Post-Deployment Fine-Tunable Semantic Communication

Peiyuan Si, *Graduate Student Member, IEEE*, Renyang Liu, *Graduate Student Member, IEEE*, Liangxin Qian, *Graduate Student Member, IEEE*, Jun Zhao, *Member, IEEE*, Kwok-Yan Lam, *Senior Member, IEEE*

***Abstract*—Semantic communication (SemCom) is an emerging way that aims to improve communication efficiency based on the semantics of content, which relies on the knowledge base (KB) and is usually dedicated to specific tasks or datasets. To improve the adaptability of SemCom systems on unknown datasets, we propose a post-deployment Fine-Tunable Semantic Communication (FTSC) system for image transmission. Towards an adaptive and efficient SemCom system, our research consists of the framework design of FTSC and its system optimization study. Firstly, the generalizability study is conducted based on a two-layer hierarchical vector quantized-variational autoencoder (VQ-VAE-2). Unlike traditional SemCom that can work on limited pretrained datasets, FTSC adapts to varied input data post-deployment, enhancing practicality in diverse communication scenarios. This system incorporates two novel fine-tuning methods: Decoder Fine-Tuning (DFT) and Latent Space-based Decoder Fine-Tuning (LSDFT). DFT updates the decoder for new images post-deployment without transmitting gradients, while LSDFT eliminates the need for raw image transmission during fine-tuning. Secondly, we study the system optimization of the proposed FTSC framework to improve the efficiency of communication resource allocation with the concern of recovery quality, time delay, and energy cost in downlink transmissions. Extensive experiments demonstrate the superiority of FTSC over Joint Photographic Experts Group (JPEG) and Joint Source-Channel Coding (JSCC) across various datasets and noise levels, and both DFT and LSDFT significantly enhance image recovery on unfamiliar datasets compared to pre-trained models.**

***Index Terms*—Semantic communication, VQ-VAE-2, wireless communication, fine-tuning, resource allocation.**

I. Introduction

With the increasing application of Internet of Things (IoT) devices and smart services, a blueprint for sixth-generation (6G) wireless networks has been proposed [2]. Such a blueprint aims to build a space-air-ground-sea integrated network offering higher data rates and wider coverage. However, the ambition of 6G faces challenges due to limited communication resources, exacerbated by the burgeoning number of user devices and the growing data demands of future mobile services, such as mobile virtual reality (VR) and augmented reality (AR) [3]–[7]. Source coding has been an efficient way to address the scarcity of communication resources, which compresses the data size during transmission and recovers it at the receiver side. Recently, the applications of Artificial Intelligence (AI) technologies in source coding have led to the proposal of a new type of source coding method known as semantic communication (SemCom) [8]–[10], which extracts semantic information from raw data and achieves a higher compression rate than traditional algorithms.

The basic framework of SemCom comprises an encoder, a decoder, and a common knowledge base deployed at both encoder and decoder [11]–[14]. This framework serves as the foundation for SemCom systems across different data types. In the context of text data, Xie *et al*. [15] conducted early but significant explorations by designing a Transformer-based model DeepSC to reduce the number of symbols required for text transmission. Building upon DeepSC, a lite distributed version, L-DeepSC, was proposed to accommodate IoT devices with less computational capacity [16]. For speech data, Weng *et al*. [17] developed DeepSC-S based on a squeeze-and-excitation network (SENet), which is tailored to extract semantic information from speech signals. Additionally, Han *et al*. [18] proposed an end-to-end deep learning-based transceiver to minimize semantic redundancy in speech signals. Regarding image data, the joint source-channel coding (JSCC) technique proposed by Bourtsoulatze *et al*. [19] integrated autoencoder-based source coding with traditional channel coding and achieved robust image compression and recovery under noisy channels. Moreover, Hu *et al*. [20] introduced semantic noise into the raw data and proposed a masked vector quantized-variational autoencoder (VQ-VAE), which improved the robustness of the model against noise by masking part of the input image and further reducing the size of transmitted data.

Existing SemCom systems have demonstrated efficacy in reducing data size for transmission [21]–[26]. However, their generalization capability remains a significant challenge. In the case of image-based SemCom, most existing works rely on training results derived from limited training datasets. This approach poses a problem for real-world communication systems, where the input often contains unknown data. As a result, the performance of pretrained models may diminish due to distribution shifts. Once the encoder and decoder are deployed at the sender and receiver, respectively, updating the model involves transmitting parameters. This process is not only costly but also counter-intuitive to the fundamental goal of SemCom. Furthermore, most autoencoder-based SemCom models use the original image as the training label. However,

The authors are all with the College of Computing and Data Science, Nanyang Technological University, Singapore. Email: peiyuan001@e.ntu.edu.sg, n2208056e@e.ntu.edu.sg, qian0080@e.ntu.edu.sg, junzhao@ntu.edu.sg, kwokyan.lam@ntu.edu.sg. Corresponding author: Jun Zhao

This research is supported by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Future Communications Research & Development Programme.

A 7-page short paper appears at the 2024 IEEE 10th International Conference on Big Data Computing and Communications (BigCom) [1]. This is allowed by the policy at https://www.comsoc.org/publications/journals/ieee-transactions-wireless-communications/conference-vs-journal

in a SemCom system, only the latent space is transmitted, rendering the original training label inaccessible to post-deployment. This limitation further complicates the real-time updating of the SemCom model.

To address the existing research gap in the generalization capabilities of Semantic Communication (SemCom) systems with efficient utilization of communication and computation resources, this paper consists of the following two parts of researches which together form a comprehensive investigation: Part 1 about framework design, and Part 2 for rigorous system optimization. In Part 1, we design a SemCom framework which achieves high generalizability in the sense that it can adapt to unknown datasets through fine-tuning after the deployment of encoder and decoder. In Part 2, we conduct performance optimization for the SemCom system proposed in Part 1 by considering not just generalizability but also energy consumption and delay. We elaborate Part 1 and Part 2 below.

In Part 1, we introduce the Fine-Tunable Semantic Communication (FTSC) framework based on VQ-VAE-2 for image data [27], [28]. FTSC facilitates two distinct post-deployment fine-tuning methodologies: Decoder Fine-Tuning (DFT) and Latent Space-based Decoder Fine-Tuning (LSDFT). **DFT** implements the traditional transfer learning in the context of SemCom, where the encoder and decoder are deployed distributively. Throughout the DFT process, we freeze parameters in the encoder and transmit a small proportion of original images as labels to update the decoder. DFT requires additional data transmission of original data but avoids the transmission overhead for gradients. **LSDFT** updates the decoder using the received latent space via indirect inference methods. During the training phase, the encoder and the codebook are frozen, and a copy of the encoder is deployed at the receiver additionally. The received latent space is processed through a decoder and the frozen encoder to get its recovered version, and the loss, which calculates the distance between these two latent variables, is used to update the decoder on the receiver side. In the ablation study of FTSC, extensive experiments are conducted to reveal the relationship between compression rate and recovery quality under different noise types and levels.

In Part 2, we formulate a mathematical optimization problem and sets the objective function as the system performance, which incorporates generalizability studied in Part 1 as well as the energy consumption and delay of the system. The rationale is that in some practical SemCom systems, although generalizability matters, the energy cost and latency may be also important; for instance, delay-critical applications (e.g., those for extended reality [29]–[31]) require small latency, and devices with limited battery may have stringent energy requirements. In our optimization problem, we tune the decision variables in order to improve the system performance as much as possible under practical constraints. The designated decision variables comprise compression rate, bandwidth, and power allocation. The adopted constraints include sum bandwidth, sum transmission power, available range of compression rate and delay limitation. To deal with the non-convex optimization problem, we utilize fractional programming and alternative optimization to remove the coupling among variables and transfer the original problem into a convex form. Simulation results validate the superiority of the FTSC framework over traditional JPEG and JSCC [19] across a variety of channel conditions and datasets.

The contributions of this paper are as follows:

- As far as we know, we are the first to study the post-deployment model updating problem in SemCom with integrate study from framework design to rigorous system optimization. To achieve this, we propose the Fine-Tunable Semantic Communication (FTSC) system in Part 1, which enables the update of the deployed SemCom model and improve the generalization capability of SemCom systems. To improve the efficiency of FTSC in practical use cases, we study its performance optimization with resource limitation in Part 2.
- To improve the generalizability of SemCom systems, we propose two post-deployment fine-tuning methods under the FTSC framework in Part 1, i.e., DFT and LSDFT, with different training cost and recovery quality. With available raw images through additional transmission, DFT is adopted to update the decoder and embedding space by the difference between the recovered image and the original image. With stricter communication resource limitations and unavailable raw images, we adopt LSDFT to update the decoder based on the received latent space and update the model via indirect inference methods.
- To improve the efficiency of FTSC in practical communication, we study the balance of computational costs and communication resources in Part 2 to maximize the utility in a typical downlink transmission case. To solve the non-convex original problem, fractional programming and alternative optimization are utilized to remove the coupling among variables and transfer the original problem into a convex optimization problem.
- Extensive empirical results on DFT and LSDFT over various datasets and channel conditions verify the superiority of the proposed methods compared to traditional algorithms, e.g., JPEG and SemCom benchmark JSCC, in terms of peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and classification accuracy. Furthermore, simulation results of the downlink transmission optimization problem reveal the impact of resource limitation and the important factors of recovery quality, time delay, and energy cost in searching for the optimal model configuration and resource allocation.

The rest of this paper is organized as follows. We review the related works in Sec. II. Sec. III provides an overview of the system model and preliminary experiment results on VQ-VAE-2 to investigate the model configuration for semantic communication. Towards a generalizable and efficient SemCom system, our work is presented in two parts. Part 1 of our research is elaborated in Sec. IV, which introduces the proposed FTSC framework, and proposes DFT and LSDFT to improve the adaptability over various datasets. Then, in Sec. V dedicated to Part 2 of our study, a joint resource allocation and model configuration optimization problem in a downlink transmission use case is formulated and solved by fractional programming-based methods. The simulation results of the

optimization algorithm are presented in Sec. V-B. Finally, Sec. VI concludes this work and discusses potential research directions.

II. Related Work

In this section, we provide a brief literature review of semantic communication, VQ-VAE-2, and transfer learning and introduce their development, features, and applications.

A. Semantic Communication

The field of semantic communication (SemCom) has its roots in semiotic studies [32], and the initial developments in SemCom trace back to the 1950s [33]. Empowered by the development of AI technologies, SemCom has been propelled to an emerging research topic in recent years. Despite being in its nascent stage, SemCom research has diverse exploration directions, e.g., semantic-oriented communication, goal-oriented communication, and semantic-aware communication. Semantic-oriented communication aims to capture the core information in source data, and reduce the data size for transmission by removing the irrelevant information [8], [34]–[36]. Different from semantic-oriented communication, goal-oriented communication focuses more on the result, which is acknowledged by the communication participants and influences semantic information extraction. Xie *et al*. [37] designed a goal-oriented semantic transceiver MU-DeepSC for visual question answering (VQA) tasks, which outperforms the benchmarks in multi-user SemCom. Semantic-aware communication refers to the SemCom that plays a role in the analysis of agent behavior and the environment, facilitating better collaboration and task achievement. Its essential difference from the other two types of SemCom is that the other two focus on the result of communication, while semantic-aware communication emphasizes understanding and utilizing the contextual and semantic meanings to enhance interactions and decision-making. An example work for the above is [38], which studied the SemCom problem in autonomous vehicle networks where SemCom provides auxiliary information that assists in the task cooperation. Zhang *et al*. [39] introduced U-DeepSC, a unified semantic communication system, featuring a uniquely designed multi-exit architecture designed to accommodate the diverse layer requirements of various tasks. In addition to the progress on SemCom models, the communication resource and user allocation problem also attracted research interests. *Xia et al*. [40] developed a novel two-stage solution for optimizing user association and bandwidth allocation in intelligent SemCom systems to enhance throughput and efficiency. Zhang *et al*. [41] introduced a deep reinforcement learning framework for optimizing image transmission in SemCom networks, which significantly enhanced the efficiency and reduced transmission latency. The neural-network-based semantic encoder/decoder in SemCom leads to more computational cost. Towards the reduction of computational energy cost, Yang *et al*. [42] focus on energy-efficient SemCom based on rate splitting and designed an alternating algorithm where the closed-form solutions for semantic information extraction ratio and computation frequency are obtained at each step.

