A Survey on Lightweight Semantic Communication for Resource-Constrained Internet Things: Technologies, Applications, and Challenges

Siyin Dai*

Nanjing University of Posts and Telecommunications, Nanjing, 210023, Jiangsu, China

* Corresponding author email: a3304809447@qq.com

Abstract

Terminal devices in the Resource-Constrained Internet of Things (RC-IoT) generally face multiple limitations in computing power, memory, bandwidth, and energy consumption, making it challenging to directly deploy semantic communication systems based on large-scale deep learning models. This paper presents a systematic survey on lightweight semantic communication research tailored for RC-IoT, establishing a unified classification and evaluation framework that encompasses three core technological categories: (1) Model compression (e.g., pruning, quantization, distillation); (2) Codec and lightweight architecture optimization (e.g., edge-side encoding, compressed sensing, distributed inference); (3) Dynamic resource allocation (e.g., event-triggered mechanisms, Value-of-Information (VoI) prioritization, deep reinforcement learning-based scheduling). A comparative analysis of representative approaches is conducted using metrics (BLEU, FID, mIoU, mAP, PSNR, bpp, latency, and energy consumption). Furthermore, adaptation strategies and engineering challenges are explored in three representative application scenarios: industrial IoT, healthcare, and smart homes. Finally, the study highlights current research gaps in robustness, privacy-performance trade-offs, cross-device semantic consistency, and standardization/reproducibility, while proposing future directions to facilitate the practical deployment of semantic communication in RC-IoT. This work aims to provide actionable insights for advancing semantic communication systems in resource-limited IoT environments.

Keywords

Resource-constrained IoT; Semantic communication; Lightweight technologies.

1. Introduction

The deep integration of the Internet of Things (IoT) into domains such as industrial manufacturing, smart healthcare, and intelligent homes is driving the deployment of massive numbers of end devices. However, in many practical scenarios (e.g., edge sensing nodes, wearable devices), these devices often face severe constraints in computational power, memory capacity, energy supply, and communication bandwidth [1-8], forming what is known as the Resource-Constrained IoT (RC-IoT). When performing real-time intelligent tasks (e.g., object recognition, status monitoring, collaborative control), such devices must also efficiently transmit information highly relevant to the task, posing unprecedented demands for low-latency, high-energy-efficiency, and robust communication. Unfortunately, traditional communication paradigms still aim at bit-level accurate reconstruction [9], failing to distinguish data importance or task relevance. This results in high transmission redundancy and excessive consumption of bandwidth and energy, making it difficult to meet the extreme efficiency requirements of RC-IoT.

Semantic communication, as an emerging paradigm, offers a promising solution to these challenges by transmitting the "meaning" (semantics) of data directly rather than raw bitstreams [10-11]. Its core lies in understanding and transmitting the semantic content relevant to the target task, theoretically enabling significant bandwidth reduction while enhancing robustness against channel interference. However, existing semantic communication schemes—especially those based on deep learning models-are typically computation-intensive and parameter-heavy [6]. Their substantial computational and storage overhead far exceeds the capabilities of typical RC-IoT devices, hindering large-scale deployment in real-world scenarios.

Therefore, the development of lightweight semantic communication technologies has become a research hotspot in recent years. The primary goal is to significantly reduce computational, storage, communication, and energy costs during model inference, feature encoding/decoding, and data transmission—while maintaining or approaching the original task performance (e.g., recognition accuracy, semantic understanding)—through techniques such as model compression, architectural optimization, efficient codec design, and intelligent resource scheduling [6][7]. This ultimately enables efficient and practical deployment of semantic communication in RC-IoT environments.

