ICT Express 11 (2025) 335–340

Contents lists available at ScienceDirect

ICT Express

journal homepage: www.elsevier.com/locate/icte

TranGDeepSC: Leveraging ViT knowledge in CNN-based semantic communication system

Tung Son Do, Thanh Phung Truong, Quang Tuan Do, Sungrae Cho *

School of Computer Science and Engineering, Chung-Ang University, Seoul 06974, South Korea

ARTICLE INFO

Keywords:
CNN
Lightweight
Semantic communication
6G

ABSTRACT

This paper introduces TranGDeepSC, a lightweight CNN-based deep semantic communication (DeepSC) system that leverages Vision Transformer (ViT) knowledge through co-training to enhance image transmission. Evaluated on CIFAR-100 across various SNRs, TranGDeepSC demonstrates competitive performance with ViTDeepSC, and outperforms SemViT and ADJSCC-V in image quality, particularly in low-SNR environments. Notably, it offers substantial gains in efficiency: 92.8% fewer parameters than ADJSCC-V, 72.0% lower energy use, and 48% faster processing than ViTDeepSC. These advantages make TranGDeepSC well-suited for resource-constrained applications in next-generation communication systems, including 6G, IoT, and real-time multimedia streaming.

1. Introduction

Semantic communication represents a paradigm shift from traditional bit-based transmission to meaning-focused data exchange. Conventional systems often struggle with inefficient bandwidth usage, poor performance in noisy environments, and inability to capture nuanced meaning [1–3]. To address these limitations, deep learning-based semantic communication (DeepSC) systems have emerged. These innovative approaches aim to bridge the gap between traditional communication theory and semantic understanding, offering more efficient and reliable data transmission in complex environments. By prioritizing essential information and adapting to context, DeepSC systems have the potential to revolutionize fields such as wireless communication, Internet of Things (IoT), and multimedia streaming [4].

1.1. Related works

Initial research in this field primarily concentrated on unimodal data transmission. In the realm of text transmission, groundbreaking work such as DeepSC [5] introduced Transformer-based architectures that simultaneously optimize semantic and channel coding. Building on this foundation, Lite-DeepSC [6] proposed a more lightweight approach to address resource constraints in Internet of Things devices. Further advancements in text semantic communication systems include the work of Jia et al. [7], who developed a lightweight joint source-channel coding (JSCC) scheme. 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.

Recent advancements in visual semantic communication have addressed various challenges in the field. Yoo et al. [8] introduced SemVit, integrating ViT and CNN architectures for enhanced performance, while Ren et al. [9] developed an asymmetric system based on Diffusion models for edge devices. Zhang et al. [10] addressed the bandwidth efficiency challenge by proposing a predictive and adaptive deep coding (PADC) framework that enables flexible code rate optimization. Their approach combines a variable code length DeepJSCC model with an Oracle Network for PSNR prediction, achieving minimal bandwidth consumption while maintaining quality constraints. Ye et al. [11] proposed a robust codebook-based system with vector-to-index transformers to mitigate noise effects. Zhang et al. [12] introduced a multi-server framework using image-to-graph semantic similarity and multi-agent RL for efficient resource allocation. Addressing semantic noise, Hu et al. [13] developed a framework incorporating adversarial training, masked VQ-VAE, and feature importance modules. Lyu et al. [14] proposed a multi-task JSCC framework supporting both image recovery and classification, with a gated design for channel adaptation. Fu et al. [15] introduced VQ-DeepSC, a digital system employing CNN-based transceivers and multi-scale semantic embedding for robust image transmission which improve semantic fidelity, noise resistance, multi-task capabilities, and resource utilization in semantic communication systems.

* Corresponding author.
E-mail addresses: tsdo@uclab.re.kr (T.S. Do), tptruong@uclab.re.kr (T.P. Truong), dqtuan@uclab.re.kr (Q.T. Do), srcho@cau.ac.kr (S. Cho).

https://doi.org/10.1016/j.icte.2025.02.010
Received 10 September 2024; Received in revised form 25 December 2024; Accepted 19 February 2025
Available online 8 March 2025
2405-9595/© 2025 The Authors. Published by Elsevier B.V. on behalf of The Korean Institute of Communications and Information Sciences. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

