Learned Image Transmission with Hierarchical Variational Autoencoder

Guangyi Zhang*, Hanlei Li*, Yunlong Cai, Qiyu Hu, Guanding Yu, Runmin Zhang

College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China
{zhangguangyi, hanleili, ylcai, qiyhu, yuguanding, runmin_zhang}@zju.edu.cn

Abstract

In this paper, we introduce an innovative hierarchical joint source-channel coding (HJSCC) framework for image transmission, utilizing a hierarchical variational autoencoder (VAE). Our approach leverages a combination of bottom-up and top-down paths at the transmitter to autoregressively generate multiple hierarchical representations of the original image. These representations are then directly mapped to channel symbols for transmission by the JSCC encoder. We extend this framework to scenarios with a feedback link, modeling transmission over a noisy channel as a probabilistic sampling process and deriving a novel generative formulation for JSCC with feedback. Compared with existing approaches, our proposed HJSCC provides enhanced adaptability by dynamically adjusting transmission bandwidth, encoding these representations into varying amounts of channel symbols. Additionally, we introduce a rate attention module to guide the JSCC encoder in optimizing its encoding strategy based on prior information. Extensive experiments on images of varying resolutions demonstrate that our proposed model outperforms existing baselines in rate-distortion performance and maintains robustness against channel noise.

Introduction

To meet the transmission requirements of heavy data traffic in future sixth-generation (6G) networks, wireless edge devices need to be equipped with higher transmission efficiency. Most contemporary systems employ a two-step strategy for data transmission: first, the raw data is compressed using a source codec, such as JPEG (Wallace 1992) and BPG (Lainema et al. 2012). Then, the encoded bits are protected with redundancy introduced by a carefully designed channel codec, such as LDPC and Polar codes (Arikan 2009). However, in many practical applications, the bit length is generally finite, making it impossible to guarantee optimality. In this context, joint source-channel coding (JSCC) has emerged as a potential solution, offering higher coding gains than the traditional separation-based coding paradigm.

With the revolutionary progress of deep learning in various fields, such as image compression (Ballé et al. 2018; He et al. 2021; Xu et al. 2022b) and generative models

(Kingma and Welling 2022; Razavi, van den Oord, and Vinyals 2019), a novel design paradigm for JSCC, called learned image transmission (LIT), has been conceived by formulating the communication pipeline as an end-to-end deep learning model (Bourtsoulatze et al. 2019; Zhang et al. 2024; Sun et al. 2023). Specifically, these methods leverage powerful neural networks to implement the encoding and decoding processes. In this approach, the whole system is viewed as an autoencoder (AE), which can be jointly learned in a data-driven manner. A notable method proposed by (Bourtsoulatze et al. 2019) employed CNNs to construct the source and channel codecs for wireless image transmission, achieving great performance by mapping the input image directly into channel symbols. Moreover, Kurka and Gündüz; Wu et al. investigate the JSCC with feedback. In this context, the transmission of these representations is divided into multiple phases, with the transmitter receiving the channel symbol vector after each phase, which simplifies the encoding process and improves the overall performance.

Beyond deterministic AEs, some studies have employed variational autoencoders (VAE) to design JSCC systems (Choi et al. 2019; Bo et al. 2024; Hu et al. 2023), where the channel symbols are generated through sampling. Part of these VAE-based methods show superior performance compared to deterministic AE-based methods, particularly under severe channel conditions. Though VAE-based methods have demonstrated remarkable performance, they experience significant performance degradation on high-resolution images. Furthermore, most existing methods only support fixed-rate coding, which contrasts with emerging works on transform coding-based image compression (Ballé et al. 2018), where the compression rate for each image is determined by the estimated entropy of its feature representation and varies with different samples. Consequently, these methods are less flexible and adaptive, potentially leading to performance penalties.

In this work, we aim to overcome the limitations of previous methods while enhancing performance. Specifically, we develop a hierarchical JSCC (HJSCC) framework based on a powerful hierarchical VAE architecture (Child 2020). Our transmitter employs both bottom-up and top-down paths to autoregressively generate multiple hierarchical representations of the original image. These representations are then mapped to channel symbols using multiple JSCC encoder

---

*These authors contributed equally.

blocks. Building upon this, we further explore the application of HJSCC in a classical scenario where a feedback link exists. By modeling transmission over a noisy channel as a probabilistic sampling process, we derive a novel generative formulation for JSCC with feedback, which achieves significantly better performance than most existing advanced schemes. While there have been attempts at variable-rate transmission (Dai et al. 2022; Song et al. 2023; Zhang et al. 2023; Yang and Kim 2022) in the realm of JSCC without feedback, the problem of rate-adaptive design for JSCC with feedback remains underexplored. Unlike existing works (Kurka and Gündüz 2020; Wu et al. 2024a; Li et al. 2024; Jiang et al. 2022), we leverage the prior distribution (which characterizes the entropy information) of each representation to generate masks that control the number of symbols for each representation. This approach allows us to dynamically adjust the transmission rate. Additionally, we introduce a rate attention module to guide the JSCC encoder in adjusting the encoding strategy according to its prior information.

In summary, our contributions are as follows:

- **HJSCC Framework**: Developing a hierarchical scheme that is able to support the transmission of high-resolution images.

- **HJSCC with Feedback**: Extending HJSCC to the case with feedback, by viewing the transmission as a sampling process and deriving a generative formulation.

- **Dynamic Rate Control:** By utilizing the entropy information of representations to dynamically control the transmission rate, this approach bridges the gap of lacking rate-adaptive design when a feedback link is present.

- **Rate Attention Module:** Proposing a spatial grouping strategy and a rate attention module to improve the overall rate-distortion performance.

- **Experimental Studies:** Providing substantial experiments to verify the effectiveness of the proposed method, demonstrating that the proposed scheme achieves better coding gain than emerging deep learning-based JSCC and separation-based digital transmission schemes.

Related Works

Varational Autoencoder VAE can be employed as deep generative models capable of generating high-dimensional data based on a low-dimensional latent space, and is a variant of autoencoder (Sø nderby et al. 2016; Cemgil et al. 2020). By sampling from the learned latent space distribution and passing these samples through the decoder network, VAE can generate new data points that resemble the training data. However, the original VAE is known to perform worse than many other generative models, particularly when applied to high-resolution images. (Vahdat and Kautz 2020; Child 2020; Yuhta Takida 2024) addressed this by proposing a deep hierarchical VAE, where the latent variable is divided into several disjoint groups, achieving significantly better performance than standard VAEs. More recently, VAE has been applied to compression tasks (Duan et al. 2023; Lu et al. 2024; Townsend, Bird, and Barber 2019; Kingma, Abbeel, and Ho 2019).
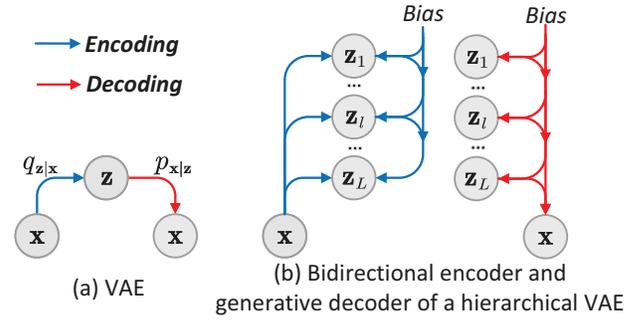

Figure 1: Probabilistic model of VAEs and hierarchical ResNet VAE. The bias is a trainable parameter.

