Available online at https://jcupt.bupt.edu.cn

The Journal of China Universities of Posts and Telecommunications

February 2024, 31(1): 1 – 11

SNR-adaptive deep joint source-channel coding scheme for image semantic transmission with convolutional block attention module

Yang Yujia, Liu Yiming (✉), Zhang Wenjia, Zhang Zhi

State Key Laboratory of Networking and Switch Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract

With the development of deep learning (DL), joint source-channel coding (JSCC) solutions for end-to-end transmission have gained a lot of attention. Adaptive deep JSCC schemes support dynamically adjusting the rate according to different channel conditions during transmission, enhancing robustness in dynamic wireless environment. However, most of the existing adaptive JSCC schemes only consider different channel conditions, ignoring the different feature importance in the image processing and transmission. The uniform compression of different features in the image may result in the compromise of critical image details, particularly in low signal-to-noise ratio (SNR) scenarios. To address the above issues, in this paper, a dual attention mechanism is introduced and an SNR-adaptive deep JSCC mechanism with a convolutional block attention module (CBAM) is proposed, in which matrix operations are applied to features in spatial and channel dimensions respectively. The proposed solution concatenates the pooling feature with the SNR level and passes it sequentially through the channel attention network and spatial attention network to obtain the importance evaluation result. Experiments show that the proposed solution outperforms other baseline schemes in terms of peak SNR (PSNR) and structural similarity (SSIM), particularly in low SNR scenarios or when dealing with complex image content.

Keywords semantic communication, joint source-channel coding, image transmission

1 Introduction

With the development of artificial intelligence and computing technologies, semantic communication focuses on the semantic level of information, with the goal of precisely conveying the intended meaning of the message, which can improve transmission efficiency and have received widespread attention. Semantic communication usually leverages JSCC technique, which can jointly design and optimize source coding and channel coding processes, effectively addressing the "cliff effect", reducing communication bandwidth and improving communication robustness. For image transmission, DL techniques are widely employed in semantic communication systems because of their capability to extract intricate features from images[1-2]. Specifically, Bourtsoulatze et al. introduced a deep

Received date: 21-11-2023

Corresponding author: Liu Yiming, E-mail: liuyiming@bupt.edu.cn

DOI: 10.19682/j.cnki.1005-8885.2024.2001

JSCC scheme[1], which involves the direct mapping of pixel values from an image to complex-valued channel input symbols. The scheme outperforms traditional transmission solutions, making it especially effective in low SNR scenarios. Yang et al. proposed a deep JSCC scheme with adaptive rate control capability for wireless image transmission[3], in which the outputs of semantic encoder are divided into selective features and non-selective features. An introduced policy network is employed to determine the transmission of selective features in conjunction with non-selective features. Experiments illustrated that the proposed deep JSCC method employing a single model can achieve comparable performance to the optimized model particularly trained with static target rate. Kurka et al. proposed a JSCC scheme that incorporates channel output feedback, which shows good performance in terms of reconstruction quality for end-to-end fixed-length image transmission, and reduces the average delay in variable-length image transmission[4].

In image processing and transmission tasks, the attention mechanism can be viewed as a dynamic selection process of crucial features input to the image, implemented through adaptive weights. Specifically, in image processing tasks, Zhang et al. proposed a hierarchical structure for extracting the semantic information captured by the encoder[5]. To address multi-task-oriented image features, this structure employs multi-attention networks to extract image features at the pixel level. Kang et al. introduced a task-oriented semantic communication framework that employs an efficient image retrieval approach[6]. In Ref. [6], a personalized attention-based mechanism is designed to achieve personalized semantic communication by enabling the differential weight encoding of triplets for crucial information based on user preferences. In image transmission tasks, Xu et al. proposed an attention-based deep JSCC (ADJSCC) scheme[7]. This scheme allows for dynamic adjustment of the transmission rate according to different SNRs during transmission, without training a number of neural networks to cover scenarios with varying SNR levels. Bao et al. proposed an ADJSCC-l architecture for image transmission[8]. This architecture incorporates an SNR-adaptive module, providing excellent resilience against the mismatch between the trained and tested channel SNRs resulting from channel variations. Simulation results show that the proposed ADJSCC-l architecture can successfully improve the reconstruction quality for wireless image transmission in low SNR and bandwidth-limited scenarios. However, in Ref. [7], the attention mechanism of the ADJSCC is essentially a squeeze excitation network that only considers the importance of features in different channels but ignores the importance of pixels at various positions, resulting in the loss of certain valuable information during the image compression and transmission process, affecting the accuracy of image transmission.

To address the above issues, in this paper, an SNR-adaptive deep JSCC framework with attention weight allocation mechanism is designed. The proposed framework aims to effectively enhance the reconstruction quality under a specific compression ratio (CR). The main contributions are outlined as follows.

1) A dual attention mechanism is introduced and an SNR-adaptive deep JSCC framework with a CBAM is proposed. The proposed framework consists of a feature extraction module, a semantic importance evaluation module, JSCC encoding and decoding modules, and an image reconstruction module. By employing semantic importance evaluation module, it can achieve better transmission performance under various channel conditions within a single trained network, saving storage space significantly.

2) The integration of channel attention and spatial attention mechanisms is introduced, taking into account the interdependence of spatial and channel characteristics. The proposed SNR-adaptive CBAM takes into account varying SNR levels and image contents, allocating distinct weights to different features by concatenating the pooling feature with the

SNR level.

