sensors

MDPI

Article

CNN and Attention-Based Joint Source Channel Coding for Semantic Communications in WSNs

Xinyue Liu, Zhen Huang, Yulu Zhang, Yunjian Jia and Wanli Wen *

School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 401331, China; 27168752@alu.cqu.edu.cn (X.L.); zhenhuang@cqu.edu.cn (Z.H.); yulu_zhang@stu.cqu.edu.cn (Y.Z.); yunjian@cqu.edu.cn (Y.J.)

* Correspondence: wanli_wen@cqu.edu.cn

Abstract: Wireless Sensor Networks (WSNs) have emerged as an efficient solution for numerous real-time applications, attributable to their compactness, cost-effectiveness, and ease of deployment. The rapid advancement of 5G technology and mobile edge computing (MEC) in recent years has catalyzed the transition towards large-scale deployment of WSN devices. However, the resulting data proliferation and the dynamics of communication environments introduce new challenges for WSN communication: (1) ensuring robust communication in adverse environments and (2) effectively alleviating bandwidth pressure from massive data transmission. In response to the aforementioned challenges, this paper proposes a semantic communication solution. Specifically, considering the limited computational and storage resources of WSN devices, we propose a flexible Attention-based Adaptive Coding (AAC) module. This module integrates window and channel attention mechanisms, dynamically adjusts semantic information in response to the current channel state, and facilitates adaptation of a single model across various Signal-to-Noise Ratio (SNR) environments. Furthermore, to validate the effectiveness of this approach, the paper introduces an end-to-end Joint Source Channel Coding (JSCC) scheme for image semantic communication, employing the AAC module. Experimental results demonstrate that the proposed scheme surpasses existing deep JSCC schemes across datasets of varying resolutions; furthermore, they validate the efficacy of the proposed AAC module, which is capable of dynamically adjusting critical information according to the current channel state. This enables the model to be trained over a range of SNRs and obtain better results.

Keywords: mobile edge computing; wireless sensor networks; semantic communications; attention mechanism; joint source channel coding; deep neural network

Citation: Liu, X.; Huang, Z.; Zhang, Y.; Jia, Y.; Wen, W. CNN and Attention-Based Joint Source Channel Coding for Semantic Communications in WSNs. *Sensors* **2024**, *24*, 957. https://doi.org/10.3390/s24030957

Academic Editor: Giovanni Pau

Received: 25 December 2023
Revised: 24 January 2024
Accepted: 28 January 2024
Published: 1 February 2024

1. Introduction

Wireless Sensor Networks (WSNs) [1,2] have emerged as highly effective solutions for a multitude of real-time applications owing to their compactness, cost-effectiveness, and ease of deployment. In recent years, the rapid development of 5G technology and mobile edge computing (MEC) [3,4] has facilitated the gradual transition of WSN devices towards large-scale deployment [5]. For instance, the deployment of numerous sensor nodes in urban areas enables real-time monitoring of environmental conditions, thereby offering crucial insights for urban planning and management. However, the resulting proliferation of data and the complexity of evolving communication environments introduce new challenges for WSN communications: (1) ensuring robust communication in poor Signal-to-Noise Ratio (SNR) conditions; (2) effectively alleviating bandwidth pressure amidst massive data transmission, particularly in high-density WSN scenarios.

Current WSN communication strategies employ traditional separated source channel coding methods, prioritizing accuracy and high fidelity at the physical layer. This often necessitates transmitting complete raw data, irrespective of content relevance, resulting in inefficiencies. Furthermore, this conventional approach is susceptible to the "cliff effect",

leading to communication failures in challenging conditions. This makes it difficult to fulfill high-precision task requirements in low SNR environments, a situation exacerbated in high-density WSNs. Deep Learning (DL)-based semantic communication offers a novel approach by employing Joint Source Channel Coding (JSCC) through Deep Neural Networks (DNNs) [6,7], focusing on transmitting semantic information rather than all data, thereby demonstrating potential in reducing data volume and enhancing communication reliability.

Recent preliminary research in the field of semantic communication has demonstrated its potential to effectively address the two aforementioned challenges in WSNs. Bourtsoulatze et al. [8] developed a JSCC scheme utilizing convolutional neural networks for wireless image semantic communication (ISC) that outperformed advanced separate source channel coding methods (JPEG+LDPC, JPEG2000+LDPC), particularly in low SNR environments. They also discovered that deep JSCC is immune to the "cliff effect". To mitigate channel distortion from noise, certain approaches [6,7,9] have adopted generalized divisive normalization (GDN) and incorporated feedback mechanisms to enhance performance. However, these methods are limited to training at a fixed SNR and exhibit suboptimal performance in varying SNR conditions. Consequently, ref. [10] introduced a design featuring a single JSCC-encoder paired with multiple JSCC-decoders, where the decoder selection is contingent on the channel SNR. This design facilitates optimal model performance across a spectrum of SNRs, albeit at the expense of significant computational and storage demands, hindering its large-scale applicability in WSN devices. Therefore, the development of a single model adaptable to a wide range of SNRs is essential.

The advent of attentional mechanisms, particularly their successful application in visual tasks, offers a novel direction to tackle the aforementioned challenges. This approach [11] enhances feature learning in pivotal regions by emulating attention allocation processes observed in biological vision, concurrently suppressing the interference from non-essential information. Current research [12–16] indicates that while non-local attention adaptively adjusts feature representation and enhances model performance, it concurrently incurs considerable computational overhead. Conversely, the window attention mechanism [17] offers an efficient alternative, applying attention within a confined scope, thereby diminishing the model's computational demands. Nevertheless, how to design a single model that can adapt to a wide range of SNR conditions remains an open research question, which will be explored in depth in this paper.

In this paper, we aim to address the problem of poor performance of a single model in semantic communication under different SNR conditions. Specifically, we propose an Attention-based Adaptive Coding (AAC) module for semantic communication and design a novel JSCC scheme based on this. The principal contributions of this paper are summarized as follows:

- We propose a flexible AAC module. Considering the resource-limited nature of WSN devices, it is able to capture the correlation between spatial neighboring elements to dynamically weight key local semantic information without sacrificing too many computational resources and is able to dynamically adjust the model output based on the current channel state information, which is capable of adapting/training a single model for a wide range of SNRs.
- We propose a novel JSCC model based on AAC modules and CNNs for ISC. Experimental results show that our model is more robust than the baseline model when compared to the current state-of-the-art methods, even in the case of channel mismatch.

Figure 1 presents a detailed overview of an ISC system, comprising a transmitter and a receiver. The transmitter extracts semantic information from the input image using the semantic encoder. To guarantee the validity of the information, it is forwarded to the channel encoder prior to transmission. The encoded information is sent via wireless channels after power normalization, such as additive white Gaussian noise (AWGN) channels or fading channels to the receiver. Upon receiving the information, the receiver processes it sequentially through the channel decoder and the semantic decoder to reconstruct the image. In this paper, DNNs are employed to collaboratively design a semantic encoder/decoder

and channel encoder/decoder for the ISC system. Furthermore, the wireless channel is conceptualized as a non-updatable layer, facilitating end-to-end optimization from the transmitter to the receiver.

