PAW: Parallel-Wavelet Attention for Lightweight Asymmetric Semantic Image Communication in 6G Space Networks

Abstract

An important challenge in semantic communication for 6G space networks is achieving reliable image reconstruction under low-SNR conditions with constrained onboard resources. Existing Swin-based schemes rely on symmetric transceivers and spatial-only attention, which are not well suited to satellite–cloud systems. To address these issues, we propose PAW, a **Pa**rallel-**W**avelet Attention-based asymmetric architecture for semantic image communication. PAW integrates discrete wavelet transform (DWT) to enable high-fidelity channel and spatial attention through parallel frequency-aware recalibration paths. Lightweight MBConv and DSConv modules are further employed to reduce encoder-side computational complexity. Experiments show that PAW reduces encoder-side GPU runtime by up to **3.5**$\times$ compared to Swin-based baselines, while maintaining competitive reconstruction performance. These results suggest that joint *channel & spatial* attention is an effective design for future resource-constrained 6G space semantic communications.

1. Introduction

With the rapid evolution of satellite constellations and 6G-oriented Space-Air-Ground Integrated Networks (SAGIN), the demand for efficient visual data transmission—from Earth observation to deep-space exploration—has grown significantly (Leonard et al., 2026). In such scenarios, large volumes of imagery must be transmitted over long distances to support time-critical missions. However, space networks exhibit a **sharp computational asymmetry**: ground stations or orbital cloud hubs have abundant processing power, whereas satellite platforms are constrained by limited onboard resources. This disparity calls for more efficient transmission paradigms under strict resource constraints (Guo et al., 2025).

Traditional communication paradigms, rooted in Shannon's bit-level framework, are increasingly ill-suited for such noise-intensive environments due to the "cliff effect" (Zhang et al., 2026b). Semantic Communication (SemCom) has emerged as a promising alternative, transmitting task-relevant features to improve bandwidth efficiency, with recent extensions to multimodal settings (Ahn et al., 2025). However, translating these advances to 6G space networks remains challenging.

Existing SemCom approaches face two major bottlenecks in this setting. First, most SemCom designs primarily focus on symmetric transceiver architectures (Bourtsoulatze et al., 2019; Dai et al., 2022; Huang et al., 2023). These models assume that the encoder and decoder possess similar computational capabilities, performing exactly reciprocal operations. This assumption ignores the intrinsic asymmetry of space systems, leading to excessive onboard complexity for the transmitter. Second, with the recent success of vision transformers, attention-based architectures (e.g., Swin Transformer) have become a dominant backbone for SemCom to enhance representation capacity (Khalid et al., 2025; Wu et al., 2024). However, these methods primarily focus on capturing spatial-domain dependencies through heavy token-shuffling or self-attention, failing to exploit the **frequency-domain characteristics** of image data. Moreover, their massive parameter scales and memory footprints are often prohibitive for real-time onboard processing (Nguyen et al., 2024; Yang et al., 2025; Zhang et al., 2026a).

Motivated by these observations, this paper addresses two fundamental challenges in semantic communication for 6G space networks: **resource asymmetry**, where spaceborne transmitters operate under strict onboard constraints, while receiver-side cloud infrastructure offers much stronger computational capability; and **representation inefficiency**, where spatial-only attention fails to capture frequency-aware channel–spatial dependencies, leading to the loss of fine-grained semantic structures. To tackle these challenges, the main contributions of this work are summarized as follows:

- **Novel channel–spatial attention for semantic representation:** We propose a **Pa**rallel-**W**avelet Attention (PAW) module that leverages discrete wavelet transform (DWT) to enable joint channel–spatial modeling, resulting in improved structural fidelity with limited complexity.
- **Lightweight asymmetric architecture:** We develop

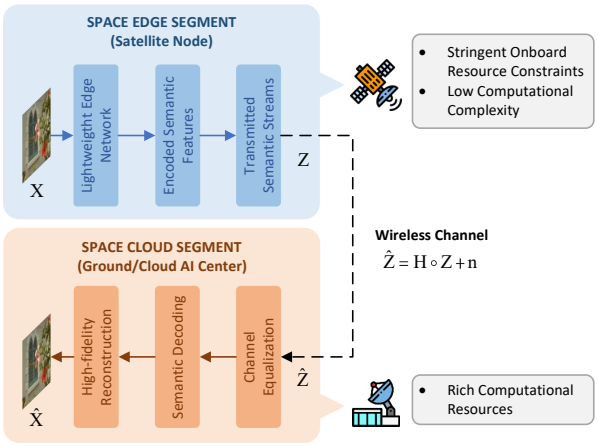

Figure 1. System model of the proposed asymmetric semantic communication framework for resource-constrained space networks.

an edge–cloud semantic communication architecture, where a lightweight encoder based on Mobile Inverted Bottleneck Convolution (MBConv) and Depthwise Separable Convolution (DSConv) reduces onboard complexity, achieving efficient complexity reduction without compromising reconstruction performance.