3) An image semantic encoder is designed by employing a residual structure, which is implemented to mitigate gradient disappearance and overfitting when training the network, enhancing the training effect of the model.

The remaining parts of this paper are organized as follows. Sect. 2 demonstrates the end-to-end semantic communication system based on attention mechanism. Sect. 3 presents the specific network of our proposed SNR-adaptive CBAM and image semantic encoder. The simulation results and analysis are illustrated and analyzed in Sect. 4. Finally, Sect. 5 concludes this paper.

2 System model

Semantic communication systems consist of three essential components: the transmitter, the channel, and the receiver. An end-to-end image semantic communication system based on the attention mechanism is proposed, as depicted in Fig. 1.

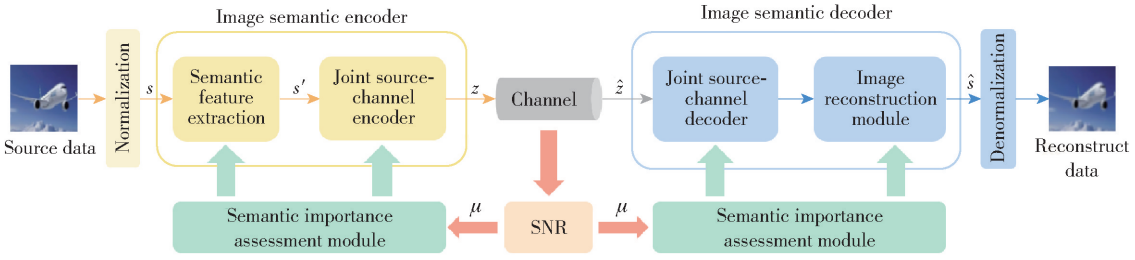


Fig. 1 End-to-end image semantic communication system based on attention mechanism

The system extracts source features through a feature extractor based on DL, and assigns different weights to different features with the help of joint training of the semantic importance evaluation module and the JSCC encoder. The JSCC decoder restores the semantic features, and utilizes the image reconstruction module to reconstruct the target image at the receiver. The proposed architecture uses a semantic importance evaluation module, enabling simultaneous consideration of both image contents and channel SNR conditions for image compression and reconstruction. This approach efficiently allocates attention weights to more critical tasks that can adaptively adjust the data rate in diverse channel environments to enable efficient semantic transmission. The specific structure of the network is as follows.

2.1 Transmitter of the proposed architecture

The transmitter of the proposed architecture consists of three parts: Semantic feature extractor, semantic importance assessment module, and JSCC encoder. The input image of size $n = H$ (the height of image) $\times$ W (the width of image) $\times$ C (the number of image channels) is represented by a vector $\boldsymbol{s} \in \mathbb{R}^n$, where $\mathbb{R}$ represents the set of real numbers.

The specific process involves several steps. Firstly, the image $\boldsymbol{s}$ undergoes preprocessing through a normalization layer, mapping image pixel value to the range of 0 to 1. Secondly, distinct semantic features of the input image are extracted through a semantic feature extractor based on convolutional neural network (CNN), combined with the channel SNR $\mu \in \mathbb{R}$. A semantic importance assessment module is used to evaluate importance of features and reassign weights to obtain the semantic attention feature vector $\boldsymbol{s}' \in \mathbb{R}^{H \times W \times C}$; Finally, the JSCC encoder is employed to encode the semantic attention feature vector $\boldsymbol{s}'$ and the channel feedback SNR μ, resulting in a vector of complex-valued channel input symbols $\boldsymbol{z} \in \mathbb{C}^k$, where k is the size of the channel input symbol, and $\mathbb{C}$ denotes the set of complex numbers. The encoding process can be expressed in Eqs. (1) and (2)

$$\boldsymbol{s}' = \boldsymbol{M}_\beta(\boldsymbol{T}_\alpha(\boldsymbol{s}), \boldsymbol{\mu}) \tag{1}$$

$$\boldsymbol{z} = \boldsymbol{M}_\beta(\boldsymbol{E}_\theta(\boldsymbol{s}'), \boldsymbol{\mu}) \tag{2}$$

where $\boldsymbol{M}_{\beta}(\cdot)$ serves as the semantic importance evaluator, with its network parameter designated as $\boldsymbol{\beta}$. $\boldsymbol{T}_{\alpha}(\cdot)$ is the semantic feature extraction network, with its network parameter denoted as α. $\boldsymbol{E}_{\theta}(\cdot)$ corresponds to the encoding function of the JSCC encoder, the network parameter is indicated as $\boldsymbol{\theta}$.

Overall, the transmitter maps an n-dimensional vector of the real-valued image $\boldsymbol{s}$ to a k-dimensional vector of the channel input samples $\boldsymbol{z}$. To adhere to the average power constraint of the JSCC encoder, a power normalization $(1/k)\boldsymbol{E}(\boldsymbol{z}\boldsymbol{z}^{*}) \leqslant 1$ is also enforced, where $\boldsymbol{z}^{*}$ represents the complex conjugate of $\boldsymbol{z}$.

