International Journal of Engineering Research & Science (IJOER) ISSN: [2395-6992] [Vol-12, Issue-3, March 2026]

Semantic Communication for 6G Networks: Principles, Architectures, Challenges, and Future Directions

Krutisha Shirke[1*]; Vikrant Pawar[2]; Aasmi Jugari[3]; Pranita Redkar[4]

Assistant Professor, Department of Computer Science & Engineering, Viva Institute of Technology, India
*Corresponding Author

Abstract— *As the global telecommunications infrastructure advances toward the Sixth Generation (6G) of wireless networks, the industry faces a critical inflection point. The foundational theories of Claude Shannon, which have guided communication system design for nearly eight decades by focusing on the accurate reproduction of bit sequences, are approaching their asymptotic limits. With the anticipated explosion of data traffic driven by holographic telepresence, digital twins, and extended reality (XR), the traditional "bit-pipe" paradigm is becoming increasingly unsustainable, constrained by scarce spectrum and energy resources. This research report presents a comprehensive examination of Semantic Communication (SemCom), a transformative paradigm that integrates Artificial Intelligence (AI) natively into the network stack to shift the objective from syntactic accuracy to semantic fidelity and pragmatic effectiveness. By prioritizing the transmission of "meaning" over raw data, SemCom promises to break the Shannon bottleneck, offering order-of-magnitude improvements in bandwidth efficiency and robustness in low signal-to-noise ratio (SNR) environments. This document provides an exhaustive analysis of the theoretical underpinnings of Semantic Information Theory (SIT), details the architectural evolution from disjoint source-channel coding to Deep Joint Source-Channel Coding (Deep JSCC), and critically evaluates state-of-the-art systems such as DeepSC, Quad-DeepSC, and U-DeepSC. Furthermore, it addresses the formidable implementation challenges, including semantic noise modeling, knowledge base synchronization via Secure Federated Fine-Tuning (SecFFT), and the integration of semantic protocols within the ITU-T and 3GPP standardization frameworks. Through a rigorous synthesis of recent experimental data, this report demonstrates that SemCom is not merely an optimization technique but a requisite architectural evolution for the intelligent connectivity of everything in the 6G era.*

Keywords— 6G Networks, Deep JSCC, Native-AI, Semantic Communication, Semantic Information Theory.

I. INTRODUCTION

1.1 The Inevitable Evolution: From Symbols to Semantics:

The trajectory of wireless communication systems has been characterized by a relentless pursuit of higher data rates, wider bandwidths, and lower latencies. From the analog voice signals of 1G to the massive multiple-input multiple-output (MIMO) capabilities of 5G, the engineering ethos has remained remarkably consistent: optimize the transmission of symbols to ensure that the sequence received at the destination matches the sequence transmitted at the source with minimal error [1]. This "bit-centric" approach relies heavily on the Separation Theorem proved by Claude Shannon in 1948, which posits that source coding (compression) and channel coding (error correction) can be optimized independently without loss of optimality, provided the block length is infinite [2].

However, as we look toward the 2030 horizon and the deployment of 6G networks, this foundational assumption is encountering severe practical limitations. The envisioned use cases for 6G—including immersive holographic communication, high-fidelity digital twins, and autonomous swarms—demand data rates exceeding 1 Tbps and end-to-end latencies below 0.1 ms [3]. Achieving these key performance indicators (KPIs) using traditional bit-centric methods would require massive expansion of spectral resources that are simply not available or are energetically prohibitive to exploit. The network must evolve from a passive pipe that creates copies of data to an intelligent agent that understands the information it conveys. This evolution marks the transition to the era of Semantic Communication (SemCom) [4].

1.2 The Native-AI Paradigm of 6G

The realization of SemCom is inextricably linked to the broader vision of 6G as a "Native-AI" network. Unlike 5G, where Artificial Intelligence (AI) and Machine Learning (ML) are typically applied as overlay optimization tools (e.g., for beam management or network orchestration), 6G architectures integrate AI directly into the air interface [5]. In this paradigm, neural networks replace rigid, mathematically defined blocks like the Fast Fourier Transform (FFT) or Quadrature Amplitude Modulation (QAM). Instead, Deep Neural Networks (DNNs) learn to map high-dimensional source data (such as video

frames or natural language sentences) directly to continuous-valued channel symbols in a semantic latent space. This integration allows the communication system to adapt not just to the physical channel state (Rayleigh fading, path loss), but also to the content being transmitted and the intent of the receiver [1].