Learned Image Transmission Unlike the separation-based design described above, recent studies have delved into the utilization of AE and its variants, e.g., VAE to design wireless image transmission systems, resulting in a number of efficient methods (Bourtsoulatze et al. 2019; Xu et al. 2022a; Zhang et al. 2024; Sun et al. 2023; Xu et al. 2022a; Yang et al. 2024; Saidutta, Abdi, and Fekri 2019). In particular, (Bourtsoulatze et al. 2019; Wu et al. 2024b) and (Yang et al. 2024) conceived of using neural networks to simultaneously finish the source encoding/decoding and channel coding/decoding, with the goal of jointly optimizing the entire system to maximize PSNR. The VAE-based methods adopted probabilistic modeling, wherein the encoding process is characterized as a stochastic procedure. In these systems, channel symbols are generated by sampling from the probability distribution conditioned on the input image (Bo et al. 2024; Saidutta, Abdi, and Fekri 2019). These approaches have demonstrated superior performance, especially under severe transmission conditions.

Proposed Methods

Background

VAE and Hierarchical VAE As stated in (Cemgil et al. 2020), VAE is a stochastic variational inference scheme that can be applied to various intelligent tasks, such as recognition, denoising, and generation. As shown in Fig. 1(a), to formulate a vision-related model, we typically start with the following premises. Let $\mathbf{x}$ denote an image intensity vector, which is drawn from a dataset $\mathcal{X}$ with distribution $p_{\mathbf{x}}$. Another variable is the latent variable $\mathbf{z}$, with a prior $p_{\mathbf{z}}$. The main target of a VAE is to learn a generative model (decoder) $p_{\mathbf{x}|\mathbf{z}}$ for sampling, and a posterior density model (encoder) $q_{\mathbf{z}|\mathbf{x}}$ for variational inference. The objective for learning a VAE model can be formulated as minimizing the (variational) upper bound on the marginal likelihood of a batch of data points, as given by

$$\mathcal{L}\left(\boldsymbol{\theta}, \boldsymbol{\phi}\right) = \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}}, \mathbf{z}\sim q_{\mathbf{z}|\mathbf{x}}}\left[D_{KL}\left(q_{\mathbf{z}|\mathbf{x}}\|p_{\mathbf{z}}\right) - \log p_{\mathbf{x}|\mathbf{z}}\right], \text{ (1)}$$

where $D_{KL}\left(q_{\mathbf{z}|\mathbf{x}}\|p_{\mathbf{z}}\right) = \log\left(q_{\mathbf{z}|\mathbf{x}}/p_{\mathbf{z}}\right)$, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ represent the parameters of the encoder $q_{\mathbf{z}|\mathbf{x}}$ and decoder $p_{\mathbf{x}|\mathbf{z}}$, respectively.

Hierarchical VAEs are a series of VAE models that partition the latent variables into several disjoint groups. The

Figure 2: Diagram of a deep learning-based JSCC system.

probabilistic diagram of a classical hierarchical VAE, the ResNet VAE, is shown in Fig. 1(b), consisting of a bidirectional encoder and a generative decoder. Specifically, the latent variables can be denoted by $\mathbf{z} \triangleq \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_L\}$, where L represents the number of groups. The prior for $\mathbf{z}$ is modeled as $p_{\mathbf{z}_{1:L}} = \prod_l p_{\mathbf{z}_l | \mathbf{z}_{<l}}$, and the approximate posterior is denoted as $q_{\mathbf{z}} = \prod_l q_{\mathbf{z}_l | \mathbf{z}_{<l}, \mathbf{x}}$, where $\mathbf{z}_{<l}$ represents $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{l-1}\}$. In general, the dimension of $\mathbf{z}_l$ is designed to be smaller than that of $\mathbf{z}_{l+1}$, fulfilling the target of capturing the coarse-to-fine nature of images. The objective for training a hierarchical VAE mode can be obtained by extending Eq. (1) for multiple latent variables, as given by

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) =$$
$$\mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}, \mathbf{z} \sim q_{\mathbf{z}|\mathbf{x}}} \left[\sum_{l=1}^{L} D_{KL} \left(q_{\mathbf{z}_l | \mathbf{z}_{<l}, \mathbf{x}} \| p_{\mathbf{z}_l | \mathbf{z}_{<l}} \right) - \log p_{\mathbf{x}|\mathbf{z}} \right],$$
(2)

where we define $\mathbf{z}_{<1}$ as an empty set, and thus $p_{\mathbf{z}_1 | \mathbf{z}_{<1}} = p_{\mathbf{z}_1}$ and $q_{\mathbf{z}_1 | \mathbf{z}_{<1}, \mathbf{x}} = q_{\mathbf{z}_1 | \mathbf{x}}$.

Proposed HJSCC

System Overview Here, we aim to give a brief overview of deep learning-based JSCC. In particular, the model of a typical JSCC is shown in Fig. 2. The transmitter employs a JSCC encoder to map the input image $\mathbf{x} \in \mathbb{R}^N$ directly into channel symbol vector $\mathbf{s} \in \mathbb{C}^K$ for transmission, where N denotes the number of pixels and K represents the number of channel symbols. This process can be expressed as $\mathbf{s} = g_e(\mathbf{x}; \boldsymbol{\theta}_e)$, where g_e signifies the encoding function and $\boldsymbol{\theta}_e$ represents its trainable parameters. Accounting for the limited transmission power, the transmitted signal should satisfy the power constraint P, implying that $\|\mathbf{s}\|_2^2 / K \leq P$. Subsequently, the channel symbol vector is transmitted through the wireless channel, as given by $\tilde{\mathbf{s}} = \mathbf{s} + \mathbf{n}$, where $\mathbf{n} \sim \mathcal{N}(0, \sigma_n^2 \mathbf{I})$ is additive white Gaussian noise (AWGN). At the receiver, the received noisy signal $\tilde{\mathbf{s}}$ is processed by a decoder g_d to obtain the reconstructed image $\bar{\mathbf{x}} = g_d(\tilde{\mathbf{s}}; \boldsymbol{\phi}_d)$, with $\boldsymbol{\phi}_d$ denoting the trainable parameters of the JSCC decoder. The optimization objective of a deep learning-based JSCC system is to minimize the difference between $\mathbf{x}$ and $\bar{\mathbf{x}}$, and thus mean-square error (MSE) can be employed as the loss function. Furthermore, to evaluate the performance of a JSCC system and ensure fairness, we define the **signal-to-noise ratio (SNR)** as SNR $= 10 \log \frac{P}{\sigma^2}$(dB), which characterizes the channel quality of the system. Then, we introduce the **channel bandwidth ratio (CBR)** to describe the **transmission rate** (overhead), which is expressed as CBR $= K/N$. Intuitively, CBR actually signifies the number of symbols for transmitting one pixel, and *a higher CBR brings a higher overhead, while usually resulting in a better system performance.*