This paper aims to systematically review recent advances in lightweight semantic communication for RC-IoT, establish a unified technical taxonomy and evaluation framework, and analyze adaptation strategies and challenges in typical application scenarios. The main contributions are as follows: (1) Proposing a unified classification and evaluation framework for lightweight semantic communication in RC-IoT, encompassing three key technical pathways: model compression, codec optimization, and dynamic resource allocation; (2) Conducting a systematic comparison and analysis of representative approaches based on unified evaluation metrics (BLEU, FID, mIoU, mAP, PSNR, bpp, latency, and energy consumption); (3) Summarizing adaptation strategies and performance characteristics in three representative application scenarios: industrial IoT, healthcare, and smart homes; (4) Identifying current research challenges in robustness, privacy-performance trade-offs, cross-device semantic consistency, and standardization /reproducibility, and proposing future research directions.

2. Key Technologies for Lightweight Semantic Communication

Deploying semantic communication in Resource-Constrained IoT (RC-IoT) environments faces stringent constraints on terminal computing power, memory, power consumption, and communication bandwidth. To overcome these limitations, model compression, codec optimization, and dynamic resource allocation constitute the three core technical pillars of lightweight semantic communication. These pillars work synergistically, significantly reducing the overhead of semantic processing and transmission while ensuring task performance and system robustness. They achieve this by addressing three critical dimensions: model simplification at the terminal, enhanced efficiency in feature representation and reconstruction, and intelligent system-level resource scheduling, respectively. The key technologies supporting these three pillars are summarized in Fig. 1 below.

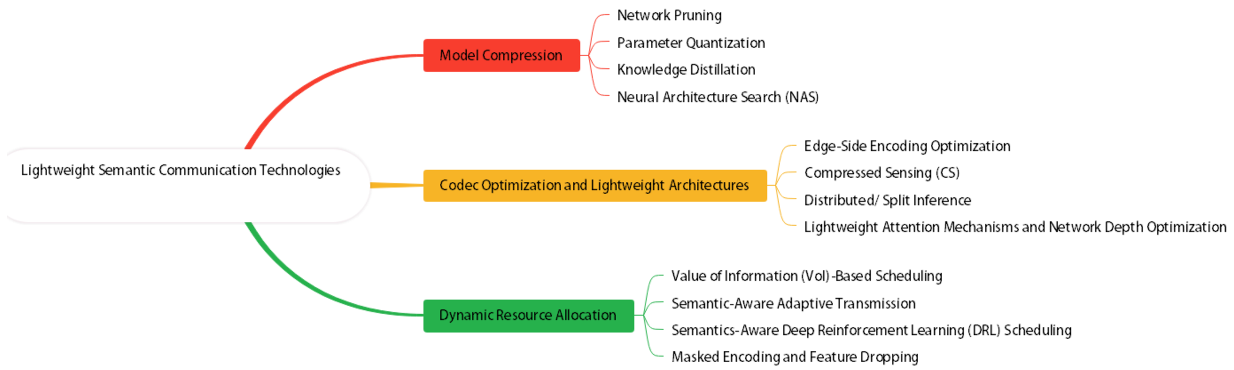

Fig. 1 Three Core Technical Pillars of Lightweight Semantic Communication

2.1. Model Compression

Model compression aims to reduce the number of model parameters, memory footprint, and computational cost of inference without significantly degrading task performance, thereby adapting models to the computational and storage limitations of end devices. Common techniques include:

Network Pruning: Removing channels or weights with minor contributions. For example, unstructured pruning can significantly reduce computational load and accelerate inference when supported by hardware [3].

Parameter Quantization: Mapping weights and activations to low-bit representations (e.g., 8-bit quantization) to reduce storage requirements and improve computational efficiency [6].

Knowledge Distillation: Training a compact "student" model to mimic the behavior of a larger "teacher" model, enabling the student to maintain high performance while drastically reducing model complexity [12].

Neural Architecture Search (NAS): Automatically searching for network architectures that meet latency and accuracy constraints, thereby optimizing performance under resource-limited conditions [13].

