CDDM: Channel Denoising Diffusion Models for Wireless Semantic Communications

Tong Wu, Zhiyong Chen, *Senior Member, IEEE*, Dazhi He, *Member, IEEE*, Liang Qian, Yin Xu, *Member, IEEE*, Meixia Tao, *Fellow, IEEE*, and Wenjun Zhang, *Fellow, IEEE*

***Abstract*— Diffusion models (DM) can gradually learn to remove noise, which have been widely used in artificial intelligence generated content (AIGC) in recent years. The property of DM for eliminating noise leads us to wonder whether DM can be applied to wireless communications to help the receiver mitigate the channel noise. To address this, we propose channel denoising diffusion models (CDDM) for semantic communications over wireless channels in this paper. CDDM can be applied as a new physical layer module after the channel equalization to learn the distribution of the channel input signal, and then utilizes this learned knowledge to remove the channel noise. We derive corresponding training and sampling algorithms of CDDM according to the forward diffusion process specially designed to adapt the channel models. We also theoretically prove that the well-trained CDDM can effectively reduce the conditional entropy of the received signal under small sampling steps. Moreover, we apply CDDM to a semantic communications system based on joint source-channel coding (JSCC) for image transmission and design a three-stage training algorithm for combining them. Extensive experimental results demonstrate that CDDM can further reduce the mean square error (MSE) after minimum mean square error (MMSE) equalizer, and the joint CDDM and JSCC system achieves better performance than the JSCC system, the traditional JPEG2000 with low-density parity-check (LDPC) code approach and other benchmarks in diverse scenarios.**

***Index Terms*— Diffusion models, wireless image transmission, semantic communications, joint source-channel coding.**

I. Introduction

DIFFUSION models (DM) [2], [3], [4] have recently achieved unprecedented success in artificial intelligence generated content (AIGC) [5], including multimodal image generation and edition [6], [7], text, and video generation [8], [9]. DM is a class of latent variable models inspired by non-equilibrium thermodynamics. It directly models the score function of the likelihood function through variational lower bounds, resulting in advanced generative performance. Compared to previous generative models such as variational auto-encoders (VAE) [10], generative adversarial networks (GAN) [11], and normalization flows (NF) [12], DM can learn fine-grained knowledge of the distribution, allowing it to generate contents with rich details. Additionally, DM is capable of generating more diverse images and has been shown to be resistant to mode collapse. The emergence of implicit classifiers endows DM with flexibility controllability, enhanced efficiency and ensuring faithful generation in conditional generation tasks.

More specifically, DM gradually adds Gaussian noise to the training data in the forward diffusion process until the data becomes pure noise. Then, in the reverse sampling process, it learns to recover the data from the noise, as shown in Fig. 1. Generally, given a data distribution $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, the forward diffusion process generates the t-th sample of $\mathbf{x}_t$ by sampling a Gaussian vector $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ as following

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{1}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ and $\alpha_i \in (0, 1)$ are hyperparameters.

In wireless communications, it is well known that the received signal y is a noisy and distorted version of the transmitted signal x, e.g., we have the following for the additive white Gaussian noise (AWGN) channel

$$y = x + n, \tag{2}$$

where n is white Gaussian noise.

Interestingly, by comparing (1) and (2), we can find that the designed process of DM and the wireless communications system are similar. DM progressively learns to effectively remove noise, thereby generating data that closely resembles the original distribution, while the receiver in the wireless communications system aims to recover the transmitted signal from the received signal. Clearly, **can DM be applied to the wireless communications system to help the receiver remove noise?**

Some previous works [13], [14], [15] have designed effective methods for denoising with advanced machine learning techniques such as convolutional neural networks (CNN) and GAN. Reference [13] proposes DnCNN for image denoising, which introduces residual learning into the image denoising task and achieves remarkable performance under AWGN

Manuscript received 25 August 2023; revised 23 January 2024 and 13 March 2024; accepted 13 March 2024. Date of publication 28 March 2024; date of current version 12 September 2024. This work was supported in part by the National Key Research and Development Project of China under Grant 2023YFB2906200, in part by the Fundamental Research Funds for the Central Universities, and in part by NSF of China under Grant 62222111 and Grant 62125108. An earlier version of this paper was presented in part at the 2023 IEEE Global Communications Conference (GLOBECOM) [DOI: 10.1109/GLOBECOM54140.2023.10436728. The associate editor coordinating the review of this article and approving it for publication was X. Chen. *(Corresponding author: Zhiyong Chen.)*

Tong Wu, Zhiyong Chen, Dazhi He, and Yin Xu are with the Cooperative Medianet Innovation Center (CMIC), Shanghai Jiao Tong University, Shanghai 200240, China, and also with Shanghai Key Laboratory of Digital Media Processing and Transmission, Shanghai 200240, China (e-mail: wu_tong@sjtu.edu.cn; zhiyongchen@sjtu.edu.cn; hedazhi@sjtu.edu.cn; xuyin@sjtu.edu.cn).

Liang Qian, Meixia Tao, and Wenjun Zhang are with the Cooperative Medianet Innovation Center (CMIC) and the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China, and also with Shanghai Key Laboratory of Digital Media Processing and Transmission, Shanghai 200240, China (e-mail: lqian@sjtu.edu.cn; mxtao@sjtu.edu.cn; zhangwenjun@sjtu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2024.3379244.

Digital Object Identifier 10.1109/TWC.2024.3379244

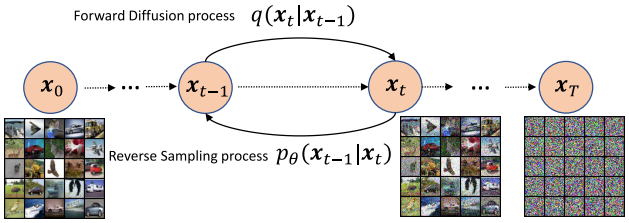


Fig. 1. The forward diffusion process with transition kernel $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ and the reverse sampling process with learnable transition kernel $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ of diffusion model in [3].

conditions. Reference [14] utilizes the residual neural network (ResNet) to establish a GAN to generate pure images based on corrupted images. The GAN is trained with a comprehensive loss function comprising of pixel loss, feature loss, smooth loss and adversarial loss. Reference [15] improves GAN performance with advanced Wasserstein GAN and achieves impressive results in computed tomography (CT) image denoising. Despite the extensive research on image denoising methods using advanced neural networks, they are limited in removing noise in image pixel space. To the best of our knowledge, there is no existing work on AI concentrating on removing noise in signal space. The effectiveness of adapting image pixel space denoising techniques for application in signal space, with the aim of enhancing the performance of the wireless communications, remains to be validated.

In contrast to the extensive research on DM in AIGC, there has been little work on DM in communications so far. Reference [16] investigates the possibility of utilizing conditional diffusion models to directly generate solutions for optimization problems, conditioned on the states, constraints and skills. Experiments show that the diffusion model-based policies outperform existing offline reinforcement learning (RL) approaches. Reference [17] designs MADiff for multi-agent scenario, a diffusion-based multi-agent offline RL framework which behaves as both a decentralized policy and a centralized controller for good coordination among agents. For incentive design, [18] proposes to motivate users to share information to facilitate computing resources of all users and designs the incentive objective based on contract theory. Diffusion models are utilized to generate the incentive mechanisms adaptively to the various characteristics of the sender and wireless channels. However, these works are limited in providing more suitable policies for decision-making and have no concentration on the performance of wireless communications.

There are only few works applying DMs into wireless communications. In [19], DM is employed to generate the wireless channel for an end-to-end communication system, achieving almost the same performance as the channel-aware case. In [20], DM with an adapted diffusion process is proposed for the decoding of algebraic block codes. Additionally, for semantic communications, [21] applies DM as the semantic decoder to generate the image under the condition of the transmitted semantic segment labels of the original image, achieving excellent mean intersection over union (mIoU) and learned perceptual image patch similarity (LPIPS) performance. Reference [22] develops a diffusion model with INN as a decoder to achieve better reconstruction quality. The INN is a neural network trained for simulating the degradation from coding and transmission. After training, the INN decomposes the image generated by the intermediate process of the diffusion model into its degraded version and the lost details. Therefore, the image reconstructed by the decoder and the lost details obtained from decomposition can be combined together to generate an implicit image to guide the next generation step of the diffusion model. In [23], a joint source-channel coding (JSCC)-based communications system is designed, which utilizes diffusion models as a decoder in the receiver. The system utilizes JSCC to transmit the low-frequency features of images, while diffusion models are responsible for generating the high-frequency details of the images, compensating for the high-frequency distortion. In [24], a redesign of the reverse sampling process based on the range-null space decomposition of the wireless channel to generate audio latent embeddings from the corrupted embeddings. However, these works employ conditional diffusion models as a decoder to generate the desired images conditioned on the received signal, without considering diffusion models as a module for denoising and purifying the received signal.

Motivated by this, in this paper, we design channel denoising diffusion models (CDDM) for communication systems to remove channel noise and purify the received signals. The proposed CDDM is conditioned on the received signal and channel estimation results to eliminate channel noise. In contrast to conventional generative models that only generate data adhering to the original data distribution, CDDM directly generates data that closely resembles the transmitted signal $\mathbf{x}$, consequently enhancing the performance of the communication systems. By employing carefully designed forward diffusion and reverse sampling processes based on an explicit conditional probabilistic model of the received signal, CDDM can adapt to diverse complex channel conditions, such as AWGN channel and Rayleigh fading channel with different signal-to-noise ratios (SNR). To leverage the received signal, we start the reverse sampling process from the received signal rather than pure noise, greatly reducing the number of reverse sampling steps and thus accelerating the process.

On the other hand, semantic communications [25], [26] have emerged as a novel paradigm that facilitates the seamless integration of information and communication technology with artificial intelligence (AI), which have been recognized as a highly promising solution for the sixth-generation (6G) wireless networks [27]. Semantic communications emphasize the transmission of valuable semantic information rather than bits, thereby guaranteeing improved transmission efficiency and reliability. One fundamental concept behind semantic communications is to bridge the source and channel components of Shannon theory [28], thereby enhancing the overall performance of end-to-end transmission. The paradigm focusing on the integrated design of source and channel coding processing is known as JSCC, which is a classical subject in the coding theory and information theory [29], [30], [31]. However, traditional JSCC techniques are predominantly rooted in complex and explicit probabilistic models, heavily relying on expert manual designs which often face challenges when dealing with complex sources. Moreover, these JSCC techniques overlook semantic aspects and lack optimization for specific tasks or human visual perception.

Many previous studies investigate deep-learning based JSCC techniques for semantic communications [32], [33], [34], [35], [36], [37]. Most studies concentrate on designing specific frameworks for different data models and have achieved better performance compared with traditional wireless transmission schemes. For wireless image transmission, [33] proposes a novel JSCC method based on attention mechanisms, which can automatically adapt to various channel conditions. In [35], an entropy model is proposed to achieve adaptive rate control for deep learning based JSCC architecture for semantic communications. In [36], the swin transformer [38] is integrated into the deep JSCC framework to improve the performance of wireless image transmission. Reference [37] develops a joint coding-modulation method and achieves end-to-end digital semantic communication system for image transmission. Generally, the deep-learning based JSCC methods have shown great performance surpassing classic separation-based JPEG2000 source coding and advanced low-density parity-check (LDPC) channel coding, especially for small size images and under human visual perception evaluation matric such as muti-scale structure similarity index measure (MSSSIM) [39].

Despite its great potential, previous studies predominantly concentrate on the development of a more sophisticated model architecture with increased capacity to enhance overall performance. The channel distortion is handled through direct end-to-end optimization. In this case, the JSCC models solely learn coding and decoding strategies by utilizing received signal samples, combating channel interference. To more effectively mitigate channel interference, we integrate the CDDM with the JSCC-based semantic communications system for wireless image transmission, where the signal after CDDM is fed into the JSCC decoder to recover the image. As previously discussed, our CDDM is specially developed to mitigate channel distortion by eliminating channel noise based on an explicit probability of the received signal, thereby improving the performance of the JSCC-based semantic communication system.

The contributions of this paper can be summarized as follows.

