Semantic Communications for 6g Networks: Theory, Architectures, and Challenges

Jie Ji

Portland Institute, Nanjing University of Posts and Telecommunications, Nanjing, 210023, China

2583184264@qq.com

Abstract. As 6G communication systems advance toward the grand vision of intelligent connectivity, the conventional bit-centric communication paradigm is encountering multiple bottlenecks in energy efficiency, spectrum utilization, and security when addressing the extreme demands of millisecond-level latency, terabit-per-second data rates, and massive connectivity. Semantic communication, as a paradigm shift from transmitting bits to conveying meaning, offers a critical pathway to overcoming these limitations. This survey provides a systematic review and forward-looking analysis of semantic communication technologies for 6G. We clarify the core value of semantic communication—namely, achieving a qualitative leap in communication efficiency and network intelligence by intelligently extracting and transmitting task-relevant semantic information. Furthermore, we examine three key technological evolution paths: interpretable approaches grounded in traditional model-based optimization; dedicated, efficient schemes driven by deep learning; and cognitive communication paradigms empowered by large-scale foundation models. For each path, we present a critical analysis of its strengths, limitations, and ongoing debates. We then summarize four representative 6G architectures enabled by semantic communication, elucidating their design goals, technical distinctions, and inherent trade-offs. Finally, we identify the open research challenges and outline promising future research directions. This survey aims to serve as a comprehensive reference for both academia and industry, fostering the deep integration of semantic communication technologies in the 6G era.

Keywords: Semantic Communication, 6G Networks, Cross-Domain Collaboration, Lightweight Models.

1. Introduction

With the accelerated pace of digitalization and intelligence, 6G networks are expected to play a pivotal role in supporting more extensive and deeper connectivity. By 2030, 6G is projected to serve billions of terminal devices, covering scenarios such as augmented reality (AR), virtual reality (VR), the metaverse, and smart industries. Compared to 5G, 6G aims not only for higher peak data rates and lower latencies but also emphasizes ubiquitous coverage, ultra-reliable connectivity, intelligent networks, and green, low-carbon capabilities. These advancements position 6G as a critical engine for powering future digital societies. However, when confronted with the massive data scale of the 6G era, traditional communication systems fall short in terms of transmission efficiency. The large volume of redundant bit-level transmission not only increases bandwidth burdens but also imposes significant pressure on computation and energy consumption, especially in resource-constrained terminal devices and edge nodes [1]. Additionally, the limitations of traditional communication architectures exacerbate control signaling and scheduling overheads in ultra-dense networks, leading to latency jitter and undermining the real-time performance of critical applications. Furthermore, the inefficient integration of multi-source information in heterogeneous networks and multimodal scenarios severely restricts end-to-end system performance [1]. Thus, there is an urgent need for innovative semantic communication architectures to enhance end-to-end transmission performance and optimize resource utilization.

Semantic communication intelligently extracts and reconstructs the intrinsic meaning of information, thereby significantly reducing redundant bit-level transmissions and retaining only the key content relevant to the task objectives. A substantial body of work has already been devoted to this emerging paradigm. As a promising and effective approach to address the aforementioned

challenges, semantic-level data compression and intelligent decision-making can markedly reduce the number of transmitted bits, thereby improving transmission efficiency, lowering bandwidth requirements, and reducing overall power consumption [2]. Moreover, by enabling context-aware and task-oriented transmission, this approach minimizes unnecessary interactions and ensures the quality of service for latency-sensitive applications [3]. In addition, the cross-modal information fusion capability of semantic communication enhances both the robustness and scalability of communication systems. The 6G network roadmap outlined in [4] explicitly identifies the deep integration of artificial intelligence and semantic technologies as a driving force for the future evolution of wireless networks. Consequently, the systematic development of semantic communication theories and architectures not only holds significant theoretical value but also promises to accelerate the comprehensive advancement of the 6G application ecosystem.