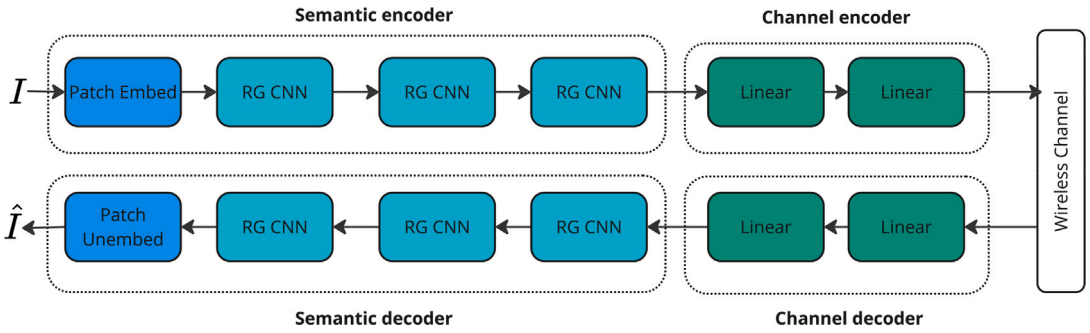


Fig. 1. The framework of the proposed TranGDeepSC includes semantic encoder, channel encoder, channel decoder and semantic decoder.

1.2. Motivations and contributions

Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) have emerged as powerful architectures for Deep Semantic Communication (DeepSC) systems, each with distinct strengths and limitations. CNNs excel at capturing local spatial patterns and hierarchical features, demonstrating robust performance even with limited datasets due to their spatial invariance and local representation learning. However, CNNs have inherent limitations that hinder their ability to capture long-range dependencies effectively. This is primarily due to their local receptive fields – the convolution kernels that only process small regions of the input data at a time. To capture distant relationships, information has to be gradually propagated through multiple convolutional layers and a very deep network is required. This process is computationally inefficient and makes it challenging to directly capture long-range interactions. Furthermore, while CNNs are effective at hierarchical feature learning, they lack an explicit mechanism to model long-range spatial relationships effectively. Conversely, ViTs leverage self-attention mechanisms to model global relationships and long-range dependencies within images, exhibiting superior performance on large datasets.

However, CNNs may struggle to capture long-range dependencies without significant depth, while ViTs are computationally intensive due to complex attention computations. Given these complementary characteristics, we propose a novel approach combining CNNs and ViTs strengths in a CNN-based DeepSC systems. The main contributions of this research are:

- We introduce TranGDeepSC, a novel CNN-based semantic communication architecture featuring global attention mechanisms. This lightweight design exhibits improved noise resilience while maintaining computational efficiency.
- We develop an innovative co-training algorithm that effectively transfers knowledge from ViT-based DeepSC to TranGDeepSC. This approach synergizes the local processing efficiency of CNNs with the global representation learning capabilities of ViTs, resulting in a more robust and versatile CNN-based model.
- We conduct comprehensive experiments demonstrating that TranGDeepSC achieves superior performance in terms of image reconstruction quality and computational efficiency compared to existing DeepSC models, particularly in challenging low-SNR environments.

The remainder of this paper is organized as follows: Section 2 presents a comprehensive overview of the proposed TranGDeepSC system, detailing its architecture and the novel co-training algorithm. Section 3 provides an in-depth analysis of the numerical results, demonstrating the performance of TranGDeepSC across various metrics and comparing it with existing state-of-the-art DeepSC systems. Finally, Section 4 concludes the paper by summarizing the key findings, discussing the implications of this research for semantic communication, and suggesting directions for future work in this field.

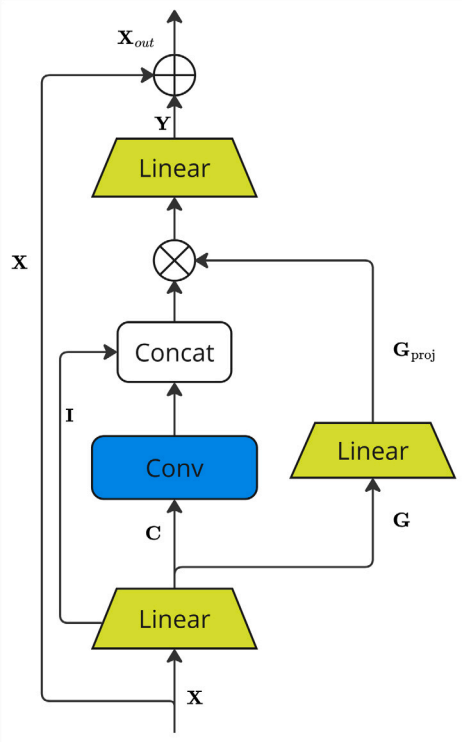


Fig. 2. The architecture of Residual Gated CNN (RG-CNN).

