Codebook-enabled Generative End-to-end Semantic Communication Powered by Transformer

Peigen Ye^{1, 2}, Yaping Sun^{2, 3}, Shumin Yao², Hao Chen², Xiaodong Xu^{4, 2}, Shuguang Cui^{3, 2}, Fellow, IEEE

¹The Shenzhen Campus of Sun Yat-sen University, Sun Yat-sen University, Shenzhen, China

²Peng Cheng Laboratory, Shenzhen, China

³FNii and SSE, The Chinese University of Hong Kong, Shenzhen, China

⁴Beijing University of Posts and Telecommunications, Beijing, China

Abstract—Codebook-based generative semantic communication attracts increasing attention, since only indices are required to be transmitted when the codebook is shared between transmitter and receiver. However, due to the fact that the semantic relations among code vectors are not necessarily related to the distance of the corresponding code indices, the performance of the codebook-enabled semantic communication system is susceptible to the channel noise. Thus, how to improve the system robustness against the noise requires careful design. This paper proposes a robust codebook-assisted image semantic communication system, where semantic codec and codebook are first jointly constructed, and then vector-to-index transformer is designed guided by the codebook to eliminate the effects of channel noise, and achieve image generation. Thanks to the assistance of the highquality codebook to the Transformer, the generated images at the receiver outperform those of the compared methods in terms of visual perception. In the end, numerical results and generated images demonstrate the advantages of the generative semantic communication method over JPEG+LDPC and traditional joint source channel coding (JSCC) methods.

Index Terms—Codebook, Generative Semantic Communication, Transformer, Semantic Knowledge Base

I. INTRODUCTION

Semantic communications have recently been recognized as a promising technology for beyond 5G (B5G) and 6G wireless networks, whereby transmitters are designed to efficiently convey semantic information to receivers, rather than reliably transmit syntactic information as in conventional wireless communication systems. Via the end-to-end (E2E) semantic information transmission design, the semantic communications are able to efficiently compress messages while preserving the essential meaning by filtering out the irrelevant information, thus significantly enhancing the communication efficiency. Semantic communications are envisioned to have abundant applications such as network intelligence and industrial automation

Artificial intelligence technology significantly contributes to the progress of communication technology [1] [2] [3]. Semantic communication relies primarily on artificial intelligence techniques to construct a series of task-specific or general-purpose semantic knowledge bases. Semantic knowledge base (SKB), as a key enabler of semantic communication, facilitates semantic encoding/decoding via providing semantic knowledge vectors, and refining a compact search space [4]. For instance, Sun et al. [5] proposed a novel SKB-enabled

multi-level feature transmission framework that significantly improves the performance of remote zero-shot transmission. Li et al. [6] introduced a medical semantic communication system, which leverages a domain knowledge with retinal fundus segmentation labels to enhance image reconstruction accuracy at the receiver.

Regarding the construction of SKBs, there is a current trend towards discrete quantized codebooks. For example, Hu et al. [7] proposed a semantic communication method with masked vector quantized-variational autoencoder (VQ-VAE) enabled codebook. This method utilizes a discrete codebook to share encoded feature representations between the transmitter and the receiver. Park et al. [8] proposed a novel DeepSC framework with federated codebook to serve multi-user scenarios [9]. The codebook of discrete coding serves as apriori knowledge, endowing the encoder with a more stable mapping target range during the encoding process. Simultaneously, it reduces the bias in information comprehension between the encoder and decoder. This not only enhances the stability of feature encoding by the encoder, but also significantly improves the transmission efficiency.

However, when indices are transmitted over wireless channel, a minor error of decoded indices at the receiver could potentially lead to catastrophic consequences for the reconstruction of the original information. This is because each index generally corresponds to a substantial amount of semantic information, and the semantic relations among code vectors are not necessarily related to the distance of the corresponding code indices. As an alternative approach, transmitting the feature map at the transmitter and subsequently performing codebook-based feature matching at the receiver could be a viable option. However, the transmission of vector features is also susceptible to the noise in the physical communication channel. Although the impact of feature contamination by noise may be relatively smaller compared to the tampering of indices, it remains a factor that requires careful consideration.