In contrast to existing surveys that concentrate on a single technological dimension—such as [5], which focuses on deep learning methods, and [6][7], which emphasize architectural taxonomies—this work adopts an innovative perspective by cross-mapping three core methodological paradigms with representative 6G architectures empowered by semantic communication. This cross-mapping provides a novel and intuitive navigational framework for understanding the applicability boundaries and optimization potentials of different approaches. To systematically review the opportunities and challenges that semantic communication brings to 6G, this survey aims to deliver a comprehensive, in-depth, and critically informed analytical framework. As illustrated in Fig. 1, we classify the core technologies, based on recent advances in semantic communication, into three categories: (i) model-based optimization, (ii) deep learning–driven schemes, and (iii) large-model–enabled paradigms. We critically examine their respective limitations, offering new insights into the intrinsic relationships and applicability boundaries across these approaches. We then extend the analysis to the application of semantic communication in 6G networks, summarizing four representative architectures, providing detailed examinations, and mapping them against the core technological categories. By critically analyzing the latest developments, open challenges, and inherent limitations—while outlining promising research directions—this survey guides the reader from the macro-level background to core technologies, then to system architectures, and finally converging on future challenges and prospects. The outcome is a logically coherent and progressively layered research framework.

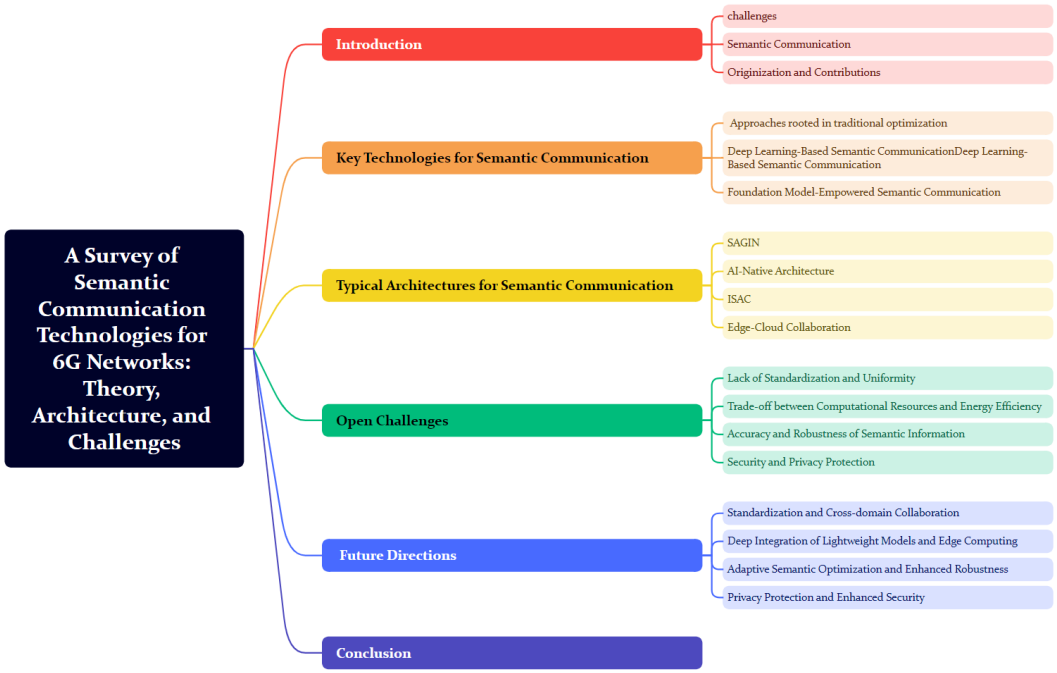

Figure 1. The Framework of the Essay

2. Key Technologies for Semantic Communication

For 6G networks, semantic communication technologies aim to significantly reduce the volume of transmitted data by focusing on the efficient conveyance of information meaning. The key technologies can be categorized into three major paradigms: traditional model-optimized semantic communication, deep learning-based semantic communication and large model-enabled semantic communication. This classification reflects a technological evolution path from foundational AI integration toward advanced model expansion and efficiency enhancement. (Table 1)

Table 1. Table of Comparison Among Key Technologies

Technological Path	Application Scenarios	Limitations	References
Traditional Model Optimization	Resource-constrained IoT and low-latency mission-critical communications (e.g., sensor networks, industrial automation)	Limited semantic representation capability	[3][5][7][9]
Deep Learning–Driven	Vertical industry–specific scenarios (e.g., autonomous driving, remote surgery, smart manufacturing)	Insufficient generalization capability	[10][11][12][13][14][15]
Large Model–Enabled	Large-scale open Internet services (e.g., content delivery networks, real-time interaction in the metaverse)	Extremely high computational resource requirements	[16][17][18]