2. Proposed system

The section outlines the proposed TrangDeepSC system which incorporates joint semantic-channel encoder and decoder.

2.1. Residual Gated CNN

The Residual Gated CNN (RG-CNN), illustrated in Fig. 2 extends the Gated CNN [16] by incorporating residual connections. This architecture enhances feature extraction and information flow while improving robustness to noise. The RG-CNN block can be described by the following equations:

$$\begin{aligned}
[\mathbf{G}, \mathbf{I}, \mathbf{C}] &= \text{Split}(\text{FC}_1(\mathbf{X})) \\
\mathbf{C}_{\text{conv}} &= \text{Conv2D}(\mathbf{C}) \\
\mathbf{G}_{\text{proj}} &= FC_G(G) \\
\mathbf{Y} &= \text{FC}_2(\mathbf{G}_{\text{proj}} \odot [\mathbf{I}, \mathbf{C}_{\text{conv}}]) \\
\mathbf{X}_{\text{out}} &= \mathbf{X} + \mathbf{Y}
\end{aligned} \tag{1}$$

where $\mathbf{X}$ is the input tensor, FC_1, FC_2, FC_G are fully connected layers, Split divides the output of FC_1 into three parts: $\mathbf{G}$ (gate), $\mathbf{I}$ (identity),

and **C** (convolution), and Conv2D is a 2D convolution operation. The RG-CNN block is first passed through a fully connected layer and split into three parts. The gating mechanism (**G**) controls the flow of information, while the identity branch (**I**) allows for direct feature propagation. The convolution branch (**C**) applies a spatial convolution to capture local patterns. The gated output is then passed through another fully connected layer FC_G, and the result G_{proj} is added to the original input via a residual connection. This structure allows the network to adaptively combine spatial and channel-wise information, leading to more effective feature extraction and improved performance in various computer vision tasks.

2.2. Proposed TranGDeepSC system

The proposed TranGDeepSC system illustrated in Fig. 1. For quick reference, Table 1 comprehensively lists the primary notations applied throughout this subsection. For image transmission tasks, the TranGDeepSC processing pipeline comprises the following stages:

1. Patch Embedding: The input image $I \in \mathbb{R}^{H\times W\times C}$ is processed using patch embedding to create a sequence of embedded patches $E \in \mathbb{R}^{N\times D}$:

$$E = \text{PatchEmbed}(I) \tag{2}$$

where N represents the number of patches, and D denotes the embedding dimension.
2. Semantic encoding: The embedded patches pass through a series of Residual Gated CNN (RG CNN) blocks to produce semantic encoded features $F_{SE} \in \mathbb{R}^{N\times D}$:

$$F_{SE} = \text{RGCNN}_3(\text{RGCNN}_2(\text{RGCNN}_1(E))) \tag{3}$$

where RGCNN_i represents the ith Residual Gated CNN block.
3. Channel encoding: The semantic encoded features F_{SE} are transformed into regulated transmitting symbols $Z \in \mathbb{R}^K$ using a two-layer linear transformation::

$$Z = \text{Linear}_2(\text{Linear}_1(F_{SE})) \tag{4}$$

where K denotes the number of transmitted symbols, Linear_1 is the feature projection from $\mathbb{R}^D$ to $\mathbb{R}^{256}$ and Linear_2 acts as learned symbol modulator which maps the intermediate feature space $\mathbb{R}^{256}$ to the channel symbol space $\mathbb{R}^K$
4. Channel transmission: The encoded symbols Z are transmitted through a wireless channel, modeled as an Additive White Gaussian Noise (AWGN) channel:

$$Y = Z + N \tag{5}$$

where N represents the channel noise.
5. Channel decoding: At the receiver, the received features $Y \in \mathbb{R}^K$ are processed via the channel decoder:

$$F_{CD} = \text{Linear}_4(\text{Linear}_3(Y)) \tag{6}$$

where Linear_3 and Linear_4 are linear transformation layers, which implements the inverse operations of channel encoding to mapping received symbols to channel decoded features.
6. Semantic decoding: The channel decoded features F_{CD} pass through a series of RG CNN blocks to reconstruct the embedded patches $E' \in \mathbb{R}^{N\times D}$:

$$\hat{E} = \text{RGCNN}_3(\text{RGCNN}_2(\text{RGCNN}_1(F_{CD}))) \tag{7}$$

where RGCNN_i represents the ith Residual Gated CNN block in the decoder.
7. Patch Unembedding: The reconstructed embedded patches $\hat{E}$ are transformed back into the image domain $\hat{I} \in \mathbb{R}^{H\times W\times C}$:

$$\hat{I} = \text{PatchUnembed}(\hat{E}) \tag{8}$$

The TranGDeepSC framework integrates semantic encoding and decoding with channel coding to achieve robust image transmission over noisy channels.

Table 1
Table of proposed system's notations.

Symbol	Description
I, $\hat{I}$	Input image, Reconstructed image
H, W	Image height, Width
C	Number of image channels
E, E'	Embedded patches, Reconstructed patches
N, D	Number of patches, Embedding dimension
F_{SE}, F_{CD}	Semantic features, Channel decoded features
Z, Y	Transmitting symbols, Received features
K, N	Number of transmitted symbols, Channel noise

2.3. Evaluation metrics

To assess TrangDeepSC's performance in image transmission and reconstruction, we employ two complementary metrics: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM).

2.3.1. Peak Signal-to-Noise Ratio (PSNR)

PSNR quantifies the ratio between the maximum possible signal power and the power of distorting noise. It is defined as:

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{\text{MAX}_I^2}{\text{MSE}}\right) \tag{9}$$

where MAX_I is the maximum possible pixel value and MSE is the Mean Squared Error between the original and reconstructed images. Higher PSNR values indicate better image quality.

2.3.2. Structural Similarity Index (SSIM)

SSIM measures the similarity of two images in terms of luminance, contrast, and structure. It is computed as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{10}$$

where μ_x, μ_y are the averages, σ_x^2, σ_y^2 are the variances, and σ_{xy} is the covariance of image windows x and y. SSIM values range from -1 to 1, with 1 indicating perfect similarity. While PSNR provides a good measure of overall noise level, SSIM better captures the preservation of structural information and aligns more closely with human visual perception. Together, these metrics offer a comprehensive evaluation of TranGDeepSC's performance in maintaining image quality across various channel conditions.

2.4. Co-training algorithm

The TranGDeepSC framework employs an innovative co-training algorithm that leverages the strengths of both CNN-based and ViT-based models into TranGDeepSC. This algorithm aims to enhance the overall performance and robustness of the semantic communication system. To teach for TranGDeepSC, we initialized a ViT-based DeepSC which basically is TranGDeepSC but replace all RG CNN blocks to standard Transformer architecture.

The co-training process utilizes the Mean Squared Error (MSE) as the primary loss function. MSE is a widely used loss function in regression tasks and image reconstruction problems. For two tensors A and B of the same shape, the MSE is defined as:

$$\text{MSE}(A, B) = \frac{1}{n}\sum_{i=1}^{n}(A_i - B_i)^2 \tag{11}$$

where n is the total number of elements in each tensor. In our context, MSE measures the average squared difference between the pixel values of the original image and the reconstructed image, providing a measure of reconstruction quality. Lower MSE values indicate better reconstruction.