In recent years, the development of Transformer [10] has been widely recognized, particularly with the emergence of large models such as ChatGPT, which has propelled artificial intelligence to new heights. The self-attention mechanism employed by Transformer allows each position in the input sequence to dynamically attend to all other positions, enabling the model to directly gather information from the entire input

sequence. This emphasis on global information over local details endows Transformer with outstanding performance in various tasks, especially in situations that require consideration of long-range dependencies. In [11], Zhou et al. addressed the matching problem between degraded image feature vector and high-quality image feature vector by introducing the Transformer. Inspired by their work [11] [12], this paper incorporates the Transformer to tackle the challenge of recovering image feature map, as transmitted information, after being contaminated by noise. The goal is to restore these features to the state of indices suitable for image reconstruction.

Main contributions of this paper lie in the following aspects.

- We build a codebook-assisted generative semantic encoding and decoding architecture. And in the receiver, a transformer with nine self-attention blocks is used to realize the mapping of feature map with errors to high-quality codebook indices based on the global information of the feature map and the assistance of codebook, thereby eliminating noise contamination.
- The proposed method adopts a two-stage training mechanism. In Stage 1, the encoder, decoder, and codebook are obtained through E2E training. In Stage 2, the vector-to-index Transformer is fully trained by fixing the upstream and downstream parameters.
- Simulation results and generated images demonstrate
 the advantages of generative semantic communication
 methods over JPEG+LDPC and traditional joint source
 channel coding (JSCC) methods. Under low signal-tonoise ratio (SNR), the proposed method can still maintain good performance on the Learned Perceptual Image
 Patch Similarity (LPIPS) metric and perform high-quality
 image transmission.

II. SYSTEM MODEL

In this section, we present the framework of codebookenabled generative E2E semantic communication system. As illustrated in Fig. 1, the input image is first encoded into a feature map by the semantic encoder at the transmitter, and then the feature map is transmitted through the wireless channel. Subsequently, the received feature map is first mapped into the corresponding indices by the vector-to-index Transformer (V2IT), and then decoded by the semantic decoder to generate a high-quality image with similar visual perception. The codebook $\mathcal C$ consists of a set of image features c_k , expressed as $\mathcal C = \left\{c_k \in \mathbb R^d\right\}_{k=0}^L$.

A. Semantic Encoder at Transmitter

As shown in Fig. 1, the transmitter is endowed with an encoder with a Convolutional Neural Network (CNN) structure, designed to extract high-dimensional spatial features from the input image. Initially, the input image $I_h \in \mathbb{R}^{H \times W \times 3}$ is fed into the network and normalized. Subsequently, within the encoder, after undergoing operations such as convolution, the image is encoded into a feature vector $\mathbf{z}_h \in \mathbb{R}^{m \times n \times q}$. This process can be represented as follows:



Fig. 1: The framework of the proposed semantic communication system.

$$\mathbf{z}_{h} = \mathbf{E}_{\theta} \left(I_{h} \right), \tag{1}$$

where $\mathbf{E}_{\theta}(\cdot)$ represents the entire semantic encoding function with respect to parameters θ .

Due to the first stage of training, where the encoder contributes to the generation of the codebook through end-to-end training, the encoded feature map \mathbf{z}_h is highly similar to the vector in the codebook. \mathbf{z}_h is composed of $m \times n$ q-dimensional feature vector $z_h^{(i,j)}$. Its overall size is $N=m\times n\times q$, which is closely related to the amount of transmitted information. In fact, the semantic encoder can be considered a generalized SKB, as it learns the knowledge of encoding the image to vectors that approximate the vectors in the codebook. The performance of the encoder also affects the recovery of the feature map in the subsequent processes.

B. Physical Communication Channel

Assuming a single communication link for image semantic communication transmission, to simulate the process of physical information transmission, we have employed the Additive White Gaussian Noise (AWGN) channel model widely used in wireless communication as the physical communication channel. \mathbf{z}_h undergoes transmission through the physical channel to reach the receiver. The signal received at the receiver is denoted as $\hat{\mathbf{z}}_h$, and this process can be represented as follows:

$$\hat{\mathbf{z}}_h = \mathbf{z}_h + \mathbf{n},\tag{2}$$

where the AWGN n has i.i.d elements with zero mean and variance σ^2 .

C. Vector-to-index Transformer, Codebook, and Semantic Decoder at Receiver

The corrupted output $\hat{\mathbf{z}}_h$ from the wireless communication physical channel is acquired by the receiver. Initially, in the receiver, vector-to-index Transformer predicts and corrects the feature map based on its global information, resulting in the semantic feature indices $\hat{\mathbf{s}}$. This process can be expressed as follows:

$$\hat{\mathbf{s}} = \mathbf{T}_{\tau} \left(\hat{\mathbf{z}}_h \right), \tag{3}$$

where $\mathbf{T}_{\tau}\left(\cdot\right)$ indicates the function of vector-to-index Transformer with respect to parameters τ .

