RDJSCC: Robust Deep Joint Source-Channel Coding Enabled Distributed Image Transmision over Severe Fading Channel

Biao Dong†, Wenkai Tian†, Bin Cao†‡, and Yu Wang*

† School of Electronics and Information Engineering, Harbin Institute of Technology, Shenzhen, China
‡ Research Center of Networks and Communications, Peng Cheng Laboratory, Shenzhen, China
* School of Telecomm. & Inform. Eng., Nanjing Univ. of Posts & Telecomm., Nanjing, China
Email: 23b952012@stu.hit.edu.cn, 22s152098@stu.hit.edu.cn, caobin@hit.edu.cn, yuwang@njupt.edu.cn

***Abstract*—In this paper, we investigate the effects of severe channel fading in the scenario of distributed deep learning-based joint source-channel coding (DJSCC) for image transmission without perfect channel state information (CSI). To tackle the challenges posed by imperfect CSI, we propose a robust DJSCC (RDJSCC) scheme that operates at three levels: modulation, encoding, and decoding, respectively. Firstly, at the modulation level, we adopt orthogonal frequency division multiplexing (OFDM) modulation for exploring the trade-off between reconstruction performance and peak-to-average power ratio (PAPR). Secondly, at the encoding level, two parameter-efficient operators are introduced to combat channel fading with low encoding complexity. Finally, at the decoding level, we divide the decoding process into two stages, i.e., denoising and recovery, aiming to maximize the correlation between the encoded representations. Theoretic analysis and simulation results show that our proposed RDJSCC can effectively alleviate the effects of severe fading with imperfect CSI, leading to an improved reconstruction performance while maintaining low PAPR and encoding complexity.**

I. Introduction

6G mobile networks are envisioned to enable intelligent transmission with extremely low end-to-end latency, even in severe fading environments (e.g., an autonomous driving scenario in Fig. 1). Deep learning (DL)-based joint source-channel coding (DJSCC) can be a potential technology for 6G [1]–[4], since DJSCC can dynamically allocate bandwidth to source or channel coding and thus exhibit a graceful performance degradation in fading environments. Further, [2]–[4] extended the DJSCC to dynamic channel environments and allocated bandwidth based on image content, achieving a better end-to-end transmission performance.

The majority of existing works on DJSCC focus on point-to-point communication and seldom consider the distributed communication among multiple nodes. For example, in a distributed sensor network for autonomous driving (described in Fig. 1), it is challenging to encode correlated sources distributed across different locations without global communication [5]. Although the theoretical framework of distributed source coding (DSC) provides a promising approach to handling data compression problems [6], practical DSC systems have not been widely used due to challenges in designing efficient DSC ystems to capture complex correlations between sources. [7] and [8] attempted to apply DJSCC to distributed scenarios of additive white Gaussian noise (AWGN) channel and Rayleigh channel with perfect channel state information (CSI), while neglecting the impacts of severe fading channel with imperfect CSI on the correlation of sources.

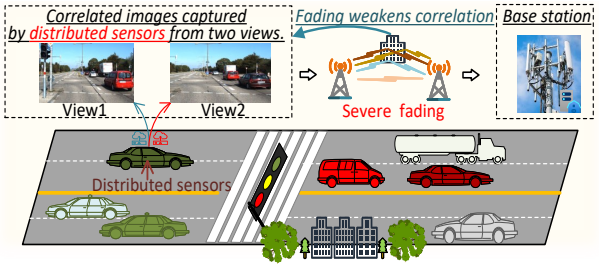


Fig. 1. An autonomous driving example for illustrating a distributed sensor network, where fast vehicle motion leads to severe channel fading and thus degrades the distributed transmission performance.

In this paper, we deploy DJSCC in a distributed sensor network for image transmission under a severe fading channel without perfect CSI. Our goal is to mitigate the effects of severe fading channel on the encoded representations and thus improve the system's robustness. To the best of our knowledge, distributed DJSCC in such a context has not been studied before. To fill this research gap, we propose **a robust DJSCC (RDJSCC) scheme**, which is specially designed for distributed sensor networks under severe fading channels. The proposed RDJSCC scheme **mitigates the effects of severe fading** at three levels: modulation, encoding, and decoding while exploring trade-offs among peak-to-average power ratio (PAPR), encoding complexity, and reconstruction performance, respectively.

II. Related Works

DJSCC: DJSCC is essentially an autoencoder over a noisy channel applying DL to JSCC as an encoder-decoder pair. [1] introduced the first DJSCC method, which proved the superiority of DJSCC in low signal-to-noise ratio (SNR) environments. To further improve the rate-distortion (RD) performance, [4] introduced a hyperprior as side information and integrated Swin Transformer as the backbone. Leveraging the advantages of digital modulation, [9], [10] incorporated orthogonal frequency division multiplexing (OFDM) into DJSCC. To adapt a single DJSCC across different channel conditions and compression ratios, adaptive strategies were also widely researched [2]–[4]. In addition, task-oriented DJSCC also attracted great attention [11], focusing on jointly optimizing DJSCC with downstream tasks in a task-oriented manner to achieve edge inference with low latency.