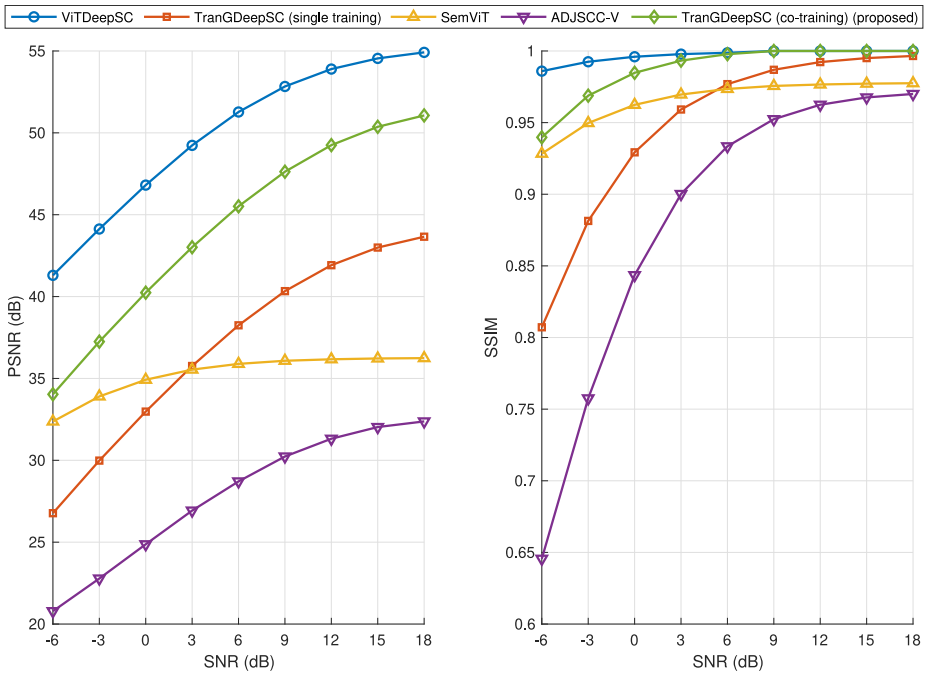


Fig. 3. Results of the peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) of TranGDeepSC versus comparison baselines.

Algorithm 1 TrangDeepSC Co-training Algorithm

Require: Training data X, TranGDeepSC Learning rate η_{TranG}, ViT-DeepSC Learning rate η_{ViT}, Number of epochs NE
Ensure: Trained TranGDeepSC
Initialize TranGDeepSC and ViTDeepSC models
for epoch = 1 to NE **do**
  for each minibatch x in X **do**
    $y_{TranG} \leftarrow \text{TranGDeepSC}(x)$
    $y_{ViT} \leftarrow \text{ViTDeepSC}(x)$
    $L_{TranG} \leftarrow \text{MSE}(y_{TranG}, x)$
    $L_{ViT} \leftarrow \text{MSE}(y_{ViT}, x)$
    $\theta_{ViT} \leftarrow \theta_{ViT} - \eta_{ViT} \nabla L_{ViT}$
    $\theta_{TranG} \leftarrow \theta_{TranG} - \eta_{TranG} \nabla L_{TranG}$
    $L_{disparity} \leftarrow \text{MSE}(y_{TranG}, y_{ViT})$
    $\theta_{TranG} \leftarrow \theta_{TranG} - \eta_{TranG} \nabla L_{disparity}$
  end for
end for
return Trained TranGDeepSC with optimized θ_{TranG}

The co-training process can be described as Algorithm 1.

The co-training algorithm leverages mutual learning and knowledge transfer between the CNN-based TranGDeepSC and the ViT-based DeepSC models. In each iteration, both models process the same input batch, but their parameter updates differ. TranGDeepSC is optimized to minimize both its own reconstruction error (L_{TranG}) and the disparity ($L_{disparity}$) between its output and that of ViTDeepSC. This disparity loss enables TranGDeepSC to learn from the global semantic features captured by the ViT architecture. ViTDeepSC, acting as a teacher model, is updated based solely on its own reconstruction error (L_{ViT}). This approach allows TranGDeepSC to benefit from the complementary strengths of both CNN and ViT architectures, potentially enhancing its performance and generalization in semantic communication tasks.

Table 2
Distribution of CIFAR-100 dataset.

Set	Total images	Images per class
Training	50,000	500
Testing	10,000	100

3. Numerical results

3.1. Dataset

In this study, we adopted CIFAR-100 dataset [17] which is a widely used computer vision dataset consisting of 60,000 32 × 32 color images in 100 classes. The Table 2 show the distributions of CIFAR-100 dataset. This standard split allows for consistent model evaluation and comparison across different machine learning approaches.

3.2. Simulation setup

This experiment is conducted using a system equipped with an Intel Core I7-14700 with 2.1 GHz and an NVIDIA GeForce RTX 4070Ti Super with 16 GB DRAM. Table 3 lists the other simulation setups

3.3. Comparison baselines

For comparison, the following models are considered

- **ViTDeepSC**: This model serves as the teacher in our co-training framework and represents a purely Vision Transformer-based approach to semantic communication. This baseline establishes the upper performance bound for semantic feature extraction and global dependency modeling in deep semantic communication systems.