- **Efficiency–performance trade-off:** The proposed PAW-based framework achieves up to **3.5**× reduction in encoder-side GPU runtime, while maintaining competitive reconstruction quality and robustness compared to Swin-based baselines.

2. System Model

In this section, we first present a general end-to-end formulation for semantic communication systems, followed by a formal analysis of existing design paradigms and their inherent limitations in satellite-to-ground scenarios.

2.1. End-to-End Semantic Communication

Consider a source image $\mathbf{X} \in \mathbb{R}^{3\times h\times w}$ captured by a space-borne device. The transmission process begins with the onboard semantic encoder $\mathcal{E}_\theta(\cdot)$, which maps the high-dimensional pixel space into a compact latent semantic representation:

$$\mathbf{Z} = \mathcal{E}_\theta(\mathbf{X}), \quad \mathbf{Z} \in \mathbb{R}^{C\times h\times w}, \tag{1}$$

where θ denotes the trainable parameters of the encoder. The extracted semantic vector $\mathbf{Z}$ is then transmitted through the wireless channel $\mathbf{H}$, where the signal is subjected to channel fading and additive white Gaussian noise (AWGN). The received signal at the ground station $\hat{\mathbf{Z}}$ is expressed as:

$$\hat{\mathbf{Z}} = \mathbf{H} \circ \mathbf{Z} + \mathbf{n}, \tag{2}$$

where $\mathbf{n} \sim \mathcal{CN}(0, \sigma^2\mathbf{I})$ represents the complex noise with variance σ^2, and $\circ$ denotes the element-wise multiplication with the fading coefficients. Upon receiving the corrupted semantic features, the ground-based decoder $\mathcal{D}_\phi(\cdot)$ performs semantic equalization and generative reconstruction to obtain the recovered image:

$$\hat{\mathbf{X}} = \mathcal{D}_\phi(\hat{\mathbf{Z}}). \tag{3}$$

By integrating these sequential operations, the entire end-to-end transmission chain can be represented as a unified mapping function:

$$\hat{\mathbf{X}} = \mathcal{D}_\phi\left(\mathcal{H}(\mathcal{E}_\theta(\mathbf{X})) + \mathbf{n}\right). \tag{4}$$

The system is jointly optimized to minimize a semantic-oriented distortion metric $d(\mathbf{X}, \hat{\mathbf{X}})$ across various channel conditions $(\mathbf{H}, \mathbf{n})$, which can be formulated as the following objective:

$$(\theta^*, \phi^*) = \arg\min_{\theta,\phi} \mathbb{E}_{\mathbf{X},\mathbf{H},\mathbf{n}}[d(\mathbf{X}, \hat{\mathbf{X}})]. \tag{5}$$

2.2. Semantic Encoder and Decoder Design

Existing semantic communication frameworks can be abstracted as

$$\mathcal{E}_{gen}(\mathbf{X}) = \mathcal{A}_{att}\big(\mathcal{B}_{block}(\mathbf{X})\big), \tag{6}$$

where $\mathcal{B}_{block}$ denotes feature extraction blocks and $\mathcal{A}_{att}$ represents the attention operator. Based on this formulation, current implementations generally exhibit two key limitations. First, $\mathcal{A}_{att}$ is typically instantiated using Transformer-based designs (e.g., Swin), which operate primarily in the spatial domain while neglecting frequency-domain characteristics, and incur high computational overhead due to stacked attention modules. Second, $\mathcal{B}_{block}$ commonly relies on convolutional backbones without explicit lightweight considerations, resulting in redundant computation. Moreover, many works assume comparable encoder–decoder complexity ($\mathcal{O}(\mathcal{E}) \approx \mathcal{O}(\mathcal{D})$), which is impractical for resource-constrained scenarios.

Motivated by these issues, we propose a novel asymmetric semantic communication architecture that jointly considers efficiency and representation capability. The framework integrates a Lightweight Basic Block (LBB) for efficient feature extraction and PAW module for joint spatial–frequency refinement.

3. PAW-Based Lightweight Network

The overall architecture follows an asymmetric encoder–decoder design. The encoder adopts LBB to extract compact semantic features with low complexity, while the decoder enhances reconstruction capability. To overcome the limitations of spatial-only attention, the PAW module is introduced to enable joint spatial–frequency feature refinement via multi-resolution wavelet decomposition.