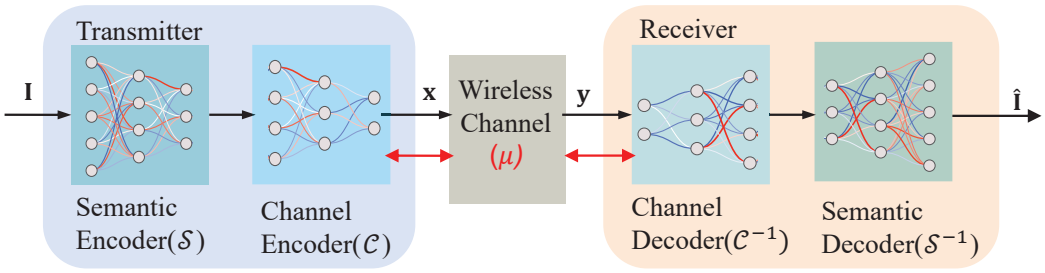


Figure 1. System model.

2. System Model

Specifically, the input image of the transmitter is denoted as $\mathbf{I} \in \mathbb{R}^{H\times W\times C}$, where $\mathbb{R}$ signifies the set of real numbers, while H, W, and C represent the height, width, and color channels of the image, respectively. During the transmission stage, $\mathbf{I}$ is sequentially mapped by the semantic encoder $\mathcal{S}$ and the channel encoder $\mathcal{C}$ into an L-dimensional complex vector $\mathbf{x} \in \mathbb{C}^L$, where $\mathbb{C}$ denotes the set of complex numbers. In this context, $\mathcal{S}$ efficiently extracts semantic information from $\mathbf{I}$, while $\mathcal{C}$ dynamically enhances this information in response to the current channel state (e.g., SNR), mitigating the adverse effects of the wireless channel. It should be noted that in this paper, we assume both communicating sides have the knowledge of the wireless channel's SNR, denoted as μ. The bandwidth compression ratio, labeled as R, can be calculated as $R = \frac{L}{HWC} \in (0,1)$, where a smaller R indicates more compression. The above encoding process can be mathematically expressed as $\mathbf{x} = \mathcal{C}(\mathcal{S}(\mathbf{I}), \mu)$.

The wireless channel transmits the vector $\mathbf{x}$, resulting in an output vector denoted as $\mathbf{y}$, i.e.,

$$\mathbf{y} = \begin{cases} \mathbf{x} + z, & \text{for AWGN channel,} \\ h\mathbf{x} + z, & \text{for fading channel,} \end{cases} \tag{1}$$

where $h \in \mathbb{C}$ represents the channel gain, which is assumed to be a circularly symmetric complex Gaussian random variable with zero mean and unit variance, i.e., $h \sim \mathcal{CN}(0,1)$, and $z \sim \mathcal{CN}(0,\sigma_{\mathrm{n}}^2)$ is the complex Gaussian noise with zero mean and variance σ_{n}^2. Based on (1), μ can be calculated as

$$\mu = \begin{cases} 10\log_{10}\left(\frac{P}{\sigma_{\mathrm{n}}^2}\right), & \text{for AWGN channel,} \\ 10\log_{10}\left(\frac{P|h|^2}{\sigma_{\mathrm{n}}^2}\right), & \text{for fading channel,} \end{cases} \tag{2}$$

where P is the transmit power of the transmitter. During the receiving stage, the receiver is equipped with a semantic decoder and a channel decoder, which are labeled as $\mathcal{S}^{-1}$ and $\mathcal{C}^{-1}$, respectively. Using the knowledge of μ, $\mathcal{S}^{-1}$ and $\mathcal{C}^{-1}$ decode $\mathbf{y}$ to a reconstructed image, denoted as $\hat{\mathbf{I}} \in \mathbb{R}^{H\times W\times C}$. Such a process can be mathematically expressed as $\hat{\mathbf{I}} = \mathcal{S}^{-1}(\mathcal{C}^{-1}(\mathbf{y}, \mu))$.

Based on the above formulations, we define the overall DNNs as $\mathcal{N} \triangleq \{\mathcal{S}, \mathcal{C}, \mathcal{C}^{-1}, \mathcal{S}^{-1}\}$. In this paper, we would like to train $\mathcal{N}$ to achieve a JSCC scheme for the ISC system. To evaluate the performance of the proposed JSCC scheme, we adopt the following two distortion metrics. The first is the PSNR metric [18] (in dB), defined as

$$\mathrm{PSNR}(\mathbf{I}, \hat{\mathbf{I}}) = \log_{10}\left(\frac{A^2}{||\mathbf{I} - \hat{\mathbf{I}}||_2^2}\right), \tag{3}$$

where A is the maximum possible value for a given pixel and $\|\cdot\|_2$ is the l_2-norm operator. The other is the SSIM metric [19], defined as

$$\mathrm{SSIM}(\mathbf{I}, \hat{\mathbf{I}}) = \left(\frac{2\mu_{\mathbf{I}}\mu_{\hat{\mathbf{I}}} + v_1}{\mu_{\mathbf{I}}^2 + \mu_{\hat{\mathbf{I}}}^2 + v_1}\right)\left(\frac{2\sigma_{\mathbf{I}}\sigma_{\hat{\mathbf{I}}} + v_2}{\sigma_{\mathbf{I}}^2 + \sigma_{\hat{\mathbf{I}}}^2 + v_2}\right), \tag{4}$$

where v_1 and v_2 are coefficients for numeric stability, and $\mu_{\mathbf{I}}$ (resp. $\mu_{\hat{\mathbf{I}}}$) and $\sigma_{\mathbf{I}}^2$ (resp. $\sigma_{\hat{\mathbf{I}}}^2$) are the mean and variance of $\mathbf{I}$ (resp. $\hat{\mathbf{I}}$), respectively.

3. The Proposed JSCC Scheme

This section is dedicated to the development of the JSCC scheme for the ISC system. The comprehensive neural network architecture, represented by $\mathcal{N}$, is illustrated in Figure 2. Initially, the semantic encoder neural network $\mathcal{S}$, along with its corresponding decoder $\mathcal{S}^{-1}$, is designed to extract and subsequently recover semantic information from the original image $\mathbf{I}$. Subsequently, the channel encoder $\mathcal{C}$ and its decoder $\mathcal{C}^{-1}$ are developed to mitigate the adverse effects typical of wireless channels. Benchmarking system performance necessitates the power normalization of encoded data before their transmission through the wireless channel. Ultimately, the comprehensive neural network $\mathcal{N}$ is realized through the integration of neural networks $\mathcal{S}$, $\mathcal{S}^{-1}$, $\mathcal{C}$, and $\mathcal{C}^{-1}$.

