

Intermediate-Step Conditional Diffusion Reconstruction for VAE-Based Semantic Communication

Naoki Omi and Shun Kojima

Graduate School of Engineering, Chiba University, Chiba, Japan

Email: naoki.omi031@gmail.com

Abstract—Diffusion models have demonstrated remarkable performance in semantic communication systems, achieving superior reconstruction quality even under extremely high compression ratios and low signal-to-noise ratio (SNR). However, diffusion models typically suffer from high computational cost and lack an explicit latent representation, which limits their controllability. To overcome these limitations, this paper proposes a semantic communication system that integrates the variational autoencoder (VAE) and intermediate-step diffusion model (IS-VAE-Diff). In the proposed framework, the transmitter employs a VAE to encode input images into channel-adaptive latent representations, while the receiver uses the VAE reconstruction as a conditional input to the diffusion model for perceptual quality enhancement. Furthermore, reverse diffusion is initiated from an intermediate step to reduce inference cost while restoring lost details using the global structures preserved by the VAE. Experimental results show that the proposed IS-VAE-Diff achieves favorable reconstruction accuracy and perceptual quality with substantially lower inference complexity than existing diffusion-based methods.

Index Terms—diffusion models, joint source-channel coding, semantic communications, variational autoencoders.

I. INTRODUCTION

Wireless communication systems have continuously improved communication quality by optimizing physical-layer performance metrics such as throughput and bit error rate. However, current technologies are already approaching the Shannon limit, and it has been pointed out that they may struggle to accommodate the exponentially increasing communication demands in the near future [1]. Semantic communications have consequently attracted considerable attention as a promising framework [2]. In semantic communications, the transmitter extracts meaningful information from the original data and eliminates redundant components that are unnecessary for the task before transmission. This enables a significant reduction in bandwidth consumption while maintaining the performance required for the end task. In particular, deep joint source-channel coding (DeepJSCC) [3] has been widely used as a representative framework for image transmission, where source and channel coding are jointly optimized using deep neural networks.

Building on such deep learning-based image transmission frameworks, the utilization of generative artificial intelligence (GAI) models has recently been regarded as a promising approach to enhance image reconstruction accuracy at the

receiver in semantic communication systems [4]–[6]. With the continued advancement of research on GAI applications, diffusion models have attracted considerable attention, particularly for image transmission [4]. Among them, the denoising diffusion probabilistic model (DDPM) [7] has demonstrated remarkable reconstruction performance in image generation tasks owing to its strong generative capability and stable training characteristics, and is widely regarded as a promising candidate for semantic communications.

S. F. Yilmaz et al. [8] proposed a method in which low-resolution images intentionally degraded by DeepJSCC are transmitted and then reconstructed at high quality using DDPM at the receiver. F. Pezone et al. [9] employed a conditional DDPM (CDDPM) at the receiver to generate high-resolution images from transmitted semantic segmentation maps and low-resolution images. C. Liang et al. [10] presented a selective image generation approach in which the transmitter extracts image features and category information, and the receiver uses a CDDPM to generate multiple candidate images, selecting the most similar to the original using a similarity metric. In these methods, data compression and transmission rely on a separate JSCC module, while the CDDPM serves as an auxiliary reconstruction module at the receiver. This modular design facilitates integration with existing DeepJSCC frameworks.

However, a standard CDDPM lacks an explicit latent space, making it challenging to design conditional inputs that reliably guide the generation toward desired outputs. Moreover, because image data are high-dimensional and processed sequentially through the diffusion steps, these models entail high computational costs and long inference times [11]. K. Pandey et al. [12] proposed a method that combines a variational autoencoder (VAE) with a CDDPM to address the lack of controllability in the latent space of CDDPM, where the reconstructed image produced by the VAE is used as a conditional input for the CDDPM to generate high-resolution images. The VAE has an explicit latent space and enables control over the latent-variable distribution [13]. By applying this property to semantic communication, it becomes possible to design robust transmission symbols adapted to channel conditions [14]. However, the reconstructed images from a VAE tend to be overly smoothed and often lose fine details. In addition, the increased computational complexity associated