While previous studies have achieved excellent rate-distortion performance based on the framework depicted in Fig. 2, it is evident that the transmission rate is solely determined by the image resolution. This limitation can result in performance degradation in overall rate-distortion due to the inability to adaptively adjust the rate for each image (He et al. 2022). To address this issue, we propose our HJSCC framework, as illustrated in Fig. 3. Given an image $\mathbf{x}$, the bottom-up path generates a set of latent features, which are subsequently passed to the top-down path to autoregressively generate the latent representations $\boldsymbol{\mu} \triangleq \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_L\}$. These representations are then fed to a set of JSCC encoders $g_e = \{g_e^1, g_e^2, \ldots, g_e^L\}$, respectively. In this way, we are able to obtain a set of channel symbol vectors $\mathbf{s} \triangleq \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_L\}$, where $\mathbf{s}_l = g_e^l(\mathbf{z}_l)$. Then, these channel symbol vectors are transmitted to the receiver through the wireless link, and the received symbol vectors are represented by $\tilde{\mathbf{s}} \triangleq \{\tilde{\mathbf{s}}_1, \tilde{\mathbf{s}}_2, \ldots, \tilde{\mathbf{s}}_L\}$. At the receiver, the noisy $\tilde{\mathbf{s}}$ undergoes processing by the JSCC decoder g_d^l, and we obtain $\tilde{\boldsymbol{\mu}} \triangleq \{\tilde{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\mu}}_2, \ldots, \tilde{\boldsymbol{\mu}}_L\}$. With $\tilde{\boldsymbol{\mu}}$ at hand, the receiver can reconstruct the image using the top-down path (decoder), and the reconstructed image is denoted by $\hat{\mathbf{x}}_{\mathrm{H}}$. Our objective is to minimize the distortion between the transmitted image $\mathbf{x}$ and $\hat{\mathbf{x}}_{\mathrm{H}}$.

Training Objective Formulation

We aim to develop a rate-adaptive JSCC model capable of adjusting the transmission rate, CBR, based on the source content. To this end, we propose an inherited training strategy motivated by (Dai et al. 2022), where the objective of minimizing $d(\mathbf{x}, \hat{\mathbf{x}}_{\mathrm{H}})$ is guided by a learned image coder. Specifically, we have designs for the posteriors, priors, and training objectives as follows (Duan et al. 2023).

Posteriors: The posteriors $q_{\mathbf{z}_l | \mathbf{z}_{<l}, \mathbf{x}}$ is set to a uniform distribution

$$q_{\mathbf{z}_l | \mathbf{z}_{<l}, \mathbf{x}}(\mathbf{z}_l | \mathbf{z}_{<l}, \mathbf{x}) \triangleq \prod_i \mathcal{U} \left(\mu_l^{(i)} - \tfrac{1}{2}, \mu_l^{(i)} + \tfrac{1}{2} \right), \quad (3)$$

where $\mathcal{U}$ denotes a uniform distribution centered on $\mu_l^{(i)}$, and $\mu_l^{(i)}$ is the i-th element of parameter $\boldsymbol{\mu}_l$, which is obtained from the l-th posterior branch.

Priors: The conditional prior distribution $p_{\mathbf{z}_l | \mathbf{z}_{<l}}$ is defined to be a Gaussian distribution convolved with a uniform distribution (Ballé et al. 2018):

$$p_{\mathbf{z}_l | \mathbf{z}_{<l}}(\mathbf{z}_l | \mathbf{z}_{<l}) \triangleq \prod_i \mathcal{N} \left(\hat{\mu}_l^{(i)}, (\hat{\sigma}_l^{(i)})^2 \right) * \mathcal{U} \left(-\tfrac{1}{2}, \tfrac{1}{2} \right), \quad (4)$$

where $*$ represents the convolution operation.

Training Loss: With the posteriors and priors defined above, the loss function (2) for an image compression model (Duan et al. 2023) can be expressed as

$$\mathcal{L} = \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}, \mathbf{z} \sim q_{\mathbf{z}|\mathbf{x}}} \left[\sum_{l=1}^{L} \log \frac{1}{p_{\mathbf{z}_l | \mathbf{z}_{<l}}(\mathbf{z}_l | \mathbf{z}_{<l})} + \lambda \cdot d(\mathbf{x}, \hat{\mathbf{x}}) \right],$$
(5)

