Semantics-Guided Contrastive Joint Source-Channel Coding for Image Transmission

Semantics-Guided Contrastive Joint Source-Channel Coding for Image Transmission

Wenhui Hua†, Dezhao Chen†, Junli Fang†, Lingyu Chen†, João F. C. Mota‡ and Xuemin Hong†
† *School of Informatics, Xiamen University*, Xiamen, China
‡ *School of Engineering and Physical Sciences, Heriot-Watt University*, Edinburgh, U.K.
Email: {huawenhui, chendezhao, junlifang}@stu.xmu.edu.cn, j.mota@hw.ac.uk, {chenly, xuemin.hong}@xmu.edu.cn

***Abstract*—Deep joint source-channel coding (D-JSCC) provides several advantages over conventional coding schemes, in which source and channel coding are designed separately. For example, D-JSCC schemes suffer from smaller delays and are more robust to rapid channel variation. However, D-JSCCs are often designed without explicit structure or insight, making them less adaptive, hard to control, and theoretically unfounded. In this paper, we propose a contrastive joint-source-channel coding (C-JSCC) design, which uses supervised contrastive learning (SCL) to make the latent space of a conventional D-JSCC more structured and meaningful. By testing on the CIFAR-10 dataset, we show that C-JSCC consistently outperforms its D-JSCC counterpart in both tasks of image reconstruction and classification. Moreover, C-JSCC is shown to output images with perceptual quality better than the classic BPG image codec in the low bits-per-pixel (bpp) region. The roles of various hyper-parameters in C-JSCC are investigated by analytical approximations, experiments, and visualization techniques.**

***Index Terms*—Contrastive learning, joint source-channel coding, semantic information, image communications.**

I. Introduction

The strategy of separating source coding and channel coding in digital communication systems has achieved tremendous success. Such a strategy, however, also has limitations. First, the separated design is optimal only when the code length approaches infinity [1]. But long codes introduce excessive delays. Second, the coding performance deteriorates rapidly when the channel condition falls below a certain threshold. This is also known as the "cliff effect" [2]. Third, significant processing power is usually required for high-performance decoding. These drawbacks make the separated coding paradigm less attractive for delay/power sensitive applications and communication scenarios that have fast varying channels or long-delay feedback.

Joint source-channel coding (JSCC) [3]–[9] aims to overcome the drawbacks of the separated design with a unified coding structure. Recently, deep neural networks (DNNs) have been applied to source coding and shown to achieve outstanding performance beyond conventional hand-crafted codecs [10]. The advantage of using DNNs becomes particularly relevant when dealing with complex high-dimensional signals such as images. Inspired by such a success, much research attention has been drawn to DNN-based JSCC (or deep JSCC, D-JSCC), especially under the context of image transmission [3]–[9]. In [3], [4], Bourtsoulatze et al. proposed an autoencoder-based D-JSCC scheme, which was shown to outperform separated systems in low signal-to-noise ratio (SNR). This scheme was further extended to channels with feedback in [5], [6]. The work in [7] proposed an adaptive D-JSCC scheme, which balances the trade-off between rate and signal quality with a policy network. Further work investigated the issue of progressive coding for channel bandwidth adaptation [8], and studied D-JSCC under the context of multi-path channels and orthogonal frequency division multiplexing (OFDM) modulation [9]. Compared with separated designs, advantages of D-JSCC include smaller delays, absence of the cliff effect, and competitive performance at low SNR. On the downside, D-JSCCs are often designed in an ad-hoc manner, without conferring any spectral structure to the encoded signal. As a result, they are less adaptive, hard to control, and theoretically unfounded.

A key feature of D-JSCC is to encode high dimensional source signals, e.g., images, directly into continenous-valued baseband signals(e.g., complex baseband [3], [4] or [9] samples). From the perspective of machine learning, finding a good encoder can be seen as representation learning [11]. The encoded baseband signals correspond to latent vectors or latent representations. The major challenge of D-JSCC is to learn a representation that is both compact and robust to channel distortions. To address this challenge, it is important to gain insight into the structure of the encoded latent space.

