International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 12 | Dec – 2025 DOI: 10.55041/ISJEM05262
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Review on Vision Transformer Based Semantic Communications for Next Generation Wireless Networks

Akhilesh[1], Mr Pradeep Nayak[2], Adithi Shetty[3], Ankitha K N[4], Anushree K[5]

Faculty, Department of Information Science and Engineering[2]

Students, Department of Information Science and Engineering[1,3,4,5]

Alva's Institute of Engineering and Technology, Mijar, Mangalore, Karnataka, India.

Email: anukiran959@gmail.com

Abstract: Semantic communications, which prioritize the transfer of semantic meaning above raw data, have the potential to completely transform data transmission in the developing 6G network landscape. data precision. This research introduces a semantic communication system based on Vision Transformers (ViTs) that has been specifically built to minimize bandwidth consumption while achieving high semantic similarity during image transmission. The suggested architecture may effectively encode images into a high semantic content at the transmitter and accurately reconstruct the images at the receiver while taking noise and real-world fading into account by using ViT as the encoder-decoder framework. Our approach beats Convolution Neural Networks (CNNs) and Generative Adversarial Networks (GANs) designed for producing such images, building on the attention processes built into ViTs. With a Peak Signal-to-noise Ratio (PSNR) of 38 dB, the architecture based on the suggested ViT network outperforms existing Deep Learning (DL) techniques in preserving semantic similarity across various communication settings. These results prove that our ViT-based method is a major advancement in semantic communications. Vision Transformer (ViT), Deep Learning (DL), 6G, semantic communication, bandwidth efficiency, and Peak Signal to Noise Ratio (PSNR) are the index terms.

Keywords: Semantic Communication, Deep Learning (DL), Image Transmission, Bandwidth Efficiency

I. OVERVIEW

A new paradigm in wireless communication called SEMANTIC communications aims to convey the core of the information rather than just the data. This technique is particularly important for 6G networks, where maintaining the connection under challenging circumstances and attaining great bandwidth efficiency are critical [1]. Conventional information transmission techniques, which primarily rely on bit-level transmissions, frequently fall short of certain requirements for efficient bandwidth usage or signal stability in the face of interference. Computer vision, image processing, and wireless communication have all significantly improved with the introduction of Deep Learning (DL) architectures. In this context, a number of research have suggested strategies for effective bandwidth utilization [2]. Compressed communication has long made use of autoencoders. Remarkable results were obtained by combining a simple autoencoder with a denoising autoencoder in a robust technique presented in [3]. Nevertheless, only the effects of quantization and Additive White Gaussian Noise (AWGN) are examined in this study. In a similar vein, [4] suggested a densely linked autoencoder structure to optimize feature extraction, which was modeled after the DenseNet architecture. To lessen distortion, they also created a structure resembling a U-Net. Even though their approach performed better than JPEG 2000, there were still some imperfections.

Better methods that preserve the conveyed semantic content while using less bandwidth than the existing standard ones are needed for the transition from 5G to 6G. Specifically, DL is well suited for semantic communication due to its ability to extract and prioritize useful information. Convolutional Neural Networks (CNNs) are particularly promising for compressing and transferring visual data in computer vision [5].

Because of their advantages in spatial hierarchies and local feature extraction, CNNs have historically performed exceptionally well at picture reduction and transmission. CNN-based applications for picture transmission in noisy channels with channel coding and denoising were investigated by the authors in [6]. On the one hand, CNNs' capacity to capture global context—a crucial component of semantic communication—is hampered by their local receptive fields. However, with improved semantic communication capabilities, Vision Transformers (ViTs) have lately become a viable substitute for CNNs. ViTs are ideal for contexts with limited bandwidth and channel interference because they can communicate more semantic data and encode global context. ViTs are centered on preserving semantic communication, in contrast to more traditional approaches.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 12 | Dec – 2025 DOI: 10.55041/ISJEM05262
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