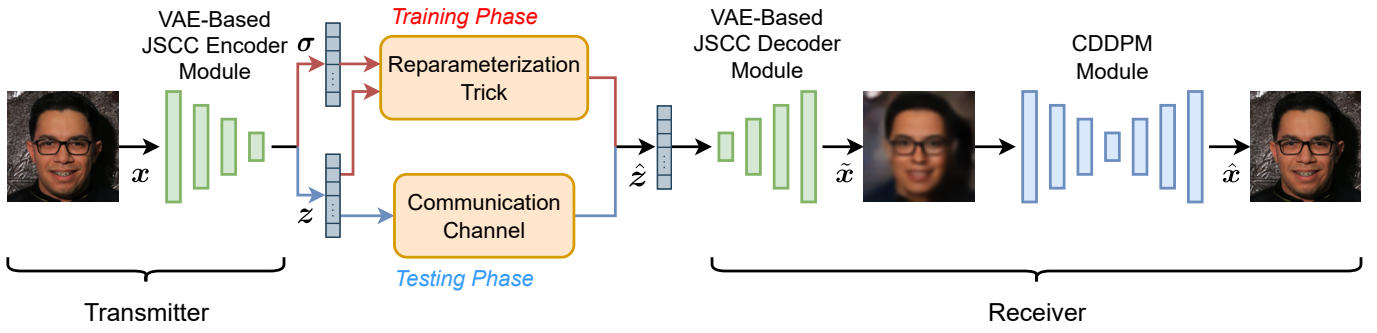


Fig. 1. The proposed semantic communication system model.

with the diffusion process poses a significant challenge.

To address these problems, this paper proposes a novel semantic communication framework, the intermediate-step VAE-diffusion model (IS-VAE-Diff), which maintains reconstruction quality while reducing computational cost by initiating the reverse diffusion process at an intermediate step. The proposed approach simultaneously ensures robustness and perceptual quality by leveraging the global structural information preserved in the VAE as a conditional input and by allowing the CDDPM to complement missing fine details.

II. SYSTEM MODEL

The proposed system model consists of a joint source-channel encoder, a wireless communication channel, a joint source-channel decoder, and a CDDPM, as illustrated in Fig. 1. The encoder f_θ maps the source image $x \in \mathbb{R}^n$ to transmitted symbols $z \in \mathbb{C}^k$:

$$z = f_\theta(x), \quad (1)$$

where n is the dimensionality of the source data and k is the dimensionality of the transmitted symbols. The bandwidth compression ratio is defined as

$$\rho = \frac{k}{n}. \quad (2)$$

The transmitted symbols output by the encoder must satisfy the following average power constraint before transmission:

$$\frac{1}{k} \mathbb{E} [\|z\|_2^2] \leq P. \quad (3)$$

Assuming a Rayleigh fading channel, the received symbols are given by

$$\hat{z} = hz + n, \quad (4)$$

where $h \sim \mathcal{CN}(0, 1)$ is the channel coefficient, and $n \sim \mathcal{CN}(0, \sigma_n^2 \mathbf{I}_k)$ is independent and identically distributed complex Gaussian noise with power σ_n^2 . The decoder g_ϕ reconstructs an intermediate image $\tilde{x}$ from $\hat{z}$:

$$\tilde{x} = g_\phi(\hat{z}). \quad (5)$$

Finally, the CDDPM r_ψ generates the reconstructed image $\hat{x}$ using $\tilde{x}$ as a conditional input:

$$\hat{x} = r_\psi(x_T, \tilde{x}), \quad (6)$$

where $x_T \sim \mathcal{N}(0, \mathbf{I}_n)$ is a noise image sampled from a standard normal distribution, and $\mathbf{I}_n$ is the $n \times n$ identity matrix.