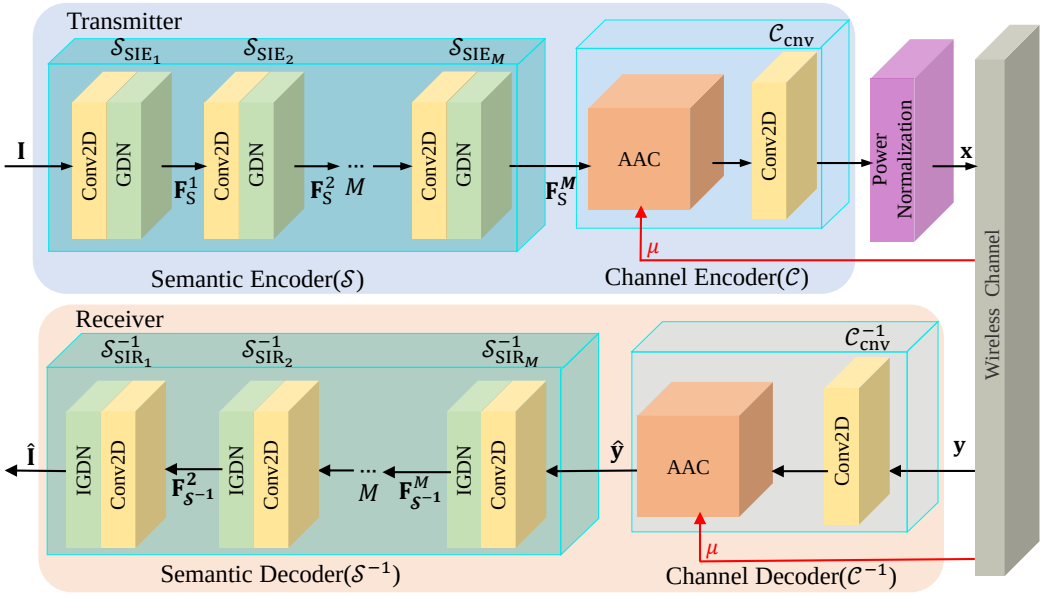


Figure 2. The neural network structure of the proposed JSCC scheme.

3.1. The Design of $\mathcal{S}$ and $\mathcal{S}^{-1}$

For the effective extraction of image semantic information, Convolutional Neural Networks (CNNs) are utilized in the design of $\mathcal{S}$ and $\mathcal{S}^{-1}$. As depicted in Figure 2, $\mathcal{S}$ comprises M Semantic Information Extraction (SIE) layers, denoted as $\mathcal{S}_{\mathrm{SIE}}$. Each $\mathcal{S}_{\mathrm{SIE}}$ layer encompasses a 2D convolutional layer, a Generalized Division Normalization (GDN) layer, and an activation function. The local perceptual capabilities of the convolutional layer enable the efficient extraction and recovery of semantic information such as colors, textures, shapes, and image contents. The nonlinear manipulation afforded by the GDN layer facilitates the extraction of more complex semantic information from the input image, ensuring spatially adaptive normalization. This aspect is pivotal in reducing spatial redundancy and more effectively capturing the image's spatial structural information. The sequential

layering of the SIE modules effectively extracts the image's semantic information. The semantic information output by the semantic encoder, denoted as $\mathbf{F}_{\mathcal{S}}^{M}$, can be represented as

$$\mathbf{F}_{\mathcal{S}}^{M} = \mathcal{S}_{\mathrm{SIE_M}}(\mathbf{F}_{\mathcal{S}}^{M-1}), \tag{5}$$

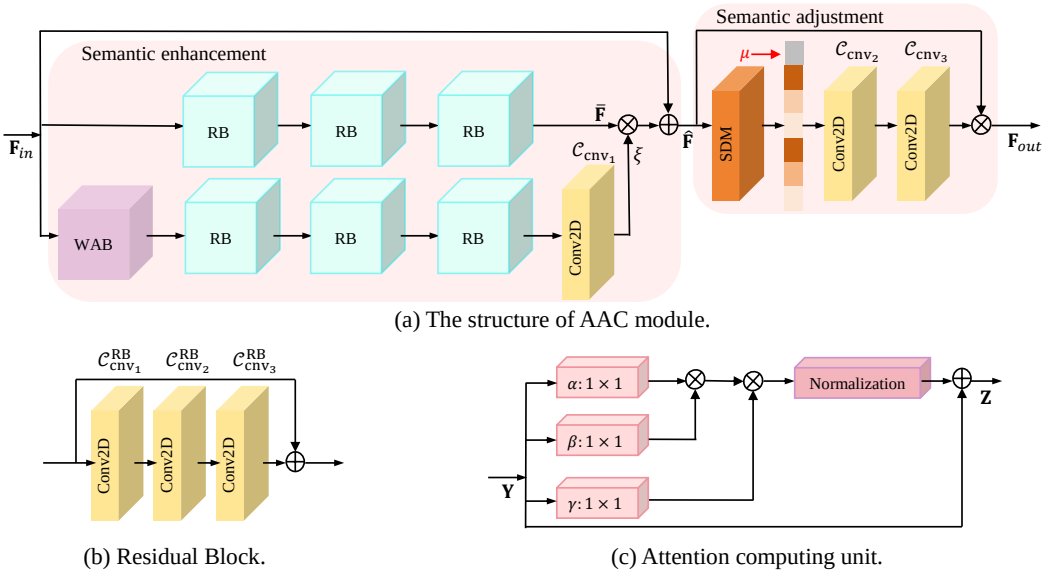
when $M = 1$, $\mathbf{F}_{\mathcal{S}}^{1} = \mathcal{S}_{\mathrm{SIE_1}}(\mathbf{I})$. Analogous to $\mathcal{S}$, $\mathcal{S}^{-1}$ comprises M Semantic Information Reconstruction (SIR) layers, denoted as $\mathcal{S}_{\mathrm{SIR}}^{-1}$. Each $\mathcal{S}_{\mathrm{SIR}}^{-1}$ layer consists of a 2D deconvolution, an Inverse Generalized Division Normalization (IGDN) layer, and an activation function. $\mathcal{S}^{-1}$ and $\mathcal{S}_{\mathrm{SIR}}^{-1}$ represent the inverse processes of $\mathcal{S}$ and $\mathcal{S}_{\mathrm{SIE}}$, respectively. In our design, the configuration of $\mathcal{S}_{\mathrm{SIE}}$ and $\mathcal{S}_{\mathrm{SIR}}^{-1}$ layers is symmetrical. The process of semantic decoding is formulated as

$$\mathbf{F}_{\mathcal{S}^{-1}}^{m} = \mathcal{S}_{\mathrm{SIR_m}}^{-1}(\mathbf{F}_{\mathcal{S}^{-1}}^{m+1}), m = 1, 2, \ldots, M. \tag{6}$$

Here, $\mathbf{F}_{\mathcal{S}^{-1}}^{m}$ denotes the output from the m-th Semantic Information Reconstruction (SIR) layer within $\mathcal{S}^{-1}$, with the output from the final SIR layer being the reconstructed image $\mathbf{I}$. When $m = M$, the output $\mathbf{F}_{\mathcal{S}^{-1}}^{M}$ equates to $\mathcal{S}_{\mathrm{SIR_M}}^{-1}(\hat{\mathbf{y}})$.