DL-based DSC: DSC is also known as the Wyner–Ziv scenario [6], which addresses the lossy source coding with correlated sources (which are also referred to as side information, while the lossless scenario is known as Slepian–Wolf coding). Inspired by the success of learning-based image compression, an early framework for learning-based distributed image compression was proposed in [12], which utilized mutual information between distributed images for efficient compression. [13] further improved this compression framework, by extracting common information rather than feeding correlated sources to the decoder directly. There have been a growing interests in utilizing the correlated sources efficiently, interested readers can refer to [5] and references therein.

III. Problem Formulation

A. Scenario

We consider the following distributed wireless image uplink transmission scenario. As shown in Fig. 2, sensor $\mathbf{s} \in \mathbb{R}^M$ and its correlated version (side information) $\mathbf{s}_{side} \in \mathbb{R}^M$ with a joint distribution $p(\mathbf{s}, \mathbf{s}_{side})$ capture two views of a same object. Two sensors independently transmit their compressed representations $(\mathbf{x}, \mathbf{x}_{side})$ to a central decoder over a severe fading channel. We define $(\hat{\mathbf{x}}, \hat{\mathbf{x}}_{side})$ as the corrupted version of $(\mathbf{x}, \mathbf{x}_{side})$ by channel. We use $\hat{\mathbf{x}}^c$ to denote the common features between $\hat{\mathbf{x}}$ and $\hat{\mathbf{x}}_{side}$, which satisfy the Markov chain $\mathbf{s} \rightarrow \hat{\mathbf{x}}^c \rightarrow \mathbf{s}_{side}$. As a remark, unlike the setting in [7], [13], where the decoder can only losslessly access one source, we assume the decoder can access both correlated sources undergoing channel fading, which reflects real-world scenarios.

According to the Wyner-Ziv theorem [6], the expected length for lossy compression of ($\mathbf{s}$, $\mathbf{s}_{side}$), i.e., joint entropy, is $R_{\mathbf{s}} + R_{\mathbf{s}_{side}} = H(\mathbf{s}, \mathbf{s}_{side})$ when separate encoding and joint decoding are employed. Based on the Wyner-Ziv setup, the rate-distortion function with a distortion metric $d(\cdot, \cdot)$ is given by $R_{\mathbf{s}|\mathbf{s}_{side}}(d) = \inf I(\mathbf{s}; \hat{\mathbf{x}} \mid \hat{\mathbf{x}}_{side})$, where the infimum is taken over all random variables. Hence, we define a reconstruction mapping under a given distortion C as $g : \hat{\mathcal{X}}_{side} \times \hat{\mathcal{X}} \rightarrow \hat{\mathcal{S}}$ with constraint $\mathbb{E}\left[d(\mathbf{s}, f(\hat{\mathbf{x}}, \hat{\mathbf{x}}_{side}))\right] \leq C$.

B. Fading Channel

Unlike AWGN channel adopted in [7], [8], the received fading signal in this paper is modeled as follows:

$$\hat{\mathbf{x}} = h * \mathbf{x} + w, \tag{1}$$

where the impulse response of the multipath channel consists of L paths, i.e., $h \in \mathbb{C}^L$. $w \sim \mathcal{CN}(0, \sigma^2 I_k)$ is the AWGN. Each h_l follows the complex Gaussian distribution with zero mean and variance σ_l^2, i.e., $h_l \sim \mathcal{CN}(0, \sigma_l^2)$ for $l = 0, 1, \ldots, L-1$. The variance σ_l^2 of each path follows the exponential decay, i.e., $\sigma_l^2 = \alpha_l e^{-\frac{l}{\gamma}}$, where γ represents the delay and α_l is a normalization coefficient. The sum of variances of all paths equals to 1, i.e., $\sum_{l=0}^{L-1} \sigma_l^2 = 1$.

C. The Optimization Goal

In this paper, we use DL as an encoder-decoder pair [1]. Specifically, the distribution of latent representation $q_{\hat{\mathbf{x}}|\mathbf{s}}(\hat{\mathbf{x}}|\mathbf{s})$ is learnt by a transform f (parameterized by $\boldsymbol{\phi}$) at the encoder as $\mathbf{x} = f(\mathbf{s}; \boldsymbol{\phi})$, $\mathbf{x} \in \mathbb{C}^{M'}$ and compression ratio is defined as $R \triangleq M'/M$. Similarly, the distribution of latent representation of correlated sources is parameterized by ψ as $\hat{\mathbf{x}}_{side} = f(\mathbf{s}_{side}; \boldsymbol{\psi})$. Finally, the reconstructed image can be derived from a DL-based decoder as $\hat{\mathbf{s}} = g(\hat{\mathbf{x}}, \hat{\mathbf{x}}_{side}; \boldsymbol{\theta})$, where θ denotes the parameters of the decoder. In this setting, we aim to approximate the joint distribution of the random variables as $p(\mathbf{s}, \mathbf{s}_{side}, \hat{\mathbf{x}}, \hat{\mathbf{x}}_{side})$, which is intractable. To acquire a tractable solution, a factored variational approximation of the posterior distribution is introduced as $q_{\boldsymbol{\phi}}(\hat{\mathbf{x}}, \hat{\mathbf{x}}_{side} \mid \mathbf{s}, \mathbf{s}_{side})$.

