OFDM-guided Deep Joint Source Channel Coding for Wireless Multipath Fading Channels

Mingyu Yang, Student Member, IEEE, Chenghong Bian, and Hun-Seok Kim, Member, IEEE

Abstract—We investigate joint source channel coding (JSCC) for wireless image transmission over multipath fading channels. Inspired by recent works on deep learning based JSCC and model-based learning methods, we combine an autoencoder with orthogonal frequency division multiplexing (OFDM) to cope with multipath fading. The proposed encoder and decoder use convolutional neural networks (CNNs) and directly map the source images to complex-valued baseband samples for OFDM transmission. The multipath channel and OFDM are represented by non-trainable (deterministic) but differentiable layers so that the system can be trained end-to-end. Furthermore, our JSCC decoder further incorporates explicit channel estimation, equalization, and additional subnets to enhance the performance. The proposed method exhibits 2.5 – 4 dB SNR gain for the equivalent image quality compared to conventional schemes that employ state-of-the-art but separate source and channel coding such as BPG and LDPC. The performance further improves when the system incorporates the channel state information (CSI) feedback. The proposed scheme is robust against OFDM signal clipping and parameter mismatch for the channel model used in training and evaluation.

Index Terms—Joint source channel coding, deep neural networks, OFDM, model-driven machine learning

I. INTRODUCTION

S HANNON’S separation theorem [1] states that, if the minimum achievable source coding rate of a given source is less than the channel capacity, then the source can be reliably transmitted through the channel. Thus, the source and channel coding can be designed separately without loss of optimality. This famous theorem is the basis for the tremendous success of practical communication systems because of the modularity it affords. For example, most modern systems for wireless image transmission first compress the image with a source coding algorithm (e.g., JPEG, WebP, BPG) and then encode the bit stream with a source-independent channel code (e.g., LDPC, Polar, Turbo, etc.) as shown in Figure 1 top. However, to reach the optimality of the separation theorem, we require infinite codeword length which is not practical for applications that impose low latency real-time communication and/or low computation complexity constraints such as autonomous driving and the Internet of Things. Moreover, when the channel quality falls below a certain threshold, the channel coding may not provide satisfying error corrections and the source coding will tend to fail catastrophically. As a result, it is beneficial to jointly optimize source coding and channel coding for relatively short messages, which is referred to as joint source-channel coding (JSCC). There are various schemes to jointly optimize different components in source and channel coding such as vector quantization and index assignment [2] [3] [4] [5]. However, these methods cannot provide satisfying performance and most systems still adopt separate coding methods.

In recent years, deep learning has been successfully applied to a wide range of areas such as computer vision [6] [7], natural language processing [8] [9] and wireless communications [10] [11] [12], achieving significant performance improvement over analytical (i.e., non-data-driven) algorithms. Its ability to extract complex features from images has led to applications of deep JSCC over simple channel models such as AWGN and Rayleigh flat fading channels. The authors in [13] first introduce deep learning (DL) to JSCC by using RNNs for text transmission, and in [14] the authors apply deep JSCC for analog data compression and storage. The first use of deep JSCC for wireless image transmission is reported in [15], where the authors incorporate both source and channel coding into an auto-encoder structure, and source compression and error correction coding are jointly learned through back-propagation optimization. The proposed deep JSCC algorithm in [15] outperforms conventional separate schemes under AWGN and Rayleigh flat fading channels. Later, this scheme is extended with larger neural networks.

M. Yang, C. Bian and H. Kim are with the Department of Electrical and Computer Engineering, University of Michigan, Ann Arbor, MI, 48109 USA e-mail: (mingyu@umich.edu; cblian@umich.edu; hunseok@umich.edu)

This work was funded in part by DARPA YFA #D18AP00076 and NSF CAREER #1942806.

The source code is available at https://github.com/mingyuyng/OFDM-guided-JSCC.
feedback [17], progressive transmission [18] or additional attention modules that deal with different SNRs [19][20]. The idea of deep JSCC is also adopted for discrete channels such as the binary symmetric channel (BSC) [21][22]. Due to the discrete nature, the authors perform end-to-end optimization in a variational manner and optimize a lower bound of mutual information between the source images and noisy channel outputs. Different from prior JSCC works, we consider JSCC in the more realistic and challenging multipath fading channels. To the best of our knowledge, there is no existing JSCC work specifically designed and optimized for the multipath fading channels.

Directly replacing the simple channel model such as AWGN with the multipath fading channel in the prior JSCC works may lead to sub-optimal performance due to the inter-symbol interference. Motivated by model-based deep learning approaches [23][24][25][26][27][28], we incorporate orthogonal frequency division multiplexing (OFDM) as our domain knowledge, where equalization is performed in frequency domain for orthogonal sub-carriers without inter-symbol interference. Instead of naively treating a neural network as a black box, our model-based deep learning framework imposes a guided structure for neural networks with analytical models and signal processing blocks designed using the expert domain knowledge. The proposed model-based structure shortens the training time and also improves the system performance.

Recently, substantial effort has been made to incorporate end-to-end learning based communication systems with OFDM as domain knowledge. The authors in [29] first introduce an end-to-end learning-based OFDM system, where an autoencoder maps the transmitted binary bits to baseband channel inputs and recovers the bit information from noisy channel output. This method can be further extended to reduce the high peak-to-average power ratio (PAPR) in OFDM [30]. In [31], the authors consider one-bit quantization of the received OFDM signal. A pilot-less end-to-end learning for OFDM is studied in [32]. However, those prior works only focus on learning modulation constellation designs for arbitrary binary bit sequence inputs without considering the source coding. Other prior works such as [10][33][34][35] limit the scope to deep learning based channel estimation and detection instead of the end-to-end OFDM system.

In this paper, we propose a deep learning based JSCC scheme for wireless image transmission under multipath fading channels. Unlike prior works [15][17][18] where the encoded images are directly transmitted through the channel, we concatenate deep neural networks with OFDM processing blocks by feeding the neural network encoded image as frequency domain OFDM baseband symbols. The multipath channel and OFDM baseband processing blocks are all represented by non-trainable but differentiable layers, and the whole system can be trained in an end-to-end manner including the multipath channel model.