where λ is the introduced weight to control the tradeoff between rate (the first term) and distortion (the second term)
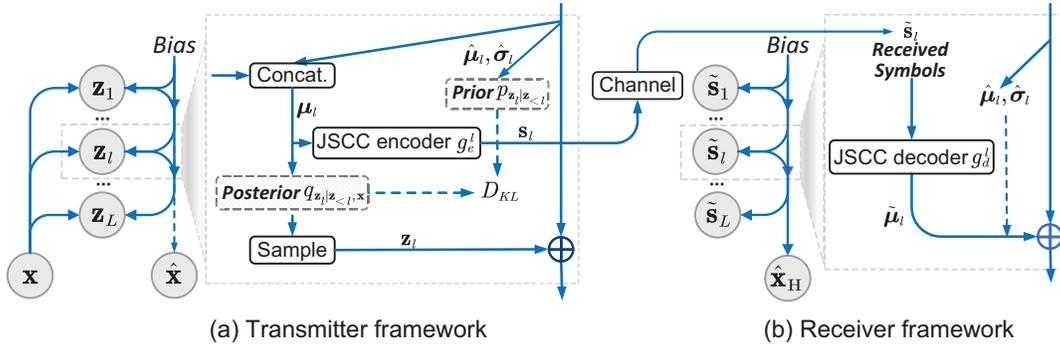
Figure 3: The probabilistic diagram of the proposed HJSCC. The transmitter employs the bottom-up and top-down paths for encoding, while the receiver reconstructs the image with the received symbols.
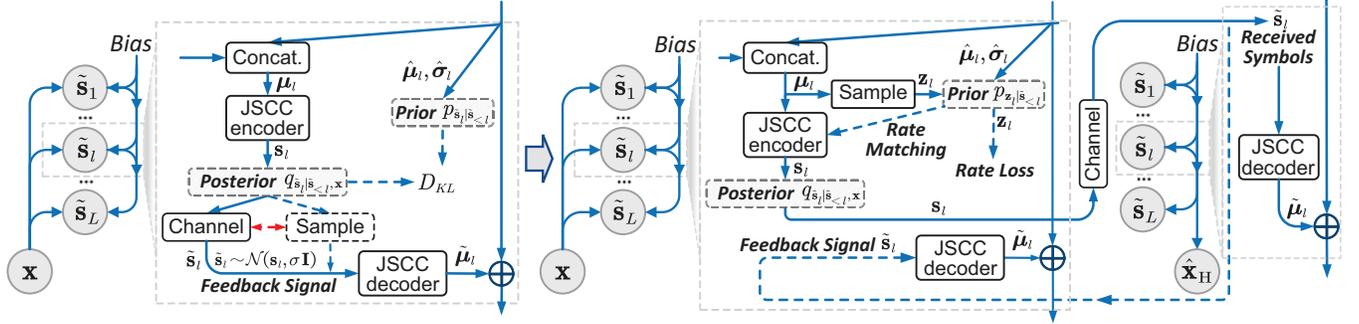

Figure 4: The probabilistic diagram of the proposed HJSCC with feedback.

$d(\mathbf{x}, \hat{\mathbf{x}})$. Thus, (5) can be utilized for image compression to achieve a trade-off between rate and distortion. Under the guidance of (5), we further develop the optimization problem for HJSCC as follows:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}, \mathbf{z} \sim q_{\mathbf{z}|\mathbf{x}}} \Big[$$

$$\underbrace{\sum_{l=1}^{L} -\alpha \log p_{\mathbf{z}_l|\mathbf{z}_{<l}}(\mathbf{z}_l|\mathbf{z}_{<l})}_{\text{transmission rate}} + \lambda \cdot \underbrace{(d(\mathbf{x}, \hat{\mathbf{x}}) + d(\mathbf{x}, \hat{\mathbf{x}}_{\mathrm{H}}))}_{\text{weighted distortion}} \Big],$$

(6)

where the calculation ways of each variable can be found in Fig. 3. Intuitively, there are two main differences from the loss function (5). The first is the additional distortion term to optimize the transmission distortion $d(\mathbf{x}, \hat{\mathbf{x}}_{\mathrm{H}})$, which helps to obtain a robust representation against noise. The second is the scaling parameter α to control the relation between the entropy of the latent $\mathbf{z}_l$ and the transmission rate for $\mathbf{s}_l$.

To flexibly adjust the CBR for different images, a natural choice for reference is the prior of latent variable $\mathbf{z}_l$ that is correlated to the bit length after entropy coding, where we assume the entropy of $\mathbf{z}_l$ is positively related to the entropy of $\boldsymbol{\mu}_l$. Thus, the CBR for encoding $\boldsymbol{\mu}_l$ is designed to be proportional to the prior $p_{\mathbf{z}_l|\mathbf{z}_{<l}}$. Moreover, noting that the channel bandwidth is determined by reducing the number of the channel symbols, we achieve rate adjustment by masking the channel symbol vector according to the transmission rate $\alpha p_{\mathbf{z}_l|\mathbf{z}_{<l}}$, which will be introduced in the following sections.

HJSCC with Feedback

Building upon previous designs, we further investigate the JSCC scenario with feedback link from the receiver to the transmitter. In this scenario, image transmission is divided into multiple phases, allowing the transmitter to use the received channel symbol vectors from previous phases when encoding channel symbols in the current phase.

As shown in the left framework of Fig. 4, we formulate our HJSCC with feedback by viewing the received symbols vector $\tilde{\mathbf{s}}_l$ as the latent variable. In this case, the loss function for training HJSCC with feedback can be written as

$$\mathcal{L} = \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}, \tilde{\mathbf{s}} \sim q_{\tilde{\mathbf{s}}|\mathbf{x}}} \left[\sum_{l=1}^{L} \log \frac{q_{\tilde{\mathbf{s}}_l|\tilde{\mathbf{s}}_{<l}, \mathbf{x}}(\tilde{\mathbf{s}}_l|\tilde{\mathbf{s}}_{<l}, \mathbf{x})}{p_{\tilde{\mathbf{s}}_l|\tilde{\mathbf{s}}_{<l}}(\tilde{\mathbf{s}}_l|\tilde{\mathbf{s}}_{<l})} - \log p_{\mathbf{x}|\tilde{\mathbf{s}}} \right].$$

(7)

Intuitively, we view the transmission over noisy channels as a process of sampling. In particular, since we consider AWGN channels, i.e., $\tilde{\mathbf{s}}_l = \mathbf{s}_l + \mathbf{n}_l$, the posteriors $q_{\mathbf{s}_l|\mathbf{s}_{<l}, \mathbf{x}}$ is actually a Gaussian distribution

$$q_{\tilde{\mathbf{s}}_l|\tilde{\mathbf{s}}_{<l}, \mathbf{x}}(\tilde{\mathbf{s}}_l|\tilde{\mathbf{s}}_{<l}, \mathbf{x}) \triangleq \prod_i \mathcal{N}\left(s_l^{(i)}, \sigma_n^2\right),$$

(8)

where $\mathbf{s}_l$ can be directly computed with known $\mathbf{x}$ and $\tilde{\mathbf{s}}_{<l}$. Besides, with this formulation, the generation of the l-th channel symbol vector $\tilde{\mathbf{s}}_l$ is conditioned on the received vectors in the former phases. It enables the HJSCC to adjust the transmitted symbols based on feedback signals. In this way,
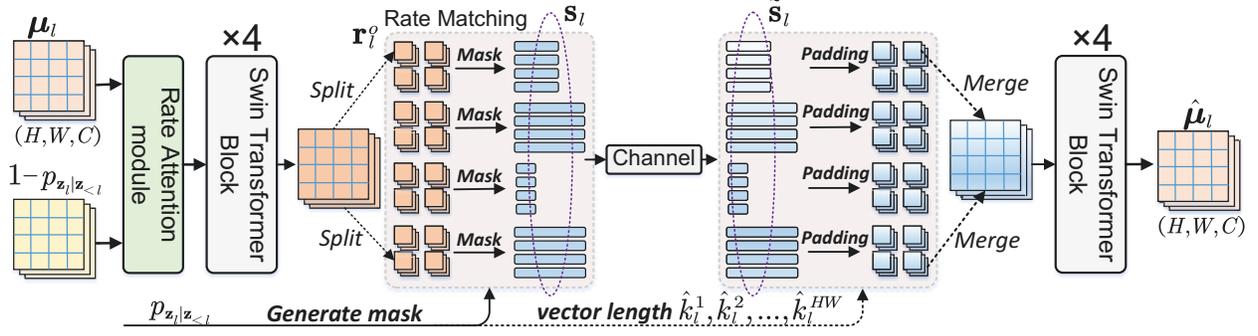
Figure 5: The illustration of the process of the JSCC encoder and JSCC decoder for transmitting latent representation $\boldsymbol{\mu}_l$.