$$\begin{aligned}
&\mathbb{E}_{\mathbf{s}, \mathbf{s}_{\text{side}} \sim p(\mathbf{s}, \mathbf{s}_{\text{side}})} D_{\mathrm{KL}}[q_{\boldsymbol{\phi}}(\hat{\mathbf{x}}, \hat{\mathbf{x}}_{\text{side}} \mid \mathbf{s}, \mathbf{s}_{\text{side}}) \| p(\hat{\mathbf{x}}, \hat{\mathbf{x}}_{\text{side}} \mid \mathbf{s}, \mathbf{s}_{\text{side}})] \\
&= \mathbb{E}_{\mathbf{s}, \mathbf{s}_{\text{side}} \sim p(\mathbf{s}, \mathbf{s}_{\text{side}})} \mathbb{E}_{\hat{\mathbf{x}}, \hat{\mathbf{x}}_{\text{side}} \sim q_{\boldsymbol{\phi}}} \Big(\log q_{\boldsymbol{\phi}}(\hat{\mathbf{x}} \mid \mathbf{s}) q_{\boldsymbol{\phi}}(\hat{\mathbf{x}}_{\text{side}} \mid \mathbf{s}_{\text{side}}) \\
&- \big(\underbrace{\log p_{\boldsymbol{\theta}}(\mathbf{s} \mid \hat{\mathbf{x}}_{\text{side}}, \hat{\mathbf{x}})}_{D_{\mathbf{s}}} + \underbrace{\log p\left(\hat{\mathbf{x}}\right)}_{R_{\hat{\mathbf{x}}}} + \underbrace{\log p\left(\hat{\mathbf{x}}_{\text{side}}\right)}_{R_{\hat{\mathbf{x}}_{\text{side}}}} \big) \Big) + \text{const.}
\end{aligned} \tag{2}$$

$$(\boldsymbol{\phi}^*, \boldsymbol{\theta}^*) = \arg\min_{\boldsymbol{\phi}, \boldsymbol{\theta}} \mathbb{E}_{\mathbf{s}} \left[d\left(\mathbf{s}, \hat{\mathbf{s}}\right) \right]. \tag{3}$$

We minimize the Kullback-Leibler (KL) divergence between the approximate density $q_{\boldsymbol{\phi}}(\hat{\mathbf{x}}, \hat{\mathbf{x}}_{side} \mid \mathbf{s}, \mathbf{s}_{side})$ and the true posterior $p(\hat{\mathbf{x}}, \hat{\mathbf{x}}_{side} \mid \mathbf{s}, \mathbf{s}_{side})$ as (2). The first term in the KL divergence can be technically dropped [4]. The term $D_{\mathbf{s}}$ represents the reconstruction distortion, and we adopt peak signal-to-noise ratio (PSNR) to qualify image transmission distortion based on the mean-squared error (MSE). $R_{\hat{\mathbf{x}}}$ and $R_{\hat{\mathbf{x}}_{side}}$ denote the compression ratios of $\hat{\mathbf{x}}$ and $\hat{\mathbf{x}}$, respectively. Our goal is to address the effects of severe fading channel on distributed image transmission performance based on a fixed compression ratio, i.e., $R_{\hat{\mathbf{x}}}$ and $R_{\hat{\mathbf{x}}_{side}}$ are constant. In this regard, we use the commonly employed MSE as the loss function, as defined in (3).

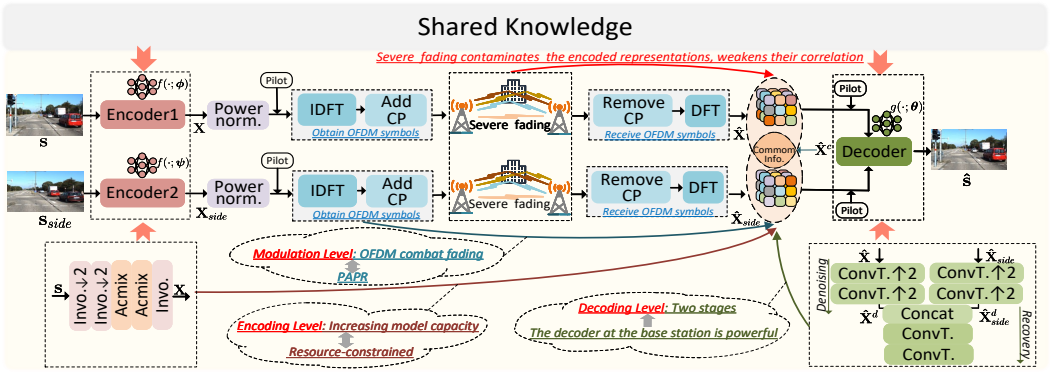


Fig. 2. Our proposed RDJSCC scheme. The encoder and decoder of RDJSCC are stacked by different operator blocks. Different operator blocks adopt a residual connections structure, where the operator can be ACmix, involution, standard convolution or transpose convolution, and ↓/↑ denotes downsampling/upsampling.