2.1. Traditional Model–Optimized Semantic Communication

Traditional model–optimized semantic communication aims to integrate semantic understanding into the classical source-channel coding framework. The core of this approach lies in utilizing classical information theory, probabilistic models, and optimization theory to abstract semantic information into structured mathematical representations, such as knowledge graphs, semantic graphs, and task utility functions [8]. Through model-driven coding compression and resource allocation strategies, efficient transmission is achieved. This approach balances transmission efficiency with semantic fidelity, offering strong theoretical interpretability and low computational overhead. It does not rely on large-scale pre-trained models, making it well-suited for resource-constrained communication scenarios.

Several pivotal works have further explored traditional model–optimized semantic communication. Extending Shannon's theory, [5] introduces the definition of semantic information, quantifying it in terms of entropy, and deriving the semantic information entropy constraint. A layered communication system model, which includes semantic encoding/decoding modules and semantic channels, is constructed, clearly distinguishing between physical channel noise and semantic noise. Based on this, methods for channel capacity analysis are provided. In terms of end-to-end joint optimization and task-oriented design, [3] extends traditional separate-source-channel coding to an end-to-end joint optimization framework, embedding task success probability or semantic distortion metrics directly into the optimization objectives. This enables dynamic bandwidth and power allocation under different channel conditions, achieving adaptive scheduling of communication systems based on upper-layer task requirements. Furthermore, [9] deepens the end-to-end design concept by addressing the interweaving of semantic signals and channel noise. A confidence-based distillation mechanism is proposed for joint semantic encoding and channel encoding optimization (JSNC). This approach utilizes internal model confidence evaluation to guide information compression and error correction strategies, significantly enhancing the system's semantic recovery robustness under harsh channel conditions. In terms of compatibility design, [7] proposes a bit-conversion-based semantic transmission framework compatible with existing wireless systems. It introduces Integer Error Rate (IER) as a new physical layer performance metric to optimize the design of semantic mapping and

coding strategies. This approach not only facilitates the deployment of the semantic layer over traditional communication links but also provides a reference for the development of subsequent performance metrics.

Most studies rely on simulated channels or controlled environments, which have the advantage of enhancing the understanding of channel capacity by explicitly distinguishing between physical channel noise and semantic noise, thereby providing a theoretical foundation for system design. However, these studies often lack robust evaluation in complex, dynamic real-world wireless environments, and their generalization ability remains questionable. Furthermore, end-to-end joint optimization and joint distillation mechanisms show significant potential in improving semantic recovery performance, especially after embedding task success probability or semantic distortion metrics directly into the optimization objective. These methods can dynamically adjust bandwidth and power allocation, achieving adaptive scheduling based on upper-layer task requirements. However, they are accompanied by high computational complexity and latency overhead, making them challenging to meet real-time requirements, especially on resource-constrained terminals. In summary, to achieve the industrial deployment of semantic communication for 6G, further research is urgently needed in areas such as real-world scenario validation and lightweight algorithms,of a unified evaluation metric system.

2.2. Deep Learning–Driven Semantic Communication

Deep learning–driven semantic communication represents a paradigm shift from the traditional bit-precise transmission model. The core of this approach lies in leveraging deep learning models to extract, compress, and reconstruct the semantic features of information—such as intent, context, and task-relevant elements—rather than transmitting the complete bitstream of raw data. This process enables the efficient encoding and recovery of semantic information from raw data through an end-to-end neural network that automatically learns the "semantic feature-to-bitstream" mapping. Deep learning models do not rely on precise channel models or manually selected features. Instead, they use large-scale data-driven methods to automatically learn semantic representations and noise-resistant transmission strategies, aiming for semantically equivalent intelligent interactions [3][5]. The primary advantage of this approach is a significant enhancement in bandwidth efficiency and interference resistance, while supporting cross-modal interactions for human cognition or machine intelligence [10].