Contrastive learning [12]–[15] is a recently proposed self-supervised learning technique that shows great promise for representation learning. The basic idea is to enforce two structural properties directly into the latent space. The first structural property is *semantic alignment*, which means that semantically-similar neighbors in the source domain remain neighbors in the latent space. For example, images with the same labels or objects should have a representation that is close in latent space. Or "For example, images whose distance in pixel space is small should have a representation that is also small in latent space". Another structural property is straightforward *uniformity*, which means encoded latent vectors are distributed uniformly in the entire feasible latent space. Using image class labels as an additional source of semantic information, supervised contrastive learning (SCL) [16] is a further enhancement to contrastive learning by encouraging the the third structural property: clustering of samples with identical labels.

Contributions. In this paper, we propose a contrastive JSCC

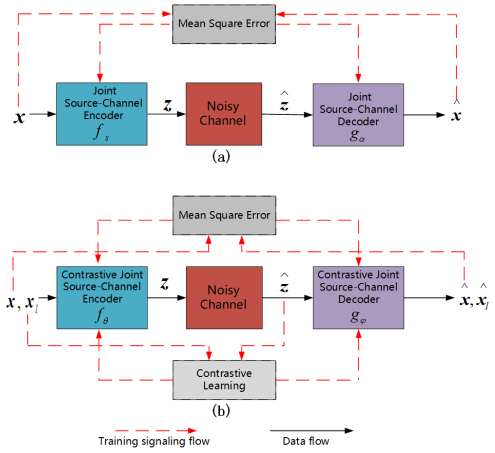


Fig. 1. Block diagram of DNN based JSCC systems: (a) Deep JSCC [3]–[9]; (b) **Our scheme:** Contrastive learning based JSCC.

(C-JSCC) scheme, which essentially applies SCL techniques to D-JSCC. The proposed C-JSCC can be seen as a generalization of the original D-JSCC proposed in [3], [4]. Although the application of SCL to D-JSCC is straight-forward, it is unclear *whether* or *how* the structural properties of SCL improve the performance of D-JSCC. To this end, our paper makes the following contributions:

- Under the popular context of image transmission, C-JSCC is shown to outperform its D-JSCC counterpart in both image reconstruction and classification. Tested on the CIFAR-10 dataset, and under Gaussian channels, the peak SNR (PSNR) of the reconstructed image improves by 2-3 dB, and the classification accuracy by 10%. In Rayleigh fading channels, the performance gains in PSNR and classification accuracy are 3dB and 15%, respectively.
- By relating the SCL loss to a more intuitive quantity, we shed new insights into how the SCL loss brings semantic structure into the latent space.
- We systematically study the impact of various C-JSCC hyper-parameters on the performance of image reconstruction and classification, and discuss the resulting trade-offs.
- We propose a new metric for measuring the information bottleneck, which upper bounds the information carrying capacity of a C-JSCC encoded representation. Using this metric as a common resource constraint, C-JSCC is shown to visually outperform the classic BPG image codec in the low bit per pixel (bpp) region.

II. Problem Formulation

A. D-JSCC under Gaussian and Reyleigh channels

Fig. 1(a) shows the block diagram of D-JSCC [3]-[9]. An encoder $\boldsymbol{f_s} : \mathbb{R}^L \to \mathbb{C}^K$, which implements a deterministic function (namely, a DNN) with trainable parameters $\boldsymbol{s}$, maps an input signal $\boldsymbol{x} \in \mathbb{R}^L$, e.g., an image, into a vector of complex-valued latent code $\boldsymbol{z}' \in \mathbb{C}^K$ to be transmitted via the channel. The last layer of $\boldsymbol{f_s}$ normalizes $\boldsymbol{z}$ such that the transmission power is fixed to $\boldsymbol{z} = \sqrt{KP/\|\boldsymbol{z}\|_2}\boldsymbol{z}'$, where $P > 0$.

The encoded vector $\boldsymbol{z}$ is sent over a physical layer channel as complex-valued symbols. The received symbol is given by $\hat{\boldsymbol{z}} = \mathbf{H}\boldsymbol{z}+\boldsymbol{n}$, where $\mathbf{H}$ is the channel gain and $\boldsymbol{n} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_K)$ is the additive Gaussian noise. We consider two types of classic channels: Rayleigh and Gaussian. The channel gain matrix $\mathbf{H}$ in Rayleigh fading and Gaussian channels are given by $\mathbf{H} \sim \mathcal{CN}(0, \mathbf{I}_K)$ and $\mathbf{H} = 1$, respectively.