B. VQ-VAE-2 and other VAE-based models

VQ-VAE-2 [27] is a variant of vector quantized-variational autoencoder (VQ-VAE) [28] that introduces a two-tier hierarchical structure and significantly elevates its capability to generate high-fidelity images. By capturing intricate details across multiple scales, VQ-VAE-2 surpasses the BigGAN-deep [43] in terms of recovered images' classification accuracy. VQ-VAE has a wide application in various research fields, including image inpainting, video generation, and speech coding [44]–[47].

Besides the VQ-VAE, many other variants of the variational autoencoder (VAE) have been developed and proposed. Conditional VAE (CVAE) proposed by Sohn *et al*. [48] conditions the generation process on additional information, e.g., labels or other relevant data, and achieved intersection over union (IoU) of 98.52% in object segmentation task on CUB database with the noise level of 25%. β-VAE [49] introduces an adjustable hyperparameter β to the original VAE loss function to balance latent space disentanglement and reconstruction accuracy and achieved a disentanglement metric score of 99.23% on the 2D shapes dataset. Nouveau VAE (NVAE) proposed by Vahdat *et al.* [50] adopted deep hierarchical VAE for image generation using depth-wise separable convolutions and batch normalization, and can produce high-quality images as large as 256×256 pixels. Among all the mentioned variants of VAE, VQ-VAE is the most suitable model for SemCom because it compresses the original data into a smaller latent space and outperforms other models in recovery quality.

C. Transfer Learning

Transfer learning is a machine learning methodology that aims to enhance the performance of models in target domains by leveraging knowledge from related but different source domains, and often leads to significant improvements in learning efficiency and prediction accuracy [51]. Transfer learning is typically classified into three categories based on the availability of training labels and the relationship between source and target tasks: Inductive, Unsupervised, and Transductive transfer learning [52], [53]. In the context of computer vision, a common approach to implement transfer learning involves freezing part of or the whole feature extractor, while fine-tuning the fully connected layers [54]. This technique leverages the common features learned by pretrained feature extractors, either assisting with or being directly applied to the target dataset. Guo *et al*. [55] proposed AdaFilter for deep learning with both pretrained and fine-tuned filters, and improved average classification accuracy on multiple datasets by 2.54% compared with traditional methods. Transfer learning is also extensively applied in the context of VAE-based models [56]–[59], which implies the feasibility of the proposed FTSC framework in this paper.

III. System Model

VQ-VAE-2 is an enhanced variant of VQ-VAE, which has been proven to perform better in image recovery [27]. To the best of our knowledge, this is the first work to implement VQ-VAE-2 in SemCom. As an exploration of this topic, we

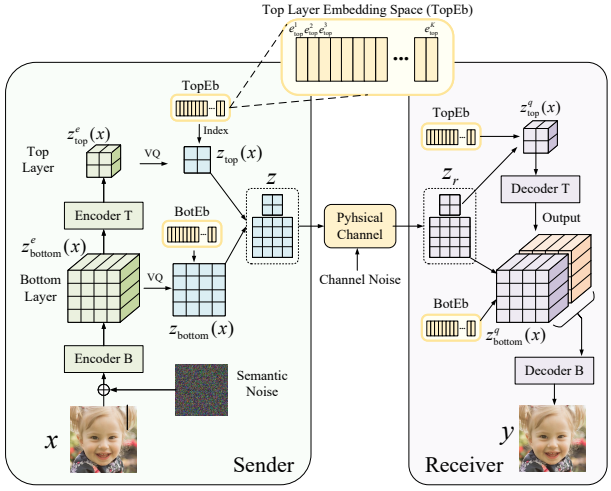


Fig. 1: Framework of FTSC.

first propose the SemCom framework based on VQ-VAE-2 and conduct preliminary experiments for feasibility investigation, denoising testing, and ablation study.

A. VQ-VAE-2-based SemCom System

The proposed VQ-VAE-2-based SemCom system is shown in Fig. 1. The deployment of this system follows the basic structure of SemCom, including encoders at the sender, decoders at the receiver, and a common embedding space as the knowledge base (KB). The sender is deployed with two encoders: an encoder for the bottom layer (Encoder B) and an encoder for the top layer (Encoder T). The input image x with semantic noise is encoded into the feature map $z^e_{\text{bottom}}(x)$ by the encoder B, and mapped into the bottom layer latent space $z_{\text{bottom}}(x)$ by vector quantization (VQ). The bottom layer feature map $z^e_{\text{bottom}}(x)$ is further encoded into the top layer feature map $z^e_{\text{top}}(x)$ by Encoder T, which generates the top layer latent space $z_{\text{top}}(x)$. The combined latent space is given by $Z = \{z_{\text{top}}(x), z_{\text{bottom}}(x)\}$. In this paper, we consider the possible injection of semantic noise on original image x, which can be injected by a malicious user who uploads the image with noise to the dataset, which can cause misclassification and degrade the image quality [20].

VQ is assisted by the embedding spaces for the top layer (TopEb) and bottom layer (BotEb), respectively. To facilitate VQ (take the top layer as an example), the feature map $z^e_{\text{top}}(x)$ is passed through a discretization bottleneck and generates the mapping index with respect to the embedding space TopEb. The discretization follows

$$q(z_{\text{top}}(x) = k|x) = \begin{cases} 1, & \text{for } k = \arg\min_j \|z^e_{\text{top}}(x) - e^j_{\text{top}}\|_2, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where $q(z_{\text{top}}(x) = k|x)$ denotes the posterior categorical probability, $z^e_{\text{top}}(x)$ denotes the output of the top layer encoder network, and e^j_{top} denotes the element with index j in the embedding space TopEb. After discretization, $z_{\text{top}}(x)$ contains the index information with respect to the embedding space. The quantization method of the bottom layer is the same as the top layer, and the difference is that the input is $z^e_{\text{bottom}}(x)$ rather than x.

The combined latent space Z is transmitted through the noisy physical channel. The received latent space is given by $Z_r = \{z^r_{\text{top}}(x), z^r_{\text{bottom}}(x)\}$, where $z^r_{\text{top}}(x)$ and $z^r_{\text{bottom}}(x)$ denote the received latent space for the top layer and the bottom layer, respectively. The feature map is recovered by mapping latent space Z to the elements in TopEb, which is given by

$$z^q_{\text{top}}(x) = e^{z^r_{\text{top}}(x)}_{\text{top}}, z^q_{\text{bottom}}(x) = e^{z^r_{\text{bottom}}(x)}_{\text{bottom}}. \tag{2}$$

The image recovery is facilitated by two steps. (1) Forwarding the top layer feature map $z^q_{\text{top}}(x)$ through the decoder for the top layer (Decoder T). (2) Forwarding both the output of Decoder T and $z^q_{\text{bottom}}(x)$ through the decoder for the bottom layer (Decoder B) to obtain the recovered image. The loss calculation is based on reconstruction mean squared error (MSE) loss and quantization loss $\mathcal{L}_q$, which are given by

$$\mathcal{L} = \text{MSE}\,(x, y) + \mathcal{L}_q, \tag{3}$$

$$\begin{aligned}\mathcal{L}_q = &\|\text{sg}[z^e_{\text{bottom}}(x)] - e^{z_{\text{top}}(x)}_{\text{bottom}}\|^2_2 + \beta\|z^e_{\text{bottom}}(x) - \text{sg}[e^{z_{\text{bottom}}(x)}_{\text{bottom}}]\|^2_2 \\ &+ \|\text{sg}[z^e_{\text{top}}(x)] - e^{z_{\text{top}}(x)}_{\text{top}}\|^2_2 + \beta\|z^e_{\text{top}}(x) - \text{sg}[e^{z_{\text{top}}(x)}_{\text{top}}]\|^2_2,\end{aligned} \tag{4}$$

where 'sg' and y denote the stop-gradient operator and recovered image, respectively. During forward computation, it functions as an identity operator, while its partial derivatives are set to zero. This design ensures that the operand acts upon remains constant and is not updated during the optimization process [28]. β is a hyperparameter set to 0.25 in our experiments. The decoder optimizes the MSE term only, and the encoder optimizes MSE and $\mathcal{L}_q$, simultaneously.

B. Preliminary Investigation and Ablation Study

The experiment results presented in Table I explore the feasibility and effectiveness of the VQ-VAE-2-based Semantic Communication (SemCom) framework. This investigation involved comprehensive experiments on eight open-source datasets,[1]

examining their performance across various noise types and levels. To assess the framework's enhancement over traditional methods, we compare it against two benchmark scenarios: traditional image compression algorithm JPEG and SemCom benchmark Joint Source-Channel Coding (JSCC) [19]. Our approach includes denoising training for both VQ-VAE and VQ-VAE-2. This process entailed introducing corresponding noise either to the input data or latent space, while using the original image as the training label. Evaluation metrics include the peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM). The compression rate ρ is set to 0.03. In the JPEG scenario, a low-density parity check (LDPC) is utilized under noisy channels to reduce the bit error rate (BER). For uniformity across different datasets, all input

[1]The datasets are accessible at **FFHQ**: https://github.com/NVlabs/ffhq-dataset, **ImageNet**: https://www.image-net.org/, **CIFAR-10** and **CIFAR-100**: https://www.cs.toronto.edu/~kriz/cifar.html, **Pokemon**: https://www.kaggle.com/datasets/vishalsubbiah/pokemon-images-and-types, **Xray**: https://www.kaggle.com/datasets/tolgadincer/labeled-chest-xray-images/data, **CityScapes**: https://www.cityscapes-dataset.com/, **Naruto**: https://www.kaggle.com/datasets/neetuk/naruto-face-dataset

TABLE I: Preliminary test on different models, noise levels, and datasets. (Metric: PSNR/SSIM; $\rho = 0.03$).