3.2. The Design of $\mathcal{C}$ and $\mathcal{C}^{-1}$

The channel encoder $\mathcal{C}$ comprises an AAC block and a 2D convolutional layer (labeled as $\mathcal{C}_{\mathrm{cnv}}$). Similarly, the channel decoder $\mathcal{C}^{-1}$ comprises an AAC block and a 2D deconvolution layer (labeled as $\mathcal{C}_{\mathrm{cnv}}^{-1}$). Here, AAC is able to dynamically adjust the semantic information of the model output (resp. reconstruction) according to the current channel state to improve the quality of coding (resp. decoding), and $\mathcal{C}_{\mathrm{cnv}}$ (resp. $\mathcal{C}_{\mathrm{cnv}}^{-1}$) is used to adjust the transmitted (resp. received) information according to the bandwidth compression ratio. The AACs within $\mathcal{C}$ and $\mathcal{C}^{-1}$ possess identical structures yet differ in their parameters. The architectural configuration of the AAC block is depicted in Figure 3a. This block comprises two components—semantic enhancement and semantic adjustment—both of which will be elucidated in detail.



(a) The structure of AAC module.

(b) Residual Block.

(c) Attention computing unit.

Figure 3. AAC module.

Semantic enhancement: The semantic enhancement component consists of a Window Attention Block (WAB) and multiple Residual Blocks (RBs). In our architecture, the WAB is initially employed to empower the neural network to prioritize elements crucial to the current task, achieved by weighting key areas within the input feature map. This approach allows the neural network to focus its resources on processing specific regions of the image, like particular textures or edges, in more detail rather than treating the entire

image uniformly. Such a mechanism enhances feature representation, particularly in scenes characterized by rich content or varied detail. Subsequently, the RBs assist the neural network in focusing on regions requiring improvement, as they are optimized for addressing reconstruction discrepancies during training. The detailed process is described below.

First of all, the input information of the AAC block, which is marked as $\mathbf{F}_{in}$, passes through three RBs of the same structure, generating statistical information $\overline{\mathbf{F}}$. As shown in Figure 3b, each RB consists of three convolutional layers (labeled as $\mathcal{C}_{\mathrm{cnv}_1}^{\mathrm{RB}}$, $\mathcal{C}_{\mathrm{cnv}_2}^{\mathrm{RB}}$, and $\mathcal{C}_{\mathrm{cnv}_3}^{\mathrm{RB}}$), which help to improve the stability of GDN. The connection of multiple RBs can enhance the features of the input image to combat channel noise. At the same time, to generate the weight factor ξ, a WAB, three RBs, and a convolutional layer denoted as $\mathcal{C}_{\mathrm{cnv}_1}$ are utilized. WAB is used to focus on high-contrast regions. Specifically, the feature map of the image is divided into windows of $Q \times Q$ in a non-overlapping way; then, it uses multi-head attention with a head number of t to compute the attention map in each window separately, where the attention computation unit is shown in Figure 3c. Suppose Y_i^q and Y_j^q are the pixels in the i-th row and j-th column of the q-th window. The i-th row output of the q-th window, which is labeled as Z_i^q, can be formulated as

$$Z_i^q = \frac{W_z \sum_{\forall j} W_\gamma Y_i^q e^{W_\alpha W_\beta Y^T Y}}{\sum_{\forall j} e^{W_\alpha W_\beta Y^T Y}} + Y_i^q, \tag{7}$$

where W_α and W_β are cross-channel transforms, and W_γ and W_z are weight matrices. In order to improve the learning ability of the model, three RBs are used after WAB; then, a convolutional layer, followed by a Sigmoid function, is used, generating factor $\xi \in [0, 1]$. Finally, by weighting and residual operations, $\hat{\mathbf{F}} \in \mathbb{R}^{H_\mathrm{s} \times W_\mathrm{s} \times C_\mathrm{s}}$ is obtained, which has clearer semantic information. Here, W_s, H_s, and C_s denote the semantically enhanced feature map width, height, and number of feature channels, respectively. The process can be expressed as $\hat{\mathbf{F}} = \mathbf{F}_{in} + \xi \overline{\mathbf{F}}$.

Semantic adjustment: In order to effectively reduce the detrimental effects of the channel on the model and to improve the robustness (i.e., communication reliability) of the model in terms of the SNR over the range, we improve the channel attention mechanism to be able to dynamically adjust the more attended information based on the channel state information (i.e., SNR) of the changing wireless channel. Specifically, the sum of standard deviation and mean $SDM(\cdot)$ is first used to capture the global information within $\hat{\mathbf{F}}$. Compared with the global average pooling operation, $SDM(\cdot)$ is able to better preserve information about relevant structures, textures, and edges, which are highly beneficial for enhancing image details (related to SSIM). For input vector $\hat{\mathbf{F}} = [\hat{\mathbf{F}}_1, \ldots, \hat{\mathbf{F}}_k, \ldots, \hat{\mathbf{F}}_{C_\mathrm{s}}]$, the output of the k-th feature channel (defined as z_k) after $SDM(\cdot)$ can be expressed as

$$z_k = SDM(\hat{\mathbf{F}}_k) = \frac{1}{H_\mathrm{s} W_\mathrm{s}} \sum_{\lambda=1}^{H_\mathrm{s}} \sum_{\nu=1}^{W_\mathrm{s}} \hat{\mathbf{F}}_k^{\lambda,\nu} + \sqrt{\frac{1}{H_\mathrm{s} W_\mathrm{s}} \sum_{\lambda=1}^{H_\mathrm{s}} \sum_{\nu=1}^{W_\mathrm{s}} (\hat{\mathbf{F}}_k^{\lambda,\nu} - \frac{1}{H_\mathrm{s} W_\mathrm{s}} \sum_{\lambda=1}^{H_\mathrm{s}} \sum_{\nu=1}^{W_\mathrm{s}} \hat{\mathbf{F}}_k^{\lambda,\nu})^2}. \tag{8}$$

After that, the obtained global information is connected with μ along the feature channel dimension, and two convolutional layers (denoted as $\mathcal{C}_{\mathrm{cnv}_2}$ and $\mathcal{C}_{\mathrm{cnv}_3}$) are used to predict the weighting factors; finally, the adjusted output is obtained by weighting $\hat{\mathbf{F}}$, which can be expressed as

$$\mathbf{F}_{out} = \hat{\mathbf{F}} \mathcal{C}_{\mathrm{cnv}_3}(\mathcal{C}_{\mathrm{cnv}_2}(Concat(z_1, \ldots, z_k, \ldots, z_{C_\mathrm{s}}, \mu))). \tag{9}$$

Here, $Concat(\cdot)$ denotes the concatenation operation, $\mathcal{C}_{\mathrm{cnv}_2}$ uses the ReLu activation function to learn the nonlinear relationship, and $\mathcal{C}_{\mathrm{cnv}_3}$ uses the Sigmoid activation function to ensure that the weight factor is between 0 and 1.

