Generative Joint Source-Channel Coding for Semantic Image Transmission

Ecenaz Erdemir, *Member, IEEE*, Tze-Yang Tung, *Member, IEEE*, Pier Luigi Dragotti, *Fellow, IEEE*, and Deniz Gündüz, *Fellow, IEEE*

***Abstract*—Recent works have shown that joint source-channel coding (JSCC) schemes using deep neural networks (DNNs), called DeepJSCC, provide promising results in wireless image transmission. However, these methods mostly focus on the distortion of the reconstructed signals with respect to the input image, rather than their perception by humans. However, focusing on traditional distortion metrics alone does not necessarily result in high perceptual quality, especially in extreme physical conditions, such as very low bandwidth compression ratio (BCR) and low signal-to-noise ratio (SNR) regimes. In this work, we propose two novel JSCC schemes that leverage the perceptual quality of deep generative models (DGMs) for wireless image transmission, namely InverseJSCC and GenerativeJSCC. While the former is an inverse problem approach to DeepJSCC, the latter is an end-to-end optimized JSCC scheme. In both, we optimize a weighted sum of mean squared error (MSE) and learned perceptual image patch similarity (LPIPS) losses, which capture more semantic similarities than other distortion metrics. InverseJSCC performs denoising on the distorted reconstructions of a DeepJSCC model by solving an inverse optimization problem using the pre-trained style-based generative adversarial network (StyleGAN). Our simulation results show that InverseJSCC significantly improves the state-of-the-art DeepJSCC in terms of perceptual quality in edge cases. In GenerativeJSCC, we carry out end-to-end training of an encoder and a StyleGAN-based decoder, and show that GenerativeJSCC significantly outperforms DeepJSCC both in terms of distortion and perceptual quality.**

***Index Terms*—Semantic communication, wireless communication, joint source-channel coding, perceptual similarity, generative adversarial networks, inverse problems, StyleGAN.**

I. Introduction

COMMUNICATION systems are designed and optimized to reliably transmit information from one point to another over noisy communication channels. They consist of three blocks: an encoder at the transmitter, a noisy channel, and a decoder at the receiver. In almost all existing communication systems, the encoding process follows the two-step approach motivated by Shannon's separation theorem [1], which decomposes the transmitter into a source encoder and a channel encoder. The former removes the redundant information from the source to allow reconstruction at the desired quality, while the latter protects the compressed information against errors introduced over the wireless channel. Similarly, the receiver consists of source and channel decoders. Shannon proved that separate source and channel coding provides theoretical optimality in the case of infinitely long code blocks. On the other hand, in practical systems that require source transmission under extreme latency and bandwidth constraints, it is known that the separation approach can be highly suboptimal [2]. Mission-critical applications such as industrial automation, remote surgery, and autonomous vehicles demand very low latency with response times measured in milliseconds or even microseconds [3]–[6]. While 5G standards introduced services for ultra-reliable low latency communications (URLLC) motivated by these use cases, particularly in cases involving the delivery of high-rate image/video sources, end-to-end latency is dominated by the time needed for compression and decompression of these sources. By combining the compression and communication into a single transformation, JSCC has the potential to meet the requirements of these critical applications in the 6G networks.

Although joint source-channel coding (JSCC) has been known to achieve better performance than separate source compression followed by channel coding, practical design of such joint coding schemes has been a long-standing challenge. Recently, deep learning (DL) based JSCC methods, e.g., DeepJSCC, have shown outstanding results due to their ability to extract complex features from the training data while incorporating the channel characteristics into their encoding implicitly [7]–[15]. However, these DL-based JSCC methods do not focus on the semantic similarities between the source signal and its reconstruction at the receiver and have mostly considered the mean squared error (MSE) or structured similarity metrics (SSIM/MS-SSIM) as the end-to-end measure of distortion, and the loss function for training. On the other hand, in semantic communications [16], particularly in the context of image/video transmission, the receiver is not necessarily interested in reconstructing the original source signal with minimal distortion. Instead, the receiver may be interested in some downstream task, such as classification [17], [18], or retrieval [19], in which case it would be sufficient to only convey the relevant features of the source signal for the prescribed task. Alternatively, for extreme image compression [20], the receiver may be interested in generating an output with the same distribution as the source signal rather than its accurate representation. In image compression literature, this aspect has recently been acknowledged as the perception

E. Erdemir, T.-Y. Tung, P. L. Dragotti, and D. Gündüz are with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. D. Gündüz is also with the Theory Lab, Central Research Institute, 2012 Labs, Huawei Technologies Co., Ltd., Hong Kong SAR, China (e-mails: (e.erdemir17, tze-yang.tung14, d.gunduz)@imperial.ac.uk)

Manuscript received August 31, 2022; revised March 22, 2023; accepted May 3, 2023.

This work received funding from the UKRI for the projects AIR (ERC-CoG, EP/X030806/1).

loss, resulting in the rate-distortion-perception trade-off [21]. In general, the perception loss is measured by the divergence between the source and reconstruction distributions, and can be minimized in practice using generative adversarial networks (GANs) [22].

In classical DeepJSCC, the goal is to reconstruct the source signal, e.g., image, at the receiver with minimum distortion by jointly optimizing the source and the channel coding via deep neural networks (DNNs). However, this often results in a significant loss of perceptual quality for the edge cases. We overcome this by following a context-aware communication approach that utilizes style-based generative adversarial network (StyleGAN) generators as well as distortion metrics that align well with the human perception in our optimizations.

Past works on DeepJSCC train the encoder/decoder networks on patches of images from a large dataset, which are then used and evaluated for the transmission of high-resolution images from the Kodak dataset [7]–[14]. It has been observed that the larger and richer the training dataset the better the performance. On the other hand, the performance improves further if the statistics of the training dataset match that of the test images. This is shown in [8] by training and testing DeepJSCC on satellite images that exhibit a particular statistic, where the gain from DeepJSCC with respect to conventional separation-based approaches becomes even more significant. In our first scenario in this work, we consider a DeepJSCC encoder/decoder pair trained on a generic training dataset, e.g., ImageNet. Later, we are given images from another dataset to be transmitted wirelessly. We assume that the encoder network remains the same, but the decoder can deduce the knowledge of the dataset to improve the perceptual quality of the reconstructed images. We propose a novel inverse problem approach for this scenario, called *InverseJSCC*. InverseJSCC performs denoising on the distorted reconstructions of DeepJSCC by solving an unsupervised inverse problem and recovers a high-quality source image with better perceptual quality. Exploiting the learned distribution for a particular dataset, e.g., face images, we use a pre-trained StyleGAN-2 [23] generator, the SotA architecture for high-quality image generation, to generate new face images that perceptually match the source image. While InverseJSCC significantly improves classical DeepJSCC perceptually, it also allows using any differentiable encoder/decoder pair to reconstruct the noisy images. We also show that, unlike DeepJSCC, InverseJSCC does not lose significant performance when the statistics of the training dataset do not match that of the test images.

Next, we consider end-to-end training of the encoder and decoder using the learned distribution. The proposed solution is called *GenerativeJSCC*, which is composed of an encoder, a non-trainable channel, and a decoder that incorporates residual DNNs and the StyleGAN-2 generator structure. We perform end-to-end training of the encoder/decoder pair of GenerativeJSCC by optimizing a distortion measure that also takes perceptual quality into account. Our results show that GenerativeJSCC outperforms classical DeepJSCC in terms of both traditional distortion metrics (e.g., mean-squared error) and perceptual similarity metrics that align with human perception well.

A. Our Contributions

We can summarize the main contributions of this work as follows:

- We study the wireless image transmission problem by considering both the end-to-end distortion and the perceptual quality of the reconstructed image. To achieve realistic high-quality image reconstruction, we employ a pre-trained generator at the receiver and treat the problem as an unsupervised image reconstruction problem. To the best of our knowledge, this is the first work to study distortion-perception trade-off in the context of JSCC.
- We first propose InverseJSCC, which uses a pre-trained GAN generator to invert the wireless communication problem in an unsupervised manner and improves the perceptual quality of the state-of-the-art (SotA) DeepJSCC by exploiting the learned distribution of the transmitted images at the receiver. To the best of our knowledge, this is the first unsupervised inverse problem approach to the wireless image communication problem.
- We then propose GenerativeJSCC, which is a supervised end-to-end solution to the wireless image communication problem using a pre-trained GAN.
- We carry out extensive simulations using ImageNet and CelebA-HQ datasets and show that InverseJSCC can help improve the perceptual quality of the reconstructed images even though the encoder remains the same. On the other hand, GenerativeJSCC significantly outperforms the SotA DeepJSCC model in edge cases in terms of both distortion and perception quality.

