This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2026.3660029

IEEE Open Journal of the Communications Society

Received XX Month, XXXX; revised XX Month, XXXX; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version 11 January, 2024.

Digital Object Identifier 10.1109/OJCOMS.2024.011100

Generative AI-Enhanced Robust Semantic Communication Architecture for UAV Image Transmission

Canpu Liu[1], Li Zhou [1], Xinfeng Deng [1], Yichi Zhang [1] , Nan Li [1], Jun Xiong [1], AND Boon-Chong Seet [2].

[1]College of Electronic Science and Engineering, National University of Defense Technology, Changsha, China
[2]Department of Electrical and Electronic Engineering, Auckland University of Technology, Auckland, New Zealand

CORRESPONDING AUTHORS: L. Zhou AND N. Li (e-mail: zhouli2035@nudt.edu.cn; li.nan@nudt.edu.cn).

ABSTRACT Unmanned aerial vehicle (UAV) wireless image transmission has gained widespread application across various fields due to its flexibility, yet it faces critical challenges such as resource constraints and degradation of reconstruction quality caused by harsh channel conditions. To address these issues, we designed a lightweight semantic communication backbone network that substantially reduces the computational and storage overhead of UAVs through codebook assistance and efficient encoder-decoder design. On this basis, to tackle severe image degradation under adverse channel conditions, we introduced a generative artificial intelligence-based (GAI) enhancement module. Specifically, we developed a semantic refinement network (SRN) that employs an innovative signal-to-noise ratio (SNR) adaptive feature-wise linear modulation (FiLM) layer to dynamically adjust its refinement strategy based on real-time channel quality, fundamentally transforming the image reconstruction paradigm from traditional signal recovery to conditional content generation. Extensive experimental results demonstrate that our proposed framework significantly outperforms the current state-of-the-art method under extreme channel conditions, highlighting its great potential for achieving robust UAV image transmission in challenging operational environments.

INDEX TERMS Semantic communication, Generative AI, UAV image transmission, lightweight model.

I. INTRODUCTION

A. Background and motivations

UNMANNED aerial vehicles (UAVs) have garnered significant attention due to their flexible deployment, low cost, and high adaptability [1]. Equipped with visual sensors, UAVs can collect image data in aerial domains that are difficult for humans to access directly. These data are then transmitted to the ground receiver to perform tasks such as image reconstruction and object detection, enabling widespread applications in post-disaster rescue, agricultural monitoring, military reconnaissance, and urban planning [2]–[4].

However, as edge mobile device, UAVs face severe challenges in complex tasks due to their inherent resource constraints. Their limited computing power and storage capacity struggle to support large-scale data transmission and processing [5]. Fortunately, the emergence of semantic communication offers a novel solution to this issue. Unlike conventional communication methods, semantic communication represents a new paradigm that focuses on the meaning of information, prioritizing the transmission of underlying significance of the data rather than the exact bitstream [6]. This approach shows great potential in enhancing compression efficiency and robustness of resource-constrained UAVs.

Although existing semantic communication systems demonstrate promising performance in data compression, their robustness in extreme environments remains notably inadequate, particularly for UAV applications with dynamic environmental changes and poor channel conditions [7]. In these challenging environments, the detailed parts of semantic information are highly susceptible to noise contamination, leading to structural artifacts in decoded images and even the loss of critical semantic features at the receiver. Such degradation in semantic integrity may have severe consequences for tasks that rely on high-precision semantic interpretation.

To overcome this bottleneck, we attempt to shift from simple signal recovery to the research of generative image reconstruction. This paradigm not only aims to denoise corrupted signals but also leverages the powerful capabilities

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2026.3660029

of advanced generative artificial intelligence (GAI) to generate high-fidelity and semantically coherent images, thereby enhancing the robustness of the system under extreme conditions [8]. The generative framework particularly addresses the critical challenge of semantic feature preservation during transmission, as it can intelligently reconstruct missing or corrupted visual elements while maintaining contextual integrity through learned prior knowledge of image semantics.

B. Related Work

Semantic Communication for UAVs: For UAV image transmission, conventional communication approaches generally suffer from two fundamental limitations of high resource consumption and insufficient environmental adaptability, making it difficult to meet the inherent resource constraints of UAVs and the increasingly complex mission requirements [9]. With the emergence of semantic communication, UAV image transmission is progressively evolving from simple data transmission to more sophisticated, semantic-aware, and intelligent applications.

For task-oriented UAV image transmission, Kang *et al.* [10] proposed a deep reinforcement learning-based semantic compression and channel-aware technique. This method offloads partial computational tasks to Mobile Edge Computing (MEC) servers, enabling collaborative computation between front-end UAVs and back-end servers to alleviate onboard processing burdens. The semantic communication framework based on federated learning proposed by Xie *et al.* [11] offers an efficient solution for bandwidth-sensitive distributed image transmission. To address the low efficiency of image transmission caused by power limitations in UAV networks, Yao *et al.* [12] designed a Soft Actor-Critic reinforcement learning algorithm with Entropy Maximization (SAC-EM) to optimize the selection of semantic information, achieving convergence speed of 80% faster than conventional communication approaches. However, this algorithm requires extensive training data, which increases hardware costs for real-time deployment on UAVs. To further compress the volume of transmitted data, Song *et al.* [13] introduced a cognitive semantic communication system based on a knowledge graph. By incorporating a multi-scale encoder-decoder architecture and a signal-to-noise ratio (SNR) adaptive module, their approach reduces transmitted data while simultaneously enhancing model robustness under dynamic channel conditions.

In search and rescue missions with stringent real-time requirements, Papi *et al.* [14] significantly reduced image transmission latency between the UAV and the ground station by employing a dual-channel datalink and directional antenna system. To address the challenge of large-scale forest safety monitoring, Jiao *et al.* [15] proposed a real-time forest fire detection method by integrating deep learning into the UAV platform. In military reconnaissance tasks, Alexan *et al.* [16] addressed the high-security transmission of surveillance images through a novel 2-layer image encryption and transmission scheme. Industrial applications such as power line inspection [17], [18] further expanded the applicability boundaries of UAV image transmission.

While existing semantic communication systems have demonstrated superior performance over conventional methods in UAV image transmission, most approaches rely on neural network-based structures that operate as a "black box," lacking theoretical foundations and interpretability. To address this issue, the authors in [19] proposed a novel text-based semantic communication system employing a shared knowledge base, introducing definitions of semantic self-information and source entropy to enhance theoretical reliability. Similarly, Zhang *et al.* [20] extended this design principle to image transmission by developing a shared semantic codebook-based encoding method. This approach constructs a Semantic-Aware Codebook (SAC) through Weighted Data-Semantic distance (WDS) to guide the processing of the semantic encoder and decoder, significantly improving image reconstruction quality and classification accuracy. However, the semantic encoder-decoder employed in this method involves substantial parameter storage and computational resources, which presents a significant challenge for practical deployment on UAVs. Consequently, how to further balance the efficiency and explainability, while concurrently exploring lightweight models to accommodate the hardware constraints of UAVs, remains an open research question.

Generative AI in Semantic Communications: Recent advances in semantic communication systems have increasingly focused on integrating GAI to address the limitations of deep learning-based approaches, particularly their poor generalization, robustness, and reasoning capabilities [21]. For example, Zhou *et al.* [22], [23] have explored the use of generative pre-trained Transformers for semantic information extraction and have investigated the application of GAI in optimizing radio resource distribution in future mobile networks. Current generative semantic communication frameworks predominantly employ classical generative models, including variational autoencoders (VAEs), generative adversarial networks (GANs), and diffusion models (DMs), aiming to improve communication efficiency and task adaptability through efficient semantic encoding and generative reconstruction.