Next, the system can retrieve the corresponding feature map from the codebook based on the feature indices $\hat{\mathbf{s}} = \left\{\hat{s}^{(i,j)}\right\}^{m \cdot n}$. This retrieval process reconstructs the complete feature map $\hat{\mathbf{z}}_c \in \mathbb{R}^{m \times n \times q}$. This process can be expressed as:

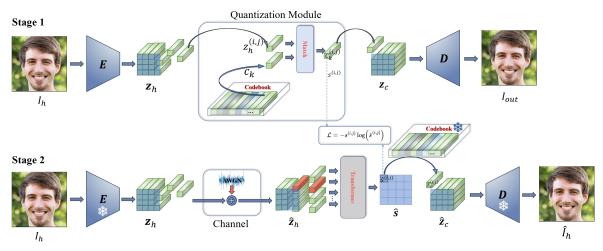


Fig. 2: The network architecture of the proposed semantic communication system, a two-stage trained neural network model. In Stage 1, a high-quality codebook, enriched with semantic details from trained images, is constructed through learning. Simultaneously, an encoder and a decoder are obtained. In Stage 2, the Transformer, as the vector-to-index Transformer, is trained by fixing its upstream and downstream parameters. The corrected feature map are utilized for image reconstruction.

$$\hat{z}_c^{(i,j)} = \mathbf{S}\left(\hat{s}^{(i,j)}\right),\tag{4}$$

where $\mathbf{S}\left(\cdot\right)$ denotes the method of retrieval (Look-up Table), $\hat{s}^{(i,j)} \in \{0,\cdots,L-1\}$, and L represents the size of the codebook.

Subsequently, the feature map $\hat{\mathbf{s}}_c$ is fed into the decoder for image reconstruction, resulting in the output \hat{I}_h . This process can be expressed as:

$$\hat{I}_h = \mathbf{D}_{\mathcal{E}} \left(\hat{\mathbf{z}}_c \right), \tag{5}$$

where $\mathbf{D}_{\xi}(\cdot)$ represents the decoder function with respect to parameters ξ , which is an image generation module.

III. THE PROPOSED METHOD

The image feature vector is contaminated by noise when it passes through the channel, causing confusion in image reconstruction. Therefore, the core objective of our work is to recover and correct the contaminated feature vector, mapping it to a vector index from the learned codebook that is more semantically related to the original feature vector.

This section introduces the specific training method of the proposed semantic communication system. The training process of the proposed method consists of two stages. In Stage 1, through end-to-end training on image dataset, it generates a codebook and an encoder-decoder pair. In Stage 2, it addresses the challenge of how the image feature map, contaminated by channel noise, can be corrected as accurately as possible by its global information and the codebook.

A. Stage 1: Joint Construction of Semantic Codec and Codebook

To obtain the codebook and the encoder-decoder pair, a vector-quantized autoencoder network [12] is constructed. As illustrated in Fig. 2, in Stage 1, a high-quality input image

 $I_h \in \mathbb{R}^{H \times W \times 3}$ is fed into the encoder \mathbf{E} , resulting in the encoded feature map $\mathbf{z}_h \in \mathbb{R}^{m \times n \times q}$. Subsequently, it is passed through the quantization module. In the quantization module, each vector $z_h^{(i,j)}$ in \mathbf{z}_h is matched to the quantized vector c_k from the codebook $\mathcal C$ using nearest-neighbor matching. Subsequently, the quantization module places the matched vector $c_k^{(i,j)}$ at the corresponding position (i,j), combines them to obtain the map \mathbf{z}_c for reconstruction, and simultaneously outputs the index $s^{(i,j)}$ ($\mathbf{s} = \left\{s^{(i,j)}\right\}^{m \cdot n}$) corresponding to $c_k^{(i,j)}$. Finally, the \mathbf{z}_c is input into the decoder (image generation module) for image reconstruction, resulting in the reconstructed image I_{out} .