3.3. The Training Algorithm

We employ mean square error (MSE) to measure the difference between the original image $\mathbf{I}$ and the reconstructed image $\hat{\mathbf{I}}$. Therefore, the loss function is given by

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} d(\mathbf{I}, \hat{\mathbf{I}}). \tag{10}$$

Here, N is the number of samples, $d(\mathbf{I}, \hat{\mathbf{I}}) = \frac{1}{n}\|\mathbf{I} - \hat{\mathbf{I}}\|^2$ is the mean squared-error distortion, and n represents the total number of pixels in the image. Despite the presence of noise and interference, the entire DNNs $\mathcal{N}$ can effectively learn and recover the transmitted information by minimizing the loss function (10).

Algorithm 1 describes the training process for the neural network $\mathcal{N}$. The first step involves initializing the neural network parameters for $\mathcal{N}$. Following this, the image $\mathbf{I}$ is input into the semantic encoder $\mathcal{S}$, which produces its semantic information $\mathbf{F}_{\mathrm{S}}^{M}$. Subsequently, the channel encoder $\mathcal{C}$ and the power normalization operation transform $\mathbf{F}_{\mathrm{S}}^{M}$ into $\mathbf{x}$, which will be transmitted through the wireless channel. Upon receiving the compressed information $\mathbf{y}$ from the wireless channel, the channel decoder $\mathcal{C}^{-1}$ outputs the information $\hat{\mathbf{y}}$. Based on this, the semantic decoding neural network $\mathcal{S}^{-1}$ reconstructs the transmitted image $\hat{\mathbf{I}}$. Finally, we employ the Adam optimizer [20] to update the parameters of $\mathcal{N}$.

Algorithm 1 Training Algorithm for $\mathcal{N}$

Input: The original image $\mathbf{I}$ and SNR μ.
Output: The neural network $\mathcal{N}$.
1: Initialize the parameters in $\mathcal{N}$.
2: **Transmitter:**
Perform semantic encoding process: $\mathcal{S}(\mathbf{I}) \rightarrow \mathbf{F}_{\mathrm{S}}^{M}$.
Perform channel encoding and power normalization process: $\mathcal{C}(\mathbf{F}_{\mathrm{S}}^{M}, \mu) \rightarrow \mathbf{x}$.
3: Transmit $\mathbf{x}$ over the wireless channel to obtain $\mathbf{y}$ in (1).
4: **Receiver:**
Perform channel decoding process: $\mathcal{C}^{-1}(\mathbf{y}, \mu) \rightarrow \hat{\mathbf{y}}$.
Perform semantic decoding process: $\mathcal{S}^{-1}(\hat{\mathbf{y}}) \rightarrow \hat{\mathbf{I}}$.
5: Compute the loss function in (10).
6: Train the neural network $\{\mathcal{S}, \mathcal{C}, \mathcal{C}^{-1}, \mathcal{S}^{-1}\}$ using Adam optimizer.

4. Simulation Results

In this section, we will give the specific parameter settings. Following that, we will assess the performance of the proposed JSCC scheme, as well as representative baselines, through the examination of simulation results.

4.1. Simulation Settings

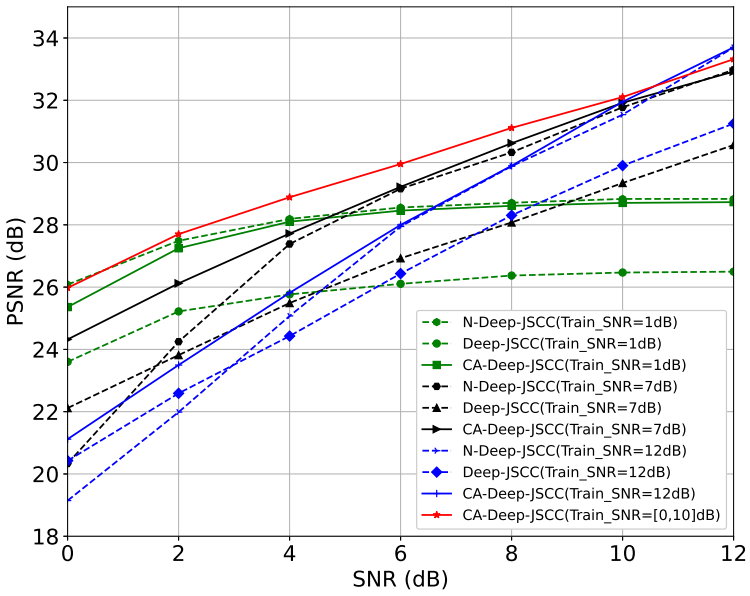
$\mathcal{N}$ consists of four SIE and four SIR blocks (i.e., $M = 4$). Table 1 details the structural parameters of $\mathcal{N}$—*input*, *output*, *k_size*, *stride*, and *Activation*, denoted as input dimensions, output dimensions, convolution kernel sizes, step sizes, and activation functions, respectively. As an additional note, the size of $\mathcal{U}$ is determined by the bandwidth compression ratio calculation [8], and the parameter settings are the same for all RBs in $\mathcal{C}$ and $\mathcal{C}^{-1}$. The proposed JSCC scheme along with two benchmark JSCC schemes [9,21], designated as CA-Deep-JSCC, Deep-JSCC, and N-Deep-JSCC, respectively, were trained and tested across varying bandwidth compression ratios ($\frac{1}{6}$ and $\frac{1}{12}$), employing datasets of different resolutions (CIFAR-10 [22], ImageNet2012 [23], and Kodak [24]). PSNR and SSIM were used as evaluation criteria to react to the reconstructed image quality (i.e., communication reliability). All experiments were conducted on a PC with an Intel Core i7-10700 CPU@2.90 GHz and an NVIDIA RTX A4000 GPU.

Table 1. The parameters of $\mathcal{N}$.

