Leveraging Stable Diffusion with Context-Aware Prompts for Semantic Communication

Long V. Nguyen*, Tuan T. Nguyen†, Octavia A. Dobre* and Trung Q. Duong*

* Memorial University, Canada, e-mail: {lnguyen, odobre, tduong}@mun.ca
† University of Greenwich, UK, e-mail: tuan.nguyen@greenwich.ac.uk

***Abstract*—Semantic communication in wireless image transmission leverages the meaning embedded in the image data, aiming to compress, transmit, and reconstruct images based on their semantic content rather than purely pixel data. This paradigm shift allows more efficient utilization of bandwidth and computational resources, focusing on extracting key features and contextual information that is critical for ensuring that the essential content of the image is preserved and accurately conveyed. In this study, we present a novel Stable Diffusion-based semantic communication (SDSC) framework that demonstrates high performance, characterized by an elevated bandwidth compression ratio (BCR) and robust noise tolerance achieved by diffusion mechanism integrating supplementary prompts. Our approach utilizes pre-trained modules of a Variational autoencoder (VAE) and a modified U-shaped network (UNet) to enable robust semantic encoding, decoding, and effective channel denoising. This scheme significantly enhances the system's ability to preserve data integrity and meaning in noisy environments. By introducing additional context-aware prompts during transmission, we improve the accuracy of received information and mitigate the adverse effects of interference and noise. Extensive simulations show that our framework outperforms previous innovative models, demonstrating superior communication fidelity and efficiency under various challenging conditions.**

I. Introduction

Wireless image transmission plays a crucial role in diverse applications, from surveillance and remote sensing to telemedicine and autonomous vehicles. The traditional approaches to data transmission are being reconsidered to enhance efficiency, reliability, and relevance. They focus primarily on the bit-level accuracy, often neglecting the semantic content that is interpreted by end-users. Recently, semantic communication (SemCom) [1] [2] has emerged as a promising approach, aiming to improve meaning-oriented accuracy and efficiency by prioritizing the semantic content of images. This approach enables significant improvements in data compression by filtering out irrelevant information and selectively transmitting essential semantic elements rather than the bit symbols.

Joint source-channel coding (JSCC) has been an active research area that combines source and channel coding to improve the overall performance of communication systems. This method has gained considerable attention in recent years due to its potential to enhance the robustness and efficiency of wireless communication systems over the separate approach [3]. The primary goal of JSCC is to optimize the transmission of information over noisy channels by jointly designing the source coding and channel coding processes. This technique is particularly crucial in wireless communication systems, where channel errors and packet loss can significantly degrade the quality of the transmitted data. However, developing these techniques has not resulted in practical applications because of their high complexity and restricted performance improvements when utilized with genuine sources and channels [4]. Deep JSCC [5], an emerging alternative approach leveraging deep neural networks, has shown promise in dealing with low signal-to-noise ratio (SNR) and narrow channel bandwidth under different channel conditions. In [6], the authors improve deep JSCC by integrating the swin transformer architecture [7] instead to achieve the WITT framework. Unlike conventional CNN-based methods, which struggle with global dependencies and high-resolution images, WITT harnesses the superior ability of the transformer architecture to extract long-range and high-level semantic features. This design results in higher fidelity textures, fewer block artifacts, and better overall performance. However, such transformer-based models lack robustness due to their reliance on massive amounts of data for large-scale training, making them sensitive to the dataset quality and generalization ability. Through end-to-end training, the above systems emphasize altering individual pixels or maintaining structural similarity rather than perceptual similarity.

Recent years have witnessed substantial advancements in generative models that produce highly realistic images with augmented perceptual quality. The intersection of generative artificial intelligence (GenAI) and SemCom presents exciting possibilities. Generative adversarial networks (GANs) [8] and diffusion models [9] exemplify this innovation, achieving state-of-the-art results in text, image, and video generation. GANs have been increasingly applied in various fields, including computer vision, natural language processing, and audio processing. In the context of SemCom, GANs have been used to generate original semantic signals from distorted semantic signals, eliminating the need for channel state information (CSI) [10]. Despite the promising results, the application of GANs faces several challenges. For instance, GANs are prone to mode collapse, where the generator produces a limited variety of outputs despite the potential diversity in the training data, and gradient disappearance, which can affect the diversity and quality of generated samples. Additionally, the training process of GANs can be unstable and require careful tuning of hyperparameters. Diffusion models, instead, have become a game-changer due to their ability to generate high-quality multimedia content while preserving semantic features. For example, a diffusion model is deployed as a decoder in [11] to build a hybrid JSCC scheme. The works [12] and [13] adopt the diffusion mechanism on the channel denoising process to improve the constructed image. However, these approaches do not directly employ the state-of-the-art generative model in the JSCC-based wireless transmission system. Hence, they have not fully exploited the strong prior knowledge of large pretrained models, leading to a missed opportunity for substantial improvement in the perceptual quality of reconstructed images. Moreover, the condional generation of genAI is not leveraged to provide further semantic context in advancing the quality of wireless transmission over noisy channels.