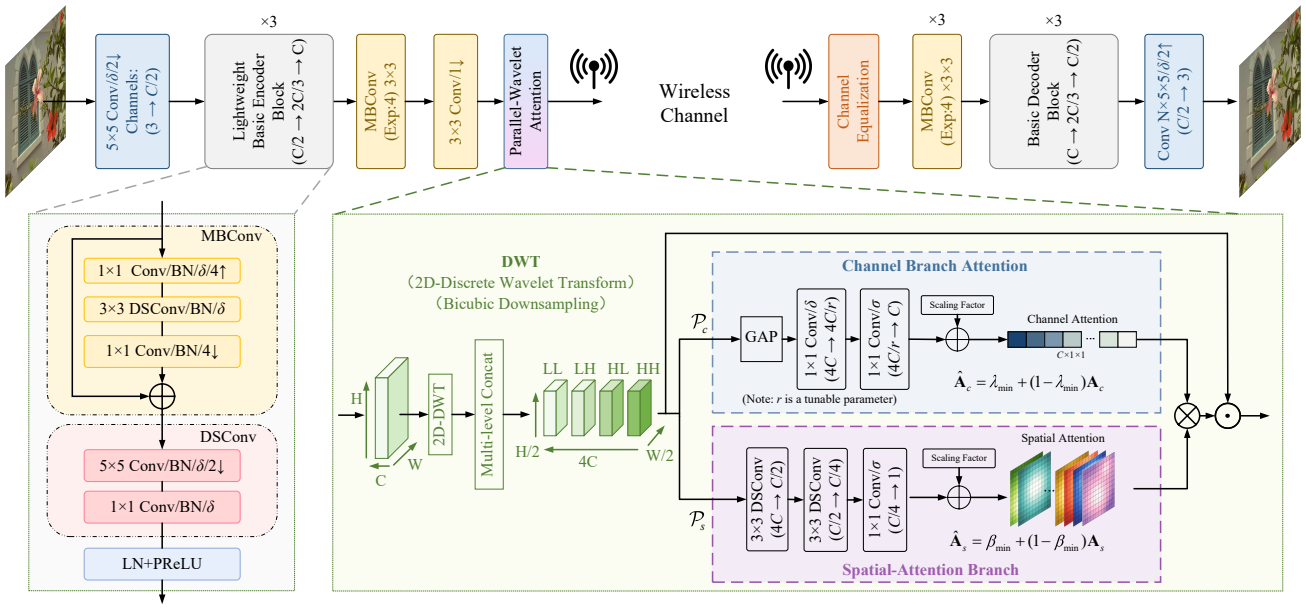

Figure 2. Overall architecture of the proposed PAW-based semantic communication. The detailed boxes highlight our key designs: the Lightweight Basic Encoder Block for efficient feature extraction, and the Parallel-Wavelet Attention module for semantic recalibration.

3.1. Parallel-Wavelet Attention Module

Unlike conventional attention modules that operate exclusively in the spatial domain, the proposed PAW module leverages 2D wavelet decomposition to perform joint recalibration across both spatial and channel domains. As shown in Fig. 2, by decomposing features into multi-resolution sub-bands, PAW effectively extracts fine-grained semantic details and global structures, which is critical for robust semantic transmission over noisy satellite channels.

3.1.1. Wavelet Decomposition

Given an input feature map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$, we first apply a 2D Haar Discrete Wavelet Transform (DWT), denoted as $\mathcal{W}(\cdot)$, to decompose it into four wavelet sub-bands:

$$\{\mathbf{F}_{LL}, \mathbf{F}_{LH}, \mathbf{F}_{HL}, \mathbf{F}_{HH}\} = \mathcal{W}(\mathbf{F}), \tag{7}$$

where $\mathbf{F}_{LL} \in \mathbb{R}^{C\times \frac{H}{2}\times \frac{W}{2}}$ represents the low-frequency approximation (global structure), and $\{\mathbf{F}_{LH}, \mathbf{F}_{HL}, \mathbf{F}_{HH}\}$ capture the horizontal, vertical, and diagonal high-frequency details, respectively. These sub-bands are concatenated along the channel dimension to form a frequency-aware feature tensor $\mathbf{F}_{freq} \in \mathbb{R}^{4C\times \frac{H}{2}\times \frac{W}{2}}$, effectively mapping the spatial information into a multi-resolution frequency space.

3.1.2. Parallel-Wavelet Attention Structure

The PAW module performs joint spatial-frequency recalibration by processing $\mathbf{F}_{freq}$ through parallel branches. This architecture independently optimizes the preservation of "what" (channel-wise) and "where" (spatial-wise) semantic information, which can be formally defined as a combination of two independent projections, $\mathcal{P}_c$ and $\mathcal{P}_s$.