Ye *et al.* [24] proposed a codebook-based image semantic communication system where the transmitter uses a vector quantized VAE (VQ-VAE) structure to discretize continuous semantic features into codebook indices, while the receiver employs a VQ-GAN decoder to reconstruct images, effectively mitigating the interference of channel noise on feature mapping. Meanwhile, Ye *et al.* [25] introduced an ultra-low-bitrate semantic communication system using a conditional generative model. At the receiver, a pre-trained conditional DM utilizes text descriptions and saliency map indices as guidance to generate high-fidelity images, significantly improving robustness against noise and inaccuracies in saliency detection. Facing the characteristics of low transmission

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2026.3660029

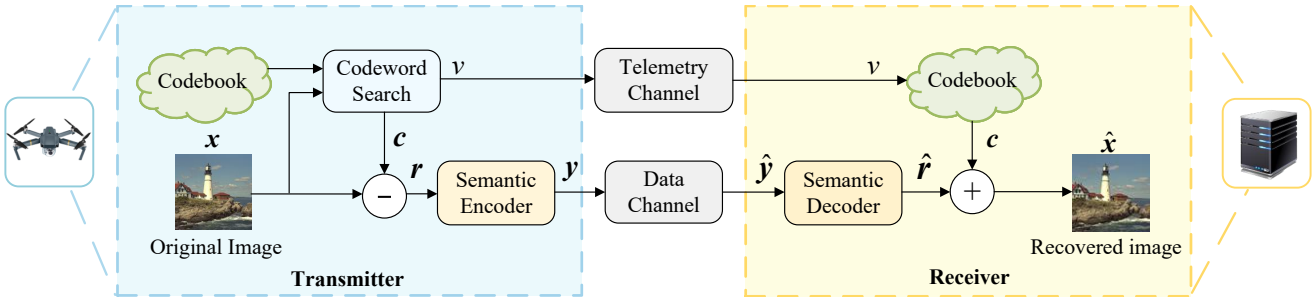


FIGURE 1. **Architecture of the lightweight codebook-assisted semantic transmission backbone.**

efficiency and unstable channels in vehicle networks, Lu *et al.* [26] introduced a DM as the semantic decoder. Through its powerful prior-based generation ability, high-fidelity semantic reconstruction and prediction in a noisy environment are achieved.

Similar to the application of GAI in vehicle networks, deeply integrating generative models into the semantic communication framework to achieve reliable UAV image transmission is a challenging problem to be solved in this paper.

C. Main Contributions

Building upon these developments, this study proposes a novel GAI-enhanced semantic communication framework for reliable UAV-to-ground server image transmission, especially in challenging channel environments. The key contributions of this work are summarized as follows:

- Considering the computational and storage constraints of UAVs, we adopt a dual-level semantic transmission strategy. We use a pre-constructed codebook, where the codeword index corresponding to the original image serves as the coarse-grained semantic information, while the residual information beyond codewords is processed through a semantic encoder as fine-grained semantic information. Notably, we specifically design a lightweight semantic encoder-decoder based on the MobileViTv3 network to satisfy the edge deployment requirements of UAVs.
- We propose a decoupled control-data transmission (DCDT) mechanism. As prior information, the codeword holds a higher priority than the fine-grained semantic information. We therefore regard it as control information and transmit through the UAV's telemetry channel, while the fine-grained semantic are conveyed over the data channel. The dual-level semantic information then fused at the receiver to obtain the preliminary reconstructed image.
- To compensate for the performance limitations of the lightweight semantic encoder-decoder and potential detail loss during the mapping process from the original image to the codeword index, we introduce a generative model at the receiver. This transforms the reconstruction task from simple signal fusion to conditional image generation, significantly enhancing system robustness under adverse channel conditions. Additionally, the learned perceptual image patch similarity (LPIPS) metric is incorporated to further evaluate the improvement of images in terms of visual perception quality.

The remainder of this paper is organized as follows. Section II details our proposed lightweight semantic communication backbone, including its dual-level transmission architecture and the efficient MobileViTv3-based encoder-decoder design. Section III introduces our core innovation, the GAI-enhanced robust reconstruction framework, providing a deep dive into the architecture of the semantic refinement network and its channel-adaptive mechanism. Section IV provides a comprehensive experimental evaluation, where we validate the superiority of our proposed framework and conduct ablation studies to analyze its key components. Finally, Section V concludes the paper and discusses potential future work.

II. LIGHTWEIGHT CODEBOOK-ASSISTED SEMANTIC TRANSMISSION BACKBONE

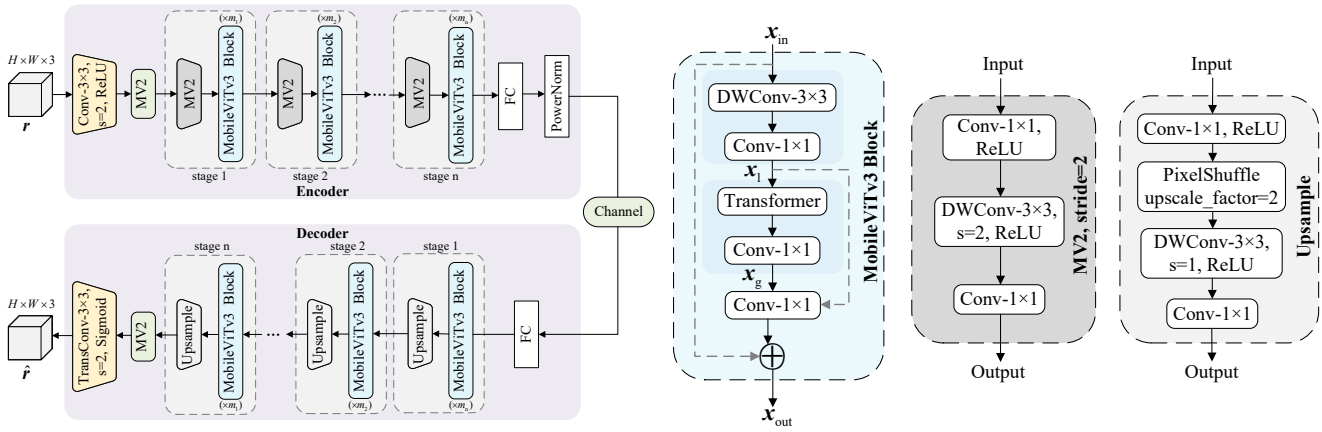
A. Overall Architecture

The proposed lightweight codebook-assisted semantic communication architecture for UAV-to-server image transmission is illustrated in Figure 1. In this system, the UAV functions as the transmitter responsible for image acquisition and encoding, while the ground server acts as the receiver tasked with decoding and reconstructing the original image, with wireless channel transmission established between these two endpoints.

Considering the dynamic environment and resource constraints of the UAV, we employ a dual-level semantic transmission strategy. Specifically, we adopt an offline codebook construction approach, where the UAV obtains a shared codebook $\mathcal{C} = \{\boldsymbol{c_1}, \boldsymbol{c_2}, \cdots, \boldsymbol{c_n}, \}$ of size N preconstructed by the server prior to deployment [20]. During operation, the UAV captures images via its sensors, and utilizes a codeword search module to find the best-matching codeword $\boldsymbol{c} \in \mathbb{R}^{H \times W \times 3}$ and its corresponding index $v \in \{1, 2, \cdots, N\}$ from the shared codebook $\mathcal{C}$, where H and W represent the height and width of the image respectively. Notably, since both transmitter and receiver share the same codebook, only the codeword index v needs to be transmitted to retrieve the corresponding codeword at the receiver. This codeword

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2026.3660029



(a) (b)

FIGURE 2. (a) The proposed model of lightweight semantic encoder-decoder based on MobileViTv3 network. (b) The detailed structure of the main modules in the semantic encoder, including MobileViTv3 block, MV2 block for downsampling, and Upsample block.

serves as coarse-grained semantic information, aiding the image reconstruction process. Furthermore, we compute the residual information $\boldsymbol{r} \in \mathbb{R}^{H\times W\times 3}$ by differencing the original image $\boldsymbol{x} \in \mathbb{R}^{H\times W\times 3}$ with the retrieved codeword $\boldsymbol{c}$, which is then processed by our semantic encoder to generate fine-grained semantic information. Finally, to ensure the reliable reception of the codeword index serving as prior, we propose a decoupled transmission mechanism that uses two independent channels to transmit the dual-level semantic information respectively.