IV. Methods

To mitigate the effects of a severe fading channel on the encoded representations in a distributed wireless image transmission system, the proposed RDJSCC employs three key techniques to enhance the robustness of image transmission: modulation level, encoding level, and decoding level, respectively.

A. Modulation Level

We apply OFDM to DJSCC, aiming to mitigate the effects of severe fading, which is inspired by [9]. As shown in Fig. 2, each encoded representation $\mathbf{x}$ is power normalized and then allocated with an OFDM packet. Each packet contains N_s information symbols and N_p pilot symbols. The pilot symbols $\mathbf{x}_p \in \mathbb{C}^{N_p \times N_c}$ are known to both the transmitter and receiver. Under the OFDM modulation setting, $\mathbf{x}$ represents the frequency domain symbol. First, $\mathbf{x}$ selects N_c subcarriers, i.e., $\mathbf{x}$ is reshaped as $\mathbf{x} \in \mathbb{C}^{N_s \times N_c}$. Then, symbols on all subcarriers are synthesized into a single time-domain symbol $\mathscr{X} \in \mathbb{C}^{N_s \times N_c}$ through inverse discrete Fourier transform (IDFT) denoted as, $\mathscr{X} = \boldsymbol{F}_{N_c}^H \mathbf{x}$, where $\boldsymbol{F}_{N_c}$ is an N_c-dimensional DFT matrix and $\boldsymbol{F}_{N_c}^H$ is an IDFT matrix. Next, a cyclic prefix (CP) of length L_{cp} is added on the time-domain symbol $\mathscr{X}$, yielding $\mathscr{X}_{\text{cp}} \in \mathbb{C}^{N_s \times (N_c + L_{\text{cp}})}$. The pilot symbol $\mathbf{x}_p$ is also performed IDFT and added CP addition, then concatenated with $\mathscr{X}_{\text{cp}}$ to form an OFDM symbol $\mathscr{X}_{\text{ofdm}} \in \mathbb{C}^{(N_s+N_p) \times (N_c + L_{\text{cp}})}$.

After adding pilots, $\mathscr{X}_{\text{ofdm}}$ is transmitted through the fading channel as (1). When the base station receives $\hat{\mathscr{X}}_{\text{ofdm}}$ (the corrupted version of $\mathscr{X}_{\text{ofdm}}$), the OFDM demodulation is performed. This involves removing the CP, applying DFT to obtain the corrupted frequency-domain symbols $\hat{\mathbf{x}}$ and the corrupted pilot symbols $\hat{\mathbf{x}}_p$. The correlated source $\mathbf{x}_{side}$ experiences the same process as $\mathbf{x}$. The procedure is also illustrated in Algorithm 1.

It should be pointed out that the proposed method utilizes discrete-time analog transmission (DTAT), where baseband complex symbols are transmitted directly after OFDM modulation, rather than using passband transmission of digital signals [10]. This approach combines the advantages of OFDM (i.e., overcoming frequency-selective fading) with DTAT. However, there is a trade-off between PAPR and performance in OFDM-based DJSCC system. [10] demonstrated that the high PAPR of DJSCC could be mitigated by incorporating clipping into the training process as,

$$\mathscr{X}_{\text{ofdm}}^{clip} = \begin{cases} \mathscr{X}_{\text{ofdm}}, & \text{if } |\mathscr{X}_{\text{ofdm}}| \le \rho \bar{\mathscr{X}}_{\text{ofdm}} \\ \rho \bar{\mathscr{X}}_{\text{ofdm}}, & \text{if } |\mathscr{X}_{\text{ofdm}}| > \rho \bar{\mathscr{X}}_{\text{ofdm}}, \end{cases} \tag{4}$$

where $\mathscr{X}_{\text{ofdm}}^{clip}$ represents the clipped signal, ρ represents the clipping ratio and $\bar{\mathscr{X}}_{\text{ofdm}}$ represents the average amplitude

Algorithm 1: The proposed RDJSCC method.