III. PROPOSED METHOD

The proposed IS-VAE-Diff consists of two main components: a VAE-based JSCC module for obtaining an intermediate reconstruction and a conditional diffusion module for perceptual refinement. The key idea is to use the VAE output not only as a conditional input but also as the basis for initializing the reverse diffusion process from an intermediate step.

A. VAE-Based JSCC

The VAE-based JSCC module directly maps an input image into channel symbols and reconstructs an intermediate image from the received symbols. The VAE is trained to minimize the reconstruction error while regularizing the latent representation. The loss function is given by

$$\mathcal{L}_{\text{VAE}} = \frac{1}{n} \|x - \tilde{x}\|_2^2 + \frac{1}{k} d_{\text{KL}}(q_\theta(\hat{z}|x) \| p(\hat{z})), \quad (7)$$

where the first term represents the reconstruction loss and the second term is the Kullback-Leibler (KL) divergence between the learned latent distribution and the prior distribution. The KL regularization enables control of the latent distribution according to the communication channel. In this paper, a channel-aware prior distribution depending on the channel noise variance is employed during training.

Since sampling from a probability distribution is non-differentiable, the reparameterization trick [13] is introduced under Rayleigh fading:

$$\hat{z} = hz + \sigma \epsilon_k, \quad (8)$$

where z is the transmitted latent representation, σ is the standard deviation output by the encoder, and $\epsilon_k \sim \mathcal{CN}(0, \mathbf{I}_k)$ is complex Gaussian noise.

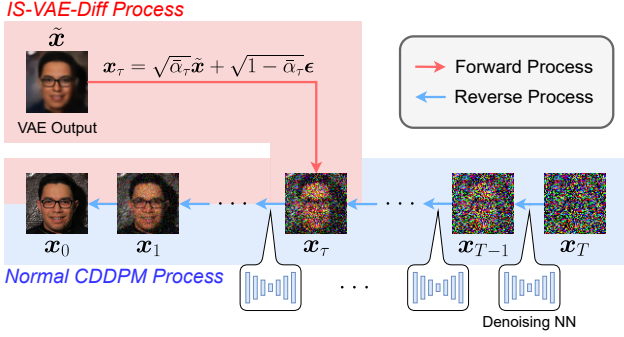


Fig. 2. Intermediate-step diffusion reconstruction.

B. Conditional Diffusion Reconstruction

To improve the perceptual quality of the VAE reconstruction, we introduce a conditional diffusion model at the receiver. In the forward diffusion process, Gaussian noise is gradually added to the original image x_0 :

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon_n, \quad (9)$$

where $\epsilon_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $t \in \{1, \dots, T\}$ is the diffusion step, and $\bar{\alpha}_t$ is determined by the noise schedule. The conditional diffusion model learns to predict the added noise using the noisy image x_t , the diffusion step t , and the VAE output $\tilde{x}$ as the conditional input. The training loss is defined as

$$\mathcal{L}_{\text{diff}} = \frac{1}{n} \|\epsilon_n - \epsilon_\psi(x_t, t, \tilde{x})\|_2^2, \quad (10)$$

where ϵ_ψ denotes the noise prediction network. In the reverse diffusion process, the model progressively removes noise while conditioning on the VAE reconstruction. Through this process, the diffusion model restores perceptually important textures and fine details lost in the VAE reconstruction.

C. Intermediate-Step Diffusion Reconstruction

Standard conditional diffusion reconstruction starts the reverse process from pure Gaussian noise x_T , requiring T reverse diffusion steps. Although this process can improve perceptual quality, it incurs high computational cost. To reduce the inference cost, IS-VAE-Diff starts the reverse diffusion process from an intermediate step τ instead of T , as illustrated in Fig. 2. The key observation is that the VAE reconstruction generally preserves global structures such as object shape and spatial arrangement, whereas diffusion models are effective at restoring fine textures and contours [7], [15]. Therefore, the proposed scheme skips the early reverse diffusion steps and performs only the latter refinement process.