Dataset	FFHQ	ImageNet	CIFAR-100	Pokemon	X-ray	Cityscapes	CIFAR-10	Naruto
Noise Level	No Noise							
VQ-VAE-2	**30.20/0.88**	**28.50/0.86**	**36.05/0.91**	**35.50/0.95**	**34.56/0.93**	**31.01/0.91**	**35.15/0.92**	**33.22/0.96**
VQ-VAE	28.14/0.87	25.05/0.71	31.00/0.91	30.50/0.90	34.15/0.91	29.56/0.89	30.81/0.87	31.71/0.91
JPEG	23.62/0.64	21.99/0.59	25.83/0.73	25.92/0.81	25.48/0.62	22.75/0.61	24.00/0.69	22.47/0.72
JSCC	24.01/0.66	22.32/0.65	25.90/0.74	26.84/0.85	26.20/0.86	24.30/0.76	24.50/0.72	23.90/0.68
Noise Level	20dB Channel Noise							
VQ-VAE-2+Denoise	**29.72/0.85**	**27.80/0.80**	**35.56/0.94**	**32.50/0.92**	**34.47/0.92**	**31.02/0.91**	**34.16/0.91**	**32.51/0.95**
VQ-VAE+Denoise	28.92/0.84	24.91/0.77	30.58/0.92	27.73/0.88	34.30/0.90	29.50/0.89	28.20/0.86	31.52/0.91
JPEG+LDPC	21.77/0.57	20.68/0.52	23.99/0.68	25.92/0.81	25.48/0.61	22.75/0.61	24.00/0.69	22.47/0.72
JSCC+Denoise	23.61/0.64	21.52/0.64	25.02/0.72	26.62/0.82	26.18/0.85	24.01/0.75	24.16/0.68	23.53/0.67
Noise Level	20dB Semantic Noise							
VQ-VAE-2+Denoise	**29.82/0.85**	**27.29/0.79**	**35.50/0.95**	**32.68/0.91**	**34.28/0.91**	**30.89/0.92**	**32.12/0.95**	**32.61/0.95**
VQ-VAE+Denoise	29.09/0.84	25.10/0.78	30.58/0.91	28.18/0.89	33.79/0.90	29.06/0.88	28.11/0.94	30.49/0.89
JPEG+LDPC	20.17/0.53	18.87/0.47	22.02/0.63	20.88/0.61	21.97/0.58	20.97/0.62	22.97/0.67	21.75/0.70
JSCC+Denoise	21.9/0.79	20.61/0.64	21.76/0.70	21.55/0.62	26.07/0.88	23.65/0.73	20.51/0.65	23.22/0.75
Noise Level	10dB Channel Noise							
VQ-VAE-2+Denoise	**29.30/0.80**	**27.30/0.73**	**34.34/0.87**	**30.05/0.90**	**34.31/0.91**	**29.52/0.87**	**32.06/0.88**	**32.22/0.91**
VQ-VAE+Denoise	28.88/0.78	23.03/0.72	27.41/0.88	27.22/0.84	34.08/0.89	28.99/0.86	29.6/0.86	30.98/0.91
JPEG+LDPC	21.75/0.56	20.66/0.52	23.97/0.68	25.88/0.81	25.46/0.61	22.71/0.61	23.98/0.69	22.46/0.71
JSCC+Denoise	21.98/0.60	21.90/0.65	24.11/0.73	26.23/0.52	26.17/0.80	23.89/0.74	22.95/0.65	23.39/0.66
Noise Level	10dB Semantic Noise							
VQ-VAE-2+Denoise	**29.53/0.80**	**25.94/0.77**	**35.41/0.88**	**31.23/0.90**	**34.29/0.91**	**30.42/0.89**	**31.51/0.89**	**32.46/0.91**
VQ-VAE+Denoise	27.59/0.77	23.32/0.72	27.61/0.85	27.19/0.84	33.28/0.88	29.18/0.85	29.57/0.86	30.94/0.89
JPEG+LDPC	15.66/0.42	15.12/0.35	17.68/0.55	17.96/0.56	18.15/0.54	17.97/0.51	18.89/0.58	16.50/0.53
JSCC+Denoise	19.47/0.72	18.32/0.58	19.34/0.64	19.16/0.56	23.18/0.80	21.02/0.66	18.23/0.59	20.64/0.68
Noise Level	0dB Channel Noise							
VQ-VAE-2+Denoise	25.02/0.71	**21.91/0.58**	29.92/0.92	24.33/**0.77**	28.77/0.81	24.74/**0.71**	**29.35/0.91**	25.32/0.80
VQ-VAE+Denoise	**25.31**/0.73	21.77/0.57	**30.15/0.93**	**24.73**/0.76	**28.78**/0.80	**24.94/0.71**	28.64/0.89	**27.15/0.85**
JPEG+LDPC	10.36/0.05	10.29/0.04	11.46/0.18	6.38/0.03	12.24/0.16	10.66/0.13	11.86/0.18	10.46/0.16
JSCC+Denoise	19.79/**0.74**	19.16/**0.58**	20.37/0.66	19.87/0.50	25.59/**0.87**	23.22/0.69	19.64/0.63	18.89/0.54

images are resized to 256×256 pixels, enabling their testing on a consistent model. The channel model considered in FTSC framwork is Additive White Gaussian Noise (AWGN) channel. The transfer function is given by $\eta_n(z) = z + \mathcal{N}_0$, where the vector $\mathcal{N}_0 \in \mathbb{C}^k$ consists of independent identically distributed (i.i.d.) samples from a circularly symmetric complex Gaussian distribution $\mathcal{N}_0 \sim \mathcal{CN}(0, \sigma^2 \mathbf{I}_k)$, where σ^2 denotes the average noise power.

As shown in Table. I, the VQ-VAE-2 and VQ-VAE-based Semantic Communication (SemCom) framework demonstrate a substantial advantage over the traditional JPEG compression algorithm and JSCC across a variety of datasets. VQ-VAE-2 achieves slightly higher PSNR and SSIM than VQ-VAE owing to its two-tier design, which better handles image details [27]. With increasing noise levels, both VQ-VAE-2 and VQ-VAE exhibited only a minor degradation in performance. JPEG has good resilience against channel noise, but is vulnerable to semantic noise due to the lack of denoising on source data. Although JSCC is inferior to VQ-VAE-2 and VQ-VAE, it displayed resilience against both channel and semantic noise. This similarity is attributed to the autoencoder network structure within its framework, which enables denoising training. In the experiment under 0dB SNR (channel noise), the performance of JPEG+LDPC drops drastically due to the rise of the bit error rate while the machine-SemCom-based methods including JSCC and VQ-VAE, are less affected. We also find that with low signal-to-noise-ratio (SNR), VQ-VAE-2 loses its advantage over VQ-VAE. The result shows that VQ-VAE-2 still beats VQ-VAE on ImageNet and CIFAR-10, but has slightly lower performance on other tested datasets. A possible reason is that the error caused by channel noise on the top layer latent space has more effect on recovery, and further investigation is needed on this issue. To focus on the fine-tuning research in this paper, our further experiments are conducted under SNR higher than 10dB, where the experiment results in Table. I verify the feasibility of the proposed VQ-VAE-2-based SemCom framework and indicate the potentiality of enhancing image recovery quality.

To reveal the relationship between compression rate ρ and image recovery quality, we tested VQ-VAE-2 and two benchmark scenarios on the FFHQ dataset with different ρ, which is shown in Fig. 2. The reciprocal of ρ, data reduction ratio $1/\rho$, is plotted on the x-axis for better visualization, and the red dotted line denotes the unavailable data reduction rate. The curves are fitted based on the real data points, which are presented as corresponding dots. Our analysis revealed that VQ-VAE-2 can achieve the highest data reduction ratio (exceeding 600), and the JPEG algorithm struggles to reach a reduction rate of 100 on the FFHQ dataset even if the quality parameter is set to its lowest value. Among all the three

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2024.3433479

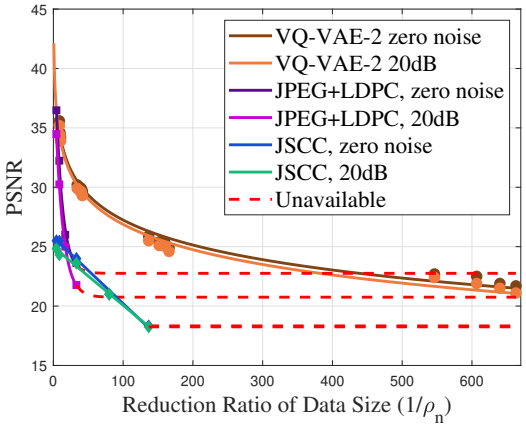


Fig. 2: Reduction rate of data size versus PSNR with different models and channel noise on the FFHQ dataset.

scenarios, VQ-VAE-2 has a close PSNR to JPEG at a low data reduction rate, but the PSNR of JPEG degrades dramatically as the data reduction ratio increases while VQ-VAE-2's PSNR decreases slower.

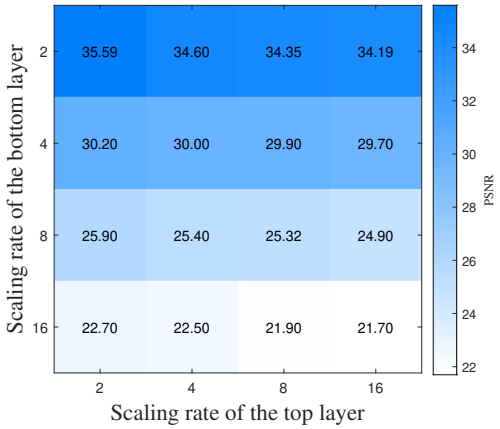


Fig. 3: The impact of scaling settings of the bottom layer and top layer on PSNR. Dataset: FFHQ without noise.

The two-tier architecture of VQ-VAE-2 allows for control of the compression rate ρ through the scaling rates of both layers. To investigate the influence of scaling rates on PSNR, we test the VQ-VAE-2 model on FFHQ without noise. The scaling rates for each layer are varied within the range of $[2, 16]$, as depicted in Fig. 3. Our findings highlight that the PSNR is predominantly influenced by the scaling rate of the bottom layer. For instance, when the scaling rate of the bottom layer is increased from 2 to 16 while maintaining the top layer's scaling rate at 2, the PSNR decreases from 35.59 to 22.70. Conversely, if we fix the bottom layer's scaling rate at 2 and increase the top layer's scaling rate from 2 to 16, the reduction in PSNR is relatively modest, amounting to only 1.0 (from 22.70 to 21.70). This phenomenon can be attributed to the encoding process in VQ-VAE-2, where the input image first passes through the bottom layer, and the encoding of the top layer is based on the output of the bottom layer. The data in Fig. 3 also follows the trend that a higher compression rate results in lower recovery quality, suggesting that there is a trade-off between these two factors.

IV. Part 1: Fine-Tunable Semantic Communication (FTSC)

TABLE II: Performance test on different models and datasets (**Metric: PSNR/SSIM**; $\rho = 0.03$; pretrained: evaluation of model pretrained on hybrid dataset formed by the pretraining group; Basic: training only on the target dataset).

Training Method	DFT	pretrained	Basic
Noise Level	No Noise		
X-ray	**38.49/0.95**	35.33/0.94	34.56/0.93
Cityscapes	**31.98/0.94**	31.20/0.92	31.01/0.91
CIFAR-10	**42.90/0.99**	42.10/0.98	35.15/0.92
Naruto	**38.5/0.97**	35.59/0.96	33.22/0.96
FFHQ	**33.7/0.94**	33.57/0.93	30.20/0.88
ImageNet	27.14/0.85	27.03/0.85	**28.50/0.86**
Noise Level	20dB Channel Noise		
X-ray	**37.36**/0.93	34.91/**0.94**	34.47/0.92
Cityscapes	**31.50/0.92**	31.30/0.90	31.02/0.91
CIFAR-10	**42.45/0.99**	41.43/0.98	34.16/0.91
Naruto	**37.40/0.96**	35.38/0.95	32.51/0.95
FFHQ	**33.50/0.94**	33.23/0.93	29.72/0.85
ImageNet	27.06/**0.85**	26.90/0.84	**27.80**/0.80
Noise Level	10dB Channel Noise		
X-ray	**36.52**/0.90	34.61/**0.92**	34.31/0.91
Cityscapes	**31.10/0.90**	30.86/0.89	29.52/0.87
CIFAR-10	**42.35/0.99**	40.77/0.98	32.06/0.88
Naruto	**36.10/0.95**	34.86/0.95	32.22/0.91
FFHQ	**32.67/0.90**	32.61/0.91	29.30/0.80
ImageNet	26.49/0.81	26.44/**0.82**	**27.30**/0.73

In the previous section, we verified the feasibility of the VQ-VAE-2-based SemCom system. Typically, in a conventional SemCom system, such a pretrained model is considered ready for deployment, i.e., deploying the encoder and the decoder at the sender and receiver ends, respectively. However, the pretrained model may face challenges from out-of-distribution data, such as images that are significantly differ from those in the training dataset. A potential solution is to apply transfer learning to the pretrained model, but this approach often incurs additional communication overhead due to the transmission of gradients and raw images. To circumvent this issue and facilitate transfer learning at a lower communication cost, we will introduce a novel transfer learning framework to establish the Fine-Tunable Semantic Communication (FTSC) system in this section, and conduct further study on its implementation in next section. The proposed FTSC framework consists of two distinct approaches: Decoder Fine-Tuning (DFT) and Latent Space-based Decoder Fine-Tuning (LSDFT), which will be introduced in the subsequent subsections.

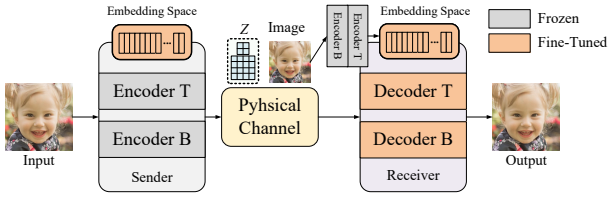


Fig. 4: Decoder fine-tuning (DFT) framework.