Channel-Attention Branch The channel attention branch determines the relative significance of frequency-aware features by co-modulating all sub-band components. This process is denoted as the channel-wise semantic projection $\mathcal{P}_c$:

$$\mathbf{A}_c = \mathcal{P}_c(\text{GAP}(\mathbf{F}_{freq})), \tag{8}$$

where $\text{GAP}(\cdot)$ signifies global average pooling. Mathematically, $\mathcal{P}_c : \mathbb{R}^{4C} \to \mathbb{R}^{C}$ represents a non-invertible, low-rank projection for which no closed-form optimal solution exists. In our framework, a multi-layer perceptron (MLP) acts as a functional approximator to learn this mapping:

$$\mathbf{A}_c = \sigma(\text{Conv}_{1\times1}(\delta(\text{Conv}_{1\times1}(\text{GAP}(\mathbf{F}_{freq}))))), \tag{9}$$

where σ and δ denote the Sigmoid and ReLU functions, respectively. While this MLP-based $\mathcal{P}_c$ provides sufficient representational power, it can be reduced to various simplified forms for computational efficiency. For instance, the uniform sub-band averaging strategy is a static, non-parametric special case of our PAW framework (Yang et al., 2023). More details of these projection variants are provided in Appendix A.

To ensure semantic stability in low Signal-to-Noise Ratio (SNR) environments, we apply a scaling constraint:

$$\hat{\mathbf{A}}_c = \lambda_{min} + (1 - \lambda_{min}) \cdot \mathbf{A}_c, \tag{10}$$

where λ_{min} is a scaling factor that prevents excessive semantic suppression under adverse channel conditions.

Spatial-Attention Branch Simultaneously, the spatial branch localizes salient semantic regions via a spatial projection $\mathcal{P}_s$. Following the channel branch's logic, we employ convolutional blocks to learn this mapping:

$$\begin{aligned}\mathbf{A}_s &= \mathcal{P}_s(\mathbf{F}_{freq}) \\ &= \mathcal{U}(\sigma(\text{Conv}_{1\times1}(\text{DSConv}(\text{DSConv}(\mathbf{F}_{freq}))))),\end{aligned} \tag{11}$$

where $\mathcal{U}(\cdot)$ denotes bilinear interpolation used to restore the original spatial resolution. This design captures spatial-frequency dependencies with linear complexity relative to the channel dimension, ensuring hardware efficiency for spaceborne platforms. Similarly, $\mathbf{A}_s$ is constrained by:

$$\hat{\mathbf{A}}_s = \beta_{min} + (1 - \beta_{min}) \cdot \mathbf{A}_s, \tag{12}$$

where β_{min} is the spatial scaling factor.

3.1.3. Feature Fusion

The final refined feature map is generated by applying the learned channel-spatial weights through a cascaded fusion process. This ensures the model prioritizes both robust frequency channels and informative spatial regions:

$$\mathbf{F}_{out} = (\hat{\mathbf{A}}_c \otimes \hat{\mathbf{A}}_s) \odot \mathbf{F}, \tag{13}$$

where $\otimes$ and $\odot$ represent channel-wise scaling and the Hadamard product, respectively.

3.2. Lightweight Basic Encoder Block Design

To balance limited onboard resources with semantic transmission performance, we design LBB as the core encoder unit. As shown in Fig. 2, LBB adopts a hierarchical structure that jointly performs feature enhancement and spatial compression. Specifically, MBConv (Sandler et al., 2018) is employed for efficient feature refinement, while DSConv (Chollet, 2017) enables low-complexity spatial extraction and downsampling. By decoupling channel-wise and spatial operations, this design effectively reduces transmitter-side complexity while preserving critical edge and texture information during resolution reduction.

At the receiver side, the decoder incorporates higher-capacity reconstruction modules to compensate for compression loss and channel-induced distortions. Accordingly, the overall encoder–decoder architecture follows an asymmetric computation principle: the satellite-side encoder remains lightweight, whereas the receiver-side decoder performs more powerful semantic reconstruction. In this way, LBB serves as an efficient backbone for PAW, achieving a favorable trade-off between onboard computational cost and reconstruction quality.

4. Results and Analysis

In this section, we evaluate the proposed PAW-based asymmetric semantic communication framework in terms of reconstruction performance and computational efficiency. We first describe the experimental setup, including the datasets, hardware platform, channel settings, and comparison baselines. Then, we analyze the model complexity and compare the proposed method with representative baseline schemes under different channel conditions.

4.1. Experiment Setup

To evaluate the proposed framework for semantic image transmission, we adopt a mixed training strategy using three widely used datasets: Flickr30K (Young et al., 2014), Liu4K (Liu et al., 2020), and the CLIC2020 (Toderici et al., 2020) training set. Testing is performed on the Kodak dataset (Company, 1993) and the CLIC2020 test set, focusing on image reconstruction quality and perceptual fidelity.

All experiments are conducted on an NVIDIA RTX A6000 GPU. The latent channel dimension is set to $C = 48$, with a total of $16\times$ down-sampling ratio. For the PAW module, we empirically set the lower-bound scaling factors $\lambda_{\min} = 0.3$ and $\beta_{\min} = 0.3$ to maintain a robust semantic feature recalibration. The framework is evaluated under both stationary Additive White Gaussian Noise (AWGN) and dynamic Rayleigh fading channels.

