A Survey on Energy-Efficient in Semantic Communication: Techniques, Challenges, and Future Directions

Thwe Thwe Win
Department of Computer Science and Engineering
Chung-Ang University
Seoul, South Korea
ttwin@uclab.re.kr

Ayalneh Bitew Wondmagen
Department of Computer Science and Engineering
Chung-Ang University
Seoul, South Korea
ayalneh@uclab.re.kr

Thi My Tuyen Nguyen
Department of Computer Science and Engineering
Chung-Ang University
Seoul, South Korea
tuyen@uclab.re.kr

Junsuk Oh
Department of Computer Science and Engineering
Chung-Ang University
Seoul, South Korea
jsoh@uclab.re.kr

Gahyun Kim
Department of Computer Science and Engineering
Chung-Ang University
Seoul, South Korea
ghkim@uclab.re.kr

Sungrae Cho
Department of Computer Science and Enginnering
Chung-Ang University
Seoul, South Korea
srcho@cau.ac.kr

***Abstract*—As communication networks evolve to accommodate the increasing demands of Artificial intelligence (AI) applications, Internet of Things (IoT) devices, and edge computing, energy efficiency has become a critical concern. By focusing on transmitting only the most important information rather than raw data, semantic communication (SemCom) presents a promising solution for reducing bandwidth usage. However, this shift introduces new challenges in managing energy consumption, particularly in dynamic and resource-constrained environments. We survey recent advances in energy-efficient techniques for semantic communication, focusing on methods such as task offloading, resource optimization, and adaptive transmission strategies. In addition, integration of energy harvesting and the deployment of semantic communication in 6G networks are explored as a promising approach to sustaining long-term energy efficiency. Emerging trends in multimodal semantic communication highlight the challenges of optimizing energy when transmitting different types of data simultaneously. Finally, we discuss the open challenges and future research directions needed to ensure the practical scalability of energy-efficient semantic communication systems in next-generation networks.**

***Index Terms*—Energy Efficiency, Resource Allocation, Semantic Communications, Task Offloading**

I. Introduction

The rapid advancement of artificial intelligence (AI) and the Internet of Things (IoT) [1], [2] has created an enormous demand for efficient and scalable communication systems. With the explosion of smart devices, edge computing, and data-driven applications, conventional communication networks are under more strain than ever before. Traditional communication systems focus on transmitting raw data with the objective of maintaining high fidelity between sender and receiver. However, this paradigm has become increasingly inefficient as the volume of data grows exponentially and energy consumption becomes a primary concern for network sustainability.

Semantic communication is a revolutionary approach that shifts the focus from transmitting raw data to sending only meaningful information. By reducing the total volume of data transmitted, it helps ease bandwidth demands and reduces energy consumption. Instead of preserving every bit of data, the system extracts and transmits only the essential information needed to complete a task. This presents an opportunity to drastically improve the efficiency of communication networks, especially in the context of emerging AI and machine learning tasks that thrive on semantic-rich data.

However, while semantic communication offers a path forward for reducing data traffic and improving performance under limited bandwidth, it also introduces new challenges, with the primary challenge centered on optimizing energy consumption. As communication networks grow in complexity, with the integration of edge computing, aerial networks, and IoT devices, the need for energy-efficient solutions becomes critical. In resource-constrained environments, such as battery-operated IoT devices or drone-assisted edge networks, optimizing energy use is a core necessity to prolong device life, reduce operational costs, and improve overall system sustainability.

Energy efficiency in semantic communication is driven by two primary factors: transmission efficiency and computational efficiency. Transmission energy refers to the energy consumed to send semantic information from one node to another, while computational energy is consumed to process data, extract relevant semantics, and make off-loading decisions. As edge and aerial networks see increased usage, managing these two

aspects of energy consumption becomes more complex, requiring sophisticated algorithms and optimization techniques. For instance, multi-user semantic communication scenarios necessitate dynamic power control, resource allocation, and the integration of energy harvesting mechanisms to achieve long-term energy sustainability.

Recent advances in edge computing and aerial-aided networks have further propelled the need for energy-efficient communication systems. These technologies enable the offloading of computationally heavy tasks from user devices to more powerful edge servers, greatly reducing local energy consumption. However, these systems are not without limitations.The energy used to offload data and maintain reliable communication links, especially in environments with fluctuating channel conditions, needs to be carefully controlled. Moreover, unmanned aerial vehicles (UAVs) are used to provide temporary network coverage, making energy efficiency a crucial metric due to the limited power reserves available on these platforms in aerial-aided networks.