2.2 Receiver of the proposed architecture

The receiver of the proposed architecture comprises three components: JSCC decoder, semantic importance assessment module and image reconstruction module. The reshaping layer reorganizes the received signal, collaborates with the semantic channel decoder to mitigate noise interference in the additive white Gaussian noise (AWGN) channel, and restores the semantic features. The image reconstruction module deeply mines semantic information through the attention mechanism, fuses semantic features and reconstructs the target image. The denormalization layer scales each element back to the value of each pixel within the range of 0 to 255. Specifically, $\boldsymbol{z}$ is transmitted to the receiver through the physical channel. The output symbol of channel $\hat{\boldsymbol{z}}$ received by the JSCC decoder is expressed in Eq. (3)

$$\hat{\boldsymbol{z}} = \boldsymbol{z} + \boldsymbol{\omega} \tag{3}$$

where the vector $\boldsymbol{\omega} \in \mathbb{C}^{k}$ is composed of independent and identically distributed (i. i. d) samples with the distribution $\mathcal{CN}(0, \sigma^{2}\boldsymbol{I})$. σ^{2} is the noise power and $\mathcal{CN}(\cdot,\cdot)$ represents a circularly symmetric complex Gaussian distribution. $\boldsymbol{I}$ is the identity matrix.

The JSCC decoder uses the decoding function $R_{\xi}(\cdot)$ to map $\hat{\boldsymbol{z}}$ and $\boldsymbol{\mu}$, ξ is the parameter set; the image reconstruction module uses the reconstruction function $R_{\eta}(\cdot)$ to reconstruct the image, η is the parameter set. The reconstructed image of the receiver is expressed in Eq. (4)

$$\hat{\boldsymbol{s}} = \boldsymbol{R}_{\eta}(\boldsymbol{R}_{\xi}(\hat{\boldsymbol{z}}, \boldsymbol{\mu}), \boldsymbol{\mu}) \tag{4}$$

where $\hat{\boldsymbol{z}} \in \mathbb{C}^{k}$ represents the signal received by the channel, and $\hat{\boldsymbol{s}} \in \mathbb{R}^{n}$ is the estimate of the original image $\boldsymbol{s}$.

2.3 Loss function of the proposed architecture

The mean square error (MSE) distribution between the original image $\boldsymbol{s}$ and the reconstructed image $\hat{\boldsymbol{s}}$ is expressed in Eq. (5)

$$d(s,\hat{s}) = \frac{1}{n}\| s - \hat{s} \|^{2} = \frac{1}{n}\sum_{i=1}^{n}(s_i - \hat{s}_i)^{2} \tag{5}$$

where s_i and $\hat{s}_i$ depict the color component intensity of each pixel related to $\boldsymbol{s}$ and $\hat{\boldsymbol{s}}$ respectively.

In this paper, a CNN is used to model the JSCC encoder and decoder. The objective of the CNN is to find optimal parameters to minimize the distortion $\boldsymbol{\theta}^{*}$ and $\boldsymbol{\xi}^{*}$. The loss function of the network is expressed in Eq. (6)

$$(\theta^{*}, \xi^{*}) = \arg\min_{\theta,\xi}\{E_{p(\mu)}E_{p(s,\hat{s})}[d(s,\hat{s})]\} \tag{6}$$

where θ^{*} represents the optimal parameter of the encoder, ξ^{*} represents the optimal parameter of the decoder, $p(s,\hat{s})$ is the joint probability distribution of s and $\hat{s}$, and $p(\mu)$ is the probability distribution of the SNR.

3 Proposed semantic encoder based on SNR-adaptive CBAM

In this section, a dual attention mechanism based on SNR-adaptive mechanism with CBAM named SNR-adaptive CBAM is proposed. The channel attention module and spatial attention module can effectively enhance the model's perception ability by prioritizing crucial features and inhibiting irrelevant features. SNR-adaptive CBAM can be incorporated into many network architectures as a plug-and-play module. Furthermore, a detailed network architecture is presented for the proposed image JSCC encoder, which is designed to efficiently extract semantic features.

3.1 Proposed SNR-adaptive CBAM

The specific structure of the SNR-adaptive CBAM is illustrated in Fig. 2. Comprising two sequentially connected modules, the proposed module is designed for integration into main layers of neural networks, such as convolutional and transposed convolutional layers.