Specifically, we evaluate three architectural configurations within our proposed PAW framework to investigate the impact of spectral representation and computational efficiency, as shown in Fig. 3:

- **PAW-Full**: A full-capacity version that combines the proposed parallel $4C$-channel wavelet attention with standard convolutions and Multi-Scale Residual Blocks (MSRB). It serves as a performance upper bound to evaluate the representation gain of PAW with a stronger backbone.
- **PAW-Lite**: A lightweight variant of PAW-Full, where standard convolutions and MSRB modules are replaced by MBConv and DSConv blocks. It targets a better efficiency–performance tradeoff for resource-constrained encoders.
- **PAW-Lite(Avg)**: A dimension-collapsed variant that averages the four wavelet sub-bands into a single C-dimensional space, as illustrated in Appendix A, to evaluate the benefit of parallel sub-band preservation.

We further compare the proposed PAW variants with the following representative baselines:

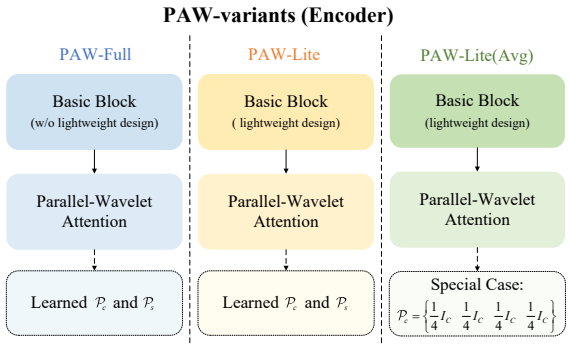

Figure 3. Architecture comparison of PAW variants: PAW-Full uses a standard backbone with parallel wavelet attention, PAW-Lite replaces it with lightweight MBConv/DSConv blocks, and PAW-Lite(Avg) simplifies PAW-Lite by averaging parallel subbands.

- **SwinJSCC** (Yang et al., 2025): A state-of-the-art semantic communication framework based on Swin Transformer blocks, with channel bandwidth ratio (CBR) set to $1/16$ for fair comparison..
- **BPG+LDPC** (Yang & Kim, 2022): A traditional separation-based scheme combining BPG image compression with LDPC channel coding (rate $2/3$).

4.2. Performance Comparison

We first compare the proposed PAW method with representative baselines in terms of image reconstruction performance, as shown in Fig. 4. The evaluation is conducted over both AWGN and Rayleigh channels, with particular focus on the low-SNR regime. The results in Fig. 4 show that PAW-Full exhibits strong robustness under severe channel impairments. When $\mathrm{SNR} \leq 1$ dB, PAW-Full consistently outperforms the baselines across different metrics. For example, under the AWGN channel, at $\mathrm{SNR} = -3$ dB and -5 dB, PAW-Full improves PSNR over SwinJSCC by 6.03 dB and 7.51 dB, respectively, and improves MS-SSIM by 0.19 and 0.18. Similar gains are observed under the Rayleigh channel, where PAW-Full improves PSNR by 3.64 dB and 7.64 dB and MS-SSIM by 0.15 and 0.09 at $\mathrm{SNR} = -3$ dB and -5 dB, respectively. Moreover, under the Rayleigh channel at $\mathrm{SNR} = -5$ dB, PAW-Full achieves a PSNR of 23.24 dB, whereas BPG+LDPC suffers from a performance collapse due to the cliff effect. This demonstrates that PAW-Full benefits from analog semantic transmission and its parallel wavelet-domain representation, which enables graceful degradation and preserves essential semantic structures under severe fading.

We further evaluate the lightweight variants, PAW-Lite and PAW-Lite(Avg). PAW-Lite achieves PSNR, MS-SSIM, and LPIPS performance close to PAW-Full over both channel types with reduced computational complexity, demonstrating robust performance. In contrast, PAW-Lite(Avg) shows

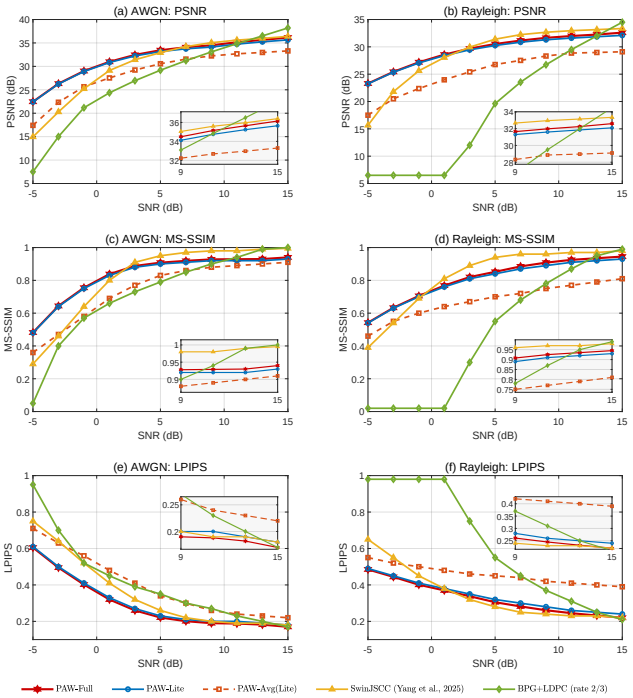

Figure 4. Performance comparison across various metrics in AWGN and Rayleigh channels.