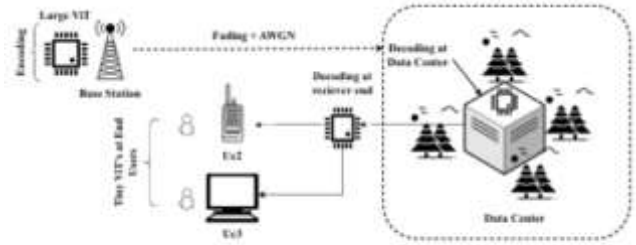


Fig. 1: System model of proposed wireless network system

By comparing ViTs' performance with CNNs and Generative Adversarial Networks (GANs) on datasets like ImageNet, CIFAR-10, and CIFAR-100, this study investigates the potential of ViTs in semantic communication. We evaluate these models' ability to withstand difficult channel conditions such AWGN, Rayleigh fading, Rician fading, and Nakagami-m fading. Motivated by the benefits of ViTs, our paper adds to the body of knowledge in the following ways:

- Outperformed the effectiveness of cutting-edge autoencoders by presenting a unique ViT architecture for effective semantic transmission.
- Performed a performance analysis using a variety of datasets, modeling several fading models to precisely mimic wireless conditions found in the real world.
- Assessed ViT's resilience in comparison to modern semantic models by comparing its accuracy against many DL models, resulting in a 72% reduction in bandwidth usage.

Keywords: Autoencoder ,Semantic Similarity, Wireless Channel Noise

II. Model of the System

We assume a system model for compressed communication, namely semantic communication, employing ViT architecture made up of big and tiny ViT models, as seen in Fig. 1. We suppose that the transmitter side processes the input images Xu. Each of the N patches created by slicing these images is restricted to a linear embedding and transformed into an appropriate vector representation. After that, a ViT-based encoder receives these embedded patches and converts them into learnable features. F.

A. Modulation of BPSK

A technique called Binary Phase Shift Keying (BPSK) modulation modulates the reference signal's phases in order to send the message [7]. In our model, a bipolar NRZ encoder was used to process binary data streams xi(t) from the ViT model, mapping 0 and 1 to voltage levels −1 and +1 to generate X(t). BPSK was used to modulate this waveform, altering the carrier phase by 0◦ for +1 and 180◦ for −1. Since the BPSK signal is a phase shift keying (PSK) signal, its phase changes based on the input bit at a certain moment.

$$X_{\text{PSK}}(t) = A\cos(\omega_c t) \quad \text{for input bit 1,} \qquad (1)$$

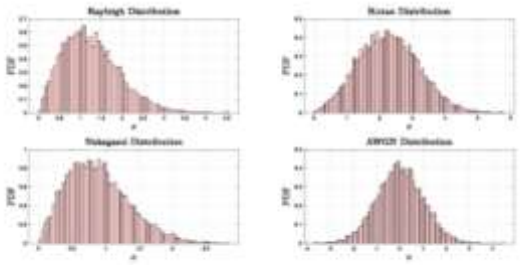


Fig. 2: Sampled probability density function (PDF's) of fadings

$$X_{\text{PSK}}(t) = A\cos(\omega_c t + \pi) \quad \text{for input bit 0.} \qquad (2)$$

It can transmit data while utilizing the least amount of bandwidth possible with just two phases, which is crucial and consistent with the processing requirements of ViTs. Here, the following formula is used to transform each bit bi into a symbol:

$$s_i = 2b_i - 1, \qquad (3)$$

where bi can be either 0 or 1. The binary features are converted into symbols that are prepared for transmission through this modulation procedure.

B. Channel of Communication

We included both Line-of-Sight (LOS) and Non-Line-of-Sight (NLOS) channel characteristics to replicate the stochastic nature of real-world communication settings [8]. Prior to sampling those distributions, their Probability Density Functions (PDFs) were plotted as seen in Fig. 2. We created complex Gaussian random variables, hI and hQ, with zero mean and unit variance for Rayleigh and Rician fading channels.

One way to express the Rayleigh channel is:

$$h(t) = \frac{1}{\sqrt{2}}\left(h_I(t) + jh_Q(t)\right), \qquad (4)$$

where fR(r) = r 2σ2 gives the PDF of the Rayleigh distributed amplitude, and the LOS component improves the channel modeling even more. The following represents the Rician channel model with fading coefficient h(t):

$$h(t) = \sqrt{\frac{K}{K+1}} + \sqrt{\frac{1}{K+1}} \cdot \frac{1}{\sqrt{2}}\left(h_I(t) + jh_Q(t)\right), \qquad (5)$$

where σ = 2K σ2 and I0(·) is the bessel function, and R(r) = r σ2 e−r2+v2 2σ2 I0 r σ2 is the PDF of the Rician distribution amplitude.