Specifically, we generate an intermediate noisy image from the VAE reconstruction as

$$\tilde{x}_\tau \sim \mathcal{N}(\sqrt{\bar{\alpha}_\tau} \tilde{x}, (1 - \bar{\alpha}_\tau) \mathbf{I}_n). \quad (11)$$

The reverse diffusion process is then performed from $t = \tau$ to $t = 0$ using $\tilde{x}$ as the condition. This reduces the number of reverse diffusion steps from T to τ . The selection of τ involves

TABLE I
NETWORK ARCHITECTURES

VAE-based JSCC module	CDDPM module
Conv. Block (32)	Conv. (64)
Conv. Block (64) ↓	Residual + Atten. Block (64) ↓
Conv. Block (128) ↓	Residual + Atten. Block (128) ↓
Conv. Block (256) ↓	Residual + Atten. Block (256) ↓
Lin. (2k) ↓ / Lin. (2k) ↓	Residual + Atten. + Residual Block (256) ↑
Reparam. Trick	Residual + Atten. Block (256) ↑
Lin. (4 × H × W) ↑	Residual + Atten. Block (128) ↑
Tr. Conv. Block (128) ↑	Residual + Atten. Block (64)
Tr. Conv. Block (64) ↑	Conv. (3)
Tr. Conv. Block (32) ↑	
Tr. Conv. Block (3)	

a tradeoff between reconstruction quality and inference cost. A smaller τ requires fewer reverse steps and preserves more information from the VAE output, while a larger τ allows stronger diffusion-based refinement but increases the inference cost. In this paper, we set $T = 1000$ and $\tau = 100$, reducing the number of reverse diffusion steps to 10% of the full T -step reverse diffusion process as a practical tradeoff between reconstruction quality and inference cost.

D. Network Architecture

Table I summarizes the architectures of the VAE and CDDPM modules. Here, the values in parentheses denote the number of output channels. The symbols ↓ and ↑ represent downsampling and upsampling operations, respectively. The VAE module consists of four convolutional blocks, one linear layer, and one normalization layer in the encoder, and one linear layer and four transposed convolutional blocks in the decoder. Each convolutional block includes a convolutional layer, a parametric rectified linear unit activation function, and a batch normalization layer. The CDDPM module adopts a U-Net architecture [16] composed of three downsampling modules, one bottleneck module, and three upsampling modules. Each module consists of residual and attention blocks, and skip connections are employed between the downsampling and upsampling modules.

IV. NUMERICAL RESULTS

This section presents the performance evaluation of the proposed scheme. The Flickr-Faces-HQ (FFHQ) dataset [17] with a resolution of 64×64 pixels is used in the simulations. To evaluate the end-to-end performance, structural similarity (SSIM) [18], learned perceptual image patch similarity (LPIPS) [19], and Fréchet inception distance (FID) [20] are employed as image reconstruction metrics. All experiments are conducted on a workstation equipped with an Intel Core i9-14900K CPU and an NVIDIA GeForce RTX 4090 GPU. To verify the effectiveness of the proposed scheme, Deep-JSCC [3], VAE [5], and the diffusion-based semantic communication system [8] are used for comparison.

A. Performance Against Varying Bandwidth Compression Ratios

Fig. 3 shows the performance at different bandwidth compression ratios ρ under Rayleigh fading. During both training