II. Related Work

DL techniques for JSCC and its various wireless communication applications have attracted a lot of attention in the past decade [7]–[14]. The first work employing DNNs in wireless image transmission, called *DeepJSCC*, was proposed in [7]. The authors used an autoencoder architecture to represent a wireless communication system that is composed of an encoder, a non-trainable channel and a decoder. The encoder/decoder pair is jointly trained such that the encoder learns a function that maps the source image directly to continuous channel inputs and the decoder learns to reconstruct the image from its noisy observations. They demonstrated in [7] that DeepJSCC outperforms classical separation schemes with compression [24] and channel coding [25], and it shows success in adapting to different source or channel types and avoiding the *cliff effect*. DeepJSCC has later been extended to different channel models [9], [12] as well as to video transmission [11]. However, previous JSCC approaches mostly focus on the rate-distortion trade-off in terms of classical pixel-wise distortion metrics, and often disregard human perception. In [12], the authors study JSCC combined with orthogonal frequency division multiplexing (OFDM) to cope with multi-path fading. The authors adopt a GAN formulation to combine the MSE loss with the adversarial loss of a discriminator network, which distinguishes whether the image is original or generated by the decoder. The generator network of [12] is jointly trained with the encoder by following the traditional

adversarial training approach. However, GAN-generated image quality relies heavily on the range of the generator. The generator network of [12] cannot achieve the perceptual image quality of the SotA StyleGAN generator, which follows the methodology of progressive growing GANs [26], and hence produces remarkably high-quality images. Moreover, it has been shown that the traditional distortion metrics, e.g., MSE, PSNR, and SSIM, also used in [12], are not as successful in representing human judgments compared to the learned perceptual image patch similarity (LPIPS) metric [27].

Deep generative models (DGMs) have recently shown immense perceptual quality, which refers to being perceived by humans as a valid (natural) sample [21]. Therefore, another line of DNN-based JSCC work has utilized well-known DGMs, such as Variational Autoencoders (VAEs), GANs, and etc., in wireless image transmission [15], [28]–[30]. The idea behind these JSCC approaches comes from the similarity between the wireless communication systems and inverse problems, in which an unknown signal, image or a multi-dimensional volume is reconstructed from its observations [31]. In this analogy, the observations are obtained from a forward process, e.g., an encoder and a noisy channel, which are then inverted by a decoder in a supervised manner to reconstruct the original signal. Hence, a wireless communication system is optimized to learn an identity function such that the reconstructed message matches the source message. This supervised inverse problem approach is similar to denoising autoencoders [32], which was incorporated into the wireless communication setting in [15]. The authors propose a VAE-based JSCC scheme for a binary symmetric channel and a binary erasure channel system, later extended to wiretap channel scenario by [30]. Despite their exceptional generative capability, GANs have only been considered in a JSCC scheme in the context of secure image transmission [29], where generative adversarial training is used for image reconstruction at the receiver. Despite utilizing DGMs, these approaches still do not incorporate the SotA DGM structures into the wireless communication problem in order to reconstruct high perceptual quality images.

Unsupervised inverse problems, on the other hand, do not rely on a matched dataset of the input images and measurements, since the input images are not available at test time [31]. Instead, they estimate the best potential input image that goes through the known forward process and matches with the observed measurements. However, prior assumptions on the input image distribution, e.g., sparsity assumption in classical compressed sensing [33] are needed. Pre-trained DGM-based approaches have recently received great attention due to their immense capacity of input data representation [34]–[36]. The first DGM-based inverse problem approach, called CSGM, was proposed in [34]. The authors solved the classical compressed sensing problem, i.e., $Ax+\text{noise} = y$, by assuming that the input image x is in the range of a pre-trained DGM. Their results show that CSGM outperforms traditional compressed sensing solutions for denoising, inpainting, and compressed sensing, when the input image is in the range of a pre-trained VAE or GAN generator. Later, [35] improved CSGM by including a regularization term that limits the latent input of the generator network to a spherical ball. PULSE [35] also used a pre-trained StyleGAN-2 generator [23]. Another work that uses GAN priors for inverse problems is proposed in [36], namely intermediate layer optimization (ILO). As its name suggests, ILO solves the classical compressed sensing objective by optimizing the intermediate layers of a pre-trained StyleGAN-2 generator. The algorithm first applies CSGM to find the best initialization for the latent input of the GAN generator, then the objective is optimized by solving projected gradient descent for each intermediate layer recursively. ILO [36] significantly outperforms the previous approaches, hence it is the SotA inverse problem solution that uses GAN priors. As it can be seen from the mentioned references, in highly complex machine learning tasks, it is a common practice to use pre-trained SotA generative models. These models are trained for extended periods using extensive resources and made open source for scientific development. Despite their high cost of training, they are becoming increasingly popular and widely available as more data and compute resources are becoming available.

To the best of our knowledge, unsupervised inverse problems, which show exceptional denoising performance, have not been studied in the context of wireless communication. For both supervised and unsupervised methods, there is a demand in the field to incorporate the generative capabilities of SotA DGMs into wireless image transmission problems.

III. Problem Statement

We consider a communication scenario in which a user wants to reliably transmit a source signal from one point to another over a noisy channel. Our goal is to maximize the semantic similarity between the source and the reconstructed signals at the receiver while minimizing the distortion caused by the channel. We measure the semantic similarity using the perceptual quality metric LPIPS introduced in [27]. In [27], the authors argue that LPIPS is a distortion measure that better aligns with human perception than the typical pixel-wise metrics. We consider a transmitter that maps the source vector $\mathbf{x} \in \mathbb{R}^m$ into a vector of complex-valued channel input symbols $\mathbf{z} \in \mathbb{C}^k$ by an encoding function $f : \mathbb{R}^m \to \mathbb{C}^k$, where m and k are source and channel bandwidth, respectively. We define the BCR as

$$\rho = \frac{k}{m}, \tag{1}$$

which represents the level of compression applied to the source signal. In the case of an image source, the input size would be $m = H \times W \times C$, where H, W, and C are the height, width, and color dimensions of the source image. An average transmit power constraint $\bar{P}$ is imposed at the output of the encoder before the signal is transmitted through the channel, such that

$$\frac{1}{k}\mathbb{E}_{\mathbf{z}}[\|\mathbf{z}\|_2^2] \leq \bar{P}, \tag{2}$$

where the expectation is over the distribution of the encoded signal. The power constraint is implemented at the immediate

output of the encoder by normalizing the encoded signal according to:

$$\mathbf{z} = \sqrt{k\bar{P}}\frac{\tilde{\mathbf{z}}}{\sqrt{\tilde{\mathbf{z}}^H\tilde{\mathbf{z}}}}, \tag{3}$$

where $\tilde{\mathbf{z}}$ is the input to the final normalization at the output of the encoder, and H refers to the Hermitian transpose. The normalized encoded signal $\mathbf{z}$ is transmitted over the noisy channel that applies a random corruption function $\eta : \mathbb{C}^k \to \mathbb{C}^k$ that turns $\mathbf{z}$ into the corrupted signal $\hat{\mathbf{z}}$. The transfer function of the AWGN channel is

$$\eta(\mathbf{z}, \sigma^2) = \mathbf{z} + \mathbf{n}_C, \tag{4}$$

where the channel noise vector $\mathbf{n}_C$ is sampled in an independent identically distributed (i.i.d.) manner from a circularly symmetric complex Gaussian distribution, i.e., $\mathbf{n}_C \sim \mathcal{CN}(0, \sigma^2 I_{k\times k})$, and σ^2 is the channel noise power known by both the transmitter and the receiver. Accordingly, the channel SNR is defined as

$$\text{SNR} = 10\log_{10}\frac{\bar{P}}{\sigma^2}\ \text{dB}. \tag{5}$$

The noisy channel output is observed by the receiver, which then decodes it to an approximate reconstruction $\hat{\mathbf{x}} \in \mathbb{R}^m$ of the source signal by a decoding function $g : \mathbb{C}^k \to \mathbb{R}^m$. Our problem employs the AWGN channel model, with the assumption that the channel model and SNR are known to both the transmitter and the receiver.

The entire pipeline is optimized by jointly designing the encoder and the decoder functions such that the average distortion between the original source $\mathbf{x}$ and its reconstruction $\hat{\mathbf{x}}$ at the output of the receiver is minimum:

$$\min_{f,g} \mathbb{E}_{p(\mathbf{x},\hat{\mathbf{x}})}\big[d(\mathbf{x}, \hat{\mathbf{x}})\big], \tag{6}$$

where $p(\mathbf{x}, \hat{\mathbf{x}})$ is the joint distribution of the source and its reconstruction, and $d(\mathbf{x}, \hat{\mathbf{x}})$ can be any measure. In a typical communication scheme, compression schemes usually try to minimize the distortion at any given bit rate, for instance, by minimizing the MSE or maximizing the PSNR, SSIM, etc. However, typical distortion measures do not necessarily imply high perceptual quality; in fact, minimizing the distortion might reduce the perceptual quality [21]. Therefore, throughout this paper, we focus on a distortion measure that takes both perceptual and pixel-wise similarities into account. Fig. 1 illustrates a simplified diagram of the communication system considered herein.