we also fortunately find that the term $\log q_{\tilde{\mathbf{s}}_l|\tilde{\mathbf{s}}_{<l},\mathbf{x}}(\tilde{\mathbf{s}}_l|\tilde{\mathbf{s}}_{<l},\mathbf{x})$ in (7) is only related to the channel noise and will not introduce gradient to the whole model. Thus, this term can be directly dropped from the loss function.

We aim to achieve rate-adaptive transmission for HJSCC in the presence of feedback link. Similar to the scenario without feedback, we are able to take the prior information $p_{\tilde{\mathbf{s}}_l|\tilde{\mathbf{s}}_{<l}}(\tilde{\mathbf{s}}_l|\tilde{\mathbf{s}}_{<l})$ as a rate indicator. However, the calculation of this prior depends on the values of $\tilde{\mathbf{s}}_l$, which is unknown before transmission. As shown in the right of Fig. 4, we address this by proposing a new prior $p_{\mathbf{z}_l|\tilde{\mathbf{s}}_{<l}}(\mathbf{z}_l|\tilde{\mathbf{s}}_{<l})$ as a substitution, where $\mathbf{z}_l$ is sampled from a uniform distribution centered on $\boldsymbol{\mu}_l$. This is inspired by the fact that $p_{\mathbf{z}_l|\tilde{\mathbf{s}}_{<l}}(\mathbf{z}_l|\tilde{\mathbf{s}}_{<l})$ is positively related to $p_{\tilde{\mathbf{s}}_l|\tilde{\mathbf{s}}_{<l}}(\tilde{\mathbf{s}}_l|\tilde{\mathbf{s}}_{<l})$, and thus can be employed as the rate term when training the model. As a result, the loss function for training the HJSCC with feedback can be written as

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}},\tilde{\mathbf{s}}\sim q_{\tilde{\mathbf{s}}|\mathbf{x}}}\left[\sum_{l=1}^{L}\log\frac{\text{const}}{p_{\mathbf{z}_l|\tilde{\mathbf{s}}_{<l}}\cdot\beta}+\lambda d(\mathbf{x},\hat{\mathbf{x}}_{\mathrm{H}}))\right],$$

(9)

where β denotes the scaling factor.

Masking and Length Information Reduction

A visualized example of our proposed masking strategy is shown in Fig. 5. Particularly, we employ the Swin Transformer blocks to implement our JSCC encoder, since we find that this architecture presents better robustness against channel noise. In this way, the shape of the output feature at the l-th layer, $\mathbf{r}_l$, can be denoted as $C\times H\times W$, which is the same as that of $\boldsymbol{\mu}_l$. We split the output of the Swin Transformer blocks, $\mathbf{r}_l$, into HW sequences $\mathbf{r}_l^o$, for $o = 1, 2, \ldots, HW$, where the length of each $\mathbf{r}_l^o$ is C. Then, we include a rate-matching layer after the Swin Transformer blocks. It accepts two inputs, $\mathbf{r}_l$ and $\alpha p_{\mathbf{z}_l|\mathbf{z}_{<l}}$, and generates the masked channel symbol vector $\mathbf{s}_l$. Specifically, the length for each vector $\mathbf{s}_l^o$ will be constrained to $k_l^o = \sum_{c=1}^{C}-\alpha\log p_{z_l^{(o,c)}|\mathbf{z}_{<l}}(z_l^{(o,c)}|\mathbf{z}_{<l})$, which actually represents the summation of entropy along all the C dimensions of $\mathbf{z}_l^o$. We adjust the length by generating a mask vector $\mathbf{m}_l^o$, which can be written as

$$\mathbf{m}_l^o = [\underbrace{1, 1, \ldots, 1}_{k_l^o}, 0, \ldots, 0],$$

(10)

indicating that only the former k_l^o elements of $\mathbf{r}_l^o$ will be used as the channel symbols, i.e., $\mathbf{s}_l^o = \mathbf{r}_l^o \odot \mathbf{m}_l^o$, where $\odot$ denotes the element-wise multiplication.

Though this enables the transmitter to determine the transmission rate adaptively for each image, this design introduces a length-matching issue. Specifically, the receiver needs to know the length of each transmitted symbol vector to identify the channel symbols from different spatial positions and layers. This requirement, however, adds an overhead of communicating this information to the receiver. To mitigate this overhead, we propose two practical designs:

Firstly, instead of considering infinite precision, we opt for a finite set of length options, comprising $\{2^{N_q}\}$ integers, where N_q is a selected integer. We define the optional length set as $\mathcal{Q} = \{q_1, q_2, \ldots, q_{2^{N_q}}\}$. Then, for each $\mathbf{s}_l^o$, the transmission rate is quantized to $\hat{k}_l^o = Q(k_l^o) = Q(\sum_{c=1}^{C}-\alpha\log p_{z_l^{(o,c)}|\mathbf{z}_{<l}}(z_l^{(o,c)}|\mathbf{z}_{<l}))$ with the optional length set $\mathcal{Q}$. Therefore, we incorporate an extra link to transmit N_q bits as side information to inform the length for each $\mathbf{s}_l^o$, where we assume this information should be transmitted without errors. The additional overhead for each $\mathbf{s}_l^o$ is $\frac{\log_2 N_q}{C_r}$, where C_r denotes the channel capacity.

Secondly, for each $\mathbf{s}_l$ we need in total HWN_q extra bits as the side information. In comparison to the transmission overhead at the l-th layer, $\sum_{o=1}^{HW}\hat{k}_l^o$, this amount of side information will be quite significant when C is small. To address this issue, we design a spatial grouping strategy. As shown in the middle part of Fig. 5, we split the $\mathbf{r}_l$ into multiple patches along the width and height dimensions, where the channel symbol sequence at each patch is assigned with the same length value (number of symbols). This grouping strategy is inspired by the findings that the allocated rates within the kind of patch are rather similar, and thus it will not lose much flexibility even if we force the vectors within a patch to have the same length. With the spatial grouping, only one length scalar is required for multiple sequences in $\mathbf{r}_l$, and thus the overhead can be mitigated.