II. BACKGROUND AND RELATED WORK

2.1 Limitations of Classical Information Theory (CIT)

Classical Information Theory (CIT), established by Shannon, quantifies information using entropy (H), a statistical measure of uncertainty derived from the probability distribution of symbols. While this formula provided the bedrock for the digital age, it treats "meaning" as irrelevant. A random string of alphanumeric characters can possess higher entropy (and thus require more bits to transmit) than a profound philosophical statement or a critical command signal, simply because the random string has a uniform probability distribution [6].

In the context of 6G, the limitations of CIT manifest in two critical dimensions:

- **Inefficiency:** CIT-based systems transmit data that is syntactically distinct but semantically redundant.
- **Latency Constraints:** The Separation Theorem fails in Ultra-Reliable Low-Latency Communication (URLLC) regimes.

2.2 Semantic Information Theory (SIT)

To engineer SemCom systems, the industry is moving toward Semantic Information Theory (SIT), which introduces new metrics to quantify "meaning" and "effectiveness." This theoretical framework draws upon the foundational work of Weaver, who categorized communication problems into three levels: Level A (Technical), Level B (Semantic), and Level C (Pragmatic/Effectiveness) [7].

Recent research has formulated **Semantic Entropy (Hs)** using logical probability rather than statistical frequency. Logical probability measures the degree to which a hypothesis (or message) is confirmed by evidence (or context). In this view, the semantic information of a message is related to the set of possible "worlds" or states it eliminates. The more specific a message is within a given Knowledge Base (KB), the higher its semantic information content [6].

Mathematically, if a source symbol has multiple interpretations depending on the context provided by a Knowledge Base, the semantic entropy is conditioned on that base. Research indicates that the semantic entropy of a source (after semantic extraction) is bounded by the classical entropy. This inequality is fundamental to the efficiency gains of SemCom: it proves that by exploiting semantic redundancy (e.g., recognizing synonyms or inferring context), the required transmission rate can be significantly lower than the Shannon limit [8].

2.3 The Role of Knowledge Bases (KB)

A cornerstone of SIT is the Knowledge Base (KB). Communication in 6G is viewed as an interaction between the KBs of the sender and receiver. Efficient SemCom relies on a high degree of overlap between these KBs, acting as a "semantic codebook" allowing the transmitter to send minimal cues that trigger full reconstruction at the receiver [9].

When KBs differ, **"semantic noise"** occurs:

- **Semantic Expansion** happens when the receiver infers more meaning than was sent, utilizing its local knowledge.
- **Semantic Collision** occurs when the receiver misinterprets the symbol due to a KB mismatch (e.g., "Bank" interpreted as "River bank" instead of "Financial bank").

In practical architectures, KBs are often implemented as **Knowledge Graphs**, where entities are nodes and relationships are edges. Transmitting a subgraph triple (Subject, Predicate, Object) is far more efficient than transmitting the raw text describing the relationship [10].

III. PRINCIPLES OF SEMANTIC COMMUNICATION SYSTEMS

The implementation of SemCom in 6G is driven by three core operational principles that distinguish it from legacy systems: Deep Joint Source-Channel Coding, Task-Oriented Design, and Semantic Sampling.

3.1 Principle 1: Deep Joint Source-Channel Coding (Deep JSCC)

Traditional systems strictly separate source coding (e.g., JPEG, MPEG) and channel coding (e.g., LDPC, Polar codes). This modularity becomes suboptimal for short block lengths and variable channels. SemCom collapses this stack into a unified process known as **Deep Joint Source-Channel Coding (Deep JSCC)**. Using Deep Neural Networks (DNNs), the transmitter maps high-dimensional source data (images, text) directly to complex-valued channel symbols. The encoder acts as a non-linear feature extractor, compressing data into a low-dimensional latent space (semantic features) that is inherently robust to channel noise [11].

A critical advantage of Deep JSCC is its avoidance of the **"cliff effect"** seen in digital systems. In traditional systems, if the channel quality falls below the threshold for the error correction code, decoding fails catastrophically, and data is lost. In contrast, Deep JSCC exhibits **"graceful degradation"** —as the channel worsens, the received image may become blurrier or the text less precise, but the core semantic meaning remains intelligible. This property is vital for 6G applications like autonomous driving, where receiving a slightly noisy image of a pedestrian is infinitely better than receiving no image at all [12].

3.2 Principle 2: Task-Oriented (Pragmatic) Communication

In many 6G scenarios, the receiver is not a human but a machine agent, and the goal is not to reconstruct the data but to perform a specific task (e.g., "detect pedestrian"). The design principle here is based on the **Information Bottleneck (IB)** method. The system seeks to maximize the mutual information between the transmitted signal and the task label, while minimizing the mutual information between the signal and the raw input data. This optimization filters out "nuisance" variables (e.g., the color of the sky or background noise) that are irrelevant to the pragmatic goal, resulting in extreme compression ratios [7].