It follows that the received signal has an average SNR $\rho = P/\sigma^2$. The decoder uses a decoding function $\boldsymbol{g_\alpha} : \mathbb{C}^K \to \mathbb{R}^L$ to directly map the received signal $\hat{\boldsymbol{z}}$ to an output (e.g., reconstructed image) $\hat{\boldsymbol{x}}$. The DNN-based decoding function is parameterized by $\boldsymbol{\alpha}$.

B. Information bottleneck as a resource constraint

The compression performance of a D-JSCC scheme is traditionally evaluated by K/L, which is the ratio of dimension reduction achieved by the encoder. This ratio, called "bandwidth compression" in the D-JSCC literature [3]–[9], has been widely used as a resource constraint for performance comparison of different coding schemes [17]. However, this constraint is incompatible with the common performance evaluation framework of separated source and channel coding, which use "bit" as a standard unit. For example, in image compression, the performance of an image codec is typically assessed via bpp.

To establish a common ground for performance comparison of D-JSCC and conventional separated coding schemes, this paper introduces a new resource constraint called "*information bottleneck measure*.". The idea is simply to use Shannon channel capacity (measured in bits) to upper bound the average amount of information that a D-JSCC symbol $\boldsymbol{z}$ can carry through a channel. More specifically, for a D-JSCC transmission with symbol dimension K and SNR ρ, in general, the information bottleneck measure (assuming i.i.d. parallel Gaussian channel) is given by

$$C = \frac{K}{2}\log_2(1+\rho). \tag{1}$$

This result comes directly from the property of Gaussian channels, in which the information per (scalar) symbol transmission is upper bounded by $\frac{1}{2}\log_2(1+\rho)$ [1]. Applying this new constraint, an image transmitted via D-JSCC can be compared against the same image encoded by a conventional codec into C bits and then decoded assuming no bit errors. This essentially means that the D-JSCC can be compared against the concatenation of a conventional source codec and a perfect channel codec.

C. Performance metrics and problem statement

The primary goal of D-JSCC for image transmission is to faithfully reconstruct the original image. In addition to pixel-wise reconstruction fidelity, classification accuracy is another widely-used performance metric of image transmission [18]–[20]. It has been suggested that the class of an image is akin to semantic information [21].

The aim of D-JSCC under a Gaussian channel is to find the optimal encoding-decoding function pair $\boldsymbol{f_s}$ and $\boldsymbol{g_\alpha}$ given

a constraint on the information bottleneck C. The encoded latent vector is used separately for image reconstruction and classification. Our goal is then to find a pair $(\boldsymbol{f_s}, \boldsymbol{g_\alpha})$ that jointly minimizes the reconstruction error while maximizing the classification accuracy.

III. Contrastive learning-based JSCC

A. Design intuition

From the perspective of representation learning, D-JSCC essentially tries to learn a noise-tolerant latent representation. The two optimization objectives of image reconstruction and classification can be translated into two desirable properties of the latent vector $\boldsymbol{z}$. First, for image reconstruction, it is desirable to maximize the (differential) entropy of $\boldsymbol{z}$, so that we can encode as much information as possible into the latent vector. A simple heuristic for maximizing entropy is to disperse the latent vector uniformly in the latent space. This property is called uniformity in the machine-learning literature [13]. Second, for image classification, images with similar semantic features (e.g., same class label) should have similar latent vectors, while images with dissimilar semantic features should have latent vectors that are far apart. This second property, called alignment, entails attributing different meanings or traits to different regions of the latent space. And to make the latent representations robust to noise, these regions should be as far apart as possible.

B. Loss function design and its interpretation

The top-level architecture of the proposed C-JSCC scheme is illustrated in Fig. 1(b). In essence, we propose a simple modification to the classical MSE loss function by combining it with an SCL loss:

$$\mathcal{L}_{sum} = (1-\lambda) * \mathcal{L}_{mse} + \lambda * \mathcal{L}_{scl}, \tag{2}$$

where

$$\mathcal{L}_{mse} = \frac{1}{N}\sum_{i=1}^{N} ||\boldsymbol{x}^i - \hat{\boldsymbol{x}}^i||^2 \tag{3}$$

is the average mean square error (MSE) between the original input image $\boldsymbol{x}^i$ and the reconstructed image $\hat{\boldsymbol{x}}^i$. Here, N is the number of randomly selected samples in a batch.