In the following section, we propose two JSCC schemes that effectively minimize the distortion between the source signal and its reconstruction, while maximizing perceptual generation quality with the help of DGMs. Inspired by the SotA DL-based JSCC approaches [7], [9]–[11], we consider DNN-based encoder/decoder pairs that also benefit from the generative capabilities of DGMs in both approaches. The first scheme can be considered as an inverse problem approach to the optimization (6), in which the encoder/decoder functions f and g are fixed and a high-quality approximation of $\mathbf{x}$ is recovered from reconstructions $\hat{\mathbf{x}}$ using DGMs. In the second scheme, given $\mathbf{x}$, Eqn. (6) is minimized by jointly optimizing f and g in a supervised manner.

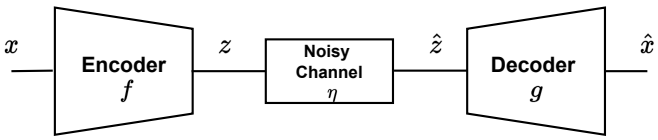


Fig. 1. Communication system with joint source-channel coding.

IV. Proposed Solutions

Recently proposed CNN-based approaches to JSCC [7]–[11] are based on the autoencoder architecture and have focused on the end-to-end distortion between the source and the reconstructed image. The encoder and decoder functions, f and g, are modeled as DNNs, which are jointly trained. The goal of the encoder is to transform the source image to a form that can be conveyed through the channel; that is, it satisfies the power and bandwidth constraints. The decoder aims at reversing the impact of the encoder as well as the channel on the input image. However, these works did not consider the perceptual quality of the reconstructed images.

A GAN-based distortion measure is considered in [12], but their decoder acts as the generator directly, which is trained as part of the encoder/decoder pair. In contrast, our work takes advantage of the outstanding generative performance of pre-trained SotA GANs as well as distortion metrics that align with human visual judgment in order to obtain high perceptual quality reconstructions. Firstly, in InverseJSCC, we improve the classical DeepJSCC model by inverting the communication pipeline with the help of a GAN generator by exploiting the learned statistics of the source images at the receiver. This requires solving an inverse problem, in which the forward operator to be inverted is represented by the encoder-channel-decoder architecture from [11]. In a sense, we optimize the input of the generator model to produce a certain image; that is, we find a latent vector, whose output transmitted by the DeepJSCC encoder through the channel and reconstructed by the DeepJSCC decoder, gives the observed reconstruction. InverseJSCC significantly improves the performance of SotA DeepJSCC schemes perceptually in cases where the training dataset statistics of the encoder/decoder pair match with the test dataset, as well as those that do not. On the other hand, in GenerativeJSCC, we propose an end-to-end training scheme for JSCC which consists of an encoder, a non-trainable channel, and a GAN-based decoder.

A. Inverse Problem Approach to Semantic Communications

In this section, we introduce InverseJSCC as an unsupervised improvement on the DeepJSCC approach by treating the end-to-end reconstruction by DeepJSCC as the FP, and model the reconstruction of the source image with better perceptual quality as an *inverse problem*. More specifically, InverseJSCC treats the reconstructions at the receiver as the *measurements*, obtained by an unknown data sample going through a typically non-invertible *forward process (FP)*. Inverse problems can be formalized as follows [31],

$$\mathbf{y} = A(\mathbf{x}) + \mathbf{n}_A, \tag{7}$$

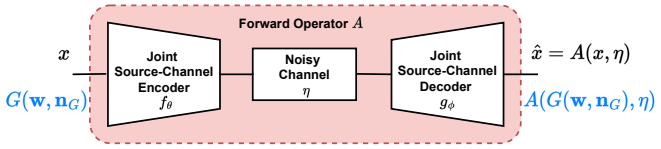


Fig. 2. DeepJSCC model as the forward operator A, where η is the AWGN channel function, $G(\cdot)$ is the generator network of a GAN, and $\mathbf{w}$ is the latent vector input to the generator.

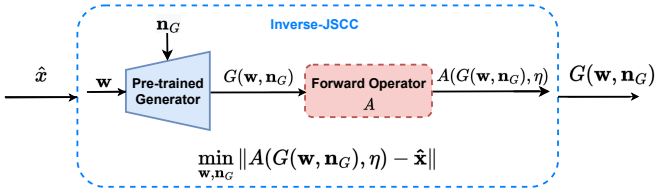


Fig. 3. InverseJSCC scheme applied to the signal reconstructed by the DeepJSCC encoder/decoder pair in an unsupervised way.

where A is the forward operator, an approximate representation of FP, and $\mathbf{n}_A$ is an additive noise. Alternatively, non-additive noise can be considered in a more general model, i.e., $\mathbf{y} = \mathcal{N}(A(\mathbf{x}))$ for the noise distribution $\mathcal{N}(\cdot)$. It is important to note that the FP and the forward operator A are not necessarily identical, in fact, the FP is often a black box and the forward operator is an approximate model of FP. A typical inverse problem can be solved by minimizing the loss function

$$\text{MSE}(A(\mathbf{x}), \mathbf{y}) = \|A(\mathbf{x}) - \mathbf{y}\|_2^2 \tag{8}$$

with respect to $\mathbf{x}$ with the help of prior assumptions on $\mathbf{x}$.

In our communication setting, the source signal $\mathbf{x}$ goes through the encoder, channel, and decoder of the SotA DeepJSCC model in [11]. We assume that the FP is only one realization of the DeepJSCC model. We have the knowledge of channel statistics, but not the noise realizations. Therefore, the FP is only partially known due to the stochastic channel noise. Our goal is to invert the FP and reconstruct the source signal $\mathbf{x}$ from the observations $\hat{\mathbf{x}}$ with high perceptual quality. We model the encoder and the decoder functions of the DeepJSCC model [11], i.e., f_θ and g_ϕ, as DNNs parameterized by θ and ϕ, respectively. Then, we represent our forward operator A as a non-linear process with AWGN noise function η, i.e.,

$$\begin{aligned} \mathbf{y} &= A(\mathbf{x}, \eta) && (9)\\ &= g_\phi(\eta(f_\theta(\mathbf{x}), \sigma^2)) && (10)\\ &= \hat{\mathbf{x}}. && (11) \end{aligned}$$

Extreme physical conditions, such as very low BCR or very low SNR in the communication pipeline introduce high distortion in the signal reconstructed by DeepJSCC, $\hat{\mathbf{x}}$. This might lead to significant loss of semantic similarity between the original signal and the reconstruction, especially when the source data distribution is complex and high dimensional (e.g., high-quality images or videos). In classical signal reconstruction methods, some prior knowledge about the properties of $\mathbf{x}$, such as sparsity, dictionary, or geometric properties, are used to minimize the loss (8) effectively. Here, our assumption is that the distribution of our input signal, i.e., images from a certain dataset, is in the range of a GAN generator function $G : \mathbb{R}^q \to \mathbb{R}^m$, which is parameterized by DNNs trained on the same context as the input distribution. In general, DGMs have recently demonstrated unprecedented visual results for image generation and have been shown to represent complex-high dimensional distributions in inverse problems successfully [31], [34], [36]. Primary examples for DGMs are VAEs [37], Diffusion Models [38] and GANs [39]. In this paper, we use the StyleGAN-2 [23] generator in the InverseJSCC approach to the semantic communication problem defined in Section III.

In InverseJSCC, we are inspired by a general inverse problem approach called intermediate layer optimization (ILO) [36], which utilizes a pre-trained StyleGAN-2 generator for various inverse problems. Although we are mainly interested in high-quality face images, the same ideas can be applied to other domains. Here, ILO solves an inverse problem objective by adaptively changing the StyleGAN-2 layer to be optimized, moving from the initial latent vector to intermediate layers closer to the output. We solve a slightly modified optimization problem:

$$\begin{aligned} \min_{\mathbf{w}, \mathbf{n}_G} \; & \lambda_1 \|A(G(\mathbf{w}, \mathbf{n}_G), \eta) - \hat{\mathbf{x}}\|_2^2 + \lambda_2 \mathcal{L}(A(G(\mathbf{w}, \mathbf{n}_G), \eta), \hat{\mathbf{x}}) \\ & + \lambda_3 \mathcal{R}(\mathbf{w}, \hat{\mathbf{x}}, G, \mathbf{n}_G), \end{aligned} \tag{12}$$

where $G(\cdot)$ is the pre-trained generator network, and $\mathbf{w}$ and $\mathbf{n}_G$ are the latent input and the noise vector of the generator, respectively. While the latent vector controls the style parts of the generated image, the noise fed to the generator determines the high-resolution details. $\mathcal{L}(\cdot)$ is the LPIPS loss function that uses VGG [27], and $\mathcal{R}$ is the regularization loss defined by

$$\mathcal{R} = \lambda_4 \mathcal{L}(G(\mathbf{w}, \mathbf{n}_G), \hat{\mathbf{x}}) + \lambda_5 \text{GEO}(\mathbf{w}). \tag{13}$$

Here, $\text{GEO}(\cdot)$ is the geodesic distance which represents the shortest path between two points on a curved surface. This term regularizes the deviation of the latent vector $\mathbf{w}$. Unlike in ILO, we use an additional LPIPS regularization term that helps make sure that the generator output does not divert far from the reconstructed input perceptually.