Input: Training data $(\mathbf{s}, \mathbf{s}_{side})$; the pilot symbols $\mathbf{x}_p \in \mathbb{C}^{N_p \times N_c}$; learning rate η_0; the training epochs of global model T; the batch size of a training epoch B;

Output: Trained parameters;

for $(\mathbf{s}, \mathbf{s}_{side}) \in \mathcal{D}_{\text{train}}$ **do**
- $\boldsymbol{\Gamma} \leftarrow$ Randomly generate uniform SNR;
- $\mathbf{x} \in \mathbb{C}^{N_s \times N_c} \leftarrow f(\mathbf{s}, \boldsymbol{\Gamma}; \boldsymbol{\phi})$; ▷Encoder
- ▷Modulation
- $\mathscr{X} \in \mathbb{C}^{N_s \times N_c} \leftarrow \boldsymbol{F}_{N_c}(\mathbf{x})$;
- $\mathscr{X}_{\text{cp}} \in \mathbb{C}^{(N_s+L_{cp}) \times N_c} \leftarrow \text{CP}(\mathscr{X}, CP)$;
- $\mathscr{X}_{\text{ofdm}} \in \mathbb{C}^{(N_s+L_{cp}) \times (N_c+N_p)} \leftarrow \text{Pilot}(\mathscr{X}_{\text{cp}}, \mathbf{x}_p)$;
- $\mathscr{X}_{\text{ofdm}}^{clip} \leftarrow \text{Clip}(\mathscr{X}_{\text{ofdm}}, \mathbf{x}_p)$;
- ▷Through the fading channel
- $\hat{\mathscr{X}}_{\text{ofdm}}^{clip} \leftarrow h * \mathscr{X}_{\text{ofdm}}^{clip} + w$;
- $\hat{\mathbf{x}} \leftarrow \hat{\mathscr{X}}_{\text{ofdm}}^{clip}$; ▷Demodulation
- ▷$\mathbf{x}_{side}$ is processed by the same procedure with $\mathbf{x}$
- $\hat{\mathbf{s}} = g(\hat{\mathbf{x}}, \hat{\mathbf{x}}_{side}, \boldsymbol{\Gamma}; \boldsymbol{\theta})$; ▷Decoder

end

Calculate the loss $d(\mathbf{s}, \hat{\mathbf{s}})$;

Update model parameters $(\boldsymbol{\phi}, \boldsymbol{\theta})$.

of $\mathscr{X}_{\text{ofdm}}$. The clipping ratio ρ should be selected properly to avoid destroying the orthogonality among subcarriers. We incorporate clipping into the training process of RDJSCC to strike the balance between PAPR and performance.

B. Encoding Level

Next, we aim to combat severe fading at the encoding level. Generally, increasing the model capacity of the encoder improves encoding performance [4]. This is because a high-capacity encoder can capture high-level semantic features for source compression and noise resistance. However, in a distributed wireless sensors network, sensors are often resource-constrained, which poses a challenge in employing a model with high capacity. Hence, there exists a trade-off between model capacity and performance in the proposed RDJSCC. Here, we investigate the above trade-off at the encoding level. Specifically, two parameter-efficient operators, namely involution and ACmix, are introduced at the encoder, as illustrated in Fig. 2.

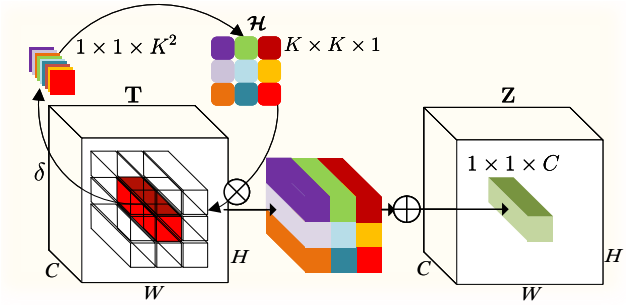


Fig. 3. An involution example when $G = 1$. C, H, and W denote the channel dimension, height, and width, respectively. Involution kernels are generated from input feature tensors. Specifically, $\boldsymbol{\mathcal{H}}_{i,j} = \delta(\mathbf{T}_{i,j})$, where $\mathbf{T}_{i,j}$ indexes a pixel and $\boldsymbol{\mathcal{H}}_{i,j}$ is conditioned on $\mathbf{T}_{i,j}$. δ is formulated as $\delta(\mathbf{T}_{i,j}) = \mathbf{W}_1\varphi(\mathbf{W}_0\mathbf{T}_{i,j})$, which is realized by two linear transformers, $\mathbf{W}_1 \in \mathbb{R}^{(K\times K\times G)\times \frac{C}{r}}$, $\mathbf{W}_0 \in \mathbb{R}^{\frac{C}{r}\times C}$, along with batch normalization with activation functions φ. Here, the reduction ratio r denotes a hyperparameter that signifies the intermediate channel dimension.