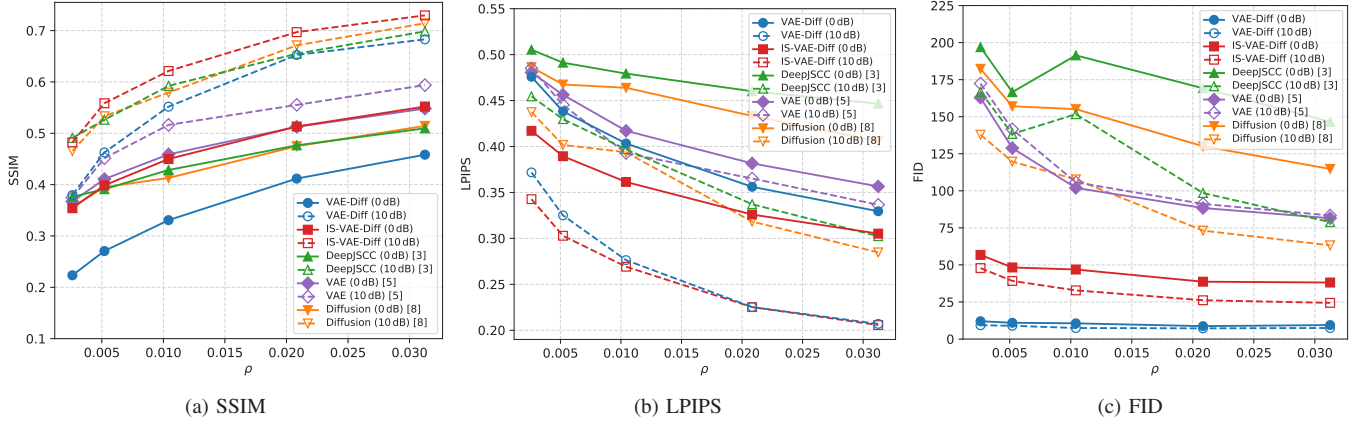


Fig. 3. Performance at various bandwidth compression ratios over the Rayleigh fading channel.

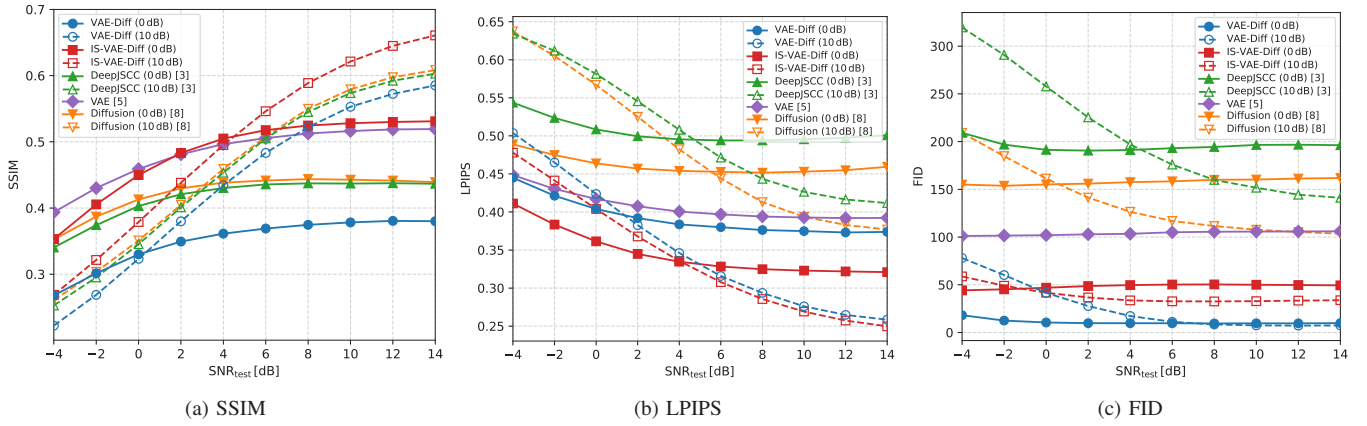


Fig. 4. Performance comparisons at various SNR levels over the Rayleigh fading channel at $\rho = 0.010$.

and validation, the SNR is fixed at 0 dB or 10 dB, and the same SNR is used in both phases to consistently assess performance across compression ratios. For the VAE, the prior distribution is set to $\mathcal{N}(\mathbf{0}, \mathbf{I})$ independently of the SNR. Here, VAE-Diff can be regarded as a special case of IS-VAE-Diff with $\tau = T$, in which the reverse diffusion process starts from the initial noise state.