To calculate $\mathcal{L}_{scl}$, each of the N original data samples is augmented to produce a "positive" sample with the same class label, resulting in a set of $2N$ samples. For example, an augmentation strategy could be rotating the image, which preserves the class label. Within the augmented batch, let $i \in I \equiv \{1, \ldots, 2N\}$ be the index of an arbitrary sample, let $A(i) \equiv I \backslash \{i\}$ be the set of samples that excludes the anchored index i. Let $Q(i) \equiv \{q \in A(i) : \boldsymbol{x}_l^q = \boldsymbol{x}_l^i\}$ be the set of indices of all samples in A that have the same class label as the anchored sample. $\boldsymbol{x}_l$ is a semantic label. The upper right is the index. $\mathcal{L}_{scl}$ is defined as [16]

$$\mathcal{L}_{scl} = \sum_{i\in I} \frac{-1}{|Q(i)|} \sum_{q\in Q(i)} \log \frac{\exp(\hat{\boldsymbol{z}}_i \cdot \hat{\boldsymbol{z}}_q/\tau)}{\sum_{a\in A(i)} \exp(\hat{\boldsymbol{z}}_i \cdot \hat{\boldsymbol{z}}_a/\tau)}, \tag{4}$$

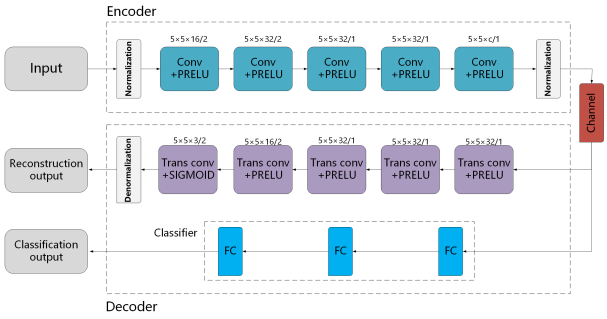


Fig. 2. DNN architecture and implementation configuration.

where $\cdot$ denotes the inner product, τ is a scalar temperature parameter, and $|\cdot|$ the cardinality of a set.

The above $\mathcal{L}_{scl}$ was originally proposed in [16] as a generalization to the contrastive learning loss to handle extra positive data samples with the same labels. In this paper, we will subsequently give a novel interpretation of the SCL loss to shed light into its insight. Our interpretation is based on an approximation to the log-sum-exp (LSE) function, which is defined as $\mathrm{LSE}(x_1, x_2, \ldots, x_m) = \log \sum_{i=1}^{m} \exp x_i$. The LSE is a smooth approximation of the maximum function and the approximation is bounded as

$$\begin{aligned} &\tfrac{1}{t}\,\mathrm{LSE}(tx) > \max\{x_1, \ldots, x_n\} \\ &\tfrac{1}{t}\,\mathrm{LSE}(tx) \le \max\{x_1, \ldots, x_n\} + \tfrac{\log(n)}{t} (t > 0). \end{aligned} \tag{5}$$

Applying the above inequality to (4) yields,

$$\Gamma < \mathcal{L}_{scl} < \Gamma + \log(2N), \tag{6}$$

where

$$\begin{aligned} \Gamma &= \tfrac{1}{\tau} \sum_{i\in I} \tfrac{1}{|Q(i)|} \sum_{q\in Q(i)} \left[\max_{a\in A(i)} (\hat{\boldsymbol{z}}_i \cdot \hat{\boldsymbol{z}}_a) - \hat{\boldsymbol{z}}_i \cdot \hat{\boldsymbol{z}}_q \right] \\ &= \tfrac{1}{\tau} \sum_{i\in I} \left[\max_{a\in A(i)} (\hat{\boldsymbol{z}}_i \cdot \hat{\boldsymbol{z}}_a) - \tfrac{1}{|Q(i)|} \sum_{q\in Q(i)} \hat{\boldsymbol{z}}_i \cdot \hat{\boldsymbol{z}}_q \right]. \end{aligned} \tag{7}$$

Furthermore, because the power of all latent vectors $\boldsymbol{z}$ is normalized to a constant P, the latent vectors are distributed over a hypersphere. The inner product terms in (4) essentially measure the cosine similarity between two vectors. The cosine similarity is measured by the cosine of the angle between two vectors and indicates whether two vectors are pointing in roughly the same direction.