A. Decoder Fine-Tuning (DFT)

As shown in Fig. 4, the decoder fine-tuning model consists of frozen encoders, tunable embedding spaces deployed at both the sender and the receiver, and the decoders deployed at receiver only. To achieve high-quality semantic image transmission with unknown datasets, the SemCom process is divided into two phases: (1) the fine-tuning phase and (2) the transmission phase. At the beginning of transmission, the encoder and decoder networks at both sides are loaded with pre-trained parameters. The tunable embedding spaces at both sides are reset to default E_d, which is the same as in the pretrained model. During the fine-tuning phase, the embedding space at the sender is updated by the quantization loss, given by

$$\begin{aligned}\mathcal{L}_q = &\|\text{sg}[z^e_{\text{bottom}}(x)] - e^{z_{\text{top}}(x)}_{\text{bottom}}\|_2^2 + \beta\|z^e_{\text{bottom}}(x) - \text{sg}[e^{z_{\text{bottom}}(x)}_{\text{bottom}}]\|_2^2 \\ &+ \|\text{sg}[z^e_{\text{top}}(x)] - e^{z_{\text{top}}(x)}_{\text{top}}\|_2^2 + \beta\|z^e_{\text{top}}(x) - \text{sg}[e^{z_{\text{top}}(x)}_{\text{top}}]\|_2^2. \end{aligned} \tag{5}$$

The original image and the latent space produced at the sender end are transmitted through the physical channel. Upon receipt of this data, the receiver utilizes the decoders to reconstruct the image, and updates the decoder networks based on the loss calculated from the reconstructed and the original images, which is given by

$$\mathcal{L} = \text{MSE}\,(x, y) + \mathcal{L}_q, \tag{6}$$

where the embedding space at the receiver is updated by the quantization loss $\mathcal{L}_q$ calculated with the assistance of the frozen encoders and the received image by the same method as in (5). As the input image, frozen encoder, and the initial value of embedding spaces are the same, the consistency of embedding space at both sender and receiver can be ensured.

With the same $\mathcal{L}_q$ calculated at both sides, the embedding spaces with the same initial value are updated and synchronized during the fine-tuning process. Once the fine-tuning process reaches convergence, indicating that the SemCom system has adequately adapted to the new dataset, the system transitions from the fine-tuning phase to the transmission phase. This approach ensures that the system remains versatile and effective, even when dealing with previously unseen data. The DFT framework avoids the communication cost of gradient transmission and embedding space initialization during the fine-tuning process. The additional communication cost for fine-tuning is mainly caused by the transmission of the original image because the data size of $\mathcal{L}_q$ is negligible.

To simulate the performance of DFT, we split the eight tested datasets into the pretraining group (FFHQ, ImageNet, CIFAR-100, and Pokemon) and the testing group (X-ray, Cityscapes, CIFAR-10, and Naruto). The training set for the fine-tuning phase is sampled from the target dataset with ratio less than 5%. The VQ-VAE-2 model is pretrained on a combined dataset of the pretraining group, and then applied to the testing group with DFT. The simulation results are presented in Table. II, where 'DFT' denotes the proposed method, 'pretrained' denotes directly applying the pretrained model on a combined dataset formed by the pretraining group to unknown datasets, and 'Basic' denotes training the VQ-VAE-2 model only on the target dataset.

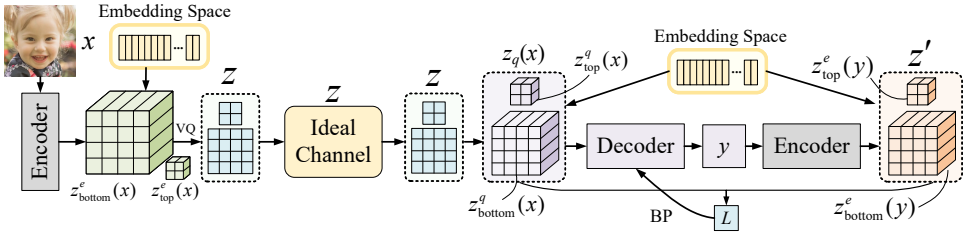
Based on the simulation results, it is evident that the proposed DFT scenario enhances the PSNR across various testing datasets under different channel noise conditions, outperforming both the pretrained model and the basic training scenario. The pretrained model in the pretraining group achieves higher PSNR compared to the basic training in the testing group because the additional training data out of the target dataset plays the role of data augmentation. We also tested two datasets in pretraining group, FFHQ and ImageNet. Results on FFHQ align with the phenomenon on testing groups, but the ImageNet shows different results. Although DFT improves the PSNR, the pretrained model achieves lower PSNR than the basic training. A possible cause is that FFHQ is a highly dedicated dataset on human faces, while ImageNet covers a wide variety of images. For FFHQ, training together with other datasets, such as ImageNet, functions as data augmentation. However, the training on ImageNet with other dedicated datasets introduces interference into the original distribution. Another observation from Table. II is that the maximum PSNR does not always appear together with the maximum SSIM. A possible reason is that our loss is calculated based on mean squared error (MSE), which directly relates to PSNR rather than SSIM. Thus, we consider PSNR as the main metric, and SSIM is considered as an auxiliary metric.

B. Latent Space-based Decoder Fine-tuning (LSDFT)

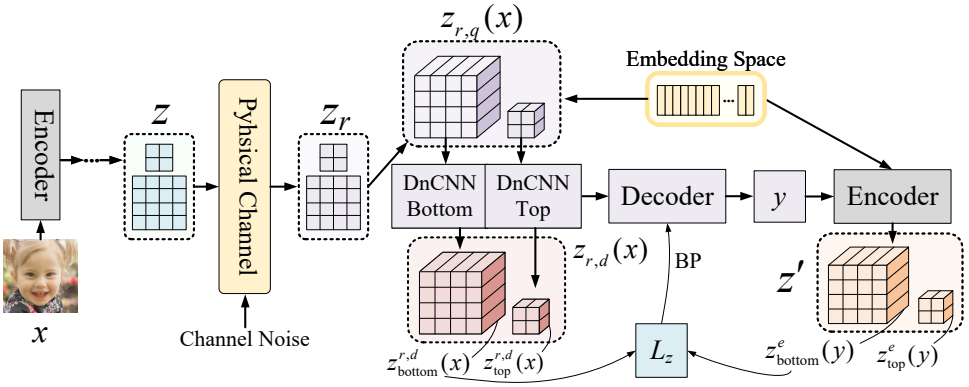
The DFT scenario allows for fine-tuning without gradient transmission, but it still requires additional transmission of original images. To further reduce the communication cost, we propose the latent space-based decoder fine-tuning (LSDFT) to enable fine-tuning without original image transmission. There are two differences between LSDFT and DFT: (i) DFT calculates loss based on raw data, but LSDFT calculates loss based on an intermediate way. (ii) The communication cost reduction ratio of LSDFT in fine-tuning phase can be the same as the compression ratio. To deal with both ideal channel and noisy channels, we propose the following two distinct LSDFT designs.

1) Design for Ideal Channel: The system model of LSDFT under an ideal channel, i.e., the channel without noise, is shown in Fig. 5 (a). At the sender end, the quantization loss $\mathcal{L}_q$ is calculated in addition to the latent space Z, and both $\mathcal{L}_q$ and Z are transmitted to the receiver. In this scenario, we assume that a frozen encoder and a fine-tunable decoder are deployed at the receiver end. The decoder generates output y from the received latent space Z. Different from traditional transfer learning that requires original image x as the label, the recovery loss of the decoder is inferred by

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2024.3433479



(a) System model of LSDFT with ideal channel.



(b) System model of LSDFT with noisy channel.

Fig. 5: System models of LSDFT under idea channel and noisy channel.

$$\mathcal{L} = \text{MSE}\left(z^q_{\text{bottom}}(x), z^e_{\text{bottom}}(y)\right). \tag{7}$$

With a well-trained embedding space, the retrieved feature map $z^q_{\text{bottom}}(x)$ is expected to closely approximate $z^e_{\text{bottom}}(x)$. As the original image x and decoder output y pass through the same encoder network to generate $z^e_{\text{bottom}}(x)$ and $z^e_{\text{bottom}}(y)$, respectively, reducing the difference between x and y is positively correlated to minimizing the inferred loss $\mathcal{L}$. Note that the reduced communication cost of LSDFT is at the cost of inaccurate loss calculation due to the non-zero quantization loss in practical implementation, which results in pixel-level deviant in the recovered images, as shown in Fig. 6. Compared to the quality degradation of JPEG, LSDFT controls the deviance of pixel blocks on a smaller scale.

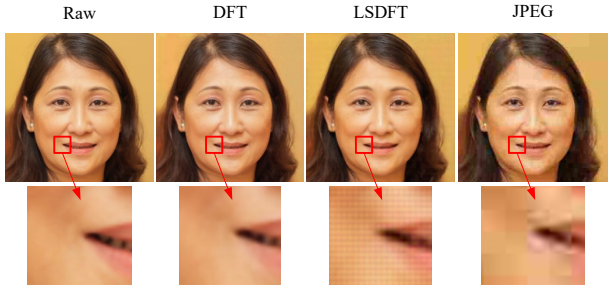


Fig. 6: Recovery detail comparison on FFHQ.

2) Design for Noisy Channel: The errorless transmission of the latent space, which functions as the label during the fine-tuning process, is critical for LSDFT. To improve the robustness against channel noise, we propose a denoising model-based system model for the case with corrupted latent space, which is shown in Fig. 5 (b).

With the assumption of detectable channel noise, we deploy pretrained DnCNN [60] networks at the receiver end to recover clean latent space before fine-tuning. Due to the impact of channel noise, the received latent is given by $z_r = z + \tilde{n}$, where $\tilde{n}$ denotes the channel noise. The retrieved feature map $z_{r,q}(x)$ is divided into the top layer and the bottom layer, which are denoised by two different pretrained DnCNN networks to obtain $z^{r,d}_{\text{bottom}}(x)$ and $z^{r,d}_{\text{top}}(x)$, respectively. The fine-tuning process takes the denoised latent space as the label, and the quantization loss is assumed to be correctly transmitted.

TABLE III: Performance comparison of receiver side fine tuning and benchmark methods(Metric: PSNR; $\rho = 0.03$; LSDFT: the proposed latent space-based decoder fine-tuning; PT: applying pretrained model on tested datasets).

Model	LSDFT	DFT	PT	JPEG	JSCC
Data Type	PSNR	Δ PSNR			
Noise Level	No Noise				
X-ray	35.89	-2.60	+1.33	+10.41	+9.69
Cityscapes	31.23	-0.75	+0.22	+8.48	+6.93
CIFAR-10	37.15	-5.75	+2.00	+13.15	+12.65
Naruto	35.67	-2.83	+2.45	+13.20	+11.77
FFHQ	29.66	-4.04	-0.54	+6.04	+5.65
ImageNet	24.93	-2.21	-3.57	+2.94	+2.61
Noise Level	20dB Channel Noise				
X-ray	34.33	-3.03	-0.14	+8.85	+8.15
Cityscapes	29.30	-2.20	-1.72	+6.55	+5.29
CIFAR-10	36.19	-6.26	+2.03	+12.19	+12.03
Naruto	33.58	-3.82	+1.07	+11.11	+10.05
FFHQ	29.46	-4.04	-0.26	+7.69	+5.85
ImageNet	23.96	-3.10	-3.74	+3.28	+2.44

3) Simulation on LSDFT: To assess the efficacy of the proposed LSDFT framework, we test the PSNR of the recovered images on multiple datasets under noisy and ideal channel assumptions. The outcomes are detailed in Table. III, where PT denotes directly applying the model pretrained on the pretraining group (FFHQ, ImageNet, CIFAR-100, and Pokemon) to the tested datasets.

