Super-Resolution Semantic Communication System for Satellite Image

1st Tung Son Do
Chung-Ang University
Seoul 06974, South Korea
tsdo@uclab.re.kr

2nd Thanh Phung Truong
Chung-Ang University
Seoul 06974, South Korea
tptruong@uclab.re.kr

3rd Quang Tuan Do
Chung-Ang University
Seoul 06974, South Korea
dqtuan@uclab.re.kr

4th Dongwook Won
Chung-Ang University
Seoul 06974, South Korea
dwwon@uclab.re.kr

5th Ayalneh Bitew Wondmagegn
Chung-Ang University
Seoul 06974, South Korea
ayalneh@uclab.re.kr

6th Sungrae Cho
Chung-Ang University
Seoul 06974, South Korea
srcho@cau.ac.kr

***Abstract*—This paper introduces SR-DeepSC, a novel super-resolution semantic communication system to serve for satellite image transmission. We propose an asymmetric architecture utilizing Vision Mamba-inspired blocks, tailored for both satellite and ground station environments. SR-DeepSC significantly outperforms traditional methods like JPEG+Cubic in image quality, maintaining high SSIM and PSNR values across various SNR levels. The system's lightweight satellite-side implementation and more complex ground station processing effectively balance computational load, reducing latency while improving reconstruction quality. Experimental results demonstrate SR-DeepSC's robustness to noise and efficiency in bandwidth-limited conditions, making it a promising solution for enhanced satellite image communication.**

***Index Terms*—semantic communication, lightweight, 6G**

I. Introduction

Satellite image transmission plays a crucial role in Earth observation, meteorology, and remote sensing applications. However, the process faces significant challenges due to bandwidth limitations, atmospheric interference, and the vast distances involved. These constraints often result in the reception of low-resolution images that may lack critical details for accurate analysis and interpretation. To address this issue, super-resolution techniques have emerged as a promising solution. These methods aim to enhance the spatial resolution of received images, reconstructing high-fidelity versions from their low-resolution counterparts.

Semantic communication, a paradigm shift from traditional Shannon-based communication systems, offers a potential avenue for optimizing satellite image transmission. By focusing on the meaning and relevance of the transmitted information rather than raw data, semantic communication systems can prioritize the transmission of essential image features. This approach may reduce bandwidth requirements while preserving the semantic content crucial for downstream applications.

A. Related Works

In the field of semantic communication, significant strides have been made, particularly in text transmission. The DeepSC [1] model pioneered the use of Transformer-based architectures, optimizing both semantic and channel coding simultaneously. This was followed by LiteDeepSC [2], which offered a more resource-efficient solution for IoT applications. Jia *et al.* [3] further advanced the field with a lightweight JSCC scheme using a DeLighT-based neural network, achieving comparable or better communication reliability than Transformer-based models while substantially reducing computational demands. Their approach employs a DeLighT-based deep neural network model, achieving comparable or superior communication reliability to Transformer-based JSCC schemes while significantly reducing computational requirements and parameter count. SemVit, introduced by Yoo et al. [4], combined ViT and CNN architectures to enhance performance. Ren et al. [5] developed an asymmetric system for edge devices based on Diffusion models.