Representative examples include L-DeepSC [3], which applies structured pruning and quantization for text transmission, and the Q-GESCO framework [6], which employs quantized encoding on the edge device for image transmission.

Model compression is particularly suitable for devices with extremely limited resources, as it greatly reduces memory and computational burden. However, excessive compression may lead to performance degradation, and there exists a trade-off between compression ratio and accuracy loss.

2.2. Codec Optimization and Lightweight Architectures

These methods optimize the end-to-end semantic communication pipeline from the perspectives of system architecture and signal processing, aiming to reduce the transmission and computational overhead of redundant features. Common approaches include:

Edge-Side Encoding Optimization: Extracting and encoding only task-relevant features to minimize the transmission of irrelevant information [1].

Compressed Sensing (CS): Leveraging the sparsity of signals to reduce the amount of sampled data directly at the acquisition stage [2].

Distributed/ Split Inference: Partitioning the model and deploying different segments across end devices, edge nodes, and cloud servers to alleviate the computational burden on the terminal device [7].

Lightweight Attention Mechanisms and Network Depth Optimization: Employing channel and spatial attention modules to selectively emphasize key features, while reducing the number of downsampling layers to significantly lower computational cost [14].

Representative examples include LICRnet [1], which adopts end-to-end joint source-channel coding for image compression, and CS-ASIC [2], which integrates compressed sensing with lightweight encoding for image semantic segmentation.

These methods typically achieve a good balance between reducing communication and computational overhead and maintaining high task performance. However, they often require task-specific and channel-aware codec design, which limits their generalization across diverse scenarios.

2.3. Dynamic Resource Allocation

Dynamic resource allocation aims to improve overall system efficiency by adaptively adjusting communication and computational resources in real time based on task requirements and channel conditions. Common approaches include:

Value of Information (VoI)-Based Scheduling: Prioritizing the transmission of data that has the greatest impact on task completion [15].

Semantic-Aware Adaptive Transmission: Dynamically selecting whether to transmit shared or private information based on semantic similarity analysis, thereby avoiding redundant data transmission [16].

Semantics-Aware Deep Reinforcement Learning (DRL) Scheduling: Dynamically optimizing the allocation of bandwidth and computational resources by jointly considering channel state, semantic value of information (e.g., timeliness, accuracy, task criticality), and end-to-end objectives (e.g., real-time reconstruction error, execution cost) [17].

Masked Encoding and Feature Dropping: Selectively discarding certain features under low-resource conditions while preserving core semantic content [15].

Representative examples include the DRL-based adaptive resource allocation scheme SAC + D-SAC (Soft Actor-Critic + Discrete SAC) for multi-task scenarios [15], and the E2E Semantics policy, which is broadly applicable to event-driven scenarios in smart homes [17].

Dynamic resource allocation can significantly enhance system adaptability and energy efficiency. However, it incurs additional overhead for state sensing and decision-making and may suffer from unstable resource allocation under extreme channel fluctuations.

3. Technical Comparison and Typical Applications

As lightweight semantic communication technologies are increasingly applied across diverse domains, different application scenarios exhibit distinct performance priorities. To enable precise technology selection and efficient deployment, this section proposes a unified evaluation framework and integrates it with three representative application scenarios—Industrial Internet of Things (IIoT), Smart Healthcare, and Smart Home—to provide a metric-weighted analysis and comparative assessment for technology adaptation.

3.1. Evaluation Framework

In current research on lightweight semantic communication, different studies often vary significantly in terms of datasets, channel conditions, model scales, and training strategies, making direct performance comparisons challenging. To enable fair and systematic evaluation of various lightweight semantic communication schemes, this section establishes a unified evaluation framework. This framework defines core metrics across three dimensions (task/communication/computational performance) with standardized datasets.