In this work, we introduce SDSC, a framework that integrates the pretrained Stable Diffusion (SD) model [14] into a semantic communication system to enhance the perceptual quality of the constructed images. The key contributions of this paper are outlined below:

- We directly adopt the generative model into the traditional JSCC system to fully exploit the strong prior knowledge of large pretrained models for efficient semantic communication.
- By leveraging the diffusion mechanism, we can diminish the effects of channel noise from corrupted data to facilitate image reconstruction.
- Our findings show that using additional context-aware prompts during transmission can elevate the perceptional quality of the reconstructed image, especially in low SNR scenarios.
- Numerical tests on the Kodak dataset [15] demonstrate that our system balances the trade-off between compression ratio and semantic preservation in image quality in challenging communication environments.

II. System Model and Evaluation Metrics

A. System Model

This section details the working principle of our semantic communication system for wireless image transmission. As in [5], we model the transmitter and receiver as neural networks using pretrained components of the Stable Diffusion (SD) model [14]. The SD-based wireless semantic image communication system model (SDSC) is illustrated in Fig. 1. An RGB image $x \in \mathbb{R}^n$, where $n = 3 \cdot H \cdot W$, is encoded by a variational autoencoder (VAE) $\mathcal{E}$ into a low-dimensional latent representation $z = \mathcal{E}(x)$. The complex latent vector $z \in \mathbb{C}^k$ is normalized to satisfy the average power constraint P_{avg} and obtain z_n

$$z_n = \sqrt{\frac{kP_{\text{avg}}}{||z||_2^2}}z \tag{1}$$

The bandwidth compression ratio (BCR) is defined as $p = k/n$ in JSCC literature, where n represents the source bandwidth and k is the channel bandwidth. After that, the achieved signal is transmitted over the noisy channel. In this study, we consider the widely adopted channel model, additive white Gaussian noise (AWGN) channel, denoted by $n \sim \mathcal{CN}(0, \sigma^2 I_k)$, where σ^2 is the average noise power. Therefore, the received signal $\tilde{z}$ at the receiver is indicated as follows:

$$\tilde{z} = z_n + n \tag{2}$$

We also define the SNR, which reflects the channel quality

$$\text{SNR} = 10 \log_{10} \left(\frac{P_{\text{avg}}}{\sigma^2} \right) \text{dB}. \tag{3}$$

To incorporate textual information, we adopt the BLIPv2 model [16] as an image captioner to extract further semantic features in the form of text prompts corresponding with each input image. These prompt is then validated thoroughly to guarantee their alignment with the visual content in hard cases. Due to the significantly higher bandwidth consumption of wireless image transmission than text transmission, we focus solely on the image, assuming that any additional text prompt is transmitted without error over the noisy channel.

We consider the transmission process over the AWGN channel at the transmitter as a forward process of the SD model at timestep t. Similarly, reconstruction of the noisy image is employed at the receiver as a reverse diffusion process in the latent space, which iteratively samples $\tilde{z}_{t-1}$ from $p(\tilde{z}_{t-1}|\tilde{z}_t)$ to obtain a noise-free latent representation $\tilde{z}_0 \sim q(z)$ through a sequence of denoising steps. As mentioned in [9], this denoising process can be expressed as

$$\tilde{z}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\tilde{z}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(\tilde{z}_t, t) \right) + \sigma_t \epsilon \tag{4}$$

where $\alpha_t \in (0,1)$, $\bar{\alpha}_t = \prod_{i=0}^{t} \alpha_i$, and $\epsilon \sim \mathcal{N}(0,1)$. $\epsilon_\theta(\tilde{z}_t, t)$ is an estimation of added noise. To sample $\tilde{z}_0$, the added noise is predicted by the pretrained UNet model [17], which is augmented with cross-attention module mechanism [18]. For the inclusion of semantic information as the text prompt y, the CLIP encoder model is leveraged to project y to prompt embedding $\tau_\theta(y)$, which is mapped via a cross-attention layer to the intermediate layers of the UNet. This conditions the prediction to produce the desired output. The network ϵ_θ can be learned through the loss function

$$L_{LDM} = \mathbb{E}_{\mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0,1), t} \left[\|\epsilon - \epsilon_\theta \left(\tilde{z}_t, t, \tau_\theta(y) \right)\|_2^2 \right] \tag{5}$$