The approximation in (6), (7) provides further insight into how the SCL loss introduces desirable properties into the latent space. First, semantic alignment is achieved by minimizing the angles (and distances) between positive sample pairs. Second, uniformity is encouraged by maximizing the angle between a sample and its nearest neighbor. Decreasing the temperature parameter τ can serve to highlight the importance of the nearest neighbor. The specific algorithmic steps are described in Algorithm 1.

Algorithm 1 Training of C-JSCC

Input: The Data sets $\boldsymbol{x}$, $\boldsymbol{x}_l$

Output: Encoder network parameters θ. Decoder network parameters $\boldsymbol{\varphi}$

while Stopping criterion is not met **do**
  Encoder encodes $\boldsymbol{x}$ into $\boldsymbol{z}$
  $\boldsymbol{z}$ is fed to the channel, which outputs $\hat{\boldsymbol{z}}$.
  Decoder decodes $\hat{\boldsymbol{z}}$ into $\hat{\boldsymbol{x}}$, $\hat{\boldsymbol{x}}_{\boldsymbol{l}}$
  Calculate loss $\mathcal{L}_{sum}$ by (2) and $\mathcal{L}_{cross-entropy}$
  Update $\boldsymbol{\theta}$ $\longleftarrow$ Gradient descent ($\boldsymbol{\theta}$, $\mathcal{L}_{sum}$)
  Update $\boldsymbol{\varphi}$ $\longleftarrow$ Gradient descent ($\boldsymbol{\varphi}$, $\mathcal{L}_{sum}$)
end while
Return the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$

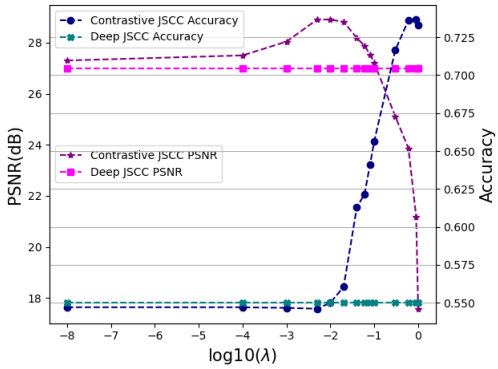


Fig. 3. Image PSNR (left) and classification accuracy (right) as functions of λ (K=500, $\tau = 0.01$, SNRtrain=SNRtest=10dB).

IV. Experiment Results and Discussions

A. DNN implementation and experiment settings

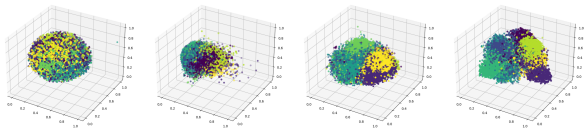
The proposed C-JSCC is a general training framework that can be applied to different types of DNNs. Without loss of generality, we adopt the same DNN implementation as in the original D-JSCC [3], to highlight the advantages of C-JSCC over the original D-JSCC. The DNN implementation is shown in Fig. 2. Readers can refer to [4] for further implementation details. The CIFAR-10 data set is used, which includes images of size 32x32 [22]. The AdamW optimization framework is used at a learning rate of 0.001. The batch size is set to 64. The peak SNR (PSNR) of reconstructed images and classification accuracy are adopted for performance evaluation of different JSCC schemes.

Similar to the training procedure of D-JSCC, C-JSCC is trained with varying SNR values, denoted as SNRtrain. This means that end-to-end DNN training is conditioned on the levels of noise imposed on the latent vector. Major hyper-parameters associated with C-JSCC training include the loss function weighting factor λ, the SCL temperature τ, the latent vector dimension K, and the training SNR. We will subsequently investigate the impacts of these hyper-parameters.

B. Impact of hyper-parameter λ

Fig. 3 shows how λ, which varies from 10^{-8} to 1, affects classification accuracy and the PSNR. The performance of conventional D-JSCC (i.e., $\lambda = 0$) is also shown for comparison. It is observed that the classification accuracy improves with increasing λ until a performance ceiling is reached around $\lambda = 1$. Moreover, increasing λ up to a certain threshold (in this case around 0.01) is shown to be beneficial to the PSNR. However, further increasing λ beyond this threshold leads to a steady degradation of PSNR. It is interesting to see that the PSNR and accuracy exhibit a clear trade-off right after the threshold. Within this trade-off region, there is an interval (around 0.01 to 0.1) where C-JSCC outperforms D-JSCC in both PSNR and accuracy.