In order to replicate all of the statistical channel fading scenarios, Nakagami-m, a more comprehensive representation of fading scenarios, was also included. R(r) = 2mm Γ(m)Ωm r2m−1e−mr2 Ω is the PDF of the Nakagami-m distributed amplitude r, where Ω = E[r2] is the spread parameter, the mean power of the fading envelope, m ≥ 1 2 is the shape parameter, and Ω is the Gamma function. Thermal noise is a prevalent occurrence in electronic communication systems. This AWGN was simulated by adding stationary noise. N(n) = 1 2πσ2 e−n2 2σ2 is the PDF of Gaussian noise n. To replicate the actual communication environment, the modulated signals were then sent over a communication channel with small-scale fading. One way to model the received signal y is as follows:

$$y = h \cdot s + n, \qquad (6)$$

where s is the transmitted BPSK signal, n is the AWGN, and h stands for the channel fading coefficients. By incorporating fading and noise into the system model, the effects can be better studied by utilizing both LOS and NLOS scenarios.

C. Conversion of Features to Bit

The possibility that the data center would receive encoded data from several sources was taken into account. At the data center, the received signal y is first demodulated to extract the encoded characteristics. A decoder then processes these features to reconstitute the original data. Floating-point numbers in binary format have been represented using IEEE 754. For this conversion, single precision (32 bits) has been considered [9]. For element-wise conversions, statistical techniques were used. By compiling the decorated functions to machine code, the vector size decorator was set up for this effective conversion from integer to floating point and floating to bit conversions, and vice versa. The mantissa M, the exponential value E, and the sign bit S make up the IEEE 754 format. The floating point's value, V, is determined by:

$$V = (-1)^S \times M \times 2^E. \quad (7)$$

D. Information Sets

We experimented with three distinct datasets to train our suggested DL models in order to obtain a thorough assessment of our models. Initially, we used Cifar-10, which consists of 32x32 colored images with labels against these classes, to train the DL models. We also included another dataset, Cifar-100, which has 100 classes, to further evaluate the robustness of our model. Lastly, using
the imagenette dataset, which has over a million photos from a thousand classes, we assessed the performance and scalability of our suggested custom model against a published baseline.

Keywords: AWGN Noise, Rayleigh Fading, Rician Fading, CIFAR-10, CIFAR-100

III. MODELS OF DEEP LEARNING

A. Architecture of Transformers

We suggest a unique ViT model made up of the following components: (i) Non-Sequential: ViTs process images in a different way than conventional CNNs, which do so sequentially.
Compared to CNNs, they require less training time because of their non-sequential nature, which permits greater parallelization and the acquisition of global context. (ii) Self Attention: Using sequential data processing to rank the similarity scores between several patches in a picture. (iii) Positional Embedding: Since ViTs are non-sequential learning models, the spatial layout of image patches is restored using positional embedding.

Algorithm 1 Algorithm for Semantic Communication

Input: Image X for encoding with $x \times y$ dimensions.
Step 1: Image Compression.
for epoch $= 1$ **to** N **do**
  Obtain positional embedding of X with dimension θ_e.
  Obtain patches of X with dimensions θ_p.
  Pass the input image X through transformer to extract global features θ_f with dimension $(\theta_e, \theta_n, \theta_l)$.
end for
Step 2: Phase Modulation and Interference Management.
Perform BPSK modulation to convert features to symbols.
Add pathloss θ_z and fading θ_w to the encoded features θ_f to mimic the wireless environment.
Step 3: Image Reconstruction.
for epoch $= 1$ **to** N **do**
  Pass the noisy features through decoder positional embedding θ_e to obtain the position of patches θ_x.
  Pass the output θ_x through the transformer block to remove the noise and reconstruct the image.
  Arrange the patches in order to get the denoised reconstructed image from the features θ_f.
end for
Output: Reconstructed denoised image $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$.

Three main modules make up our Denoising Autoencoder Vision Transformer model (DAE-ViT): (a) encoder, (b) decoder, and (c) additional utility (fading channels). We created the ViT base model in order to explore with the ViT model; Table I displays its scales.
However, by modifying the hyperparameters to adapt the model complexity of various conditions by adhering to the method, alternative models can also be implemented in different ways. 1. Encoder Patch units of images are the input that ViT receives. Using Multi-Head Self Attention (MHSA) operations, these patches are converted into a two-dimensional sequence known as embedded patches X ∈ R(N+1)×D in order to understand their reciprocity. In essence, MHSA is a component of the transformer model that enables the model to concurrently focus on various input sequence segments [10]. The patch layer first divides input images into patches of 16x16 processing pixels, with each image dimensions set to 224 x 224 pixels.