The loss function is minimized gradually until the training process converges. Ultimately, the denoising scheduler can subtract the predicted noises successfully to obtain the pure latent representation $\tilde{z}_0$ after t denoising steps. Subsequently, the noise-free latent signal $\tilde{z}_0$ is forward through the VAE decoder $\mathcal{D}$ to reconstruct the original image $\tilde{x}$

$$\tilde{x} = \mathcal{D}(\tilde{z}_0) \tag{6}$$

B. Evaluation metrics

In our evaluation, we utilize widely recognised distortion metrics such as peak signal-to-noise ratio (PSNR) and multi-scale structural similarity index (MS-SSIM). We also consider learned perceptual image patch similarity (LPIPS) [19] and Fréchet inception distance (FID) scores [20], which have been

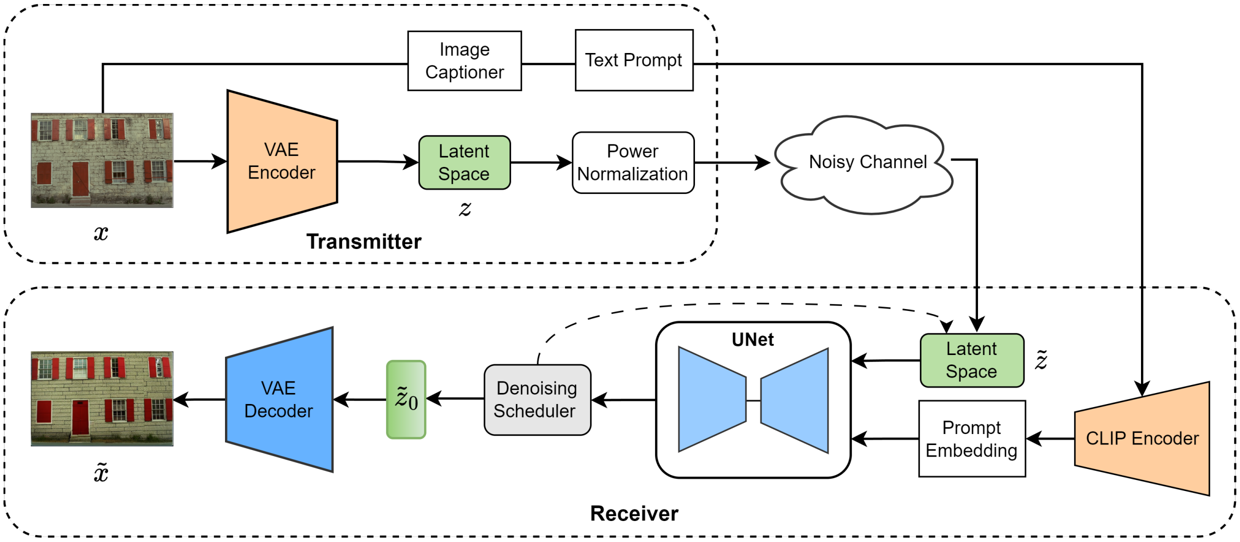

Fig. 1: Illustration of the wireless image transmission system with Stable Diffusion (SDSC) in semantic communication. The pretrained components are leveraged to encode and decode the transmitted data within latent space, optimizing the trade-off between compression ratio and semantic preservation. Channel noise is iteratively refined through a diffusion mechanism, while context-aware prompts are incorporated to provide additional semantic features for the reconstruction process.

shown to enhance the perceptual quality of constructed images. The models exhibit improved performance with higher PSNR and MS-SSIM values, while lower values are better for the latter two metrics.

To prevent randomness and maintain objectivity, each image is transmitted ten times, and the evaluation scores are averaged to determine the final results.

