst, A. E. Ashikhmin, and R. Zhang, "An overview of massive MIMO: benefits and challenges," *IEEE J. Select. Top. Signal Processing*, vol. 8, no. 5, pp. 742–758, Apr. 2014.

[5] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4157–4170, Jun. 2019.

[6] K. Sayood, *Introduction to Data Compression*. Morgan Kaufmann, 2017.

[7] A. Goldsmith, *Wireless Communications*. Cambridge university press,

Fig. 11. Part of results for the VQA task and visualized attention for layer-wise Transformer at 18 dB over Rician Channels.

2005.

[8] M. Kountouris and N. Pappas, "Semantics-empowered communication for networked intelligent systems," *IEEE Commun. Mag.*, vol. 59, no. 6, pp. 96–102, Jul. 2021.

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

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

[11] H. Xie and Z. Qin, "A lite distributed semantic communication system for internet of things," *IEEE J. Select. Areas Commun.*, vol. 39, no. 1, pp. 142–153, Jan. 2021.

[12] 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, Sept. 2019.

[13] D. B. Kurka and D. Gündüz, "DeepJSCC-f: Deep joint source-channel coding of images with feedback," *IEEE J. Select. Areas Inf. Theory*, vol. 1, no. 1, pp. 178–193, May 2020.

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

[15] C. Lee, J. Lin, P. Chen, and Y. Chang, "Deep learning-constructed joint transmission-recognition for internet of things," *IEEE Access*, vol. 7, pp. 76 547–76 561, Jun. 22019.

[16] M. Jankowski, D. Gündüz, and K. Mikolajczyk, "Wireless image retrieval at the edge," *IEEE J. Select. Areas Commun.*, vol. 39, no. 1, pp. 89–100, Jan. 2021.

[17] Z. Weng, Z. Qin, and G. Y. Li, "Semantic communications for speech recognition," *arXiv preprint arXiv:2107.11190*, 2021.

[18] H. Xie, Z. Qin, and G. Y. Li, "Task-oriented semantic communications for multimodal data," *arXiv preprint arXiv:2108.07357*, 2021.

[19] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Processing Mag.*, vol. 34, no. 6, pp. 96–108, Nov. 2017.

[20] A. Vakil, J. Liu, P. Zulch, E. Blasch, R. Ewing, and J. Li, "A survey of multimodal sensor fusion for passive rf and eo information integration," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 36, no. 7, pp. 44–61, Jul. 2021.

[21] C. Ruiz, J. Falcao, and P. Zhang, "Autotag: Visual domain adaptation for autonomous retail stores through multi-modal sensing," in *ACM adjunct Proc. Int. Joint Conf. Pervasive Ubiquitous Comput. and Proc. Int. Symp. Wearable Comput.*, London, UK, Sept. 2019, pp. 518–523.

[22] B. L. R. Stojkoska and K. V. Trivodaliev, "A review of internet of things for smart home: Challenges and solutions," *J. Clean. Prod.*, vol. 140, pp. 1454–1464, Jan. 2017.

[23] H. Zou, J. Yang, H. P. Das, H. Liu, Y. Zhou, and C. J. Spanos, "Wifi and vision multimodal learning for accurate and robust device-free human activity recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. Workshops (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 426–433.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Advances Neural Inf. Processing Systems (NIPS)*, Long Beach, CA, USA, Dec. 2017, pp. 5998–6008.

[25] W. Chen, Y. Liu, W. Wang, E. Bakker, T. Georgiou, P. Fieguth, L. Liu, and M. S. Lew, "Deep image retrieval: A survey," *arXiv preprint arXiv:2101.11282*, 2021.

[26] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of cnn activations," in *Proc. Int'l. Conf. Learn. Representations (ICLR)*, San Juan, Puerto Rico, May 2016, pp. 1–12.

[27] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "Deep image retrieval: Learning global representations for image search," in *Proc. Eur. Conf. Comput. Vision (ECCV)*, Amsterdam, The Netherlands, Oct. 2016, pp. 241–257.

[28] A. Babenko and V. Lempitsky, "Aggregating local deep features for image retrieval," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1269–1277.

[29] C. Huang, S. Yang, Y. Pan, and H. Lai, "Object-location-aware hashing for multi-label image retrieval via automatic mask learning," *IEEE Trans. Image Processing*, vol. 27, no. 9, pp. 4490–4502, May 2018.

[30] E. W. Teh, T. DeVries, and G. W. Taylor, "ProxyNCA++: Revisiting and revitalizing proxy neighborhood component analysis," in *Proc. Eur. Conf. Comput. Vision (ECCV)*, Glasgow, UK, Aug. 2020, pp. 448–464.

[31] A. El-Nouby, N. Neverova, I. Laptev, and H. Jégou, "Training vision transformers for image retrieval," *arXiv preprint arXiv:2102.05644*, 2021.

[32] Y. Gu, S. Wang, H. Zhang, Y. Yao, W. Yang, and L. Liu, "Clustering-driven unsupervised deep hashing for image retrieval," *Neurocomputing*, vol. 368, pp. 114–123, Nov. 2019.

[33] N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models," in *Proc. Conf. Empir. Methods Natural Language Processing (EMNLP)*, Seattle, Washington, USA, Oct. 2013, pp. 1700–1709.

[34] F. Meng, Z. Lu, M. Wang, H. Li, W. Jiang, and Q. Liu, "Encoding source language with convolutional neural network for machine translation," in *Proc. of Annual Meeting of the Assoc. for Computational Linguistics and Int'l Joint Conf. Natural Language Processing of the Asian Fed. of Natural Language Processing*, Beijing, China, Jul. 2015, pp. 20–30.

[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[36] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using

RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empir. Methods Natural Language Processing (EMNLP)*, Doha, Qatar, Oct. 2014, pp. 1724–1734.

[37] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, "Simple baseline for visual question answering," *arXiv preprint arXiv:1512.02167*, 2015.

[38] J. Kim, K. W. On, W. Lim, J. Kim, J. Ha, and B. Zhang, "Hadamard product for low-rank bilinear pooling," in *Int'l Conf. Learn. Representations (ICLR)*, Toulon, France, Apr. 2017.

[39] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 21–29.

[40] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 6077–6086.

[41] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 6281–6290.

[42] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 4004–4012.

[43] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011.

[44] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *Proc. IEEE Int'l Conf. Comput. Vision Workshops (ICCV)*, Sydney, Australia, Dec. 2013, pp. 554–561.

[45] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1096–1104.

[46] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. B. Girshick, "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1988–1997.