(1) Task Performance Metrics

Task performance measures the effectiveness of a semantic communication system in completing specific tasks (e.g., text transmission, image reconstruction, object detection, semantic segmentation). It directly reflects task completion quality and usability and is often the primary concern in practical applications. Commonly used metrics include [1-3,6,13-16]:

BLEU (Bilingual Evaluation Understudy): Measures accuracy in text reconstruction, widely used in semantic communication for machine translation-like tasks.

FID (Fréchet Inception Distance): Evaluates the distributional difference between generated and real images, reflecting perceptual quality.

mIoU (Mean Intersection over Union): Measures accuracy in semantic segmentation.

mAP (Mean Average Precision): Used for object detection tasks.

PSNR (Peak Signal-to-Noise Ratio): Assesses reconstruction quality for images or videos.

F1-score (Harmonic Mean of Precision and Recall): Balances false negatives and false positives, particularly important in medical or safety-critical applications.

Semantic Similarity: Quantifies the degree of meaning preservation in textual content.

(2) Communication Performance Metrics

Communication performance evaluates the transmission efficiency of the semantic communication link under bandwidth constraints and dynamic channel conditions. It determines the system's performance in terms of latency and efficiency, where low delay and low bandwidth usage are particularly critical for industrial control and interactive applications. Key metrics include [1-3,6,13-16]:

bpp (bits per pixel): Number of bits transmitted per pixel, indicating compression efficiency.

End-to-End Latency: Total time from data acquisition at the source to final result delivery at the destination.

Bandwidth Consumption: Actual amount of network bandwidth used during transmission.

Bandwidth Requirement: The maximum data volume the network must support within a given time window.

Bandwidth Efficiency: Measures how effectively the available bandwidth is utilized for transmitting task-relevant semantic information.

(3) Computational Performance Metrics

Computational performance quantifies the computational and memory overhead of models on end devices, edge nodes, or cloud servers. This is especially critical in resource-limited scenarios such as embedded systems and wearable devices. Main metrics include [1-3,6,13-16]:

FLOPs (Floating Point Operations): A direct measure of computational complexity.

Memory Usage: Storage space required during model inference (e.g., model size, activation memory).

Energy Consumption: Total energy required to execute a single inference or communication task.

(4) Datasets and Channel Settings

To ensure fair comparison, experiments are typically conducted using publicly available datasets and standardized channel models, such as:

Commonly used datasets: Cityscapes (urban scene segmentation), MIT-BIH (arrhythmia detection), COCO (general object detection), ImageNet (image classification), and others.

Channel models: Additive White Gaussian Noise (AWGN) channel, Rayleigh Fading Channel, Rician Fading Channel, wireless erasure channel, among others.

Signal-to-Noise Ratio (SNR) settings: Typically ranging from−5 dB to 20 dB, to evaluate the robustness of methods under low, medium, and high SNR conditions.

Rician K-factor: Typically set between 0 and 10+, to assess the robustness of communication systems in environments ranging from no line-of-sight to strong line-of-sight conditions.

3.2. Comparative Analysis of Typical Lightweight Semantic Communication Techniques

Table 1. Comparison of Representative Lightweight Semantic Communication Techniques