B. MobileViT-Based Semantic Encoder-Decoder

Figure 2(a) illustrates the structure of the lightweight semantic encoder-decoder based on MobileViTv3 network. Specifically, we construct a hierarchical architecture that enables the model to learn semantic features ranging from low-level to high-level. The encoder consists of a 3×3 convolution block, an MobileNetV2 (MV2) block, and several stages, which can gradually compress the image while extracting semantic features.

The encoder receives the residual information r of the original image, performs the first downsampling and feature extraction through a 3×3 convolution and an MV2 block, yielding the preliminary feature map as

$$\boldsymbol{y_0} = \mathrm{MV2}_{s=1}(\mathrm{ReLU}(\mathrm{Conv}_{3\times 3,s=2}(\boldsymbol{r}))). \tag{1}$$

The feature map $\boldsymbol{y_0}$ is then fed into the n-stage encoder backbone. In each stage, the feature map is first downsampled by an MV2 block with a stride of 2 to reduce its spatial resolution. As illustrated in Figure 2(b), the downsampling MV2 block employs depthwise separable convolution (DWConv) for efficient local feature extraction, demonstrating significantly lower computational cost compared to standard convolution. These feature maps are then processed by several MobileViTv3 blocks to learn the global dependencies among them, thereby extracting higher-level semantic information from the image. This process can be represented as

$$\boldsymbol{y_i} = (\mathrm{MViT}^{(i)} \times m_i)(\mathrm{MV2}_{s=2}{}^{(i)}(\boldsymbol{y_{i-1}})), \tag{2}$$

where $\boldsymbol{y_i}$ and $\boldsymbol{y_{i-1}}$ represent the output and input feature maps at the i-th stage respectively, and m_i represents the number of MobileViTv3 blocks at stage i.

The detailed structure of the MobileViTv3 block is presented in Figure 2(b), which serves as the core component in the lightweight semantic encoder-decoder for efficiently learning both global representations and local details. It replaces the computationally expensive global self-attention mechanism in the standard ViT by innovatively combining the local feature extraction ability of CNNs with the ability of Transformers to capture long-range dependencies. The mathematical representation of the processing procedure in the MobileViTv3 block is:

$$\boldsymbol{x}_\mathrm{l} = \mathrm{Conv}_{1\times 1}(\mathrm{DWConv}_{3\times 3}(\boldsymbol{x}_\mathrm{in})), \tag{3a}$$

$$\boldsymbol{x}_\mathrm{g} = \mathrm{Conv}_{1\times 1}(\mathrm{LGL}_\mathrm{Attention}(\boldsymbol{x}_\mathrm{l})), \tag{3b}$$

$$\boldsymbol{x}_\mathrm{out} = \mathrm{Conv}_{1\times 1}([\boldsymbol{x}_\mathrm{g}; \boldsymbol{x}_\mathrm{l}]) + \boldsymbol{x}_\mathrm{in}, \tag{3c}$$

where $\boldsymbol{x}_\mathrm{in}$, $\boldsymbol{x}_\mathrm{out}$, $\boldsymbol{x}_\mathrm{l}$ and $\boldsymbol{x}_\mathrm{g}$ denote the input, output, local and global feature map of the MobileViTv3 block, respectively. $\mathrm{LGL}_\mathrm{Attention}(\cdot)$ refers to the local-global-local attention mechanism of the Transformer in MobileViTv3 block, and $[\boldsymbol{x}_\mathrm{g}; \boldsymbol{x}_\mathrm{l}]$ indicates the concatenation of these two features.

It is noteworthy that the number of encoder stages n, and the quantity of MobileViTv3 blocks within each stage m_i, are flexibly adjusted according to the resolution of the original image. Typically, higher-resolution images require more stages.

In the final stage of the encoder, a FC layer maps the high-level feature map $\boldsymbol{y_n}$ into a latent semantic representation $\boldsymbol{y}$ with dimension of K, which is then power-normalized and transmitted over the wireless channel. The compression ratio of the image during this progress is calculated as

$$CR = \frac{K}{H \times W \times 3}. \tag{4}$$

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2026.3660029

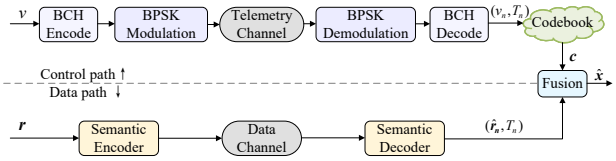


FIGURE 3. The architecture of proposed DCDT mechanism.

Within the decoder, the Upsample block is designed as the inverse operation of the MV2 downsampling block in encoder, as shown in Figure 2(b). The decoding process culminates with a standard MV2 block followed by a 3×3 transposed convolution (TransConv) layer to map the high-level semantic features back into the pixel space, obtaining the reconstructed residual image as

$$\hat{\boldsymbol{r}} = \text{Sigmoid}(\text{TransConv}_{3\times3,s=2}(\text{MV2}_{s=1}(\boldsymbol{y_0}))). \quad (5)$$

C. Decoupled Control-Data Transmission Mechanism

The dual-level semantic information can be analogized to a jigsaw puzzle assembly, where coarse-grained semantics represent the puzzle outline and fine-grained semantics provide detailed components. Consequently, coarse-grained semantics assume greater criticality than fine-grained semantics. Crucially, any transmission error in the codeword index would lead to retrieval of completely erroneous codewords at the receiver, resulting in catastrophic failure in image reconstruction.

Leveraging the inherent multi-channel capability of UAVs, we propose a decoupled control-data transmission (DCDT) mechanism, whose architecture is illustrated in Figure 3. The DCDT mechanism uses two independent wireless channels to process different information in parallel, and combines the coarse-grained and fine-grained semantic information through a fusion module at the receiver to obtain the reconstructed image.

Specifically, for a codebook of size N, the index v can be represented by only $\log_2(N)$ bits. Given that the critical importance of this index for the entire reconstruction process, our design prioritizes its transmission reliability over further compression. We therefore regard the index as a digital signal and employ powerful BCH codes with multi-bit error correction capability for channel coding. We argue that sacrificing a negligible potential compression gain in exchange for near-perfect reliability of this critical information is the optimal trade-off for ensuring overall system robustness. To ensure power-efficient transmission suitable for UAVs, we adopt BPSK digital modulation to convert the index into an analog waveform for transmission over the telemetry channel, which we term as control transmission.

However, since we utilize two independent channels to transmit the dual-level semantic information separately, these two data streams may exhibit different latencies at the receiver. This creates a critical challenge for data synchronization. To address this issue, we assume a data synchronization mechanism based on timestamps.

The transmitter assigns each image a unique timestamp T_n, which is simultaneously appended to both the index and semantic features. At the receiver, the two data streams are decoded separately, then a fusion module is utilized to complete data synchronization, the details are outlined in Algorithm 1. This module maintains two buffers: one for the decoded index from the telemetry channel $\mathcal{B}_v$ and another for semantic features from the data channel $\mathcal{B}_r$. A synchronization matching mechanism within this module continuously monitors both buffers, triggering pixel-level superposition to produce the reconstructed image $\hat{\boldsymbol{x}}$ when timestamp-matched pairs are detected. (this study assumes ideal synchronization matching performance).