Semantic communication combined with energy-efficient techniques is expected to play a key role in enabling next-generation communication systems such as 6G networks. By employing deep learning-based semantic encoding and decoding mechanisms, these systems can further reduce transmission costs by compressing raw data into semantic features. This directly reduces the energy needed for transmission since fewer bits are sent over the air. Furthermore, novel energy-aware resource allocation techniques, such as those that leverage deep reinforcement learning (DRL) and multiagent systems, are proving effective in managing the energy consumption of semantic communication systems.

The rapid advancement of artificial intelligence (AI) and the Internet of Things (IoT) [1], [2] has created an enormous demand for efficient and scalable communication systems. With the explosion of smart devices, edge computing, and data-driven applications, conventional communication networks are under more strain than ever before. Traditional communication systems focus on transmitting raw data with the objective of maintaining high fidelity between sender and receiver. However, this paradigm has become increasingly inefficient as the volume of data grows exponentially and energy consumption becomes a primary concern for network sustainability.

Semantic communication is a revolutionary approach that shifts the focus from transmitting raw data to sending only meaningful information. By reducing the total volume of data transmitted, it helps ease bandwidth demands and reduces energy consumption. Instead of preserving every bit of data, the system extracts and transmits only the essential information needed to complete a task. This presents an opportunity to drastically improve the efficiency of communication networks, especially in the context of emerging AI and machine learning tasks that thrive on semantic-rich data.

However, while semantic communication offers a path forward for reducing data traffic and improving performance under limited bandwidth, it also introduces new challenges, with the primary challenge centered on optimizing energy consumption. As communication networks grow in complexity, with the integration of edge computing, aerial networks, and IoT devices, the need for energy-efficient solutions becomes critical. In resource-constrained environments, such as battery-operated IoT devices or drone-assisted edge networks, optimizing energy use is a core necessity to prolong device life, reduce operational costs, and improve overall system sustainability.

Energy efficiency in semantic communication is driven by two primary factors: transmission efficiency and computational efficiency. Transmission energy refers to the energy consumed to send semantic information from one node to another, while computational energy is consumed to process data, extract relevant semantics, and make off-loading decisions. As edge and aerial networks see increased usage, managing these two aspects of energy consumption becomes more complex, requiring sophisticated algorithms and optimization techniques. For instance, multi-user semantic communication scenarios necessitate dynamic power control, resource allocation, and the integration of energy harvesting mechanisms to achieve long-term energy sustainability.

Recent advances in edge computing and aerial-aided networks have further propelled the need for energy-efficient communication systems. These technologies enable the offloading of computationally heavy tasks from user devices to more powerful edge servers, greatly reducing local energy consumption. However, these systems are not without limitations. The energy used to offload data and maintain reliable communication links, especially in environments with fluctuating channel conditions, needs to be carefully controlled. Moreover, unmanned aerial vehicles (UAVs) are used to provide temporary network coverage, making energy efficiency a crucial metric due to the limited power reserves available on these platforms in aerial-aided networks.

Semantic communication combined with energy-efficient techniques is expected to play a key role in enabling next-generation communication systems such as 6G networks. By employing deep learning-based semantic encoding and decoding mechanisms, these systems can further reduce transmission costs by compressing raw data into semantic features. This directly reduces the energy needed for transmission since fewer bits are sent over the air. Furthermore, novel energy-aware resource allocation techniques, such as those that leverage deep reinforcement learning (DRL) and multiagent systems, are proving effective in managing the energy consumption of semantic communication systems.

This survey aims to provide a comprehensive overview of the state-of-the-art in energy-efficient semantic communication. We explore key techniques used to optimize energy consumption in both transmission and computation, including task offloading, aerial-assisted networks, and resource optimization. In addition, we address the emerging trends in energy harvesting and deep learning-driven resource allocation, highlighting how these innovations are paving the way for more sustainable and energy-conscious communication networks.

