Edge Diffusion Block: Dynamic Model with Enhanced Resolution in Semantic Communication

Brian Estadimas Arfeto, Fadhel Hariz Dzulfikar, Shehbaz Tariq, Shin Hyundong
Department of Electronics and Information Convergence Engineering
Kyung Hee University, Yongin, South Korea
hshin@khu.ac.kr

***Abstract*—This paper proposes a semantic communication system for image transmission, harnessing the capabilities of diffusion models within the context of generative AI. The system comprises three core components: a transmitter-receiver system with JSCC for image transmission, a channel layer simulating realistic noise and compression challenges, and an image reconstruction block that incorporates a lightweight edge diffusion model with zero-shot capabilities, enabling the system to restore high-fidelity images closely aligned with visual semantics.**

***Index Terms*—semantic communication (SC), diffusion model, zero-shot dynamic architecture**

I. Introduction

As we approach the limits defined by Shannon's theory, traditional methods for improving transmission rates, such as increasing power, expanding bandwidth, and adding antennas, are becoming less effective. This is mainly due to spectrum limitations and high power consumption, as discussed in Luo et al.'s work [1]. Consequently, there is a growing need to explore the semantic domain to enhance communication efficiency significantly. Semantic communication involves representing source information at a semantic level, retaining only the task-relevant information. This compact representation can then be used for various purposes, such as data recovery [2] or intelligent tasks like classification and object detection [3]. The robustness of semantic communication holds great promise for various 6G applications, including the metaverse, mobile broadband, digital twins, and fully autonomous driving. By ensuring reliability and efficiency, semantic communication can establish a strong foundation for 6G systems [4], [5], supporting seamless interactions, content delivery optimization, real-time data exchange, and precise control.

One such progressive avenue involves the integration of generative AI into the communication framework. Generative AI, a subset of artificial intelligence, has revolutionized various domains by enabling machines to mimic, generate, and synthesize data that resembles human-created content. Leveraging generative AI in communication technologies can transform the landscape by facilitating semantic communication—an approach focused on conveying information at a higher semantic level, retaining only the most pertinent data for a given task. The paper contributions are organized as follows.

- First, we introduce the concept of semantic communication and its advantages. We also review the state-of-the-art techniques and challenges in semantic communication, such as joint source-channel coding (JSCC) [6].
- Subsequently, we propose a novel generative AI framework for semantic communication based on diffusion model [7] and dynamic architecture-based JSCC.
- Then, we evaluate the performance with metrics and demonstrate that our framework results in terms of image quality, semantic preservation, and robustness.

II. Previous Works

The pioneering work in semantic communication was the deep joint source-channel coding for image transmission [6], which proposed directly mapping image pixel values to channel inputs. This approach demonstrated improved accuracy metrics performance compared to conventional compression techniques such as JPEG and JPEG2000 [6] under noisy channel conditions. Subsequently, an alternative approach, the Wireless Image Transmission Transformer for Semantic Communications (WITT) [2], was introduced, leveraging a different neural network architecture based on transformers [3]. This approach, demonstrating robustness, further enhanced the quality. Nevertheless, both methodologies perform semantic encoding and decoding of the entire image, potentially including irrelevant information.

To address this, Joint Source-Channel Coding (JSCC) was proposed. JSCC refers to the combined process of processing data (source coding) and protecting it for transmission through a communication channel (channel coding). It aims to optimize the overall efficiency of the communication system by integrating source coding and channel coding to minimize the transmission and storage requirements while maintaining data integrity. The advantage of the JSCC is that it can take advantage of the statistical properties of the source data and the characteristics of the communication channel, optimizing the coding strategy for the specific characteristics of the data and the transmission channel.