Algorithm 1 Synchronization and Fusion Mechanism

Initialize: Index buffer $\mathcal{B}_v \leftarrow \emptyset$, Semantic features buffer $\mathcal{B}_r \leftarrow \emptyset$;
Input: Shared codebook $\mathcal{C}$;
Function ReceiveIndex(v_n, T_n):
  Add (v_n, T_n) to $\mathcal{B}_v$;
  TriggerFusion().
Function ReceiveFeatures($\hat{\boldsymbol{r}}_n, T_n$):
  Add $(\hat{\boldsymbol{r}}_n, T_n)$ to $\mathcal{B}_r$;
  TriggerFusion().
Function TriggerFusion():
  for each (v_i, T_i) in $\mathcal{B}_v$ **do**
    if a pair $(\hat{\boldsymbol{r}}_j, T_j)$ exists in $\mathcal{B}_r$ where $T_j = T_i$ **then**
      $\boldsymbol{c} \leftarrow \mathcal{C}[v_i]$; {Retrieve codeword using index}
      $\hat{\boldsymbol{x}} \leftarrow \hat{\boldsymbol{r}}_j + \boldsymbol{c}$; {Fuse to reconstruct image}
      Remove (v_i, T_i) from $\mathcal{B}_v$;
      Remove $(\hat{\boldsymbol{r}}_j, T_j)$ from $\mathcal{B}_r$;
      return $\hat{\boldsymbol{x}}$.
    end if
  end for

III. GAI-ENHANCED ROBUST RECONSTRUCTION

In our proposed lightweight codebook-assisted semantic communication architecture, image reconstruction relies on the pixel-wise superposition of decoded residual information $\hat{\boldsymbol{r}}$ and codeword $\boldsymbol{c}$. However, under extremely poor channel conditions, the residual information may suffer severe distortion due to the channel noise. In such cases, simple pixel superposition would propagate these noise components into the final reconstructed image, resulting in image blurring, artifacts, and even the loss of key semantic details. In addition, the operation of mapping original images to codewords inherently constitutes a lossy quantization representation, creating an intrinsic "fidelity gap" that cannot be bridged through simple pixel-level fusion.

To mitigate the distortion introduced by the channel and compensate for the quantization losses inherent in codebook representations, we introduce a GAI module at the receiver to transform the image reconstruction from conventional signal recovery to conditional content generation. This goal

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2026.3660029

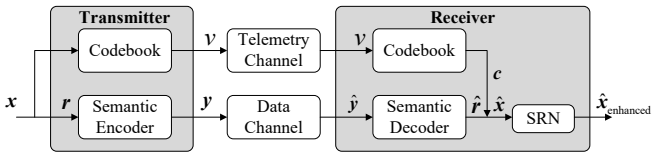


FIGURE 4. Architecture of the GAI-enhanced image reconstruction system.

is implemented through our designed Semantic Refinement Network (SRN).

A. Architecture of the Semantic Refinement Network

As illustrated in Figure 4, we integrate the SRN module into the original receiver architecture. Specifically, the SRN is a lightweight GAI model that treats the preliminary reconstructed image $\hat{\boldsymbol{x}}$ as a compressed and noise-corrupted semantic prompt. Leveraging its powerful prior learning ability, the module subsequently generates a perceptually realistic and high-definition image as the final output of the system.

The SRN is implemented by a conditional generative adversarial network (cGAN), a powerful framework for image-to-image translation. Building upon the standard cGAN architecture, we design two core components, the channel-adaptive generator G and the pair-aware discriminator D, which are trained in an adversarial manner. Through this competitive learning paradigm, the generator progressively acquires the capability to produce increasingly realistic and detail-enhanced images while preserving the fundamental semantic content from the preliminary reconstructed images $\hat{\boldsymbol{x}}$. Concurrently, the discriminator becomes more adept at identifying even subtle artifacts in the generated outputs $\hat{\boldsymbol{x}}_{\text{enhanced}}$. The detailed structure and data flow of the these components are illustrated in Figure 5.

1) Channel-Adaptive Generator

The generator G in the SRN module adopts a U-Net encoder-decoder architecture [27], a symmetric structure proven effective for image refinement tasks. Unlike conventional conditional generation models, our generator receives dual conditional inputs, namely, the preliminary reconstructed image $\hat{\boldsymbol{x}}$ from pixel superposition and the channel SNR value s. This innovative design enables the network to dynamically adjust its refinement strategy based on transmission conditions.

The encoder path gradually extracts multi-scale feature representations from $\hat{\boldsymbol{x}}$ through a series of downsampling convolution blocks, forming a feature pyramid from coarse to fine. Our core innovation lies in the SNR-adaptive modulation mechanism introduced at the bottleneck layer, as shown in the blue box in Figure 5. We design a feature-wise linear modulation (FiLM) layer to achieve this, and the detailed operation is as follows.

First, the scalar SNR value s is processed by a two-layer multilayer perceptron (MLP) serving as an SNR encoder. This MLP composed of two linear layers with a ReLU activation in between, performs a non-linear mapping from the scalar input to a high-dimensional embedding vector. The MLP output dimension is deliberately set as twice the channel count of the bottleneck feature map $\boldsymbol{z}$. This vector is subsequently split along its feature dimension into two parameter vectors, the scaling factors $\boldsymbol{\gamma}$ and the offset factors $\boldsymbol{\beta}$. Finally, these vectors are reshaped to match the channel dimension of the feature map and perform a channel-wise affine transformation facilitated by the broadcasting mechanism [28] in deep learning. The modulation operation is formulated as:

$$\boldsymbol{z}' = \boldsymbol{z} \odot (1 + \boldsymbol{\gamma}(s)) + \boldsymbol{\beta}(s), \tag{6}$$

where $\odot$ denotes element-wise multiplication, $\boldsymbol{\gamma}(s)$ and $\boldsymbol{\beta}(s)$ are effectively expanded to match the spatial dimensions of $\boldsymbol{z}$, applying a learned scaling and offsetting to each channel.

Through this affine transformation, the network can dynamically adjust the amplitude and bias of the feature representation according to the channel conditions, thereby achieving an adaptive enhancement strategy: adopting an aggressive denoising strategy under low SNR conditions and conservatively retaining detailed information under high SNR conditions. Such channel-aware capability is not available in traditional fixed-strategy enhancement methods.

The decoder path progressively restores the spatial resolution of the image through transposed convolution operations. The key design lies in the skip connection mechanism, where each decoder layer establishes direct links with its corresponding encoder layer through feature concatenation. This multi-scale feature fusion strategy offers dual advantages. Firstly, it enables the network to directly propagate undistorted useful information from the superimposed output $\hat{\boldsymbol{x}}$, avoiding unnecessary information loss. Secondly, it provides additional pathways for gradient backpropagation, effectively mitigating the vanishing gradient problem in the training of deep networks.

2) Pair-Aware Discriminator

We employ the PatchGAN architecture for the discriminator D and redesign its conditional input. For a given preliminary image $\hat{\boldsymbol{x}}$, the discriminator needs to evaluate whether the image generated by the generator is a reasonable, high-quality, and semantically consistent enhanced version of that input. Therefore, we use the image pair concatenated in the channel dimension as the input of the discriminator. This paired-input design, which we term "pair-aware", enables the discriminator to learn complex conditional dependencies between the input prompt and the refined output.