- We design a CDDM module based on the U-Net framework in wireless communications, which lies after the channel equalization (or without channel equalization) over the Rayleigh channel (or AWGN channel). The CDDM module learns the distribution of the channel input signal to predict the channel noise and remove it. The model is trained through the forward diffusion process specially designed to adapt the channel models, requiring no knowledge of the current channel state. After training, the CDDM addresses the received signal after equalization with the corresponding sampling algorithm, succeeding in eliminating the channel noise.
- We derive the explicit condition probability of the received signal after equalization according to the channel model and the equalization algorithm, which instructs us to design the corresponding forward diffusion process to match the conditional distribution. The training of the proposed CDDM is accomplished by maximizing the variational lower bound of the logarithm maximum likelihood function, which is relaxed by introducing a series of latent variables in the forward diffusion process. Furthermore, we decompose the variational lower bound into multiple components associated with the latent variables and derive the final loss function using re-parameterization and re-weighted techniques to optimize these components respectively. By utilizing the Bayesian conditional posterior probability, we obtain a sampling algorithm that successfully and effectively mitigates the channel noise.
- We derive the sufficient condition for the reverse sampling algorithm reducing the conditional entropy of the received signal. Through Monte Carlo experiments, we discover the magnitude of the reduction in the upper bound of the conditional entropy differs from various sampling steps, providing insights for selecting the maximum sampling steps.
- We apply the CDDM to a semantic communication system based on the JSCC technique for wireless image transmission, called the joint CDDM and JSCC system and propose a three-stage training algorithm to combine the whole system. Experiments on the mean square error (MSE) between the transmitted signal and the received signal after CDDM prove that compared to the system without CDDM, the system with CDDM has a smaller MSE performance for both Rayleigh fading channel and AWGN channel, indicating that the proposed CDDM can effectively reduce the impact of channel noise through learning. Finally, we conduct numerical experiments on various data distributions, different channels, channel noise levels, channel estimation accuracy, and different evaluation matrices. These results provide compelling evidence that the proposed CDDM outperforms all the benchmarks in dynamic, real-world environments.

The rest of this paper is organized as follows. The system model is introduced in Section II. The detail of the proposed CDDM is presented in Section III. The joint CDDM and JSCC system for semantic communications is introduced in Section IV. Finally, extensive experimental results are presented in Section V, and conclusions are drawn in Section VI.

II. System Model

In this section, we describe the system where the proposed CDDM is employed after the channel equalization as shown in Fig. 2. Let $\mathbf{x} \in \mathbb{R}^{2k}$ be the real-valued symbols. Here, k is the number of channel uses. $\mathbf{x_c} \in \mathbb{C}^k$ is the complex-valued symbols which can be transmitted through the wireless channel, and the i-th transmitted symbol of $\mathbf{x_c}$ can be expressed as $x_{c,i} = x_i + jx_{i+k}$, for $i = 1, \ldots, k$. Thus, the i-th received symbol of the received signal $\mathbf{y_c}$ is

$$y_{c,i} = h_{c,i}x_{c,i} + n_{c,i}, \tag{3}$$

where $h_{c,i} \sim \mathbb{CN}(0,1)$ are independent and identically distributed (i.i.d.) Rayleigh fading gains, $\mathbf{x_c}$ has a power constraint $\mathbb{E}[||\mathbf{x_c}||_2^2] \leq 1$, and $n_{c,i} \sim \mathbb{CN}(0, 2\sigma^2)$ are i.i.d. AWGN samples.

$\mathbf{y_c}$ is then addressed by equalization as $\mathbf{y_{eq}} \in \mathbb{C}^k$, following a normalization-reshape module outputting a real vector $\mathbf{y_r} \in$

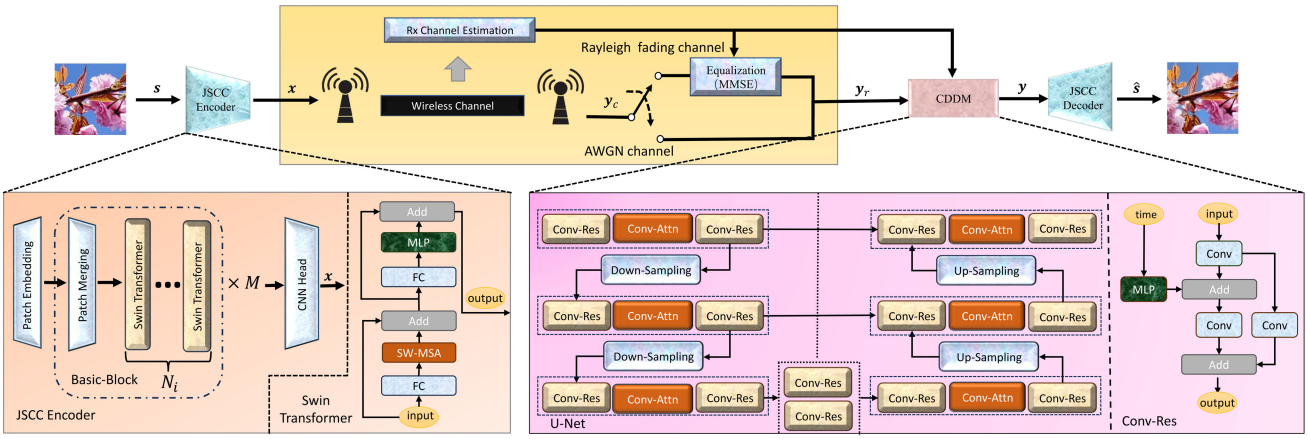


Fig. 2. Architecture of the joint CDDM and JSCC system.

$\mathbb{R}^{2k}$. We consider that the receiver can obtain the channel state $\mathbf{h_c} = [h_{c,1}, \ldots, h_{c,k}]$ through channel estimation and in this paper, we apply minimum mean square error (MMSE) as the equalizer. Therefore, we can derive the conditional distribution of $\mathbf{y_r}$ with known $\mathbf{x}$ and $\mathbf{h_c}$, which can be formulated to instruct the forward diffusion and reverse sampling processes of CDDM.

Proposition 1: With MMSE equalizer, the conditional distribution of $\mathbf{y_r}$ with known $\mathbf{x}$ and $\mathbf{h_c}$ under Rayleigh fading channel is

$$p(\mathbf{y_r}|\mathbf{x}, \mathbf{h_c}) \sim \mathcal{N}(\mathbf{y_r}; \frac{1}{\sqrt{1+\sigma^2}}\mathbf{W_s}\mathbf{x}, \frac{\sigma^2}{1+\sigma^2}\mathbf{W_n^2}), \tag{4}$$

where $\mathbf{H_r} = diag(\mathbf{h_r})$, $\mathbf{h_r} = \begin{bmatrix} |\mathbf{h_c}| \\ |\mathbf{h_c}| \end{bmatrix} \in \mathbb{R}^{2k}$, and

$$\mathbf{W_s} = \mathbf{H_r^2}(\mathbf{H_r^2} + 2\sigma^2\mathbf{I})^{-1}, \mathbf{W_n} = \mathbf{H_r}(\mathbf{H_r^2} + 2\sigma^2\mathbf{I})^{-1}. \tag{5}$$

Proof: Based on the definition, $\mathbf{W_s}$ and $\mathbf{W_n}$ are diagonal matrices, where the i-th and $(i+k)$-th diagonal elements are

$$\begin{aligned} W_{s,i} = W_{s,i+k} &= \frac{|h_{c,i}|^2}{|h_{c,i}|^2 + 2\sigma^2}, \\ W_{n,i} = W_{n,i+k} &= \frac{|h_{c,i}|}{|h_{c,i}|^2 + 2\sigma^2}. \end{aligned} \tag{6}$$

The i-th output of MMSE equalizer $y_{eq,i}$ can be expressed as

$$y_{eq,i} = \frac{|h_{c,i}|^2 x_{c,i} + h_{c,i}^H n_{c,i}}{|h_{c,i}|^2 + 2\sigma^2}. \tag{7}$$

Based on (6), we have

$$\frac{|h_{c,i}|^2 x_{c,i}}{|h_{c,i}|^2 + 2\sigma^2} = W_{s,i} x_{c,i}. \tag{8}$$

With the resampling trick, the conditional distributions of real part and imaginary part of $\frac{h_{c,i}^H n_{c,i}}{|h_{c,i}|^2+2\sigma^2}$ are

$$\begin{aligned} p(Re(\frac{h_{c,i}^H n_{c,i}}{|h_{c,i}|^2 + 2\sigma^2})|h_{c,i}) &\sim \mathcal{N}(0, \sigma^2(\frac{|h_{c,i}|}{|h_{c,i}|^2 + 2\sigma^2})^2) \\ &= \mathcal{N}(0, \sigma^2 W_{n,i}^2), \end{aligned} \tag{9}$$

$$p(Im(\frac{h_{c,i}^H n_{c,i}}{|h_{c,i}|^2 + 2\sigma^2})|h_{c,i}) \sim \mathcal{N}(0, \sigma^2 W_{n,i}^2). \tag{10}$$

Accordingly, we can rewrite $\mathbf{y_r}$ as

$$\mathbf{y_r} = \frac{1}{\sqrt{1+\sigma^2}}(\mathbf{W_s}\mathbf{x} + \mathbf{n_r}), \tag{11}$$

and the distribution $p(\mathbf{n_r}|\mathbf{h_c})$ is $\mathcal{N}(0, \sigma^2\mathbf{W_n^2})$.

Therefore, we have

$$p(\mathbf{y_r}|\mathbf{x}, \mathbf{h_c}) \sim \mathcal{N}(\mathbf{y_r}; \frac{1}{\sqrt{1+\sigma^2}}\mathbf{W_s}\mathbf{x}, \frac{\sigma^2}{1+\sigma^2}\mathbf{W_n^2}). \tag{12}$$

■

Similarly, we have the following proposition for AWGN channel.

Proposition 2: Under AWGN channel, the conditional distribution of $\mathbf{y_r}$ with known $\mathbf{x}$ is

$$p(\mathbf{y_r}|\mathbf{x}) \sim \mathcal{N}(\mathbf{y_r}; \frac{1}{\sqrt{1+\sigma^2}}\mathbf{W_s}\mathbf{x}, \frac{\sigma^2}{1+\sigma^2}\mathbf{W_n^2}), \tag{13}$$

where $\mathbf{W_s}$ and $\mathbf{W_n}$ both become $\mathbf{I}_{2k}$ under AWGN channel.

Proposition 1 and Proposition 2 demonstrate that the channel noise after equalization and normalization-reshape can be re-sampled using $\epsilon \sim \mathcal{N}(0, \mathbf{I}_{2k})$. Additionally, the noise matrix $\mathbf{W_n}$ is related to the modulo form of $\mathbf{h_c}$. As a result, $\mathbf{y_r}$ can be expressed as

$$\mathbf{y_r} = \frac{1}{\sqrt{1+\sigma^2}}\mathbf{W_s}\mathbf{x} + \frac{\sigma}{\sqrt{1+\sigma^2}}\mathbf{W_n}\epsilon. \tag{14}$$

Therefore, the proposed CDDM is trained to obtain $\epsilon_\theta(\cdot)$, which is an estimation of ϵ. Here, θ is all parameters of CDDM. By using $\epsilon_\theta(\cdot)$ and $\mathbf{W_n}$, a sampling algorithm is proposed to obtain $\mathbf{y}$ with the aim to recover $\mathbf{W_s}\mathbf{x}$, which will be described in the next section.

III. Channel Denoising Diffusion Models

The whole structure of the CDDM forward diffusion and reverse sampling process is illustrated in Fig. 3. In this section, we first describe the training algorithm and sampling algorithm of the proposed CDDM. We then derive the sufficient condition for the reverse sampling algorithm reducing the conditional entropy of the received signal.

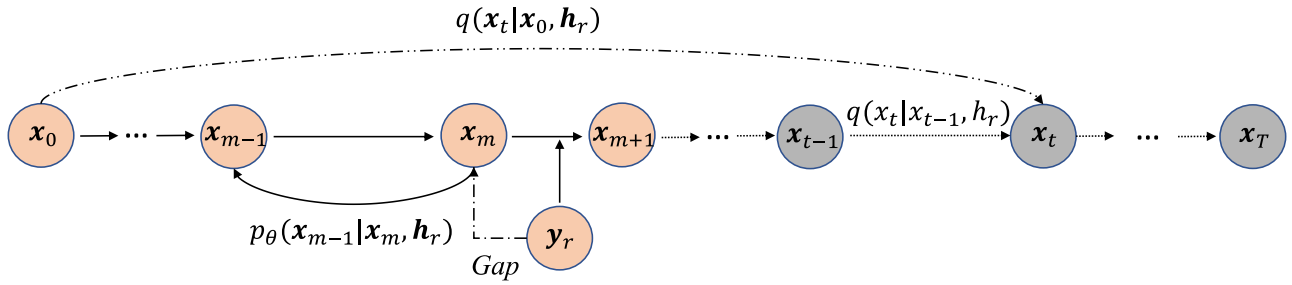


Fig. 3. The forward diffusion process and reverse sampling process of the proposed CDDM.