III. System Model

In this section, a generative AI model (diffusion) based semantic communication system is proposed for image transmission. The system model is shown in Fig 1. It can be broken down into three parts. The transmitter extracts the semantic features of their structured representations from the original image and then extracts their compressed semantic features;

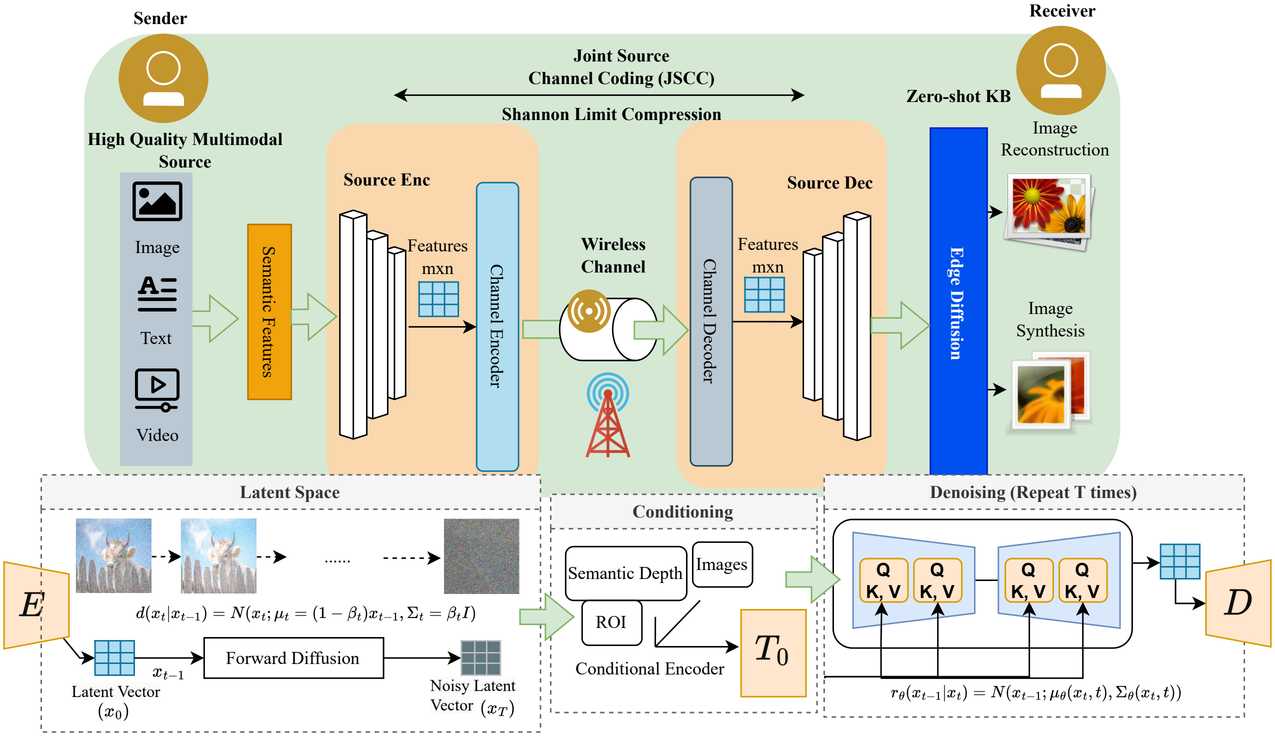

Fig. 1: This framework proposes an architecture for a zero-shot knowledge base employing a diffusion model. The architecture involves extracting semantic features from input data and encoding them for transmission through a wireless channel. The encoding and decoding processes make use of the JSCC technique. In the reconstruction with the edge diffusion process, the whole process consists of forward diffusion, conditioning, and reverse diffusion (or denoising).

the channel is used to simulate realistic noise and heavy compression; We employ a diffusion model at the receiver side after the DNN-based semantic encoder to reconstruct the original image from the noisy, compressed features.

A. Image Transmission with JSCC

We employ a Vision Transformer (ViT)-based architecture, comprised of an encoder F_{enc} and a decoder F_{enc}, dedicated solely to the source-channel coding. The architecture of the model is illustrated in Fig 2. The semantic features, denoted as X_s, are obtained by encoding the input X_{in}, which could be either the input image or mask. Mathematically, this process can be expressed as:

$$x_s = F_{enc}(x_{in}), \tag{1}$$

A noteworthy feature of this architecture is its ability to compress the transmitted signal size, given by:

$$CR = \frac{C_o \times H_o \times W_o}{C_{in} \times H_{in} \times W_{in}}, \tag{2}$$

where C_o, H_o, W_o represent the output channels, height, and width from the semantic encoder, respectively. Conversely, C_{in}, H_{in}, W_{in} denote their corresponding values for the input image.