The introduction of deep learning methods provides a more efficient solution for semantic communication in the context of 6G. Notably, significant progress has been made in general end-to-end semantic modeling, particularly through the development of encoder-decoder architectures. The DeepSC framework, introduced in [10], is the first to apply Transformer models to text-based semantic communication, demonstrating that, under low signal-to-noise ratio (SNR) conditions, it can significantly reduce the bit rate while maintaining sentence-level semantic correctness compared to traditional layered communication. Studies [11][12] showcase how deep neural networks can replace iterative algorithms, enabling rapid inference for wireless link and power allocation problems. These advances provide optimized wireless link scheduling and resource allocation strategies for 6G, facilitating efficient semantic transmission and control in complex, multi-user, dynamic channel environments and enhancing the intelligence and adaptability of communication networks. In task-oriented and multi-user semantic communication, [13] extends single-user deep semantic communication to multi-user scenarios by proposing task-oriented resource allocation strategies. The system assigns different levels of importance to semantic bits based on the user's task (e.g., sentiment analysis or question-answering systems). Through deep network learning, the system enables channel sharing and power allocation, maximizing task performance across users. This approach supports efficient parallel communication among multiple users in 6G networks, improving task performance and surpassing traditional fairness or maximum throughput allocation strategies. Addressing the computational and energy constraints of Internet of Things (IoT) devices, [14] introduces a lightweight semantic encoder that retains only keyword-level features and significantly reduces

model size through parameter pruning and quantization techniques. A distributed aggregation mechanism is also designed, where edge servers perform joint training and updates, achieving low-latency semantic transmission across multiple channel conditions with a substantial reduction in model parameters. In multi-modal semantic communication scenarios, [15] extends the aforementioned text-based semantic communication to image-based scenarios. A convolutional autoencoder is employed to extract high-level semantic features from images, which are then jointly encoded with channel feature maps (nn-Block). This approach improves peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) compared to traditional JPEG + LDPC pipeline architectures, even under the same channel resources.

Despite the significant advances that deep learning–driven semantic communication has brought to end-to-end modeling, task-oriented resource allocation, and multimodal extensions for 6G, several critical bottlenecks and challenges remain to be addressed. First, although existing methods achieve high reconstruction accuracy and perform well on traditional signal-to-noise ratio (SNR) metrics, most studies still lack dedicated semantic evaluation criteria capable of directly measuring downstream task performance or user-perceived quality. In practical 6G scenarios, task-oriented semantic performance is more important than mere signal reconstruction; thus, accurately assessing and optimizing semantic quality remains a key research challenge. Second, the majority of current models rely on offline training and lack adaptability, particularly when facing unknown channel conditions or sudden interference. This limitation makes it difficult for these approaches to meet the real-time communication requirements of highly dynamic and multi-scenario 6G networks. Third, large-scale models still encounter multiple constraints—computational resources, energy consumption, and communication latency—when deployed at IoT network edges, especially on resource-limited nodes, making it difficult to ensure continuous online learning. Future research should focus on: (i) establishing a quantifiable and interpretable semantic evaluation framework to enable multi-dimensional performance assessment from the bit level to the semantic level; (ii) designing lightweight semantic models that support online incremental updates to accommodate heterogeneous networks and real-time communication demands; (iii) enhancing model robustness and privacy protection mechanisms against adversarial examples and malicious interference; and (iv) exploring joint optimization strategies for task-awareness and channel-awareness, leveraging causal inference and learnable collaborative modules to drive semantic communication in 6G toward greater efficiency, security, and intelligence.

2.3. Large-Model–Driven Semantic Communication

Large-model–driven semantic communication represents an innovative paradigm that introduces large pre-trained models into communication links. This paradigm shifts the focus from simple bit transmission to the extraction and sharing of high-level, task-relevant semantic information from vast datasets. The core principle of this approach is the establishment of a shared semantic knowledge base between the sender and receiver. By leveraging pre-trained or online learning methods, this technology gains deep representation capabilities for multimodal data, enabling more efficient semantic compression and reconstruction within limited bandwidth. The method combines the advantages of end-to-end optimization with dynamic knowledge update capabilities, which may overcome the performance limitations of traditional models in complex scenarios, enhancing the robustness and accuracy of communication systems in the face of semantic loss [16][17].