Our analysis reveals that the performance of LSDFT is intermediate between DFT and PT on the testing group (X-ray, Cityscapes, CIFAR-10, and Naruto) under the assumption of errorless latent space transmission, i.e., ideal channel without noise. These findings suggest that LSDFT is a viable option for fine-tuning pretrained models without the transmission of original images. However, it is important to note that as a trade-off for reduced data size during the fine-tuning phase, LSDFT's PSNR is lower than that of DFT. This is attributed to the accumulated error on $z_{r,q}(x)$, which is caused by the non-zero quantization loss. Within the scope of the pretrained datasets (FFHQ and ImageNet), LSDFT does not demonstrate additional performance enhancements. Therefore, the choice of fine-tuning approach should be contingent upon the data source. In scenarios involving a noisy channel, LSDFT shows an improvement in PSNR over the pretrained model on datasets like CIFAR-10 and Naruto, but experiences performance degradation on X-ray and Cityscape datasets. This pattern suggests that LSDFT is vulnerable to noise even if a denoising technique is adopted. The noise on latent space affects not only the recovery quality but also impacts the loss calculation. In a broader context, both LSDFT and DFT exhibit superior performance compared to JPEG and JSCC on unknown datasets and enable post-deployment fine-tuning for SemCom systems.

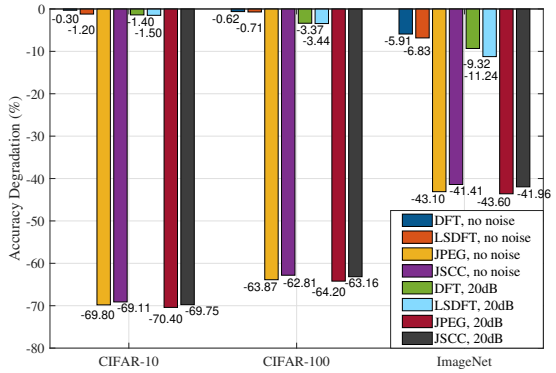


Fig. 7: Top-1 Accuracy test.

To verify the performance of LSDFT in image classification tasks, we tested the degradation of top-1 classification accuracy compared to the original data on CIFAR-10, CIFAR-100, and ImageNet, as shown in Fig. 7. With the compression rate of 0.03, JPEG and JSCC show obvious accuracy drops on all three datasets, while DFT and LSDFT maintain performance close to those of the original images on CIFAR-10 and CIFAR-100. It is important to note that the proposed LSDFT and DFT frameworks do experience a recognizable decrease in accuracy

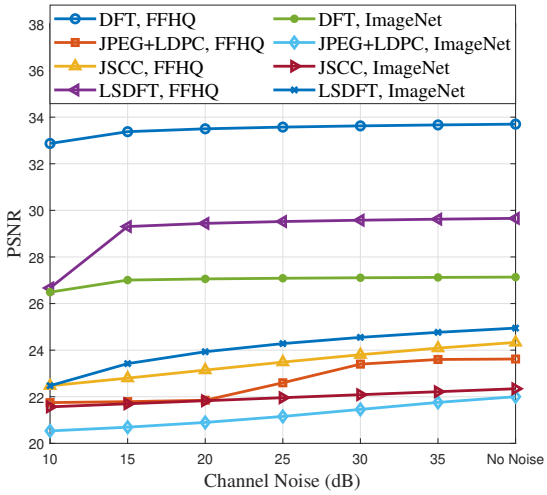


Fig. 8: Channel noise versus PSNR with different models and datasets.

on the ImageNet dataset. This reduction can be attributed to the vast number of classes in ImageNet, which renders the model more sensitive to even minor alterations in the image.

The performance of the proposed LSDFT, DFT, and two benchmark scenarios under varying levels of channel noise is depicted in Fig. 8. As channel noise decreases, all four scenarios exhibit an increase in PSNR for both FFHQ and ImageNet datasets, suggesting that while the impact of noise on image recovery quality can be mitigated, it cannot be eliminated. The DFT scenario is regarded as the theoretical upper limit for LSDFT performance because its training labels are the original images. The results from the simulations show that DFT consistently surpasses LSDFT and other benchmark scenarios in all test cases. Although LSDFT does not reach the same PSNR as DFT, it still outperforms both JSCC and JPEG combined with LDPC, confirming its superiority over traditional SemCom scenarios. Additionally, it's observed that all tested scenarios demonstrate varied performance across different datasets, indicating that the actual effectiveness of a SemCom system is likely influenced by the characteristics of the data.

The image recovery test cases of the proposed DFT and LSDFT on FFHQ, CIFAR-10, and X-ray datasets are presented in Fig. 10. We can find that the image recovered by DFT is very close to the raw image. The recovered image by JPEG has obvious compression artifacts due to the small compression rate 0.03. The recovery quality of LSDFT is lower than DFT but higher than JPEG.

The convergence analysis of DFT and LSDFT are presented in Fig. 11. The experiments are conducted on Xray dataset and CIFAR-10, where a batch of 64 images are fed into the network during each iteration. We can find that both proposed algorithms converge within 600 iterations, with a small fluctuation of PSNR value caused by the difference among batches. To reduce the communication cost during the fine-tuning phase, we sample 5% of images from the aimed dataset as training sets, and the transmitted image or latent

space can be reused during fine-tuning. The computational complexity analysis can be found in Section V.

To summarize the FTSC framework with the design of DFT and LSDFT, DFT facilitates fine-tuning in Semantic Communication (SemCom) without the need for gradient transmission, and LSDFT further reduces additional data transmission through the implementation of intermediate loss calculation. However, this conservation of communication resources comes at the expense of decreased recovery quality and a less robust training process. A similar trade-off is also evident in the relationship between the compression rate and recovery quality, where a lower compression rate typically results in diminished recovery quality, as previously presented in Fig. 2. Therefore, it is crucial to understand when and how to effectively utilize SemCom systems such as FTSC with given constraints of computation and communication resources, which will be discussed in the next section.

V. Part 2: Performance Optimization for the SemCom System Proposed in Part 1

The FTSC framework proposed in the Part 1 focuses on the generalizability of SemCom systems, but its effectiveness is based on the assumption of adequate computing resources. This limitation exists not only in FTSC but also other autoencoder-based SemCom systems due to the computation overhead of the encoder and decoder network. In some practical communication applications, there may be practical resource limitations which constraints the utilization of SemCom. To find the balance between additional computational cost and recovery quality of SemCom systems such as the proposed FTSC in practical communication tasks, this section studies the joint scaling rate and communication resource allocation optimization problem in a typical downlink image SemCom task with a single sender and N receivers, as shown in Fig. 9. The receiver set is denoted by $\mathcal{N} := \{1, 2, \ldots, N\}$. To accommodate extremely limited communication resources, we specify the SemCom scenario as LSDFT with fine-tuned encoder and decoders. To investigate its application, we study a joint resource allocation and model configuration problem in the following subsections.

A. Problem Formulation and Solution

A closed-form relationship between PSNR and ρ on the FFHQ dataset can be simulated through experimental data points, which is given by

$$Q_n = a_n^q + b_n^q \ln(\rho_n), \tag{8}$$

where Q_n denotes the PSNR of the n-th receiver, a_q and b_q are fitted parameters. The fitting is based on the "fit(·)" function in Matlab, with fit type $a+b\ln x$. For the case without channel noise, $a_q = 41.22$ and $b_q = 3.035$ (Root Mean Square Error (RMSE) = 0.4959, R-square = 0.9904). For the case with 20dB channel noise, $a_q = 41.03$, $b_q = 3.073$ (RMSE = 0.4582, R-square = 0.9920). The input data size D_0 is given by

$$D_0 = 3 \times H \times W, \tag{9}$$

where H and W denote the height and width of the image, respectively. The number of channels is given by 3 because the datasets contain RGB images. The compression rate ρ is calculated according to the scaling rate in each layer, e.g., with 512 entries in the embedding space, applying scaling rate [4,2] to a 256×256 RGB image results in a bottom layer latent space with shape 64×64 and a top layer latent space with shape 32×32. In this case, the compression rate is given by

$$\rho_{[4,2]} = \frac{64 \times 64 + 32 \times 32}{256 \times 256 \times 3} \times \frac{9\text{bits}}{8\text{bits}} \approx 0.03, \tag{10}$$

where the ratio of bits term is based on the required bits to save each pixel value. For an RGB image, the pixel value in each channel varies from 0 to 255, which requires 8 bits of storage. For the latent space, the information in each pixel is represented by an index within $[1, 512]$, which requires 9 bits storage.

The computational complexity can be approximated by the complexity of convolution layers, which takes the major part of the calculation. The computational complexity of a general convolution layer is given by $\mathrm{O}(M^2k^2C_{\text{in}}C_{\text{out}})$, where k^2 denotes the kernel size, M^2 denotes the size of output feature map, C_{in} denotes the number of input channels, and C_{out} denotes the number of output channels. In the case of VQ-VAE-2 with scaling rate $[2, 2]$, the convolution layer complexity of the encoder/decoder is given by

$$C_{[2,2]} = C_0 - C_\Delta + \frac{1}{4}C_0, \tag{11}$$

where C_0 and C_Δ are given by

$$\begin{aligned} C_0 &= \frac{1}{4}HWk^2{C_{\text{ch}}}^2, \\ C_\Delta &= \frac{1}{4}HWk^2C_{\text{ch}}(C_{\text{ch}} - 3), \end{aligned} \tag{12}$$

where C_{ch} denotes the number of filters. With the increase of scaling rate at the bottom layer, the computational complexity follows a geometric sequence, and the closed-form relationship between compression rate and scaling rate s of the bottom layer is given by

$$C = -C_\Delta + \frac{4C_0(1 - 0.25^{\log_2(s)+1})}{3}, \tag{13}$$

where the scaling rate of the top layer is fixed as 2. The compression rate of n-th receiver with respect to s is given by

$$\rho_n = 16\rho_{[2,2]} \times 0.25^{\log_2(s)+1}. \tag{14}$$

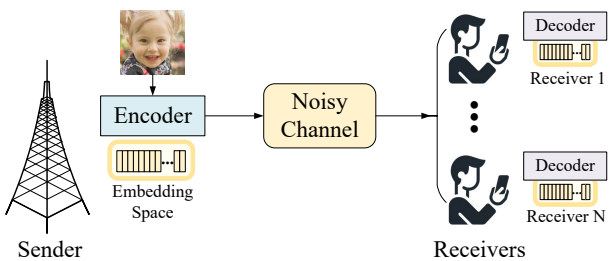


Fig. 9: Downlink image transmission task with single sender and multiple receivers.