In the trained autoencoder bottleneck architecture, given an input image represented by a sequence of flattened patches $x_i \in \mathbb{R}^{d_{\text{patch}}}$, where i indexes the patches and d_{patch} is the dimensionality of each patch, the ViT-B-16 architecture processes the input through linear projections to obtain query q_i, key k_i, and value v_i representations:

$$q_i, k_i, v_i = x_i W^Q, x_i W^K, x_i W^V \tag{3}$$

where W^Q, W^K, and W^V are the parameter matrices for queries, keys, and values, respectively. The self-attention mechanism of the transformer is then applied:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{4}$$

where Q, K, and V are the matrices formed by stacking the query, key, and value vectors, respectively, and d_{head} is the dimensionality of the query and key vectors. This process is repeated for multiple heads, and the outputs are concatenated and linearly transformed. With patch embedding Z_0 and transformer encoder Z'_ℓ described as follows

$$Z_0 = [X_p E] + E_{pos}, E \in \mathbb{R}^{(P^2 \cdot C) \times D}, E_{pos} \in \mathbb{R}^{(N+1) \times D} \tag{5}$$

$$Z'_\ell = \text{MSA}(\text{LN}(Z_{\ell-1})) + Z_{\ell-1}, \text{ for } \ell = 1 \dots L \tag{6}$$

Here, patch embeddings, denoted as Z_0, are defined as the sum of the product of the input patch matrix X_pE and a learnable embedding matrix E, augmented by a positional embedding matrix E_{pos}. The dimensions of E are in $\mathbb{R}^{(P^2 \cdot C)\times D}$, and E_{pos} has dimensions $\mathbb{R}^{(N+1)\times D}$, where P represents the patch size, C is the number of channels, and D is the embedding dimension. Additionally, the transformer encoder blocks, denoted as Z'_ℓ for the ℓ-th layer, follow a recurrent structure. Each block involves multi-head self-attention (MSA) applied to the layer-normalized input $Z_{\ell-1}$, summed with $Z_{\ell-1}$ itself. This process is repeated for L layers ($\ell = 1 \ldots L$), where L is the total number of layers. After the process, a multi-layer perceptron (MLP) is applied to each position independently:

$$\text{MLP}(x) = \text{PReLU}(W_2\text{PReLU}(W_1x + b_1) + b_2) \tag{7}$$

where W_1, W_2, b_1, and b_2 are learnable parameters. The ViT-B-16 model consists of a stack of such self-attention and MLP layers, contributing to its ability to capture complex patterns and long-range dependencies in images.

B. Channel Layer

To establish a communication channel, we consider the presence of noise in the environment. Although real physical noise is more complex to model for simplicity, we consider two common types of noise, additive white Gaussian noise (AWGN) and Rayleigh fading (RF), that are more widely adopted in several studies [2], [6]. AWGN represents random noise, while RF accounts for signal attenuation due to multipath propagation.

C. Image Reconstruction Block

The noisy received signal $y \in \mathbb{R}^{H_o\times W_o\times C_o}$ is subsequently input to the CNN decoder, which reconstructs the original image, represented as $x'_{in} \in \mathbb{R}^{H_{in}\times W_{in}\times C_{in}}$. This process is formalized as:

$$x'_{in} = F_{dec}(y). \tag{8}$$

To refine the process of image reconstruction block, we are employing a lightweight edge diffusion model with zero-shot capabilities. Diffusion models are a class of latent variable models that learn to generate data by reversing a Markov chain that gradually adds noise to the data. The Markov chain is defined by a score function that measures the likelihood of the data given the noise level.