To gain further insight into how λ modulates the latent space, Fig. 4 applies the t-distributed stochastic neighbor embedding (t-SNE) technique [23] to visualize the structure of latent vectors $\boldsymbol{z}$. Different colors are used to differentiate different image classes. At $\lambda = 0$, the latent vectors are mixed without obvious structure. However, when λ increases, latent vectors of the same class tend to aggregate into distinct clusters. Fig. 4 partly explains the underlying cause of the PSNR-accuracy trade-off previously shown in Fig. 3: while semantic clustering helps to improve the classification accuracy by forming semantic constellations, well-formed clusters could be detrimental to the PSNR due to a reduction of the overall entropy.



(a) $\lambda = 0$ (b) $\lambda = 0.01$ (c) $\lambda = 0.6$ (d) $\lambda = 1.0$

Fig. 4. Visualization of the C-JSCC latent vector structure using t-SNE (λ = [0, 0.01, 0.6, 1.0], $\tau = 0.01$, K=500, and SNRtrain=SNRtest=10dB) .

C. Impact of hyper-parameter τ

Similar to Fig. 3, Fig. 5 shows the impact of temperature parameter τ on the PSNR and accuracy. As discussed in [13], increasing the value of τ tends to improve alignment (i.e., produce semantic clusters) in the latent space. This means increasing τ has a similar effect as increasing λ. The results in Fig. 5 mirrors our previous findings in Fig. 3: the PSNR and accuracy both improve initially with increasing τ, but eventually degrade after τ exceeds a certain threshold value. Figs. 3 and 5 both suggest that a small degree of semantic clustering is beneficial to the PSNR.

D. Impacts of training SNR

The training SNR is another important parameter that has an overall impact on the D-JSCC performance. Setting lambda = tau = 0.01, so that we are in the trade-off region, Fig. 6 compares the PSNR performance of C-JSCC against the conventional D-JSCC [3], [4] with varying training SNR values ranging from 1dB to 25 dB. In all values of training

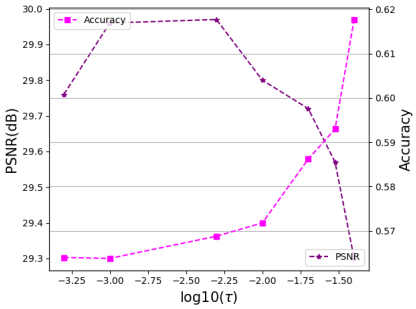
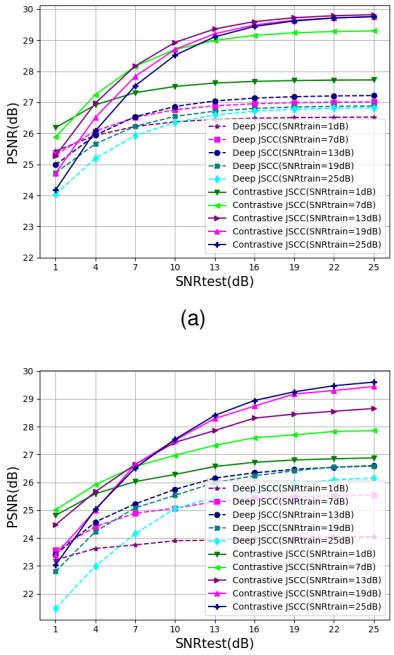


Fig. 5. Image PSNR (left) and classification accuracy (right) as functions of τ (K=500, $\lambda = 0.01$, SNRtrain=SNRtest=10dB)



(a)

(b)

Fig. 6. Image PSNR as a function of SNRtest, with C-JSCC models trained under different values of SNRtrain: (a) Gaussian channel, (b) Rayleigh channel($\lambda = 0.01$, $\tau = 0.01$, $K = 500$).

SNR, C-JSCC is shown to consistently outperform D-JSCC by 2 to 3 dB.