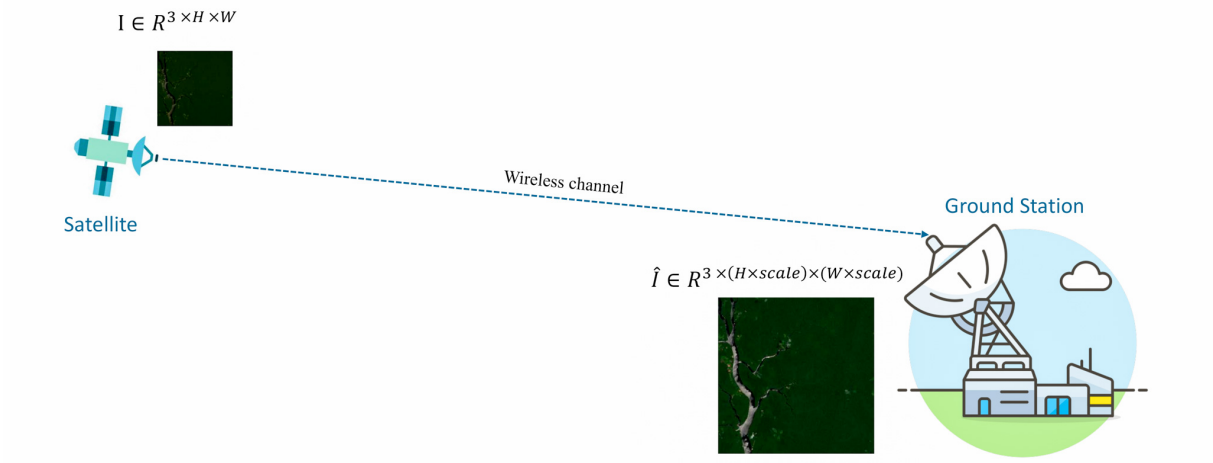


Fig. 1: Overview about Super-Resolution Satellite Image Transmission

Ye et al. proposed a robust codebook-based system utilizing vector-to-index transformers to combat noise effects. Zhang et al. [6] introduced a multi-server framework leveraging image-to-graph semantic similarity and multi-agent reinforcement learning for efficient resource allocation. In the realm of multimodal semantic communication, Do et al. [7] presented a Mamba-based multi-user multi-modal DeepSC to improve multi-modal data transmission efficiency. transmission.

B. Contributions

The main contributions of this research are:

- We propose SR-DeepSC, a novel super-resolution semantic communication architecture designed to generate high-resolution images from low-resolution inputs received over constrained wireless channels. This system addresses the unique challenges of satellite image transmission in bandwidth-limited environments.
- We introduce two innovative variants of Vision Mamba-inspired blocks, specifically tailored for satellite and ground station environments. These adaptations enhance the efficiency and effectiveness of image processing in the context of long-range, low-bandwidth satellite communications.

II. Proposed System

Fig. 1 illustrates the overview of the Super-Resolution Image Transmission system, wherein a satellite transmits low-resolution images to a ground station, which subsequently processes and outputs high-resolution images.

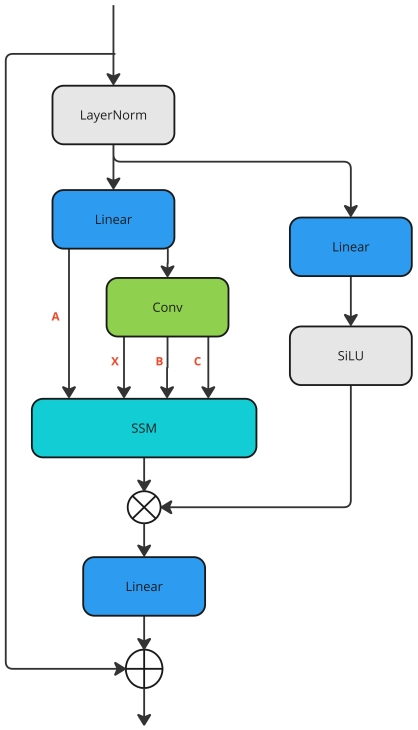


Fig. 2: The architecture of Vision Mamba-2 (VM-2)

A. Vision Mamba blocks

The core component of the proposed architecture is the Vision Mamba-2 (VM-2) block, which is fundamentally derived from the Mamba-2 module [8]. However, it incorporates modifications to the State Space Model (SSM) to accommodate 2D data processing. Fig. 2 depicts the structure of the Vision Mamba-2.