1) Involution: Involution addresses the limitations of adaptability related to different spatial positions and inter-channel redundancy. By leveraging spatial-specific and channel-agnostic approaches, involution achieves parameter-efficient feature extraction [14]. Involution process can be formulated as (5), where $\boldsymbol{\mathcal{H}} \in \mathbb{R}^{H\times W\times K\times K\times G}$ is a involution kernel, $k \in \{0, 1, \cdots, C_{out}\}$, C_{out} represents the number of output channels, and G represents the number of groups of channels, where channels within the same group share the involution kernel, resulting in a channel-agnostic characteristic. Fig. 3 provides an example when $G = 1$.

$$\mathbf{Z}_{i,j} = \sum_{p,q} \boldsymbol{\mathcal{H}}_{i,j,p,q,\lceil kG/C_{in}\rceil}\mathbf{T}_{i+p-\lfloor K/2\rfloor, j+q-\lfloor K/2\rfloor, k}. \quad (5)$$

2) ACmix: As illustrated in Fig. 4, ACmix is introduced to capture the dependencies of encoded representations in a parameter-efficient manner. It combines convolution and self-attention [15], as illustrated in Fig. 4. ACmix partitions both convolution and self-attention into two stages. Note that convolution and self-attention share the same standard 1×1 convolutions during the partitioning process, which helps reducing computational overhead.

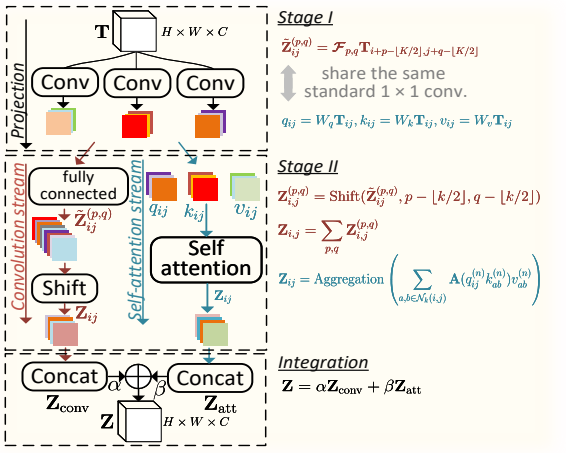


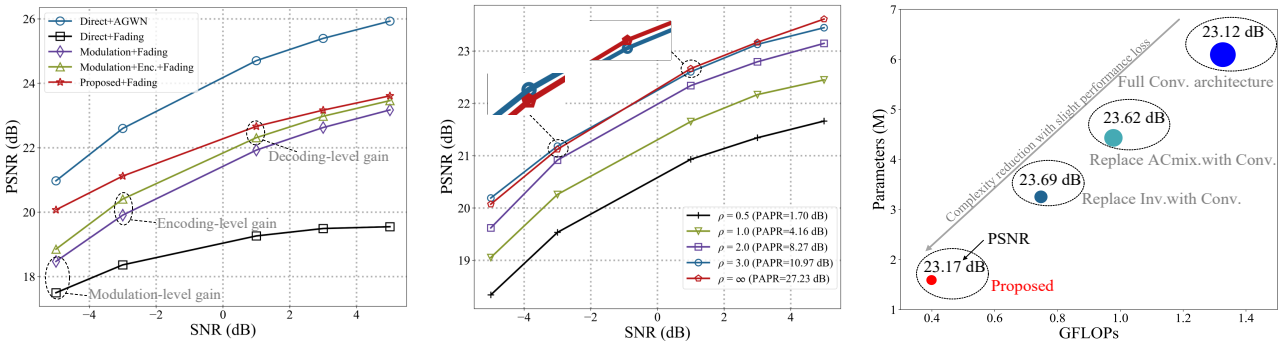
Fig. 4. An example of ACmix. Specifically, ACmix partitions the convolution process into two stages. At Stage I, projections are realized using a standard 1×1 convolution. And then, the output feature tensors are fed into a fully connected layer to generate K^2 feature tensors $\tilde{\mathbf{Z}}_{ij}^{(p,q)}$. At Stage II, the feature tensors $\tilde{\mathbf{Z}}_{ij}^{(p,q)}$ are shifted, specifically $\tilde{\mathbf{Z}}_{i,j} = \mathbf{Z}_{i+\Delta x, j+\Delta y}, \forall i, j$, where Δx and Δy denotes horizontal and vertical displacements, respectively. Finally, the shifted feature tensors are aggregated to form $\mathbf{Z}_{i,j}$. Meanwhile, ACmix partitions the self-attention into two stages. Stage I is similar to that of convolution, where $W_{q,k,v}$ are parameter matrices used for projection. At Stage II, attention weights are calculated, and the value matrices are aggregated. Finally, ACmix merges the convolution and attention branches and combines them with two learnable hyperparameters, α and β, resulting in $\mathbf{Z} = \alpha\mathbf{Z}_{\text{conv}} + \beta\mathbf{Z}_{\text{att}}$.

3) Additional Details: Different operators all adopt a residual connection as [3]. Meanwhile, we introduce an SNR-adaptive strategy that utilizes a single $(\boldsymbol{\phi}, \boldsymbol{\theta})$ pair across different SNRs. This approach has been widely shown to be beneficial for channels with varying SNRs [2], [3], [8].