Numerical experiments show that combining JSCC with OFDM in the proposed way outperforms the direct channel transmission used in prior works [15][17][18]. Moreover, we demonstrate that adopting more domain-knowledge-driven structures such as channel estimation and equalization with optional channel state information (CSI) feedback further improves the system performance. Our results show that the proposed JSCC-OFDM method outperforms the baseline separate source and channel scheme with a state-of-the-art image compression method and error correction code applied to OFDM. Furthermore, enlightened by recent works in image compression [36][37], we demonstrate that our scheme can be combined with adversarial training to produce more visually appealing results. Finally, we show that our proposed method is robust to additional signal clipping-based PAPR reduction [38], a wide range of noise power levels, and various channel conditions that the neural network has never seen during training.

Main contributions of this paper are summarized as follows:

- To the best of our knowledge, it is the first to extend deep learning based JSCC to the challenging multipath channels. We jointly train deep neural networks with explicit OFDM blocks to combat the multipath fading and enhance the system performance.
- We design a decoder structure that combines deep neural networks with additional non-trainable blocks such as PAPR reduction signal-clipping, channel estimation and channel equalization to exploit the domain expert knowledge. We conduct detailed ablation studies and verify that this elaborate structure outperforms a blackbox network that solely relies on data-driven training.
- The proposed JSCC is evaluated for wireless image transmission compared to a traditional modularized scheme using a state-of-the-art image compression codec and separate error correction code. It is shown that the proposed JSCC significantly (2.5 - 4dB SNR gain) outperforms the baseline.
- We evaluate a CSI-feedback based pre-coding scheme to allocate the power for each subcarrier. Our results quantify the gain from the proposed pre-coding that outperforms conventional water-filling algorithms applied to the baseline.
- We evaluate a new adversarial training scheme for JSCC and demonstrate that it enhances the image perceptual quality. It is also shown that the proposed JSCC method is robust to various channel conditions such as noise power, number of multipaths that the neural networks have not seen during training.

II. PROPOSED METHOD

A. Deep JSCC in multipath fading channels

The general framework for deep learning based JSCC is shown in Figure[1], bottom. In our approach, a pair of encoder $E_\theta$ and decoder $D_\phi$ directly map the real-valued source (image) input $x$ to complex-valued (modulated) baseband samples $y$ and estimate the source signal $\hat{x}$ from noisy multipath fading channel outputs $\hat{y}$. Here, $\theta$ and $\phi$ represent the parameters for the encoder and decoder neural networks respectively. The channel can be represented with the conditional probability $p(\hat{y}|y;\epsilon)$ where $\epsilon$ denotes the parameters that describe the channel statistics. The channel model (1) is defined in the
discrete time domain assuming interpolation and Nyquist sampling for the channel inputs and outputs.

We parameterize the multipath fading channel using a discrete channel transfer function:

\[
y = h(y; \sigma_0^2, \ldots, \sigma_{L-1}^2, \sigma^2) = h * y + w,
\]

where \(*\) denotes the convolution operation, \(h \in \mathbb{C}^L\) denotes the sample space channel impulse response, and \(L\) is the number of multipaths. \(w \sim \mathcal{CN}(0, \sigma^2 I_k)\) represents the additive Gaussian noise. Each path experiences independent Rayleigh fading satisfying \(h_l \sim \mathcal{CN}(0, \sigma_l^2)\) for \(l = 0, 1, \ldots, L - 1\). The power for each path follows an exponential decay profile \(\sigma_l^2 = \alpha_l e^{-\frac{L}{\gamma}},\) where \(\alpha_l\) is a normalization coefficient to satisfy \(\sum_{l=0}^{L-1} \alpha_l^2 = 1\). \(\gamma\) is the time decay (or delay spread) constant. Note that the channel transfer function is fully differentiable for each \(y\) given the realizations of random channel parameters. This means that the gradients from the decoder can propagate back to the encoder through a particular realization of the multipath fading channel \(h\) when the end-to-end system is trained with gradient descent methods.

### B. JSCC with OFDM extension

We extend the JSCC framework to an OFDM-based system to efficiently mitigate the multipath fading channel with simple single-tap frequency domain equalization. We assume that each source image \(x\) is transmitted in a single OFDM packet that contains \(N_p\) pilot symbols and \(N_s\) information symbols. For channel estimation, we adopt block-type pilots by sending known symbols on all subcarriers. For simplicity, we assume that the transmitter and receiver are synchronized in time (within the cyclic-prefix length \(L_{c,p}\)) without any carrier frequency offset.

The OFDM extension to the JSCC framework is shown in Figure 2. Given the input image dimension of \(H \times W\), the encoder output dimension is \(C \times \frac{H_d}{d} \times \frac{W_d}{d}\), where \(d\) is the down-sampling factor and \(C\) is the number of output channels. At this point, the encoder output is real-valued. Then, the encoder output is converted to complex numbers by mapping one half to the real part and the other half to the imaginary part. Then it is reshaped to \(Y \in \mathbb{C}^{N_N \times L_{f,f} \times C}\), where \(L_{f,f}\) denotes the number of subcarriers in an OFDM symbol and \(Y\) represents the frequency domain complex-valued symbols to be fed to the OFDM transmitter. Note that \(N_s = \frac{d}{2} \times N_p L_{f,f}\), which is the number of OFDM information symbols. In all of our experiments, the parameters are carefully chosen so that \(N_s\) is an integer. The pilot symbols \(Y_p \in \mathbb{C}^{N_N \times L_{f,f \times C}}\) are known to both the transmitter and receiver. In this paper, we assume that the \(N_p\) pilot symbols are identical without loss of generality. Next, we apply the inverse discrete Fourier transform (IDFT) to each OFDM symbol and append the cyclic-prefix (CP). After that, the transmit signal \(y \in \mathbb{C}^{(N_s+N_p)(L_{f,f}+L_{c,p})}\) propagates through the multipath channel described in (1). Once the receiver obtains the noisy channel output \(\hat{y}\), it removes the CP, and applies DFT to produce the frequency domain complex-valued symbols \(\hat{Y}_p\) and data symbols \(\hat{Y}\). Thus, the following equations hold for pilot and information symbols:

\[
\hat{Y}_p[i,k] = H[k] Y_p[i,k] + W[i,k],
\]

\[
\hat{Y}[j,k] = H[k] Y[j,k] + V[j,k],
\]

where \(H[k]\) denotes the channel frequency response for the \(k_{th}\) subcarrier, and both \(W\) and \(V\) denote noise samples.
follows: channel estimation method used in [29], which is as
and equalization are included as domain knowledge. In this
we propose the decoder structure shown in Figure 2 where
learn the underlying relationships in (2) and (3) to perform the
network. This approach relies on the deep neural network to
\( \hat{C} \).