During the training process, the discriminator will receive two types of samples. One is the real sample pair, which is formed by concatenating the original image $\boldsymbol{x}$ with its preliminary reconstructed version $\hat{\boldsymbol{x}}$. This sample pair establishes a positive constraint for the discriminator, enabling

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2026.3660029

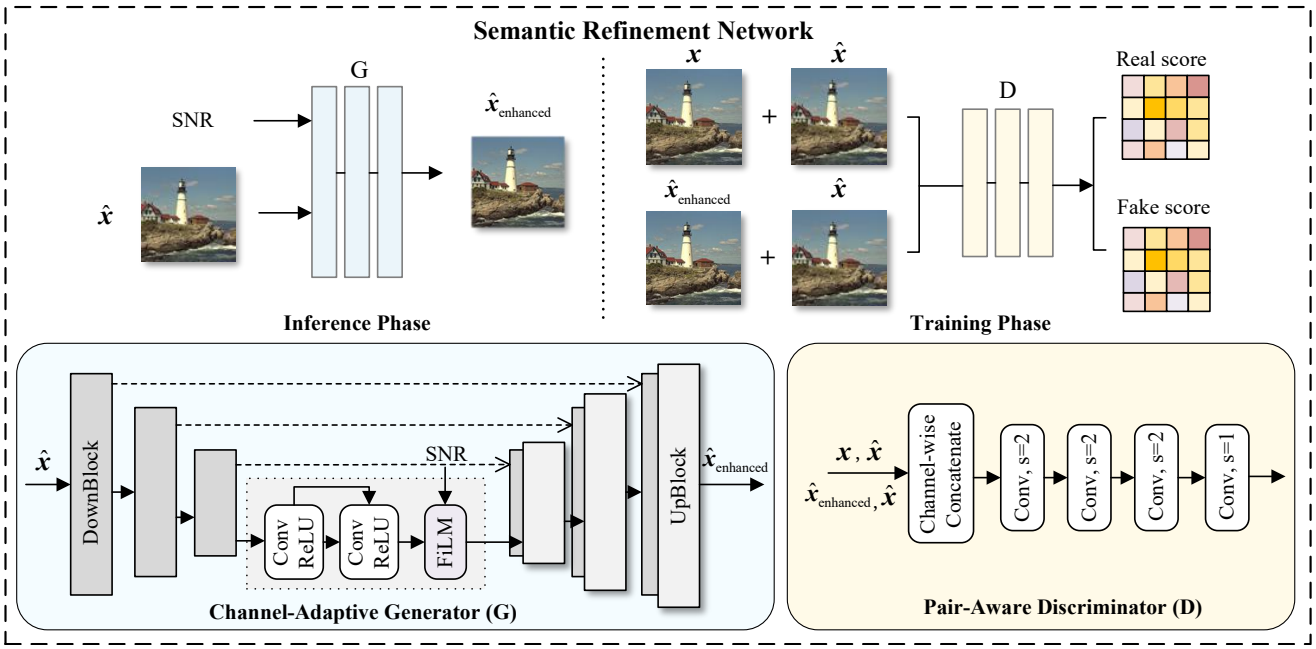


FIGURE 5. **Detailed structure and operational phases of the semantic refinement network (SRN).**

it to learn the standard refinement pattern from low-quality input to high-quality output. The other is the generated sample pair, constructed by concatenating the refined image $\hat{\boldsymbol{x}}_{\text{enhanced}}$ enhanced by the generator and the preliminary reconstructed image $\hat{\boldsymbol{x}}$. This sample pair provides a negative constraint, allowing the discriminator to learn to identify the unreasonable detail addition in the generated image. It should be emphasized that both types of sample pairs contain the preliminary reconstructed image, which ensures that the discriminator's judgment is based on the same conditional input.

Through bidirectional constraints imposed by both real and generated sample pairs, the discriminator can progressively master two critical abilities. First, it learns to identify images that appear visually clear but contain false details that do not exist in the input. Second, it can detect excessive enhancements that destroy the original semantic content. These two abilities jointly constrain the generator, ensuring its outputs not only improve visual quality but also strictly maintain consistency with the input conditions and preserve semantic integrity.

B. Optimization Objective and Loss Functions

To guide the SRN towards generating images that are pixel-accurate, perceptually realistic, and semantically correct, we design a composite loss function $\mathcal{L}_{\text{G}}$ to train the generator. This loss function consists of four complementary components, each constraining the reconstruction quality at different levels:

$$\mathcal{L}_{\text{G}} = \lambda_1 \mathcal{L}_{\text{L1}} + \lambda_2 \mathcal{L}_{\text{adv}} + \lambda_3 \mathcal{L}_{\text{perc}} + \lambda_4 \mathcal{L}_{\text{sem}}, \tag{7}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are hyperparameter weights used to balance different loss terms.

The pixel-level loss $\mathcal{L}_{\text{L1}}$ provides a fundamental spatial domain constraint, ensuring that the output image of the generator is accurately aligned with the ground truth $\boldsymbol{x}$ at the pixel level. This term prevents the generator from producing outputs that deviate excessively from the original content and offers robust, stable gradients during the initial training phase:

$$\mathcal{L}_{\text{L1}} = \mathbb{E}_{\boldsymbol{x},\hat{\boldsymbol{x}},s} \left[\| \boldsymbol{x} - \text{G}(\hat{\boldsymbol{x}}, s) \|_1 \right]. \tag{8}$$

The adversarial loss $\mathcal{L}_{\text{adv}}$ comes from the feedback of the discriminator, driving the generator to produce outputs that are statistically indistinguishable from the real data. This component aims to enhance the visual realism of generated images, making their statistical distribution close to the manifold of real images:

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{\hat{\boldsymbol{x}},s} \left[-\log \text{D}(\text{G}(\hat{\boldsymbol{x}}, s), \hat{\boldsymbol{x}}) \right]. \tag{9}$$

The perceptual loss $\mathcal{L}_{\text{perc}}$ ensures that the generator produces images consistent with human visual perception. Leveraging a pre-trained VGG-19 [29] network $\phi(\cdot)$, we minimize the distance between high-level features of the generated and original images to preserve similar textures, styles, and perceptual content:

$$\mathcal{L}_{\text{perc}} = \mathbb{E}_{\boldsymbol{x},\hat{\boldsymbol{x}},s} \left[\| \phi(\boldsymbol{x}) - \phi(\text{G}(\hat{\boldsymbol{x}}, s)) \|_2^2 \right]. \tag{10}$$

The semantic loss $\mathcal{L}_{\text{sem}}$ is calculated in the high-dimensional feature space of the pre-trained ResNet18 [30] network $\psi(\cdot)$, aiming to supervise the generation of SRN and prevent the generator from making catastrophic semantic errors when refining details:

$$\mathcal{L}_{\text{sem}} = \mathbb{E}_{\boldsymbol{x},\hat{\boldsymbol{x}},s} \left[\| \psi(\boldsymbol{x}) - \psi(\text{G}(\hat{\boldsymbol{x}}, s)) \|_2^2 \right], \tag{11}$$

For the discriminator, its goal is to maximize the distance between the real and fake samples. We adopt the more stable

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2026.3660029

Wasserstein GAN with gradient penalty (GP) objective and the loss function for the discriminator $\mathcal{L}_{\mathrm{D}}$ is defined as:

$$\mathcal{L}_{\mathrm{D}} = \mathbb{E}_{\hat{\boldsymbol{x}},s}[\mathrm{D}(\mathrm{G}(\hat{\boldsymbol{x}},s),\hat{\boldsymbol{x}})] - \mathbb{E}_{\boldsymbol{x},\hat{\boldsymbol{x}}}[\mathrm{D}(\boldsymbol{x},\hat{\boldsymbol{x}})] + \lambda_{\mathrm{GP}}\mathcal{L}_{\mathrm{GP}}, \quad (12)$$

where the gradient penalty term $\mathcal{L}_{\mathrm{GP}}$ enforces the 1-Lipschitz constraint, which is crucial for stabilizing training. It is calculated on interpolated samples $\tilde{\boldsymbol{x}}$, which are formed by a random convex combination of real and fake image pairs:

$$\mathcal{L}_{\mathrm{GP}} = \mathbb{E}_{\hat{x}}\left[\left(\|\nabla_{\hat{\boldsymbol{x}}}\mathrm{D}(\hat{\boldsymbol{x}},\tilde{\boldsymbol{x}})\|_2 - 1\right)^2\right], \quad (13a)$$

$$\text{where } \tilde{\boldsymbol{x}} = \epsilon\boldsymbol{x} + (1-\epsilon)\mathrm{G}(\hat{\boldsymbol{x}},s) \text{ with } \epsilon \sim U[0,1]. \quad (13b)$$

IV. EXPERIMENT EVALUATION

A. Experimental Setup

Dataset and Preprocessing. We use the widely-adopted VisDrone dataset, a large-scale benchmark collected by UAVs that contains thousands of images with diverse scenes and resolutions. To prepare the data for training, all images are first scaled by a certain ratio, and then regions of 256×256 pixels are extracted through random cropping as training samples. The dataset is divided into training, validation, and testing sets.