Furthermore, as the JSCC encoder is required to encode the images into symbols of different numbers, we devise a rate attention module by incorporating the prior information in the encoding process. The rate attention involves two inputs, the latent representation $\boldsymbol{\mu}_l$ of shape (H, W, C) and the
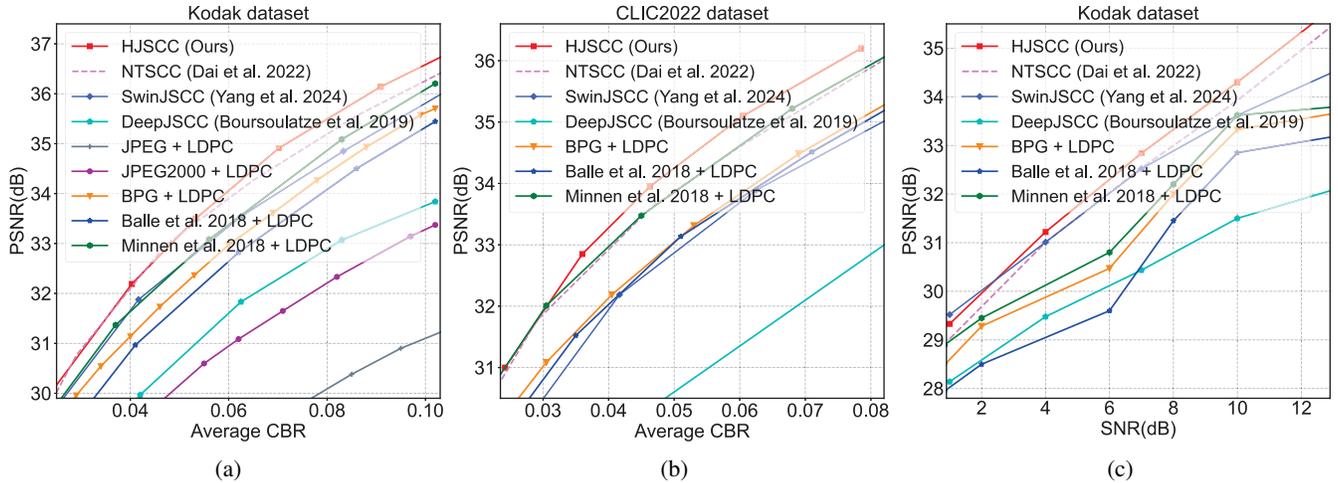
Figure 6: The end-to-end distortion performance versus over different datasets. The results are evaluated on (a) Kodak and (b) CLIC2022 datasets, at SNR = 10 dB. (c) The end-to-end distortion performance on Kodak dataset versus the SNR.

prior information $p_{\mathbf{z}_l|\mathbf{z}_{<l}}$. We calculate the length value for different vectors in $\boldsymbol{\mu}_l$, obtaining in total HW real length values. Then, we calculate the merged length values after the spatial grouping. As we allocate the same length value to the vectors within a patch, the merged value is calculated by averaging the real length values in a sample patch. The rate attention operation accepts two inputs. The first is the concatenated information from the representation and prior information, while the second is the concatenated matrix of the real length value vector and the merged length value vector. Through this operation, the index information can be fused to the encoding process, enabling the JSCC encoder block to adaptively adjust the encoding process.

Experiments

Metrics and Test Datasets For performance evaluation, we consider the pixel-wise peak signal-to-noise ratio (PSNR)[1]. We quantify the performance by considering the following datasets of different resolutions with necessary preprocessing. **Kodak** (Kodak 1993): The dataset consists of 24 images of resolution 512×768 or 768×512. **CLIC2022 Test** (Toderici et al. 2022): The test set contains 30 images up to size 1365×2048.

Benchmarks To verify the performance of our image transmission models, we compare them with a range of benchmarks. First, we consider emerging deep learning-based schemes, including DeepJSCC (Bourtsoulatze et al. 2019), SwinJSCC (Yang et al. 2024), and the nonlinear transform source-channel coding (NTSCC) (Dai et al. 2022). Second, we compare our methods with separation-based schemes widely used in real transmission systems. These include powerful image codecs such as BPG, JPEG, and JPEG2000, combined with a practical LDPC code (Shahid and Yahampath 2013), labeled as "BPG + LDPC", "JPEG

+ LDPC", and "JPEG2000 + LDPC", respectively. In addition to these hand-crafted image codecs, we also consider the learning-based image codecs combined with LDPC, (Ballé et al. 2018) and (Minnen, Ballé, and Toderici 2018). For the feedback JSCC schemes, we consider the most advanced JSCCformer-f (Wu et al. 2024a) and the classical DeepJSCC-f (Kurka and Gündüz 2020). In our experiments, we test various schemes across different CBRs and SNRs under AWGN channels. We take the architecture in (Duan et al. 2023) as the backbone model.

Comparisons and Results Analysis

In Fig. 6 (a) and (b), we evaluate the transmission performance at different CBRs under AWGN channels, with the test SNR set to 10 dB. To ensure a reliable transmission link, we adopt 16-order quadrature amplitude modulation (16QAM) combined with an LDPC rate of $2/3$, in accordance with the 3GPP standard. The results indicate that our proposed HJSCC significantly outperforms the fixed-rate schemes, DeepJSCC and SwinJSCC. Compared with NTSCC, the proposed scheme achieves comparable performance. The performance gap widens with increasing image resolution and CBR. Additionally, compared to hand-crafted schemes, the proposed method achieves substantially better PSNR performance. This demonstrates the potential of utilizing HJSCC in practical wireless communication systems.

Fig. 6 (c) demonstrates the transmission performance across varying channel SNR levels. To ensure fairness, the CBR for these schemes is constrained to 0.0625. For DeepJSCC and NTSCC, the training SNR equals the testing SNR to achieve optimal performance. For the separation-based methods, we test these schemes across different channel coding rates and modulation orders to determine the optimal settings[2], and then the image codec needs to compress the source with the resulted bits per pixel (bpp) value.

---

[1]Codes are given in *https://github.com/zhang-guangyi/HJSCC*.

[2]Given a bpp value, the CBR ρ can be calculated as $\rho =$
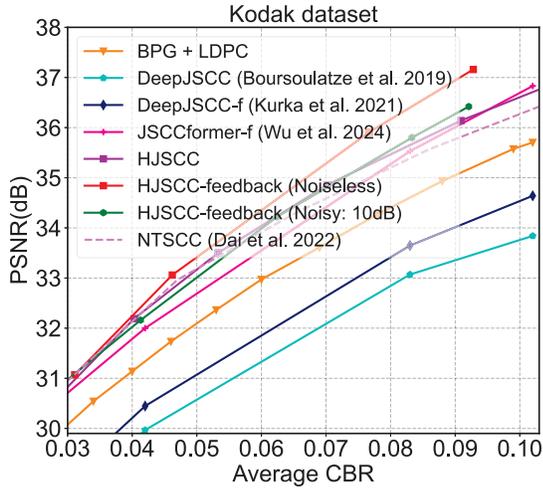
Figure 7: PSNR performance with feedback links.