III. Simulation Results and Discussions

A. Simulation Setting

To illustrate the advanced performance of the proposed system, we compare our method SDSC with two recent innovative methods: Deep JSCC [5] and WITT [6]. As previous research has already demonstrated the superior performance of DeepJSCC over traditional separation-based methods, we do not include those approaches as baselines for comparison in this study. The SDSC with the textual prompt input (SDSC + Prompt) is also provided as a baseline to examine the efficiency of additional semantic context.

For the introduced models, we implement the pretrained Stable Diffusion model version 1.5, which was trained on 512x512 images from the vast database LAION-5B [21]. The Deep JSCC and WITT models are trained on the DIV2K dataset [22] with 800 high-resolution images, and the Kodak dataset [15] containing 24 images of 768 × 512 dimensions is utilized for evaluation. During the training process, all the images are randomly cropped in the same input dimension of the proposed model. Experiments are conducted on channel SNRs of $\{1, 4, 7, 10, 13\}$ dB, the power constraint $P_{\text{avg}} = 1$, and the compression ratio $p = 1/48$, corresponding to complicated communication situations. The SNR values for training are uniformly sampled within this range. Through empirical experimentation, we set the strength noise hyperparameter at 0.4 for SNR = 1 dB, 0.2 for SNR = 4 dB, and 0.1 for all remaining SNR values. Since we use wireless channel noise instead, the strength noise hyperparameters are adopted in this case to control the magnitude of the sampling timestep t. We apply the DDPM sampling method to the denoising process. The number of sampling steps is important in determining the visual quality of reconstructed output images. To opt for the proper steps, our choices are based on the LPIPS score with pretrained neural network VGG to capture high-level perceptual features at various layers of the compared images. We investigate multiple values of sampling steps for different SNRs at $p = 1/48$ to balance realism and distortion, as shown in Fig. 2. The optimal steps selected are $\{25, 25, 125, 25, 25\}$ steps for each corresponding SNR value.

With the experiments involving additional text prompts, we define the classifier-free guidance scale, which denotes how strongly the stable diffusion model should follow the guidance of the input prompt, at the medium value of 7.5. The other settings are the same as the SDSC model without prompts.

B. Numerical Results and Discussions

Fig. 3 presents the performance comparison between our proposed methods and baseline schemes for varying SNR values on the Kodak dataset with $p = 1/48$. Across four evaluation metrics, Deep JSCC illustrates the worst performance, indicating poorer reconstructed image quality given the conditions. At low SNR values, our proposed models show lower PSNR yet become superior to the WITT model at higher SNR values from 7 dB. Regarding the MS-SSIM metric, WITT performs better than our models in all cases. The reason is that PSNR and MS-SSIM are two traditional metrics measuring the quality of the reconstructed images based on structural

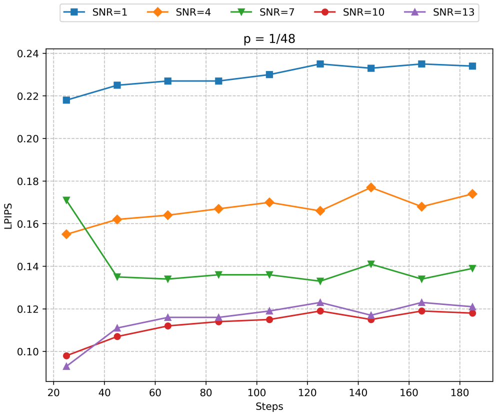

Fig. 2: Effect of sampling steps for various SNRs at $p = 1/48$. The optimal steps are selected based on the lowest LPIPS score for each SNR value.

information rather than perceptual similarity. At the same time, our systems are designed to capture and maintain the semantic features of the transmitted images. With the presence of channel noise, preserving subtle structural information does not necessarily ensure maintaining visual quality, which is pivotal in semantic communication. Therefore, regarding LPIPS and FID metrics, both SDSC and SDSC + Prompt significantly surpass the baseline models by a large margin. Especially, SDSC + Prompt consistently outperforms SDSC in most scenarios in terms of both structural and perceptual similarity, verifying the efficacy of additional prompts in enhancing image fidelity.

Fig. 4 shows samples of reconstructed images for visual comparison regarding perceptual quality. The Deep JSCC method struggles to recover fine details, while WITT performs better; however, the presence of noise artifacts remains significant at $p = 1/48$. In contrast, our proposed SDSC model, incorporating the diffusion model's denoising mechanism, effectively restores key features of the original images. Furthermore, leveraging the additional semantic context of text prompt, SDSC + Prompt further enhances reconstruction quality, producing results more faithful to the source images, despite minor deviations. Notably, the final image sample demonstrates the model's ability to significantly improve the reconstruction of parrots with more realistic body parts (beak and eyes) given the additional text prompt "two colorful parrots standing next to each other." The similar experiment in [23] strengthens further our approach.