Recall that the forward operator A is the DeepJSCC model in Fig. 2, which includes a pre-trained encoder/decoder pair (f_θ, g_ϕ) and the channel model $\eta(\cdot, \sigma^2)$, and is a partial description of the FP due to the stochasticity of the channel. Fig. 3 shows the InverseJSCC scheme that reconstructs $\mathbf{x}$ from its distorted measurements $\hat{\mathbf{x}}$ with the help of the pre-trained generator. The forward operator A can be trained on any dataset regardless of the source dataset, whereas StyleGAN-2 generator G is pre-trained on the same context data as the source dataset. The generator G can be decomposed into multiple layers as

$$G = G_4 \circ G_3 \circ G_2 \circ G_1, \text{ where } \begin{cases} G_1 : & \mathbb{R}^q \to \mathbb{R}^{t_1} \\ G_2 : & \mathbb{R}^{t_1} \to \mathbb{R}^{t_2} \\ G_3 : & \mathbb{R}^{t_2} \to \mathbb{R}^{t_3} \\ G_4 : & \mathbb{R}^{t_3} \to \mathbb{R}^{m}. \end{cases} \tag{14}$$

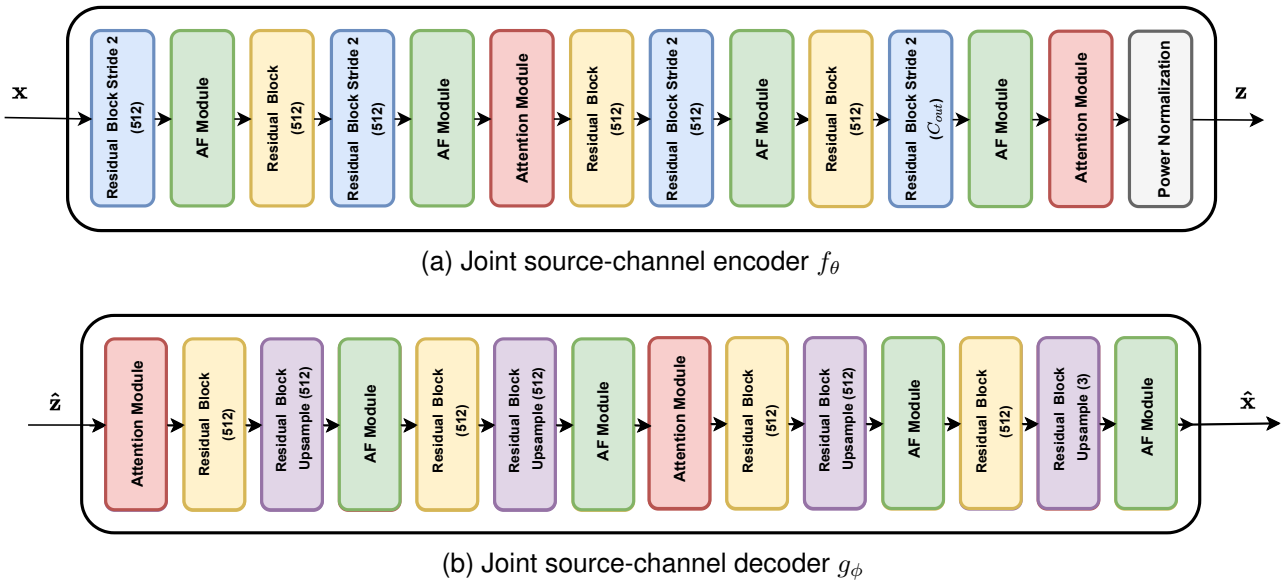


Fig. 4. DNN architecture of (a) the encoder and (b) the decoder of the forward operator A.

Similarly to ILO, the optimization in (12) is performed over $\mathbf{w}$ and $\mathbf{n}_G$. The optimized latent vector $\mathbf{w}^*$, which is the input to the first layer of the generator G_1, is used for the intermediate latent representation of the input to G_2, i.e., $\hat{\mathbf{w}} = G_1(\mathbf{w}^*)$. This recursive optimization continues until the intermediate latent inputs are optimized for all the intended number of layers. The noise vector $\mathbf{n}_G$, on the other hand, is optimized up to a pre-determined layer. Note that the weights of the pre-trained A and G functions are fixed, and only the latent input and noise vectors are optimized.

InverseJSCC has the same encoder/decoder structure used for the key frame structure in [11], which is shown in Fig. 4. We also incorporate the same residual blocks, attention modules, and attention feature (AF) modules proposed in [14]. AF module allows training a single model that works for a range of SNRs by randomly choosing the channel SNR during training and providing the AF modules with the current SNR. A single model trained with AF modules is shown by [14] to perform as well as the models trained for each SNR individually. This is further extended to an OFDM and MIMO channels in [40] and [41], respectively. Moreover, StyleGAN-2 generator structure and the recursive optimization of the intermediate latent vectors are the same as ILO [36]. We apply our own objective as well as the forward operator as shown in Eqn. (12) and Fig. 3.

Our forward operator A lies within the class of non-linear and stochastic forward operators in the inverse problem taxonomy [31], which is only partially known during test time due to the stochastic noise function η. The novelty of our InverseJSCC approach is that inverse problem solution has not been considered in the context of wireless communication, and partially known forward operators have not been investigated in this context yet. In the numerical results section, we show for the image domain that InverseJSCC provides perceptually high-quality images that better preserve the semantics of the source image despite communication over the noisy channel. This is achieved by exploiting the remarkable representation quality that DGMs have achieved in the image domain in the past few years. In particular, the improvements that InverseJSCC provides compared to DeepJSCC become more significant as the physical condition of the channel deteriorates. Moreover, we also show in Section V that InverseJSCC allows transferability by generating high-quality images even though the forward operator is trained on a different dataset. This is due to the fact that the inverse problem approach is based on successfully inverting the FP with the help of function A, which gives the flexibility to use various forward operators.

B. End-to-end Semantic Communication

InverseJSCC, presented in the previous section, improves the perceptual quality of a pre-trained DeepJSCC by exploiting the generative capability of a pre-trained GAN. Moreover, it allows flexibility to use encoder/decoder pairs trained on datasets that are not the test set, since it solves an inverse problem, which is only sensitive to the forward operator A. On the other hand, in a scenario where we can train the whole encoder/decoder structure with the same dataset as the test dataset, the best approach would indeed be training an encoder/decoder pair in a supervised way. While InverseJSCC solves an optimization problem over the layers of the generator network in test time, the alternative GenerativeJSCC approach involves an end-to-end optimization during training, not testing.

Similarly to InverseJSCC, our goal in the end-to-end design is to maximize the semantic similarity of the reconstructed signal to the source image while also retaining its perceptual quality in the presence of extreme channel conditions. Accordingly, the GenerativeJSCC scheme is an end-to-end strategy based on jointly training the encoder and the decoder functions represented by DNNs parameterized by θ and ψ. The source image $\mathbf{x}$ is sent to the encoding function f_θ, which has the same structure as shown previously in Fig. 4a, and normalized

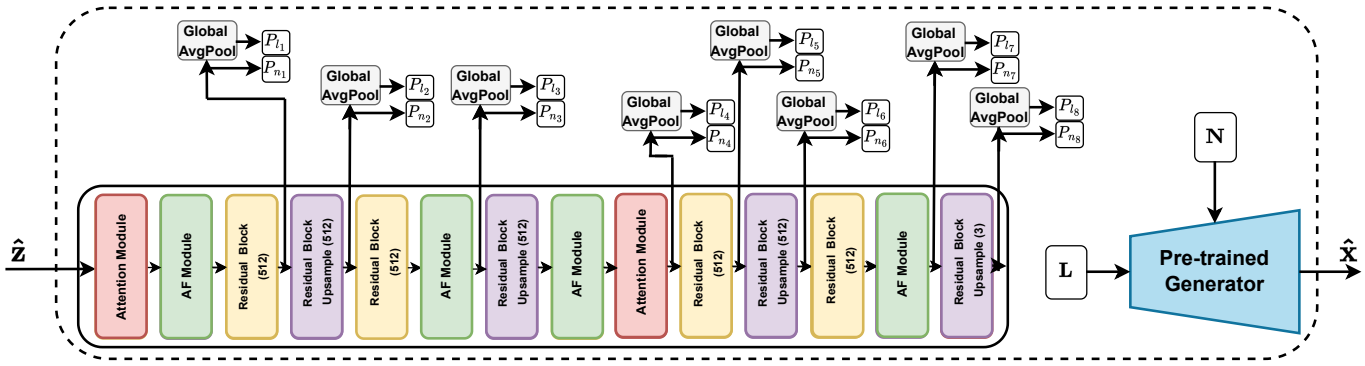


Fig. 5. DNN architecture of the decoder function g_ψ of GenerativeJSCC.

by (3) at the last layer to satisfy the power constraint. Then encoded signal $\mathbf{z}$ is transmitted over the AWGN channel with the transfer function (4). The receiver observes the corrupted signal $\hat{\mathbf{z}}$, and decodes it into an approximation of the source signal as follows:

$$\hat{\mathbf{x}} = g_\psi(\eta(f_\theta(\mathbf{x}), \sigma^2)), \tag{15}$$

where the decoding function g_ψ is represented by a DNN with parameters ψ. As we mentioned in the previous section, AF modules allow end-to-end training of a model that can be adapted for a range of channel SNRs. The current SNR is known by the encoder and the decoder, and it is provided to the AF modules of both models.