Applying large models to semantic communication can significantly improve the quality of end-to-end semantic representations and enhance noise resilience. Jiang et al. [16] utilized a pre-trained large language model to extract deep semantic features at the sender, and a lightweight decoder to reconstruct the information at the receiver. Experimental results demonstrate a significant reduction in bit error rates under low signal-to-noise ratio (SNR) conditions. Chaccour et al. [18] proposed a knowledge-graph-driven intelligent scheduling framework, combined with network slicing technology, to dynamically allocate bandwidth and computational resources, thereby reducing end-to-end latency. Xie et al. [17] designed a multi-stage decoder with a channel feedback loop to maintain

semantic recovery accuracy even in fluctuating channel conditions. Chaccour also deployed large model inference engines at the network edge to optimize task classification and resource allocation [18]. Taken together, large-model–driven semantic communication holds great potential for widespread application in large-scale commercial networks, promoting the intelligent and efficient development of 6G communication systems.

However, despite the significant performance advantages of large-model–driven semantic communication in low-SNR environments, several challenges and limitations remain in practical applications. First, the lack of interpretability in large pre-trained models remains one of their biggest challenges. As "black-box" structures, deep learning models, while exhibiting powerful performance in certain tasks, lack transparency in the process of semantic extraction and reconstruction. This makes performance diagnosis complex and may impact system security and compliance requirements. Second, although knowledge-graph-driven scheduling frameworks theoretically achieve a dynamic balance between computational power and bandwidth, the lack of unified standards and mature mechanisms for fine-grained management and fairness of network slicing in multi-user and complex network environments remains a challenge. Furthermore, the conflict between model size and computational resource demands, along with the generalization ability of these models, requires urgent attention. Future research should focus on improving model interpretability, reducing deployment costs, enhancing robustness, strengthening privacy protection, and developing comprehensive evaluation frameworks.

3. Typical 6G Architectures Enabled by Semantic Communication

With the evolution of 6G networks, future communication architectures will advance toward greater intelligence, flexibility, and efficiency. Conventional architectures can no longer meet increasingly complex and diversified application demands, particularly under the stringent requirements of massive connectivity, ultra-low latency, high reliability, and energy efficiency. In this context, semantic communication—through semantic extraction, encoding, and reasoning—has the potential to drive architectural innovation and enhance the overall performance of multiple representative 6G network architectures.

2.4. Space–Air–Ground–Sea Integrated Network

The space–air–ground–sea integrated architecture leverages the coordinated operation of satellites, near-Earth platforms, terrestrial base stations, and maritime stations to provide wide-area, all-weather network coverage. Its strengths lie in handling severe channel variations, latency and bandwidth constraints, and ensuring robustness and flexibility under heterogeneous node capabilities [4]. In such networks, semantic communication—through cross-layer collaboration and intelligent processing—can significantly enhance overall system performance and resource utilization. By employing large-scale pre-trained models for deep semantic extraction and encoding of multimodal data, satellites and high-altitude platforms can perform preliminary semantic analysis at the source, transmitting only high-value information elements to substantially reduce backhaul load while preserving critical information integrity. On the encoding side, complexity-aware model compression ensures that semantic encoders can operate efficiently even on resource-constrained platforms such as maritime buoys or small unmanned surface vessels. Multi-hop semantic routing, guided by nodes' semantic awareness, selects paths based on semantic similarity, prioritizing nodes that preserve the most complete semantic chain to ensure end-to-end task quality. For instance, in maritime rescue scenarios, airborne UAV relays can rapidly forward key semantic segments, eliminating redundant transmissions and avoiding service interruptions caused by single-path failures.

2.5. AI-Native Architecture

The AI-native architecture integrates deep learning and large-model capabilities across all layers of the network, seamlessly blending communication, computation, control, and reasoning. The

network is capable of autonomously sensing its environment, predicting demands, and dynamically optimizing resource allocation. In this framework, semantic communication not only provides intelligent support and resource optimization for 6G's AI-native architecture but also plays a crucial role in low-latency, high-efficiency data transmission. Through causal reasoning and contextual modeling, semantic communication enables semantic-aware routing and scheduling, which improves communication efficiency and optimizes the AI decision-making process. In an AI-native architecture, intelligent routing dynamically adjusts data transmission paths based on real-time contextual information, thereby enhancing overall system decision-making efficiency and adaptive responsiveness.