C. Decoder design with domain knowledge

The decoder is trained to estimate transmitted source \( \hat{x} = D_\phi(Y_p, \hat{Y}, \hat{Y}_p) \) given \( Y_p, \hat{Y}_p \) and \( \hat{Y} \). The reconstruction loss can then be expressed as follows:

\[
L_{rec} = \mathbb{E}_{p(x,\hat{Y},\hat{Y}_p)}[||D_\phi(Y_p,\hat{Y},\hat{Y}_p) - x||^2_2],
\]

(4)

where \( \mathbb{E}_p[\ ] \) is the expected value over distribution \( p \) and \( D_\phi(Y_p, \hat{Y}, \hat{Y}_p)(= \hat{x}) \). Here, we use mean squared error (MSE) loss as our distortion function.

It must be noted that including OFDM signal processing in the proposed JSSC framework is possible because IDFT/DFT and CP removal/insertion can be treated as linear layers (i.e., matrix vector multiplications) with fixed parameters in the neural network. During back-propagation, gradients can pass through these (I)DFT/CP linear layers.

D. OFDM with PAPR reduction

One notable drawback of OFDM is the high PAPR that causes excessive power consumption at the power amplifier [39]. For the TX signal \( y[n] \), the PAPR is defined as:

\[
\text{PAPR}(y) = \frac{\max_n |y[n]|^2}{\mathbb{E}[|y[n]|^2]}.
\]

(9)

A number of solutions have been proposed to reduce the PAPR [30]. In this paper, we investigate the signal clipping technique [38] [39], which can be easily combined with our JSSC scheme.

Signal clipping is one of the simplest and extensively studied PAPR reduction methods. Suppose we have the original TX sample \( y[n] = A[n]e^{j\phi[n]} \), where \( A[n] \) is the amplitude and \( \phi[n] \) is the phase. The clipping operation can be expressed as follows:

\[
y_{\text{clip}}[n] = \begin{cases} 
A[n]e^{j\phi[n]} & \text{when } |A[n]| \leq \rho \sqrt{P_s} \\
\rho \sqrt{P_s}e^{j\phi[n]} & \text{otherwise},
\end{cases}
\]

(10)

where \( \rho \) denotes the clipping ratio (CR) and \( P_s \) is the input signal power. In the proposed JSSC, this clipping operation can be considered as introducing an additional non-linear activation function to the time domain OFDM signal. After clipping, we re-scale (i.e., normalize) the clipped signal so that the signal power is unchanged. According to [9], the normalization operation does not affect the PAPR. Gradient calculation during the training can be back-propagated through those normalization and non-linear clipping functions such that the encoder \( E_\phi \) and decoder \( D_\phi \) are trained for PAPR reduction via signal clipping.
E. Adaptive power allocation with CSI feedback

It is known that although CSI feedback does not increase the capacity of a memoryless channel \([40]\), it can significantly improve the error rate because of the increased error exponent using dynamic power allocation algorithms (e.g., water-filling [41]) and adaptive modulation [42]. For the system scenario where the CSI knowledge is available, we design an optional (i.e., only enabled when CSI is available) pre-coding block for the proposed JSCC to evaluate the impact of exploiting the CSI knowledge. The structure of our optional pre-coding block is shown in Figure 3. The pre-coding network takes the encoder output \(Y\) and (perfect) channel frequency response \(H \in \mathbb{C}^{N_c \times L_{f, f}}\) as inputs, and generates weights \(W \in \mathbb{R}^{N_c \times L_{f, f}}\). The weight \(W\) mimics the power distribution in traditional power allocation methods and thus it has the same size as \(Y\). The OFDM frequency-domain symbols are obtained by doing element-wise multiplication between \(W\) and \(Y\).

F. Combination with adversarial loss

Most prior works on JSCC wireless image transmission optimize the traditional distortion metric of PSNR. However, when the image is encoded with a low rate, PSNR tends to lose their importance as it favors pixel-wise error minimization over preservation of texture and content of the image. Thus, aggressive image compression maximizing PSNR could result in blurry images and the obtained PSNR often does not align well with human perception quality [43] [44]. Recently, minimizing adversarial loss [45] was proposed as a promising solution to capture the global semantic information and local textures, yielding visually-appealing generated images [46] [47] [48] [49]. In this section, we extend our model to an adversarial setting where an additional discriminator neural network is introduced to distinguish whether the image is original (uncompressed) or generated by the decoder (shown in Figure 3). We adopt the formulation of LSGAN [48] because of its simple formulation and improved stability than the original GAN [45]. The adversarial loss is defined in (11) where \(\text{Dis}(\cdot)\) is the discriminator network output for input image \(x\) (desired output is 1 for uncompressed images and 0 for images generated by the decoder).

\[
L_{\text{GAN}} = \max_{\text{Dis}} \mathbb{E}[(\text{Dis}(\hat{x}) - 1)^2] + \mathbb{E}[(\text{Dis}(x))^2].
\] (11)

When the adversarial loss is adopted, the training process becomes a minmax optimization and the objective function is modified from (8) to (12), where \(\lambda_g\) denotes the weight for the adversarial loss.