The main contributions of this survey are as follows:

- **Comprehensive Review:** We provide a detailed review of energy-efficient techniques in semantic communication, focusing on task offloading, resource optimization, and adaptive transmission strategies.
- **Integration with Emerging Technologies:** We discuss the integration of energy-efficient semantic communication with 6G networks, energy harvesting mechanisms, and deep learning-based resource allocation.
- **Focus on Multi-Modal Communication:** We highlight the challenges and opportunities of energy efficiency in multi-modal semantic communication, which involves transmitting multiple data types such as text, video, and sensor data.
- **Identification of Open Challenges:** We identify key open challenges, such as achieving scalability in multi-user scenarios and ensuring synchronization in real-time applications.
- **Future Directions:** We propose promising research directions to enhance energy efficiency and scalability in next-generation semantic communication systems.

The remainder of this paper is organized as follows. In Section II, we discuss various energy-efficient approaches in semantic communication systems, including task offloading, resource allocation, and integration with emerging technologies such as 6G networks. Section III highlights the challenges and future directions in achieving energy efficiency in semantic communication, with a particular focus on multi-modal communication and real-time applications. Finally, Section IV concludes the paper by summarizing the key findings and outlining potential areas for further research.

II. Related Works

Over the past few years, extensive research has been conducted on energy efficiency in semantic communication systems. Several studies have focused on task offloading, resource optimization, and adaptive transmission strategies to enhance energy efficiency in resource-constrained environments. In [3], Ji and Qin proposed a proximal policy optimization-based multi-agent reinforcement learning algorithm for optimizing both computation and communication resources in semantic-aware networks. This approach demonstrated significant energy savings but lacked consideration of multi-modal semantic communication scenarios.

Zhang et al. in [4] investigated energy-efficient techniques for aerial-aided edge networks, emphasizing the role of semantic compression and resource allocation. While their work addressed the integration of energy harvesting, it did not explore the challenges associated with multi-user scenarios and task synchronization. Yang et al. [5] explored the role of semantic communication in 6G networks, focusing on AI-driven resource management for reducing transmission energy. However, the study primarily emphasized single-modal data, leaving gaps in understanding the complexities of multi-modal communication.

Several recent works have also introduced the concept of energy-aware resource allocation in semantic communication. For instance, [6] proposed the DACODE framework, which optimizes energy efficiency in heterogeneous wireless sensor networks. While effective in industrial IoT applications, its applicability to real-time multi-modal scenarios remains limited. Despite these advances, there remains a lack of comprehensive research on multi-modal semantic communication, particularly in scenarios involving heterogeneous data types, real-time synchronization, and scalability in multi-user networks. This survey addresses these gaps by providing an extensive review of state-of-the-art energy-efficient techniques, identifying open challenges, and proposing future research directions.

III. Energy-Efficient Approaches in Semantic Communication Systems

A. Energy Efficiency in Semantic Communications

Energy efficiency plays a crucial role in the successful deployment of semantic communication systems, especially in resource-constrained environments such as IoT networks, edge computing, and aerial-assisted networks [8], [9]. Although semantic communication has the promise of significantly reducing data traffic by transmitting only meaningful information, the computational burden associated with the extraction and processing of semantics introduces new challenges in energy consumption [3]. For example, the encoding and decoding of semantic data, often powered by deep learning models, can be computationally expensive, leading to higher local energy consumption, especially in devices with limited computational power such as IoT sensors and edge devices. Moreover, semantic communication systems must balance the trade-off between data compression and the fidelity of the transmitted information. As more semantic features are compressed to save transmission energy, the risk of losing critical information increases, which can result in costly retransmissions or the need for additional processing, further exacerbating energy consumption.

This balance is particularly critical in dynamic and mobile environments, such as UAV-assisted networks, where maintaining energy-efficient communication links is essential to extend operational time and reduce the need for frequent battery recharging. Furthermore, in multi-user scenarios, energy efficient techniques are needed to optimize resource allocation while ensuring that semantic communication remains robust and scalable [3], [10]. Hence, energy efficiency is not just an optimization goal, but a fundamental requirement for ensuring the practicality and sustainability of that systems, particularly as these systems scale to support increasingly complex AI-driven tasks in real-time environments.