A. Training Algorithm of CDDM

For the forward process of the proposed CDDM, the original source $\mathbf{x}_0$ is

$$\mathbf{x}_0 = \mathbf{W_s x}. \tag{15}$$

Let T be a hyperparameter. Similar to (1), for all $t \in \{1, 2, \ldots, T\}$, we define

$$\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\mathbf{W_n}\epsilon, \tag{16}$$

and then it can be re-parametered as

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\mathbf{W_n}\epsilon, \tag{17}$$

such that the distribution $q(\mathbf{x}_t|\mathbf{x}_0, \mathbf{h_r})$ is

$$q(\mathbf{x}_t|\mathbf{x}_0, \mathbf{h_r}) \sim \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{W}_\mathbf{n}^2). \tag{18}$$

Based on (4) and (18), if $\bar{\alpha}_m = \frac{1}{1+\sigma^2}$, the Kullback-Leibler (KL) divergence is

$$D_{KL}(q(\mathbf{x}_m|\mathbf{x}_0, \mathbf{h_r})||p(\mathbf{y_r}|\mathbf{x}_0, \mathbf{h_c})) = 0, \tag{19}$$

This indicates that through defining a forward diffusion process, we progressively generate a signal following the same distribution as the one passed through the real channel and equalizer. In this way, **CDDM can be trained on $\mathbf{x}_m$ instead of $\mathbf{y_r}$**. $\mathbf{x}_m$ is defined by m steps as (16) such that in the sampling process, the predicted distribution by CDDM can be decomposed into m small steps and each of them is $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{h_r})$ for $t \in \{1, 2, \ldots, m\}$.

The goal of CDDM is to recover $\mathbf{x}_0$ by learning the distribution of $\mathbf{x}_0$ and removing the channel noise. Therefore, the training of CDDM is performed by optimizing the variational bound on negative log likelihood L. The variational bound of L is formed by $\mathbf{x}_{0:m}$ and $\mathbf{y_r}$, which is given by

$$\begin{aligned} L &= \mathbb{E}\ [-\log\ p_\theta(\mathbf{x}_0|\mathbf{h_r})] \le \mathbb{E}_q\left[-\log(\frac{p_\theta(\mathbf{x}_{0:m}, \mathbf{y_r}|\mathbf{h_r})}{q(\mathbf{x}_{1:m}, \mathbf{y_r}|\mathbf{x}_0, \mathbf{h_r})})\right] \\ &= \mathbb{E}_q\ \underbrace{[D_{KL}(q(\mathbf{y_r}|\mathbf{x}_0, \mathbf{h_r})||p(\mathbf{y_r}|\mathbf{h_r}))}_{L_y} \underbrace{- \log p_\theta(\mathbf{x}_0|\mathbf{x}_1, \mathbf{h_r})}_{L_0} \\ &\quad + \underbrace{D_{KL}(q(\mathbf{x}_m|\mathbf{y_r}, \mathbf{x}_0, \mathbf{h_r})||p_\theta(\mathbf{x}_m|\mathbf{y_r}, \mathbf{h_r}))}_{L_m} \\ &\quad + \sum_{t=1}^{m} \underbrace{D_{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0, \mathbf{h_r})||p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{h_r}))}_{L_{t-1}}], \end{aligned} \tag{20}$$

where L_m instructs to select the hyperparameter m. In this paper, we select m by

$$m = \underset{m_s \in \{1,2,\ldots,T\}}{\operatorname{argmin}} |2\sigma^2 - \frac{1-\bar{\alpha}_{m_s}}{\bar{\alpha}_{m_s}}|. \tag{21}$$

Similar to the process in [3], L_{t-1} can be calculated in closed-form using the Rao-Blackwellized method. The optimization objective of L_{t-1} can be simplified by adopting re-parameterization and re-weighting methods as following

$$\mathbb{E}_{\mathbf{x}_0, \epsilon, \mathbf{h_r}}\left[||\mathbf{W_n}\epsilon - \mathbf{W_n}\epsilon_\theta(\mathbf{x}_t, \mathbf{h_r}, t)||_2^2\right], \tag{22}$$

where $\epsilon_\theta(\mathbf{x}_t, \mathbf{h_r}, t)$ is the output of CDDM. Moreover, it is obvious that $\mathbf{W_n}$ is a random matrix related to $\mathbf{h_r}$. We can derive the expectation of the i-th and $(i+k)$-th diagonal elements of $\mathbf{W_n}$ by following (6):

$$\mathbb{E}_{h_{r,i}}[W_{n,i}] = \mathbb{E}_{h_{r,i}}[W_{n,i+k}] = \mathbb{E}_{h_{r,i}}\left[\frac{h_{r,i}}{h_{r,i}^2 + 2\sigma^2}\right] \triangleq E_{W_n} \tag{23}$$

where $h_{r,i}, i = 1, 2 \ldots, k$ are i.i.d Rayleigh random variables such that all the diagonal elements of $\mathbf{W_n}$ share the same expectation E_{W_n}, indicating that $\mathbb{E}_{\mathbf{h_c}}[\mathbf{W_n}] = E_{W_n}\mathbf{I}$, which is an identical matrix with a coefficient. Therefore, inspired by the reweighting skill in [3], we can ignore $\mathbf{W_n}$ in the loss to reweight the loss as

$$\mathbb{E}_{\mathbf{x}_0, \epsilon, \mathbf{h_r}}\left[||\epsilon - \epsilon_\theta(\mathbf{x}_t, \mathbf{h_r}, t)||_2^2\right]. \tag{24}$$

Finally, to optimize (24) for all $t \in \{1, 2, \ldots, T\}$, the loss function of the proposed CDDM is expressed as follows

$$L_{CDDM}(\theta) = \mathbb{E}_{\mathbf{x}_0, \epsilon, \mathbf{h_r}, t}\left[||\epsilon - \epsilon_\theta(\mathbf{x}_t, \mathbf{h_r}, t)||_2^2\right]. \tag{25}$$

Benefitting from the training algorithm, our CDDM is capable for estimating noise because this is equivalent to utilizing the parameterized distribution $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{h_r})$ to appropriate the real posterior distribution $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0, \mathbf{h_r})$. Neural networks are proficient in distribution approximation. The training procedures of the proposed CDDM are summarized in Algorithm 1.

B. Sampling Algorithm of CDDM

To reduce the time consumption of the sampling process and recover the transmitted signal accurately, (20) implies that selecting m according to (21) and setting $\mathbf{x}_m = \mathbf{y_r}$ is a promising way. By utilizing the received signal $\mathbf{y_r}$, only m steps are needed to be executed. For each time step $t \in \{1, 2, \ldots, m\}$, the trained CDDM outputs $\epsilon_\theta(\mathbf{x}_t, \mathbf{h_r}, t)$,

Algorithm 1 Training Algorithm of CDDM

Input: Training set S, hyper-parameter T and $\bar{\alpha}_t$.
Output: The trained CDDM.

1: **while** the training stop condition is not met **do**
2: Randomly sample $\mathbf{x}$ from S
3: Randomly sample t from $Uniform(\{1,\dots,T\})$
4: Sapmle $|\mathbf{h_c}|$ and compute $\mathbf{H_r}$, $\mathbf{W_s}$ and $\mathbf{W_n}$
5: Randomly sample ϵ from $\mathcal{N}(0,\mathbf{I}_{2k})$
6: Take gradient descent step according to (16) and (25) $\nabla_\theta(||\epsilon-\epsilon_\theta(\mathbf{x}_t,\mathbf{h_r},t)||_2^2)$
7: **end while**

Algorithm 2 Sampling Algorithm of CDDM

Input: $\mathbf{y_r}$,$\mathbf{h_r}$,hyperparameter m
Output: $\mathbf{y}$

1: $\mathbf{x}_m=\mathbf{y_r}$
2: **for** $t=m,\dots,2$ **do**
3: $\mathbf{z}=\mathbf{W_n}\epsilon_\theta(\mathbf{x}_t,\mathbf{h_r},t)$
4: $\mathbf{x}_{t-1}=\sqrt{\bar{\alpha}_{t-1}}(\frac{\mathbf{x}_t-\sqrt{1-\bar{\alpha}_t}\mathbf{z}}{\sqrt{\bar{\alpha}_t}})+\sqrt{1-\bar{\alpha}_{t-1}}\mathbf{z}$
5: **end for**
6: $t=1$
7: $\mathbf{z}=\mathbf{W_n}\epsilon_\theta(\mathbf{x}_1,\mathbf{h_r},1)$
8: $\mathbf{y}=\frac{\mathbf{x}_1-\sqrt{1-\bar{\alpha}_1}\mathbf{z}}{\sqrt{\bar{\alpha}_1}}$

which attempts to predict ϵ from $\mathbf{x}_t$ without knowledge of $\mathbf{x}_0$. A sampling algorithm are required to sample $\mathbf{x}_{t-1}$. The process is executed for m times such that $\mathbf{x}_0$ can be computed out finally.

We first derive the distribution of $\mathbf{x}_{t-1}$ conditioned on $\mathbf{x}_t$, $\mathbf{x}_0$ and $\mathbf{h_r}$ through Bayes rule

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0,\mathbf{h_r}) \sim \mathcal{N}(\mathbf{x}_{t-1};\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_{t-1}}\frac{\mathbf{x}_t-\sqrt{\bar{\alpha}_t}\mathbf{x}_0}{\sqrt{1-\bar{\alpha}_t}},0), \quad (26)$$

where $\mathbf{x}_0$ is acquired by re-writing (17) as following

$$\mathbf{x}_0=\frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t-\sqrt{1-\bar{\alpha}_t}\mathbf{W_n}\epsilon). \quad (27)$$

However, only $\epsilon_\theta(\mathbf{x}_t,\mathbf{h_r},t)$ is available for sampling. $\mathbf{x}_0$ only can be estimated by replacing ϵ with $\epsilon_\theta(\mathbf{x}_t,\mathbf{h_r},t)$ as following

$$\hat{\mathbf{x}}_0=\frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t-\sqrt{1-\bar{\alpha}_t}\mathbf{W_n}\epsilon_\theta(\mathbf{x}_t,\mathbf{h_r},t)). \quad (28)$$

As a result, we are only capable to estimate $q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0,\mathbf{h_r})$ with $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t,\hat{\mathbf{x}}_0,\mathbf{h_r})$. Therefore, a sample of $\mathbf{x}_{t-1}$ can be obtained as

$$\mathbf{x}_{t-1}=\sqrt{\bar{\alpha}_{t-1}}\underbrace{(\frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t-\sqrt{1-\bar{\alpha}_t}\mathbf{W_n}\epsilon_\theta(\mathbf{x}_t,\mathbf{h_r},t)))}_{estimate\ \mathbf{x}_0} + \underbrace{\sqrt{1-\bar{\alpha}_{t-1}}\mathbf{W_n}\epsilon_\theta(\mathbf{x}_t,\mathbf{h_r},t)}_{sample\ \mathbf{x}_{t-1}}. \quad (29)$$

Note that for the last step $t=1$, we only predict $\mathbf{x_0}$ such that the sampling process is taken as

$$\mathbf{y}=\frac{1}{\sqrt{\bar{\alpha}_1}}(\mathbf{x}_1-\sqrt{1-\bar{\alpha}_1}\mathbf{W_n}\epsilon_\theta(\mathbf{x}_1,\mathbf{h_r},1)). \quad (30)$$

The sampling process is summarized in Algorithm 2.

C. Analysis on the Conditional Entropy

To explain the denoising ability of the CDDM, we compare the conditional entropy between $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$, where $\mathbf{x}_t$ is considered as the received signal because (19) has shown that $\mathbf{x}_t$ can belong to the same conditional distribution as the received signal.

For all $t\in\{1,2,\dots,T\}$, $\mathbf{x}_t$ is acquired as (17). According to (18), we can get the conditional entropy of the i-th element of $\mathbf{x}_t$ as $H(x_{t,i}|\mathbf{x}_0,\mathbf{h_c})=\frac{1}{2}\ln(W_{n,i}^2(1-\bar{\alpha}_t))+C$, $i=1,2,\dots,2k$. Here, C is a constant. $\mathbf{x}_{t-1}$ is sampled as (29). However, $\mathbf{x}_t$ is unknown in $H(x_{t-1,i}|\mathbf{x}_0,\mathbf{h_c})$. We can reparameterize (29) with (17) and obtain

$$\mathbf{x}_{t-1}=\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0+\beta_t\mathbf{W_n}\epsilon-\beta_t\mathbf{W_n}\epsilon_\theta(\cdot)+\gamma_{t-1}\mathbf{W_n}\epsilon_\theta(\cdot), \quad (31)$$

where $\beta_t=\frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\alpha_t}}$ and $\gamma_t=\sqrt{1-\bar{\alpha}_t}$. $\epsilon\sim\mathcal{N}(0,\mathbf{I})$ and thus $\mathbf{x}_{t-1}$ is a random variable with respect to ϵ with unknown distribution.

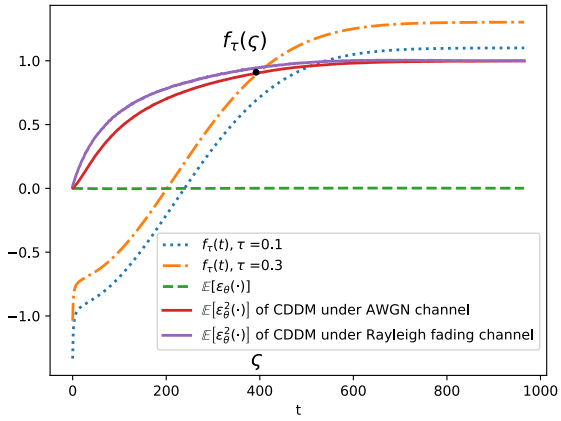


Fig. 4. Experiment results of $\mathbb{E}[\epsilon_\theta(\cdot)]$ and $\mathbb{E}[\epsilon_\theta^2(\cdot)]$ with theoretical values of $f_\tau(t)$ versus sampling step t. The black dot marked the maximum sampling step, below which the model satisfies the sufficient condition under AWGN channel.