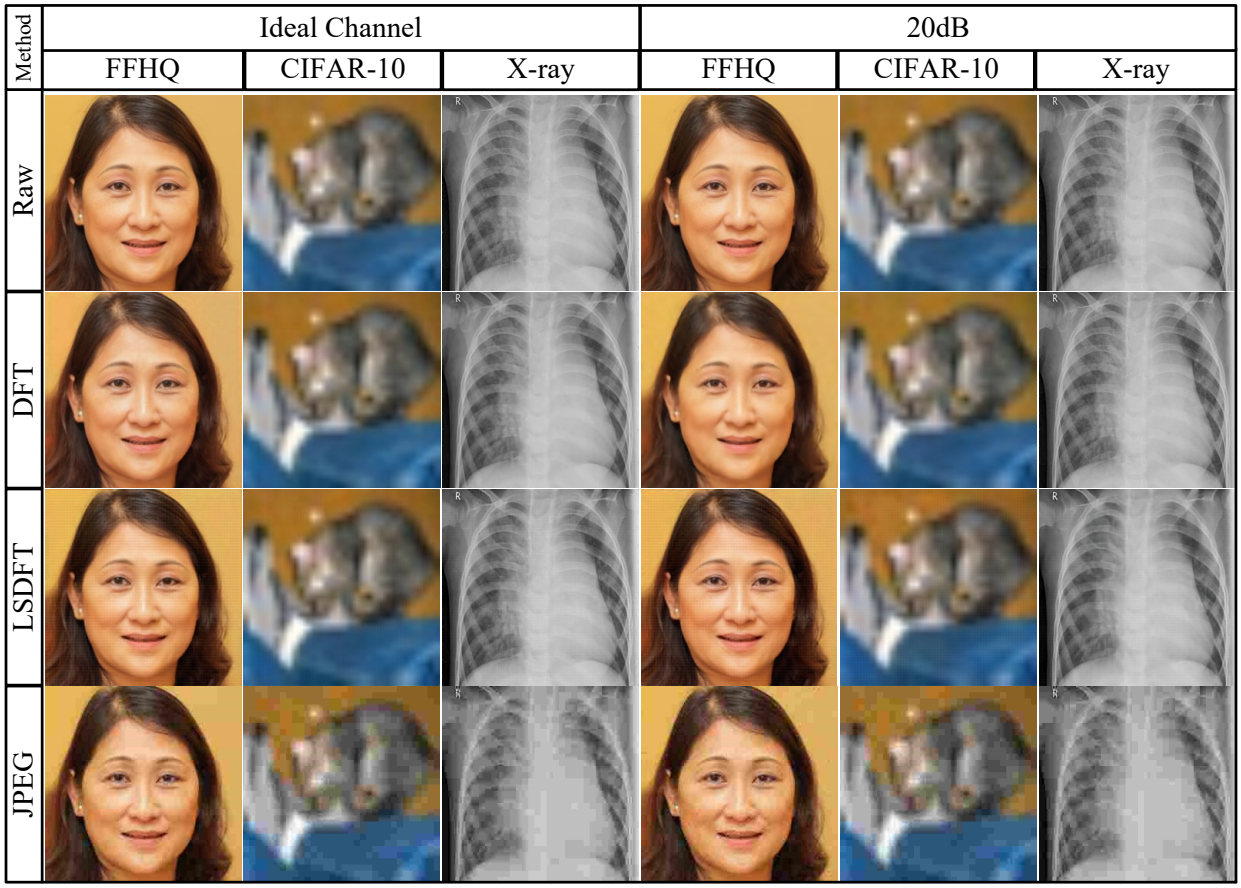


Fig. 10: Image recovery test cases of the proposed DFT and LSDFT on FFHQ, CIFAR-10, and X-ray datasets.

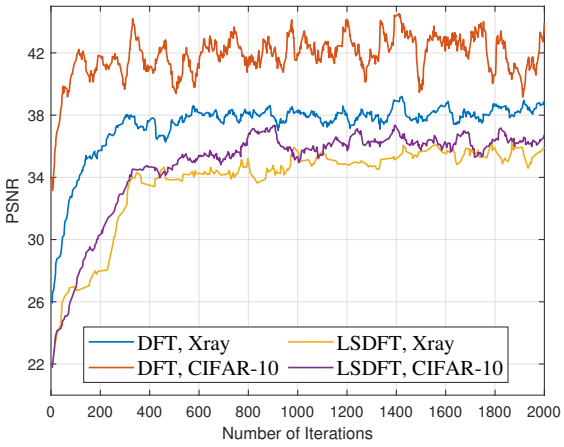


Fig. 11: Convergence analysis of the proposed algorithms

Substituting (14) into (13), the computational complexity can be written as

$$C = -C_\Delta + \frac{4}{3}C_0\left(1 - \frac{\rho_n}{\rho_0}\right), \tag{15}$$

where $\rho_0 = 16\rho_{[2,2]}$. The required time for latent space transmission is given by

$$T_n^{\text{trans}} = \frac{\rho_n D_0}{b_n \log_2\left(1 + \frac{p_n g_n}{b_n \sigma^2}\right)}, \tag{16}$$

where b_n denotes the allocated bandwidth for receiver n, σ^2 denotes the noise power spectral density, and g_n denotes the channel power gain of the n-th receiver. The required time for computation is approximated by

$$T_n^{\text{comp}} = \frac{-C_\Delta + \frac{4}{3}C_0\left(1 - \frac{\rho_n}{\rho_0}\right)}{f}, \tag{17}$$

where f denotes the GPU frequency. Thus, the time delay of n-th receiver is given by

$$T_n = T_n^{\text{trans}} + T_n^{\text{comp}}. \tag{18}$$

The energy cost is calculated by

$$E_{\text{total}} = E_t + E_c, \tag{19}$$

where E_t denotes the sum energy cost for data transmission, and E_c denotes the sum energy cost for computation, which

are given by

$$E_{\mathrm{t}} = \sum_{n\in\mathcal{N}} \frac{p_n \rho_n D_0}{b_n \log_2\left(1+\frac{p_n g_n}{b_n \sigma^2}\right)}, \tag{20}$$

$$E_c = \sum_{n\in\mathcal{N}} \omega_c f^2 \left(-C_\Delta + \frac{4}{3} C_0 \left(1-\frac{\rho_n}{\rho_0}\right)\right), \tag{21}$$

where ω_c denotes the computational efficiency of GPU. Data rate r_n is given by

$$r_n = b_n \log_2\left(1+\frac{p_n g_n}{b_n \sigma^2}\right), \tag{22}$$

where

$$g_n = \beta_0 d_n^{-\alpha} \left\| \sqrt{\frac{K}{K+1}}\hat{g}_n + \sqrt{\frac{1}{K+1}}\tilde{g}_n \right\|^2, \tag{23}$$

where β_0 denotes the channel gain at the reference distance $d_0 = 1\mathrm{m}$, α denotes the path loss exponent (in this paper we assume that $\alpha = 2$). g_n denotes the deterministic LoS channel component with $|\hat{g}_n| = 1$, and $\widetilde{g}_n$ denotes the random scattered component. The Rician factor is denoted by K. Thus, the utility function is given by

$$\begin{aligned} U &= \lambda_1 AQ - \lambda_2 B T_{\max} - \lambda_3 C E_{\text{total}} \\ &= A\left(a_q - b_q \ln\left(\frac{1}{\rho_n}\right)\right) \\ &- B\max\left(\frac{\rho_n D_0}{r_n} + \frac{-C_\Delta + \frac{4}{3}C_0\left(1-\frac{\rho_n}{\rho_0}\right)}{f}\right) \\ &- C\sum_{n\in\mathcal{N}}\left(\frac{p_n\rho_n D_0}{r_n} + \omega_c f^2\left(-C_\Delta + \frac{4}{3}C_0\left(1-\frac{\rho_n}{\rho_0}\right)\right)\right), \end{aligned} \tag{24}$$

where λ_1, λ_2, and λ_3 are hyperparameters that denote the importance of data recovery quality, time delay, and total energy cost, respectively. Based on the utility function, the optimization problem is formulated as $P1$:

$$\begin{aligned} P1: &\max_{\rho_n, b_n, p_n, T} \lambda_1\left(a_q - b_q\ln\left(\frac{1}{\rho_n}\right)\right) - \lambda_2 T \\ &- \lambda_3 \sum_{n\in\mathcal{N}}\left(\frac{p_n\rho_n D_0}{r_n} + \omega_c f^2\left(-C_\Delta + \frac{4C_0}{3}\left(1-\frac{\rho_n}{\rho_0}\right)\right)\right) \\ s.t.\ & C1: \sum_{n\in\mathcal{N}} b_n \le b_{\max},\ C2: \sum_{n\in\mathcal{N}} p_n \le p_{\max}, \\ & C3: 0 < \rho_n \le \rho_{[2,2]}, n\in\mathcal{N}, \\ & C4: \frac{\rho_n D_0}{r_n} + \frac{-C_\Delta + \frac{4}{3}C_0\left(1-\frac{\rho_n}{\rho_0}\right)}{f} \le T, \forall n\in\mathcal{N}, \end{aligned} \tag{25}$$

where the maximization in (24) is transferred into constraint $C4$. $b_{\max}$ and $p_{\max}$ denote the sum bandwidth and transmission power limitation, respectively. The optimization problem $P1$ is non-convex due to the term $\frac{p_n\rho_n D_0}{r_n}$. To address this issue, we introduce a set of auxiliary variables $\theta_n, n\in\mathcal{N}$, given by

$$\theta_n = \frac{1}{2 p_n \rho_n r_n}, \tag{26}$$

where the non-convex term is transferred as

$$\frac{p_n\rho_n}{r_n} = (p_n\rho_n)^2\theta_n + \frac{1}{4\theta_n r_n^2}. \tag{27}$$

With this auxiliary variable, the coupling between $p_n\rho_n$ and r_n is removed, and the problem can be solved in an alternating manner. With given $p_n^{(i-1)}$, $T^{(i-1)}$, $b_n^{(i-1)}$, $\rho_n^{(i-1)}$ from the previous iteration, the optimization problem with respect to θ_n at i-th iteration is given by

$$\begin{aligned} P1.1: &\max_{\theta_n^{(i)}} \lambda_1 Q^{(i-1)} - \lambda_2 T^{(i-1)} \\ &- \lambda_3 \sum_{n\in\mathcal{N}} D_0\left(\left(p_n^{(i-1)}\rho_n^{i-1}\right)^2\theta_n^{(i)} + \frac{1}{4\theta_n^{(i)}\left(r_n^{(i-1)}\right)^2}\right) \\ &- \lambda_3 \sum_{n\in\mathcal{N}} \omega_c f^2\left(-C_\Delta + \frac{4}{3}C_0\left(1-\frac{\rho_n^{(i-1)}}{\rho_0}\right)\right) \\ s.t.\ & \theta_n^{(i)} > 0, n\in\mathcal{N}. \end{aligned} \tag{28}$$

Optimization problem $P1.1$ is convex with respect to θ. The obtained $\theta^{(i)}$ in i-th iteration is substituted into the optimization problem with respect to the rest of the variables, which formulates

$$\begin{aligned} P2.1: &\max_{\rho_n^{(i)}, b_n^{(i)}, p_n^{(i)}, T^{(i)}} \lambda_1 \sum_{n\in\mathcal{N}}\left(a_q - b_q\ln\left(\frac{1}{\rho_n^{(i)}}\right)\right) - \lambda_2 T^{(i)} \\ &- \lambda_3 \sum_{n\in\mathcal{N}} D_0\left(\left(p_n^{(i)}\rho_n^{(i)}\right)^2\theta_n^{(i)} + \frac{1}{4\theta_n^{(i)}\left(r_n^{(i)}\right)^2}\right) \\ &- \lambda_3 \sum_{n\in\mathcal{N}} \omega_c f^2\left(-C_\Delta + \frac{4}{3}C_0\left(1-\frac{\rho_n^{(i)}}{\rho_0}\right)\right) \\ s.t.\quad & C1 - C3 \\ & C4: \frac{\rho_n D_0}{r_n} + \frac{-C_\Delta + \frac{4}{3}C_0\left(1-\frac{\rho_n}{\rho_0}\right)}{f} \le T, \forall n\in\mathcal{N}. \end{aligned} \tag{29}$$

Problem $P2.1$ is still non-convex due to the coupling between $p_n^{(i)}$ and $\rho_n^{(i)}$. The variable $\rho_n^{(i)}$ also couples with data rate r_n in $C4$. To address this issue, we divide the variables into two groups, $\rho_n^{(i)}$ and $\{p_n^{(i)}, b_n^{(i)}, T^{(i)}\}$ and optimize them alternatively by $P2.2$ and $P2.3$, which are given by

$$\begin{aligned} P2.2: &\max_{\rho_n^{(i)}} \lambda_1 \sum_{n\in\mathcal{N}}\left(a_q - b_q\ln\left(\frac{1}{\rho_n^{(i)}}\right)\right) - \lambda_2 T^{(i-1)} \\ &- \lambda_3 \sum_{n\in\mathcal{N}} D_0\left(\left(p_n^{(i-1)}\rho_n^{(i)}\right)^2\theta_n^{(i)} + \frac{1}{4\theta_n^{(i)}\left(r_n^{(i-1)}\right)^2}\right) \\ &- \lambda_3 \sum_{n\in\mathcal{N}} \omega_c f^2\left(-C_\Delta + \frac{4}{3}C_0\left(1-\frac{\rho_n^{(i)}}{\rho_0}\right)\right) \\ s.t.\ & C3, \\ & \frac{\rho_n^{(i)} D_0}{r_n^{(i-1)}} + \frac{-C_\Delta + \frac{4}{3}C_0\left(1-\frac{\rho_n^{(i)}}{\rho_0}\right)}{f} \le T^{(i-1)}, \forall n\in\mathcal{N} \end{aligned} \tag{30}$$

and

$$P2.3: \max_{p_n^{(i)}, b_n^{(i)}, T^{(i)}} \lambda_1 \sum_{n\in\mathcal{N}} \left(a_q - b_q \ln\left(\frac{1}{\rho_n^{(i)}}\right)\right) - \lambda_2 T^{(i)} - \lambda_3 \sum_{n\in\mathcal{N}} D_0 \left(\left(p_n^{(i)}\rho_n^{(i)}\right)^2 \theta_n^{(i)} + \frac{1}{4\theta_n^{(i)}\left(r_n^{(i)}\right)^2}\right) - \lambda_3 \sum_{n\in\mathcal{N}} \omega_c f^2 \left(-C_\Delta + \frac{4}{3} C_0 \left(1 - \frac{\rho_n^{(i)}}{\rho_0}\right)\right)$$
$$s.t. \quad C1, C2$$
$$\frac{\rho_n^{(i)} D_0}{r_n^{(i)}} + \frac{-C_\Delta + \frac{4}{3} C_0 \left(1 - \frac{\rho_n^{(i)}}{\rho_0}\right)}{f} \leq T^{(i)}, \forall n \in \mathcal{N}. \tag{31}$$

Problems $P2.2$ and $P2.3$ are convex optimization problems, which can be solved by CVX. By iteratively solving problems $P2.1$, $P2.2$, and $P2.3$, a sub-optimal solution can be obtained. The whole process is presented in **Algorithm** 1.