In GenerativeJSCC, we employ a weighted distortion metric as follows:

$$\min_{f_\theta, g_\psi} \gamma_1 \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \gamma_2 \mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}), \tag{16}$$

where γ_1 and γ_2 are the weights of the MSE and LPIPS loss $\mathcal{L}(\cdot)$. Here, we impose our requirement to maximize the semantic content of the source at the reconstruction by measuring it with the learned perception, i.e., LPIPS, and optimizing it together with MSE. Finally, given the distortion metric defined by Eqn. (16) between the original source signal and its reconstruction, the network weights (θ, ψ) are updated via backpropagation with respect to the gradient of the distortion.

We focus on the transmission of face images when designing the GenerativeJSCC architecture. While we use the same structure shown in Fig. 4a for the encoder, we propose a novel decoder structure that uses the StyleGAN-2 [23] generator architecture, and residual blocks introduced in [11] as well as AF modules.

The GenerativeJSCC decoder contains residual networks and the StyleGAN-2 generator that accepts two types of inputs: (1) the mapped latent vector $\boldsymbol{L}$, which determines the style parts of the generated image, and (2) the stochastic noise $\boldsymbol{N}$, which is typically drawn from a normal distribution and determines the high-resolution details. We model our decoder such that the noisy signal at the receiver input is transformed into the noise vector $\boldsymbol{N}$ and the initial latent vector $\tilde{\boldsymbol{L}}$, which is then mapped to $\boldsymbol{L}$ by a multi-layer linear network with Leaky ReLU activation. Fig. 5 shows the decoder architecture in more detail. While the channel output $\hat{\mathbf{z}}$ goes through the attention modules, AF modules and residual blocks introduced in [11]; the output of each residual block is split into three parts. The first split is followed by a global average pooling and a projection function P_l: a 1×1 convolution that produces parts of the initial latent vector $\tilde{\boldsymbol{L}}$. The second branch directly goes to a projection P_n: a 1×1 convolution that is used to produce the noise vector in the form of single channel images. The third split is followed by the next residual block. Then, the projections are concatenated as $\tilde{\boldsymbol{L}} = \text{Concat}\{P_{l_1}, P_{l_2}, \ldots, P_{l_8}\}$ and as $\boldsymbol{N} = \text{Concat}\{P_{n_1}, P_{n_2}, \ldots, P_{n_8}\}$, where subscripts represent the order of the intermediate layer outputs. Finally, the initial latent vector is mapped to the latent vector $\boldsymbol{L}$, and it is fed to the generator together with the noise vector to generate high-resolution face images. Note that the weights of the StyleGAN-2 generator are fixed during the training of the encoder/decoder pair of GenerativeJSCC.

We use a two-stage training scheme for learning the latent vector and the noise to maximize the perceptual quality of the reconstructions. The first stage only involves the reconstruction from the latent vector and disables learning of the noise maps. The second stage fine-tunes the model by training the layers learning the noise maps, which improves the details of the generated images. We implement this strategy by end-to-end training the model with the same learning rate for the P_ℓ layers and the backbone of the model, whereas we do not train the P_n layers initially. In the second stage, we train all layers of the model but we adjust the learning rate of the P_n layers to the initial learning rate of the backbone, whereas the learning rate for the backbone and P_ℓ is divided by 100.

V. Numerical Results

In this section, we present our experimental results to demonstrate and compare the performance of InverseJSCC and GenerativeJSCC. Both schemes are tested over an AWGN channel with the transfer function (4), while the generator network refers to StyleGAN-2 generator. StyleGAN-2 was trained on 70000 high-quality PNG images from the Flickr-Faces-HQ (FFHQ) dataset at a resolution of 1024x1024 over a week using eight Tesla V100 GPUs. We also use 512×512 CelebA-HQ dataset [26], which consists of 30000 high-quality celebrity images, and a subset of ImageNet dataset with 1050000 samples of size 256×256. Both datasets are split

as $8:1:1$ for training, validation, and testing, respectively. Our forward model A, DeepJSCC, is trained on the CelebA-HQ dataset over 18 hours using four NVIDIA GeForce RTX 3090 GPUs.

A. Performance Metrics

We consider various metrics to measure the distortion between the generated images and the input image throughout the numerical results: peak signal-to-noise-ratio (PSNR), multi-scale structural similarity index measure (MS-SSIM) and LPIPS. They are defined as follows:

$$\text{PSNR}(\mathbf{x},\hat{\mathbf{x}}) = 10\log_{10}\left(\frac{255^2}{\text{MSE}(\mathbf{x},\hat{\mathbf{x}})}\right) \text{ dB}. \tag{17}$$

$$\text{MS-SSIM}(\mathbf{x},\hat{\mathbf{x}}) = [l_M(\mathbf{x},\hat{\mathbf{x}})]^{\alpha_M}\prod_{j=1}^{M}[c_j(\mathbf{x},\hat{\mathbf{x}})]^{\beta_j}[s_j(\mathbf{x},\hat{\mathbf{x}})]^{\gamma_j}, \tag{18}$$

where

$$l_M(\mathbf{x},\hat{\mathbf{x}}) = \frac{2\mu_{\mathbf{x}}\mu_{\hat{\mathbf{x}}}+c_1}{\mu_{\mathbf{x}}^2+\mu_{\hat{\mathbf{x}}}^2+c_1}, \tag{19}$$

$$c_j(\mathbf{x},\hat{\mathbf{x}}) = \frac{2\sigma_{\mathbf{x}}\sigma_{\hat{\mathbf{x}}}+c_2}{\sigma_{\mathbf{x}}^2+\sigma_{\hat{\mathbf{x}}}^2+c_2}, \tag{20}$$

$$s_j(\mathbf{x},\hat{\mathbf{x}}) = \frac{\sigma_{\mathbf{x}\hat{\mathbf{x}}}+c_3}{\sigma_{\mathbf{x}}+\sigma_{\hat{\mathbf{x}}}+c_3}. \tag{21}$$

Here, $\mu_{\mathbf{x}}, \sigma_{\mathbf{x}}^2, \sigma_{\mathbf{x},\hat{\mathbf{x}}}^2$ are the mean and variance of $\mathbf{x}$, and the covariance between $\mathbf{x}$ and $\hat{\mathbf{x}}$, respectively. Moreover, α_M, β_j and γ_j are the weights of the components; while c_1, c_2 and c_3 are coefficients for numeric stability. Subscript j represents different downsampling scale of $\{\mathbf{x},\hat{\mathbf{x}}\}$. Similar to [11], we use the default parameters for $(\alpha_M, \beta_j, \gamma_j)$ in the original paper [42]. MS-SSIM has been shown to approximate human visual perception well on various image and video databases.

LPIPS loss was proposed to measure the perceptual similarity loss between two images [27], and it has been shown to match human perception well. It essentially computes the similarity between the activations of two image patches for some pre-defined network, such as VGG or AlexNet. Note that, a lower LPIPS score is better since it means that image patches are perceptually more similar.

B. InverseJSCC Results

We consider the scenario in which the channel SNR is in the range $[-5,5]$ dB while the BCR is $\rho=\{0.0013, 0.0052\}$. The DeepJSCC model reconstructs source images with high distortion due to the extreme physical conditions considered in this scenario. Note that the DeepJSCC models used as forward operators here are trained by optimizing the objective $d(\mathbf{x},\hat{\mathbf{x}}) = \text{MSE}(\mathbf{x},\hat{\mathbf{x}}) + \text{LPIPS}(\mathbf{x},\hat{\mathbf{x}})$. Our goal is to achieve a better perception quality in the reconstructed images without sacrificing the distortion performance with respect to the existing alternatives. Here, we follow an unsupervised approach and apply InverseJSCC to the distorted reconstructions at the output of the DeepJSCC decoder.