\[
\min_{E,D} L_{\text{total}} = L_{\text{rec}} + \lambda_c L_{\text{cha}} + \lambda_g L_{\text{GAN}}. \] (12)

III. TRAINING AND EVALUATION

We first test our proposed method with images in CIFAR-10, CIFAR-100, and CelebA datasets for wireless communication. CIFAR-10 and CIFAR-100 contain 60,000 32x32-pixel images and we adopt reflection padding and random cropping for data augmentation. CelebA contains more than 200,000 celebrity images. As the pre-processing, we scale and crop CelebA images to 64x64-pixels. For testing, we take the first 10,000 images from each test dataset (unused for training) and transmit each image 5 times through different random realizations of the multipath channel. In section III-A, III-B, III-C, and III-F, we only consider the image reconstruction loss and channel reconstruction loss, and we use both PSNR and SSIM to evaluate the image reconstruction quality. The effect of including an adversarial loss is evaluated in section III-E. Despite the small image size described here, our JSCC scheme can also be applied to high-resolution images. More details are introduced in section III-G.

The encoder and decoder networks for CIFAR-10 and CIFAR-100, which follow the design principles in [42], are shown in Figure 4. For the encoder, each image is first normalized within the range of \([-1, 1]\). Then, it is passed through a series of down-scaling convolutional layers and residual blocks. The output of the final layer consists of \(C\) channels and it is followed by a power normalization layer to enforce unit power. Note that we can change \(C\) to control the compression rate. The structure of the generator network is almost symmetric (in the reverse order) to the encoder network, except that the final activation function is set to Sigmoid function to enforce a valid dynamic range of image output pixels. Networks for CelebA have one more downsampling module at the encoder and one more up-sampling module at the generator network. The structure of the CE subnet and EQ subnet is kept the same for all datasets. Here, we use convolutional layers instead of fully connected layers because of their efficiency in terms of network size (number of parameters). Although the input signals for the CE subnet and EQ subnet are 1D in nature, we reshape them to be 2D so that we can use a small kernel to obtain the full receptive field covering the entire signal. For example, assume the 3 inputs (i.e., \(Y_p, Y_p,\) and \(H_{\text{MMSE}}\)) for the CE subnet have a size of \(2 \times 64\). We could reshape and stack them to form an input of \(6 \times 8 \times 8\). Then, three 2D convolutional layers with a kernel size of \(5 \times 5\) are sufficient to guarantee a full receptive field. The same logic applies to the EQ subnet.

We optimize the loss function in equation (8) and set \(\lambda_c\)
to 0.5 for all experiments in this subsection. We use a batch size of 128 for CIFAR-10 and CIFAR-100 and a batch size of 64 for CelebA. All neural networks are trained using ADAM with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and an initial learning rate of $5 \times 10^{-4}$. We train for 300 epochs and 30 epochs for CIFAR-10/CIFAR-100 and CelebA respectively. For all datasets, we apply linear learning rate decaying for the second half of the training process.

As for the OFDM system, the parameters are set to $L_{fft} = 64$, $L_{cp} = 16$, $L = 8$, and $\gamma = 4$. Note that the coding rate of our method depends on the number of samples after encoding. For a source image of the size $H \times W \times C$ pixels ($C$ is the number of color channels), the coding rate is $\frac{(N_p+N_c)(L_{fft}+L_{cp})}{HW_{PC}}$ in channel-use per pixel (CPP). The length of CP has to be larger than the delay spread of the channel to avoid inter-symbol interference.

Note that all evaluations in the following subsections except for Section III-D are performed without the CSI feedback (i.e., the pre-coding network in Figure 3 is unused), and we use instantaneous SNR defined at the receiver, which is:

$$\text{SNR} = 10 \log_{10} \frac{\| \hat{\mathbf{y}} \|^2}{\sigma^2} (\text{dB}). \quad (13)$$

### A. Effect of domain knowledge and residual connections

In this section, we quantify the effect of incorporating different amounts of domain knowledge into our framework and the effect of additional subnets. For each method, we evaluate the PSNR, SSIM and channel estimation mean square error (CE MSE) if available. In Table I and II, the first row represents a simple JSCC scheme where symbols are transmitted directly through the multipath channel without any explicit OFDM layers (as in Figure 1 bottom). In the second row, we introduce the OFDM blocks (DFT, CP, etc.) as our domain knowledge. However, for the decoder network, we directly concatenate $Y_p, \hat{Y}_p$ and $\hat{Y}$ and feed them to a generator network without any explicit channel estimation and equalization. Next, in the third row, we further introduce per-channel MMSE channel estimation and MMSE equalization as extra domain knowledge. However, we do not include any residual connections for refinement. Next, for the fourth row, we add CE subnet and the channel reconstruction loss to refine the channel estimation. Finally, in the fifth row, we evaluate the proposed system with the second residual connection and EQ subnet. All of the above methods are tested with CIFAR10, CIFAR-100 and CelebA under two different SNRs (5dB and 15dB). For the last four rows, where the OFDM system is included, we set $N_s = 7$ and $N_p = 1$, which provides a CPP rate of 0.208 for CIFAR-10/CIFAR-100 and a CPP rate of 0.052 for CelebA. For the direct transmission scheme, we keep the same CPP by properly sizing the encoder output dimension for a fair comparison.

From row 1 and row 2, we can see that when we introduce OFDM processing blocks to alleviate the multipath fading, the performance is greatly improved for both SNRs and across all datasets, especially for the 15dB SNR case where there is over 3dB gain in PSNR and 0.1 gain in SSIM for all tested datasets. When we further add explicit channel estimation and equalization, we observe a small improvement in the high SNR case. However, for the 5dB SNR case, both PSNR and SSIM drop instead. We think the main reason is that, when SNR is low, the error of per-channel MMSE channel estimation is so large that the benefits derived from domain knowledge is cancelled out. Whereas when the SNR is high, the channel estimation error is relatively small, and therefore the benefit from domain knowledge is evident. As we add CE subnet and the channel estimation loss to refine our channel estimation, we observe better channel estimation quality and we finally...
get better reconstruction quality for both 5dB and 15dB SNR. Lastly, as we further add EQ subnet and complete our final proposed scheme, we get the best performance among all the variants.