Algorithm 1 Joint optimization of ρ, b, p, T for utility maximization.

Initialize: $\rho_n^{(0)}$, $p_n^{(0)}$, $b_n^{(0)}$, and $T^{(0)}$. Iteration index $i = 1$.
while $|U^{(i)} - U^{(i-1)}| \geq \epsilon$ **do**
  Solve problem $P2.1$ to obtain $\theta_n^{(i)}$ by substituting $\rho_n^{(i-1)}$, $p_n^{(i-1)}$, $b_n^{(i-1)}$, and $T^{(i-1)}$
  Solve problem $P2.2$ to obtain $\rho_n^{(i)}$ with given $\theta_n^{(i)}$, $p_n^{(i-1)}$, $b_n^{(i-1)}$, and $T^{(i-1)}$
  Solve problem $P2.3$ to obtain $p_n^{(i)}$, $b_n^{(i)}$, and $T^{(i)}$
  $i = i + 1$
end while

B. Simulation Results of the Utility Maximization Problem in Downlink SemCom

The simulation results of the proposed optimization algorithm for utility maximization of the LSDFT SemCom system are presented in this section. In the simulation, the number of receivers is fixed at $N = 5$, and the noise power spectral density is given by -174dBm/Hz. The computational efficiency ω_c and GPU frequency f are given by $\omega_c = 10^{-38}$ and $f = 1350$MHz, respectively [61]. The default maximum sum transmission power and sum bandwidth are given by $p_{\max} = 50$W and $b_{\max} = 5000$Hz, respectively. The default distance vector for single sender and multiple receivers is randomly generated from 1 to 1000 meters through a uniform distribution. Benchmark scenarios include

- **JPEG+LDPC:** The required computation for JPEG is much less than SemCom scenarios including JSCC and our proposed methods. Thus, the computational cost term is removed from the utility calculation in this benchmark, which makes the optimization problem easy to solve.
- **JSCC:** A SemCom benchmark based on neural network, and its neural network structure is given in [19]. The first four layers in the encoder and all the layers in the decoder are with fixed size. The compression rate only influences the size of the fifth layer with description "Conv $5 \times 5 \times c/1$", where parameter c decides the compression rate. Thus, its computational complexity can be written as $C = A + B\rho_n$ The optimization for this scenario is easy to solve because the computation cost is a linear term.

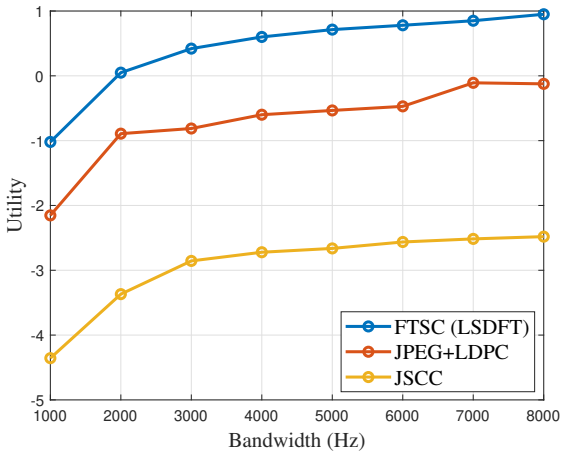


Fig. 12: Impact of total bandwidth on utility.

The utility across various total bandwidth limits, denoted as $b_{\max}$, is depicted in Fig. 12. For this simulation, the hyperparameters are set to $\lambda_1 = \lambda_2 = \lambda_3 = 0.33$, reflecting the equal weighting of recovery quality, time delay, and energy cost. The total bandwidth constraint varies from 1000 to 8000 to simulate scenarios under extremely limited communication resources. It is apparent that the proposed LSDFT scenario surpasses both JSCC and JPEG combined with LDPC within the examined bandwidth limitation, demonstrating its effectiveness. Notably, the JPEG combined with the LDPC scenario achieves higher utility than JSCC, attributable to its lower computational demands while maintaining close recovery quality, particularly at a compression rate of 0.03.

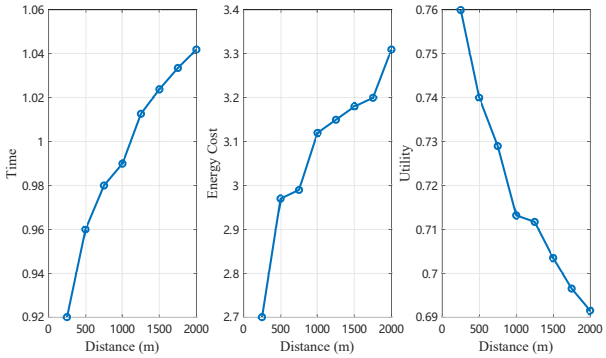


Fig. 13: Impacts of distance between sender and receivers on time, energy cost, and utility.

The impact of the sender-receiver distance on time delay, energy cost, and overall utility is illustrated in Fig. 13. In this simulation, distances are systematically generated by scaling the default distance vector to minimize the influence of random

variation. It is observed that an increase in distance leads to a marginal rise in time delay. The time delay is determined by both computing time and transmission time, and this effect mitigates the impact of reduced channel gain. In contrast, energy costs escalate more significantly with increased distance, as transmitting over longer distances demands greater energy. The utility curve clearly demonstrates that greater distances between the sender and receivers diminish the overall utility of the SemCom system. This reduction in utility is primarily due to lower channel efficiency, resulting in prolonged transmission times and larger energy costs.

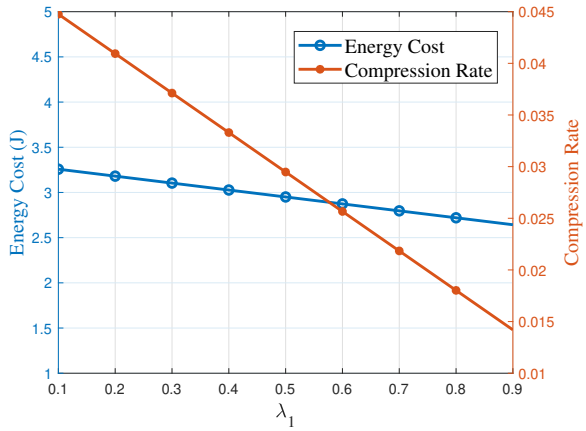


Fig. 14: Impact of λ_1 on energy cost and compression rate.

The simulation of system performance, considering varying weights of recovery quality, time delay, and energy cost, is illustrated in Fig. 14. In the objective function, recovery quality contributes positively, whereas time delay and energy cost have negative impacts. To study the impact of hyperparameters, we adjust the value of λ_1, and calculated the other two hyperparameters by $\lambda_2 = \lambda_3 = \frac{1-\lambda_1}{2}$. The results reveal that the compression rate notably decreases as λ_1 increases. This is because a higher compression rate can effectively reduce the time and energy costs associated with data transmission. It is important to note that the energy cost is influenced by both computational and communication factors. However, in our specific case study, the energy cost of communication is the predominant factor. Consequently, the energy cost displayed on the left y-axis also decreases as the compression rate decreases.

VI. Conclusion and Discussion

A. Summarizing the Paper

We designed a post-deployment fine-tunable semantic communication (FTSC) framework, which is the first semantic communication framework that can adjust the knowledge base to accommodate unknown datasets after the deployment of the encoder and decoder. The proposed FTSC is based on the VQ-VAE-2 model and supports two fine-tuning modes, i.e., DFT and LSDFT. The DFT is a fine-tuning scenario without gradient transmission from the receiver end to the sender end but still requires additional original image transmission. The LSDFT scenario further reduces communication overhead by intermediate loss calculation at the cost of acceptable degradation of performance. In the simulation, both two scenarios in FTSC have shown an advantage in image recovery quality over traditional methods on multiple datasets.

B. Considering Models beyond VQ-VAE-2

It is possible to extend FTSC to other neural networks with encoder-decoder structures, such as variational autoencoder (VAE) [62], [63], which is widely used for semantic communication.

To implement such an extension, we need the following two changes: (1) remove the operations on the embedding space because VAE does not have it; (2) deal with the data type difference on the latent space because the latent space of VQ-VAE-2 is saved with quantized integers while that of VAE is saved with float numbers.

To implement the extension in the proposed decoder fine-tuning (DFT) scenario, we need to freeze the VAE encoder at the sender side. The latent space and original images need to be transmitted to fine-tune the decoder at the receiver side. With a traditional encoder-decoder VAE structure, the proposed method can be extended smoothly without additional operations on embedding space. The extension in the proposed latent space-based decoder fine-tuning (LSDFT) is similar to that in DFT. The difference lies in the loss calculation at the receiver side because the shape of the latent space of VQ-VAE-2 and VAE are different.

To summarize, we can extend FTSC to other base neural network models with encoder-decoder structure [64]–[68] with proper modifications, but the performance and computational cost need to be verified through experiments.

C. Future Directions

In addition to Section VI-B discussed above, additional future research can be facilitated in the following two directions. Firstly, improvement in the encoder and decoder network design can improve the fundamental performance of the semantic communication system. Secondly, the errorless transmission of the latent space under noisy channels is expected to improve the recovery quality. Furthermore, the study on balancing the computational cost and the communication cost under more practical and complicated use cases is necessary. Thirdly, crucial fine-tuning of the embedding space results in non-negligible overhead. To overcome this issue, further study, including better training methods that accelerate convergence and balancing the size of embedding space and performance according to specific tasks, can be considered. Towards practical applications, more factors can be considered, e.g., imperfect channel estimation, more types of channel models, and multi-antenna users.

References

[1] P. Si, R. Liu, L. Qian, J. Zhao, and K.-Y. Lam, "Fine-tunable semantic communication for image transmission," in *IEEE 10th International Conference on Big Data Computing and Communications (BigCom)*, 2024.

[2] W. Jiang, B. Han, M. A. Habibi, and H. D. Schotten, "The road towards 6G: A comprehensive survey," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 334–366, 2021.

[3] Y. Lu and X. Zheng, "6G: A survey on technologies, scenarios, challenges, and the related issues," *Journal of Industrial Information Integration*, vol. 19, p. 100158, 2020.

[4] L. Qian, P. Yang, M. Xiao, O. A. Dobre, M. Di Renzo, J. Li, Z. Han, Q. Yi, and J. Zhao, "Distributed learning for wireless communications: Methods, applications and challenges," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 3, pp. 326–342, 2022.