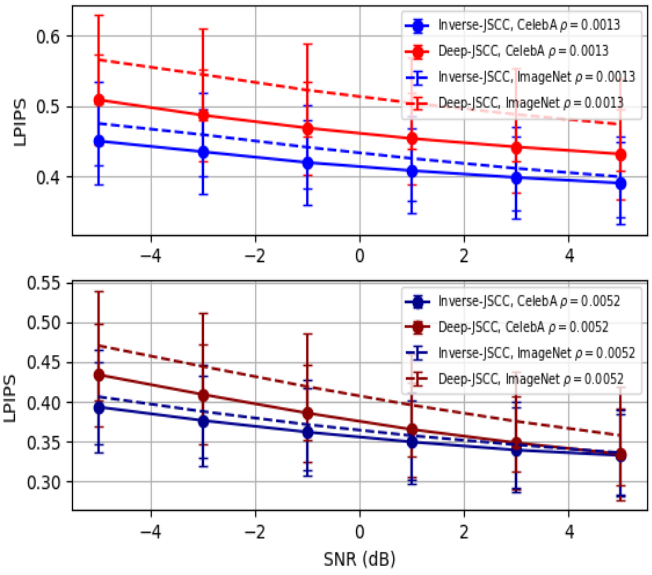
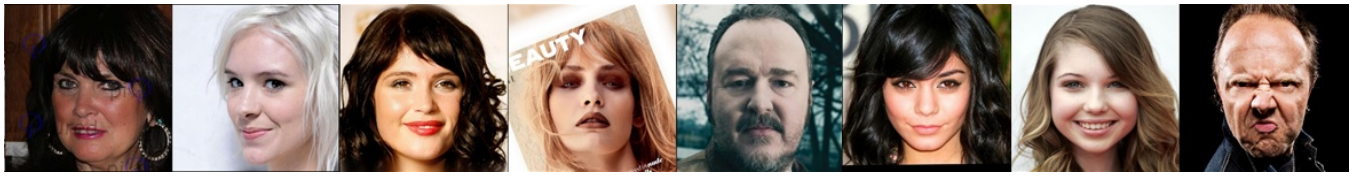


Fig. 6. LPIPS loss between the source and the reconstructed image w.r.t. channel SNR for DeepJSCC (trained with CelebA-HQ), InverseJSCC (A is trained with CelebA-HQ), DeepJSCC (trained with ImageNet) and InverseJSCC (A is trained with ImageNet) for $\rho=\{0.0013, 0.0052\}$.

We represent both the FP and the forward operator A of InverseJSCC with the SotA DeepJSCC architecture shown in Fig. 4. However, they are not identical due to the stochastic channel noise. We perform experiments for two major cases: First, each encoder/decoder pair representing the FP and A is pre-trained with CelebA-HQ dataset for the given ρ and SNR values, and the InverseJSCC is tested against the same dataset. In the second case, the FP and the forward operator A are both pre-trained with ImageNet dataset, and the InverseJSCC is tested against the CelebA-HQ dataset.

All DeepJSCC models are trained for $\text{SNR}_{\text{Train}} \in \{-5,-4,\dots,4,5\}$ dB using PyTorch framework [43]. To train the encoder and decoder networks of the FP and the forward operator A, we utilize Adam optimizer [37] with $\beta_1=0.9$, $\beta_2=0.99$ and an initial learning rate $l_r=0.0001$, which is multiplied by a factor 0.8 when the validation error does not improve in 4 epochs. We also set $\lambda_1=1, \lambda_2=1, \lambda=1$ and $\lambda_5=0.01$ as given in ILO, and we selected $\lambda_4=0.1$.

Fig. 6 shows the performances of DeepJSCC and InverseJSCC models in terms of LPIPS loss with respect to channel SNR (dB). The top plot depicts the results for $\rho=0.0013$ and the bottom for $\rho=0.0052$. "DeepJSCC, CelebA-HQ" and "DeepJSCC, ImageNet" represent the encoder/decoder pairs with the structure shown in Fig. 4 and trained with CelebA-HQ and ImageNet datasets, respectively. "InverseJSCC, CelebA-HQ" and "InverseJSCC, ImageNet" represent our proposed method described in Fig. 3 with the forward operators trained with CelebA-HQ and ImageNet datasets, respectively. These models represent the FP's and A's of the two major cases we described above. In Fig. 6, smaller LPIPS loss implies better perceptual similarity between the ground-truth and the generated images in the presented schemes. As expected, we observe better LPIPS score for both DeepJSCC and InverseJSCC models when the encoder/decoder networks are trained and



(a) Original images from the CelebA-HQ dataset.

(b) First row: Reconstructions by DeepJSCC(CelebA-HQ), Second row: InverseJSCC(CelebA-HQ), Third row: DeepJSCC(ImageNet) and Fourth row: InverseJSCC(ImageNet), respectively, for SNR$= 1$.

(c) First row: Reconstructions by DeepJSCC(CelebA-HQ), Second row: InverseJSCC(CelebA-HQ), Third row: DeepJSCC(ImageNet) and Fourth row: InverseJSCC(ImageNet), respectively, for SNR$= 5$.

Fig. 7. Original and reconstructed CelebA-HQ images, when DeepJSCC model and the forward operator A of InverseJSCC are trained on ImageNet for $\rho = 0.0013$ and $SNR = \{1, 3, 5\}$. Inverse problem solution in InverseJSCC allows A trained with a different dataset.

tested with the CelebA-HQ dataset. However, in both cases, InverseJSCC improves the perceptual similarity (measured by LPIPS) of the reconstructed images by the DeepJSCC model significantly, even when there is a domain mismatch.

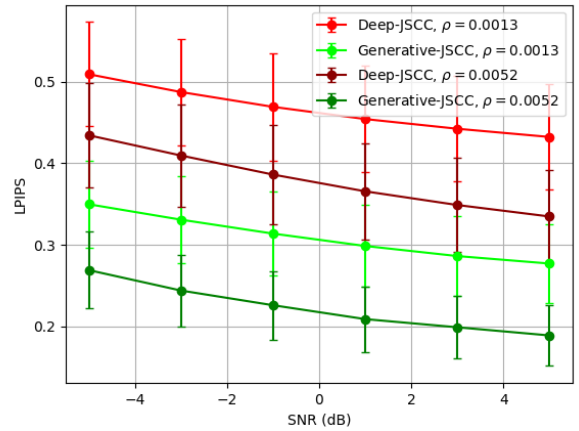
The noteworthy effect of InverseJSCC on the reconstructed images can be seen in Fig. 7 clearly. Fig. 7a shows the ground-truth images from CelebA-HQ test-set; Fig.7b and Fig.7c both show the reconstructed images at the output of Deep-JSCC (trained with CelebA-HQ), InverseJSCC (A is trained with CelebA-HQ), DeepJSCC (trained with ImageNet) and InverseJSCC (A is trained with ImageNet) from top to bottom when the channel SNR is 1 dB and 5 dB, respectively. The facial details introduced by InverseJSCC crucially improve the perceptual quality of the images reconstructed by DeepJSCC. Hence, the generated images look as realistic as the ground truth. Moreover, InverseJSCC generates considerably higher quality face images even when the available encoder/decoder networks are trained on a different domain, i.e., ImageNet. This can be attributed to two reasons: (1) InverseJSCC inverts the FP regardless of its training set, and (2) the source distribution is in the range of the StyleGAN-2 generator, therefore, the $G(\cdot)$ function is able to generate realistic-looking face images that also look similar to the ground-truth.

C. GenerativeJSCC Results

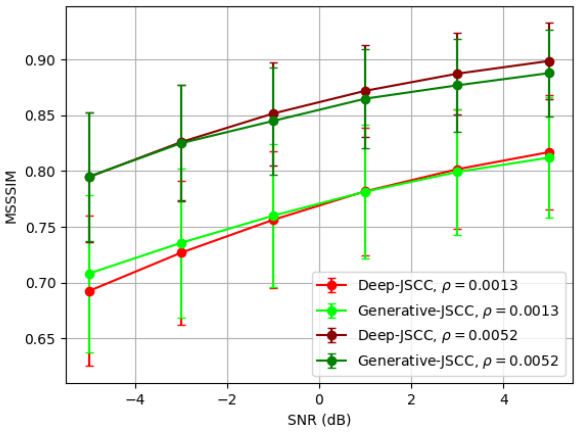
Here again, we consider the channel with SNR $[-5, 5]$ dB and the BCR of $\rho = \{0.0013, 0.0052\}$. We jointly train the encoder, f_θ, and the decoder, g_ψ, of the GenerativeJSCC model with CelebA-HQ dataset by optimizing the weighted sum of MSE and LPIPS losses between the ground-truth and the generated images (16). As before, we train the networks for a range of SNR utilizing the attention modules, i.e., $\text{SNR}_{Train} \in \{-5, -4, \dots, 4, 5\}$ dB, using PyTorch and Adam optimizer with the same parameters as in Section V-B.