TABLE I: Configuration of DAE-ViT Models

Configuration	ViT Small	ViT Base	ViT Large
Patch Size	16	16	16
Embedding Dimension	384	768	1024
Encoder Layers	12	12	24
Encoder Heads	6	12	16
Decoder Layers	8	8	16
Decoder Heads	8	16	16

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 12 | Dec – 2025 DOI: 10.55041/ISJEM05262
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

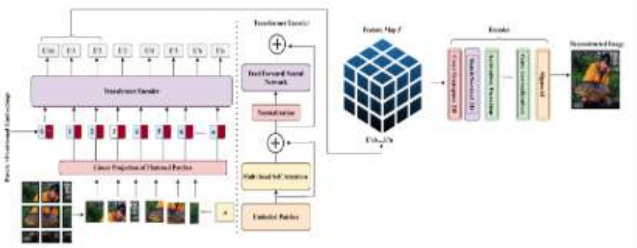


Fig. 3: Framework of proposed ViT architecture

Softmax(z)i is defined as softmax(z)i = ezi j ezj. Each patch is then embedded into a vector with a given dimension of 768. Small, non-overlapping patches of size (S × S) are created from the input photos Xu; the number of patches is N = w S2, and w = l×Nc. For (1 < j ≤ N), the patches are converted into vectors xpu,j ∈ RS2, which are then added to the model dimension d via a linear projection E ∈ RS2× d. Patch embedding is the term for the patch's output. The embedded patches contain a class token called xcls. The input sequence is then encrypted using positional embedding Epos ∈ R(N+1)× d. They have the same number of places as the number of patches when they are first initialized as a learnable parameter [11]. The patch and position encapsulating Z0's output is provided by:

$$Z_0 = [x_{cls}; x_{p_{u,1}}E; x_{p_{u,2}}E; \ldots; x_{p_{u,N}}E] + E_{pos}. \quad (8)$$

The information aggregation process is guided by the attention weight matrix (A), which computes attention scores between tokens. The attention scoreAij between query qi and key kj was computed using the compatibility function [12]. The attention information in the current embedding is represented by the weighted sum of all elements value v of patch embedding X. The query, key, and value vectors [q,k,v] that have been concatenated are provided by:

$$[q, k, v] = XU, \quad U \in \mathbb{R}_h^{D\times 3D}. \quad (9)$$

The learned weight matrix Uqkv links the vectors x to the query q, key k, and value v with dimensions D×3Dh, where 3Dh represents the concatenated dimensions of q, k, and v and D represents the dimensionality of the input embedding. The model's input flow and processing are guided by the attention weights computation [10], which was ultimately determined by:

$$A = \text{softmax}\left(\frac{qk^\top}{\sqrt{D_h}}\right), \quad A \in \mathbb{R}^{(N+1)\times(N+1)}, \quad (10)$$

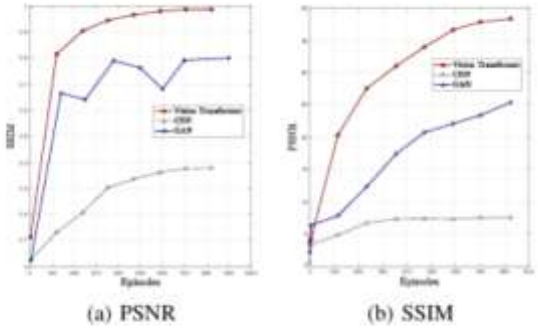
In order to further process it through a sequence of transformer blocks, each of which is made up of MHSA and feed-forward layers, the integrated embedding, patch embedding, positional embedding, and class token were rearranged. This allowed the model to learn to capture both local and global dependencies throughout the image. The MHSA equation is provided by:

$$\text{MHSA}(X) = [SA_1(X); SA_2(X); \ldots; SA_k(X)]U_{\text{msa}}, \quad U_{\text{msa}} \in \mathbb{R}^{k\cdot D_h\times D}. \quad (11)$$

When MHSA(X) is applied to X, each SAi(X) individually calculates self-attention. Umsa is a learnt weight matrix with dimensions k · Dh ×D that aggregates the outputs of all attention heads into a final output. Before obtaining the output, each transformer block's output underwent layer normalization. Because it guarantees that the permutations of the variance and the mean are the same across the embedding dimensions, this normalizing step facilitates stability and training procedure improvements.