Evaluation Metrics. We adopt the peak signal-to-noise ratio (PSNR) and the multi-scale structural similarity (MS-SSIM) as performance metrics. Here, PSNR is measured in dB, while MS-SSIM ranges from 0 to 1. For both metrics, higher values indicate superior model performance. In addition, we introduce the LPIPS metric to evaluate the improvement effect of the SRN module on the visual perception performance of images. It is highly correlated with human judgment of image quality, with lower values indicating superior performance.

Comparison Schemes. We conducted comparative experiments under identical experimental settings to verify the performance of the lightweight semantic communication backbone. The benchmark methods include the classical separate source and channel coding scheme BPG + LDPC, the codebook-assisted JSCC approach based on Swin Transformer CB-SwinJSCC [20], and the proposed lightweight backbone without the codebook MobileViTv3 JSCC (w/o Codebook).

Channel Models. To evaluate the robustness of the proposed framework, we conduct experiments under two standard wireless channel models, the AWGN channel and the Rayleigh fading channel. The received signal $\hat{\boldsymbol{y}}$ is generally modeled as $\hat{\boldsymbol{y}} = h\boldsymbol{y} + \boldsymbol{n}$, where $\boldsymbol{y}$ is the transmitted latent feature vector, h is the channel fading coefficient, and $\boldsymbol{n}$ is the additive noise vector. The AWGN channel models thermal noise in static or line-of-sight communication scenarios, where the channel coefficient $h = 1$. The noise vector $\boldsymbol{n}$ is modeled as complex Gaussian noise with zero mean and variance σ^2, i.e., $\boldsymbol{n} \sim CN(0, \sigma^2\boldsymbol{I})$. The Rayleigh fading channel models the severe signal amplitude variations caused by multi-path propagation in non-line-of-sight and mobile environments, which is highly relevant for UAV applications. In this channel, the fading coefficient h is modeled as a complex Gaussian random variable with zero mean and unit variance, i.e., $h \sim CN(0,1)$. The noise $\boldsymbol{n}$ follows the same distribution as in the AWGN channel.

Implementation Details. The experiments were implemented in PyTorch and trained on an NVIDIA GeForce RTX 3090 GPU, using the Adam optimizer. For the lightweight DCDT-MobileViTv3 backbone, we utilizes a pre-constructed fixed codebook of size $N = 16$ and the corresponding 4-bit codeword index v. The number of stages in the semantic encoder is set to $n = 4$, adopting $[m_1, m_2, m_3, m_4] = [2, 2, 6, 3]$ MobileViTv3 blocks per stage respectively.

Then we froze all the parameters of the trained lightweight semantic communication backbone to specifically train the SRN module. In the first stage, a composite loss function without an adversarial term $\mathcal{L}_{\mathrm{adv}}$ was used to pre-train the generator with a learning rate of 2e-4. In the second stage, the entire SRN was fine-tuned through adversarial training, with the learning rate of the generator set to 1e-5 and the learning rate of the discriminator set to 2e-4. The loss weighting parameters were empirically set to $\lambda_1 = 50, \lambda_2 = 10, \lambda_3 = 10, \lambda_4 = 0.5$ and $\lambda_{\mathrm{GP}} = 50$.

B. Performance of the Lightweight Semantic Transmission Backbone

In this section, we evaluate the performance of the lightweight DCDT-MobileViTv3 backbone, which serves as the foundational platform for the subsequent generative enhancement.

Figure 6 and Figure 7 shows the performance of the backbone under varying SNR and CR conditions over both AWGN and Rayleigh fading channels, respectively.

As observed across both channel models, the JSCC-based semantic communication methods significantly outperform the traditional BPG+LDPC scheme. The performance gap is particularly pronounced in lower SNR conditions, highlighting the inherent robustness of transmitting semantic features over raw pixel data. Furthermore, compared to the JSCC method without codebook, the proposed DCDT-MobileViTv3 JSCC method demonstrates superior performance across all SNR and CR ranges, with the advantage being more pronounced under the Rayleigh fading channel. This comparison between the codebook-assisted and non-codebook methods provides the core quantitative justification for the dual-level semantic transmission strategy. These results confirm that separating the semantics into the coarse-grained and fine-grained information leads to a demonstrable improvement in overall performance and robustness.

However, it is also evident that there is still a gap between our method and the SOTA method CB-SwinJSCC, which is an inherent trade-off between performance and efficiency in lightweight network design. Additionally, the performance of all methods begins to degrade sharply as the SNR drops, indicating that the lightweight backbone has a performance bottleneck under poor channel conditions. These limitations

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2026.3660029

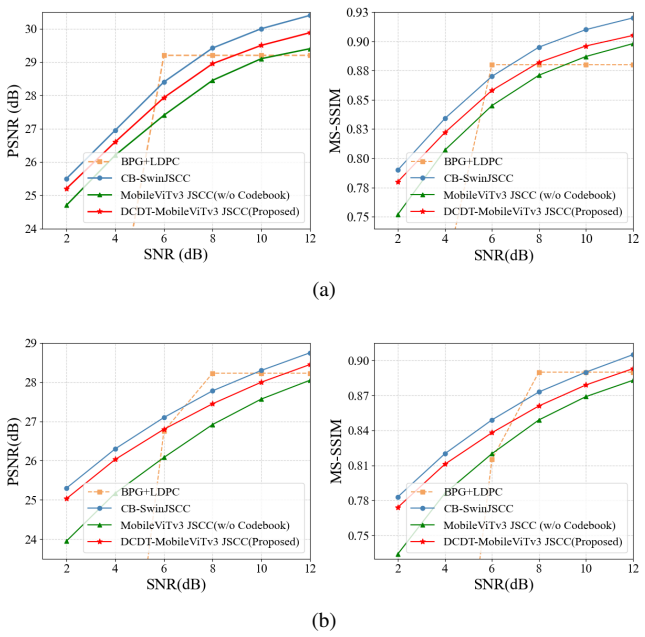


FIGURE 6. (a) PSNR and MS-SSIM performance under different SNRs over the AWGN channel with CR=0.0625. (b) PSNR and MS-SSIM performance under different SNRs over the Rayleigh fading channel with CR=0.0625.

motivate us to introduce a generative enhancement module, which is specifically designed to address this challenge.

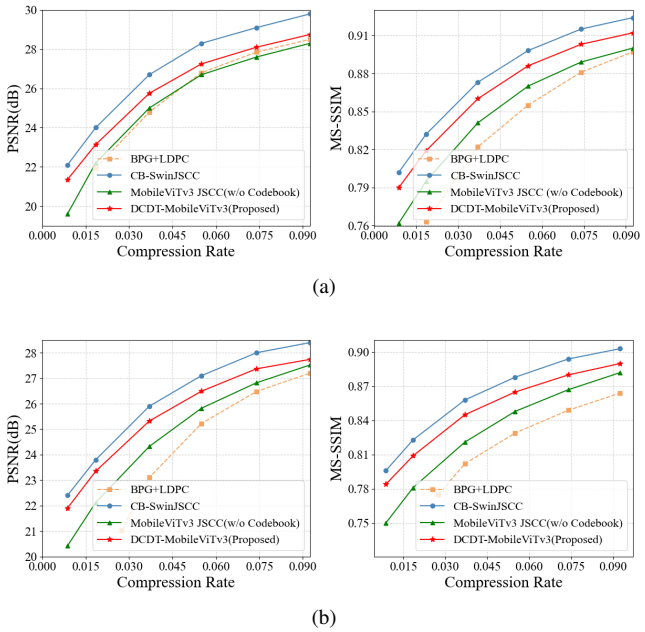


FIGURE 7. (a) PSNR and MS-SSIM performance under different SNRs over the AWGN channel with CR=0.0625. (b) PSNR and MS-SSIM performance under different SNRs over the Rayleigh fading channel with CR=0.0625.