Fig. 8 shows the comparison between DeepJSCC and GenerativeJSCC in terms of average LPIPS loss, MS-SSIM and PSNR with respect to channel SNR (dB). We observe that, for lower BCR, GenerativeJSCC outperforms DeepJSCC not only in terms of perceptual quality but also in terms of pixel-wise distortion. Considering that both models are trained to optimize the same objective given in (16), this shows the superior reconstructing capabilities of GenerativeJSCC through our novel decoder architecture. On the other hand, GenerativeJSCC outperforms DeepJSCC only in terms of perceptual similarity loss LPIPS in the larger BCR regime. This is in line with our claim that generative models in JSCC provide higher perceptual quality; and thus, more semantic similarity, than pixel-wise similarity, particularly in the edge cases with highly limited bandwidth and power resources.

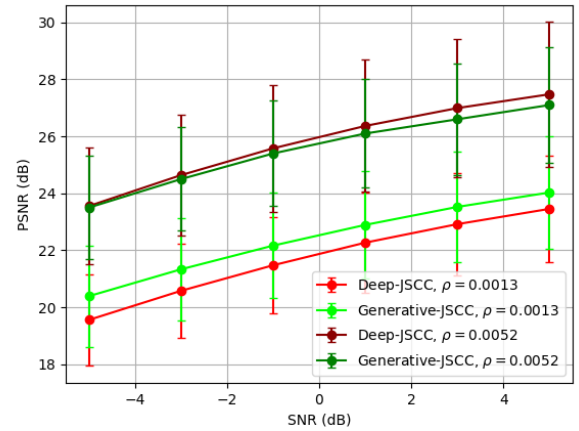
We show a visual comparison between the reconstructed images from both models in Fig. 9. When the channel SNR is as low as -5 dB, the performance difference between DeepJSCC and GenerativeJSCC is much more evident. This shows that the GAN-based decoder of GenerativeJSCC generates higher quality images for human perception, whereas DeepJSCC output distorts facial attributes significantly. As the channel SNR increases to 5 dB, the image quality difference between two models becomes less obvious, however, GenerativeJSCC still captures the image color and face details impressively better than DeepJSCC.



(a) LPIPS loss versus SNR.



(b) MS-SSIM versus SNR.



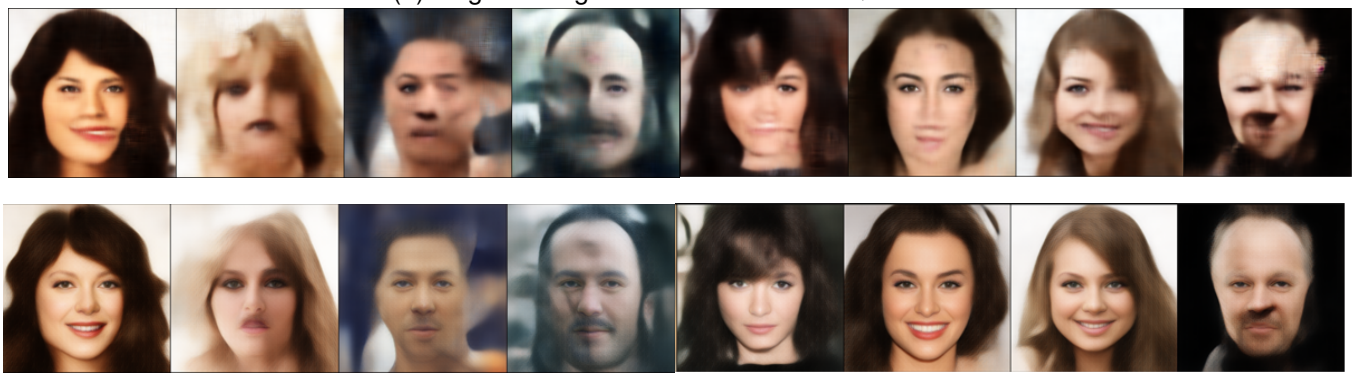
(c) PSNR versus SNR.

Fig. 8. Performance comparison between DeepJSCC and GenerativeJSCC in terms of (a) LPIPS, (b) MS-SSIM and (c) PSNR metrics w.r.t channel SNR for $\rho = \{0.0013, 0.0052\}$.

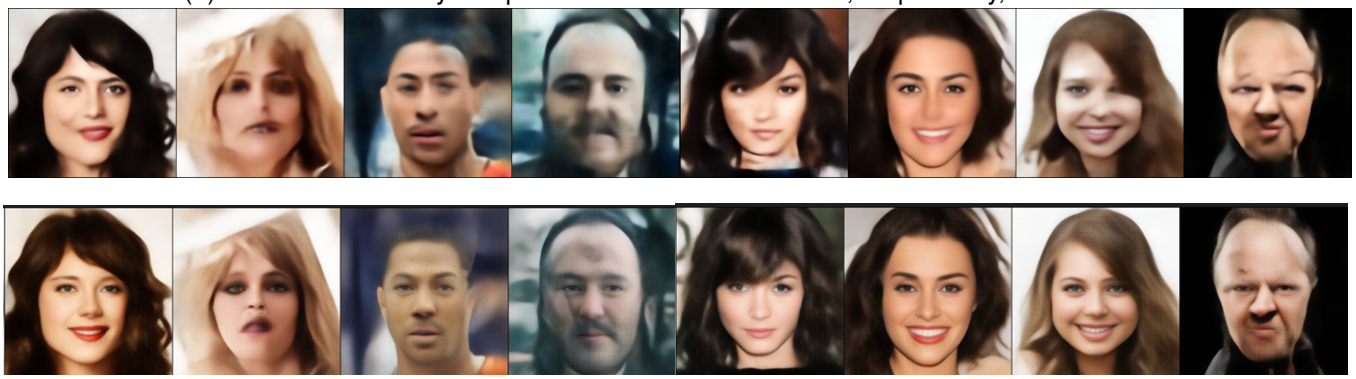
VI. Conclusion

In this paper, we presented two DL-based JSCC schemes that incorporate SotA DGMs to improve the perceptual quality

(a) Original images from the CelebA-HQ dataset.



(b) Reconstructions by DeepJSCC and GenerativeJSCC, respectively, for SNR$= -5$.



(c) Reconstructions by DeepJSCC and GenerativeJSCC, respectively, for SNR$= -1$.

(d) Reconstructions by DeepJSCC and GenerativeJSCC, respectively, for SNR$= 5$.

Fig. 9. Original and reconstructed CelebA-HQ images by DeepJSCC and GenerativeJSCC models for $\rho = 0.0013$ and $SNR = \{-5, -1, 5\}$.

of reconstructed images. Both approaches are proposed to tackle edge cases such as low bandwidth and low channel SNR, in which the conventional DeepJSCC scheme typically results in reconstructed images with poor perceptual quality. Our first scheme, InverseJSCC, takes a novel inverse problem approach to the wireless image transmission problem, and improves the classical DeepJSCC scheme by recovering the source image with the help of a StyleGAN-2 generator. It maximizes the semantic similarity between the input and the generated images in terms of the LPIPS metric, which is widely accepted to capture human perception well. Additionally, InverseJSCC performs considerably well when there is a mismatch between the image distributions in training and inference times. Our second approach, GenerativeJSCC, is an end-to-end scheme consisting of an encoder, a non-trainable channel, and a StyleGAN-based decoder. Our results show that GenerativeJSCC outperforms DeepJSCC in terms of perceptual quality metrics learned from human judgments, as well as the distortion metrics used in classical communications.

REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[2] V. Kostina and S. Verdú, "Lossy joint source-channel coding in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 59, no. 5, pp. 2545–2575, 2013.

[3] J. Park, S. Samarakoon, H. Shiri, M. K. Abdel-Aziz, T. Nishio, A. Elgabli, and M. Bennis, "Extreme URLLC: vision, challenges, and key enablers," *CoRR*, vol. abs/2001.09683, 2020.

[4] Z. Li, M. A. Uusitalo, H. Shariatmadari, and B. Singh, "5G URLLC: Design challenges and system concepts," in *2018 15th International Symposium on Wireless Communication Systems (ISWCS)*, pp. 1–6, 2018.

[5] A. Destounis, D. Tsilimantos, M. Debbah, and G. S. Paschos, "Learn2mac: Online learning multiple access for URLLC applications," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 1–6, 2019.

[6] A. Azari, M. Ozger, and C. Cavdar, "Risk-aware resource allocation for URLLC: Challenges and strategies with machine learning," *IEEE Communications Magazine*, vol. 57, no. 3, pp. 42–48, 2019.

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

[8] D. B. Kurka and D. Gündüz, "Joint source-channel coding of images with (not very) deep learning," in *International Zurich Seminar on Information and Communication (IZS 2020). Proceedings*, pp. 90–94, ETH Zurich, 2020.