Decoder: The ViT reconstructs the features F produced by the encoder into an image F ∈ RN1/2× N1/2× D. These attributes were processed by transformer blocks in the decoder, which subsequently produced the pictures [13]. Positional embedding of vector size 768 was then added for 196 distinct picture patch locations as well as one for the class token. The purpose of this phase was to preserve the spatial information of the image's patches.

Additionally, the combined embedding was followed by transformer blocks, where the MHSA technique was used to capture the local and global dependencies in the data. This output was a summary of the first input. After applying a feed-forward neural network to each point, the neural network's output was normalized. The class token was then eliminated once the characteristics were reassembled into sequences. A linear layer is used to transform each feature vector into a 768-pixel patch-sized vector. As seen in Fig. 7, these patches were adjusted to create the whole image: đX •RH×W×C. Finally, the sigmoidgate function was used to minimize noise and artifacts while maintaining high fidelity.



(a) PSNR (b) SSIM

Fig. 4: Image quality under different fading scenarios

B. CNNs, or convolutional neural networks

In particular, learnt features s in X •RB×F×2N, where B is the batch size and F is the number of frames, were transformed using conventional CNN layers with 2DCNN modules. The suggested model uses Principal Component Analysis (PCA) to reduce dimensionality, break down the feature matrices, and minimize the mean square error:

$$\|(XW)\hat{W} - X\|_2^2 \rightarrow_{W,\hat{W}} \min. \quad (12)$$

In the initial samples, the dense layer was a completely linked layer with linear activation, whereas direct W—Rm×d and reverse W—Rd×m transformations were used:

$$f(X) = W \cdot X + \vec{b}. \quad (13)$$

Dense layers were used for encoding, convolution, and pooling layers. Prior to being added to the dense layer, the output was flattened [14]. Transpose convolution was used to aggregate picture patches for semantic decoding. Channel fading and AWGN noises were present in the input data.

C. GANs, or Generative Adversarial Networks

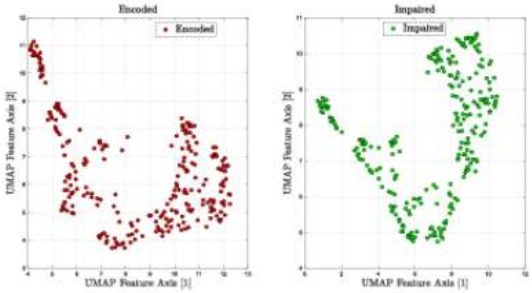
Because GANs can create rich visuals, they were used [15]. The generator, G, facilitates semantic reconstruction by mapping samples w from a fixed known distribution pw to an unknown joint distribution px|s. The discriminator D distinguishes between produced inputs (G(◦w,s),s) and

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 12 | Dec – 2025 DOI: 10.55041/ISJEM05262
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

genuine inputs (x,s). The following conditional functions control the architecture:

$$L_G = \mathbb{E}_{\hat{w}\sim p_{\hat{w}}}\left[-\log(D(G(\hat{w},s),s))\right], \quad (14)$$

$$L_D = -\mathbb{E}_{\hat{w}\sim p_{\hat{w}}}\log(1-D(G(\hat{w},s),s)) - \mathbb{E}_{x\sim p_{x|s}}\log D(x,s) \quad (15)$$



(a) Encoded (b) Impaired

Fig. 5: UMAP feature space visualization

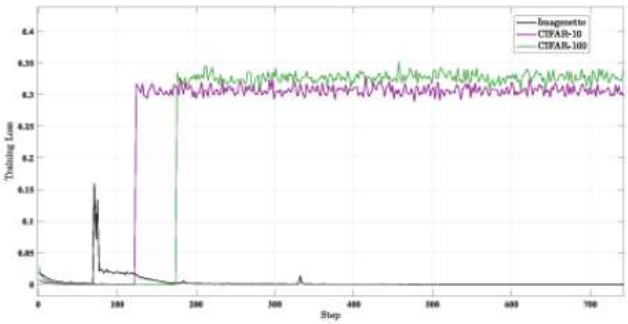


Fig. 6: Train-loss curves on proposed data sets

The distribution loss inherent to GANs is represented by the adversarial loss LGre. Furthermore, the distortion can be measured using the pictures' L1 losses:

$$\mathbb{E}[\|G(\hat{w},s)-x\|_1]. \quad (16)$$

Keywords: Vision Transformer (ViT) Architectur , Denoising Autoencoder Vision Transformer (DAE-ViT), Self-Attention Mechanism, Multi-Head Self-Attention (MHSA), Patch Embedding