	λ	CBR	CBR (s)	CBR ($\hat{k}$)	PSNR
Grouping	64	0.0403	0.0378	0.0025	32.03
No grouping	64	0.0590	0.0390	0.0200	32.24
Grouping	16	0.0249	0.0224	0.0025	29.96
No grouping	16	0.0430	0.0230	0.0200	30.13

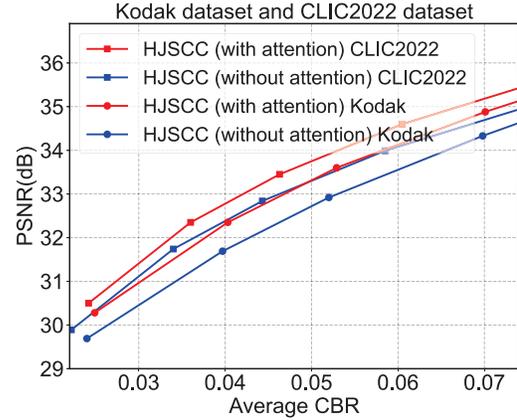Table 1: Ablation analysis on the spatial grouping strategy.

Figure 8: Ablation analysis on the rate attention module.

Our results indicate that the proposed HJSCC significantly outperforms other schemes. Additionally, HJSCC achieves a substantial performance gain compared to the emerging method (Minnen, Ballé, and Toderici 2018), with the performance gap becoming more pronounced at SNR levels above 10 dB. Furthermore, separation-based systems are known to suffer from the cliff effect, where reliable transmission cannot be maintained when the channel coding and modulation schemes fail. In contrast, the proposed HJSCC provides a graceful degradation as the SNR decreases, demonstrating its robustness in varying channel conditions.

Fig. 7 presents the PSNR performance of different schemes, where JSCCformer-f (Wu et al. 2024a) and DeepJSCC-f (Kurka and Gündüz 2020) are JSCC schemes with a feedback link. Compared with them, our proposed HJSCC shows much better PSNR performance, with a gain of about 1 dB at high CBR region. Besides, the introduction of the feedback link also provides significantly larger gain, especially at high average CBR values. This stems from rate-adaptive capability of HJSCC-feedback as well as its generative formulation, making it the state-of-the-art JSCC schemes in the presence of a feedback link. In addition, when the feedback is noisy, we can see a slight degradation in PSNR performance. Besides, the performance improves with the feedback link quality, where a higher feedback SNR can produce better PSNR performance.

Ablation Studies

Ablation on the spatial grouping. In this work, we propose a spatial grouping strategy to reduce the transmission overhead to inform the receiver of the vector length for rate matching. To show the effectiveness, we report the performance on CBR saving. Particularly, we compare the CBRs for transmitting the channel symbols **s** and the vector length

information $\hat{k}$ in Table 1. All the models are optimized on ImageNet dataset. From the results, the overall CBR can be significantly reduced by the spatial merging strategy, at the cost of a slight performance degradation.

Ablation on rate attention module. Moreover, we also verify the effect of the proposed rate attention module, as shown in Fig. 8. We compare the performance of using and not using this module on the Kodak and CLIC2022 datasets, over the AWGN channels, where the SNR is set to 10 dB. From the results, we find that the proposed rate attention module significantly improves the PSNR performance over different average CBR values. This performance gain stems from the ability of the rate attention module in guiding the JSCC encoder to encode the latent representation into channel symbol vectors of different length values, demonstrating the effectiveness of this module.

Conclusion & Discussion

In this work, we introduced a high-efficiency JSCC framework for wireless image transmission based on hierarchical VAE. Unlike conventional methods, our proposed scheme learns a hierarchical latent representation and employs multiple JSCC encoder/decoder pairs to transmit these latent representations. We formulate a novel generative formulation for HJSCC with feedback by viewing the transmission as a sampling process. By leveraging the learned prior in the JSCC encoder, our proposed HJSCC can dynamically adjust the transmission rate according to the data distribution, making it a rate-adaptive scheme compared to existing solutions.

$\frac{K}{C \times H \times W} = \frac{\text{bpp}}{C \times \log_2 M \times R_c} = \frac{\text{bpp}}{C \times \log_2 M \times R_c}$, where M is the selected modulation order and R_c denotes the channel coding rate. For example, $\rho = \text{bpp}/8$ when 16QAM and rate 2/3 are selected for SNR = 10 dB.

Acknowledgments

References

Arikan, E. 2009. Channel Polarization: A Method for Constructing Capacity-Achieving Codes for Symmetric Binary-Input Memoryless Channels. *IEEE Transactions on Information Theory*, 55(7): 3051–3073.

Ballé, J.; Minnen, D.; Singh, S.; Hwang, S. J.; and Johnston, N. 2018. Variational image compression with a scale hyperprior. arXiv:1802.01436.

Bo, Y.; Duan, Y.; Shao, S.; and Tao, M. 2024. Joint Coding-Modulation for Digital Semantic Communications via Variational Autoencoder. *IEEE Transactions on Communications*, 1–1.

Bourtsoulatze, E.; Burth Kurka, D.; Gunduz, D.; and Gunduz, D. 2019. Deep Joint Source-Channel Coding for Wireless Image Transmission. *IEEE Transactions on Cognitive Communications and Networking*, 5(3): 567–579.

Cemgil, T.; Ghaisas, S.; Dvijotham, K.; Gowal, S.; and Kohli, P. 2020. The Autoencoding Variational Autoencoder. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 15077–15087. Curran Associates, Inc.

Child, R. 2020. Deep VAEs Generalize Autoregressive Models and Can Outperform Them on Images. *arXiv preprint arXiv:2011.10650*.

Choi, K.; Tatwawadi, K.; Weissman, T.; and Ermon, S. 2019. NECST: Neural Joint Source-Channel Coding. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 1182–1192.

Dai, J.; Wang, S.; Tan, K.; Si, Z.; Qin, X.; Niu, K.; and Zhang, P. 2022. Nonlinear transform source-channel coding for semantic communications. *IEEE Journal on Selected Areas in Communications*, 40(8): 2300–2316.

Duan, Z.; Lu, M.; Ma, Z.; and Zhu, F. 2023. Lossy Image Compression With Quantized Hierarchical VAEs. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 198–207.