In Fig. 3 (a), VAE achieves the highest SSIM at SNR = 0 dB, while IS-VAE-Diff performs best at SNR = 10 dB. This indicates that VAE-Diff contributes more to perceptual quality improvement than to structural similarity under low SNR conditions. In the LPIPS results shown in Fig. 3 (b), both VAE-Diff and IS-VAE-Diff outperform the other methods. In Fig. 3 (c), VAE-Diff again achieves the best FID result, confirming that an increased number of reverse diffusion steps tends to improve perceptual quality.

B. Performance Against Varying SNR

In this subsection, we evaluate the robustness of the proposed scheme under channel conditions different from those used during training. Specifically, models trained at a fixed SNR are tested under multiple SNR conditions, and reconstruction accuracy is measured to assess their stability under

SNR variation. The bandwidth compression ratio is set to $\rho = 0.010$, assuming a Rayleigh fading channel model.

The SSIM curves in Fig. 4 (a) indicate that VAE attains the highest scores in the low-SNR regime, whereas IS-VAE-Diff consistently provides the best performance at high SNR. These observations imply that, unlike the conventional VAE, which cannot adapt its learning to the channel noise variance, the proposed scheme can effectively exploit the noise statistics, yielding substantial gains, particularly under high-SNR conditions. For LPIPS in Fig. 4 (b), both VAE-Diff and IS-VAE-Diff produce the lowest values over the entire SNR range, confirming their consistent advantage in terms of perceptual quality. For FID in Fig. 4 (c), VAE-Diff achieves the best performance, followed by IS-VAE-Diff.

C. Visualization

Fig. 5 shows the reconstructed images over the Rayleigh fading channel with SNR = 10 dB and $\rho = 0.031$. In Fig. 5 (b), the images reconstructed by DeepJSCC exhibit pronounced global distortions, where large-scale warping significantly degrades the overall facial structure. The VAE results in Fig. 5 (c) show fewer distortions compared to DeepJSCC, but

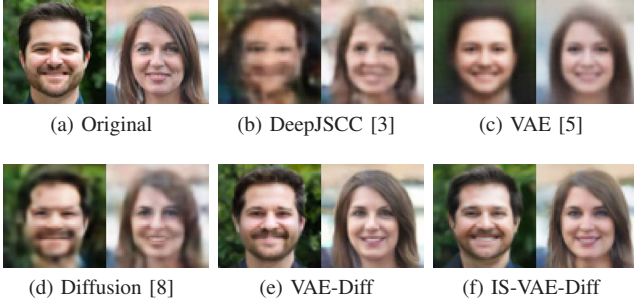


Fig. 5. Reconstructed images over the Rayleigh fading channel with SNR = 10 dB at $\rho = 0.031$.

TABLE II
COMPUTATIONAL COMPLEXITY OF EACH METHOD.

Method	FLOPs	Time [s]	LPIPS	FID
DeepJSCC [3]	49.41 M	1.69×10^{-3}	0.3970	151.69
VAE [5]	816.5 M	1.81×10^{-3}	0.3928	105.74
Diffusion [8]	7.295 T	5.43	0.3937	107.73
VAE-Diff	7.296 T	5.87	0.2764	7.44
IS-VAE-Diff	0.7303 T	0.521	0.2693	23.30

the overall images appear excessively smoothed. The Diffusion results in Fig. 5 (d) alleviate the smoothing effect observed in the VAE results; however, some local distortions still remain. In contrast, VAE-Diff in Fig. 5 (e) suppresses over-smoothing and reconstructs clearer facial features. Finally, as shown in Fig. 5 (f), IS-VAE-Diff maintains reconstruction quality comparable to that of VAE-Diff while substantially reducing the computational cost of inference.

D. Computational Complexity Analysis

Table II compares the computational complexity, processing time, and reconstruction performance of each method under Rayleigh fading with SNR = 10 dB and $\rho = 0.010$. The processing time corresponds to the average time required from encoding to image reconstruction. DeepJSCC and VAE achieve low computational cost but show inferior perceptual quality. Although VAE-Diff improves reconstruction performance, it incurs significantly higher computational cost and processing time. In contrast, IS-VAE-Diff substantially reduces floating-point operations (FLOPs) and processing time while maintaining high reconstruction quality. In particular, IS-VAE-Diff achieves the best LPIPS performance, demonstrating an effective tradeoff between inference efficiency and perceptual quality.