B. Energy-Efficient Task Offloading

One of the most effective energy-efficient techniques for semantic communication is task offloading, particularly in edge computing environments. In a semantic-aware task offloading system, local devices with limited computational power offload their resource-intensive tasks to edge servers, which are capable of performing these tasks more efficiently [8], [3]. This offloading reduces the computational energy required on

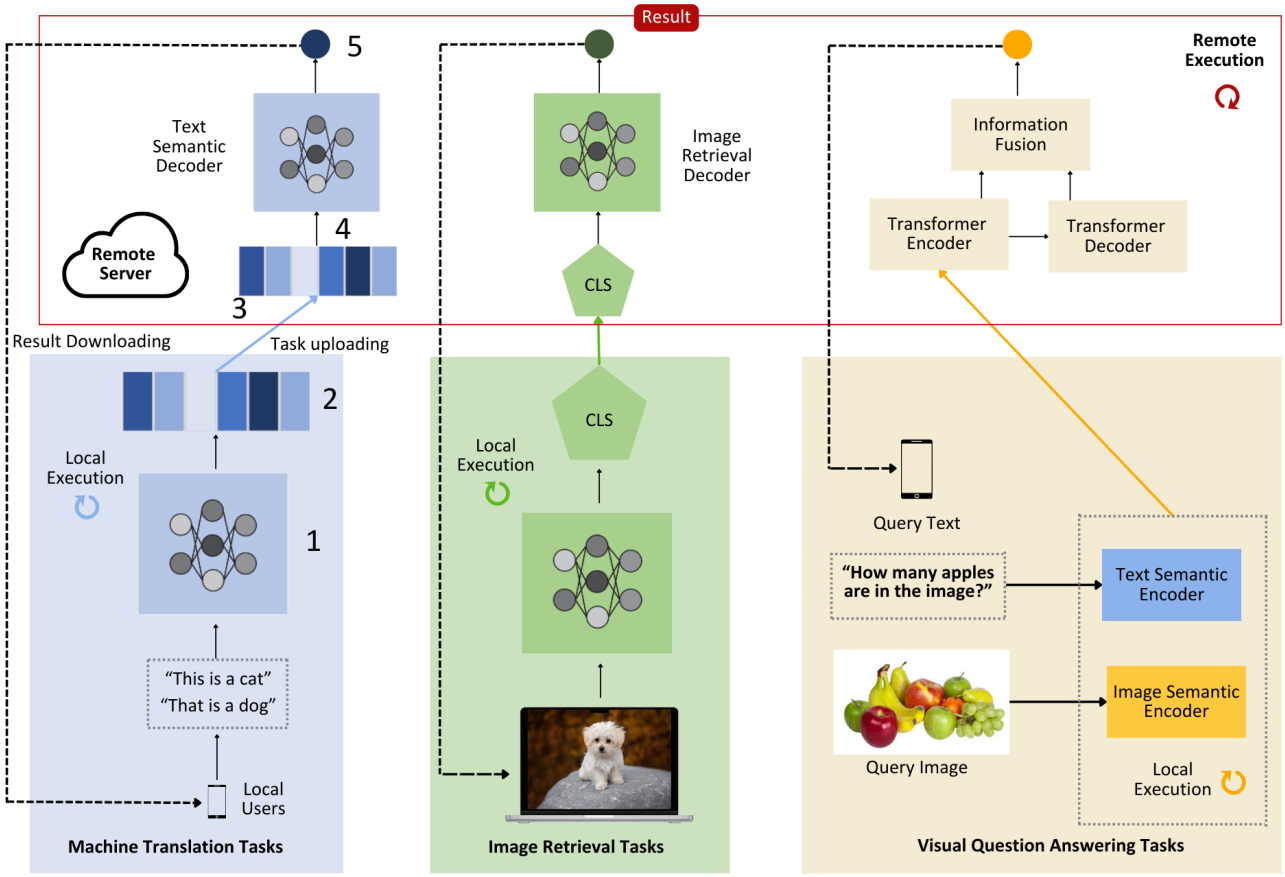

Fig. 1. System model for semantic task offloading process [3], [7]

the user's device while leveraging the edge server's superior processing capacity [11]. However, a critical challenge in this context is the efficient management of transmission energy. Since offloading involves transferring semantic data between the user device and the edge server, optimizing transmission power and bandwidth is crucial to minimizing energy consumption.

For instance, the paper by Ji and Qin proposes a proximal policy optimization-based multi-agent reinforcement learning (MAPPO) algorithm to optimize both computation and communication resources [3]. This method ensures distributed resource management and reduces the complexity of online algorithms, resulting in significant energy savings compared to traditional methods. Furthermore, edge servers can process semantic information using techniques like deep learning-based encoding, which ensures that only the most essential features of the task are transmitted, thereby reducing the energy required for both transmission and computation. Similarly, multi-agent reinforcement learning techniques, such as those proposed in [12], have been employed in cognitive radio networks to optimize energy sensing thresholds and improve energy efficiency in distributed environments.