He, D.; Yang, Z.; Peng, W.; Ma, R.; Qin, H.; and Wang, Y. 2022. ELIC: Efficient Learned Image Compression With Unevenly Grouped Space-Channel Contextual Adaptive Coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5718–5727.

He, D.; Zheng, Y.; Sun, B.; Wang, Y.; and Qin, H. 2021. Checkerboard context model for efficient learned image compression. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 14771–14780.

Hu, Q.; Zhang, G.; Qin, Z.; Cai, Y.; Yu, G.; and Li, G. Y. 2023. Robust Semantic Communications with Masked VQ-VAE Enabled Codebook. *IEEE Trans. Wireless Commun.*, 22(12): 8707–8722.

Jiang, P.; Wen, C.-K.; Jin, S.; and Li, G. Y. 2022. Deep Source-Channel Coding for Sentence Semantic Transmission With HARQ. *IEEE Transactions on Communications*, 70(8): 5225–5240.

Kingma, D. P.; and Welling, M. 2022. Auto-Encoding Variational Bayes. arXiv:1312.6114.

Kingma, F.; Abbeel, P.; and Ho, J. 2019. Bit-Swap: Recursive Bits-Back Coding for Lossless Compression with Hierarchical Latent Variables. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 3408–3417. PMLR.

Kodak. 1993. Kodak PhotoCD dataset. http://r0k.us/graphics/kodak/.

Kurka, D. B.; and Gündüz, D. 2020. DeepJSCC-f: Deep Joint Source-Channel Coding of Images With Feedback. *IEEE Journal on Selected Areas in Information Theory*, 1(1): 178–193.

Lainema, J.; Bossen, F.; Han, W.-J.; Min, J.; and Ugur, K. 2012. Intra Coding of the HEVC Standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12): 1792–1801.

Li, X.; Bi, S.; Wang, S.; Li, X.; and Zhang, Y.-J. A. 2024. Digital Semantic Device-Edge Co-Inference With Task-Oriented ARQ. *IEEE Transactions on Vehicular Technology*, 73(9): 13986–13990.

Lu, M.; Duan, Z.; Zhu, F.; and Ma, Z. 2024. Deep Hierarchical Video Compression. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(8): 8859–8867.

Minnen, D.; Ballé, J.; and Toderici, G. D. 2018. Joint Autoregressive and Hierarchical Priors for Learned Image Compression. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Razavi, A.; van den Oord, A.; and Vinyals, O. 2019. Generating Diverse High-Fidelity Images with VQ-VAE-2. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32. Curran Associates, Inc.

Saidutta, Y. M.; Abdi, A.; and Fekri, F. 2019. Joint Source-Channel Coding for Gaussian Sources over AWGN Channels using Variational Autoencoders. In *IEEE International Symposium on Information Theory (ISIT)*, 1327–1331.

Shahid, I.; and Yahampath, P. 2013. Distributed Joint Source-Channel Coding Using Unequal Error Protection LDPC Codes. *IEEE Transactions on Communications*, 61(8): 3472–3482.

Sø nderby, C. K.; Raiko, T.; Maalø e, L.; Sø nderby, S. r. K.; and Winther, O. 2016. Ladder Variational Autoencoders. In

Lee, D.; Sugiyama, M.; Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems (NIPS)*, volume 29. Curran Associates, Inc.

Song, M.; Ma, N.; Dong, C.; Xu, X.; and Zhang, P. 2023. Deep Joint Source-Channel Coding for Wireless Image Transmission with Adaptive Models. *Electronics*, 12(22).

Sun, L.; Yang, Y.; Chen, M.; Guo, C.; Saad, W.; and Poor, H. V. 2023. Adaptive Information Bottleneck Guided Joint Source and Channel Coding for Image Transmission. *IEEE Journal on Selected Areas in Communications*, 41(8): 2628–2644.

Toderici, G.; Shi, W.; Timofte, R.; Theis, L.; Balle, J.; Agustsson, E.; Johnston, N.; and Mentzer, F. 2022. Workshop and challenge on learned image compression (clic2022). In *CVPR*.

Townsend, J.; Bird, T.; and Barber, D. 2019. Practical lossless compression with latent variables using bits back coding. *International Conference on Learning Representations*.

Vahdat, A.; and Kautz, J. 2020. NVAE: A Deep Hierarchical Variational Autoencoder. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 19667–19679. Curran Associates, Inc.

Wallace, G. 1992. The JPEG still picture compression standard. *IEEE Transactions on Consumer Electronics*, 38(1): xviii–xxxiv.

Wu, H.; Shao, Y.; Ozfatura, E.; Mikolajczyk, K.; and Gündüz, D. 2024a. Transformer-aided Wireless Image Transmission with Channel Feedback. *IEEE Transactions on Wireless Communications*, 1–1.

Wu, H.; Shao, Y.; Ozfatura, E.; Mikolajczyk, K.; and Gündüz, D. 2024b. Transformer-aided Wireless Image Transmission with Channel Feedback. *IEEE Transactions on Wireless Communications*, 1–1.

Xu, J.; Ai, B.; Chen, W.; Yang, A.; Sun, P.; and Rodrigues, M. 2022a. Wireless image transmission using deep source channel coding with attention modules. *IEEE Transactions On Circuits And Systems For Video Technology*, 32(4): 2315–2328.

Xu, T.; Wang, Y.; He, D.; Gao, C.; Gao, H.; Liu, K.; and Qin, H. 2022b. Multi-Sample Training for Neural Image Compression. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 1502–1515. Curran Associates, Inc.

Yang, K.; Wang, S.; Dai, J.; Qin, X.; Niu, K.; and Zhang, P. 2024. SwinJSCC: Taming Swin Transformer for Deep Joint Source-Channel Coding. *IEEE Transactions on Cognitive Communications and Networking*, 1–1.

Yang, M.; and Kim, H.-S. 2022. Deep Joint Source-Channel Coding for Wireless Image Transmission with Adaptive Rate Control. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5193–5197.

Yuhta Takida, T. S., Yukara Ikemiya. 2024. HQ-VAE: Hierarchical Discrete Representation Learning with Variational Bayes. *arXiv preprint arXiv:2401.00365*.

Zhang, G.; Hu, Q.; Qin, Z.; Cai, Y.; Yu, G.; and Tao, X. 2024. A Unified Multi-Task Semantic Communication System for Multimodal Data. *IEEE Transactions on Communications*, 72(7): 4101–4116.

Zhang, W.; Zhang, H.; Ma, H.; Shao, H.; Wang, N.; and Leung, V. C. M. 2023. Predictive and Adaptive Deep Coding for Wireless Image Transmission in Semantic Communication. *IEEE Transactions on Wireless Communications*, 22(8): 5486–5501.