Now, we introduce two assumptions for the following analysis.

Assumption 1: There exists a constant bound $\tau>0$ on the element-wise loss function:

$$\mathbb{E}_\epsilon\left[(\epsilon_i-\epsilon_{\theta,i}(\cdot))^2\right]\le\tau. \quad (32)$$

This reasonable and necessary assumption is derived from the fact that the network is optimized sufficiently, meaning the loss function $\mathbb{E}_\epsilon\left[(\epsilon-\epsilon_\theta(\cdot))^2\right]\le\chi$, which can be written into element-wise form as (32).

Assumption 2: The expectation of network output is 0, i.e.,

$$\mathbb{E}_\epsilon\left[\epsilon_{\theta,i}(\cdot)\right]=0. \quad (33)$$

This assumption will be verified through Monte-Carlo in the following. Thus, we have the following theorem.

Theorem 1: Based on the two assumptions mentioned above, for all $t\in\{1,2,\dots,T\}$ and $i=1,2,\dots,2k$, the sufficiency condition of

$$H(x_{t-1,i}|\mathbf{x}_0,\mathbf{h_c})\le H(x_{t,i}|\mathbf{x}_0,\mathbf{h_c}) \quad (34)$$

is

$$\mathbb{E}_{\epsilon}\left[\epsilon_{\theta,i}^2(\cdot)\right] \geq \frac{1-\bar{\alpha}_t-\beta_t\gamma_{t-1}}{\gamma_{t-1}^2-\beta_t\gamma_{t-1}} - \frac{\beta_t^2-\beta_t\gamma_{t-1}}{\gamma_{t-1}^2-\beta_t\gamma_{t-1}}\tau. \quad (35)$$

Proof: According to Assumption 1, we can derive the cross-correlation coefficient of the two random variables ϵ_i and $\epsilon_{\theta,i}(\cdot)$ as following

$$\mathbb{E}_{\epsilon}\left[(\epsilon_i-\epsilon_{\theta,i}(\cdot))^2\right] = \mathbb{E}\left[\epsilon_i^2 - 2\epsilon_i\epsilon_{\theta,i}(\cdot)+\epsilon_{\theta,i}^2(\cdot)\right] \leq \tau. \quad (36)$$

We then have

$$2\mathbb{E}\left[\epsilon_i\epsilon_{\theta,i}(\cdot)\right] \geq 1-\tau+\mathbb{E}\left[\epsilon_{\theta,i}^2(\cdot)\right]. \quad (37)$$

Let $\pi_{t-1,i}^2$ be the variance of $x_{t-1,i}$. According to (31), (37) and Assumption 2, we have

$$\begin{aligned}
&\pi_{t-1,i}^2 \\
&= \mathbb{E}\left[x_{t-1,i}^2\right] - \mathbb{E}^2\left[x_{t-1,i}\right] \\
&= W_{n,i}^2\mathbb{E}\left[\beta_t^2\epsilon_i^2 + (\beta_t-\gamma_{t-1})^2\epsilon_{\theta,i}^2(\cdot) - 2\beta_t(\beta_t-\gamma_{t-1})\epsilon_i\epsilon_{\theta,i}(\cdot)\right] \\
&\leq W_{n,i}^2(\beta_t^2 + (\beta_t-\gamma_{t-1})^2\mathbb{E}\left[\epsilon_{\theta,i}^2(\cdot)\right] \\
&\quad - \beta_t(\beta_t-\gamma_{t-1})(1-\tau+\mathbb{E}\left[\epsilon_{\theta,i}^2\right])) \\
&= W_{n,i}^2((\gamma_{t-1}^2-\beta_t\gamma_{t-1})\mathbb{E}\left[\epsilon_{\theta,i}^2(\cdot)\right] \\
&\quad + \beta_t\gamma_{t-1} + (\beta_t^2-\beta_t\gamma_{t-1})\tau). \quad (38)
\end{aligned}$$

Let $u_\tau(t,\mathbf{h_c})$ be the upper bound of $H(x_{t-1,i}|\mathbf{x}_0,\mathbf{h_c})$. With the maximum entropy principle, we have

$$\begin{aligned}
H(x_{t-1,i}|\mathbf{x}_0,\mathbf{h_c}) &\leq \frac{1}{2}\ln(\pi_{t-1,i}^2) + C \\
&\leq \frac{1}{2}\ln(W_{n,i}^2((\gamma_{t-1}^2-\beta_t\gamma_{t-1})\mathbb{E}\left[\epsilon_{\theta,i}^2(\cdot)\right] \\
&\quad + \beta_t\gamma_{t-1} + (\beta_t^2-\beta_t\gamma_{t-1})\tau)) + C \\
&\triangleq u_\tau(t,\mathbf{h_c}). \quad (39)
\end{aligned}$$

Here, we have $\gamma_{t-1}^2 - \beta_t\gamma_{t-1} < 0$. Therefore, it is easy to obtain the necessity and sufficiency conditional for the inequalities $u_\tau(t,\mathbf{h_c}) \leq H(x_{t,i}|\mathbf{x}_0,\mathbf{h_c})$ as following

$$\mathbb{E}\left[\epsilon_{\theta,i}^2(\cdot)\right] \geq \frac{1-\bar{\alpha}_t-\beta_t\gamma_{t-1}}{\gamma_{t-1}^2-\beta_t\gamma_{t-1}} - \frac{\beta_t^2-\beta_t\gamma_{t-1}}{\gamma_{t-1}^2-\beta_t\gamma_{t-1}}\tau \triangleq f_\tau(t). \quad (40)$$

Taking the necessity and sufficiency condition into (39), we can get the sufficiency condition as the theory. ■

In Fig. 4, the dashed line represents the Monte Carlo results of $\mathbb{E}\left[\epsilon_\theta(\cdot)\right]$ approaching zero, which proves that Assumption 2 holds in the proposed model. It also demonstrates that there exists a limitation ς. If $t \leq \varsigma$, the condition (40) holds. This suggests that the number of sampling steps should be limited in order to achieve performance improvements. Fig. 5 shows the value of $H(x_{t,i}|\mathbf{x}_0,\mathbf{h_c}) - u_\tau(t,\mathbf{h_c})$ at $\tau = 0.3$ versus sampling step t. It is observed that the curve initially exhibits a sharp decline and subsequently levels off rapidly. Considering the two figures together, the sampling step of CDDM can not be determined utilizing (21) when the channel noise power is excessively high, as it would exceed the threshold ς. Furthermore, even if the sampling step is below ς, the gradient becomes very small when it falls within the flattened region. This can lead to the conditional entropy remaining stagnant, resulting in no performance improvement. On the other hand, if the sampling step is too small, the channel noise may not be eliminated sufficiently. Based on the analysis above, we recommend to set $t_{max} = 93$ as shown in the Fig. 5 with star. Correspondingly, (21) is revised into

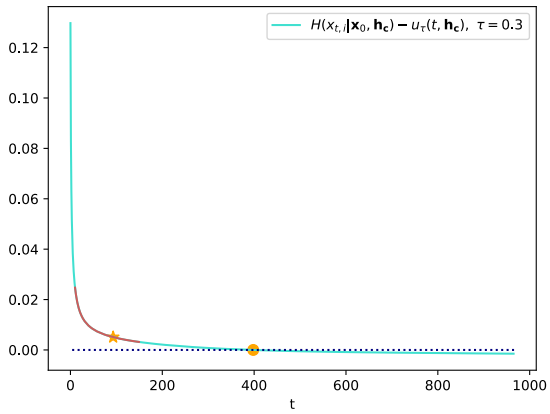


Fig. 5. Experiment results of $H(x_{t,i}|\mathbf{x}_0,\mathbf{h_c}) - u_\tau(t,\mathbf{h_c})$ at $\tau = 0.3$.

$$m = \min(t_{max}, \underset{m_s\in\{1,2,\ldots,T\}}{\operatorname{argmin}} |2\sigma^2 - \frac{1-\bar{\alpha}_{m_s}}{\bar{\alpha}_{m_s}}|). \quad (41)$$

IV. The Joint CDDM and JSCC for Semantic Communications

In this section, the proposed CDDM is applied into a semantic communication system based on JSCC for wireless image transmission. Moreover, we propose a three-stage training algorithm to combine the CDDM and JSCC system.

A. System Structure

An overall architecture of the joint CDDM and JSCC system is shown in Fig. 2. An RGB source image $\mathbf{s}$ is encoded as transmitted signal $\mathbf{x} \in \mathbb{R}^{2k}$ by a JSCC encoder. In this paper, the JSCC is built upon the Swin Transformer [38] backbone, which has a more powerful expression ability than vision transformer by replacing the standard multi-head self-attention in vision transformer with a shift window multi-head self-attention. $\mathbf{x}$ is then transmitted and processed into $\mathbf{y_r}$ at the receiver, as described in Section II. At the receiver, the proposed CDDM removes the channel noise from $\mathbf{y_r}$ using Algorithm 2. Following this, the output of CDDM is fed into the JSCC decoder to reconstruct the source image $\hat{\mathbf{s}}$.

B. Training Algorithm

The entire training algorithm of the joint CDDM and JSCC system consists of three stages. In the first stage, the JSCC encoder and decoder are trained jointly through the channel shown in Fig. 2, except for the CDDM module, to minimize the distance $d(\mathbf{s},\hat{\mathbf{s}})$. Therefore, the loss function for this stage is given by

$$L_1(\phi,\varphi) = \mathbb{E}_{\mathbf{s}\sim p_\mathbf{s}}\mathbb{E}_{\mathbf{y_r}\sim p_{\mathbf{y_r}|\mathbf{s}}}\left[d(\mathbf{s},\hat{\mathbf{s}})\right]. \quad (42)$$

where ϕ and φ encapsulate all parameters of JSCC encoder and decoder respectively.

In the second stage, the parameters of the JSCC encoder are fixed such that CDDM can learn the distribution of $\mathbf{x}_0$ via Algorithm 1. The training process is not affected by

Algorithm 3 Training Algorithm of the Joint CDDM and JSCC

Input: Training set $\mathbf{S}$, hyper-parameter T, $\bar{\alpha}_t$, and the channel estimation results $\mathbf{h_c}$ and σ^2.

Output: The well-trained joint CDDM and JSCC system.

while the training stop condition of stage one is not met **do**
  Randomly sample $\mathbf{s}$ from S
  Perform forward propagation through channel without CDDM.
  Compute $L_1(\phi, \varphi)$ and update ϕ, φ
end while
while the training stop condition of stage two is not met **do**
  Randomly sample $\mathbf{s}$ from S
  Compute $\mathbf{s}$ as $\mathbf{x}$
  Train CDDM with Algorithm 1.
end while
while the training stop condition of stage three is not met **do**
  Randomly sample $\mathbf{s}$ from S
  Perform forward propagation through channel with noise power σ^2 with the trained CDDM
  Compute $L_3(\varphi)$ and update φ
end while

the channel noise power because Algorithm 1 has a special forward diffusion process, and the process has been designed specially to simulate the distribution of channel noise. Benefitting from this, CDDM is designed for handling various channel conditions and requires only one training process.

In the third stage, the JSCC decoder is re-trained jointly with the trained JSCC encoder and CDDM to minimize $d(\mathbf{s}, \hat{\mathbf{s}})$. The entire joint CDDM and JSCC system is performed through the real channel, while only the parameters of the decoder are updated. The loss function is derived as

$$L_3(\varphi) = \mathbb{E}_{\mathbf{y}\sim p_{\mathbf{y}|\mathbf{s}}}\left[d(\mathbf{s}, \hat{\mathbf{s}})\right]. \tag{43}$$

The training algorithm is summarized in Algorithm 3.

C. Model Structure

The schematic of the JSCC encoder and the U-Net structure in CDDM are illustrated in Figure 2. In the JSCC encoder, the initial module is the patch embedding, responsible for partitioning the source image into non-overlapping patches. Subsequently, M basicblocks are employed to extract the semantic features from the source image. The i-th basicblock consists of a patch merging module and N_i swim transformers, where $i = 1, 2 \ldots M$. After addressed by a basicblock, the height and width of the features are halved, while the channel dimensions are increased to P_i. Finally, a convolution head (Conv Head) layer is adopted to compute the features as transmitted signal $\mathbf{x}$. The structure of the JSCC decoder is identical to that of the JSCC encoder, with the exception that the downsample modules in the JSCC encoder are replaced with upsample modules.