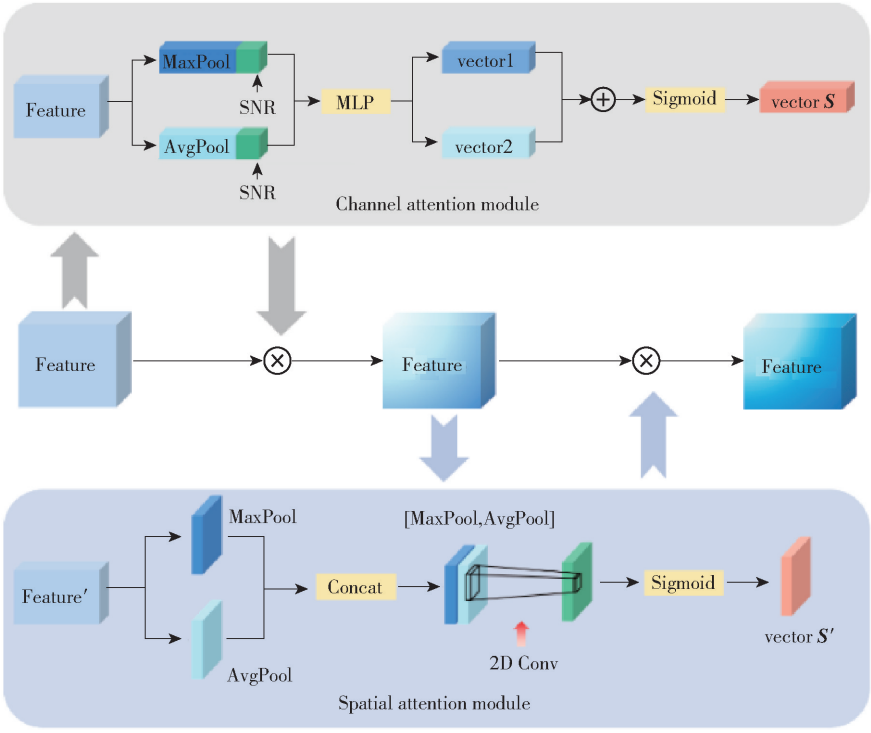


Fig. 2 Specific structure of SNR-adaptive CBAM

The SNR-adaptive CBAM is implemented by assessing the importance weights of the features. Specifically, for the feature map $\boldsymbol{F} \in \mathbb{R}^{H\times W\times C}$ obtained from the feature extraction module, the SNR-adaptive CBAM module derives the attention map across channel and spatial dimensions based on SNR levels. It derives a one-dimensional (1D) channel attention map $\boldsymbol{M}_c \in \mathbb{R}^{1\times 1\times C}$ and a two-dimensional (2D) spatial attention map $\boldsymbol{M}_s \in \mathbb{R}^{H\times W\times 1}$, which are combined with the input feature map. After multiplication, a new feature map $\boldsymbol{F}'' \in \mathbb{R}^{H\times W\times C}$ is generated for joint encoding and decoding. The specific process is expressed in Eq. (7)

$$\begin{aligned}\boldsymbol{F}' &= \boldsymbol{M}_c(\boldsymbol{F})\otimes\boldsymbol{F}\\ \boldsymbol{F}'' &= \boldsymbol{M}_s(\boldsymbol{F}')\otimes\boldsymbol{F}'\end{aligned} \tag{7}$$

where $\otimes$ denotes element-wise multiplication operation. Throughout the multiplication process, attention weights are appropriately spread. Channel attention weights are elongated across the spatial dimension, whereas spatial attention weights are elongated across the channel dimension. The particulars of each attention module are elucidated below.

1) Channel attention module

The channel attention module mainly focuses on the inter-channel relationships of features. To infer more refined channel attention, both average pooling and max-pooling are considered during image compression in the spatial dimension. Average pooling gives feedback for all pixels in the feature map, whereas max-pooling only obtains feedback for the location with the highest pixel value when calculating gradient backpropagation. Initially, spatial information is consolidated from the feature map through the utilization of both average pooling and max-pooling operations. This process yields features derived from

average pooling $\boldsymbol{F}_{\text{avg}}^{c} \in \mathbb{R}^{1\times1\times C}$ and max-pooling $\boldsymbol{F}_{\max}^{c} \in \mathbb{R}^{1\times1\times C}$. These obtained features are concatenate with the channel feedback SNR μ to generate contextual information $\boldsymbol{F}_{\text{avg}}^{c\prime}$ and $\boldsymbol{F}_{\max}^{c}$ as depicted in Eqs. (8) and (9)

$$\boldsymbol{F}_{\text{avg}}^{c\prime} = [\boldsymbol{F}_{\text{avg}}^{c}, \mu] \in \mathbb{R}^{C+1} \tag{8}$$

$$\boldsymbol{F}_{\max}^{c\prime} = [\boldsymbol{F}_{\max}^{c}, \mu] \in \mathbb{R}^{C+1} \tag{9}$$

After reshaping, the features individually fed into identical multi-layer perceptrons (MLPs) with hidden layers to generate attention map $\boldsymbol{M}_c \in \mathbb{R}^{1\times1\times C}$. To minimize parameter overhead, the hidden activation size is set to $\mathbb{R}^{1\times1\times(C/r)}$, where r is the reduction ratio (set to 16 in this paper). The output features are consolidated using element-wise summation to generate the channel attention map $\boldsymbol{M}_c(F) \in \mathbb{R}^{1\times1\times C}$. The channel attention is calculated as shown in Eq. (10)

$$\begin{aligned}\boldsymbol{M}_c(F) &= \sigma(\text{MLP}(\boldsymbol{F}_{\text{avg}}^{c\prime}) + \text{MLP}(\boldsymbol{F}_{\max}^{c\prime})) = \\ &\sigma(\boldsymbol{W}_1\delta(\boldsymbol{W}_0(\boldsymbol{F}_{\text{avg}}^{c\prime})) + \boldsymbol{W}_1\delta(\boldsymbol{W}_0(\boldsymbol{F}_{\max}^{c\prime})))\end{aligned} \tag{10}$$

where δ and σ represent the activation functions rectified linear unit (ReLU) and Sigmoid function respectively. The two inputs of the network share the MLP weight parameters $\boldsymbol{W}_0$ and $\boldsymbol{W}_1$, $\boldsymbol{W}_0 \in \mathbb{R}^{C/r\times C}$, and $\boldsymbol{W}_1 \in \mathbb{R}^{C\times C/r}$.