The operational sequence of the Vision Mamba can be described as follows:

- Layer Normalization: The input tensor $i \in \mathbb{R}^{L \times D}$ is normalized as $i_{\text{norm}} = \text{LayerNorm}(i)$, where L represents the sequence length and D the feature dimension.
- Input Projection: Two linear projections are applied to the normalized input

$$\begin{aligned} x &= \text{Linear}(i_{\text{norm}}, \theta_x) \in \mathbf{R}^{L \times (E \times D)} \\ z &= \text{Linear}(i_{\text{norm}}, \theta_z) \in \mathbf{R}^{L \times (E \times D)} \end{aligned} \tag{1}$$

  where θ_x and θ_z denote the weight matrices for the main and gating branch linear projections, respectively, and E represents the expansion factor of the Mamba module.
- Depthwise Convolution: A depthwise convolution is applied to the main branch:

$$x_c = Conv(x, \theta_{\text{Conv}}) \in \mathbf{R}^{L \times (E \times D)} \tag{2}$$

  where θ_{Conv} is the weight matrix of Convolutional layer
- State Space Model (SSM) Processing:

$$y = SSM_{A,B,C,\Delta}(x_c) \in \mathbf{R}^{L \times (E \times D)} \tag{3}$$

  where A, B, C, and Δ are the learnable parameters of the SSM.
- Gating Mechanism: The SSM output is modulated by the gating branch:

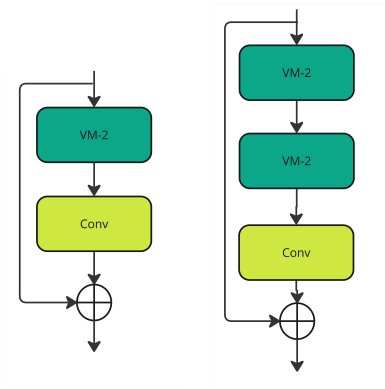
$$y_{\text{merged}} = y \cdot SiLU(z) \in \mathbf{R}^{L \times (E \times D)} \tag{4}$$

- Output Projection and Residual Connection: The final output is computed as:

$$out = Linear(y_{\text{merged}}, \theta_o) + i \in \mathbf{R}^{L \times D} \tag{5}$$

  where θ_o is the weight matrix of output projection layer.

Given the computational constraints of satellites, we propose two variants of the Vision Mamba-inspired block, tailored for both satellite and ground station applications: the Compact Residual Vision Mamba-inspired Block (CRVMB) and the Residual Vision Mamba-inspired Block (RVMB), respectively. Fig. 5 illustrates the architectures of these two variants. The primary distinction between the two variants lies in their structural complexity. The CRVMB incorporates a single Vision Mamba-2 (VM-2) block, whereas the RVMB employs two such blocks in sequence. Consequently, the CRVMB exhibits lower computational requirements compared to the RVMB, making it more suitable for satellite-based processing. The operations of two modules can express as:



(a) CRVMB (b) RVMB

Fig. 3: Architectures of two variants: Residual Vision Mamba-inspired Block (RVMB) for ground station side and Compact Residual Vision Mamba-inspired Block (CRVMB) for satellite side

$$\begin{aligned} \text{RVMB}(x) &= Conv(\text{VM-2}^2(x, \theta_{VM}), \theta_C) + x \\ \text{CRVMB}(x) &= Conv(\text{VM-2}(x, \theta_{VM}), \theta_C + x \end{aligned} \tag{6}$$

where θ_{VM}, θ_C represents for the weight matrix of Vision Mamba-2 and Convolutional layer, respectively.

B. Framework Architecture

The framework architecture, as illustrated in Fig. 4, comprises two primary components: the Satellite side and the Ground Station side. Given the computational constraints of satellite systems, we have designed the satellite-side processing to be as lightweight as possible. On the Satellite side, the process begins with a low-resolution input image $I_{LR} \in \mathbb{R}^{3 \times H \times W}$. This image undergoes initial processing through a convolutional layer followed by patch embedding to extract shallow features and divide the image into smaller patches:

$$P = \text{PatchEmbedding}(\text{Conv}(I_{LR}, \theta_{\text{shallow}}), \theta_{\text{PE}}) \tag{7}$$

where $P \in \mathbf{R}^{(H \times W) \times D}$ represents for image patches; H, W represents for image height and width, respectively; θ_{PE} is the weight matrix for patch embedding. Subsequently, these image patches are processed through two CRVMB blocks to extract semantic features:

$$\text{SF} = \text{CRVMB}^2(P, \theta_{\text{CRVMB}}) \tag{8}$$

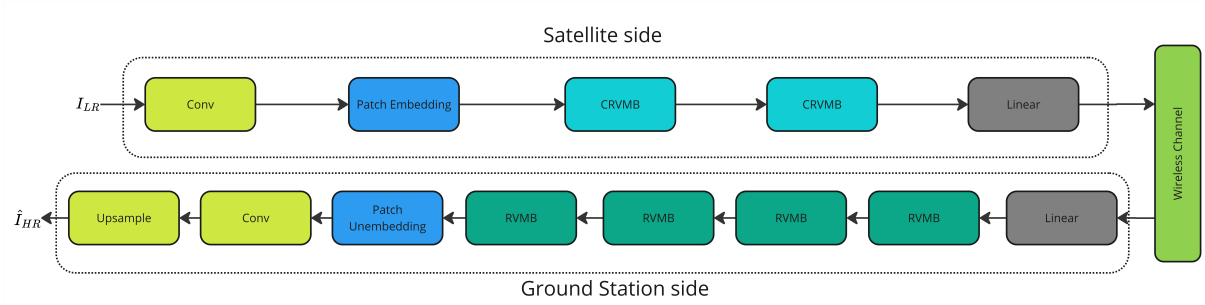


Fig. 4: The proposed framework architecture include the satellite side and the ground station side

where SF $\in \mathbf{R}^{(H\times W)\times D}$ represents for extracted semantic features; θ_{CRVMB} represents for the weight matrices of CRVMB blocks. The extracted features then undergo channel projection to transform them into regulated transmitting symbols:

$$Y = Linear(\text{SF}, \theta_{CE}) \tag{9}$$

where $Y \in \mathbf{R}^{(D\times W)\times CS}$ is channel encoded features; CS are the number of channel symbols; θ_{CE} is the weight matrix of channel encoder. The encoded symbols Y are transmitted through a wireless channel, modeled as an Additive White Gaussian Noise (AWGN) channel:

$$Z = Y + N \tag{10}$$

where $Z \in \mathbf{R}^{(D\times W)\times CS}$ is received feature at ground station side; N denotes for AWGN channel noise. On the Ground Station side, the received features Z are processed via the channel decoder:

$$\text{CDF} = Linear(Z, \theta_{\text{CD}}) \tag{11}$$

where CDF $\in \mathbf{R}^{D\times H\times W}$ represents for channel decoded features; θ_{CD} represents for the weight matrix of channel decoder projection. The channel-decoded features CDF are then processed through a series of four RVMB blocks to reconstruct the embedded patches:

$$\hat{P} = \text{RVMB}^4(\text{CDF}, \theta_{\text{RVMB}}) \tag{12}$$

where $\hat{P} \in \mathbf{R}^{D\times H\times W}$ represents for reconstructed embedded patches; θ_{RVMB} represents for weight matrix of the RVMB blocks. Finally, $\hat{P}$ is processed through a series of three convolutional layers for unembedding and upscaling:

$$\hat{I}_{HR} = Conv(Conv(Conv(\hat{P}, \theta_{\text{unembed}}), \theta_{\text{postconv}}), \theta_{up}) \tag{13}$$

where $\hat{I}\text{HR} \in \mathbb{R}^{3\times(H\cdot s)\times(W\cdot s)}$ represents the reconstructed high-resolution image, s denotes the upscale ratio, and θunembed, θ_{postconv}, and θ_{up} represent the weight matrices for unembedding, feature refinement, and upsampling, respectively.