Training Objectives. For the E2E network in Stage 1, based on the network's input and output, we can establish a result-based loss function \mathcal{L}_{result} :

$$\mathcal{L}_{result} = \left\| I_h - I_{out} \right\|_1 + \left\| \phi \left(I_h \right) - \phi \left(I_{out} \right) \right\|_2^2, \quad (6)$$

where $\|I_h - I_{out}\|_1$ represents the L1 loss, which measures the mean absolute difference between the output I_{out} and the ground truth I_h . $\|\phi(I_h) - \phi(I_{out})\|_2^2$ denotes the perceptual loss, reflecting the differences between the output and the ground truth in a high-dimensional space, where $\phi(\cdot)$ represents the deep feature extractor, utilizing the VGG19 [13].

Due to the insufficient constraints of the above loss functions on the quantization module, the loss function \mathcal{L}_{in} reflecting the intermediate layers of the network is adopted:

$$\mathcal{L}_{in} = \left\| \mathbf{sg}\left(\mathbf{z}_{h}\right) - \mathbf{z}_{c} \right\|_{2}^{2} + \beta \left\| \mathbf{z}_{h} - \mathbf{sg}\left(\mathbf{z}_{c}\right) \right\|_{2}^{2}, \tag{7}$$

where $sg(\cdot)$ represents the stop-gradient operator, and β is the weight parameter for the update rates of the encoder and the codebook.

As an adversarial model is employed as the generator module, the overall loss function needs to include an adversarial loss function \mathcal{L}_{GAN} :

$$\mathcal{L}_{Stage1} = \mathcal{L}_{result} + \mathcal{L}_{in} + \mathcal{L}_{GAN}, \tag{8}$$

B. Stage 2: Vector-to-index Transformer

In Stage 2, our goal is to obtain the vector-to-index Transformer, which can utilize the codebook and the global information of the contaminated feature map to correct feature map, to obtain vector indices from the learned codebook that are most semantically related to the original image features. The network architecture of Stage 2 was introduced in the Section II-C.

Training Objectives. The input of Transformer is the contaminated image feature map, and the output is the indices from the codebook. Based on the training results of Stage 1, we can formulate the loss function for Stage 2 as follows:

$$\mathcal{L}_{Stage2} = \lambda \sum_{i=0}^{mn-1} \left(-s^{(i,j)} \log \left(\hat{s}^{(i,j)} \right) \right) + \|\hat{\mathbf{z}}_h - \mathbf{sg} \left(\mathbf{z}_c \right) \|_2^2,$$
(9)

where $s^{(i,j)}$ and \mathbf{z}_c are both derived from Stage 1. They are considered as the ground truths for the corresponding variables in the Stage 2 network. λ represents a weight parameter. During the training process of Stage 2, the parameters of the codebook, encoder and decoder are kept unchanged, and only the vector-to-index Transformer is trained. Towards the end of the training, all parameters can be unfrozen to fine-tune the network.

IV. EXPERIMENT AND EVALUATION

A. Datasets and Settings

The training dataset: it consists of 68,659 high-quality images sourced from the FFHQ dataset [14]. The native resolution of images in the FFHQ dataset is 1024×1024 . During dataset preparation, the images are preprocessed, and their sizes are resized to 512×512 .

Two test datasets: the FFHQ-test comprises 1,094 high-quality images from the FFHQ dataset (non-overlapping with the training dataset), and the CelebA-test includes 304 high-quality images from the CelebA dataset [15]. The sizes of all images in both datasets are adjusted to 512×512 .

Settings: the size of the Codebook is set to 1024 (i.e., L=1024). The length of vector in the codebook is set to 256 (i.e., q=256). For all training stages, we employ the Adam optimizer with a batch size of 4. The learning rates for Stage I and Stage II are set to 7×10^{-5} and 1×10^{-4} , respectively. The training iterations for Stage 1 and Stage 2 are set to 415×10^3 and 200×10^3 , respectively. The proposed method is implemented based on the PyTorch framework and trained using one GeForce RTX 3090 GPU.

B. Simulation and Evaluation

In this section, the proposed method will be validated and compared with two other excellent methods: JPEG+LDPC and JSCC.

JPEG+LDPC. JPEG, established by the Joint Photographic Experts Group, is a widely used traditional image coding standard. LDPC, a powerful error correction coding method, is commonly employed in channel coding for digital communication systems. Communication systems combining JPEG image coding and LDPC channel coding are generally considered to have high reliability in image transmission, especially in communication environments affected by noise. In the following experiments, the bit rate is set as 1/2.