Table 3
Simulation setups.

Parameter name	Value
Batch size	64
ViTDeepSC learning rate	1.00E−03
TranGDeepSC learning rate	1.00E−04
Optimizer	AdamW
Training SNR	12
Training epoch	60
Model hidden size	128
Number of transmitted symbols	16
Patch size	2

- **SemViT** [8]: A hybrid semantic communication system that integrates both Vision Transformer and CNN architectures to leverage their respective strengths. SemViT semantic encoder/decoder sections adopt 2 CNN layers followed by 1 ViT layers to take the advantages of two architectures. This baseline provides a direct comparison between architectural fusion and our proposed knowledge transfer approach in combining CNN and ViT capabilities.
- **ADJSCC-V** [10]: An adaptive deep joint source-channel coding framework that enables flexible code rate optimization based on channel conditions and image content. It uses a variable code length enabled DeepJSCC model combined with an Oracle Network for PSNR prediction and code rate optimization. This baseline evaluates the performance trade-offs between our fixed-rate transmission strategy and state-of-the-art adaptive rate optimization methods.
- **TranGDeepSC (single training)**: This variant of our proposed model is trained independently, without the co-training algorithm. This baseline demonstrates the performance impact of our co-training strategy through direct ablation comparison with conventional single-model training.

3.4. Image quality results

The Fig. 3 illustrates the image quality in PSNR and SSIM metrics. The proposed TranGDeepSC with co-training demonstrates remarkable performance improvements over its single training counterpart across all SNR levels, as evidenced by both PSNR and SSIM metrics. While ViTDeepSC remains the top performer in most scenarios, TranGDeepSC (co-training) shows competitive results, particularly at higher SNR levels. In terms of PSNR, although not surpassing ViTDeepSC, the proposed method exhibits substantial improvements over other approaches, including SemViT and the single-training version of TranGDeepSC. At 18 dB SNR, ViTDeepSC achieved 54.92 dB, while TranGDeepSC reached a competitive 51.06 dB, surpassing both SemViT's 36.24 dB and ADJSCC-V's 32.18 dB by significant margins. Notably, ADJSCC-V's adaptive rate strategy, while offering flexibility in bandwidth usage, demonstrates lower reconstruction quality compared to our fixed-rate approach across all SNR regimes. The SSIM results are particularly impressive, with TranGDeepSC (co-training) reaching convergence as quickly as ViTDeepSC at 6 dB SNR. At this critical SNR point of 6 dB, TranGDeepSC achieves an SSIM of 0.92, outperforming ADJSCC-V's 0.85 and demonstrating superior robustness in challenging channel conditions. In contrast, the version without co-training only approached convergence at 18 dB SNR, and SemViT maxed out at 0.9775 SSIM at 18 dB SNR. The performance gap between our proposed method and ADJSCC-V becomes particularly pronounced in the high SNR regime (15–18 dB), where TranGDeepSC maintains consistent high-quality reconstruction while ADJSCC-V shows limited improvement despite its rate adaptation capabilities. These results underscore the effectiveness of the co-training approach in TranGDeepSC, positioning it as a highly competitive method in the field.

Table 4
Model size and Energy consumption per image (in mJ) of models.

Model	Paramters	Energy consumption per image (mJ)
ViTDeepSC	1,300,508	147.676
SemViT	8,696,343	138.272
ADJSCC-V	12,757,579	310.382
TranGDeepSC (proposed)	908,047	86.816

Table 5
Processing latency versus methods (in ms).

Model	Mean ± Std	Min	Max	Median	P95
ViTDeepSC	3.1088 ± 0.9525	1.8461	9.2952	2.6907	5.5911
SemViT	4.2965 ± 1.5452	2.7361	13.9055	3.538	8.2072
ADJSCC-V	8.3387 ± 1.8763	6.6035	16.9727	7.5762	12.3093
TrangDeepSC (proposed)	1.6042 ± 0.4298	1.2732	3.8979	1.4476	2.3817