[9] D. B. Kurka and D. Gündüz, "DeepJSCC-f: Deep joint source-channel coding of images with feedback," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 178–193, 2020.

[10] T.-Y. Tung, D. B. Kurka, M. Jankowski, and D. Gündüz, "Deepjscc-q: Constellation constrained deep joint source-channel coding," *IEEE Journal on Selected Areas in Information Theory*, pp. 1–1, 2022.

[11] T.-Y. Tung and D. Gündüz, "Deepwive: Deep-learning-aided wireless video transmission," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 9, pp. 2570–2583, 2022.

[12] M. Yang, C. Bian, and H.-S. Kim, "OFDM-guided deep joint source channel coding for wireless multipath fading channels," *IEEE Transactions on Cognitive Communications and Networking*, 2022.

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

[14] 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 Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 2315–2328, 2021.

[15] K. Choi, K. Tatwawadi, A. Grover, T. Weissman, and S. Ermon, "Neural joint source-channel coding," in *International Conference on Machine Learning*, pp. 1182–1192, PMLR, 2019.

[16] D. Gündüz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE Journal on Selected Areas in Communications*, pp. 1–1, 2022.

[17] M. Jankowski, D. Gündüz, and K. Mikolajczyk, "Joint device-edge inference over wireless links with pruning," in *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 1–5, IEEE, 2020.

[18] C.-H. Lee, J.-W. Lin, P.-H. Chen, and Y.-C. Chang, "Deep learning-constructed joint transmission-recognition for internet of things," *IEEE Access*, vol. 7, pp. 76547–76561, 2019.

[19] M. Jankowski, D. Gündüz, and K. Mikolajczyk, "Wireless image retrieval at the edge," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 89–100, 2020.

[20] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. V. Gool, "Generative adversarial networks for extreme learned image compression," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 221–231, 2019.

[21] Y. Blau and T. Michaeli, "Rethinking lossy compression: The rate-distortion-perception tradeoff," in *International Conference on Machine Learning*, pp. 675–685, PMLR, 2019.

[22] M. Tschannen, E. Agustsson, and M. Lucic, "Deep generative models for distribution-preserving lossy compression," *Advances in neural information processing systems*, vol. 31, 2018.

[23] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. CVPR*, 2020.

[24] C. A. Christopoulos, A. N. Skodras, and T. Ebrahimi, "The JPEG2000 still image coding system: An overview," *IEEE Trans. Consumer Electron.*, vol. 46, pp. 1103–1127, 2000.

[25] R. G. Gallager, "Low-density parity-check codes," *IRE Transactions on Information Theory*, vol. 8, pp. 21–28, Jan. 1963.

[26] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[27] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

[28] Y. M. Saidutta, A. Abdi, and F. Fekri, "VAE for joint source-channel coding of distributed Gaussian sources over AWGN MAC," in *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 1–5, IEEE, 2020.

[29] T. Marchioro, N. Laurenti, and D. Gündüz, "Adversarial networks for secure wireless communications," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8748–8752, IEEE, 2020.

[30] E. Erdemir, P. L. Dragotti, and D. Gündüz, "Privacy-aware communication over a wiretap channel with generative networks," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2989–2993, IEEE, 2022.

[31] G. Ongie, A. Jalal, C. A. Metzler, R. G. Baraniuk, A. G. Dimakis, and R. Willett, "Deep learning techniques for inverse problems in imaging," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 39–56, 2020.

[32] A. Mousavi, A. B. Patel, and R. G. Baraniuk, "A deep learning approach to structured signal recovery," in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, p. 1336–1343, IEEE Press, 2015.

[33] D. L. Donoho, "Compressed sensing," *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[34] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, "Compressed sensing using generative models," in *International Conference on Machine Learning*, pp. 537–546, PMLR, 2017.

[35] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, "PULSE: Self-supervised photo upsampling via latent space exploration of generative models," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 2437–2445, 2020.

[36] G. Daras, J. Dean, A. Jalal, and A. Dimakis, "Intermediate layer optimization for inverse problems using deep generative models," vol. 139, pp. 2421–2432, 18–24 Jul 2021.

[37] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[38] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[39] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[40] H. Wu, Y. Shao, K. Mikolajczyk, and D. Gündüz, "Channel-adaptive wireless image transmission with ofdm," *IEEE Wireless Communications Letters*, vol. 11, no. 11, pp. 2400–2404, 2022.

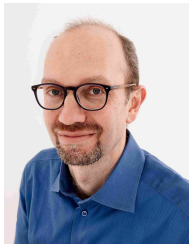
[41] H. Wu, Y. Shao, C. Bian, K. Mikolajczyk, and D. Gündüz, "Vision transformer for adaptive image transmission over mimo channels," *ArXiv*, vol. abs/2210.15347, 2022.

[42] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2, pp. 1398–1402, Ieee, 2003.

[43] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," 2017.

Ecenaz Erdemir (Member, IEEE) is an Applied Scientist at Amazon Web Services (AWS) AI, New York, NY, U.S.A. She received her B.S. and M.S. degrees in Electrical and Electronics Engineering from Middle East Technical University (METU), Ankara, Turkey, in 2014 and 2017, respectively. She received her Ph.D. degree in the Department of Electrical and Electronic Engineering at Imperial College London (ICL), London, U.K., in 2022. Her current research interests are information security, cybersecurity, data privacy, machine learning, information theory and decision theory.

Tze-Yang Tung (Member, IEEE) is a Member of Technical Staff at Nokia Bell Labs. He received his PhD and BEng degree from the department of Electrical and Electronic Engineering at Imperial College London. He also received his MSc degree from the University of Southern California in Electrical and Computer Engineering. His current research lies primarily in the junction between machine learning and communications that enable machine learning. In particular, he is interested in the emerging semantic and goal oriented communications paradigm and its relations to machine learning applications.



Pier Luigi Dragotti (Fellow, IEEE) received the Laurea degree (summa cum laude) in electronic engineering from the University of Naples Federico II, Naples, Italy, in 1997, and the master's degree in communications systems and the Ph.D. degree from the Swiss Federal Institute of Technology of Lausanne (EPFL), Switzerland, in 1998 and in April 2002, respectively. He has held several visiting positions, in particular, he was a Visiting Student at Stanford University, Stanford, CA, USA, in 1996; a Summer Researcher at Bell Labs, Lucent Technologies, Murray Hill, NJ, USA, in 2000; a Visiting Scientist with the Massachusetts Institute of Technology (MIT) in 2011; and a Visiting Scholar at Trinity College, Cambridge, UK, in 2020. Before joining Imperial College London in November 2002, he was a Senior Researcher at EPFL working on distributed signal processing for the Swiss National Competence Center in Research on Mobile Information and Communication Systems. He is currently Professor of Signal Processing with the Department of Electrical and Electronic Engineering, Imperial College London. His research interests include sampling theory and its applications, computational imaging, and sparsity-driven signal processing. Dr. Dragotti was an Elected Member of the IEEE Image, Video and Mul- tidimensional Signal Processing Technical Committee as well as an Elected Member of the IEEE Signal Processing Theory and Methods Technical Com- mittee and the IEEE Computational Imaging Technical Committee. In 2011, he was awarded the Prestigious ERC Starting Investigator Award (consolidator stream). He was also IEEE SPS Distinguished Lecturer (2021-2022), the Editor-in-Chief of the IEEE TRANSACTIONS ON SIGNAL PROCESSING (2018–2020), the Technical Co-Chair of the European Signal Processing Conference in 2012, and an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING from 2006 to 2009.

Deniz Gündüz (Fellow, IEEE) received the B.S. degree in electrical and electronics engineering from METU, Turkey in 2002, and the M.S. and Ph.D. degrees in electrical engineering from NYU Tandon School of Engineering (formerly Polytechnic University) in 2004 and 2007, respectively. Currently, he is a Professor of Information Processing in the Electrical and Electronic Engineering Department at Imperial College London, UK, where he also serves as the deputy head of the Intelligent Systems and Networks Group. He has held visiting/part-time positions at the University of Modena and Reggio Emilia, University of Padova, Princeton University, Stanford University and CTTC. His research interests lie in the areas of communications and information theory, machine learning, and privacy. Dr. Gündüz is a Fellow of the IEEE, and a Distinguished Lecturer for the IEEE Information Theory Society (2020-22). He serves in editorial roles for the IEEE Transactions on Information Theory, IEEE Transactions on Communications, IEEE Journal on Selected Areas in Communications (JSAC), and the IEEE Transactions on Wireless Communications. He is the recipient of the IEEE Communications Society - Communication Theory Technical Committee (CTTC) Early Achievement Award in 2017, Starting (2016) and Consolidator (2022) Grants of the European Research Council (ERC), and several best paper awards.