JSCC. JSCC [16] is a classical joint source-channel coding technique in the field of semantic communication, which parameterizes the encoder and decoder functions by two CNNs and trains them jointly.

About the experimental evaluation, we not only calculate traditional image quality assessment metrics such as PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity), but also include the LPIPS (Learned Perceptual Image Patch Similarity). Due to employing a pre-trained deep feature extractor, LPIPS is more agreeable with human perceptual judgments [6].

Fig. 3 illustrates the variations in the PSNR, SSIM, and LPIPS metrics for the three image transmission methods under different SNR on FFHQ-test and CelebA-test datasets. Fig. 4 depicts the actual input and output images corresponding to the respective situations.

From each subplot in Fig. 3, it is evident that the JPEG+LDPC method, representing traditional communication approaches, exhibits a drastic performance decline in image transmission under low SNR environments. This decline is observed across all metrics (where LPIPS ranges from 0 to 1, with smaller values indicating higher perceptual similarity between two images). In Fig. 4, it is more visually apparent that JPEG+LDPC suffers severe information loss in low SNR scenarios, to the extent that image decoding becomes practically impossible. The gray images in Fig. 4 are contingent on our a specific operation. In simulation experiments, we specifically provided lossless data associated with the JPEG decoding protocol to JPEG+LDPC at the receiver. Without this operation, even the gray images would not be obtainable.

From Fig. 3a, Fig. 3b, and Fig. 3e, it can be observed that when SNR is high (e.g., SNR > 12dB), our proposed method exhibits slightly lower PSNR and SSIM compared to the JSCC method. However, our method consistently outperforms the comparison method in terms of LPIPS. Moreover, as depicted output results in Fig. 4, the obtained images using our method visually outperform those of the comparison methods significantly. This is because PSNR and SSIM are metrics based on the pixel level, evaluating images in a low-dimensional space. Such comparisons are not very suitable for semantic communication tasks, especially generative semantic communication. This is because the characteristic of generative communication network, i.e., the unavoidable randomness

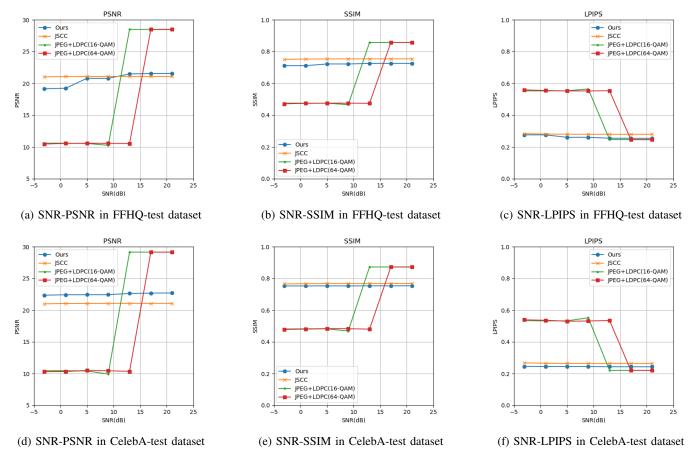


Fig. 3: The performance metrics, including PSNR, SSIM, and LPIPS, for three methods across the SNR range of -3 to 21. The evaluation is conducted on the FFHQ-test dataset and CelebA-test dataset.

within a certain range, makes it naturally less advantageous for pixel-wise or low-dimensional structural comparisons like PSNR and SSIM. However, for the majority of semantic communication tasks, the focus is on transmitting crucial semantic information required for the task, rather than pursuing absolute accuracy for each pixel. Hence, metrics resembling per-pixel comparisons may not hold significant importance in semantic communication.

V. CONCLUSION

This paper proposes a robust codebook-assisted image semantic communication method, via transmitting the feature map directly to the receiver, and leveraging Transformer to predict and correct feature map with errors based on the global information of the feature map and the assistance of codebook, achieving excellent source image reconstruction. The use of a codebook significantly enhances the performance of vector-to-index Transformer, thereby improving communication stability and efficiency. Furthermore, images generated based on high-quality map from the learned codebook surpass those of comparative methods in terms of visual aesthetics. Additionally, we conduct a thorough analysis of the characteristics of generative semantic communication in image transmission, elucidating

the superiority of generative semantic communication methods. This work further substantiates the promising prospects and value of generative network structures in the field of semantic communication.