The model structure of CDDM is predominantly based on the convolutional improved U-Net architecture [40]. Initially, $\mathbf{y}_r$ undergoes a convolution layer and then serves as the input of the U-Net. Subsequently, the output of U-Net is further processed by another convolutional layer to generate the final output $\mathbf{y}$. The U-Net is comprised of various components, including convolutional residual (Conv-Res) blocks [41], convolutional attention (Conv-Attn) blocks, down-sampling blocks, and up-sampling blocks. A down-sampling block is a convolutional layer that performs down-sampling and maintains the same number of input and output channels. The up-sampling block consists of an interpolation layer followed by a convolutional layer. The Conv-Attn is an attention block commonly adopted in classic transformer [42], but with the notable distinction of employing convolutional layers as a replacement for fully-connected (FC) layers. The structure of Conv-Res is depicted in Fig. 2. In comparison to the classic residual block, the Conv-Res block substitutes FC layers with convolutional layers. Moreover, an additional convolutional layer is incorporated into the residual path to adjust the data dimension and enhance the model's capacity. The sampling step t is addressed as an embedding vector and then embedded into the middle layer of the Conv-Res block. Multiple instances of these blocks are sequentially connected incorporating two additional residual paths, ultimately forming the U-Net architecture.

V. Experiments Results

In this section, we provide a detailed description of the experimental setup and present a considerable amount of experimental results, which comprehensively demonstrate the influence of hyperparameters and the effectiveness of our proposed CDDM in complex practical scenarios.

A. Experiment Setup

1) Datesets: To obtain comprehensive and universally applicable results, we train and evaluate the proposed joint CDDM and JSCC system on two image datasets with different resolutions. CIFAR10 [43] dataset is employed for low-resolution images with dimensions of 32×32, comprising of 50000 color images for training and 10000 images for testing. The high-resolution images are obtained from DIV2K dataset [44], which includes 800 images for training and 100 images for testing. These images are collected from a wide range of real-world scenes and have a uniform resolution of 2K. During the training process, the images with high resolution are randomly cropped into patches with a size of 256×256.

2) Comparison Schemes: We conduct a comparative analysis between the proposed joint CDDM and JSCC system and four other systems: the JSCC system, the joint DnCNN and JSCC system, the joint GAN and JSCC system and the classical handcrafted separation-based source and channel coding system. More specifically, for the joint DnCNN and JSCC system as well as the joint GAN and JSCC system, we apply the original DnCNN proposed in [13] and GAN proposed in [15] to the signal space. The JSCC modules in the three systems share an identical structure. We train

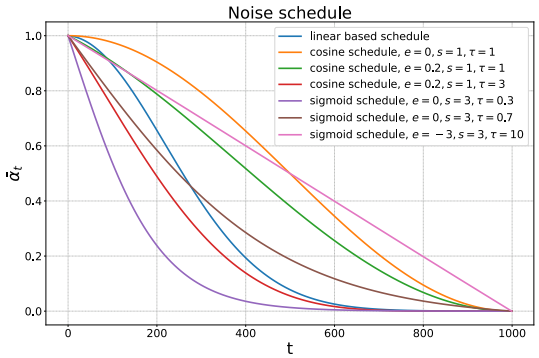


Fig. 6. Values of $\bar{\alpha}_t$ of noise schedules based on different functions with various hyperparameters.

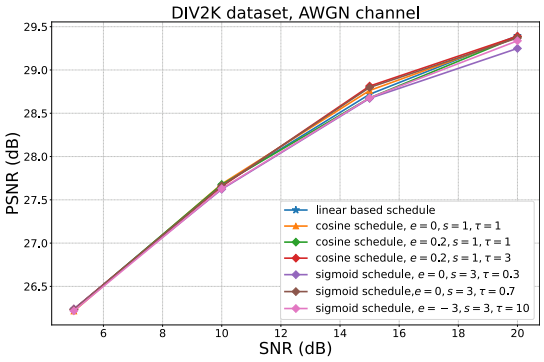


Fig. 7. PSNR performance versus SNRs under AWGN channel with various noise schedules. The CBR is set to 3/128.

these benchmarks with the same training configuration as the joint CDDM and JSCC system. It is worth emphasizing that in the event of a change in channel SNR, all systems undergo retraining to optimize their performance under the specific SNR condition. For the classical system, we employ the JPEG2000 codec for compression and LDPC [45] codec for channel coding, marking as "JPEG2000+LDPC". Here, we consider DVB-T2 LDPC codes with a block length of 64800 bits for different coding rates and quadrature amplitude modulations (QAM) adapted to the channel conditions.

3) Evaluation Metrics: We qualify the performance of all three schemes with both PSNR and MSSSIM. PSNR is a widely-used pixel-wise metric that measures the visibility of errors between the reconstructed image and the reference image. A higher PSNR value indicates a smaller loss in the image quality. In this case, we adopt MSE to calculate $d(\cdot)$ during optimizing our networks. MSSSIM is a perceptual metric that specially concentrates on the structural similarity and content of images, which aligns more closely with the evaluation results of the human visual system (HVS). The multi-scale design allows it to demonstrate consistent performance across images with varying resolutions. The value of MSSSIM ranges from 0 to 1, where a higher value indicates a higher similarity to the reference image. Also in this case, we adopt 1-MSSSIM to calculate $d(\cdot)$ during optimizing our networks. When testing the performance, we convert MSSSIM into the form of dB for more intuitive observation and comparison. The formula is $MSSSIM\ (dB) = -10\ \log_{10}(1 - MSSSIM)$.

4) Hyperparameters Design: The hyperparameters in diffusion models, such as noise schedules α_t and t_{max} may significantly influence the performance of CDDM, as illustrated in [46] and [47]. To comprehensively understand the influence of hyperparameters on our CDDM, we employ seven noise schedules based on linear function, cosine function and sigmoid function and evaluate their performance on the joint CDDM and JSCC system. For the linear function-based noise schedule, we set α_t to constants decreasing linearly from an initial value of $\alpha_1 = 0.9999$ to a final value $\alpha_T = 0.9800$. For the cosine-based schedules, the function of $\bar{\alpha}_t$ versus t can be formulated as

$$\bar{\alpha}_t = \frac{cos(\frac{\pi}{2}e)^{2\tau} - cos(\frac{\pi}{2}((e-s)\frac{t}{T}+s))^{2\tau}}{cos(\frac{\pi}{2}e)^{2\tau} - cos(\frac{\pi}{2}s)^{2\tau}},\ t \in (0, T], \tag{44}$$

For the sigmoid-based schedules, the function can be formulated as

$$\bar{\alpha}_t = \frac{sigmoid(\frac{e}{\tau}) - sigmoid(\frac{(e-s)\frac{t}{T}+s}{\tau})}{sigmoid(\frac{e}{\tau}) - sigmoid(\frac{s}{\tau})},\ t \in (0, T]. \tag{45}$$

where e, s and τ are hyperparameters, determining the shape of the functions. In our experiments, we set $[e, s, \tau] = [0, 1, 1], [0.2, 1, 1], [0.2, 1, 3]$ for cosine-based schedules. and $[e, s, \tau] = [0, 3, 0.3], [0, 3, 0.7], [-3, 3, 10]$ for sigmoid-based noise schedules, respectively. The graphical representations of all seven noise schedules are shown in the Fig. 6. These noise schedules encompass diverse noise distributions, providing a comprehensive understanding of sensitivity of our CDDM to noise schedules.

For the hyperparameter t_{max}, we also select it based on the setting of SNR as the approach we select t mentioned in Section II. In our experiments, we set the SNR to $[5, 7, 8, 10, 15, 20]$ dB for t_{max} such that the corresponding t_{max} is $[162, 131, 117, 93, 52, 27]$, respectively.

5) Training Details: For the CDDM training and sampling algorithms, T is set to 1000. During training CDDM, we employ an Adma optimizer [48] and implement a cosine warm-up learning rate schedule [49] with an initial learning rate of 0.0001. In terms of the JSCC structure, the number of basic-blocks and patches varies depending on the dataset. For CIFAR10 dataset, the number of Basicblocks, denoted as M, is set to 2, Swin Transformer numbers $[N_1, N_2] = [2, 4]$ and channel dimensions $[P_1, P_2] = [128, 256]$. On the other hand, for DIV2K dataset comprising high-resolution images, M is set to 4, Swin Transformer numbers $[N_1, N_2, N_3, N_4] = [2, 2, 6, 2]$ and channel dimensions $[P_1, P_2, P_3, P_4] = [128, 192, 256, 320]$. We employ Adam optimizer with learning rate 0.0001 to optimize the JSCC [36].

B. Hyperparameters Influence

We first evaluate the influence of hyperparameters and the approximation in loss function on the performance of our joint CDDM and JSCC system through experiments. The results are obtained from DIV2K dataset and the channel bandwidth ratio (CBR) is set to $3/128$. Fig. 7 shows the PSNR performance of the joint CDDM and JSCC system over AWGN channel with various noise schedules.

In low SNR regimes, the joint CDDM and JSCC systems under different noise schedules exhibit nearly same PSNR

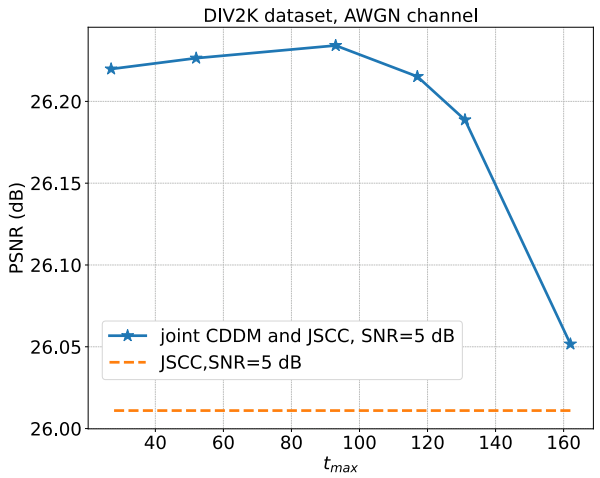


Fig. 8. PSNR performance of DIV2K dataset versus t_{max} at SNR= 5 dB. The CBR is set to $3/128$.

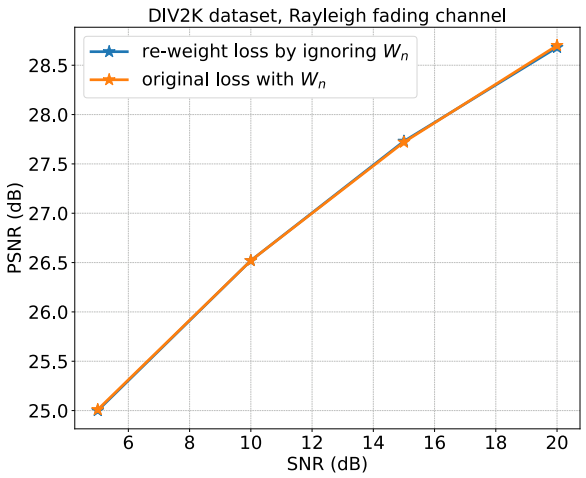


Fig. 9. PSNR performance of DIV2K dataset versus t_{max} at SNR= 5 dB. The CBR is set to $3/128$.

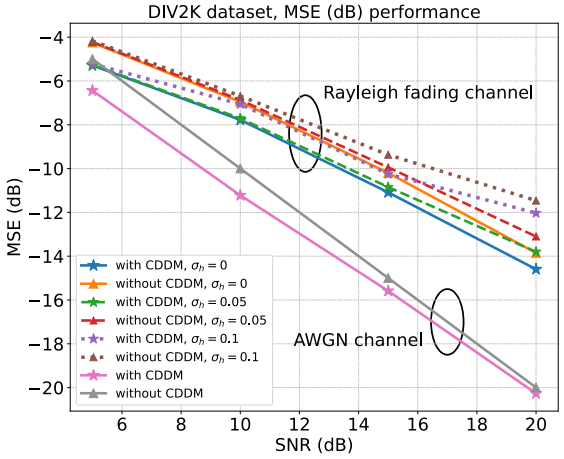


Fig. 10. MSE performance of DIV2K versus SNRs under AWGN and Rayleigh fading channel with or without channel estimation errors. The CBR is $3/128$.

performance. For example, at SNR= 5 dB, the PSNR performance ranges from a maximum of 26.239 dB, achieved by sigmoid-based noise schedule with $[e, s, \tau] = [0, 3, 0.3]$, to a minimum of 26.241 dB. While at higher SNRs, distinct noise schedules result in slight performance differences. At SNR= 15 dB, the PSNR performance ranges from a maximum of 28.814 dB, achieved by the cosine-based noise schedule with $[e, s, \tau] = [0, 1, 1]$ to 28.672 dB and the linear-based noise schedule achieves the optimal performance of 29.396 dB at SNR= 20 dB while the worst performance is 29.249 dB. These experimental results demonstrate that noise schedules only have slight influence on our system, with a maximum gap of 0.14 dB in PSNR observed at SNR= 15 dB. Moreover, the optimal noise schedule varies under different SNR conditions. This variation is attributed to the joint training of the encoder and channel in the first stage. In this training process, the output of encoder follows different distributions at varying channel states, leading to different learning tasks for CDDM. Therefore, the optimal noise schedules vary when the tasks change. In conclusion, we discover that noise schedules only have slight impact on the PSNR performance of the joint CDDM and JSCC system, and the optimal noise schedules vary under different channel conditions. Therefore, this paper selects a noise schedule that performs relatively well under all conditions and achieves the best performance at SNR= 20 dB.