We provide the complexity analysis for the networks used for CIFAR-10 as shown in Table III. OFDM processing complexity compared to the generator network is insignificant and the all subnets are much smaller than the generator network. Hence all methods listed in Table III have comparable complexity. In fact, slight complexity reduction from 'Direct' method to OFDM-based methods (OFDM+CE+EQ) is achieved as OFDM pilot symbols do not require any neural network processing. Compared to 'OFDM+CE+EQ', the CE and EQ subnet in our proposed method (Figure 2) only takes $\approx 1.2\%$ additional parameters to achieve significant performance enhancement.

**B. Comparison with a separate coding scheme**

In this section, we compare our proposed JSSC method with a baseline separate source and channel coding scheme. The baseline separate coding scheme uses the state-of-the-art BPG as the image codec and LDPC in the IEEE 802.11n WiFi standard as the channel code. We test with three different LDPC codes, (972, 1944), (1296, 1944), (1458, 1944), which correspond to rates $1/2$, $2/3$ and $3/4$, respectively. We use BPSK, QPSK, 16QAM, and 64QAM for OFDM modulation applied to BPG- and LDPC-coded bit sequences. Then, we enumerate all possible modulation and coding rate combinations to identify the optimal configuration for the baseline separate coding scheme. We consider both perfect channel state information (CSI) and the per-channel MMSE channel estimation used in our proposed scheme. The estimated channel is then used to calculate the log likelihood ratio (LLR) for LDPC decoding in the baseline separate coding scheme. Note that the perfect CSI gives an (unfair) advantage to the baseline method and serves as an upper bound for all channel estimation methods.

We compare PSNR and SSIM of the proposed JSSC and the separate coding baseline at the same/similar rate and SNR. Because BPG cannot achieve arbitrary compression rates, we exhaustively search for the BPG parameter to obtain a rate closest to the target. For fair comparison, we remove the BPG headers in the BPG rate computation. When the decoded BPG bit sequence contains bit errors, either 1) it is decompressed to 2)

---

1. https://bellard.org/bpg/
2. https://github.com/tavildar/LDPC
an image with large quality degradation or 2) it fails to decode any valid image due to catastrophic failures in the decoder algorithm. When the latter happens, we assume (although it does not produce any valid image) that the corrupted BPG bit sequence corresponds to the average PSNR/SSIM of images recovered in the aforementioned case 1.

Figure 5 compares the performance of deep JSCC schemes and baseline approaches with respect to the number of information symbols $N_s$ in different SNR regimes using the CIFAR-10 dataset. A smaller $N_s$ indicates a more aggressive (lower) rate to send an entire image with a lower number of data symbols. For comparison, we show the performance of baseline separate coding schemes with 1) perfect CSI for error-free channel estimation, and 2) per-channel MMSE channel estimation with different numbers of pilots. We only show the performance of the best combination of source and channel coding, which is BPG+1/2LDPC+QPSK for 5dB and BPG+1/2LDPC+64QAM for 15dB.

From Figure 5a, 5b, we can see that the proposed JSCC scheme consistently outperforms the baseline separate coding scheme with perfect CSI, especially when $N_s$ is small. When the channel estimation accuracy drops due to fewer number of pilots ($N_p$) for per-channel MMSE estimation, our JSCC method outperforms the baseline method by a larger margin although it uses only $N_p = 1$ pilot symbol. Even when provided with perfect CSI, the separated coding schemes could not outperform the JSCC method. In practice, the baseline scheme may need a substantial number of pilot symbols to approach near-optimal channel estimation accuracy, which can significantly decrease the transmission rate. As shown in Figure 5c, 5d, when tested with 15dB SNR, the baseline separate coding scheme with perfect CSI approaches the performance of the proposed JSCC method in PSNR when $N_s$ is large. This observation aligns with prior works [15] [17] and shows that the JSCC method achieves the biggest advantages compared with separated coding schemes in the low SNR and low rate region. Further, it is observed that our JSCC method consistently outperforms the baselines although channel estimation is non-optimal with only $N_p = 1$, especially in the low SNR condition. Also, our JSCC method is consistently better than all baselines in SSIM that is better aligned with human perception than PSNR.
Fig. 8: PAPR vs PSNR for different clipping ratios for both 5dB (blue) and 15dB (red) SNRs on CIFAR-10 dataset with $N_s = 6, N_p = 1$ (CPP = 0.182). The corresponding clipping ratio is shown on the left of each marker. Solid line: network training is done with clipping. Dash line: clipping is applied after training while neural networks were not exposed to clipping during training.

Figure 6a and 6b show the trade-off between the pilot symbols and information symbols given a fixed total transmission symbol budget ($N_s + N_p = 7$, CPP = 0.182) for the constant packet length. On one hand, we prefer more pilot symbols so that we can achieve better channel estimation accuracy. On the other hand, it is beneficial to use more information symbols to carry the transmitted images with more redundancy in coding. Our results with both 5dB and 15dB SNRs show that as JSCC uses more pilot symbols, both the channel estimation MSE and PSNR improve. It implies that using only 1 pilot symbol for channel estimation ($N_p = 1$) and remaining symbols for data $N_s = 6$ is optimal for the JSCC method as shown in Figure 6a and 6b. Although less pilots tend to decrease channel estimation accuracy, the overall image reconstruction quality improves by the gain of using more data symbols to encode the transmitted image.

Next, in Figure 7, we fix the total transmission budget to $N_s + N_p = 7$ (CPP = 0.182) and evaluate the image reconstruction quality for a variety of SNRs. For the baseline method, we choose at each point the best combination of the OFDM modulation QAM size and LDPC rate. Since we need at least one pilot symbol for channel estimation, the curve for $N_p = 1$ and perfect CSI is essentially an upper bound of all the baselines considered in this paper. Our proposed method consistently outperforms the baseline method’s upper bound for both PSNR and SSIM across all SNRs. When we increase the number of pilot symbols (e.g., $N_p = 3$) for better CSI estimation in the baseline method, the performance gap increases between the baseline and the proposed JSCC. Furthermore, for the optimal case of $N_p = 1$, the proposed JSCC outperforms baseline methods more significantly when SNR is lower.