2) Spatial attention module

The spatial attention module mainly focuses on the spatial relationship of features. The feature map output by the channel attention module serves as the input feature map for this module. Initially, average pooling and max-pooling operations are applied along the channel axis to aggregate the channel information of the feature map, generating two 2D feature maps: $\boldsymbol{F}_{\text{avg}}^{s} \in \mathbb{R}^{H\times W\times1}$ and $\boldsymbol{F}_{\max}^{s} \in \mathbb{R}^{H\times W\times1}$. After concatenating them to generate effective feature descriptors, a convolutional layer and sigmoid function layer are applied to reduce dimensionality, producing a 2D spatial attention map $\boldsymbol{M}_s(\boldsymbol{F}') \in \mathbb{R}^{W\times H\times1}$. The specific calculation of spatial attention is expressed in Eq. (11)

$$\boldsymbol{M}_s(\boldsymbol{F}') = \sigma(f^{7\times7}([\boldsymbol{F}_{\text{avg}}^{s\prime}, \boldsymbol{F}_{\max}^{s\prime}])) \tag{11}$$

where σ represents the sigmoid function and $f^{7\times7}$ represents a convolution operation with the kernel size of 7×7.

3.2 Joint source-channel encoder for image transmission

The proposed image semantic encoder is made up of two modules: A semantic feature extraction network and a JSCC encoder, both of which are composed of multiple nonlinear layers. It serves as source encoder and channel encoder that the input of the image semantic encoder is source values and the output of the image semantic encoder is channel symbols. The main intention is to utilize a CNN to extract different features within the image and subsequently transmit them to the channel after encoding.

As shown in Fig. 3, a generate feature block (GFB) comprises a convolutional layer, a generalized divisive normalization (GDN) layer[9] and a parametric ReLU (PReLU) layer[10]. Among these components, the convolution layer is specified by parameters $m\times m\times C|\downarrow s_t$, which means that the convolution kernel size is m and there are C output channels. The symbol $\downarrow$ represents downsampling operations, and the parameter s_t represents the stride in the convolution layer. Residual convolution block (RCB) is composed of a residual network incorporating two convolutional layers, two GDN layers and two PReLU layers. This configuration is implemented to mitigate gradient disappearance and overfitting during network training, enhancing the training effect of the model[11]. Each GFB and RCB is followed by an SNR-adaptive CBAM to assign feature weights based on image contents and channel information. In traditional JSCC, batch normalization (BN) introduces varying mean and standard deviation for each processed batch, which is equivalent to the adding noise. However, this approach is unsuitable for generative models like image reconstruction and compression.

Therefore, the GDN layer is introduced to normalize the feature, making it especially well-suited for image

reconstruction. The GDN layer expands the channel of each module in the convolutional layer, which is effective in addressing image compression and transmission tasks[9]. Additionally, the paper incorporates the pre-activation residual neural network (pre-activation ResNet)[12], which enhances the training effect and transmission accuracy of the model by altering the position of the normalization layer in the residual network. The image semantic decoder performs the opposite operation sequentially.

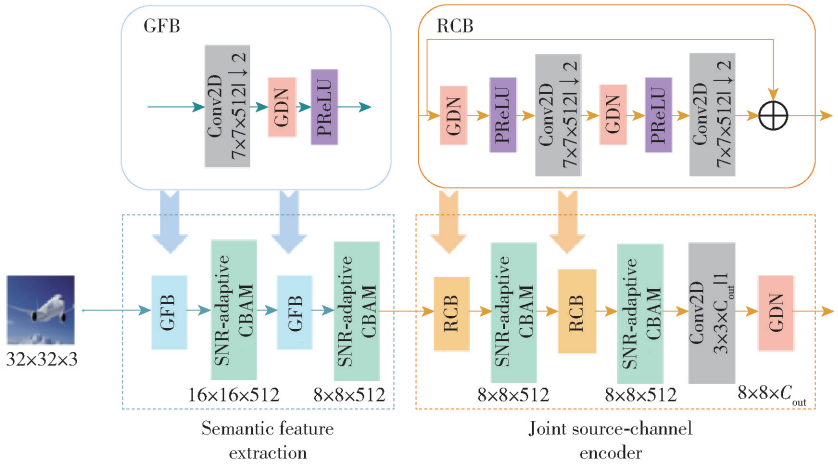


Fig. 3 Image semantic encoder components

4 Simulation results and analysis

In this section, the simulation parameters in the experiments are presented. Subsequently, simulation results are conducted to evaluate the performance of the proposed scheme.

4.1 Simulation parameters

This paper utilizes the canadian institute for advanced research (CIFAR)-10 dataset[13] to train and evaluate the proposed scheme. The CIFAR-10 dataset comprises 60 000 color images, each consisting of 32 × 32 pixels. The training dataset and testing dataset encompass 50 000 and 10 000 images respectively. The proposed scheme is implemented using TensorFlow[14] and Keras, which is a high-level application programming interface (API) designed for constructing and training DL models. An adaptive moment (Adam) estimation optimizer[15] is chosen for optimization, in which the learning rate is configured as 0.000 1 and the batch size is set to 128. The experiments maintain a fixed training epoch equal to 1 024 to measure training efficiency. In this paper, the number of transmit channel is set to 16, and the minimum training loss is set to 10^8.