C. Decoding Level

Unlike the existing methods that directly feed the encoding representations into the decoder for optimization [7], [8], we divide the decoding process into two stages: denoising and recovery. At the stage of denoising, we aim to eliminate the influence of fading noise on the encoding representations to maximize the correlation between source $\hat{\mathbf{x}}$ and its correlated version $\hat{\mathbf{x}}_{side}$. As stated in Selection III-C, MSE is adopted for recovering $\mathbf{s}$ under a fixed compression ratio. Hence, the optimization goal of $\hat{\mathbf{x}}_{side}$ is to help $\hat{\mathbf{X}}$ recovering. In other words, the correlation between the two representations (i.e., $\hat{\mathbf{x}}^d, \hat{\mathbf{x}}_{side}^d$ in Fig. 2) increases after denoising. At the stage of recovery, two encoding representations are concatenated at the channel dimension and jointly optimized by minimizing the term $\log p_{\boldsymbol{\theta}}(\mathbf{s} \mid \hat{\mathbf{x}}_{side}, \mathbf{x}_{side})$ in (2).

Different from resource-constrained encoder, the decoder at the base station is considered as computationally powerful. In other words, the design of the decoder is not constrained



(a) (b) (c)

Fig. 5. Validating the effectiveness of the proposed RDJSCC: (a) PSNR performance of different methods when $R = 1/16$. (b) PAPR for different clipping ratios ρ . (C) Model capacity of different encoders.

by model capacity. Therefore, both denoising and recovery stages employ convolutional neural networks.

V. Experiments

Next, we validate the performance of the proposed RDJSCC. First, we present the simulation settings. Then, we evaluate and analyze the performance of the proposed method.

A. Simulation Settings

1) Dataset: The KITTI dataset is adopted, which consists of stereo image pairs, with each pair captured by two cameras simultaneously. We use 1,576 pairs for training, 790 pairs for validation, and 790 pairs for testing. Each image, originally with a resolution of 375 × 1242 pixels, is center-cropped and downsampled to 128 × 256 pixels [13].

2) Implementation Details: The modulation and channel parameters are set as $N_p = 2$, $N_s = 3$, $N_c = 2048$, $L_{\text{cp}} = 16$, $L = 8$ and $\gamma = 4$. The number of reduction ratio r of involution is empirically set to 4. The number of groups G is set to 1, 8, and 2 in the order of stacking to meet the requirements for different input and output channel dimensions. All experiments are optimized based on Adam with a learning rate of 10^{-4}. The batch size is set as 5, and the number of training epochs is set to 200.

B. Performance Evaluation

To validate the effectiveness of our proposed framework, we assess it from three perspectives: performance analysis, PAPR reduction, and model capacity analysis, which correspond to the trade-offs discussed in Section IV.

1) Performance Analysis: As shown in Fig. 5(a), we present the PSNR performance from different levels. Notably, all methods are trained with under mixed SNR $\mathbf{\Gamma} \in \mathcal{U}(-5, 5)$, which is the same as [2], [3], [8]. "Direct+AWGN" indicates direct transmission using the original model over an AWGN channel without additional techniques to counter fading, serving as an upper bound. "Direct+Fading" signifies direct transmission over the fading channel. The distributed encoder and the joint decoder are the same as [3].

Comparing "Direct+AWGN" with "Direct+Fading", we observe a performance degradation in the latter, indicating that severe fading contaminates the encoded representation, thereby impairing the correlation among distributed sources. "Modulation+Fading" corresponds to the improvement at the modulation level described in in Section IV-A. Compared with "Direct+Fading", "Modulation+Fading" shows a noticeable performance gain, particularly at high SNR. This is because, in high SNR environments, the correlation between modulated sources is enhanced. Note that "Direct+Fading" also introduce pilots, but the key difference is that "Modulation+Fading" uses OFDM modulation. The performance improvement observed in "Modulation+Fading" can be attributed to the nature of OFDM, which operates on multi-carrier transmission principles. Unlike single-carrier systems, OFDM can approximate a frequency-selective fading channel as multiple frequency-flat fading channels. This characteristic allows OFDM to better handle fading effects, resulting in enhanced performance. Meanwhile, the orthogonality among subchannels avoids inter-carrier interference.

Next, we further analyze the gains at the encoding and decoding levels. "Modulation+Enc.+Fading" incorporates encoding techniques (described in Selection IV-B) on top of "Modulation+Fading". As depicted in Fig. 5(a), there is an obvious gain in comparison with "Modulation+Fading". One potential explanation is that the encoder based on parameter-efficient operators can capture more complex dependency relationships. Finally, combining "Modulation+Enc.+Fading" with decoding level yields our proposed RDJSCC. Compared with "Modulation+Enc.+Fading", RDJSCC has a performance gain, especially at low SNR, because the denoising stage of the decoder plays a significant role in low SNR environments. As SNR increases, the denoising performance tends to saturate.

2) PAPR Reduction: Here, we explore the trade-off between PAPR and performance by clipping. As shown in Fig. 5(b), it is obvious that the reconstruction performance decreases as clipping ratio decreases. This is attributed to clipping disrupting the orthogonality among subchannels.