V. CONCLUSION

This paper proposed IS-VAE-Diff, a semantic communication system integrating a VAE and a CDDPM for high-quality image transmission. The proposed scheme employs intermediate-step reverse diffusion to improve inference efficiency while restoring fine details from the VAE reconstruction. Numerical results demonstrated that the proposed scheme

achieves superior perceptual quality and reconstruction performance compared with conventional methods while reducing the computational cost of inference.

ACKNOWLEDGMENTS

This work was supported in part by JSPS KAKENHI (25K01227), Toyota Physics and Chemistry Research Institute, and Konica Minolta Science and Technology Foundation.

REFERENCES

- [1] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, and M. Zorzi, "Toward 6G networks: use cases and technologies," *IEEE Commun. Mag.*, vol. 58, no. 3, pp. 55–61, Mar. 2020.
- [2] I. Choi and S. Kojima, "Stacked intelligent metasurfaces-based joint source-channel coding for image transmissions," *IEEE Open J. Commun. Soc.*, vol. 7, pp. 2825–2837, 2026.
- [3] E. Bourtsoulatz, D. Burth Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Trans. Cognit. Commun. Networking*, vol. 5, no. 3, pp. 567–579, Sept. 2019.
- [4] L. Xia, Y. Sun, C. Liang, L. Zhang, M. A. Imran, and D. Niyato, "Generative AI for semantic communication: architecture, challenges, and outlook," *IEEE Wireless Commun.*, vol. 32, no. 1, pp. 132–140, Feb. 2025.
- [5] S. Ma *et al.*, "Task-oriented explainable semantic communications," *IEEE Trans. Wireless Commun.*, vol. 22, no. 12, pp. 9248–9262, Dec. 2023.
- [6] N. Omi *et al.*, "VAE-GAN-based semantic communication for image transmission: enhancing perceptual and semantic similarity," *Proc. IEEE Asia Pacific Conf. Wireless Mobile (APWiMob)*, Nov. 2025, pp. 12–17.
- [7] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 6840–6851, 2020.
- [8] S. F. Yilmaz, X. Niu, B. Bai, W. Han, L. Deng, and D. Gündüz, "High perceptual quality wireless image delivery with denoising diffusion models," *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, May 2024, pp. 1–5.
- [9] F. Pezone, O. Musa, G. Caire, and S. Barbarossa, "Semantic-preserving image coding based on conditional diffusion models," *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP)*, 2024, pp. 13501–13505.
- [10] C. Liang, D. Li, Z. Lin, and H. Cao, "Selection-based image generation for semantic communication systems," *IEEE Commun. Lett.*, vol. 28, no. 1, pp. 34–38, Jan. 2024.
- [11] C. Liang *et al.*, "Generative AI-driven semantic communication networks: architecture, technologies, and applications," *IEEE Trans. Cognit. Commun. Netw.*, vol. 11, no. 1, pp. 27–47, Feb. 2025.
- [12] K. Pandey, A. Mukherjee, P. Rai, and A. Kumar, "Diffusevae: efficient, controllable and high-fidelity generation from low-dimensional latents," 2022, *arXiv:2201.00308*.
- [13] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2013, *arXiv:1312.6114*.
- [14] N. Omi, S. Kojima, and C. -J. Ahn, "VAE-GAN-based semantic communication for high-quality image transmission," *IEEE Trans. Wireless Commun.*, vol. 25, pp. 9799–9813, 2026.
- [15] A. Camuto and M. Willetts, "Variational autoencoders: a harmonic perspective," 2022, *arXiv:2105.14866*.
- [16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," 2015, *arXiv:1505.04597*.
- [17] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.
- [18] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
- [19] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit.*, Jun. 2018, pp. 586–595.
- [20] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6626–6637.