The average PSNR and SSIM[16] are used for quality measurement to evaluate performance. PSNR is a measure of the ratio between the maximum potential power of a signal and the detrimental noise power that impacts its representation accuracy, determined by the MSE. The average MSE of N images is defined in Eq. (12)

$$\overline{X}_{\mathrm{MSE}} = \frac{1}{N}\sum_{i=1}^{N} d(s^{(i)}, \hat{s}^{(i)}) \tag{12}$$

where $s^{(i)}$ and $\hat{s}^{(i)}$ respectively denote the ith original image and the its reconstructed counterpart. Correspondingly, PSNR is computed in Eq. (13)

$$X_{\mathrm{PSNR}} = 10\lg\frac{X_{\max}^2}{X_{\mathrm{MSE}}} \tag{13}$$

where $X_{\max}$ denotes the maximum pixel value of the image (if each pixel in the image is 8 bit binary, the $X_{\max}$ is 255). The X_{MSE} denotes the value of MSE. The PSNR for each image is initially calculated and subsequently averaged over all test images. SSIM quantifies the similarity between two digital images. It

employs three criteria to measure images: Luminance, contrast and structure. The SSIM between two given images s and $\hat{s}$ is defined in Eq. (14)

$$X_{\mathrm{SSIM}}(s,\hat{s}) = \frac{(2\mu_s\mu_{\hat{s}} + C_1)(2\sigma_{s\hat{s}} + C_2)}{(\mu_s^2 + \mu_{\hat{s}}^2 + C_1)(\sigma_s^2 + \sigma_{\hat{s}}^2 + C_2)} \tag{14}$$

where μ_s and $\mu_{\hat{s}}$, σ_s and $\sigma_{\hat{s}}$ are the mean and standard deviation of s and $\hat{s}$, $\sigma_{s\hat{s}}$ is the covariance of s and $\hat{s}$, C_1, C_2 are all constants employed for preserving the stability of luminance, contrast and structure. The expression of the SNR μ of the proposed scheme is shown in Eq. (15)

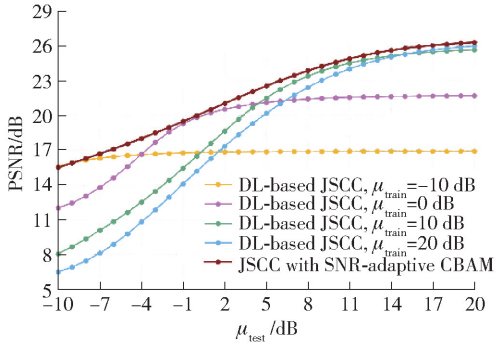
$$\mu = 10\lg\frac{P_s}{P_n} \tag{15}$$

where P_s represents the power of signal and P_n represents the power of noise.

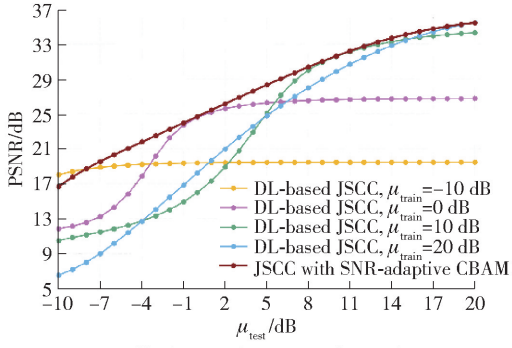
4.2 Simulation results

To validate the effectiveness of our proposed SNR-adaptive CBAM scheme, the DL-based JSCC architecture utilized in Ref. [17] is selected as a benchmark for comparison. Training the JSCC scheme under a specific SNR can produce comparable performance to the separate source channel coding (SSCC) solution using joint photographic experts group 2000 (JPEG2000) for source coding and low density parity check (LDPC) code for channel coding. The paper trains the DL-based JSCC scheme under μ_{train} = −10 dB, 0 dB, 10 dB, and 20 dB respectively, where μ_{train} is training SNR value. The SNR-adaptive CBAM architecture is trained with a distinct distribution covering the range of $\mu_{\mathrm{train}} \in [-10\ \mathrm{dB}, 20\ \mathrm{dB}]$. The performance of the proposed SNR-adaptive CBAM scheme and the DL-based JSCC scheme is evaluated under $\mu_{\mathrm{test}} \in [-10\ \mathrm{dB}, 20\ \mathrm{dB}]$. To alleviate the influence of channel noise randomness on the test results, all images in the testing dataset undergo transmission 10 times across the AWGN channel.

Fig. 4(a) and Fig. 4(b) show the results of the proposed JSCC with SNR-adaptive CBAM scheme and the DL-based JSCC scheme shown in Ref. [1] when the R_C (i.e. CR) is 1/12 and 1/3 respectively. Among them, the CR is defined as $R_C = k/n$, where k represents the number of pixels required for the image after compression, n represents the number of pixels required for the original image. A smaller CR indicates that the source occupies fewer channel resources, demanding higher performance from the model.