more noticeable degradation, but still outperforms SwinJSCC in the low-SNR regime. For instance, under the AWGN channel at $\mathrm{SNR} = -5$ dB, PAW-Lite(Avg) reduces PSNR and MS-SSIM by 5dB and 0.12 compared with PAW-Lite, but still improves them by 2.46 dB and 0.07 over SwinJSCC, respectively. These results indicate that simple dimensional averaging weakens robustness, whereas preserving parallel frequency sub-band modeling is important for reliable semantic transmission.

4.3. Complexity Analysis

To assess the efficiency of the proposed PAW framework, we compare the encoder-side complexity of different configurations under a fixed input size of 256×256, focusing on the transmitter-side modules (i.e., encoder and importance modulation).

Table 1. Encoder-side complexity comparison of different models under an input size of 256×256.

Method	Params (M) ↓	FLOPs (G) ↓	GPU Runtime (ms) ↓
SwinJSCC	14.103	34.462	6.89
PAW-Full	3.489	23.735	4.58
PAW-Lite	0.724	5.821	1.95
PAW-Lite(Avg)	0.683	5.821	1.93

"↓" indicates that lower values are better.

As shown in Table 1, the PAW family is significantly more efficient than the Transformer-based SwinJSCC. Specifi-

Table 2. Ablation study under AWGN and Rayleigh channels on Kodak and CLIC datasets ($SNR = 7$dB).

Method	AWGN						Rayleigh					
	Kodak			CLIC			Kodak			CLIC		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
w/o PAW	26.124	0.6812	0.44	27.215	0.7014	0.48	21.345	0.4856	0.68	24.120	0.6128	0.62
with PAW-Lite(Avg)	28.542	0.7428	0.38	30.156	0.7852	0.39	24.218	0.5892	0.58	27.452	0.7245	0.54
with PAW-Lite	30.125	0.8154	0.28	32.458	0.8624	0.25	26.845	0.6854	0.48	29.754	0.7956	0.42

"↑" and "↓" indicate that higher and lower values are better, respectively.

cally, PAW-Full reduces the number of parameters from 14.103M to 3.489M and FLOPs from 34.462G to 23.735G, corresponding to reductions of 75.3% and 31.1%, respectively. Its GPU runtime is also reduced from 6.89 ms to 4.58 ms, indicating improved computational efficiency enabled by the parallel wavelet-domain design.

PAW-Lite further reduces the complexity by adopting depthwise separable convolutions and MBConv blocks. Compared with SwinJSCC, it reduces the number of parameters to 0.724M and FLOPs to 5.821G, achieving reductions of 94.9% and 83.1%, respectively. Its GPU runtime is further reduced to 1.95 ms, demonstrating a favorable trade-off between reconstruction performance and computational cost.

PAW-Lite(Avg) slightly reduces the parameter count to 0.683M with nearly the same computational cost as PAW-Lite. However, its reconstruction performance degrades noticeably, suggesting that the dimension-collapsed design is mainly suitable for highly complexity-sensitive scenarios.

Overall, PAW significantly reduces encoder-side complexity compared with SwinJSCC, among which PAW-Lite achieves the best efficiency–performance balance.

4.4. Ablation Studies

The ablation results in Table 2 quantify the contribution of the proposed PAW design under both AWGN and Rayleigh channels. Compared to the baseline without PAW, all PAW variants achieve substantial performance gains across datasets and metrics. For instance, under AWGN on the Kodak dataset, PSNR improves from 26.12 dB to 30.41 dB, accompanied by consistent gains in SSIM and reductions in LPIPS. This verifies the effectiveness of introducing wavelet-domain attention for semantic feature refinement.

Comparing different PAW variants further highlights the importance of parallel spectral modeling. The dimension-collapsed version, PAW-Lite(Avg), provides moderate improvements over the baseline but remains significantly inferior to PAW-Lite, with a gap of nearly 1.6 dB PSNR on Kodak under AWGN. This indicates that simple frequency aggregation cannot fully exploit the discriminative information across sub-bands. In contrast, explicitly preserving parallel frequency branches enables more effective feature representation, leading to consistent gains across all metrics.