REFERENCES

- [1] Xuemin Shen, Jie Gao, Wen Wu, Mushu Li, Conghao Zhou, and Weihua Zhuang. Holistic network virtualization and pervasive network intelligence for 6g. *IEEE Communications Surveys & Tutorials*, 24(1):1–30, 2022.
- [2] Wen Wu, Mushu Li, Kaige Qu, Conghao Zhou, Xuemin Shen, Weihua Zhuang, Xu Li, and Weisen Shi. Split learning over wireless networks: Parallel design and resource management. *IEEE Journal on Selected Areas in Communications*, 41(4):1051–1066, 2023.
- [3] Jianhao Huang, Dongxu Li, Chuan Huang, Xiaoqi Qin, and Wei Zhang. Joint task and data oriented semantic communications: A deep separate source-channel coding scheme. *IEEE Internet of Things Journal*, pages 1–1, 2023.
- [4] Jinke Ren, Zezhong Zhang, Jie Xu, Guanying Chen, Yaping Sun, Ping Zhang, and Shuguang Cui. Knowledge base enabled semantic communication: A generative perspective. arXiv preprint arXiv:2311.12443, 2023.
- [5] Yaping Sun, Hao Chen, Xiaodong Xu, Ping Zhang, and Shuguang Cui. Semantic knowledge base-enabled zero-shot multi-level feature transmission optimization. *IEEE Transactions on Wireless Communications*, 2023.

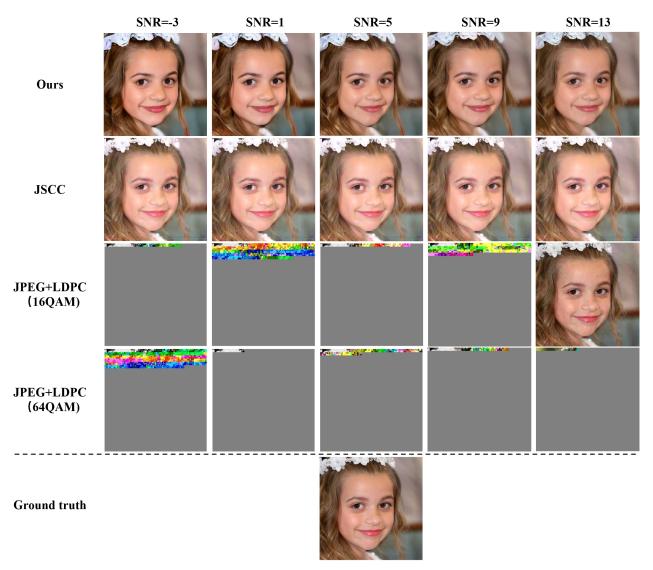


Fig. 4: The output results of image transmission for three methods under different SNR values (-3, 1, 5, 9, 13). The images represent the output of image transmission for each method. The gray images in the JPEG+LDPC method column represent cases where information loss is severe, making it challenging to decode normal images.

- [6] Aini Li, Xiaohong Liu, Guangyu Wang, and Ping Zhang. Domain knowledge driven semantic communication for image transmission over wireless channels. *IEEE Wireless Communications Letters*, 12(1):55–59, 2022
- [7] Qiyu Hu, Guangyi Zhang, Zhijin Qin, Yunlong Cai, Guanding Yu, and Geoffrey Ye Li. Robust semantic communications with masked vq-vae enabled codebook. *IEEE Transactions on Wireless Communications*, 2023
- [8] Chae-Hoon Park, Jinhyuk Choi, Jihong Park, and Seong-Lyun Kim. Federated codebook for multi - user deep source coding. In 2022 13th International Conference on Information and Communication Technology Convergence (ICTC), pages 994–996, 2022.
- [9] Dan Wang, Ran Li, Chuan Huang, Xiaodong Xu, and Hao Chen. User association and power allocation for user-centric smart-duplex networks via tree-structured deep reinforcement learning. *IEEE Internet of Things Journal*, 10(22):20216–20229, June 2023.
- [10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.
- [11] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy.

- Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35:30599–30611, 2022.
- [12] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF* conference on computer vision and pattern recognition, pages 12873– 12883, 2021.
- [13] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
- [14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.
- [15] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196, 2017.
- [16] Eirina Bourtsoulatze, David Burth Kurka, and Deniz Gündüz. Deep joint source-channel coding for wireless image transmission. *IEEE Transactions on Cognitive Communications and Networking*, 5(3):567–579, 2019.