(a) PSNR performance with R_C=1/12



(b) PSNR performance with R_C=1/3

Fig. 4 PSNR performance comparison with different CRs on AWGN channels

In Fig. 4(a), when $R_C = 1/12$, the proposed JSCC with SNR-adaptive CBAM scheme outperforms the DL-based JSCC model trained under a specific SNR across all intervals. As test SNR value μ_{test} increases, the PSNR of all schemes gradually increases. For the DL-based JSCC scheme, optimal results are achieved when the μ_{test} is similar to its corresponding μ_{train}. Even when $\mu_{\mathrm{test}} = \mu_{\mathrm{train}}$, the performance of our scheme remains superior to the DL-based JSCC scheme, if μ_{test} deviates from μ_{train}, the advantages of the proposed scheme are more obvious. Specifically, when training the DL-based JSCC scheme under a SNR of −10 dB, its performance is superior compared to other training methods under low SNR conditions. However, in a

favorable SNR environment, the performance of the DL-based JSCC scheme is relatively poor due to significant differences between the testing and training environments. In Fig. 4 (b), when $R_C = 1/3$, the PSNR of each scheme is higher, indicating better overall model performance. While the performance gain of the proposed scheme gradually disappears under specific SNR conditions, it still demonstrates robust performance in the overall analysis.

A comparative analysis is conducted by evaluating the proposed scheme against the baseline scheme on the CIFAR-10 dataset. The corresponding accuracy values under various CRs are presented in Table 1 and Table 2.

Table 1 Accuracy performance comparison with a CR of 1/3

Strategy name	μ_{train}/dB	Accuracy			
		$\mu_{test} = -10$ dB	$\mu_{test} = 0$ dB	$\mu_{test} = 10$ dB	$\mu_{test} = 20$ dB
JSCC with SNR-adaptive CBAM	[-10,20]	0.628 8	0.781 1	0.858 2	0.879 5
DL-based JSCC	-10	0.652 7	0.688 4	0.690 4	0.690 6
DL-based JSCC	0	0.549 4	0.779 1	0.810 6	0.813 4
DL-based JSCC	10	0.464 7	0.737 5	0.856 8	0.871 8
DL-based JSCC	20	0.404 8	0.679 7	0.845 3	0.881 9

Table 2 Accuracy performance comparison with a CR of 1/12

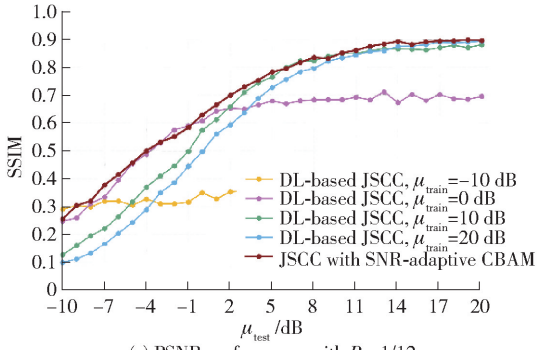
Strategy name	μ_{train}/dB	Accuracy			
		$\mu_{test} = -10$ dB	$\mu_{test} = 0$ dB	$\mu_{test} = 10$ dB	$\mu_{test} = 20$ dB
JSCC with SNR-adaptive CBAM	[-10,20]	0.581 3	0.704 5	0.783 0	0.802 5
DL-based JSCC	-10	0.592 7	0.639 4	0.645 0	0.645 6
DL-based JSCC	0	0.544 5	0.700 6	0.738 1	0.742 2
DL-based JSCC	10	0.478 5	0.673 0	0.776 8	0.796 9
DL-based JSCC	20	0.453 6	0.646 6	0.773 3	0.804 5

Table 1 and Table 2 clearly indicates that, irrespective of the R_C being 1/3 or 1/12, the proposed scheme consistently achieves the highest accuracy when $\mu_{test} = -10$ dB, 0 dB and 10 dB. Remarkably, even under challenging conditions such as an extremely poor SNR (e.g., -10 dB) or exceptional conditions such as an excellent SNR (e.g., 20 dB), our scheme exhibits only marginal performance degradation compared to the baseline scheme when $\mu_{test} = \mu_{train}$. Significantly, our scheme exhibits a pronounced advantage when μ_{test} deviates from μ_{train}. This is attributed to the effective integration of information pertaining to varying channel conditions and image contents by the SNR-adaptive CBAM, enabling the dynamic allocation of attention weights. Consequently, our scheme proves robust for image classification and transmission across a diverse range of channel conditions.

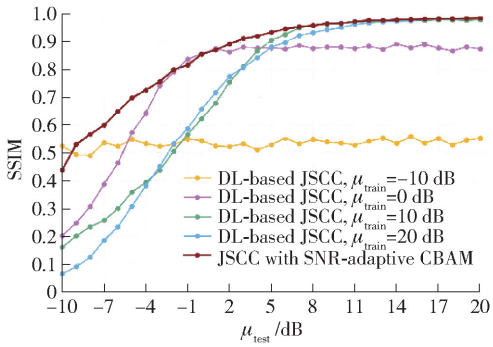
The perceptual SSIM is also adopted for evaluating image quality. Fig. 5 (a) and Fig. 5 (b) present the performance comparison of different schemes under the SSIM evaluation criterion. The proposed scheme exhibits high SNR-adaptive performance, irrespective of the $R_C = 1/3$ or 1/12, outperforming other schemes across various SNRs. Even when $\mu_{test} = \mu_{train}$, it is still able to achieve similar performance comparable to the DL-based JSCC scheme.