3.5. Models' size and energy consumption

TranGDeepSC demonstrates impressive efficiency in both model size and energy consumption. With only 908,047 parameters, it is significantly more compact than its competitors. This parameter count is approximately 30% smaller than ViTDeepSC (1,300,508 parameters), 92.8% smaller than ADJSCC-V (12,757,579 parameters), and dramatically less than SemViT (8,696,343 parameters). The compact nature of TranGDeepSC suggests it could be more suitable for deployment in resource-constrained environments or devices with limited memory. In terms of energy efficiency, TranGDeepSC also shows remarkable performance. It consumes only 86.816 mJ per image, which is substantially lower than both ViTDeepSC (147.676 mJ) and SemViT (138.272 mJ). Most notably, it achieves a 72.0% reduction in energy consumption compared to ADJSCC-V (310.382 mJ), despite ADJSCC-V's adaptive rate optimization capabilities. Specifically, TranGDeepSC uses about 41% less energy than ViTDeepSC and 37% less than SemViT per image processed. This significant reduction in energy consumption could translate to longer battery life in mobile devices or reduced operational costs in large-scale deployments. TranGDeepSC strikes an optimal balance between performance and efficiency, offering competitive or superior results despite its smaller size and lower energy consumption. This makes it particularly well-suited for applications with limited computational resources or energy constraints (see Table 4).

3.6. Models' latency analysis

TranGDeepSC demonstrates superior performance in terms of processing latency across all measured indicators. With a mean latency of 1.6042 ms (±0.4298), it significantly outpaces both ViTDeepSC (3.1088 ± 0.9525 ms) and SemViT (4.2965 ± 1.5452 ms). The proposed model achieves an 80.8% reduction in mean processing time compared to ADJSCC-V (8.3387 ± 1.8763 ms), demonstrating the computational efficiency of our fixed-rate approach over adaptive rate optimization strategies. This represents a reduction in average processing time of about 48% compared to ViTDeepSC and 63% compared to SemViT (see Table 5).

The proposed model also shows consistency in its performance. TranGDeepSC's minimum latency (1.2732 ms) is lower than the other models, and its maximum latency (3.8979 ms) is significantly less than ViTDeepSC (9.2952 ms), ADJSCC-V (16.9727 ms) and SemViT (13.9055 ms). This suggests that TranGDeepSC not only processes faster

on average but also maintains more stable performance under varying conditions.

The median latency for TranGDeepSC (1.4476 ms) further underscores its efficiency, being nearly half that of ViTDeepSC (2.6907 ms), less than half of SemViT (3.538 ms) and a fifth of ADJSCC-V (7.5762 ms). Additionally, the 95th percentile (P95) latency for TranGDeepSC is only 2.3817 ms, compared to 5.5911 ms for ViTDeepSC and 8.2072 ms for SemViT, with ADJSCC-V exhibiting the highest P95 latency at 12.3093 ms, highlighting our model's robust worst-case performance characteristics.

These latency metrics, combined with the previously discussed model size and energy efficiency, position TranGDeepSC as a highly efficient model. Its ability to process images quickly and consistently makes it particularly suitable for real-time applications or scenarios where rapid response times are crucial.

4. Conclusion

We presented TranGDeepSC, a CNN-based semantic communication system that effectively incorporates ViT strengths through co-training. Our approach demonstrates marked improvements in image transmission over noisy channels, showing competitive performance with ViT-based methods across all SNR levels. Compared to current state-of-the-art approaches, TranGDeepSC achieves remarkable efficiency gains with 92.8% fewer parameters and 72.0% lower energy consumption than ADJSCC-V, while offering 48% faster processing than ViTDeepSC, highlighting its potential for resource-constrained and real-time applications. Future work could explore multi-modal applications and further optimizations of the co-training process.

CRediT authorship contribution statement

Tung Son Do: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Thanh Phung Truong:** Writing – review & editing, Writing – original draft, Visualization. **Quang Tuan Do:** Writing – review & editing, Supervision, Software, Data curation, Conceptualization. **Sungrae Cho:** Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the IITP (Institute of Information & Coummunications Technology Planning & Evaluation) - ITRC (Information Technology Research Center) (IITP-2025-RS-2024-00436887, 50%) grant funded by the Korea government (Ministry of Science and ICT), in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00453301), and in part by the Chung-Ang University Young Scientist Scholarship in 2024.

References

[1] M.C. Ho, A.T. Tran, D. Lee, J. Paek, W. Noh, S. Cho, A DDPG-based energy efficient federated learning algorithm with SWIPT and MC-NOMA, ICT Express 10 (3) (2024) 600–607, http://dx.doi.org/10.1016/j.icte.2023.12.001.