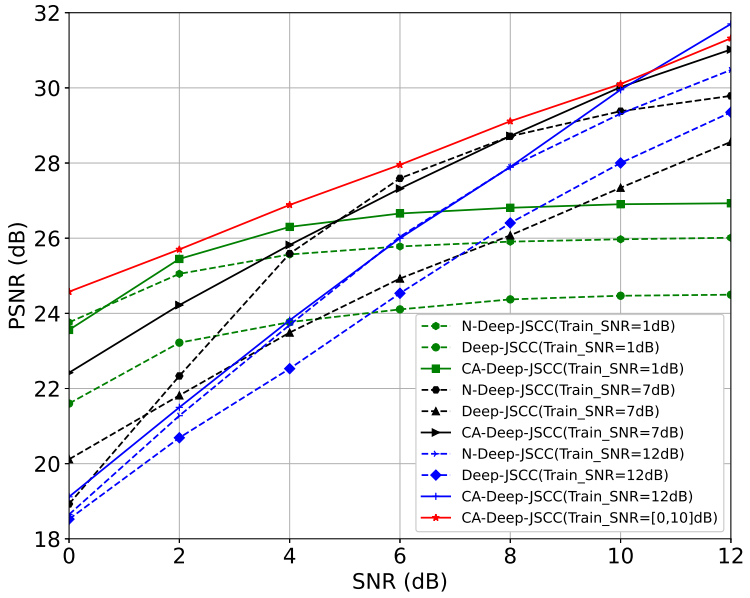
Module	Layer		Input	Output	K_Size	Stride	Activation
$\mathcal{S}(\mathcal{S}^{-1})$	$\mathcal{S}_{\text{SIE}_1}(\mathcal{S}^{-1}_{\text{SIR}_1})$		3(256)	256(3)	9(9)	2(2)	PReLU(PReLU)
	$\mathcal{S}_{\text{SIE}_2}(\mathcal{S}^{-1}_{\text{SIR}_2})$		256(256)	256(256)	5(5)	2(2)	PReLU(PReLU)
	$\mathcal{S}_{\text{SIE}_3}(\mathcal{S}^{-1}_{\text{SIR}_3})$		256(256)	256(256)	5(5)	1(1)	PReLU(PReLU)
	$\mathcal{S}_{\text{SIE}_4}(\mathcal{S}^{-1}_{\text{SIR}_4})$		256(256)	256(256)	5(5)	1(1)	PReLU(PReLU)
$\mathcal{C}(\mathcal{C}^{-1})$	$\mathcal{C}_{\text{cnv}}(\mathcal{C}^{-1}_{\text{cnv}})$		256($\mathcal{U}$)	$\mathcal{U}$(256)	3	1	PReLU(PReLU)
	AAC	$\mathcal{C}^{\text{RB}}_{\text{cnv}_1}$	256	128	1	1	GELU
		$\mathcal{C}^{\text{RB}}_{\text{cnv}_2}$	128	128	3	1	GELU
		$\mathcal{C}^{\text{RB}}_{\text{cnv}_3}$	128	256	1	1	None
		$\mathcal{C}_{\text{cnv}_1}$	256	256	1	1	Sigmoid
		$\mathcal{C}_{\text{cnv}_2}$	257	256	1	1	ReLU
		$\mathcal{C}_{\text{cnv}_3}$	256	256	1	1	Sigmoid
		WAB	$\mathcal{Q} = 4, t = 8$				

4.2. Performance Evaluation

CIFAR-10 Performance Evaluation: We use the CIFAR-10 dataset for training and testing, which contains 50,000 training images and 10,000 test images, and the image sizes are all 32×32. We set the batch_size to 128; the learning rate to 10^{-4}; and use the Adam optimizer to train the models with SNRs of 1, 7, and 12, respectively, under bandwidth compression ratios of 1/6 and 1/12 for the model. Figure 4 shows the PSNR performance of CA-Deep-JSCC, Deep-JSCC, and N-Deep-JSCC for $R = \frac{1}{6}$ and $R = \frac{1}{12}$, where CA-Deep-JSCC is represented by the solid line and baselines by the dashed line. The red solid line shows the results of training on SNRs ranging from 0 to 10 intervals of 2. These results can be summarized as follows. Firstly, the PSNR of all schemes generally increases with the rising test SNR, attributable to the enhanced image reconstruction quality concurrent with signal quality improvement. Secondly, across all training instances with identical SNR and R values, the PSNR of the CA-Deep-JSCC scheme consistently exceeds that of the two baseline models, notably in the low SNR range. This finding indicates greater robustness of CA-Deep-JSCC under varied channel conditions. This is attributed to the fact that the proposed AAC module employs a window focusing mechanism, which enhances the semantic information that is focused on more. Furthermore, all schemes generally exhibit higher PSNR values at a compression ratio of $R = \frac{1}{6}$ compared to $R = \frac{1}{12}$, implying that higher compression ratios (or higher bandwidths) facilitate the transmission of more information through the channel, thereby improving image reconstruction quality. Finally, when trained within the SNR [0, 10] dB range, the proposed JSCC scheme demonstrates outstanding performance across all tested SNR levels. The superior performance is ascribed to the proposed AAC module's capability to dynamically adjust semantic features in sync with real-time SNR, enabling CA-Deep-JSCC to be specifically trained within a certain SNR range, thereby markedly bolstering its robustness across different SNR scenarios.



(**a**)

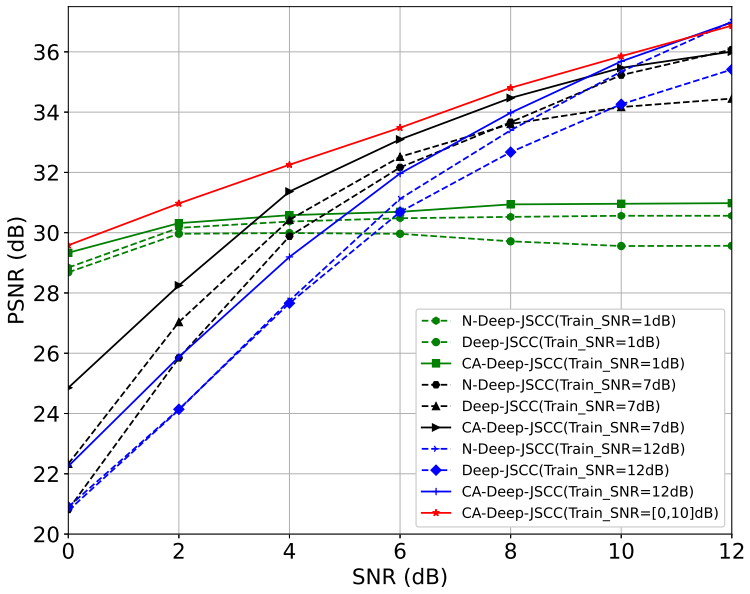


(**b**)

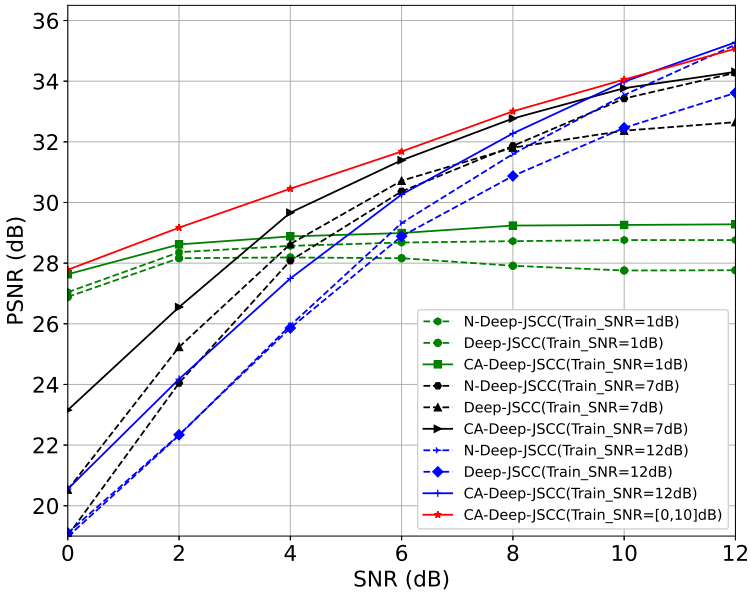
Figure 4. Model performance on CIFAR-10. Each curve of Deep-JSCC and N-Deep-JSCC was trained at a specific SNR. The curves of CA-Deep-JSCC were trained at 1 dB, 7 dB, and 12 dB with 0–10 intervals of 2 SNR. (**a**) Reconstruction distortion of Deep-JSCC, N-Deep-JSCC, and CA-Deep-JSCC on CIFAR-10, $R = \frac{1}{6}$. (**b**) Reconstruction distortion of Deep-JSCC, N-Deep-JSCC, and CA-Deep-JSCC on CIFAR-10, $R = \frac{1}{12}$.