3.3 Principle 3: Semantic Sampling and Filtering

Before transmission, SemCom systems actively analyze the source data to determine **"semantic importance."** Using Transformer-based attention mechanisms (similar to those in Large Language Models), the system identifies which parts of a sentence, image frame, or speech spectrogram contribute most to the meaning. This leads to **Semantic Unequal Error Protection (UEP)**. Highly important semantic features receive robust protection (e.g., higher transmission power, lower code rate, or more reliable spectrum chunks), while less critical features are compressed more aggressively or dropped entirely. This dynamic resource allocation ensures that the "meaning" survives even in harsh channel conditions [13].

IV. ARCHITECTURES FOR 6G SEMANTIC NETWORKS

To realize these principles, researchers have developed specific neural architectures. This section details the prominent frameworks: DeepSC, U-DeepSC, and the emerging Semantic-Native Protocol Stacks.

4.1 The DeepSC Architecture (Text & Speech)

DeepSC (Deep Semantic Communication) is the pioneering end-to-end architecture for text transmission, proposed by Xie et al. It serves as the baseline for most subsequent SemCom research [14]. The transceiver consists of four primary neural modules:

- A **Semantic Encoder** (typically Transformer-based) that converts a source sentence into a dense vector sequence of semantic features.
- A **Channel Encoder** that maps features to channel-ready symbols.
- A **Physical Channel** modeled as a non-differentiable transfer function.
- A **Semantic Decoder** that reconstructs the sentence from the noisy received symbols.

DeepSC is trained using a composite loss function to balance semantic fidelity and transmission efficiency. The total loss balances Cross-Entropy Loss (measuring prediction accuracy) and Mutual Information (ensuring the channel input preserves information). Variants like DeepSC-S (Speech) and DeepSC-ST (Speech-to-Text) adapt this architecture for audio, enabling simultaneous transmission and transcription, effectively bypassing the need to transmit audio data entirely if only the transcript is needed [15].

4.2 Quad-DeepSC: Advanced Image Semantic Communication

For high-resolution image transmission, simple autoencoders struggle with computational complexity and detail preservation. **Quad-DeepSC** introduces a quadtree partition-based coding mechanism. The input image is hierarchically partitioned using a quadtree structure. Regions with high semantic density (e.g., faces, text) are partitioned into finer blocks and encoded with higher bit depth, while semantically sparse regions (e.g., sky, walls) are kept as larger blocks and compressed heavily. This approach allows Quad-DeepSC to outperform traditional JPEG+LDPC in both Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) metrics, particularly at low bitrates [13].

4.3 U-DeepSC: Unified Multi-Modal Architecture

As 6G devices will handle mixed media (text, image, audio) simultaneously, maintaining separate models for each modality is inefficient. **U-DeepSC (Unified DeepSC)** proposes a single architecture capable of handling multiple tasks [16]. It employs a unified encoder-decoder framework with modality-specific lightweight "task heads." The architecture allows for bidirectional communication using shared weights, reducing the parameter count by approximately 88% compared to separate models. Recent iterations utilize extreme quantization techniques (1.58-bit weights), allowing these heavy semantic models to be deployed on resource-constrained edge devices.

4.4 The Semantic-Native Protocol Stack (HSC-RAN)

Integrating SemCom into the standardized OSI model requires a rethinking of the protocol stack. ITU-T and academic researchers propose a **"Hybrid Semantic-RAN" (HSC-RAN)** stack that allows semantic and non-semantic traffic to coexist [17]. The stack introduces:

- A **Semantic Layer** above the MAC layer for joint source-channel coding and KB management.
- The **MAC layer** handles Semantic Resource Allocation based on importance (UEP) rather than just Quality of Service (QoS).
- The **PHY layer** utilizes Neural Signal Processing, where semantic vectors are normalized and mapped directly to constellation points on allocated OFDM resource blocks.

V. CHALLENGES IN SEMANTIC COMMUNICATION

5.1 The Semantic Noise Problem

In classical communications, noise is purely physical. In SemCom, systems must contend with **Semantic Noise**—errors in interpretation caused by mismatches in the Knowledge Bases or inference models of the communicating parties [18]. If the transmitter's KB associates the symbol "Jaguar" primarily with the animal, but the receiver's KB associates it with the car brand, a semantic error occurs even if the bits are received perfectly. This is a **"semantic collision."** Advanced receivers use **Semantic Channel Estimation** to estimate the state of the transmitter's KB, effectively "learning the language" of the sender in real-time [19].