IV. Conclusion

By leveraging the pretrained Stable Diffusion model, we have created an innovative wireless semantic image transmission system called SDSC to significantly improve the perceptual quality of received images. Transmitted data is compressed into latent space and reconstructed with VAE blocks to balance the trade-off between compression ratio and semantic preservation. The channel denoising process is facilitated by the diffusion mechanism. Specifically, adding context-aware prompts to the given image improves the accuracy of the reconstructed information and diminishes the negative impacts of noise over the wireless channel. Through the experiments on the Kodak dataset, our approaches show the superiority in delivering a visually high-quality image with a low compression ratio across various SNRs. We believe generative AI in general, and the diffusion model in particular hold huge potential in semantic communication to enable substantial improvements regarding efficiency, relevance, and adaptability in wireless networks. In the future, we will explore further the adaptation of the diffusion model and the integration of context-aware prompts to fully realize the potential of this innovative approach. The synergy between semantic communication and generative AI could lead to revolutionary changes in how we perceive and design communication systems.

Acknowledgment

This work was supported in part by the Canada Excellence Research Chair (CERC) Program CERC-2022-00109 and in part by the Canada Research Chair Program CRC-2022-00187.

References

[1] D. Gündüz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 5–41, Jan. 2023.

[2] W. Yang, H. Du, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. Shen, and C. Miao, "Semantic communications for future internet: Fundamentals, applications, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 213–250, Feb. 2023.

[3] F. Zhai, Y. Eisenberg, and A. K. Katsaggelos, "Joint source-channel coding for video communications," *Handbook of Image and Video Processing*, pp. 1065–1082, 2005.

[4] O. Y. Bursalioglu, G. Caire, and D. Divsalar, "Joint source-channel coding for deep-space image transmission using rateless codes," *IEEE Trans Commun.*, vol. 61, no. 8, pp. 3448–3461, Aug. 2013.

[5] 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, Sep. 2019.

[6] K. Yang, S. Wang, J. Dai, K. Tan, K. Niu, and P. Zhang, "WITT: A wireless image transmission transformer for semantic communications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process (ICASSP)*, Rhodes, Greek, Jun. 2023, pp. 1–5.

[7] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *IEEE/CVF Int. Conf. on Comput. Vision (ICCV)*, Oct. 2021, pp. 9992–10 002.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Adv. in Neural Inf. Processing Syst. (NIPS)*, Montreal, Canada, Dec. 2014, pp. 2672–2680.

[9] J. A. Ho, Jonathan and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33. Curran Associates, Inc., Dec. 2020, pp. 6840–6851.

[10] J. Mao, K. Xiong, M. Liu, Z. Qin, W. Chen, P. Fan, and K. B. Letaief, "A gan-based semantic communication for text without csi," *arXiv preprint arXiv:2312.16909*, 2023.

[11] X. Niu, X. Wang, D. Gündüz, B. Bai, W. Chen, and G. Zhou, "A hybrid wireless image transmission scheme with diffusion," in *Proc. IEEE Signal Process. Adv. Wireless Commun. (SPAWC)*, Lucca, Italy, Sep. 2024, pp. 86–90.

[12] T. Wu, Z. Chen, D. He, L. Qian, Y. Xu, M. Tao, and W. Zhang, "CDDM: Channel denoising diffusion models for wireless communications," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Kuala Lumpur, Malaysia, Dec. 2023, pp. 7429–7434.

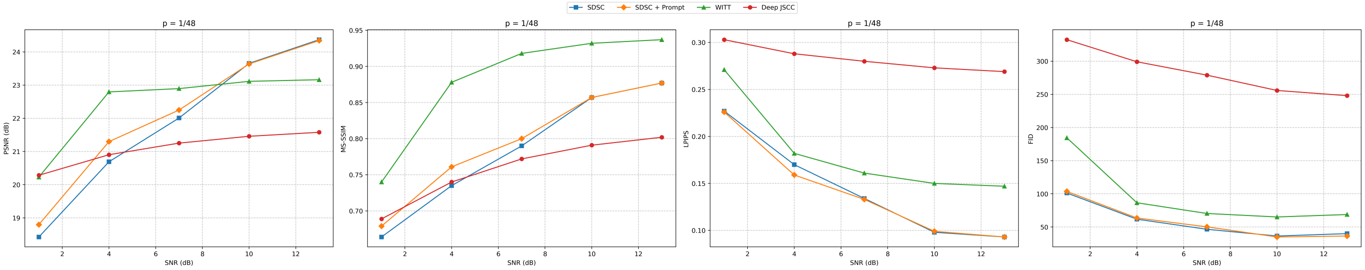

Fig. 3: Performance comparison between the proposed methods and the baseline for various SNRs at $p = 1/48$ on the Kodak dataset.