IV. OUTCOMES

The results of studies utilizing ViTs, GANs, and CNNs on the ImageNette, Cifar-10, and Cifar-100 datasets under various fading and noise circumstances, such as Rayleigh, Rician, Nakagami-mfading, and AWGN, were provided in this part. The ViTbase model used in these tests has 12 transformer layers in both the encoder and decoder, a 16x16 patch size, and a 768 embedding dimension. The decoder featured 16 self-attention heads, whereas the encoder employed 12. Additionally, a Rayleigh channel model with a noise factor of 0.2 was included in the DAE-ViT architecture to simulate channel degradation. We evaluated the Structural Similarity Index (SSIM) and Peak Signal to Noise Ratio (PSNR) based on the training losses throughout several episodes. Figures 4a and 4b show the PSNR and SSIM findings, respectively, where the ViT outperformed the CNN and GAN models across all datasets, attaining a PSNR of about 38dB after 900 episodes. On the other hand, the CNN and GAN models had leveled off at about 10dB and 25dB, respectively. The hugegapin PSNR values demonstrate that the ViT fared better than the GAN and CNN models, demonstrating the ViT's superior capacity for semantic information reconstruction in images. The improved efficiency of the ViT is also supported by SSIM data.

TABLE II: Performance Metrics for Various Fading and Noise Conditions

	Rayleigh Fading	Rician Fading	Nakagami-m Fading	Additive White Gaussian Noise
Imagenette - Transformer	98.53	97.89	98.01	98.13
Cifar-10 - Transformer	95.23	97.16	97.83	97.63
Cifar-100 - Transformer	95.83	96.71	96.43	97.81
Imagenette - GAN	95.21	94.22	94.71	94.83
Cifar-10 - GAN	92.64	93.33	94.11	94.29
Cifar-100 - GAN	92.01	92.56	92.97	93.17
Imagenette - CNN	92.56	93.47	93.54	93.79
Cifar-10 - CNN	89.77	90.7	90.52	90.38
Cifar-100 - CNN	89.24	90.23	90.16	90.01

While the GAN and CNN models obtained lower SSIM values, the ViT attained an SSIM value close to 1.0, indicating that the fraction of structural similarity between the transmitted and recovered images was nearly perfect. The ViT's ability to successfully preserve semantic information even under demanding channel conditions is demonstrated by the Uniform Manifold Approximation and Projection (UMAP) feature space visualization, which clearly distinguishes between encoded and impaired features as shown in Figs. 5a and 5b, respectively. An overview of the ViT model's performance metrics under various fading and noise scenarios is shown in Table II. Compared to the CNN and GAN models, the ViT model is more reliable. On the ImageNette dataset, it obtained a PSNR of 98.53 dB under Rayleigh fading, while the CNN and GAN models obtained PSNRs of 89.77 dB and 95.21 dB, respectively. The Cifar-10 and Cifar-100 datasets showed comparable patterns. Figure 6 Compared to Imagenette, the loss curves for CIFAR-10 and CIFAR-100 show greater fluctuation, indicating dataset-specific variations that can affect training stability. While the GAN and CNN models struggled to retain comparable accuracy, particularly on the Cifar-100 dataset, the ViT base model performed remarkably well under Nakagami-m fading circumstances, reaching PSNR values close to or over 97 dB on all datasets. The output was rebuilt without noise, as seen in Fig. 7. Our findings show that the suggested ViT design uses 72% less bandwidth while achieving a higher PSNR of up to 38 dB and SSIM values near 1.

Keywords: AWGN Noise, UMAP Feature Visualization, Semantic Image Reconstruction, Bandwidth Reduction (72%), Training Stability, Channel Degradation Effect

V. FINAL THOUGHTS AND UPCOMING WORK

The use of ViTs in semantic communication is demonstrated in this paper, along with how it differs from more conventional models like CNNs and GANs. With a focus on noise robustness and efficiency under various noisy conditions, we present a unique ViT architecture in this study

that provides higher performance metrics, especially in terms of PSNR and SSIM. Additionally, the suggested ViT design uses 72% less bandwidth while achieving a higher PSNR of up to 38 dB and SSIM values near 1. We want to investigate hybrid models with optimization and noise reduction in the future. ViTs are excellent at semantic integrity, but their resource requirements prevent them from being used in contexts with limited resources. The deployment of edge devices can be made possible by optimization approaches.