Additionally, offloading techniques can be further optimized by integrating energy harvesting mechanisms. By allowing devices to harvest ambient energy (such as solar or RF energy), the system can sustain task offloading for extended periods without exhausting battery reserves. Figure 1 demonstrates the system model for semantic task offloading, showing how edge servers and local devices interact to optimize energy consumption and computational efficiency during task offloading processes. A binary offloading scheme, as proposed by Yang et al., effectively reduces the execution latency and energy consumption of IoT devices while supporting task offloading to multiple edge servers [8]. This combination of task offloading, semantic communication, and energy harvesting not only ensures real-time performance in resource-constrained environments but also significantly enhances the energy efficiency of modern communication networks [7], [13].

C. Energy-Efficient Resource Allocation

In semantic communication systems, efficient resource allocation is essential to ensure that energy consumption is minimized while maintaining high-quality service. As these

systems often operate under constrained resources such as bandwidth, power, and computation, proper resource allocation becomes a key determinant of overall energy efficiency. Recent research has shown that traditional resource allocation methods, which primarily focus on maximizing system throughput or capacity, are insufficient in the context of semantic communication. Instead, these systems require intelligent task-oriented resource allocation that focuses on the significance of the transmitted information rather than merely the quantity of data [4], [14]. A DDPG-based approach, as proposed in [15], optimizes resource allocation by incorporating energy-efficient techniques like SWIPT and MC-NOMA, enabling simultaneous information and power transfer in federated learning systems. Similarly, the DACODE framework [6] provides a distributed adaptive communication model for optimizing energy efficiency in IIoT-based heterogeneous wireless sensor networks, demonstrating the effectiveness of intelligent resource management in industrial applications. The work by Zhang et al. introduces a two-tier deep reinforcement learning (DRL) framework to optimize resource allocation for task-oriented semantic communication, particularly in scenarios where energy harvesting is integrated with cognitive radio (CR) and non-orthogonal multiple access (NOMA) networks [4]. This method balances transmission power, time-slot division, and semantic compression ratios, achieving significant energy savings while maintaining user quality of experience (QoE).

D. Integration with 6G Networks

Integrating semantic communication into 6G networks is a key development in modern communication systems, as 6G strives to deliver ultra-low latency, high data rates, and massive device connectivity, all with a strong focus on energy efficiency [5]. Unlike previous generations, 6G will require AI-driven resource management systems capable of optimizing power, bandwidth, and computational resources dynamically in real-time. semantic communication fits well into this vision by reducing data transmission requirements and energy consumption [16]. However, integrating it into the complex and highly dynamic 6G environment will require the development of joint optimization frameworks that maintain energy efficiency even in large-scale networks. This trend highlights the future role of semantic communication in achieving sustainable, high-performance networks as 6G becomes a reality [17].

E. Energy Efficiency in Multi-Modal Semantic Communication

The transition to multi-modal semantic communication, where multiple types of data (such as video, text, and sensor information) are transmitted simultaneously, presents both opportunities and challenges for energy efficiency [4], [5]. Multi-modal systems require adaptive compression algorithms and energy-aware resource allocation to ensure energy savings across different data types, as video data, for instance, demands more energy than text [18]. Ensuring synchronization and minimizing energy wastage will be critical, particularly for applications like autonomous vehicles and smart cities, which depend on real-time, multi-modal communication. As the importance of multi-modal data grows, energy-efficient algorithms will be key to reducing overall system consumption while maintaining high-quality service [19].

IV. Conclusion

In this paper, we have provided a concise survey of energy-efficient techniques for semantic communication (SemCom), focusing on critical areas such as task offloading, resource allocation, and adaptive transmission strategies in IoT networks, edge computing, and UAV-assisted systems. Key strategies such as deep learning-driven encoding, efficient resource management, and energy harvesting were identified as vital for minimizing energy use in dynamic settings. As semantic communication integrates into future 6G networks, optimizing energy efficiency and real-time performance will be vital, particularly for supporting multi-modal data and AI-driven tasks.