Kodak Performance Evaluation: To ensure the validity of higher resolution image data, we use the ImageNet2012 training set to train the model and evaluate it on Kodak, where the Kodak dataset contains 24 images of 512×768. The ImageNet2012 training set contains 1.3 million images of various resolutions; we filter the images whose size is larger than 128×128 (about 1.25 million images), which are then cropped to a patch of 128×128; set the batch_size to 128 and the learning rate to 10^{-4}; and use the Adam optimizer for training. Similar to CIFAR-10, we train both models at $R = \frac{1}{6}$ and $R = \frac{1}{12}$, respectively, with fixed SNRs of 1, 7, and 12, and train CA-Deep-JSCC on a range of SNRs of 2 at intervals of 0 to 10. Figure 5 shows the reconstruction quality of both images on Kodak. First, the CA-Deep-JSCC scheme demonstrates a higher PSNR than the Deep-JSCC and N-Deep-JSCC

schemes at nearly all SNR points under both bandwidth compression ratios, suggesting that the proposed scheme may be more effective in feature extraction and information transfer. Secondly, with a training SNR range of $[0, 10]$ dB, the proposed JSCC scheme exhibits the best performance at all test SNR points, which verifies the effectiveness of the AAC module on images of different resolutions. Finally, compared to the performance on the CIFAR-10 dataset, the model trained on the ImageNet2012 dataset demonstrates a higher PSNR on the Kodak dataset. This enhanced performance can be attributed to the following: (1) the diversity and complexity of the ImageNet2012 dataset, enabling the model to learn a broader feature representation; (2) the larger feature map size providing the model with more detailed information, which is important in ISC.



(**a**)



(**b**)

Figure 5. Model performance on Kodak. Each curve of Deep-JSCC and N-Deep-JSCC was trained at a specific SNR. The curves of CA-Deep-JSCC were trained at 1 dB, 7 dB, and 12 dB with 0–10 intervals of 2 SNR. (**a**) Reconstruction distortion of Deep-JSCC, N-Deep-JSCC, and CA-Deep-JSCC on Kodak, $R = \frac{1}{6}$. (**b**) Reconstruction distortion of Deep-JSCC, N-Deep-JSCC, and CA-Deep-JSCC on Kodak, $R = \frac{1}{12}$.

To facilitate an intuitive visual comparison, a comparative analysis is presented between the CA-Deep-JSCC model, the Deep-JSCC model, and the N-Deep-JSCC model using sample images from the Kodak dataset in Figure 6. It can be observed that the quality of the reconstructed images of both the schemes improves as the SNR increases, which is quantified by the increase in PSNR and SSIM. The CA-Deep-JSCC scheme exhibits higher values of PSNR and SSIM in all SNR conditions, which indicates its superiority in the task of semantic communication of images.

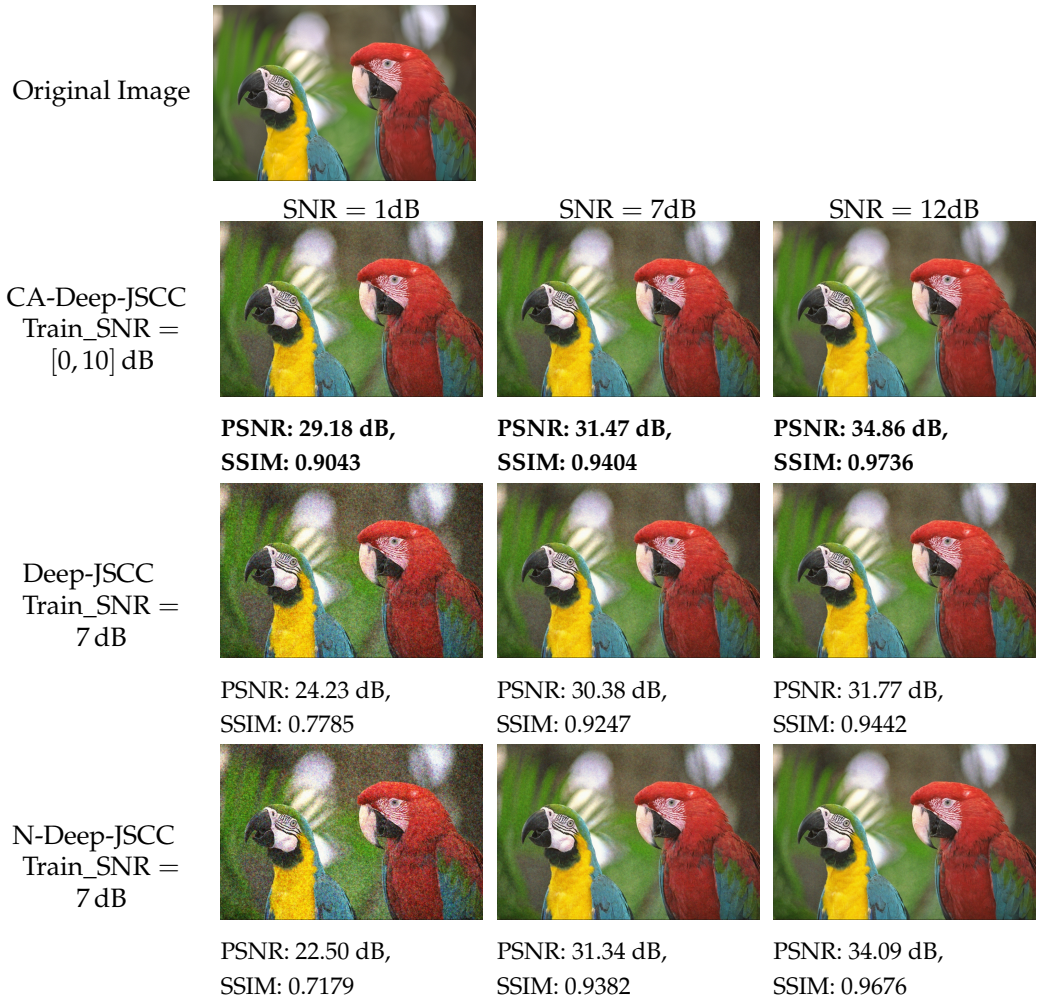


Figure 6. Visual comparison of three models (SNR = 1 dB, 7 dB, and 12 dB) for sample images of Kodak dataset under $R = \frac{1}{12}$.