[2] J. Oh, D. Lee, D.S. Lakew, S. Cho, DACODE: Distributed adaptive communication framework for energy efficient industrial IoT-based heterogeneous WSN, ICT Express 9 (6) (2023) 1085–1094, http://dx.doi.org/10.1016/j.icte.2023.02.009.

[3] T.T.H. Pham, W. Noh, S. Cho, Multi-agent reinforcement learning based optimal energy sensing threshold control in distributed cognitive radio networks with directional antenna, ICT Express 10 (3) (2024) 472–478, http://dx.doi.org/10.1016/j.icte.2024.01.001.

[4] T.S. Do, T.P. Truong, T. Do, H.P. Van, S. Cho, Lightweight multiuser multimodal semantic communication system for multimodal large language model communication. http://dx.doi.org/10.22541/au.172479430.09168922/v1.

[5] H. Xie, Z. Qin, G.Y. Li, B.-H. Juang, Deep learning enabled semantic communication systems, IEEE Trans. Signal Process. 69 (2021) 2663–2675, http://dx.doi.org/10.1109/TSP.2021.3071210.

[6] H. Xie, Z. Qin, A lite distributed semantic communication system for Internet of Things, IEEE J. Sel. Areas Commun. 39 (1) (2021) 142–153, http://dx.doi.org/10.1109/JSAC.2020.3036968.

[7] Y. Jia, Z. Huang, K. Luo, W. Wen, Lightweight joint source-channel coding for semantic communications, IEEE Commun. Lett. 27 (12) (2023) 3161–3165, http://dx.doi.org/10.1109/LCOMM.2023.3329533.

[8] H. Yoo, L. Dai, S. Kim, C.-B. Chae, On the role of ViT and CNN in semantic communications: analysis and prototype validation, IEEE Access 11 (2023) 71528–71541, http://dx.doi.org/10.1109/ACCESS.2023.3291405.

[9] T. Ren, H. Wu, Asymmetric semantic communication system based on diffusion model in IoT, in: 2023 IEEE 23rd International Conference on Communication Technology, ICCT, 2023, pp. 1–6, http://dx.doi.org/10.1109/ICCT59356.2023.10419809.

[10] W. Zhang, H. Zhang, H. Ma, H. Shao, N. Wang, V.C.M. Leung, Predictive and adaptive deep coding for wireless image transmission in semantic communication, IEEE Trans. Wirel. Commun. 22 (8) (2023) 5486–5501, http://dx.doi.org/10.1109/TWC.2023.3234408.

[11] P. Ye, Y. Sun, S. Yao, H. Chen, X. Xu, S. Cui, Codebook-enabled generative end-to-end semantic communication powered by transformer, in: IEEE INFOCOM 2024 - IEEE Conference on Computer Communications Workshops, INFOCOM WKSHPS, 2024, pp. 1–6, http://dx.doi.org/10.1109/INFOCOMWKSHPS61880.2024.10620755.

[12] W. Zhang, Y. Wang, M. Chen, T. Luo, D. Niyato, Optimization of image transmission in cooperative semantic communication networks, IEEE Trans. Wirel. Commun. 23 (2) (2024) 861–873, http://dx.doi.org/10.1109/TWC.2023.3282906.

[13] Q. Hu, G. Zhang, Z. Qin, Y. Cai, G. Yu, G.Y. Li, Robust semantic communications with masked VQ-VAE enabled codebook, IEEE Trans. Wirel. Commun. 22 (12) (2023) 8707–8722, http://dx.doi.org/10.1109/TWC.2023.3265201.

[14] Z. Lyu, G. Zhu, J. Xu, B. Ai, S. Cui, Semantic communications for image recovery and classification via deep joint source and channel coding, IEEE Trans. Wirel. Commun. (2024) http://dx.doi.org/10.1109/TWC.2023.3349330, 1–1.

[15] Q. Fu, H. Xie, Z. Qin, G. Slabaugh, X. Tao, Vector quantized semantic communication system, IEEE Wirel. Commun. Lett. 12 (6) (2023) 982–986, http://dx.doi.org/10.1109/LWC.2023.3255221.

[16] Y.N. Dauphin, A. Fan, M. Auli, D. Grangier, Language modeling with gated convolutional networks, 2017, http://dx.doi.org/10.48550/arXiv.1612.08083, arXiv:1612.08083.

[17] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images, 2009.