Based on the above analysis, it shows that JSCC with SNR-adaptive CBAM still exhibits good PSNR and SSIM performance in a challenging channel environment (such as SNR is 0 dB) a or a low CR (such as $R_C = 1/12$). By utilizing the attention

mechanism, it demonstrates outstanding SNR adaptive characteristics, enabling adaptation to different SNR levels based on a single training model. Moreover, SNR-adaptive CBAM significantly reduces model complexity and the total number of training parameters, effectively alleviating storage pressure and improving the efficiency of image transmission.



(a) PSNR performance with R_C=1/12



(b) PSNR performance with R_C=1/3

Fig. 5 SSIM performance comparison with different CRs on AWGN channels

5 Conclusions

In this paper, an SNR-adaptive deep JSCC mechanism with CBAM is proposed to achieve better transmission performance under various channel conditions. A semantic encoder model is designed using a residual structure, effectively extracting semantic features to facilitate image transmission tasks. Additionally, an SNR-adaptive CBAM is proposed to combine the varying channel conditions and image contents, enabling the allocation of attention weights. Extensive simulation results demonstrate that the proposed SNR-adaptive image semantic transmission framework achieves better performance in terms of PSNR and SSIM.

In future research, a potential direction is to collaboratively design a channel estimation module which can effectively extend the proposedscheme to a real channel environment. It can also incorporate certain anti-interference operations, potentially enhancing reconstruction performance and facilitating practical applications.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (62293481), in part by the Young Elite Scientists Sponsorship Program by CAST (2023QNRC001), in part by the National Natural Science Foundation for Young Scientists of China (62001050), and in part by the Fundamental Research Funds for the Central Universities (2023RC95).

References

[1] BOURTSOULATZE E, KURKA D B, GÜNDÜZ D. Deep joint source-channel coding for wireless image transmission. IEEE Transactions on Cognitive Communications and Networking, 2019, 5(3): 567 – 579.

[2] KURKA D B, GÜNDÜZ D. Successive refinement of images with deep joint source-channel coding. Proceedings of the IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC'19), 2019, Jul 2 – 5, Cannes, France. Piscataway, NJ, USA: IEEE, 2019: 1 – 5.

[3] YANG M Y, KIM H S. Deep joint source-channel coding for wireless image transmission with adaptive rate control. Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'22), 2022, May 23 – 27, Singapore. Piscataway, NJ, USA: IEEE, 2022: 5193 – 5197.

[4] KURKA D B, GÜNDÜZ D. Deep joint source-channel coding of images with feedback. Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'20), 2020, May 4 – 8, Barcelona, Spain. Piscataway, NJ, USA: IEEE, 2020: 5235 – 5239.

[5] ZHANG Z G, YANG Q Q, HE S B, et al. Semantic communication approach for multi-task image transmission. Proceedings of the IEEE 96th Vehicular Technology Conference (VTC-Fall'22), 2022, Sept 26 – 29, London, UK. Piscataway, NJ, USA: IEEE, 2022: 1 – 2.

[6] KANG J W, DU H Y, LI Z H, et al. Personalized saliency in

task-oriented semantic communications: Image transmission and performance analysis. IEEE Journal on Selected Areas in Communications, 2023, 41(1): 186 – 201.

[7] XU J L, AI B, CHEN W, et al. Wireless image transmission using deep source channel coding with attention modules. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(4): 2315 – 2328.

[8] BAO X W, JIANG M, ZHANG H. ADJSCC-l: SNR-adaptive JSCC networks for multi-layer wireless image transmission. Proceedings of the 7th International Conference on Computer and Communications (ICCC'21), 2021, Dec 10 – 13, Chengdu, China. Piscataway, NJ, USA: IEEE, 2021: 1812 – 1816.

[9] BALLÉ J, LAPARRA V, SIMONCELLI E P. Density modeling of images using a generalized normalization transformation. arXiv Preprint, arXiv: 1511.06281, 2016.

[10] HE K M, ZHANG X Y, REN S Q, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV'15), 2015, Dec 7 – 13, Santiago, Chile. Piscataway, NJ, USA: IEEE, 2015: 1026 – 1034.

[11] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16), 2016, Jun 27 – 30, Las Vegas, NV, USA. Piscataway, NJ, USA: IEEE, 2016: 770 – 778.

[12] HE K M, ZHANG X Y, REN S Q, et al. Identity mappings in deep residual networks. Computer Vision: Proceedings of the 14th European Conference on Computer Vision (ECCV'16), 2016, Oct 11 – 14, Amsterdam, Netherlands. LNIP 9908. Berlin, Germany: Springer, 2016: 630 – 645.

[13] KRIZHEVSKY A. Learning multiple layers of features from tiny images. Corpus ID: 18268744. Toronto, Canada: University of Toronto, 2009. https: // www. cs. toronto. edu/~ kriz/learning-features-2009-TR. pdf.

[14] ABADI M, AGARWAL A, BARHAM P, et al. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. arXiv Preprint, arXiv: 1603.04467, 2016.

[15] KINGMA D P, BA J. Adam: A method for stochastic optimization. arXiv Preprint, arXiv: 1412.6980, 2014.

[16] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, 2004, 13(4): 600 – 612.

[17] KURKA D B, GÜNDÜZ D. DeepJSCC-f: Deep joint source-channel coding of images with feedback. IEEE Journal on Selected Areas in Information Theory, 2020, 1(1): 178 – 193.

(Editor: Ai Lisha)