Finally, the advantage of PAW becomes more pronounced under Rayleigh fading. While all methods experience performance degradation, PAW-Lite and PAW-Full maintain significantly higher reconstruction quality compared to both the baseline and PAW-Lite(Avg). For example, on Kodak, PSNR improves from 21.35 dB (w/o PAW) to 26.85 dB (PAW-Lite), demonstrating strong robustness to channel fluctuations. The relatively small gap between PAW-Lite and PAW-Full further suggests that the proposed lightweight design already captures most of the representation benefits. Overall, these results confirm that parallel wavelet-based feature decomposition not only enhances reconstruction quality but also improves resilience to complex channel conditions.

5. Conclusions and Future Work

In this work, we proposed PAW, a lightweight asymmetric semantic image communication framework for resource-constrained 6G space networks. Although wavelet transforms are classical signal-processing tools, their role in semantic communications remains underexplored. To the best of our knowledge, this work is among the first to showcase how wavelet-domain decomposition can be explicitly aligned with asymmetric semantic communication, where fine-grained *channel & spatial* attentions are exploited for better semantic feature extraction.

By combining PAW with an edge–cloud model partition, our framework shifts computationally intensive reconstruction to AI data centers while keeping the satellite-side encoder lightweight. Experiments under Rayleigh fading channels demonstrate that PAW reduces encoder-side GPU runtime by up to **3.5**$\times$ while improving reconstruction quality and robustness over Swin-based baselines. These results highlight that parallel *channel&spatial* learning is an effective paradigm for practical space semantic communications.

Beyond wavelet-based modeling, our results motivate future studies on other structured representation techniques, such as multi-resolution transforms, graph-based representations, and frequency-domain neural operators, together with asymmetric edge–cloud co-design for robust and deployable semantic communication in future 6G space networks.

References

Ahn, G., Seo, J., and Kang, J. VLF-MSC: Vision-language feature-based multimodal semantic communication system. In *NeurIPS 2025 Workshop on AI and ML for Next-Generation Wireless Communications and Networking (AI4NextG)*, 2025.

Bourtsoulatze, E., Burth Kurka, D., and Gündüz, D. Deep Joint Source-Channel Coding for Wireless Image Transmission. *IEEE Transactions on Cognitive Communications and Networking*, 5(3):567–579, September 2019. ISSN 2332-7731.

Chollet, F. Xception: Deep Learning With Depthwise Separable Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1251–1258, 2017.

Company, E. K. Kodak lossless true color image suite. http://r0k.us/graphics/kodak/, 1993.

Dai, J., Wang, S., Tan, K., Si, Z., Qin, X., Niu, K., and Zhang, P. Nonlinear Transform Source-Channel Coding for Semantic Communications. *IEEE Journal on Selected Areas in Communications*, 40(8):2300–2316, August 2022. ISSN 1558-0008.

Guo, B., Du, M., He, X., Zhang, Z., Li, B., and Xiong, Z. Semantic Communication in 6G Mega-Satellite Networks: System Design and Standardization. *IEEE Communications Standards Magazine*, 9(4):65–71, December 2025. ISSN 2471-2833.

Huang, D., Gao, F., Tao, X., Du, Q., and Lu, J. Toward Semantic Communications: Deep Learning-Based Image Semantic Coding. *IEEE Journal on Selected Areas in Communications*, 41(1):55–71, January 2023. ISSN 1558-0008.

Khalid, R. A. B., Freire, P., Turitsyn, S. K., and Prilepsky, J. E. Efficient and Robust Semantic Image Communication via Stable Cascade. In *ICML 2025 Workshop on Machine Learning for Wireless Communication and Networks (ML4Wireless)*, 2025.

Leonard, C., Stober, D., and Schulz, M. FPGA-Enabled Machine Learning Applications in Earth Observation: A Systematic Review. *ACM Comput. Surv.*, 58(11):283:1–283:36, April 2026. ISSN 0360-0300.

Liu, J., Liu, D., Yang, W., Xia, S., Zhang, X., and Dai, Y. A Comprehensive Benchmark for Single Image Compression Artifact Reduction. *IEEE Transactions on Image Processing*, 29:7845–7860, 2020. ISSN 1941-0042.

Nguyen, L. X., Tun, Y. L., Tun, Y. K., Nguyen, M. N. H., Zhang, C., Han, Z., and Seon Hong, C. Swin Transformer-Based Dynamic Semantic Communication for Multi-User With Different Computing Capacity. *IEEE Transactions on Vehicular Technology*, 73(6):8957–8972, June 2024. ISSN 1939-9359.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018.

Toderici, G., Shi, W., Theis, L., Johnston, N., Ballé, J., Agustsson, E., Minnen, D., van Rozendaal, T., and Timofte, R. Workshop on learned image compression (clic). In *CVPR Workshops*, 2020.

Wu, T., Chen, Z., He, D., Qian, L., Xu, Y., Tao, M., and Zhang, W. CDDM: Channel Denoising Diffusion Models for Wireless Semantic Communications. *IEEE Transactions on Wireless Communications*, 23(9):11168–11183, September 2024. ISSN 1558-2248.