C. Evaluation Metrics

To evaluate SR-DeepSC's efficacy in transmitting and reconstructing images, we utilize two complementary quantitative measures: the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index (SSIM). The PSNR metric quantifies image quality by comparing the maximum possible signal power to the power of corrupting noise. It is calculated as:

$$PSNR = 10 * log10(\frac{MAX_I^2}{MSE}) \tag{14}$$

Where MAX_I represents the maximum attainable pixel intensity, and MSE denotes the Mean Squared Error between the source and reconstructed images. Larger PSNR values correspond to superior image fidelity. The SSIM index assesses the visual similarity between two images by examining their luminance, contrast, and structural characteristics. Its formulation is:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{15}$$

Here, μ_x and μ_y denote the mean intensities, σ_x^2 and σ_y^2 represent the variances, and σxy is the covariance of image patches x and y. The SSIM score ranges from -1 to 1, with unity indicating perfect correspondence. While PSNR effectively gauges overall noise levels, SSIM provides insight into the preservation of structural information and correlates more closely with human visual perception. In combination, these metrics offer a holistic assessment of SR-DeepSC's

capacity to maintain image integrity across diverse channel conditions.

III. Simulation Results

A. Simulation Setup

This experiment is conducted using a system equipped with an Intel Core I7-14700 with 2.1GHz and an NVIDIA GeForce RTX 4070Ti Super with 16GB DRAM. Table I lists the other simulation setups. The adopted datasets is amalgamated from Landsat 8 and Sentinel-2, include 659 265x265 training images and 117 795x795 test images. To compare with SR-DeepSC, we adopt the JPEG+Cubic in which JPEG is for image source coding and Cubic to upscale the image.

TABLE I: Simulation Setups

Name	**Value**
Batch Size	4
Learning rate	1.00E-04
Training epochs	10
Optimizer	AdamW
Loss function	MSE
Upscale ratio	3
Hidden size	60
Patch size	1

B. Image results

The SSIM results, illustrated in Fig. 5a reveals stark differences between the two methods. SR-DeepSC maintains extremely high SSIM values across all SNR levels, indicating near-perfect structural similarity to the original image. In contrast, JPEG + Cubic shows very poor SSIM values, starting at 0.008888 at -6 dB SNR and improving to only 0.100829 at 18 dB SNR. While the proposed method's SSIM remains stable, JPEG + Cubic's SSIM improves with increasing SNR, but remains far inferior throughout the range of SNR values tested.

The PSNR results, illustrated in Fig. 5b further underscores the performance difference between the two methods. SR-DeepSC demonstrates consistently high PSNR values, ranging from 23.52 dB at -6 dB SNR to 26.53 dB at 18 dB SNR. In contrast, JPEG + Cubic shows much lower PSNR values, starting at 6.91 dB and improving to 17.62 dB as SNR increases. While both methods show improvement in PSNR as SNR increases, SR-DeepSC maintains a significant advantage throughout the entire SNR range, with a performance gap of over 8 dB even at the highest SNR level tested.

Through two above metrics we can see that SR-DeepSC exhibits strong robustness to noise, maintaining high performance even at low SNR levels. As a result, SR-DeepSC show the potential for satellite image transmission where the noise environment is extreme

TABLE II: The model size of SR-DeepSC and JPEG+Cubic in satellite and ground station side

Methods	**Satellite**	**Ground Station**
SR-DeepSC (proposed)	132,968	402,942
JPEG+Cubic	-	-