2.6. Integrated Sensing and Communication

The integrated sensing and communication (ISAC) architecture combines wireless sensing (such as radar, imaging, and positioning) with communication functions in a unified design, where the same waveform simultaneously carries both data and environmental sensing information. This enables the network not only to transmit information but also to sense and process environmental data in real-time, supporting smarter decision-making and optimization. In the ISAC framework, semantic communication directly encodes sensing results (such as target class or position semantics) into communication packets, thereby avoiding redundant transmission of raw echo data. Furthermore, by leveraging causal reasoning and contextual modeling, semantic communication enables intelligent routing and scheduling. For example, semantic communication can dynamically adjust the semantic priority of sensing and communication based on upper-layer task requirements (e.g., autonomous driving, XR), optimizing resource allocation [16]. Finally, the hierarchical design of semantic communication enhances communication optimization in large-scale collaborative scenarios. In a 6G-oriented ISAC architecture, the network must transmit diverse information across different layers. Semantic communication achieves this by hierarchically processing information and transmitting only the essential elements, effectively reducing network congestion and improving resource utilization and system responsiveness [10][17].

2.7. Edge-Cloud Collaboration

In the distributed edge-cloud collaboration architecture, resources across terminals, edge nodes, and the cloud are seamlessly orchestrated, with models collectively executing tasks. Computational resources are intelligently allocated based on task complexity and latency sensitivity, enabling significant computation offloading to edge nodes while meeting both response time and processing power requirements. Powered by semantic communication, this architecture exchanges only semantic representations (such as feature vectors or intent identifiers) between nodes, significantly reducing uplink/downlink bandwidth demands. Additionally, semantic communication integrates sensing data with semantic information, allowing the edge-cloud collaborative framework to continuously sense the environment and make intelligent decisions. For example, edge devices can process and analyze key information in real-time video streams via semantic communication, sending only high-level semantic data to the cloud for further decision-making and analysis. This approach reduces latency and enhances decision-making efficiency [2][10]. Finally, the integration of semantic communication enables tighter collaboration between cloud and edge computing. Through semantic understanding and reasoning of data, the cloud can more accurately schedule computational resources, while edge devices efficiently handle local tasks, forming a flexible collaborative computing network. The goal-oriented nature and contextual awareness of semantic communication ensure the system's ability to adapt to dynamic environmental changes and evolving demands [3][17].

4. Open Challenges:

Despite the promising results demonstrated by semantic communication in 6G networks, several key challenges must be addressed to facilitate its widespread adoption. Overcoming these barriers will pave the way for more robust, efficient, and intelligent communication systems.

1) **Lack of Standardization and Uniformity** Currently, semantic communication technologies lack a unified standard and evaluation framework. Research efforts are often focused on specific scenarios or technical pathways, with insufficient cross-domain collaboration and standardization. The efficient deployment of semantic communication in 6G networks requires standardized interfaces, protocols, and performance metrics to ensure compatibility and interoperability between different devices and network nodes. Current research tends to focus on specific tasks or environments, leading to poor transferability across various use cases.

2) **Trade-off Between Computational Resources and Energy Efficiency** As semantic communication technology evolves, especially in deep learning- and large model-based frameworks, the trade-off between computational resource demands and energy efficiency becomes increasingly prominent. While large-scale deep learning models demonstrate significant advantages in semantic extraction and information reconstruction, their reliance on computational resources and energy consumption remains a major concern. 6G networks require terminal devices and edge nodes with high computational power and energy efficiency to handle vast amounts of data and low-latency challenges. However, in resource-constrained environments such as IoT and smart terminals, balancing computational load, reducing energy consumption, and maintaining performance remain significant challenges.

3) **Accuracy and Robustness of Semantic Information** Semantic communication faces challenges in ensuring the accuracy and robustness of semantic information in practical applications. Due to the complexity and uncertainty of communication environments, semantic extraction and reconstruction can be easily influenced by noise, interference, and channel conditions. Particularly in adverse wireless environments, ensuring high-precision recovery of semantic information and preventing distortion remains an unresolved issue. Although existing deep learning models and large-model approaches can theoretically address this challenge, practical deployment often struggles to maintain stable performance due to limitations in device and computational resources.