C. Performance with PAPR reduction

In this section, we evaluate our proposed method with additional signal clipping as our PAPR reduction technique. We first compute the average PAPR and PSNR obtained by different clipping ratios and the results are shown in Figure 8. We train our model with four clipping ratios $\rho = 0.8, 1, 1.2, 1.4$ for both SNR=5dB (blue markers) and SNR=15dB (red markers). For all points in that figure, we set $N_s = 6, N_p = 1$ (CPP = 0.182). To show the effect of end-to-end training with clipping, we compare it (solid line) with a simple scenario where the clipping operation is applied after training neural networks without exposing them to clipping during the training (dash line).

Figure 9 demonstrates the trade-off between image reconstruction quality and PAPR. When no PAPR reduction is applied ($\rho = \infty$), we obtain the highest PSNR but also the highest PAPR around 10dB for both noise levels, which puts a significant amplifier power efficiency burden for practical systems [38]. As we gradually put constraints on the PAPR (smaller $\rho$), we obtain lower PAPR as desired, but we gradually lose the representation power of our model and the image...
quality drops as well. With carefully designed clip ratios, we are able to achieve a low PAPR under 2dB for both 5dB and 15dB SNR. We can also observe that our model is more resilient to the clipping noise in 5dB than 15dB. When the SNR is 5dB, the channel noise dominates the total impairment; thus, the clipping noise has little effect on the image quality. However, when the SNR is 15dB, the clipping noise dominates the total distortion and hence has a large impact on the quality drop. When we directly apply the clipping operation without training with it, the additional non-linear noise has never been seen by the neural networks and thus leads to a larger performance degradation, which is shown by the dash lines in Figure 8. Similarly, the clipping noise has a larger impact for the high SNR case.

Figure 9 compares the effect of clipping on the performance of the proposed JSCC and baseline method with various clipping ratios (CR) of $\rho = 1, 1.4, \text{ and } \infty$ as defined in (10). Note that for the baseline, we adopt the methods in [38] [50] to calculate the LLR including the non-linear distortion introduced by clipping. It is observed that our JSCC approach provides a more gradual degradation in PSNR/SSIM than is observed for the baseline method as more aggressive clipping is applied. Similar to that reported in the previous section, our method outperforms the baseline approach for all SNRs tested in the presence of clipping.

D. Performance with CSI feedback

In this section, we evaluate the performance of the proposed system when the CSI feedback is available. We assume the perfect CSI is available at both transmitter and receiver. Hence, we remove the channel estimation module and CE subnet shown in Figure 2 and activate the optional pre-coding network shown in Figure 3. The structure of the pre-coding network follows the same structure as the subnet in Figure 4. We compare our JSCC method with/without CSI feedback to a baseline separated scheme that uses a water-filling power allocation algorithm [41], which assigns the optimal power for each subcarrier to attain the channel capacity when the CSI feedback is available. Instead of devising an optimal modulation and channel coding scheme (which is non-trivial) for the baseline, we make a simplifying assumption that it achieves the channel capacity given CSI. It corresponds to an upper bound performance for the baseline using BPG for image coding. The results are shown in Figure 10 where $N_x = 6$. Notice that the instantaneous (receiver) SNR defined in (13) is no longer compatible when the transmitter performs the water-filling power allocation based on the CSI feedback. Thus, when we evaluate the CSI feedback performance, we adopt the frequency domain average SNR defined as:

$$\text{Average SNR} = 10 \log_{10} \frac{1}{L_{ff}} \sum_{l=0}^{L_{ff}-1} \frac{\mathbb{E}|H[l]|^2 \sigma^2}{\sigma^2}$$

$$= 10 \log_{10} \frac{\mathbb{E}|H|^2 P}{\sigma^2} = 10 \log_{10} P \sum_{l=0}^{L} \frac{\sigma_{l-1}^2}{\sigma^2}$$

(14)

where $P$ is the signal power in frequency domain. As can be seen in Figure 10, exploiting the feedback of perfect CSI improves the performance for both PSNR and SSIM. Compared with the upper bound of the baseline method that fully attains the channel capacity, our proposed JSCC method provides similar PSNR and consistently better SSIM. The gain from the CSI feedback is larger for the proposed JSCC than that of the baseline separate scheme.

E. Performance with adversarial loss

In this section, we train our JSCC method with the adversarial loss (12) and we call this method JSCC-adv. To better evaluate the perceptual quality or recovered images, we adopt an additional metric called classification accuracy (CA) that evaluate the classification performance of the generated images based on a network pre-trained on original images. For CIFAR-10, we pre-train a ResNet50 network [51] as the classifier and it achieves a $93.62\%$ classification accuracy on the original CIFAR-10 test set. Some numerical results on CIFAR-10 is shown in Table IV. For all compared methods, we set $N_x = 6$ and $N_p = 1$ (CPP = 0.182). The baseline method uses 1/2LDPC+QPSK for 5dB SNR and 1/2LDPC+64QAM for 15dB SNR, and the perfect CSI is assumed to be known.
TABLE IV: Evaluation of Adversarial Loss on CIFAR-10

| Metric | 5dB SNR | CA | 15dB SNR | CA |
|--------|---------|----|----------|----|
| Original | ∞ | 1 | 93.62% | ∞ | 1 | 93.62% |
| BPG+LDPC | 23.433 | 0.75 | 32.3% | 29.047 | 0.919 | 81.9% |
| JSCC | 25.068 | 0.845 | 64.42% | 30.126 | 0.948 | 87.21% |
| JSCC-adv | 24.129 | 0.826 | 74.22% | 29.338 | 0.937 | 87.74% |

For the JSCC-adv method that combines the adversarial loss as in (11), we set \( \lambda_g = 2 \times 10^{-5} \) for 5dB SNR and \( \lambda_g = 5 \times 10^{-5} \) for 15dB SNR. For both SNRs, the original JSCC method achieves the highest PSNR and SSIM whereas the JSCC-adv method achieves the highest CA. That means the images recovered by JSCC-adv tends to better align with the distribution of real images although the additional adversarial loss term does slightly degrades the MSE objective thus sacrifices PSNR/SSIM accordingly. This perception-distortion trade-off aligns with prior works such as [52]. Although JSCC-adv appears lower in traditional distortion measures PSNR and SSIM (due to the adversarial loss term in addition to MSE), it still outperforms the baseline method. It can also be noticed that the perceptual improvement is more obvious for low SNR scenarios.