Future work in this area should focus on addressing the trade-offs between latency, energy efficiency, and data fidelity, particularly in multi-user scenarios. Additional research is needed to develop advanced techniques for synchronization in real-time multi-modal communication systems and explore the integration of energy harvesting technologies in large-scale semantic communication networks. Furthermore, scalability in heterogeneous environments, such as smart cities and autonomous vehicles, remains an open challenge that demands innovative solutions. By tackling these challenges, future research can enable more sustainable, efficient, and robust semantic communication systems for next-generation networks.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2024-00453301).

References

[1] A. Ghasempour, "Internet of things in smart grid: Architecture, applications, services, key technologies, and challenges," *Inventions*, vol. 4, no. 1, p. 22, 2019.

[2] IBM, "What is the iot?" https://www.ibm.com/topics/internet-of-things, accessed: 2025-01-07.

[3] Z. Ji and Z. Qin, "Energy-efficient task offloading for semantic-aware networks," in *ICC 2023-IEEE International Conference on Communications*. IEEE, 2023, pp. 3584–3589.

[4] G. Zheng, Q. Ni, K. Navaie, H. Pervaiz, A. Kaushik, and C. Zarakovitis, "Energy-efficient semantic communication for aerial-aided edge networks," *IEEE Transactions on Green Communications and Networking*, 2024.

[5] Z. Yang, M. Chen, Z. Zhang, and C. Huang, "Energy efficient semantic communication over wireless networks with rate splitting," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 5, pp. 1484–1495, 2023.

[6] J. Oh, D. Lee, D. S. Lakew, and S. Cho, "Dacode: Distributed adaptive communication framework for energy efficient industrial iot-based heterogeneous wsn," *ICT Express*, vol. 9, no. 6, pp. 1085–1094, 2023.

[7] Z. Ji, Z. Qin, X. Tao, and Z. Han, "Resource optimization for semantic-aware networks with task offloading," *IEEE Transactions on Wireless Communications*, 2024.

[8] H. Zhang, H. Wang, Y. Li, K. Long, and V. C. Leung, "Toward intelligent resource allocation on task-oriented semantic communication," *IEEE Wireless Communications*, vol. 30, no. 3, pp. 70–77, 2023.

[9] Z. Qin, X. Tao, J. Lu, W. Tong, and G. Y. Li, "Semantic communications: Principles and challenges," *arXiv preprint arXiv:2201.01389*, 2021.

[10] H. Xie, Z. Qin, X. Tao, and K. B. Letaief, "Task-oriented multi-user semantic communications," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 9, pp. 2584–2597, 2022.

[11] T. Zhou, D. Qin, X. Nie, X. Li, and C. Li, "Energy-efficient computation offloading and resource management in ultradense heterogeneous networks," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 12, pp. 13 101–13 114, 2021.

[12] T. T. H. Pham, W. Noh, and S. Cho, "Multi-agent reinforcement learning based optimal energy sensing threshold control in distributed cognitive radio networks with directional antenna," *ICT Express*, 2024.

[13] M. Min, L. Xiao, Y. Chen, P. Cheng, D. Wu, and W. Zhuang, "Learning-based computation offloading for iot devices with energy harvesting," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, pp. 1930–1941, 2019.

[14] G. Zhang, Q. Hu, Z. Qin, Y. Cai, G. Yu, and X. Tao, "A unified multi-task semantic communication system for multimodal data," *IEEE Transactions on Communications*, 2024.

[15] M. C. Ho, A. T. Tran, D. Lee, J. Paek, W. Noh, and S. Cho, "A ddpg-based energy efficient federated learning algorithm with swipt and mc-noma," *ICT Express*, vol. 10, no. 3, pp. 600–607, 2024.

[16] W. Yang, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. Cao, and K. B. Letaief, "Semantic communication meets edge intelligence," *IEEE Wireless Communications*, vol. 29, no. 5, pp. 28–35, 2022.

[17] K. B. Letaief, Y. Shi, J. Lu, and J. Lu, "Edge artificial intelligence for 6g: Vision, enabling technologies, and applications," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 1, pp. 5–36, 2021.

[18] Z. Weng, Z. Qin, X. Tao, C. Pan, G. Liu, and G. Y. Li, "Deep learning enabled semantic communications with speech recognition and synthesis," *IEEE Transactions on Wireless Communications*, vol. 22, no. 9, pp. 6227–6240, 2023.

[19] Y. Wang, M. Chen, W. Saad, T. Luo, S. Cui, and H. V. Poor, "Performance optimization for semantic communications: An attention-based learning approach," in *2021 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2021, pp. 1–6.