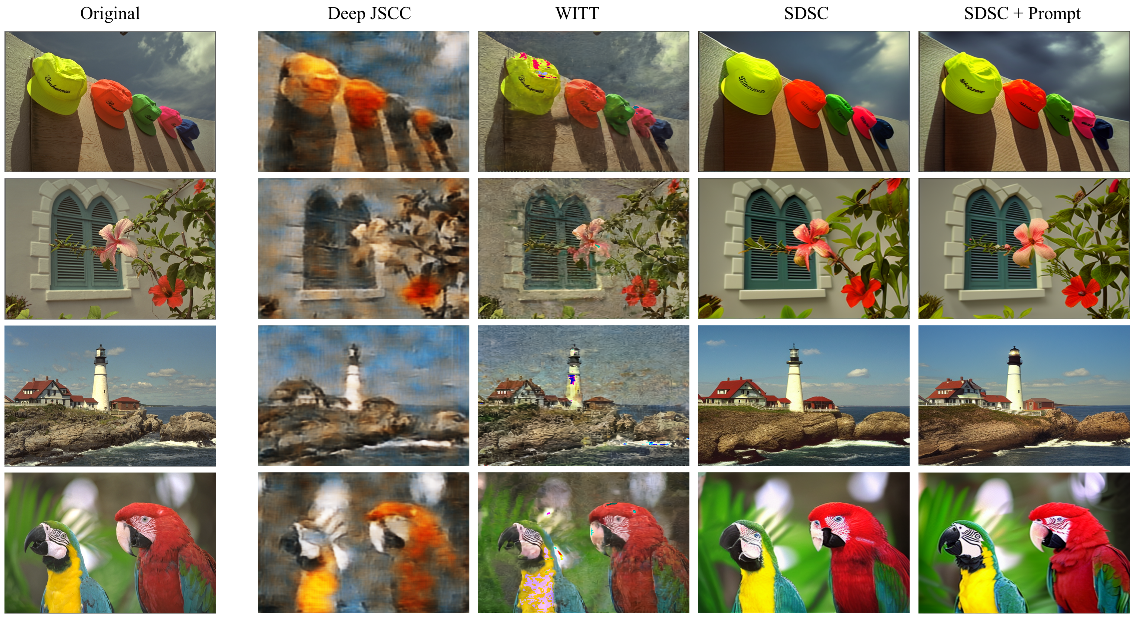

Fig. 4: Visual comparison of the reconstructed images between the proposed methods and the baseline at SNR = 1dB and $p = 1/48$ on the Kodak dataset.

[13] B. Xu, R. Meng, Y. Chen, X. Xu, C. Dong, and H. Sun, "Latent semantic diffusion-based channel adaptive de-noising semcom for future 6G systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Kuala Lumpur, Malaysia, Dec. 2023, pp. 1229–1234.

[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recogn. (CVPR)*, Jun. 2022, pp. 10 684–10 695.

[15] "Kodak lossless true color image suite," http://r0k.us/graphics/kodak/, 1993.

[16] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. conf. Mach. Learn. (PMLR)*, 2023, pp. 19 730–19 742.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Med. Imag. Comput. and Computer-Assisted Interv. (MICCAI)*. Munich, Germany: Springer, Oct. 2015, pp. 234–241.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Adv. in Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017.

[19] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recogn. (CVPR)*, 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," *Adv. in Neural Inf. Process. Syst. (NIPS)*, vol. 30, 2017.

[21] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *Adv. in Neural Inf. Processing Syst. (NeurIPS)*, vol. 35, pp. 25 278–25 294, 2022.

[22] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn. Workshop (CVPRW)*, Jul. 2017, pp. 126–135.

[23] M. Yang, B. Liu, B. Wang, and H.-S. Kim, "Diffusion-aided joint source channel coding for high realism wireless image transmission," *arXiv preprint arXiv:2404.17736*, Apr. 2024.