Fig. 8 shows the PSNR performance of the joint CDDM and JSCC system versus t_{max} under AWGN channel with SNR= 5 dB. It can be observed that performance degradation occurs when t_{max} is too large and our system achieves the best PSNR performance at $t_{max} = 93$, as recommended in Fig. 5. These experiments prove that setting an upper bound for the sampling steps according to Theorem 1 is necessary and reasonable for performance enhancement.

We also conduct experiments to illustrate the effect of reweighting loss function by ignoring $\mathbf{W_n}$. We trained two joint CDDM and JSCC systems under Rayleigh fading channel. One of the CDDM is trained using the reweighted loss as equation (25), while the other is trained with the original loss as equation (22). Fig. 9 shows their performance versus SNRs and they almost perform identically across all SNRs, proving that the approximation in loss function is adoptable.

Based on the analysis and experimental results on hyperparameters, in the following experiments, we adopt the linear based noise schedule, $t_{max} = 93$ and the reweighted loss function to train and evaluate our CDDM.

C. MSE Performance and Visualization Results

Fig. 10 illustrates the MSE performance of CDDM in different SNR regimes. The results are based on DIV2K dataset with JSCC trained for maximizing PSNR and CBR is set to $3/128$. In the case of using CDDM, we calculate the MSE between $\mathbf{x}$ and $\mathbf{y}$, while in the case of not using CDDM, we calculate the MSE between $\mathbf{x}$ and $\mathbf{y}_r$. As shown in Fig. 2, $\mathbf{y}_r$ and $\mathbf{y}$ are the input and output of CDDM, respectively. The solid line in Fig. 10 shows that the system with CDDM performs much better than the system without CDDM in all SNR regimes under both AWGN and Rayleigh fading channels. For example, for AWGN channel, the proposed CDDM reduces the MSE by 0.27 dB at SNR= 20 dB. Meanwhile, it can be seen that as the SNR decreases, the gain of CDDM in MSE increases. This indicates that as the SNR decreases, i.e., the channel noise increases, the proposed CDDM is easier to remove more noise, e.g. 1.44 dB gain at SNR= 5 dB for AWGN channel. Moreover, it is important to note that under

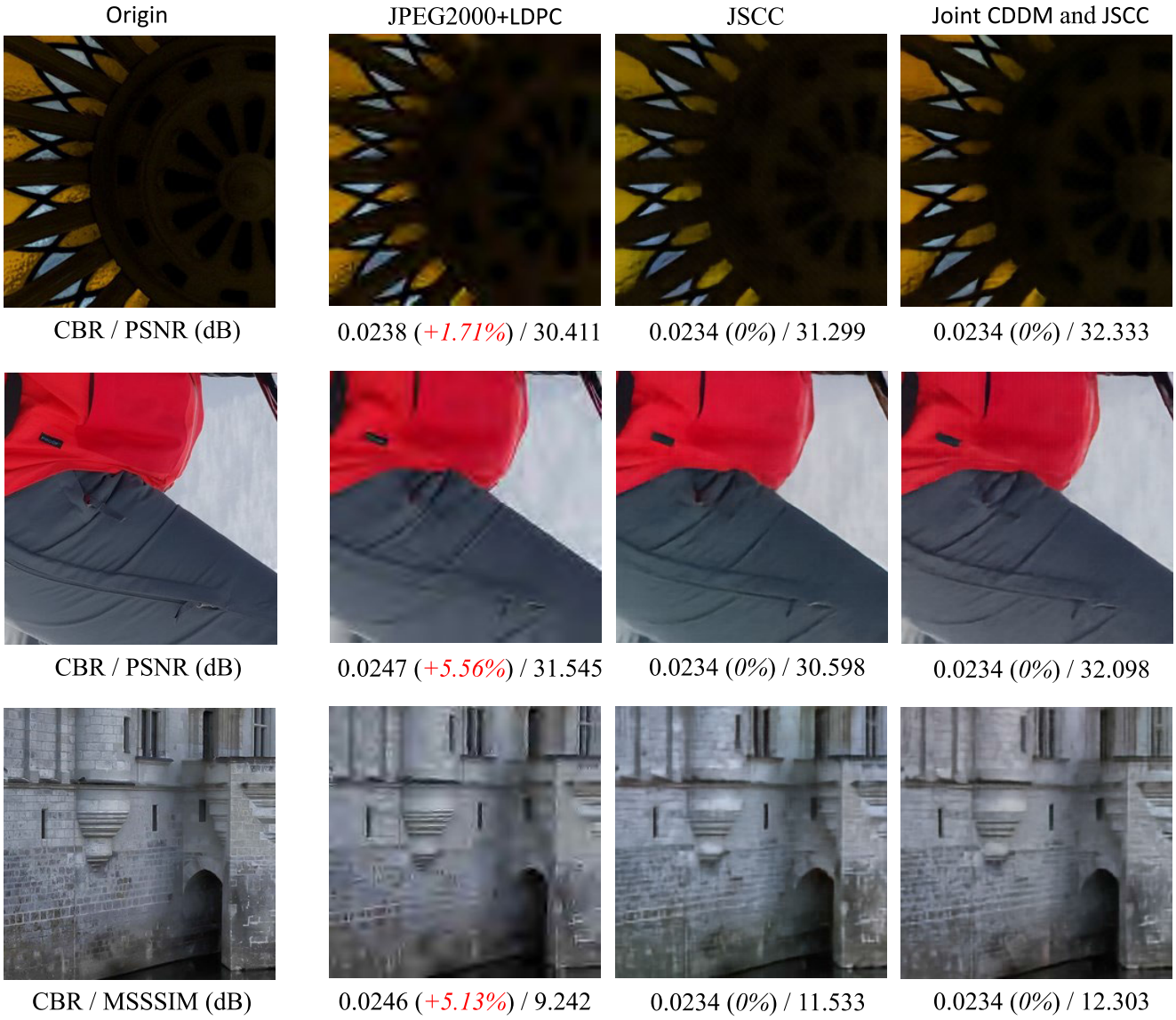


Fig. 11. Examples of visualization results under Rayleigh fading channel at SNR= 10 dB. The four columns display the original images and the reconstructed images obtained from their respective systems. The red number corresponds to the percentage of additional bandwidth cost in comparison to the joint CDDM and JSCC system.

Rayleigh fading channel, MMSE equalizer has theoretically minimized the MSE, but CDDM can further reduce the MSE after MMSE equalization. The reason for this fact is that CDDM can learn the distribution of $\mathbf{x}_0 = \mathbf{W_s x}$, and utilizes this learned knowledge to remove the noise thereby further reducing the MSE.

Additionally, to conduct a more comprehensive evaluation of our model, we assess the robustness of the proposed CDDM under Rayleigh fading channel with the presence of channel estimation errors. The receiver obtains a noisy estimation of $\mathbf{h}$, denoted as $\hat{\mathbf{h}}$ which is formulated as $\hat{\mathbf{h}} = \mathbf{h} + \Delta\mathbf{h}$, where $\Delta\mathbf{h} \sim \mathbb{CN}(0, \sigma_h^2\mathbf{I})$. In Fig. 10, the dashed lines correspond to lower estimation errors with $\sigma_h = 0.05$ and the dotted lines represent more estimation errors with $\sigma_h = 0.1$. It is observed that under $\sigma_h = 0.05$, the joint CDDM and JSCC system maintains gains relative to perfect channel estimation across all SNR ranges. However, as σ_h increases to 0.1, the gains tend to decrease. This reduction is particularly notable at SNRs of 10 and 20 dB.

Fig. 11 visualizes the reconstructions generated by the three systems. The results are obtained under Rayleigh fading channel with perfect channel estimation and an SNR of 10 dB. It can be observed clearly that both JSCC-based systems outperform JPEG2000+LDPC in terms of visual quality, despite a slightly lower CBR. However, the reconstructed images obtained from the JSCC system demonstrate significant color aberrations when compared to their corresponding original images. For example, the first image exhibits a lean towards a pale yellow hue, while the second and third images tend to lean towards a cyan color tone. On the contrary, our joint CDDM and JSCC system simultaneously demonstrates superior color consistency and better visual quality.

D. PSNR Performance

Fig. 12 illustrates the PSNR performance for DIV2K dataset versus SNR under AWGN channel. The CBR is configured to $3/128$. Our joint CDDM and JSCC system demonstrates

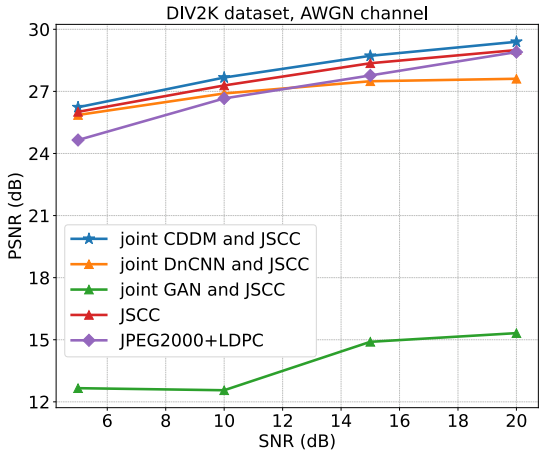


Fig. 12. PSNR performance of DIV2K dataset versus SNR under AWGN channel with various denoising methods. The CBR is set to $3/128$.

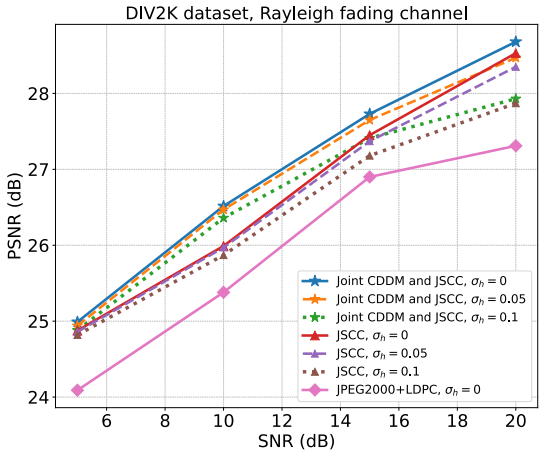


Fig. 13. PSNR performance of DIV2K versus SNR under Rayleigh fading channel with or without channel estimation errors. The CBR is $3/128$.

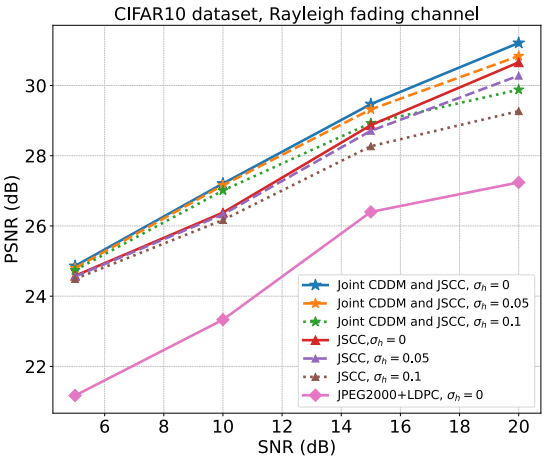


Fig. 14. PSNR performance of CIFAR10 versus SNR under Rayleigh fading channel with or without channel estimation errors. The CBR is $1/8$.

superior performance compared to all the benchmarks across a range of SNRs from 5 to 20 dB. For example, the joint CDDM and JSCC system achieves 0.4 dB gain in PSNR at SNR$=$ 20 dB compared with the JSCC system. However, it can be observed clearly that the joint DnCNN and JSCC system performs even worse than the JSCC system. This is because DnCNN only employs a single step to predict and subsequently eliminate noise. The estimation is not entirely accurate and can further corrupt the received signals. In contrast, CDDM utilizes an iterative process to predict and eliminate noise. The estimation error can be gradually corrected through the iterations. Therefore, CDDM can improve the performance, whereas DnCNN, in contrast, tends to weaken it. For the joint GAN and JSCC system, its PSNR performance is extremely poor, more than 10 dB worse than other systems. This is because GAN generates signals directly based on the received signals, potentially leading to a significant loss of useful information. Therefore, the JSCC decoder is unable to effectively reconstruct the images. According to these experiments, we can draw the conclusion that CDDM is indispensable for noise removal from received signals and for enhancing the performance of the semantic communication systems. Furthermore, the joint CDDM and JSCC system achieves significantly better performance when compared to the JPEG2000+LDPC system. Specifically, at an SNR of 20 dB, the performance of the JPEG2000+LDPC system is comparable to that of the JSCC system, but still exhibits a 0.5 dB inferiority compared to our joint CDDM and JSCC system. These experimental results shown in Fig. 12 demonstrate that CDDM can be applied to remove noise and it is necessary to utilize CDDM to remove noise in received signals.