[5] D. Wu, Z. Yang, P. Zhang, R. Wang, B. Yang, and X. Ma, "Virtual-reality inter-promotion technology for metaverse: A survey," *IEEE Internet of Things Journal*, vol. 10, no. 18, pp. 15 788–15 809, 2023.

[6] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE network*, vol. 34, no. 3, pp. 134–142, 2019.

[7] Z. Xu, D. Liu, W. Liang, W. Xu, H. Dai, Q. Xia, and P. Zhou, "Online learning algorithms for offloading augmented reality requests with uncertain demands in MECs," in *IEEE 41st International Conference on Distributed Computing Systems (ICDCS)*, 2021, pp. 1064–1074.

[8] W. Yang, H. Du, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. S. Shen, and C. Miao, "Semantic communications for future Internet: Fundamentals, applications, and challenges," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 213–250, 2022.

[9] G. Shi, Y. Xiao, Y. Li, and X. Xie, "From semantic communication to semantic-aware networking: Model, architecture, and open problems," *IEEE Communications Magazine*, vol. 59, no. 8, pp. 44–50, 2021.

[10] Z. Weng and Z. Qin, "Semantic communication systems for speech transmission," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2434–2444, 2021.

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

[12] R. Liu, J. Zhang, H. Li, J. Zhang, Y. Wang, and W. Zhou, "AFLOW: developing adversarial examples under extremely noise-limited settings," in *International Conference on Information and Communications Security (ICICS)*, vol. 14252. Springer, 2023, pp. 502–518.

[13] T. Yuan, L. Mi, W. Wang, H. Dai, and X. Fu, "AccDecoder: Accelerated decoding for neural-enhanced video analytics," in *IEEE Conference on Computer Communications (INFOCOM)*, 2023, pp. 1–10.

[14] P. K. Sangdeh and H. Zeng, "DeepMux: Deep-learning-based channel sounding and resource allocation for IEEE 802.11 ax," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2333–2346, 2021.

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

[16] H. Xie and Z. Qin, "A lite distributed semantic communication system for Internet of Things," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 142–153, 2020.

[17] Z. Weng and Z. Qin, "Semantic communication systems for speech transmission," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2434–2444, 2021.

[18] T. Han, Q. Yang, Z. Shi, S. He, and Z. Zhang, "Semantic-preserved communication system for highly efficient speech transmission," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 245–259, 2022.

[19] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 567–579, 2019.

[20] Q. Hu, G. Zhang, Z. Qin, Y. Cai, G. Yu, and G. Y. Li, "Robust semantic communications with masked VQ-VAE enabled codebook," *IEEE Transactions on Wireless Communications*, vol. 22, no. 12, pp. 8707–8722, 2023.

[21] B. Guler, A. Yener, and A. Swami, "The semantic communication game," in *IEEE International Conference on Communications (ICC)*, 2016, pp. 1–6.

[22] ——, "The semantic communication game," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 4, pp. 787–802, 2018.

[23] Z. Lyu, G. Zhu, J. Xu, B. Ai, and S. Cui, "Semantic communications for image recovery and classification via deep joint source and channel coding," *IEEE Transactions on Wireless Communications*, 2024.

[24] M. Zhang, Y. Li, Z. Zhang, G. Zhu, and C. Zhong, "Wireless image transmission with semantic and security awareness," *IEEE Wireless Communications Letters*, 2023.

[25] S. Seo, J. Park, S.-W. Ko, J. Choi, M. Bennis, and S.-L. Kim, "Toward semantic communication protocols: A probabilistic logic perspective," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 8, pp. 2670–2686, 2023.

[26] J. Choi, J. Park, S.-W. Ko, J. Choi, M. Bennis, and S.-L. Kim, "Semantics alignment via split learning for resilient multi-user semantic communication," *IEEE Transactions on Vehicular Technology*, 2024.

[27] A. Razavi, A. Van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with VQ-VAE-2," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[28] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[29] C. Wang, Y. Li, F. Gao, D. Deng, J. Xu, Y. Liu, and W. Wang, "Adaptive semantic-bit communication for extended reality interactions," *IEEE Journal of Selected Topics in Signal Processing*, 2023.

[30] Y. Zhang, L. Jiao, J. Yan, and X. Lin, "Dynamic service placement for virtual reality group gaming on mobile edge cloudlets," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 8, pp. 1881–1897, 2019.

[31] L. Wang, L. Jiao, T. He, J. Li, and M. Mühlhäuser, "Service entity placement for social virtual reality applications in edge computing," in *IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 2018, pp. 468–476.

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

[33] R. Carnap, Y. Bar-Hillel *et al.*, "An outline of a theory of semantic information," 1952.

[34] H. Tong, Z. Yang, S. Wang, Y. Hu, W. Saad, and C. Yin, "Federated learning based audio semantic communication over wireless networks," in *2021 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2021, pp. 1–6.

[35] D. Gündüz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 5–41, 2022.

[36] Z. Lyu, G. Zhu, J. Xu, B. Ai, and S. Cui, "Semantic communications for joint image recovery and classification," in *2023 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2023, pp. 1579–1584.

[37] H. Xie, Z. Qin, and G. Y. Li, "Task-oriented multi-user semantic communications for VQA," *IEEE Wireless Communications Letters*, vol. 11, no. 3, pp. 553–557, 2021.

[38] H. Bista, I.-L. Yen, F. Bastani, M. Mueller, and D. Moore, "Semantic-based information sharing in vehicular networks," in *2018 IEEE International Conference on Web Services (ICWS)*. IEEE, 2018, pp. 282–289.

[39] G. Zhang, Q. Hu, Z. Qin, Y. Cai, and G. Yu, "A unified multi-task semantic communication system with domain adaptation," in *GLOBECOM 2022-2022 IEEE Global Communications Conference*. IEEE, 2022, pp. 3971–3976.

[40] L. Xia, Y. Sun, X. Li, G. Feng, and M. A. Imran, "Wireless resource management in intelligent semantic communication networks," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2022, pp. 1–6.

[41] W. Zhang, Y. Wang, M. Chen, T. Luo, and D. Niyato, "Optimization of image transmission in semantic communication networks," in *GLOBECOM 2022-2022 IEEE Global Communications Conference*. IEEE, 2022, pp. 5965–5970.

[42] Z. Yang, M. Chen, Z. Zhang, and C. Huang, "Energy efficient semantic communication over wireless networks with rate splitting," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 5, pp. 1484–1495, 2023.

[43] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[44] J. Peng, D. Liu, S. Xu, and H. Li, "Generating diverse structure for image inpainting with hierarchical VQ-VAE," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 775–10 784.

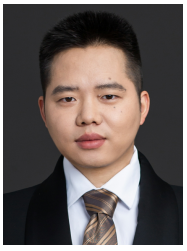
[45] W. Yan, Y. Zhang, P. Abbeel, and A. Srinivas, "VideoGPT: Video generation using VQ-VAE and transformers," *arXiv preprint arXiv:2104.10157*, 2021.

[46] Y. Hu, C. Luo, and Z. Chen, "Make it move: controllable image-to-video generation with text descriptions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 219–18 228.

[47] C. Gârbacea, A. van den Oord, Y. Li, F. S. Lim, A. Luebs, O. Vinyals, and T. C. Walters, "Low bit-rate speech coding with VQ-VAE and

a WaveNet decoder," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 735–739.
[48] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in Neural Information Processing Systems*, vol. 28, 2015.
[49] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations*, 2016.
[50] A. Vahdat and J. Kautz, "NVAE: A deep hierarchical variational autoencoder," *Advances in Neural Information Processing Systems*, vol. 33, pp. 19 667–19 679, 2020.
[51] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.
[52] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.
[53] Y. Yuan, L. Jiao, K. Zhu, X. Lin, and L. Zhang, "AI in 5G: The case of online distributed transfer learning over edge networks," in *IEEE Conference on Computer Communications (INFOCOM)*, 2022, pp. 810–819.
[54] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: A literature review," *BMC medical imaging*, vol. 22, no. 1, p. 69, 2022.
[55] Y. Guo, Y. Li, L. Wang, and T. Rosing, "AdaFilter: Adaptive filter fine-tuning for deep transfer learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 4060–4066.
[56] Z. Chai, C. Zhao, B. Huang, and H. Chen, "A deep probabilistic transfer learning framework for soft sensor modeling with missing data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 7598–7609, 2021.
[57] H. Akrami, A. A. Joshi, J. Li, S. Aydore, and R. M. Leahy, "Brain lesion detection using a robust variational autoencoder and transfer learning," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 786–790.
[58] O. Özdenizci, Y. Wang, T. Koike-Akino, and D. Erdoğmuş, "Transfer learning in brain-computer interfaces with adversarial variational autoencoders," in *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2019, pp. 207–210.
[59] M. Ye, J. Chen, F. Xiong, and Y. Qian, "Learning a deep structural subspace across hyperspectral scenes with cross-domain VAE," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.
[60] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.
[61] L. Qian and J. Zhao, "User association and resource allocation in large language model based mobile edge computing system over 6G wireless communications," in *2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring)*. IEEE, 2024, pp. 1–7.
[62] Y. Bo, Y. Duan, S. Shao, and M. Tao, "Joint coding-modulation for digital semantic communications via variational autoencoder," *IEEE Transactions on Communications*, 2024.
[63] J. Wang, G. Wang, X. Zhang, L. Liu, H. Zeng, L. Xiao, Z. Cao, L. Gu, and T. Li, "PATCH: A plug-in framework of non-blocking inference for distributed multimodal system," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 7, no. 3, pp. 1–24, 2023.
[64] E. Erdemir, T.-Y. Tung, P. L. Dragotti, and D. Gündüz, "Generative joint source-channel coding for semantic image transmission," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 8, pp. 2645–2657, 2023.
[65] D. B. Kurka and D. Gündüz, "DeepJSCC-f: Deep joint source-channel coding of images with feedback," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 178–193, 2020.
[66] D. Luan and J. S. Thompson, "Channelformer: Attention based neural solution for wireless channel estimation and effective online training," *IEEE Transactions on Wireless Communications*, vol. 22, no. 10, pp. 6562–6577, 2023.
[67] L. Wang, X. Wang, D. Zhang, X. Ma, Y. Zhang, H. Dai, C. Xu, Z. Li, and T. Gu, "Knowing your heart condition anytime: User-independent ECG measurement using commercial mobile phones," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 7, no. 3, pp. 1–28, 2023.
[68] Z. Yang, Y. Zhang, K. Qian, and C. Wu, "SLNet: A spectrogram learning neural network for deep wireless sensing," in *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2023, pp. 1221–1236.

Peiyuan Si (Graduate Student Member, IEEE) received bachelor's and master's degrees in communication engineering from Zhejiang University of Technology of China, Zhejiang, China, in 2018 and 2021, respectively. He is currently working toward his Ph.D. at the College of Computing and Data Science, Nanyang Technological University, Singapore.



Renyang Liu (Graduate Student Member, IEEE) received his B.E. degree in Computer Science from Northwest Normal University in 2017 and his Ph.D. in Information Science and Engineering from Yunnan University in 2024. He is currently a research fellow at the Institute of Data Science, National University of Singapore.

Liangxin Qian (Graduate Student Member, IEEE) received bachelor's and master's degrees in communication engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2019 and 2022, respectively. He is currently working toward his Ph.D. at the College of Computing and Data Science, Nanyang Technological University, Singapore.

Jun Zhao (Member, IEEE) (S'10-M'15) is currently an Assistant Professor in the College of Computing and Data Science (CCDS) at Nanyang Technological University (NTU) in Singapore. He received a PhD degree in May 2015 in Electrical and Computer Engineering from Carnegie Mellon University (CMU), and a bachelor's degree in July 2010 from Shanghai Jiao Tong University in China.

Kwok-Yan Lam (Member, IEEE) received the B.Sc. degree from the University of London, London, U.K., in 1987 and the Ph.D. degree from the University of Cambridge, Cambridge, U.K., in 1990. He is the Associate Vice President (Strategy and Partnerships) and Professor in the College of Computing and Data Science at the Nanyang Technological University (NTU), Singapore.