E. Impacts of latent vector dimension K

For a thorough comparison, Fig. 7 extends Fig. 6 by treating K as a new variable. For each value of K, different SNRtest-SNRtrain pairs are tested for C-JSCC and D-JSCC. The resulting PSNR is shown in Fig. 7 using a standard box plot. Again, C-JSCC is shown to consistently outperform D-JSCC in all parameter settings. Moreover, it is shown that the PSNR gain tends to increase with larger values of K.

F. Comparison of classification accuracy

Table. I shows the classification performance of C-JSCC and D-JSCC. Here, "K=average" represents the average performance over $K = \{200, 500, 1000, 1500, 2000, 2500\}$. Awgn

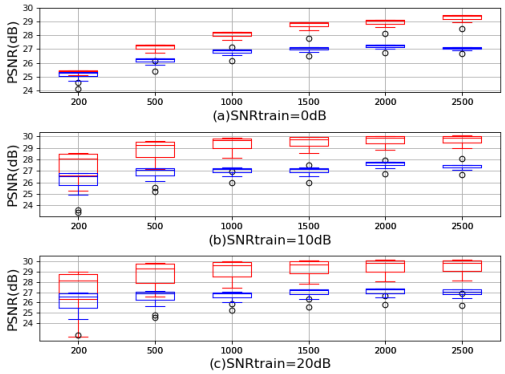


Fig. 7. Image PSNR as a function of K ($\lambda = 0.01$, $\tau = 0.01$, SNRtrain=[0dB,10dB,20dB], SNRtest=[1dB-25dB], red for C-JSCC and blue for D-JSCC).

TABLE I
IMAGE CLASSIFICATION PERFORMANCE

Method	SNRtest=[0dB,10dB,20dB]
AWGN	
CJSCC($\tau = 0.01$,K=average)	0.491,0.556,0.569
CJSCC($\tau = 0.07$,K=500)	0.554,0.618,0.646
CJSCC($\tau = 0.13$,K=500)	0.608,0.652,0.661
CJSCC($\tau = 0.19$ K=500)	0.646,**0.682**,0.685
CJSCC($\tau = 0.25$,K=500)	**0.649**,0.681,**0.692**
DJSCC(K=average)	0.486,0.551,0.553
DJSCC(K=500)	0.481,0.551,0.557
Rayleigh	
CJSCC($\tau = 0.01$,K=average)	0.439,0.546,0.574
CJSCC($\tau = 0.07$,K=500)	0.491,0.591,0.627
CJSCC($\tau = 0.13$,K=500)	0.542,0.636,0.664
CJSCC($\tau = 0.19$,K=500)	0.566,0.664,0.689
CJSCC($\tau = 0.25$,K=500)	**0.590**,**0.670**,**0.698**
DJSCC(K=average)	0.433,0.532,0.562
DJSCC(K=500)	0.408,0.514,0.554

and Rayl represent the Gaussian and Rayleigh channels respectively. We set $\lambda = 0.01$. Again, C-JSCC is shown to consistently outperform D-JSCC with an average improvement of 10 %. As expected, increasing τ leads to improved classification accuracy.

G. Visual comparison with BPG

Finally, Fig. 8 shows the visual comparison of reconstructed images yielded by the proposed C-JSCC scheme and the classic BPG codec. As explained previously in Section II, the information bottleneck is used as a resource constraint for C-JSCC. This allows us to assign proper values for K and SNR in C-JSCC given a target bpp. It can be seen that C-JSCC significantly outperforms BPG at low bpp values, making it a promising coding scheme for image transmission over adverse communication channels.

V. CONCLUSIONS

In this paper, we propose C-JSCC as a generalization and enhancement to the conventional D-JSCC. Using the CIFAR-10 dataset, the proposed C-JSCC is shown to outperform D-JSCC by 2-3 dB of PSNR for image reconstruction, and by 10% for image classification under Gaussion channels. In

Fig. 8. Visual comparison of reconstructed images from C-JSCC and BPG at different bpp (columns (b-d) are outputs of C-JSCC, SNRtrain=20dB, $\lambda = \tau = 0.01$), the text below each image indicate parameters K/SNRtest/bpp.