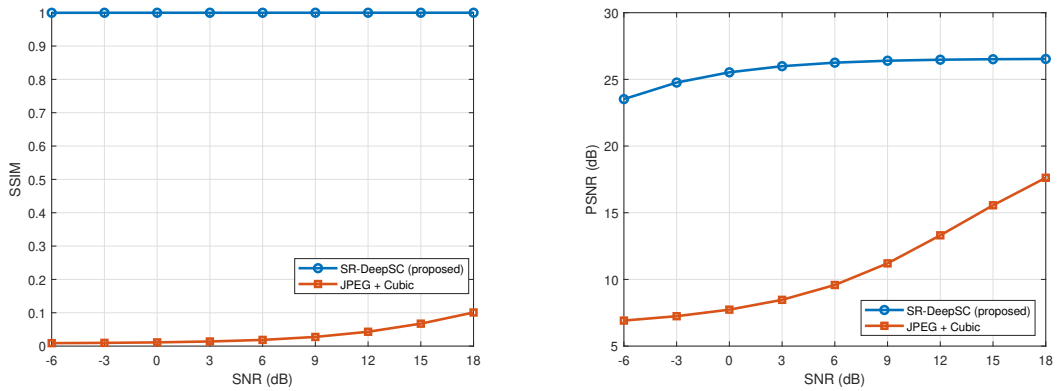
C. Model Size and Latencies

The Table II show the model size of SR-DeepSC and JPEG+Cubic in satellite and ground station side. The SR-DeepSC model employs an asymmetric architecture, which is particularly advantageous for satellite-to-ground communication systems. On the satellite side, the model has only 132,968 parameters, enabling efficient on-board processing within the limited, expensive computational resources of satellites. In contrast, the ground station component has 402,942 parameters, approximately three times larger, allowing for more complex processing and potentially better reconstruction of received signals. This asymmetric design intelligently distributes the computational load, placing the bulk of the processing burden on the ground station where resources are more abundant. JPEG+Cubic is a conventional method so it does not have any trainable parameters.

TABLE III: The mean latencies of SR-DeepSC, JPEG+Cubic method in satellite and ground Station side in millisecond (ms)

Methods	**Satellite**	**Ground Station**
SR-DeepSC (proposed)	3.033	10.308
JPEG+Cubic	13.372	13.218

The Table. III shows the comparsion between SR-DeepSC and JPEG+Cubic method. SR-DeepSC demonstrates superior performance on the satellite side, with a mean latency of only 3.033 ms compared to JPEG+Cubic's 13.372 ms. This substantial difference of over 10 ms is crucial in satellite applications where rapid processing is essential due to limited on-board resources and power constraints. On the ground station side, SR-DeepSC maintains its efficiency with



(a) SSIM (b) PSNR

Fig. 5: Results of the peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) for SR-DeepSC (proposed) and JPEG+Cubic

a latency of 10.308 ms, while JPEG+Cubic shows a similar performance to its satellite-side operation at 13.218 ms. The asymmetric nature of SR-DeepSC's architecture is evident in its latency distribution, with more processing time allocated to the ground station where computational resources are more abundant. Despite this, SR-DeepSC still outperforms JPEG+Cubic on the ground.

IV. Conclusion

In conclusion, SR-DeepSC presents a novel super-resolution semantic communication system for satellite image transmission. Leveraging asymmetric Vision Mamba-inspired blocks, it significantly outperforms traditional methods in image quality, computational efficiency, and latency. The system's robust performance across various SNR levels, coupled with its lightweight satellite-side implementation, makes it particularly well-suited for the challenging constraints of satellite communications. SR-DeepSC thus represents a promising advancement in efficient, high-quality satellite image transmission and reconstruction.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2024-00453301)

References

[1] 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.

[2] H. Xie and Z. Qin, "A Lite Distributed Semantic Communication System for Internet of Things," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 142–153, Jan. 2021.

[3] Y. Jia, Z. Huang, K. Luo, and W. Wen, "Lightweight Joint Source-Channel Coding for Semantic Communications," *IEEE Communications Letters*, vol. 27, no. 12, pp. 3161–3165, Dec. 2023.

[4] H. Yoo, L. Dai, S. Kim, and C.-B. Chae, "On the Role of ViT and CNN in Semantic Communications: Analysis and Prototype Validation," *IEEE Access*, vol. 11, pp. 71 528–71 541, 2023.

[5] T. Ren and H. Wu, "Asymmetric Semantic Communication System Based on Diffusion Model in IoT," in *2023 IEEE 23rd International Conference on Communication Technology (ICCT)*, Oct. 2023, pp. 1–6.

[6] W. Zhang, Y. Wang, M. Chen, T. Luo, and D. Niyato, "Optimization of Image Transmission in Cooperative Semantic Communication Networks," *IEEE Transactions on Wireless Communications*, vol. 23, no. 2, pp. 861–873, Feb. 2024.

[7] T. S. Do, T. P. Truong, T. Do, H. P. Van, and S. Cho, "Lightweight Multiuser Multimodal Semantic Communication System for Multimodal Large Language Model Communication."

[8] T. Dao and A. Gu, "Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality," May 2024.