However, we can strike a trade-off between performance and PAPR by setting an appropriate clipping ratio. For example, the reconstruction performance with $\rho = 3.0$ is nearly identical to the performance without clipping ($\rho = \infty$). The results are consistent with [9], [10]. This indicates that the performance gain brought by OFDM against fading and the low PAPR can coexist in a distributed image transmission system.

3) Model Capacity Analysis: Next, we examine the trade-off between model capacity and performance at the encoding level, as shown in Fig. 5(c). We evaluate complexity using two commonly used metrics: the number of floating-point operations (FLOPs) and parameters. By substituting ACMix and involution with convolution separately, we observe changes in performance and complexity. Additionally, we include the full convolutional architecture as a baseline encoder [3]. Results indicate that the proposed method is lightweight with a slight performance loss (within 0.5 dB). Although the proposed cannot enjoy the PSNR gains from both ACMix and involution, the complexity decreases significantly. Compared with [3], we achieve a better trade-off between model capacity and performance. This indicates that the encoder based on parameter-efficient operators can achieve an approximate solution with fewer computational resources.

Fig. 6. Examples of the reconstructed images when testing at $R = 1/16$ and $\mathbf{\Gamma} = -3$ dB.

4) Visual Comparison: In Fig. 5(a), we have already shown the reconstruction performance of the proposed method under a severe fading channel. To visually demonstrate the effects of fading channel, Fig. 6 presents the examples of the reconstructed images. Results show that the proposed method achieves a better recovery quality compared with direct transmission. We also observe that the recovered image contains some noise points in our proposed method, especially in the sky region (while areas such as roads and cars have fewer noise points). The noise originates from both compression and channel fading. During transmission, these sky regions may have weaker signal intensity, making them more prone to interference. Conversely, areas with object tend to retain more details, rendering them more resilient to interference. Compared with direct transmission, the proposed method exhibits significant improvement in noise reduction.

VI. Conclusion

In this paper, we propose a novel RDJSCC scheme, specifically designed for distributed sensor networks under severe fading conditions with imperfect CSI. We aim to mitigate the effects of such imperfections at three levels: modulation, encoding, and decoding, respectively. Additionally, we explore the trade-offs between PAPR, reconstruction performance, and complexity in practical applications. We find that it is feasible to balance the gains from OFDM in combating fading with a low PAPR by setting an appropriate clipping threshold. Meanwhile, deploying parameter-efficient operators at the encoder can strike a balance between reconstruction performance and complexity. We believe that the proposed RDJSCC scheme plays a pivotal role in advancing distributed systems toward practical deployment.

References

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

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

[3] W. Zhang, H. Zhang, H. Ma, H. Shao, N. Wang, and V. C. M. Leung, "Predictive and adaptive deep coding for wireless image transmission in semantic communication," *IEEE Trans. Wireless Commun.*, vol. 22, no. 8, pp. 5486–5501, Aug. 2023.

[4] J. Dai, S. Wang, K. Tan, Z. Si, X. Qin, K. Niu, and P. Zhang, "Nonlinear transform source-channel coding for semantic communications," *IEEE J. Select. Areas Commun.*, vol. 40, no. 8, pp. 2300–2316, Aug. 2022.

[5] E. Özyilkan and E. Erkip, "Distributed compression in the era of machine learning: A review of recent advances," in *Proc. 58th Annu. Conf. Inf. Sci. Syst. (CISS)*, 2024, pp. 1–6.

[6] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. 22, no. 1, pp. 1–10, Jan. 1976.

[7] S. F. Yilmaz, E. Özyilkan, D. Gündüz, and E. Erkip, "Distributed deep joint source-channel coding with decoder-only side information," *arXiv preprint arXiv:2310.04311*, 2023.

[8] S. Wang, K. Yang, J. Dai, and K. Niu, "Distributed image transmission using deep joint source-channel coding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2022, pp. 5208–5212.

[9] M. Yang, C. Bian, and H.-S. Kim, "OFDM-guided deep joint source channel coding for wireless multipath fading channels," *IEEE Trans. Cogn. Commun. Netw.*, vol. 8, no. 2, pp. 584–599, Jul. 2022.

[10] Y. Shao and D. Gündüz, "Semantic communications with discrete-time analog transmission: A PAPR perspective," *IEEE Wireless Commun. Lett.*, vol. 12, no. 3, pp. 510–514, Mar. 2023.

[11] J. Shao, Y. Mao, and J. Zhang, "Learning task-oriented communication for edge inference: An information bottleneck approach," *IEEE J. Select. Areas Commun.*, vol. 40, no. 1, pp. 197–211, Jan. 2022.

[12] J. Liu, S. Wang, and R. Urtasun, "DSIC: Deep stereo image compression," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 3136–3145.

[13] N. Mital, E. Özyilkan, A. Garjani, and D. Gündüz, "Neural distributed image compression using common information," in *Proc. IEEE Data Compression Conf. (DCC)*, Mar. 2022, pp. 182–191.

[14] D. Li *et al.*, "Involution: Inverting the inherence of convolution for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12316–12325.

[15] X. Pan *et al.*, "On the integration of self-attention and convolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 805–815.