By Figure 11 shows visualization examples of recovered images for both CIFAR-10 and CelebA. It can be observed that the JSCC method provides much better recovered images than the baseline method. However, the recovered images are blurry due to optimization of the MSE loss. On the contrary, combined with adversarial loss, JSCC-adv produces images with more high-frequency details such as human hair and background. This improvement of human perception aligns with the improvement of CA in Table IV.

Fig. 11: Visualizations of CIFAR-10 (top, CPP = 0.182) and CelebA (bottom, CPP = 0.111).

For the JSCC-adv method, the proposed network is trained with 6 multipath channels and a CPP = 0.182. As shown in Figure 12, the JSCC model trained with 5dB SNR performs better at low SNR values, whereas the model trained with 15dB SNR performs better at high SNR values. However, both of them still outperform the baseline (with perfect CSI) for most cases. The same trend can be observed from the channel estimation error on the right. Notice that even when the model is tested with unseen noise levels, its channel estimation consistently benefits from the additional residual connection and outperforms the basic per-channel MMSE method.

We also test the JSCC model with mismatched multipath number \( L \). The JSCC method in Figure 13 is trained with an \( L = 8 \) multipath channel model and tested with a wide range of \( L \) but the same noise level. The system performance is hardly affected with less than 8 paths for both SSIM and channel estimation error. Then the channel has more than 8 paths, the channel becomes more difficult and degradation on both channel estimation and SSIM gradually appears. However,
within a certain range of $L$, the JSCC method is robust to channel mismatch and still outperforms the baseline separated methods with perfect CSI.

G. Extension to larger images and qualitative results

In this section, we show that our JSCC scheme can be generalized to high-resolution images. We train our model with 200,000 natural images from Open Image Dataset V6.\footnote{https://storage.googleapis.com/openimages/web/index.html} Similar with prior JSCC works\cite{15}\cite{17}, we train on $256 \times 256$ random patches with a batch size of 16. Compared with the network structure in Figure 4, this model contains one more down-scale convolutional layer and two more residual blocks for the encoder and one more upscale convolutional layer and two more residual blocks for the generator network. Same with the networks for small images, we train our model using ADAM with $\beta_1 = 0.5$, $\beta_2 = 0.999$ and an initial learning rate of $5 \times 10^{-4}$. We train our model for 40 epochs and apply linear learning rate decaying for the last 20 epochs.

Thanks to the fully convolutional structure applied to a patch of large images, we can test our model on multiple input sizes not limited to $256 \times 256$. We evaluate our model with the Kodak dataset, which consists of 24,768 $\times$ 512 high-resolution images and serves as a popular benchmark for image compression. In Figure 14, we show the comparison between our JSCC and conventional schemes in PSNR and MS-SSIM\cite{53}. For our JSCC scheme, we set $N_s = 576$ and $N_p = 2$, which corresponds to a CPP rate of 0.039. We transmit each tested image 10 times through random channel realizations. The baseline method uses 1/2 rate LDPC and QPSK for the same CPP of 0.039.

As can be seen from Figure 14, our method provides similar PSNR in the low SNR region but worse PSNR in the high SNR region. However, when evaluated with MS-SSIM, our method greatly outperforms the baseline scheme. It is worth noting that PSNR is a less reliable (subjective) quality measure for high resolution images in Kodak dataset. For visualization, we present some examples from the 5dB SNR case in Figure 15. As shown in Figure 14, the images generated by the JSCC method have noticeably better visual quality as indicated by the (MS-)SSIM metric (even when the measured PSNR is lower). When we adopt the adversarial loss (i.e., JSCC-adv), the visual quality is further improved and the reconstructed images successfully recover fine details such as the bush, grass, and ground texture although the obtained PSNR, SSIM

![Fig. 13: Robustness test for multipath channel model parameter ($L$) mismatch scenarios. The JSCC is trained with $N_s = 6$, $N_p = 1$ (CPP = 0.182), and 5dB SNR. (a): SSIM, (b): channel estimation MSE.](image13.png)

![Fig. 14: Performance of the proposed method on Kodak dataset. (a): PSNR, (b): MS-SSIM.](image14.png)
and MS-SSIM metrics are lower. Our experiment shows that JSCC can be successfully trained with the adversarial loss for large images and the adversarial loss aligns better with human perception.

IV. Conclusion

In summary, to the best of our knowledge, we present the first implementation of an OFDM-assisted deep JSCC method for multipath fading channels. We combine the deep JSCC scheme with explicit OFDM layers and carefully design the decoder using both expert domain knowledge (such as channel estimation and equalization) and refinement neural networks. Our results demonstrate that such usage of domain knowledge such as OFDM, explicit channel estimation and equalization greatly boost the performance of our model compared to unstructured black box networks. Through extensive experimental simulations, we show that our deep JSCC method outperforms the combination of state-of-the-art image compression codecs, conventional high performing channel codes and OFDM systems, even with perfect CSI. The margin is larger for low SNR and low rate regimes. When combined with adversarial loss, our method tends to generate more realistic images. Moreover, the JSCC framework incorporates deliberate signal clipping during the training process to significantly reduce the PAPR with gradual performance degradation. Furthermore, our approach is shown to be robust when the evaluation channel model characteristics are mismatched to those used during the training.