C. Robustness Enhancement with Semantic Refinement Network

Building upon the lightweight DCDT-MobileViTv3 JSCC backbone, this section aims to evaluate the performance improvement provided by the generative SRN module for image reconstruction, particularly under extremely challenging low-SNR conditions where traditional methods fail. Ablation studies of the SRN moudule and the key component were performed over the AWGN channel to quantify the performance contributions of different components, with SNR ranging from -5 dB to 5 dB.

As shown in Figures 8(a) and 8(b), the baseline model without SRN exhibits significant degradation in both PSNR and MS-SSIM as the SNR decreases, reflecting the performance bottleneck of conventional systems where channel noise is directly transmitted to the output. In contrast, the proposed method (Full) consistently outperforms all other methods across the entire SNR range, achieving a remarkable PSNR of over 24 dB under the challenging condition of -5 dB, which represents an improvement of more than 6 dB compared to the baseline method.

The ablation study on the FiLM module reveals a crucial insight. As the results shown, when the SNR is approximately 1 dB, the baseline model without SRN and the CB-SwinJSCC method both begin to surpass the model without FiLM on both PSNR and MS-SSIM metrics. This phenomenon occurs because the model without FiLM employs a static generator that lacks SNR perception ability. Our full model, empowered by the SNR-adaptive FiLM module, resolves this issue by endowing the generator with channel-aware capability, enabling a dynamic transition from aggressive generation to conservative refinement. This adaptive characteristic allows it to consistently outperform all comparison methods across the entire SNR range.

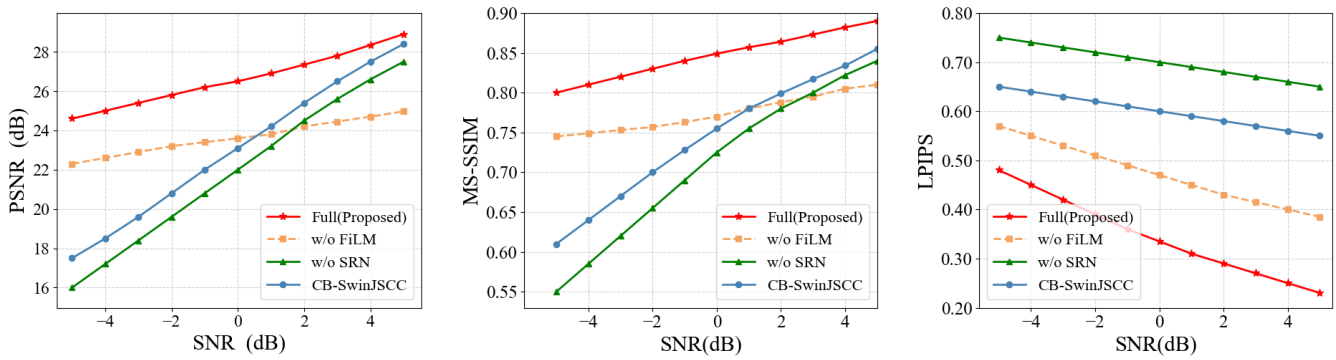
To further investigate the improvement in semantic recovery and visual perception quality afforded by the SRN, we employ the LPIPS metric for a more comprehensive evaluation. Figure 8(c) presents the quantitative performance across the SNR range, while Figure 9 provides visual examples under the AWGN channel with SNR of 0 dB. The results indicate that although the approach without SRN module attempts to maintain pixel-level accuracy, its high LPIPS value reveals the presence of severe noise and blurring, rendering the reconstructed images perceptually poor. As visualized in Figure 9, the model without FiLM shows a significant decrease in LPIPS score, which successfully recovers the basic structure of the image but still lacks fine details and color fidelity, resulting in suboptimal visual quality. In sharp contrast, the proposed model demonstrates a remarkable capability for high-fidelity reconstruction with clear details and good visual perception quality. The visual examples show significantly lower LPIPS scores of 0.28 and 0.32 respectively, confirming the powerful ability of SRN to generate high-fidelity content from severely impaired information.

D. Complexity and Latency Analysis

To address practical deployment concerns for UAV applications, particularly regarding computational complexity

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2026.3660029



(a) PSNR vs. SNR (b) MS-SSIM vs. SNR (c) LPIPS vs. SNR

FIGURE 8. Ablation study to understand the influence of our proposed SRN module and FiLM module on the enhancement performance of the image. Note that all experiments are conducted under the AWGN channel, and "w/o" means "without".

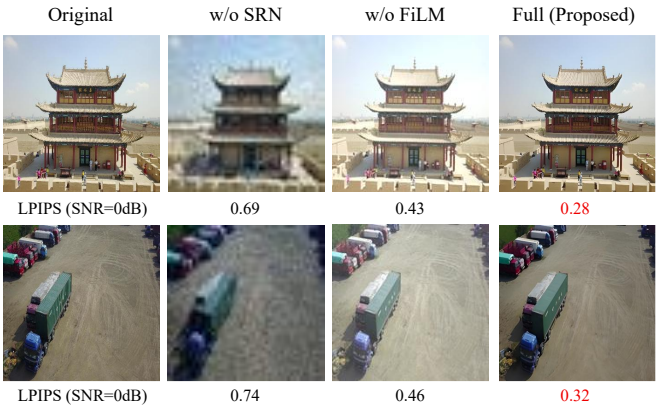


FIGURE 9. Visualization of the ablation results of the SRN under the AWGN channel with SNR of 0 dB.

and processing delay, we provide a quantitative analysis of our framework. Our design principle aims to minimize the overhead on the resource-constrained UAV transmitter while leveraging the powerful computational resources of the ground receiver for high-fidelity reconstruction. Table 1 presents a comparison of model complexity and inference latency against the SOTA SwinJSCC baseline. Complexity is measured in terms of total parameters and floating-point operations (FLOPs). Latency was evaluated by measuring the average inference time for a single 256×256 image on an NVIDIA GeForce RTX 3090 GPU.

TABLE 1. Comparison of complexity and latency under different models used as backbone networks.

Backbone	Component	Parameter (M)	FLOPs (G)	Latency (ms)
SwinJSCC	Encoder	14.8	4.5	32.7
	Receiver	15.6	5.3	60.8
	Total	*30.4*	*9.8*	*93.5*
Ours	**Encoder**	**3.8**	**0.8**	**6.5**
	Receiver	24.3	8.4	76.7
	Total	*28.1*	*9.2*	*83.2*

The results in Table 1 show that, as the main component for on-board UAV deployment, the encoder of the proposed model exhibits exceptional lightweight characteristics. Compared to the SwinJSCC encoder, the number of its parameters and FLOPs are reduced by 74.3% and 82.2% respectively, and the inference speed is increased by approximately 5 times. This significant on-device advantage is achieved by strategically offloading the computational burden to the ground server. To ensure high-quality reconstruction, our model incorporates the powerful SRN at the receiver, which leads to a higher complexity and latency compared to the SwinJSCC.

Remarkably, the total complexity and latency of the overall framework are both lower than the SwinJSCC baseline. This analysis validates that the proposed framework strikes an excellent balance between performance, complexity, and practical deployability, demonstrating great potential for deployment in UAV communication scenarios.

V. CONCLUSION

This paper proposes a novel GAI-enhanced semantic communication framework to address the critical challenges of robust UAV image transmission in low SNR environments. By shifting the reconstruction paradigm from signal recovery to conditional content generation, the proposed SRN effectively overcomes the performance limitations of conventional semantic systems. The key to our success lies in the channel-adaptive generator empowered by the SNR-adaptive FiLM module, which can intelligently customize its generation strategy according to real-time channel conditions. Experimental results show that our method significantly outperforms the traditional separate source-channel coding method and semantic communication baselines, especially in terms of perceptual quality under severe channel degradation.