$$d(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \mu_t = 1 - \beta_t x_{t-1}, \Sigma_t = \beta_t I\right) \tag{9}$$

With the expression for $d(x_{1:T} \mid x_0)$ is given by:

$$d(x_{1:T} \mid x_0) = \prod_{t=1}^{T} d(x_t \mid x_{t-1}) \tag{10}$$

The diffusion models are trained beforehand to learn the reverse process and thus produce high-fidelity image data from semantic text descriptions by iterating backward over this chain. This step consists of denoising U-Net, which is a segmentation model. In utilizing diffusion models to restore images from vision semantics, the score function employs these semantics as conditioning information and guides the reverse Markov chain as follows:

$$r_\theta(x_{t-1} \mid x_t) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\right) \tag{11}$$

This processing restores high-fidelity images that align closely with visual semantics.

IV. Case Study and Discussion

Diffusion models, such as our DiffRes and DiffRefiner, represent an innovative approach in noisy transmission channels. These models function by iteratively refining an image, adding and then removing noise to enhance the image's fidelity. DiffRes introduces resilience in the initial stages of the diffusion process, optimizing the image for transmission to the expected noise level. DiffRefiner, on the other hand, operates at the receiving end, refining the transmitted image to reduce any residual noise and recover the original quality to the greatest extent possible.

In our investigation of the DiffRes and DiffRefiner frameworks, we conducted simulations for image transmission, explicitly comparing their performance to DeepJSCC. The dataset used featured military and civilian vehicles. We employed two key metrics to assess the models' effectiveness: Structural Similarity Index Measure (SSIM) [8] and Learned Perceptual Image Patch Similarity (LPIPS) [9]. SSIM evaluates the structural information between images, with higher values indicating a closer resemblance to the original image and superior image quality. In contrast, LPIPS, a perceptual metric leveraging deep learning, measures the perceptual difference between images, aligning more closely with human visual judgment. Lower LPIPS values signify a perceptually closer match to the original, signifying successful recovery of the transmitted image.

In our presented results, the efficacy of our proposed models, DiffRes and DiffRefiner, in maintaining image integrity over noisy channels is evident. Notably, both models exhibit improved SSIM values compared to DeepJSCC. Illustrated in the accompanying figure, DiffRes achieves the highest SSIM value, showcasing superior similarity to the original, undistorted images, particularly as SNR levels improve. While DiffRefiner also outperforms DeepJSCC in SSIM, its results are marginally lower compared to DiffRes. This implies that under increasingly favorable transmission conditions—characterized by reduced noise—DiffRes excels at preserving visual details with greater effectiveness.

Furthermore, the LPIPS results provide additional support for the effectiveness of our proposed models. Both DiffRes and DiffRefiner exhibit a notable reduction in LPIPS values as SNR levels increase, showcasing their proficiency in mitigating distortions. This trend holds consistently across AWGN and Rayleigh Fading channel conditions, highlighting the models' resilience against various types of noise encountered during image transmission. Specifically, under the AWGN channel, both proposed methods achieve a lower LPIPS value