5. Conclusions

This paper addresses the challenge of robustness in WSN communication through semantic communication solutions. We propose an AAC module for flexible integration into the ISC system, comprising a semantic enhancement component, utilizing the window attention mechanism, and a semantic adjustment component based on the channel attention mechanism. This module dynamically adjusts the focus on regions and semantic contents in response to the current channel state. Subsequently, in order to prove the effectiveness of AAC, a novel DNNs model, integrating CNN and AAC, is designed for ISC. Extensive simulations demonstrate that our proposed JSCC scheme surpasses existing state-of-the-art schemes in performance and can be trained across a range of SNRs, thereby enabling a single model to adapt to various SNR scenarios. This research advances the practical application of ISC schemes within the realm of WSN communication. Future research endeavors will focus on extending the JSCC scheme's design to additional modalities—including text, audio, and video—and explore semantic-level data transfer mechanisms as

well as model lightening techniques within WSNs with the objective of implementing a holistic semantic communication framework in WSNs.

Author Contributions: Conceptualization, Z.H. and W.W.; methodology, Z.H.; software, Y.Z.; validation, X.L., Y.Z., and Y.J.; formal analysis, X.L.; investigation, W.W.; resources, W.W.; data curation, Z.H.; writing—original draft preparation, Z.H.; writing—review and editing, W.W.; visualization, Y.Z.; supervision, W.W.; project administration, W.W.; funding acquisition, W.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under Grant 62201101; the Project funded by the China Postdoctoral Science Foundation under Grant 2022M720020; the Natural Science Foundation of Chongqing, China under Grant cstc2021jcyj-msxmX0458; and the Chongqing Technology Innovation and Application Development Special Key Project under Grant CSTB2022TIAD-KPX0059.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: CIFAR-10 dataset (https://www.cs.toronto.edu/~kriz/cifar.html, accessed on 12 July 2023); Kodak dataset (https://www.kaggle.com/datasets/sherylmehta/kodak-dataset, accessed on 16 October 2023); ImageNet2012 dataset (https://www.image-net.org/challenges/LSVRC/2012/, accessed on 16 October 2023).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Compton, M.; Henson, C.; Neuhaus, H.; Lefort, L.; Sheth, A. A Survey of the Semantic Specification of Sensors. In Proceedings of the 2nd International Workshop on Semantic Sensor Networks, at 8th International Semantic Web Conference, Chantilly, VA, USA, 25–29 October 2009; pp. 17–32. [CrossRef]
2. Akyildiz, I.; Su, W.; Sankarasubramaniam, Y.; Cayirci, E. A survey on sensor networks. *IEEE Commun. Mag.* **2002**, *40*, 102–114. [CrossRef]
3. Wen, W.; Cui, Y.; Zheng, F.C.; Jin, S.; Jiang, Y. Random caching based cooperative transmission in heterogeneous wireless networks. *IEEE Trans. Commun.* **2018**, *66*, 2809–2825. [CrossRef]
4. Wen, W.; Fu, Y.; Quek, T.Q.; Zheng, F.C.; Jin, S. Joint uplink/downlink sub-channel, bit and time allocation for multi-access edge computing. *IEEE Commun. Lett.* **2019**, *23*, 1811–1815. [CrossRef]
5. Yarinezhad, R.; Hashemi, S.N. A sensor deployment approach for target coverage problem in wireless sensor networks. *J. Ambient. Intell. Humaniz. Comput.* **2023**, *14*, 5941–5956. [CrossRef]
6. Wang, J.; Wang, S.; Dai, J.; Si, Z.; Zhou, D.; Niu, K. Perceptual learned source-channel coding for high-fidelity image semantic transmission. In Proceedings of the GLOBECOM 2022–2022 IEEE Global Communications Conference, Rio de Janeiro, Brazil, 4–8 December 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 3959–3964. [CrossRef]
7. Wang, S.; Dai, J.; Liang, Z.; Niu, K.; Si, Z.; Dong, C.; Qin, X.; Zhang, P. Wireless deep video semantic transmission. *IEEE J. Sel. Areas Commun.* **2022**, *41*, 214–229. [CrossRef]
8. Bourtsoulatze, E.; Burth Kurka, D.; Gündüz, D. Deep Joint Source-Channel Coding for Wireless Image Transmission. *IEEE Trans. Cogn. Commun. Netw.* **2019**, *5*, 567–579. [CrossRef]
9. Kurka, D.B.; Gündüz, D. DeepJSCC-f: Deep joint source-channel coding of images with feedback. *IEEE J. Sel. Areas Inf. Theory* **2020**, *1*, 178–193. [CrossRef]
10. Ding, M.; Li, J.; Ma, M.; Fan, X. SNR-adaptive deep joint source-channel coding for wireless image transmission. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1555–1559. [CrossRef]
11. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.U.; Polosukhin, I. Attention is All you Need. In *Proceedings of the Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
12. Jia, Y.; Huang, Z.; Luo, K.; Wen, W. Lightweight Joint Source-Channel Coding for Semantic Communications. *IEEE Commun. Lett.* **2023**, *27*, 3161–3165. [CrossRef]
13. Xie, H.; Qin, Z.; Li, G.Y.; Juang, B.H. Deep Learning Enabled Semantic Communication Systems. *IEEE Trans. Signal Process.* **2021**, *69*, 2663–2675. [CrossRef]
14. Yao, S.; Niu, K.; Wang, S.; Dai, J. Semantic coding for text transmission: An iterative design. *IEEE Trans. Cogn. Commun. Netw.* **2022**, *8*, 1594–1603. [CrossRef]
15. Weng, Z.; Qin, Z. Semantic communication systems for speech transmission. *IEEE J. Sel. Areas Commun.* **2021**, *39*, 2434–2444. [CrossRef]

16. Bian, C.; Shao, Y.; Gunduz, D. Wireless Point Cloud Transmission. *arXiv* **2023**, arXiv:2306.08730. [CrossRef]
17. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-Local Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018. [CrossRef]
18. Horé, A.; Ziou, D. Image Quality Metrics: PSNR vs. SSIM. In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 2366–2369. [CrossRef]
19. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]
20. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980. [CrossRef]
21. Huang, X.; Chen, X.; Chen, L.; Yin, H.; Wang, W. A Novel Convolutional Neural Network Architecture of Deep Joint Source-Channel Coding for Wireless Image Transmission. In Proceedings of the 2021 13th International Conference on Wireless Communications and Signal Processing (WCSP), Changsha, China, 20–22 October 2021; pp. 1–5. [CrossRef]
22. Krizhevsky, A.; Hinton, G. Learning Multiple Layers of Features from Tiny Images. Available online: http://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf (accessed on 18 January 2024).
23. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F.-F. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009 ; IEEE: Piscataway, NJ, USA , 2009; pp. 248–255. [CrossRef]
24. Franzen, R. Kodak Lossless True Color Image Suite. 1999; Volume 4, p. 9. Available online: http://r0k.us/graphics/kodak (accessed on 18 January 2024).