In the following, we present numerical experimental results in a great number of complex scenarios, including various data distributions, different channels, channel noise levels, channel estimation accuracy, and different evaluation matrices, to illustrate that our CDDM can enhance the performance in a wide range of dynamic, real-world environments, which is critical to the deployment in practical semantic communication systems. Figs. 13 and 14 illustrate the PSNR performance for both DIV2K and CIFAR10 datasets under Rayleigh fading channel. The CBR is $3/128$ for DIV2K and $1/8$ for CIFAR10. The solid line, dashed line and dotted line represent that σ_h is 0, 0.05 and 0.1, respectively. It can be observed that, the joint CDDM and JSCC system consistently outperforms the JSCC system across both datasets and all SNRs, i.e. 0.83 dB for CIFAR10 dataset and 0.53 dB for DIV2K dataset at SNR$=$ 10 dB with perfect channel estimation. Meanwhile, it is worth noting that the gain in PSNR performance for DIV2K dataset tends to decrease as the SNR increases when $\sigma_h = 0.1$, which aligns with the decrease in MSE performance gain. The experimental results under both datasets, conducted at a channel estimation error level of $\sigma_h = 0.1$, highlight the lack of natural robustness in our system when exposed to high channel estimation errors and high SNR conditions. This finding underscores the need to devise a specialized framework to mitigate the influence of channel estimation errors and enhance the system robustness in the future.

Figs. 15 and 16 show the PSNR performance for DIV2K dataset in different CBRs under AWGN and Rayleigh fading channel, respectively. The SNR is set to 10 dB. It is evident that our joint CDDM and JSCC system maintains effectiveness for complex high-resolution DIV2K dataset across various CBRs despite that the performance gain decreases as the CBR increases. This phenomenon can be attributed to the increase in the dimensionality of the transmitted signal x when the CBR increases, thereby leading to a notable augmentation in the

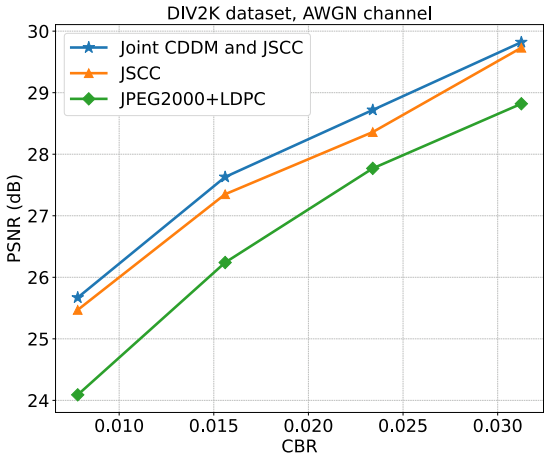


Fig. 15. PSNR performance of DIV2K dataset versus CBR under AWGN channel. The SNR is 10 dB.

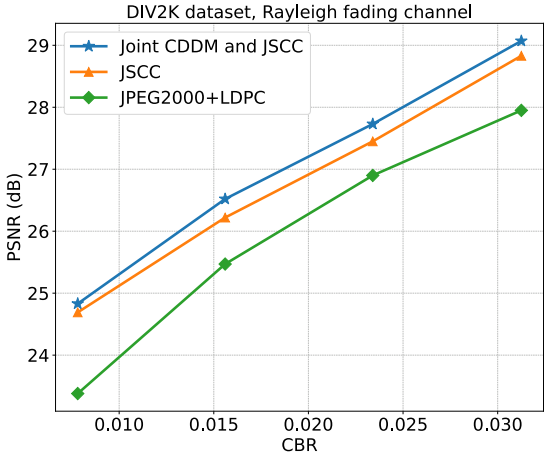


Fig. 16. PSNR performance of DIV2K dataset versus CBR under Rayleigh fading channel. The SNR is 10 dB.

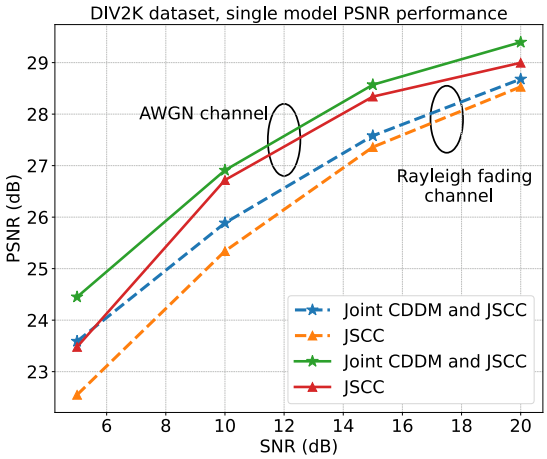


Fig. 17. PSNR performance, trained at a SNR of 20 dB, for DIV2K versus SNR under both AWGN and Rayleigh fading channels.

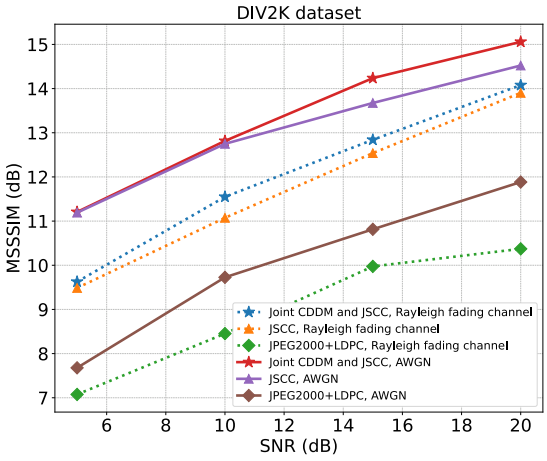


Fig. 18. MSSSIM performance of DIV2K versus SNR under both AWGN and Rayleigh fading channels. The CBR is $3/128$.

complexity of the learned distribution. However, to maintain experimental fairness, the structure and depth of the CDDM remain unchanged for different CBRs, consequently impeding the model's capacity to effectively learn the complex distribution and leading to a decline in performance gain.

Fig. 17 illustrates the PSNR performance versus SNR for DIV2K dataset over both AWGN and Rayleigh fading channel. In this experiment, the joint CDDM and JSCC system, as well as the JSCC system, are trained at a fixed SNR of 20 dB and evaluated across various SNR values. It is evident that our joint CDDM and JSCC system consistently outperforms the JSCC system. More importantly, the performance gain becomes more pronounced as the SNR decreases in the Rayleigh fading channel. We attribute this phenomenon to the training of our CDDM utilizing Algorithm 1, which encompasses a wide range of SNRs. Consequently, when the SNR varies, our CDDM still effectively reduces noise by adjusting the sampling step m, leading to enhanced performance. In contrast, the performance of the JSCC system deteriorates rapidly as the SNR decreases. This observation validates the adaptability of our joint CDDM and JSCC system to different SNRs.

E. MSSSIM Performance

Fig. 18 shows the MSSSIM performance versus SNR for DIV2K dataset over both AWGN channel and Rayleigh fading channel. The solid lines represent performance under AWGN channel and the dotted lines represent performance under Rayleigh fading channel. The results demonstrate that under AWGN channel, our joint CDDM and JSCC system achieves a notable improvement in MSSSIM performance at SNRs of 15 dB and 20 dB particularly, i.e. 0.6 dB at SNR$=$ 15 dB. At lower SNRs, we can still achieve an enhanced performance albeit with a quite small magnitude. Under Rayleigh fading channel, we achieve significant improvement across all SNRs. Fig. 19 demonstrates the MSSSIM performance for CIFAR10 dataset over Rayleigh fading channel. It can be observed that the joint CDDM and JSCC system outperforms both the JSCC system and the JPEG2000+LDPC system across all SNRs.

Fig. 20 demonstrates the MSSSIM performance versus CBR for DIV2K under both AWGN channel and Rayleigh fading channel respectively. The results demonstrate that our joint CDDM and JSCC system outperforms the JSCC system under all examined conditions. Analogous to the PSNR performance, the magnitude of gain decreases when the CBR is large due to the same reason. Moreover, all the experiment results conducted with MSSSIM performance show the consistent phenomenon that the MSSSIM performance of the JPEG2000+LDPC system is remarkably poor across all experimental configurations, showcasing a substantial disparity

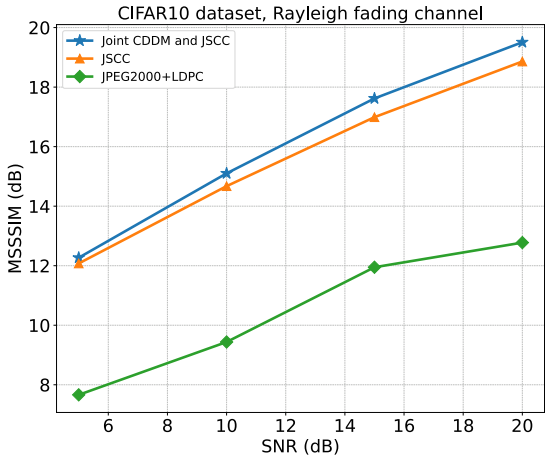


Fig. 19. MSSSIM performance of CIFAR10 versus SNR under Rayleigh fading channel. The CBR is 1/8.

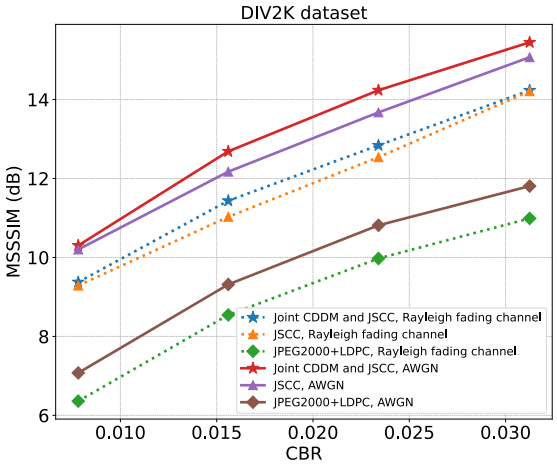


Fig. 20. MSSSIM performance of DIV2K dataset versus CBR under both AWGN and Rayleigh fading channels. The SNR is 10 dB.

TABLE I

MACs OF THE CDDM AND JSCC ON DIV2K DATASET WITH DIFFERENT CBRs

	MACs	
CBR	CDDM	JSCC
0.0078	6.769 G	32.726 G
0.0156	27.076 G	32.736 G
0.0234	60.922 G	32.746 G
0.0313	108.30 G	32.755 G

compared to both JSCC-based systems. These phenomenons prove that when considering the HVS, the JSCC system exhibits a dominant advantage over the JPEG2000+LDPC system. Furtherly, in this scenario, our joint CDDM and JSCC system can still enhance the performance.

Further, Table I presents the computational complexity of the CDDM and the JSCC in DIV2K dataset with different CBRs, given a comprehensive understanding of the trade-off between performance improvement and computational overhead. The computational complexity is measured by multiply-accumulate operations (MACs), a widely used metric for quantifying the number of multiply-accumulate operations required for neural network inference for once. It is worth noting that the MACs are dependent on the CBR because the dimensions of input data influence the model scale. We can observe that when CBR is low, the MACs required for one inference in CDDM are significantly lower than those required for JSCC. However, as CBR increases, the MACs required for inference in CDDM increase, while the MACs for JSCC remain relatively constant. This is because the increase in CBR leads to a significant increase in the input dimension for CDDM. Therefore, even though CDDM employed with high CBR have the same structure as those employed with low CBR, they require wider hidden layers to accommodate the increased input dimensions, resulting in an overall increase in computational complexity. However, for JSCC, the increase in CBR only results in an increase in the channel numbers of the CNN head modules in the encoder and decoder, while the other parts of the model remain unchanged. As a result, there is a smaller increase in MACs for JSCC. Another noteworthy point is that diffusion models require multiple inferences to obtain the final output. Therefore, the MACs required for CDDM to obtain the final output are equal to the MACs required for one inference multiplied by the number of the sampling steps. For example, with our proposed accelerating sampling algorithm, our CDDM only needs 27 steps to obtain the final output at SNR$=$ 20 dB.

The experiments conducted consistently demonstrate the efficacy of our joint CDDM and JSCC system, surpassing the performance of all benchmarks, such as the JSCC system and the JPEG2000+LDPC system across a wide range of conditions. The numerical experimental results in complex scenarios, including different data distributions, different channels, channel noise levels, channel estimation accuracy, and different evaluation metrics, strongly proving that our CDDM can enhance the performance of semantic communication systems in dynamic, real-world scenarios.