Rayleigh fading channels, the performance gains in PSNR and classification accuracy are 3dB and 15%, respectively. An analytical investigation into the SCL loss function shows that the temperature hyper-parameter can be used to adjust attention between local and global samples in the latent space. Through experiments, we show that uniformity and alignment properties in the latent space have a direct impact on the image reconstruction and classification performance, respectively. For image reconstruction, introducing semantic structure into the latent space is initially beneficial but detrimental after a certain threshold. Using bpp as a common constraint, C-JSCC is shown to outperform BPG perceptually at low rates. We conclude that the proposed C-JSCC is a promising candidate for low-rate image transmission over adverse communication channels.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China(Grant No.62077040), the Science and Technology Key Project of Fujian Province, China(No.2019HZ020009), and the Xiamen Special Fund for Marine and Fishery Development.

REFERENCES

[1] T. M. Cover and J. A. Thomas, "Elements of information theory," *Wiley*, 1991.

[2] S. Kokalj-Filipovi and E. Soljanin, "Suppressing the cliff effect in video reproduction quality," *Bell Labs Tech. J.*, vol. 16, no. 4, pp. 171–185, Apr 2012.

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

[4] ——, "Deep joint source-channel coding for wireless image transmission," in *Proc. IEEE Int'l. Conf. Acoustics Speech Signal Process. (ICASSP'19)*, Brighton, UK, May. 2019, pp. 4774–4778.

[5] D. B. Kurka and D. Gündüz, "Deep joint source-channel coding of images with feedback," in *Proc. IEEE Int'l. Conf. Acoustics Speech Signal Process. (ICASSP'20)*, Bienvenido, Bassac, May 2020, pp. 5235–5239.

[6] ——, "Deep jscc-f: Deep joint source-channel coding of images with feedback," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 178–193, Dec 2020.

[7] M. Yang and H.-S. Kim, "Deep joint source-channel coding for wireless image transmission with adaptive rate control," in *Proc. IEEE Int'l. Conf. Acoustics Speech Signal Process. (ICASSP'22)*, May 2022.

[8] D. B. Kurka and D. Gündüz, "Bandwidth-agile image transmission with deep joint source-channel coding," *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 8081–8095, Dec 2021.

[9] M. Yang, C. Bian, and H.-S. Kim, "OFDM-guided Deep Joint Source Channel Coding for Wireless Multipath Fading Channels," *IEEE Trans. Cogn. Commun. Networking.*, pp. 1–1, 2022.

[10] S. Kim, D. Min, B. Ham, S. Jeon, S. Lin, and K. Sohn, "Fcss: Fully convolutional self-similarity for dense semantic correspondence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR'17)*, Honolulu, HI, USA, July 2017.

[11] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions. Pat. Mach.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[12] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul 2020, pp. 1597–1607.

[13] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul 2020, pp. 9929–9939.

[14] O. Henaff, "Data-efficient image recognition with contrastive predictive coding," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul 2020, pp. 4182–4192.

[15] F. Wang and H. Liu, "Understanding the behaviour of contrastive loss," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV'21)*, Jan 2021, pp. 2495–2504.

[16] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS'20)*, Vancouver, Canada, 2020, pp. 18 661–18 673.

[17] N. V. B. N. Thomos and M. G. Strintzis, "Optimized transmission of jpeg2000 streams over wireless channels," *IEEE Trans. Image process.*, vol. 15, no. 1, pp. 54–67, Jan. 2006.

[18] L. D. Chamain, F. Racapé, J. Bégaint, A. Pushparaja, and S. Feltman, "End-to-end optimized image compression for machines, a study," in *Proc. IEEE Data Compress. Conf. (DCC)*, Mar 2021, pp. 163–172.

[19] Q. Wang, L. Shen, and Y. Shi, "Recognition-driven compressed image generation using semantic-prior information," *IEEE Signal Process. Lett.*, vol. 27, pp. 1150–1154, 2020.

[20] Y. Matsubara, R. Yang, M. Levorato, and S. Mandt, "Supervised compression for resource-constrained edge computing systems," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV'22)*, Waikoloa, HI, USA, 2022, pp. 2685–2695.

[21] S. Zheng and G. Gupta, "Semantic-guided zero-shot learning for low-light image/video enhancement," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV'22)*, Waikoloa, HI, USA, 2022, pp. 581–590.

[22] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[23] L. V. D. Maaten and G. Hinton, "Visualizing data using t-sne." *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov 2008.