One possible direction of future works is to generalize this work to other types of sources such as audio and video. Different types of sources may require specific model architectures and training strategies to achieve the maximum performance. Another future direction is to add more details to the OFDM system such as carrier frequency offset (CFO) and packet detection so that the system is more practical. One could also design novel pilot patterns to achieve the best efficiency for both channel estimation and source reconstruction.

APPENDIX A

ABLATION STUDIES ON $\lambda_c$ AND $\lambda_g$

An ablation study of $\lambda_c$ and $\lambda_g$ was conducted on CIFAR-10 to better illustrate their effect. For the ablation study of $\lambda_c$, we set $\lambda_g = 0, N_t = 6, N_p = 1$ and SNR = 5dB. The results of PSNR and channel estimation MSE are shown in Table V. Obtained PSNR is not sensitive to $\lambda_c$ although setting $\lambda_c = 0$ results in significantly larger channel estimation error. Using channel estimation subnet $\Phi_{ce}$ enhances the quality of the image as shown in Table V. Although the image quality does not significantly depend on $\lambda_c$ as long as $\Phi_{ce}d$ is included in the structure, applying $\lambda_c > 0$ produces more interpretable trained results from the channel estimation subnet.

For the ablation study of $\lambda_g$, we set $\lambda_c = 0.5, N_t = 6, N_p = 1$ and SNR = 5dB. As we gradually increase $\lambda_g$, the reconstruction PSNR drops while the classification accuracy increases. When $\lambda_g$ gets too large, however, the neural network tends to create (fake) contents that are differ from original images, degrading both PSNR and classification accuracy. To visualize it, we show examples from CIFAR-10 in Figure 16.

APPENDIX B

COMPARISON WITH OTHER JSCC STRUCTURES

We compare our basic neural network structure without OFDM and subnets (i.e., ‘Direct w/o OFDM’ in Table I) with those proposed in [15] and [16]. We conduct experiments on CIFAR-10 using AWGN and multipath fading channels. As we exclude OFDM from our scheme, the difference only comes from the neural network structure itself. For each 32 × 32 image, we assign 384 channel usings, yielding a CPP of 0.375. The evaluation results are shown in Figure 17. Our structure (even without OFDM and proposed subnets) constantly outperforms the other structures in both AWGN channel and multipath fading channels whereas our structure uses 23% fewer parameters than the structure in [16]. Our performance further improves in the multipath fading channel as we incorporate OFDM blocks as well as proposed subnet structures (‘OFDM+CE+EQ+\Phi_{ce}+\Phi_{ce}’). As quantified in Table IV and III.

REFERENCES

[1] C. E. Shannon, “A mathematical theory of communication,” The Bell system technical journal, vol. 27, no. 3, pp. 379–423, 1948.
[2] A. Vosoughi, P. C. Cosman, and L. B. Milstein, “Joint source-channel coding and unequal error protection for video plus data,” IEEE Signal Processing Letters, vol. 22, no. 1, pp. 31–34, 2014.
[3] S. Heinen and P. Vary, “Transactions papers source-optimized channel coding for digital transmission channels,” IEEE transactions on communications, vol. 33, no. 4, pp. 592–600, 2005.
[4] D. Goodman and T. Moulsey, “Using simulated annealing to design digital transmission codes for analogue sources,” Electronics letters, vol. 24, no. 10, pp. 617–619, 1988.
[5] V. Bozantizis and F. Ali, “Combined vector quantisation and index assignment with embedded redundancy for noisny channels,” Electronics Letters, vol. 36, no. 20, pp. 1711–1713, 2000.
[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097–1105.
[7] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.
[8] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using rnn encoder-decoder for statistical machine translation,” arXiv preprint arXiv:1406.1078, 2014.
[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in neural information processing systems, 2017, pp. 5998–6008.
[10] H. Ye, G. Y. Li, and B.-H. Juang, “Power of deep learning for channel estimation and signal detection in ofdm systems,” IEEE Wireless Communications Letters, vol. 7, no. 1, pp. 114–117, 2017.
| Method          | PSNR (dB) | SSIM | MS-SSIM |
|-----------------|-----------|------|---------|
| BPG+LDPC        | 21.515    | 0.615| 0.874   |
| JSCC            | 21.659    | 0.659| 0.902   |
| JSCC-adv        | 20.072    | 0.589| 0.864   |

Fig. 15: Visual comparison between images achieved from the baseline separated coding schemes and the proposed JSCC method on Kodak, with CPP = 0.039. The performance is associated with each recovered image in a format of PSNR (dB)/SSIM/MS-SSIM.

| Method          | PSNR (dB) | SSIM | MS-SSIM |
|-----------------|-----------|------|---------|
| BPG+LDPC        | 21.092    | 0.42 | 0.772   |
| JSCC            | 21.175    | 0.505| 0.844   |
| JSCC-adv        | 20.208    | 0.464| 0.821   |

| Method          | PSNR (dB) | SSIM | MS-SSIM |
|-----------------|-----------|------|---------|
| BPG+LDPC        | 27.334    | 0.67 | 0.867   |
| JSCC            | 27.057    | 0.709| 0.895   |
| JSCC-adv        | 25.75     | 0.655| 0.86    |

| Method          | PSNR (dB) | SSIM | MS-SSIM |
|-----------------|-----------|------|---------|
| BPG+LDPC        | 24.089    | 0.632| 0.857   |
| JSCC            | 23.68     | 0.687| 0.9     |
| JSCC-adv        | 22.607    | 0.631| 0.874   |

| Method          | PSNR (dB) | SSIM | MS-SSIM |
|-----------------|-----------|------|---------|
| BPG+LDPC        | 21.534    | 0.522| 0.817   |
| JSCC            | 22.512    | 0.636| 0.882   |
| JSCC-adv        | 21.059    | 0.56 | 0.849   |

Fig. 16: Visualizations of CIFAR-10 with different $\lambda_g$.
Fig. 17: Comparison of different neural network structures in (a) AWGN and (b) Multipath channel.
[52] Y. Blau and T. Michaeli, “The perception-distortion tradeoff,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6228–6237.

[53] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, vol. 2. IEEE, 2003, pp. 1398–1402.