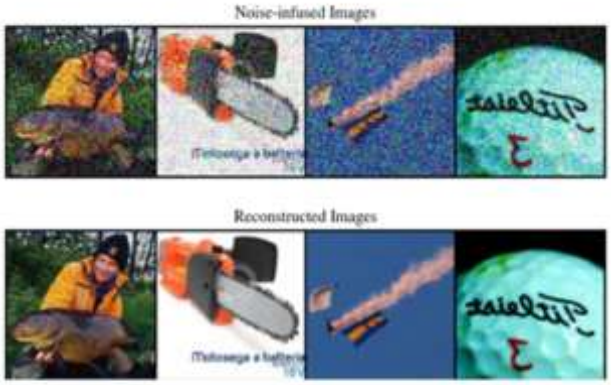


Fig. 7: Output of reconstructed image

Keywords: Vision Transformer (ViT), Semantic Communication, Noise Robustness, PSNR and SSIM Performance, Bandwidth Efficiency

REFERENCES:

[1] M. A. Jamshed, A. Kaushik, M. Dajer, A. Guidotti, F. Parzysz, E. Lagunas, M. Di Renzo, S. Chatzinotas, and O. A. Dobre, "Non-terrestrial networks for 6g: Integrated, intelligent and ubiquitous connectivity," arXiv preprint arXiv:2407.02184, 2024.

[2] M. Umer, M. A. Mohsin, A. Kaushik, Q.-U.-A. Nadeem, A. A. Nasir, and S. A. Hassan, "Reconfigurable intelligent surface-assisted aerial nonterrestrial networks: An intelligent synergy with deep reinforcement learning," IEEE Vehicular Technology Magazine, vol. 20, no. 1, pp. 55 64, 2025.

[3] F. N. Khan and A. P. T. Lau, "Robust and efficient data transmission over noisy communication channels using stacked and denoising autoen coders," China Communications, vol. 16, no. 8, pp. 72–82, 2019.

[4] S. Zebang and K. Sei-Ichiro, "Densely connected autoencoders for image compression," in Proceedings of the 2nd International Conference on Image and Graphics Processing, pp. 78–83, 2019.

[5] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," IEEE Transactions on Signal Processing, vol. 69, pp. 2663–2675, 2021.

[6] H. Saber, H. Hatami, and J. H. Bae, "List autoencoder: Towards deep learning based reliable transmission over noisy channels," in GLOBE COM 2022-2022 IEEE Global Communications Conference, pp. 1454 1459, IEEE, 2022.

[7] F. V. Dayana and V. C. Ernesto, "Analysis of bpsk modulation using the ni elvis iii communications module," in International Conference on Innovation and Research, pp. 33–49, Springer, 2021.

[8] M. Umer, M. A. Mohsin, S. A. Hassan, H. Jung, and H. Pervaiz, "Performance analysis of star-ris enhanced comp-noma multi-cell networks," in 2023 IEEE Globecom Workshops (GC Wkshps), pp. 2000–2005, 2023.

[9] M. Brain, C. Tinelli, P. Ruemmer, and T. Wahl, "An automatable formal semantics for ieee-754 floating-point arithmetic," in 2015 IEEE 22nd Symposium on Computer Arithmetic, pp. 160–167, 2015.

[10] A. Vaswani, "Attention is all you need," arXiv preprint arXiv:1706.03762, 2017.

[11] D. Alexey, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv: 2010.11929, 2020.

[12] B. Li, J. Liang, and J. Han, "Variable-rate deep image compression with vision transformers," IEEE Access, vol. 10, pp. 50323–50334, 2022.

[13] Y. Wu, "Generalized model to enable zero-shot imitation learning for versatile robots," 2024.

[14] K. Simonyan, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[15] M. U. Lokumarambage, V. S. S. Gowrisetty, H. Rezaei, T. Sivalingam, N. Rajatheva, and A. Fernando, "Wireless end-to-end image transmission system using semantic communications," IEEE Access, vol. 11, pp. 37149–37163, 2023.