4) **Security and Privacy Protection** In semantic communication frameworks, the transmission of semantic information not only involves the content itself but also its context, background, and other sensitive elements. Without proper security mechanisms during transmission, issues such as data leakage and privacy breaches may arise. Additionally, with the introduction of large models, the privacy and controllability of model training data present new challenges.

5. Future Directions

To address the challenges outlined above, semantic communication technology requires deep innovation across multiple domains.

1) **Standardization and Cross-Domain Collaboration** In the multi-layered, multi-domain 6G network, semantic communication must be closely integrated with AI-native architectures, edge computing, and distributed cloud architectures. This necessitates standards that cover not only communication protocols but also data formats, interface definitions, and interaction mechanisms. For instance, semantic communication technologies based on large models must use standardized interfaces to seamlessly connect with various network components, ensuring optimal performance across different network architectures. Furthermore, standardization in cross-domain collaboration will enable different vendors to innovate within a unified technical framework, share data and computational resources, and thereby enhance overall system efficiency and performance. Future research should focus on maintaining efficient transmission of semantic communication across different network environments, especially in multi-modal networks, heterogeneous terminals, and low-bandwidth, high-latency conditions.

2) **Lightweight Models and Deep Integration with Edge Computing** Future semantic communication technologies will make breakthroughs in lightweight models and edge computing architectures. Key technologies for reducing energy consumption and enhancing performance include model compression,

knowledge distillation, and low-power hardware design. By offloading computational tasks to the network edge, significant reductions in data transmission bandwidth and latency can be achieved, particularly in edge devices such as IoT and smart terminals. A core challenge for future research is how to process vast amounts of data on edge devices using streamlined semantic communication algorithms combined with local computational resources, while ensuring real-time processing and transmission efficiency.

3) **Adaptive Semantic Optimization and Robustness Enhancement** Future research will focus on maintaining high robustness in semantic communication in dynamic, complex environments. Through contextual awareness and intelligent inference mechanisms, systems can dynamically adjust semantic information transmission strategies to adapt to changing channel conditions, network loads, and task requirements. For instance, reinforcement learning-based adaptive scheduling can intelligently select optimal coding and scheduling strategies under unstable channel conditions, ensuring the accurate transmission of semantic information across the network. Furthermore, task-oriented semantic optimization methods will precisely control the flow of information based on task importance and latency requirements, ensuring system stability and accuracy.

4) **Privacy Protection and Enhanced Security** In 6G networks, where much task-related information may involve personal privacy, federated learning-based distributed training and encrypted transmission technologies will become key solutions for privacy protection. Through distributed processing, sensitive data can remain on terminal devices, avoiding privacy risks associated with centralized storage. Additionally, encrypted transmission within semantic communication frameworks can ensure that data is not intercepted or tampered with during transmission. In the future, semantic communication will also need to integrate explainable models to enhance security, particularly when dealing with task-specific data. The ability to restore semantic information in an explainable manner and implement privacy protection mechanisms will ensure data integrity, accuracy, and user privacy. Cross-platform security certification and privacy protection standards will also become key research areas to ensure unified privacy mechanisms across heterogeneous devices and platforms.

6. Conclusion

This survey presents a systematic analysis of the 6G vision and semantic communication technologies, summarizing their core concepts and development trajectories. As one of the keys enabling technologies for 6G, semantic communication carries profound theoretical significance and represents a transformative shift in the communications field. Despite numerous technical and implementation challenges, its developmental pathway is becoming increasingly clear, with substantial potential for advancement. At present, we stand at the early stage of this paradigm shift, where future breakthroughs will emerge not solely within communications or artificial intelligence, but through their deep integration—driving the creation and deployment of novel technologies.