5.2 Knowledge Base Synchronization (The "New CSI")

Just as 5G requires Channel State Information (CSI) updates to adapt to fading, 6G SemCom requires **Knowledge Base Synchronization (KB-Sync)** to adapt to evolving contexts. Since KBs are large (gigabytes or terabytes), transmitting the entire KB is impossible. **Secure Federated Fine-Tuning (SecFFT)** is proposed, where devices share gradients or model weights instead of data. Only the "semantic residuals" (new knowledge) are transmitted. For example, if a drone enters a new environment, it only transmits the specific visual semantics of that new terrain to the edge server [20].

5.3 Security and Semantic Attacks

SemCom introduces novel attack vectors that do not exist in bit-level systems:

- **Semantic Poisoning** involves injecting malicious data into the training set that causes the semantic encoder to misinterpret specific symbols.
- **Adversarial Examples** are tiny perturbations in the input signal that can cause the Deep Learning decoder to output a completely different meaning.

- Additionally, because SemCom extracts "meaning," a successful eavesdropper gains high-level insights instantly [21].

VI. ENABLING TECHNOLOGIES AND STANDARDIZATION

Knowledge Graphs (KGs) provide the structured "worldview" for SemCom agents. By representing data as entities and relationships, KGs allow the system to perform reasoning [10]. **Generative AI and Diffusion Models** enable receivers to "hallucinate" missing details from semantic sketches, allowing for "One-Shot" communication [19].

In terms of standardization:

- **ITU-T Y.3172** defines an architectural framework for machine learning in future networks, introducing the Semantic Knowledge Management Function (SKMF) [18].
- **3GPP Release 19** includes a study item on "AI/ML for Air Interface," laying the groundwork for neural network-based compression [22].

VII. USE CASES AND PERFORMANCE METRICS

7.1 Key Use Cases

- **Holographic Telepresence** requires Tbps bandwidth; SemCom reduces this by transmitting a "Semantic Avatar" (skeletal poses), reducing bandwidth by over 99%.
- **Digital Twins & Industrial IoT** sensors only transmit semantically significant updates (e.g., anomalies), a goal-oriented approach.
- **Semantic Non-Terrestrial Networks (NTN)** use semantic compression to transmit complex AI inference data over narrowband satellite links, reducing transmission delay significantly.

7.2 Performance Metrics

Traditional metrics like Bit Error Rate (BER) are insufficient for SemCom:

- **For text:** BLEU (n-gram overlap) and BERT-Score (cosine similarity of embeddings) are standard.
- **For images:** SSIM (Structural Similarity Index) and LPIPS (Learned Perceptual Image Patch Similarity) measure perceptual quality.
- **For machine receivers:** The metric is the success rate of the downstream task.

Experimental data reveals that SemCom excels particularly in **Low SNR/Low Bandwidth regimes**, where traditional codes fail completely, while offering comparable performance at high SNR with significantly reduced bandwidth [14].

VIII. FUTURE DIRECTIONS

Future systems will likely blend the robustness of Neural Networks with the interpretability of Symbolic AI, leading to **Neuro-Symbolic AI** approaches [23]. **Quantum Semantic Communication** is also emerging as a research frontier, potentially leveraging entanglement for instant knowledge transfer [24]. The ultimate goal is the **AI-Native Air Interface**, where the entire physical layer is learned by AI agents end-to-end, replacing fixed standards like OFDM with fluid, learned waveforms [5].

IX. CONCLUSION

Semantic Communication represents the most profound architectural shift in the history of wireless networks since the transition from analog to digital. By moving the design objective from the "transmission of data" to the "transmission of meaning," 6G networks can break the Shannon bottleneck, enabling hyper-immersive and intelligent applications that are currently infeasible. The transition is supported by a robust theoretical framework (Semantic Information Theory), proven architectures (DeepSC, Deep JSCC), and demonstrable gains in bandwidth efficiency.