Yang, K., Wang, S., Dai, J., Qin, X., Niu, K., and Zhang, P. SwinJSCC: Taming Swin Transformer for Deep Joint Source-Channel Coding. *IEEE Transactions on Cognitive Communications and Networking*, 11(1):90–104, February 2025. ISSN 2332-7731.

Yang, M. and Kim, H.-S. Deep Joint Source-Channel Coding for Wireless Image Transmission with Adaptive Rate Control. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5193–5197. IEEE, 2022.

Yang, Y., Jiao, L., Liu, X., Liu, F., Yang, S., Li, L., Chen, P., Li, X., and Huang, Z. Dual Wavelet Attention Networks for Image Classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(4):1899–1910, April 2023. ISSN 1558-2205.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, February 2014. ISSN 2307-387X.

Zhang, M., Wu, H., Zhu, G., Jin, R., Chen, X., and Gündüz, D. Semantics-Guided Diffusion for Deep Joint Source-Channel Coding in Wireless Image Transmission. *IEEE Transactions on Wireless Communications*, 25:1547–1564, 2026a. ISSN 1558-2248.

Zhang, P., Niu, K., Liang, Z., Wang, C., Wu, J., Liu, Y., Xu, W., Ma, N., Xu, X., and Zhang, R. Beyond Shannon: Semantic Information Theory and Methodology. *IEEE Transactions on Network Science and Engineering*, 13: 8062–8079, 2026b. ISSN 2327-4697.

A. PAW Channel Projection Variants and Averaging Special Case

In this section, we provide a formal derivation to show that the uniform sub-band averaging strategy (specifically instantiated as PAW-Lite(Avg) in our experiments) is a non-parametric special case of our generalized framework.

A.1. PAW-Lite(Avg) as a Uniform Sub-Band Projection

In previous work (Yang et al., 2023), the channel-wise importance weight W_c is calculated by summing the spatial components across all frequency sub-bands:

$$W_c = \sum_{i=0}^{M/2} \sum_{j=0}^{N/2} (LL_{i,j} + LH_{i,j} + HL_{i,j} + HH_{i,j}), \tag{14}$$

where $LL, LH, HL, HH \in \mathbb{R}^{C \times \frac{H}{2} \times \frac{W}{2}}$. To align this with our notation, we define the frequency-aware feature tensor as $\mathbf{F}_{freq} = [LL, LH, HL, HH] \in \mathbb{R}^{4C \times \frac{H}{2} \times \frac{W}{2}}$. The Global Average Pooling (GAP) operation on $\mathbf{F}_{freq}$ yields a concatenated vector $\mathbf{v} \in \mathbb{R}^{4C}$, which can be partitioned into four sub-vectors $\{\mathbf{v}_{LL}, \mathbf{v}_{LH}, \mathbf{v}_{HL}, \mathbf{v}_{HH}\}$, each in $\mathbb{R}^C$.

If we define the semantic projection matrix as $\mathcal{P}_{avg} = [\frac{1}{4}\mathbb{I}_C, \frac{1}{4}\mathbb{I}_C, \frac{1}{4}\mathbb{I}_C, \frac{1}{4}\mathbb{I}_C] \in \mathbb{R}^{C \times 4C}$, where $\mathbb{I}_C$ is the $C \times C$ identity matrix, the resulting channel attention $\mathbf{A}_c$ becomes:

$$\mathbf{A}_c = \mathcal{P}_{avg} \cdot \text{GAP}(\mathbf{F}_{freq}) = \frac{1}{4}(\mathbf{v}_{LL} + \mathbf{v}_{LH} + \mathbf{v}_{HL} + \mathbf{v}_{HH}). \tag{15}$$

This formulation is mathematically equivalent to the normalized version of Eq. (14). Thus, the strategy in (Yang et al., 2023), i.e. PAW-Lite(Avg), represents a static, uniform dimensional collapse of the frequency information.

A.2. Why Learnable Channel Projection Matters

The projection $\mathcal{P}_c : \mathbb{R}^{4C} \to \mathbb{R}^C$ performs a significant dimensionality reduction. From a linear algebra perspective, since $\text{rank}(\mathcal{P}_c) \leq C < 4C$, the mapping $\mathcal{P}_c$ possesses a non-trivial null space, rendering it inherently non-invertible. Consequently, there is no unique closed-form optimal solution for $\mathcal{P}_c$ that perfectly preserves the most salient semantic information across all communication scenarios.

While the static PAW-Lite(Avg) provides a robust baseline by treating all frequency components equally, it lacks the flexibility to adapt to dynamic channel conditions or varying source content. Our framework directly treats $\mathcal{P}_c$ as a learnable function (e.g., an MLP-based approximator), allowing the network to adaptively navigate this high-dimensional projection space. Identifying more sophisticated architectures to better approximate $\mathcal{P}_c$ remains a promising direction for further enhancing the performance and efficiency of wavelet-integrated semantic communication.