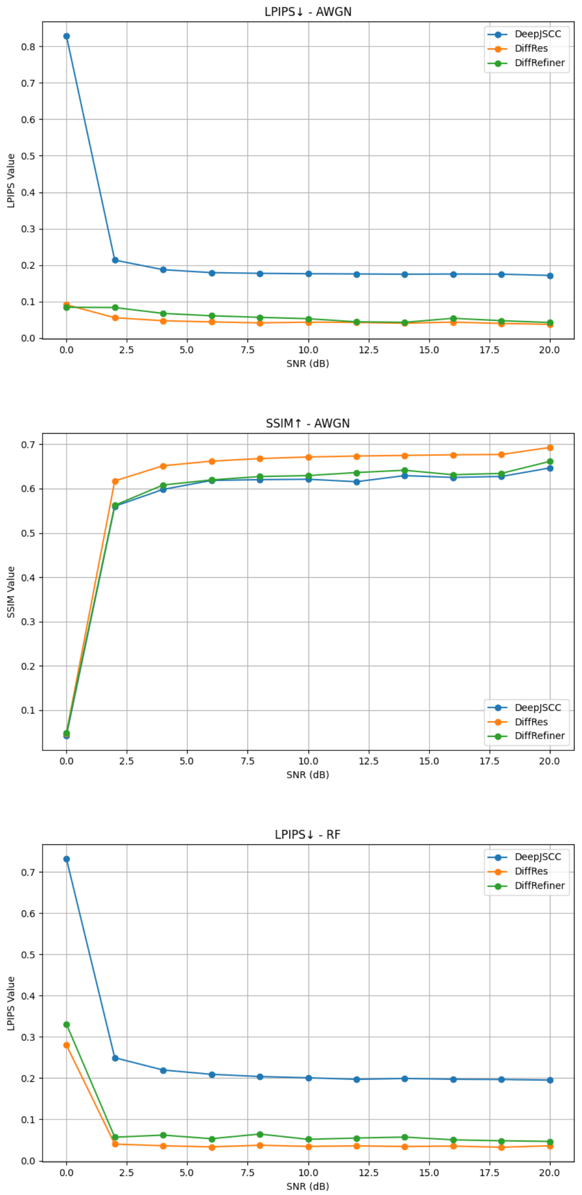

Fig. 2: Comparison results depict the performance of our proposed frameworks, DiffRes and DiffRefiner, along with DeepJSCC, under AWGN and Rayleigh fading channels at various SNR levels. Evaluation metrics include LPIPS and SSIM to assess image quality and similarity.

of 0.1 compared to DeepJSCC. In the case of the Rayleigh Fading channel, the LPIPS values for DiffRes and DiffRefiner are lower than those of the DeepJSCC model, registering around 0.15 LPIPS. This underscores the models' adaptability and superior performance in diverse noisy environments.

V. Conclusion

In this paper, we proposed DiffRes and DiffRefiner frameworks within a generative AI context, which demonstrate commendable performance in image transmission over noisy channels. Leveraging diffusion models and dynamic architecture-based Joint Source-Channel Coding (JSCC), our system excels in preserving image integrity, particularly in scenarios with lower Signal-to-Noise Ratios (SNRs). Notably, DiffRes takes a proactive stance in optimizing images for initial transmission, surpassing the performance of DeepJSCC with higher SSIM values. Moreover, both DiffRes and DiffRefiner exhibit substantial reductions in (LPIPS) values as SNR levels increase, emphasizing their efficacy in mitigating distortions under diverse noise conditions. This research contributes to advancing semantic communication, offering promising applications in 6G technologies and beyond, where reliability and efficiency are crucial for seamless interactions and various intelligent tasks. While our focus in this paper was on the impact of semantic information on image recovery in semantic coding models, future research could explore its role in task achievability. Additionally, the potential of other lightweight, sustainable, and Channel State Information (CSI)-based architectures could be leveraged for further investigation.

References

[1] X. Luo, H.-H. Chen, and Q. Guo, "Semantic communications: Overview, open issues, and future research directions," *IEEE Wirel. Commun.*, vol. 29, no. 1, pp. 210–219, Jan. 2022.

[2] K. Yang, S. Wang, J. Dai, K. Tan, K. Niu, and P. Zhang, "Witt: A wireless image transmission transformer for semantic communications," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[3] H. Xie, Z. Qin, X. Tao, and K. B. Letaief, "Task-oriented multi-user semantic communications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2584–2597, Jul. 2022.

[4] C. D. Alwis, A. Kalla, Q.-V. Pham, P. Kumar, K. Dev, W.-J. Hwang, and M. Liyanage, "Survey on 6G frontiers: Trends, applications, requirements, technologies and future research," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 836–886, Apr. 2021.

[5] E. C. Strinati and S. Barbarossa, "6G networks: Beyond shannon towards semantic and goal-oriented communications," *Comput. Netw.*, vol. 190, p. 107930, May 2021.

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

[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," Jun. 2022, pp. 10 684–10 695.

[8] W. Zhou, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

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