This work validates the substantial potential of deeply integrating GAI with communication systems to develop highly robust and intelligent wireless transmission solutions, and we identify several avenues for future research. First,

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2026.3660029

future work could explore evaluating the framework's performance under more realistic, time-varying channel models that capture the rapid SNR fluctuations caused by UAV mobility and interference. Second, to address the limitations of the static codebook in highly dynamic or entirely new operational domains, developing online adaptation or incremental learning mechanisms for the codebook is a crucial next step to enhance scalability and performance. Finally, future research directions can extend our general semantic communication architecture to advanced visual tasks such as object detection and image classification. This direction is crucial for narrowing the gap between high-quality reconstruction and executable machine intelligence in UAV applications.

REFERENCES

[1] L. Gupta, R. Jain, and G. Vaszkun, "Survey of important issues in UAV communication networks," *IEEE Commun. Surv. Tutorials*, vol. 18, no. 2, pp. 1123–1152, 2016.

[2] W. Chen, Z. Su, Q. Xu, T. H. Luan, and R. Li, "VFC-based cooperative UAV computation task offloading for post-disaster rescue," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, 2020, pp. 228–236.

[3] K. Liu and J. Zheng, "UAV trajectory optimization for time-constrained data collection in UAV-enabled environmental monitoring systems," *IEEE Internet Things J.*, vol. 9, no. 23, pp. 24 300–24 314, 2022.

[4] H. Xu, L. Wang, W. Han, Y. Yang, J. Li, Y. Lu, and J. Li, "A survey on UAV applications in smart city management: Challenges, advances, and opportunities," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 16, pp. 8982–9010, 2023.

[5] H. Chen, H. Cui, J. Wang, P. Cao, Y. He, and M. Guizani, "Computation offloading optimization for UAV-based cloud-edge collaborative task scheduling strategy," *IEEE Trans. Cognit. Commun. Networking*, vol. 11, no. 6, pp. 4240–4253, 2025.

[6] G. Shi, Y. Xiao, Y. Li, and X. Xie, "From semantic communication to semantic-aware networking: Model, architecture, and open problems," *IEEE Commun. Mag.*, vol. 59, no. 8, pp. 44–50, 2021.

[7] L. Xia, Y. Sun, C. Liang, L. Zhang, M. A. Imran, and D. Niyato, "Generative AI for semantic communication: Architecture, challenges, and outlook," *IEEE Wireless Commun.*, vol. 32, no. 1, pp. 132–140, 2025.

[8] C. Liang, H. Du, Y. Sun, D. Niyato, J. Kang, D. Zhao, and M. A. Imran, "Generative AI-driven semantic communication networks: Architecture, technologies, and applications," *IEEE Trans. Cognit. Commun. Networking*, vol. 11, no. 1, pp. 27–47, 2025.

[9] N. Cheng, S. Wu, X. Wang, Z. Yin, C. Li, W. Chen, and F. Chen, "AI for UAV-assisted IoT applications: A comprehensive review," *IEEE Internet Things J.*, vol. 10, no. 16, pp. 14 438–14 461, 2023.

[10] X. Kang, B. Song, J. Guo, Z. Qin, and F. R. Yu, "Task-oriented image transmission for scene classification in unmanned aerial systems," *IEEE Trans. Commun.*, vol. 70, no. 8, pp. 5181–5192, 2022.

[11] B. Xie, Y. Wu, Y. Shi, D. W. K. Ng, and W. Zhang, "Communication-efficient framework for distributed image semantic wireless transmission," *IEEE Internet Things J.*, vol. 10, no. 24, pp. 22 555–22 568, 2023.

[12] X. Yao, J. Zheng, X. Zheng, H. Dai, and X. Yang, "Optimization of image transmission in UAV-enabled semantic communication networks," in *Proc. 9th Int. Conf. Comput. Commun. (ICCC)*, 2023, pp. 647–652.

[13] X. Song, F. Zhou, R. Ding, Z. Qu, Y. Li, Q. Wu, and N. Al-Dhahir, "UAV cognitive semantic communications enabled by knowledge graph for robust object detection," *IEEE Trans. Commun.*, vol. 73, no. 8, pp. 6052–6067, 2025.

[14] V. Papi, P. Oli, A. Milan, S. Gotovac, and M. Poli, "High-resolution image transmission from UAV to ground station for search and rescue missions planning," *Appl. Sci.*, vol. 11, no. 5, p. 2105, 2021.

[15] Z. Jiao, Y. Zhang, L. Mu, J. Xin, S. Jiao, H. Liu, and D. Liu, "A yolov3-based learning strategy for real-time UAV-based forest fire detection," in *Proc. Chin. Control Decis. Conf. (CCDC)*, 2020, pp. 4963–4967.

[16] W. Alexan, L. Aly, Y. Korayem, M. Gabr, D. El-Damak, A. Fathy, and H. A. A. Mansour, "Secure communication of military reconnaissance images over UAV-assisted relay networks," *IEEE Access*, vol. 12, pp. 78 589–78 610, 2024.

[17] Z. Li, Q. Wang, T. Zhang, C. Ju, S. Suzuki, and A. Namiki, "UAV high-voltage power transmission line autonomous correction inspection system based on object detection," *IEEE Sens. J.*, vol. 23, no. 9, pp. 10 215–10 230, 2023.

[18] Y. Zhang, B. Li, J. Shang, X. Huang, P. Zhai, and C. Geng, "DSA-net: An attention-guided network for real-time defect detection of transmission line dampers applied to UAV inspections," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–22, 2024.

[19] P. Yi, Y. Cao, X. Kang, and Y.-C. Liang, "Deep learning-empowered semantic communication systems with a shared knowledge base," *IEEE Trans. Wireless Commun.*, vol. 23, no. 6, pp. 6174–6187, 2024.

[20] H. Zhang, M. Tao, Y. Sun, and K. B. Letaief, "Improving learning-based semantic coding efficiency for image transmission via shared semantic-aware codebook," *IEEE Trans. Commun.*, vol. 73, no. 2, pp. 1217–1232, 2025.

[21] J. Ren, Y. Sun, H. Du, W. Yuan, C. Wang, X. Wang, Y. Zhou, Z. Zhu, F. Wang, and S. Cui, "Generative semantic communication: Architectures, technologies, and applications," *Engineering*, 2025.

[22] L. Zhou, X. Deng, Z. Wang, X. Zhang, Y. Dong, X. Hu, Z. Ning, and J. Wei, "Semantic information extraction and multi-agent communication optimization based on generative pre-trained transformer," *IEEE Trans. Cognit. Commun. Networking*, vol. 11, no. 2, pp. 725–737, 2025.

[23] L. Zhou, X. Deng, Z. Ning, H. Zhao, J. Wei, and V. C. M. Leung, "When generative AI meets semantic communication: Optimizing radio map construction and distribution in future mobile networks," *IEEE Network*, vol. 39, no. 3, pp. 47–55, 2025.

[24] P. Ye, Y. Sun, S. Yao, H. Chen, X. Xu, and S. Cui, "Codebook-enabled generative end-to-end semantic communication powered by transformer," in *Proc. IEEE INFOCOM Conf. Comput. Commun. Workshops (WKSHPS)*, 2024, pp. 1–6.

[25] K. Ye, M. Gong, S. Wang, and D. Feng, "Low-rate semantic communication with codebook-based conditional generative models," 2025, *arXiv:2504.04977*.

[26] J. Lu, W. Yang, Z. Xiong, C. Xing, R. Tafazolli, T. Q. Quek, and M. Debbah, "Generative AI-enhanced multi-modal semantic communication in internet of vehicles: System design and methodologies," 2024, *arXiv:2409.15642*.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. (MICCAI)*. Springer, 2015, pp. 234–241.

[28] C. R. Harris, K. J. Millman, S. J. Van Der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith *et al.*, "Array programming with numpy," *nature*, vol. 585, no. 7825, pp. 357–362, 2020.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/