References

[1] Zawish, M., Davy, S., & Abraham, L. (2022). Complexity-driven CNN compression for resource-constrained edge AI. arXiv. https://arxiv.org/abs/2208.12816

[2] Xu, W., Yang, Z., Ng, D. W. K., Levorato, M., Eldar, Y. C., & Debbah, M. (2023). Edge learning for B5G networks with distributed signal processing: Semantic communication, edge computing, and wireless sensing. IEEE Journal of Selected Topics in Signal Processing, 17(1), 1–37. https://doi.org/10.1109/JSTSP.2023.3239189

[3] Gunduz, D., Qin, Z., Aguerri, I. E., Dhillon, H. S., Yang, Z., Yener, A., Wong, K. K., & Chae, C.-B. (2023). Beyond transmitting bits: Context, semantics, and task-oriented communications. IEEE Journal on Selected Areas in Communications, 41(1), 5–41. https://doi.org/10.1109/JSAC.2022.3223408

[4] Letaief, K. B., Chen, W., Shi, Y., Zhang, J., & Zhang, Y.-J. A. (2019). The roadmap to 6G: AI empowered wireless networks. IEEE Communications Magazine, 57(8), 84–90. https://doi.org/10.1109/MCOM.2019.1900271

[5] Qin, Z., Tao, X., Lu, J., Tong, W., & Li, G. Y. (2023). Semantic communications: Principles and challenges. IEEE Journal on Selected Areas in Communications, 41(1), 1–32. https://doi.org/10.1109/JSAC.2022.3223408

[6] Shi, G., Xiao, Y., Li, Y., & Xie, X. (2021). From semantic communication to semantic-aware networking: Model, architecture, and open problems. IEEE Communications Magazine, 59(8), 41–49. https://doi.org/10.1109/MCOM.2021.9509448

[7] Luo, X., Chen, H.-H., & Guo, Q. (2022). Semantic communications: Overview, open issues, and future research directions. IEEE Wireless Communications, 29(1), 210–220. https://doi.org/10.1109/MWC.001.2100642

[8] Shao, Y., Cao, Q., & Gunduz, D. (2024). A theory of semantic communication. arXiv. https://arxiv.org/abs/2401.00635

[9] Lu, K., Zhou, Q., Li, R., Zhao, Z., Chen, X., Wu, J., & Zhang, H. (2022). Rethinking modern communication from semantic coding to semantic communication. IEEE Wireless Communications. https://doi.org/10.1109/MWC.013.2100642

[10] Xie, H., Qin, Z., Li, G. Y., & Juang, B.-H. (2021). Deep learning enabled semantic communication systems. IEEE Transactions on Signal Processing, 69, 2663–2675. https://doi.org/10.1109/TSP.2021.3071210

[11] Lee, H., Lee, S. H., & Quek, T. Q. S. (2019). Deep learning for distributed optimization: Applications to wireless resource management. IEEE Journal on Selected Areas in Communications, 37(10), 2257–2266. https://doi.org/10.1109/JSAC.2019.2933890

[12] Sun, H., Chen, X., Shi, Q., Hong, M., Fu, X., & Sidiropoulos, N. D. (2017). Learning to optimize: Training deep neural networks for wireless resource management. In 2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC) (pp. 1–6). IEEE. https://doi.org/10.1109/SPAWC.2017.8227772

[13] Xie, H., Qin, Z., Tao, X., & Letaief, K. B. (2022). Task-oriented multi-user semantic communications. IEEE Journal on Selected Areas in Communications, 40(9), 2584–2597. https://doi.org/10.1109/JSAC.2022.3184567

[14] Xie, H., & Qin, Z. (2021). A lite distributed semantic communication system for Internet of Things. IEEE Journal on Selected Areas in Communications, 39(1), 142–153. https://doi.org/10.1109/JSAC.2020.3036955

[15] Lokumarambage, M. U., Gowrisetty, V. S. S., Rezaei, H., Sivalingam, T., Rajatheva, N., & Fernando, A. (2023). Wireless end-to-end image transmission system using semantic communications. IEEE Access, 11, 37149–37163. https://doi.org/10.1109/ACCESS.2023.3266656

[16] Jiang, F., Peng, Y., Dong, L., et al. (2023). Large AI model-based semantic communications. arXiv. https://arxiv.org/abs/2307.03492v2

[17] Xie, H., Qin, Z., Tao, X., et al. (2024). Large model empowered semantic communications. arXiv. https://arxiv.org/abs/2402.13073v2

[18] Chaccour, C., Saad, W., Debbah, M., Han, Z., & Poor, H. V. (2025). Less data, more knowledge: Building next-generation semantic communication networks. IEEE Communications Surveys & Tutorials, 27(1), 38–74. https://doi.org/10.1109/COMST.2024.3412852