VI. CONCLUSION

In this paper, we have proposed the channel denoising diffusion models to eliminate the channel noise under Rayleigh fading channel and AWGN channel. CDDM is trained utilizing a specialized noise schedule adapted to the wireless channel, which permits effective elimination of the channel noise via a suitable sampling algorithm in the reverse sampling process. Further, we derived the sufficient condition under which our CDDM can reduce the conditional entropy of the received signal and demonstrate that the well-trained model satisfies this condition for smaller sampling steps through Monte Carlo experiments. CDDM is then applied into the semantic communications system based on JSCC. Extensive experimental results show that our joint CDDM and JSCC system outperforms all the benchmarks in terms of MSE, PSNR and MSSSIM in dynamic complex scenarios.

REFERENCES

[1] T. Wu et al., "CDDM: Channel denoising diffusion models for wireless communications," in *Proc. IEEE Global Commun. Conf.*, Dec. 2023, pp. 1–5.

[2] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using non-equilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2256–2265.

[3] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851.

[4] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–22.

[5] L. Yang et al., "Diffusion models: A comprehensive survey of methods and applications," 2022, *arXiv:2209.00796*.

[6] M. Chenlin, H. Yutong, and S. Yang, "SDEdit: Guided image synthesis and editing with stochastic differential equations," in *Proc. Int. Conf. Learn. Represent*, 2022, pp. 1–33.

[7] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, "ILVR: Conditioning method for denoising diffusion probabilistic models," in *Proc. IEEE/CVF ICCV*, Aug. 2021, pp. 14347–14356.

[8] L. Zheng, J. Yuan, L. Yu, and L. Kong, "A reparameterized discrete diffusion model for text generation," 2023, *arXiv:2302.05737*.

[9] S. Yu, K. Sohn, S. Kim, and J. Shin, "Video probabilistic diffusion models in projected latent space," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 18456–18466.

[10] D. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014, pp. 1–14.

[11] I. J. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 2672–2680.

[12] D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proc. 32nd Int. Conf. Int. Conf. Mach. Learn.*, pp. 1530–1538, 2015.

[13] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[14] A. Alsaiari, R. Rustagi, A. Alhakamy, M. M. Thomas, and A. G. Forbes, "Image denoising using a generative adversarial network," in *Proc. IEEE 2nd Int. Conf. Inf. Comput. Technol. (ICICT)*, Mar. 2019, pp. 126–132.

[15] X. Wang, J. Wang, and B. Li, "Low dose ct image denoising method based on improved generative adversarial network," in *Proc. 7th Int. Conf. Autom., Control Robot. Eng. (CACRE)*, 2022, pp. 199–203.

[16] A. Ajay, Y. Du, A. Gupta, J. B. Tenenbaum, T. S. Jaakkola, and P. Agrawal, "Is conditional generative modeling all you need for decision making?" in *Proc. 11th Int. Conf. Learn. Represent.*, 2023, pp. 1–24.

[17] Z. Zhu et al., "MADiff: Offline multi-agent learning with diffusion models," 2023, *arXiv:2305.17330*.

[18] H. Du, J. Wang, D. Niyato, J. Kang, Z. Xiong, and D. I. Kim, "AI-generated incentive mechanism and full-duplex semantic communications for information sharing," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 9, pp. 2981–2997, Sep. 2023.

[19] M. Kim, R. Fritschek, and R. F. Schaefer, "Learning end-to-end channel coding with diffusion models," in *Proc. 26th Int. ITG Workshop Smart Antennas 13th Conf. Syst. Commun. Coding (WSA SCC)*, 2023, pp. 1–6.

[20] Y. Choukroun and L. Wolf, "Denoising diffusion error correction codes," in *Proc. 11th Int. Conf. Learn. Represent.*, 2023, pp. 1–15.

[21] E. Grassucci, S. Barbarossa, and D. Comminiello, "Generative semantic communication: Diffusion models beyond bit recovery," 2023, *arXiv:2306.04321*.

[22] J. Chen, D. You, D. Gündüz, and P. Luigi Dragotti, "CommIN: Semantic image communications as an inverse problem with INN-guided diffusion models," 2023, *arXiv:2310.01130*.

[23] S. F. Yilmaz, X. Niu, B. Bai, W. Han, L. Deng, and D. Gunduz, "High perceptual quality wireless image delivery with denoising diffusion models," 2023, *arXiv:2309.15889*.

[24] E. Grassucci, C. Marinoni, A. Rodriguez, and D. Comminiello, "Diffusion models for audio semantic communication," 2023, *arXiv:2309.07195*.

[25] Q. Lan et al., "What is semantic communication? A view on conveying meaning in the era of machine intelligence," *J. Commun. Inf. Netw.*, vol. 6, no. 4, pp. 336–371, 2021.

[26] J. Choi and J. Park, "Semantic communication as a signaling game with correlated knowledge bases," in *Proc. IEEE 96th Veh. Technol. Conf.*, Sep. 2022, pp. 1–5.

[27] W. Yang et al., "Semantic communications for future internet: Fundamentals, applications, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 213–250, 1st Quart., 2023.

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

[29] M. Fresia, F. Perez-Cruz, H. V. Poor, and S. Verdu, "Joint source and channel coding," *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 104–113, Nov. 2010.

[30] A. Guyader, E. Fabre, C. Guillemot, and M. Robert, "Joint source-channel turbo decoding of entropy-coded sources," *IEEE J. Sel. Areas Commun.*, vol. 19, no. 9, pp. 1680–1696, 2001.

[31] C. Chen, L. Wang, and F. C. M. Lau, "Joint optimization of protograph LDPC code pair for joint source and channel coding," *IEEE Trans. Commun.*, vol. 66, no. 8, pp. 3255–3267, Aug. 2018.

[32] E. Bourtsoulatze, D. Burth Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Trans. Cognit. Commun. Netw.*, vol. 5, no. 3, pp. 567–579, Sep. 2019.

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

[34] J. Xu, T.-Y. Tung, B. Ai, W. Chen, Y. Sun, and D. Gunduz, "Deep joint source-channel coding for semantic communications," 2022, *arXiv:2211.08747*.

[35] J. Dai et al., "Nonlinear transform source-channel coding for semantic communications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 8, pp. 2300–2316, Aug. 2022.

[36] K. P. Yang, S. Wang, J. Dai, K. Tan, K. Niu, and P. Zhang, "WITT: A wireless image transmission transformer for semantic communications," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[37] Y. Bo, Y. Duan, S. Shao, and M. Tao, "Joint coding-modulation for digital semantic communications via variational autoencoder," *IEEE Trans. Commun.*, early access, 2024, doi: 10.1109/TCOMM.2024.3386577.

[38] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[39] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, 2003, pp. 1398–1402.

[40] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2015, pp. 234–241.

[41] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 1–15.

[42] A. Vaswani, N. Shazeer, and N. Parmar, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[43] A. Krizhevsky, "Learning multiple layers of features from tiny images," M.S. thesis, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 2009.

[44] R. Timofte et al., "NTIRE 2017 challenge on single image super-resolution: Methods and results," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1110–1121.

[45] *Frame Stucture Channel Coding and Modulation for the Second Generation Digital Terrestrial Television Broadcasting System (DVB-T2)*, DVB document A122, 2008.

[46] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, M. Meila and T. Zhang, Eds. Jul. 2021, pp. 8162–8171. [Online]. Available: https://proceedings.mlr.press/v139/nichol21a.html

[47] T. Chen, "On the importance of noise scheduling for diffusion models," 2023, *arXiv:2301.10972*.

[48] K. Diederik and B. Jimmy, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[49] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–16.

Tong Wu received the B.S. degree in electronic and information engineering from Dalian University of Technology, Dalian, China, in 2022. He is currently pursuing the Ph.D. degree with Shanghai Jiao Tong University. His research interests include semantic communications and machine learning for wireless networks.

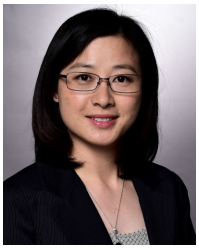
Zhiyong Chen (Senior Member, IEEE) received the Ph.D. degree from the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2011. From 2009 to 2011, he was a Visiting Ph.D. Student with the Department of Electronic Engineering, University of Washington, Seattle, USA. He is currently a Professor with the Cooperative Medianet Innovation Center, Shanghai Jiao Tong University (SJTU), Shanghai, China. His research interests include mobile communications-computing-caching (3C) networks and mobile AI systems. He served as the Student Volunteer Chair for the IEEE ICC 2019, the Publicity Chair for the IEEE/CIC ICCC 2014, and a TPC member for major international conferences. He was a recipient of the IEEE Asia–Pacific Outstanding Paper Award in 2019.

Dazhi He (Member, IEEE) is currently a Professor with the Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, and an Associate Editor of IEEE TRANSACTIONS ON BROADCASTING. He has had in-depth research experience for the technology and standardization of terrestrial digital TV in China (DTMB), direct satellite broadcasting in China (ABS-S), and the newest global digital TV standard (ATSC3.0). He has published more than 50 articles in the IEEE journals and applied more than 70 patents (including 20 PCT patents). His current research interests include 5G broadcasting, media communications, and heterogeneous networks. He ever won Second Prize of National Scientific and Technological Progress Award in China, First Prize of the Technology Innovation Award of National Radio and Television Administration in China, and First Prize of Science and Technology Progress Award in Shanghai.

Liang Qian received the B.S. degree in information and communication engineering from Shanghai Jiao Tong University (SJTU), Shanghai, China.

He is currently a Visiting Scholar with the Institute of Information Processing, Karlsruhe University, Germany. He is also the Executive Vice Director of the Wireless Communication Research Institute, Shanghai Jiao Tong University. He has published 37 domestic articles, applied for 21 patents, served as a TPC judge for five IEEE international conferences, and authored two professional works. His research interests include civil and military application, including the IoT, information processing, internet connected optical information processing, development of mobile intelligent terminal applications, key technologies and equipment for industrial informatization and supply chain finance, and other information interdisciplinary sciences. In terms of theoretical research, he works with IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, and IEE EL; and others have published 21 English articles with the highest impact factor of IF=11. He has won the First Prize of Shanghai Natural Science Progress Award, the Second Prize of Shanghai Marine Science and Technology Progress Award, and the Third Prize of China Shipbuilding Science and Technology Progress Award. He has also been awarded the title of Excellent Level Talent in the "Tianfu Talent Plan" in Sichuan Province and the title of "Ganjiang Innovative Talent" in the Ganjiang National New Area of Jiangxi Province.

Yin Xu (Member, IEEE) received the B.Sc. degree in information science and engineering from Southeast University, China, in 2009, and the master's and Ph.D. degrees in electronics engineering from Shanghai Jiao Tong University in 2011 and 2015, respectively. He is currently an Associate Professor and a Ph.D. Supervisor. His research interests include key technologies, prototype system development, and standardization of various communication systems, including 5G, broadcast, and short-range communication. His contributions have been incorporated into several domestic and international standards, including 3GPP 5G, ATSC3.0, HINOC2.0, and SparkLink. He serves as a TPC member, the co-chair, the session chair, and a keynote speaker of different major international conferences. Furthermore, he serves as a Guest Editor for IEEE TRANSACTIONS ON BROADCASTING.



Meixia Tao (Fellow, IEEE) received the B.S. degree in electronic engineering from Fudan University, Shanghai, China, in 1999, and the Ph.D. degree in electrical and electronic engineering from The Hong Kong University of Science and Technology in 2003. She is currently a Professor with the Department of Electronic Engineering, Shanghai Jiao Tong University, China. Her current research interests include wireless edge learning, coded caching, reconfigurable intelligence surfaces, and semantic communications.

She received the 2019 IEEE Marconi Prize Paper Award, the 2013 IEEE Heinrich Hertz Award for Best Communications Letters, the IEEE/CIC International Conference on Communications in China (ICCC) 2015 Best Paper Award, and the International Conference on Wireless Communications and Signal Processing (WCSP) 2012 and 2022 Best Paper Awards. She also received the 2009 IEEE ComSoc Asia–Pacific Outstanding Young Researcher Award. She is an Associate Editor of IEEE TRANSACTIONS ON INFORMATION THEORY and an Editor-at-Large of the IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY. She served as a member of the Executive Editorial Committee of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2015 to 2019. She was also on the editorial board of several other journals as an Editor or a Guest Editor, including IEEE TRANSACTIONS ON COMMUNICATIONS and IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS. She also served as the TPC Co-Chair for IEEE ICC 2023.

Wenjun Zhang (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 1984, 1987, and 1989, respectively. After three years' working as an Engineer with Philips, Nuremberg, Germany, he went back to his Alma Mater in 1993 and became a Full Professor in electronic engineering in 1995. He was one of the main contributors of the Chinese DTTB Standard (DTMB) issued in 2006. He holds 238 patents and published more than 130 papers in international journals and conferences. He is the Chief Scientist of Chinese Digital TV Engineering Research Centre (NERC-DTV), an industry/government consortium in DTV technology research and standardization, and the Director of the Cooperative Media Network Innovation Centre (CMIC), an excellence research cluster affirmed by Chinese Government. His main research interests include video coding and wireless transmission, multimedia semantic analysis, and broadcast/broadband network convergence.