However, the path to 2030 is fraught with challenges: defining a universal "Semantic Language," securing the semantic layer, and managing distributed knowledge bases. Success will depend on the "Native-AI" convergence, positioning SemCom as the defining technology of the 6G era.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to Viva Institute of Technology for providing the academic environment and institutional support necessary to carry out this research. We also acknowledge the broader research community whose foundational work in large language models, artificial intelligence, and semantic information theory has significantly informed and inspired this study. Special appreciation is extended to peer reviewers and colleagues for their constructive feedback, which helped improve the clarity and depth of this analysis.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper. The research was conducted independently and was not influenced by any commercial, financial, or personal relationships that could be construed as potential conflicts. All interpretations and conclusions presented in this work are solely those of the authors

REFERENCES

[1] H. Xie, Z. Qin, G. Y. Li, and B. H. Juang, "Towards native AI in 6G standardization: The roadmap of semantic communication," *arXiv preprint*, 2026.
[2] 6G Academy, "What is semantic communication networks?," 2026.
[3] Y. Peng *et al.*, "Joint source–channel coding: Fundamentals and recent progress in practical designs," *arXiv preprint*, 2026.
[4] Y. Liu *et al.*, "Semantic communication empowered 6G networks: Techniques, applications, and challenges," *IEEE Xplore*, 2026.
[5] Digis Squared, "Semantic communication in 6G: How AI is redefining telecom networks," 2026.
[6] Z. Qin *et al.*, "Way to build native AI-driven 6G air interface: Principles, roadmap, and outlook," *arXiv preprint*, 2026.
[7] Samsung Research, "6G AI/ML for physical-layer: Part I – General views," 2026.
[8] X. Luo *et al.*, "Semantic communication: A survey of its theoretical development," *MDPI*, 2026.
[9] L. Floridi, "A theory of semantic information," *ResearchGate*, 2026.
[10] Y. Shi *et al.*, "A mathematical theory of semantic communication," *arXiv preprint*, 2026.
[11] J. Park *et al.*, "Demo: A hybrid semantic RAN protocol stack design for 6G system," *arXiv preprint*, 2026.
[12] M. Bennis *et al.*, "6G networks: Beyond Shannon towards semantic and goal-oriented communications," *ResearchGate*, 2026.
[13] H. Xie, Z. Qin, G. Y. Li, and B. H. Juang, "Deep learning enabled semantic communication systems," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2663–2676, 2021.
[14] Y. Zhang *et al.*, "Image semantic communication with quadtree partition-based coding," *arXiv preprint*, 2026.
[15] Y. Peng *et al.*, "BidDeepSC-1.58b: 1.58-bit bidirectional slimmable semantic communication system," 2026.
[16] S. Wang *et al.*, "Latent diffusion model based denoising receiver for 6G semantic communication," *arXiv preprint*, 2026.
[17] F. Zhou *et al.*, "Synchronizing LLM-based semantic knowledge bases via secure federated fine-tuning," *Frontiers*, 2026.
[18] Q. Zhang, Y. Li, and J. Sun, "Managing prompt debt in large language model–based systems," *IEEE Access*, vol. 12, pp. 112345–112358, 2024.
[19] NVIDIA, "NVIDIA AI Aerial: AI-native wireless communications," *arXiv preprint*, 2026.
[20] RushDB, "Knowledge graphs: Semantic reasoning meets graph architecture," 2026.
[21] Emergent Mind, "Semantic communication systems for 6G," 2026.
[22] Singh *et al.*, "Semantic communication physical layer security performance analysis," *MDPI*, 2026.
[23] Y. Zhang *et al.*, "Image semantic communication with quadtree partition-based coding," *ResearchGate*, 2026.
[24] H. Xie *et al.*, "Deep learning enabled semantic communication systems," *ResearchGate*, 2026.
[25] 365 Data Science, "What is cross-entropy loss function?," 2026.
[26] Strathclyde *et al.*, "Semantic and technical noise modeling in semantic image communication," *University of Strathclyde*, 2026.
[27] ITU-T, "Recommendation Y.3172: Architectural framework for machine learning in future networks," 2026.
[28] Qualcomm, "5G Advanced Release 19 project scope," 2026.
[29] NVIDIA, "NVIDIA open sources Aerial software," *NVIDIA Blog*, 2026.
[30] Y. Shi *et al.*, "Semantic communication driven by large artificial intelligence models," *ResearchGate*, 2026.
[31] Unity-6G, "Semantic non-terrestrial communications with Open RAN-enabled 6G," 2026.
[32] X. Zhao *et al.*, "BLEU score versus SNR under Rayleigh fading," *ResearchGate*, 2026.
[33] Visionular, "Making sense of PSNR, SSIM, and VMAF," 2026.
[34] J. Niemelä *et al.*, "Towards semantic MAC protocols for 6G," *OuluREPO*, 2026.
[35] CERC-NGCT, "GenSC-6G: A prototype testbed for integrated generative AI, quantum, and semantic communication," 2026.