Technology Type	Typical Method	Task Type	Dataset / Channel Model	Compression Ratio[a]	Task Performance Change	Communication Performance Change	Computational Performance Change	Applicable Scenario	Reference
Model Compression	L-DeepSC	Text Transmission	European Parliament Corpus; AWGN (SNR = 12 dB)	2.5×	BLEU: 0.995 → 1.0 (↓0.5%)	bpp: 1.65% → 0.15% (↓91%)	Model size: 12.3 → 1.28 MB (↓89.6%)	Industrial IoT	[3]
	Q-GESCO	Image Semantic Transmission	Cityscapes; AWGN (SNR = 10 dB)	4×	FID: 18.2 → 19.4 (↑6.6%)	bpp: 1.55% → 0.02% (↓98.7%)	FLOPs: 100% → 21% (↓79%)	Intelligent Transportation	[6]
	ECG-SC-DARTS	Real-time Physiological Signal Classification and Monitoring	MIT-BIH Arrhythmia; AWGN / Rayleigh / Rician (SNR = −15 to 20 dB)	2.05×	F1-score ↑ to 92.1%	Bandwidth consumption ↓50%	FLOPs ↓30–40%; Memory usage ↓25%	Smart Healthcare	[13]
Codec Optimization & Lightweight Architectures	LICRnet	Image Semantic Compression and Reconstruction	ImageNet + CLIC; AWGN + Rician (K = 2.8)	4.0×	PSNR ↑0.12 dB; MS-SSIM ↑1.2%	bpp ↓98%	FLOPs ↓79%; Memory ↓75%	Industrial IoT	[1]
	CS-ASIC	Image Semantic Compression and Task Analysis	Cityscapes / KITTI; AWGN (SNR = 10 dB)	3.6×	mIoU ↑1.13%; mAP ↑1.15%	bpp ↓98.5%	FLOPs ↓99%; Model size ↓	Industrial IoT	[2]
	LSSC	Image Semantic Segmentation	In-house Portrait Dataset; AWGN (SNR = 1–20 dB)	2.2×	mIoU ↓3% (vs. UNet)	Bandwidth requirement ↓55%	FLOPs ↓14%; Model size ↓55%	Smart Home	[14]
Dynamic Resource Allocation	SBQ-DRL	Text Semantic Transmission	European Parliament Corpus; OFDM + Rician Fading Channel (SNR = −5 to 20 dB)	Dynamic adjustment	Semantic similarity ↑50%	Bandwidth efficiency ↑24%	Real-time decision-making	Industrial IoT	[15]
	MSS+ESC	Image Transmission and Reconstruction	COCO / Cityscapes; Rayleigh / Rician (SNR = −5 to 20 dB)	10.0×	PSNR ↑1.2 dB; mIoU ↓1.3%	bpp ↓55%; Bandwidth consumption ↓58%	Client-side FLOPs ↓40%; Memory ↓45%	Smart Healthcare	[16]

Note [a]: Compression Ratio: the ratio of the number of parameters in the original model to that in the compressed model.

As shown in Table 1, different technical approaches exhibit complementary advantages in terms of performance metrics.

Model compression methods (e.g., L-DeepSC, Q-GESCO) significantly reduce computational complexity and storage requirements by streamlining model architectures, while maintaining stable task performance. These methods are particularly suitable for deployment of resource-constrained end devices

Codec optimization methods (e.g., LICRnet, CS-ASIC) excel in data compression efficiency, achieving substantial reductions in bandwidth consumption, thereby providing effective solutions for bandwidth-limited or high-latency communication environments.

Dynamic resource allocation approaches (e.g., MSS+ESC, E2E Semantics policy) demonstrate strong environmental adaptability, enabling real-time adjustment of resource allocation based on channel conditions and task demands, thus sustaining efficient communication performance in complex and dynamic network environments.

Overall, these different technical pathways can be combined and tailored according to specific application requirements.

3.3. Typical Application Scenarios

Lightweight semantic communication technologies demonstrate significant value across diverse application scenarios. Different scenarios prioritize trade-offs between task performance, communication performance, and computational performance differently, necessitating the selection of appropriate technical solutions based on specific requirements. The following sections provide a review and analysis of three typical scenarios: Industrial Internet of Things (IIoT), Smart Healthcare, and Smart Home.

(1) Industrial Internet of Things (IIoT).

IIoT environments typically deploy numerous sensors and actuator terminals, requiring low-latency, highly reliable data transmission and processing under noisy interference and bandwidth constraints [4]. Consequently, the primary demand for drivers include:

Edge Collaboration: Requires device-edge-cloud collaborative processing to alleviate pressure on central nodes.

Low Latency & High Reliability: Production monitoring and predictive maintenance necessitate millisecond-level response times.

Adaptability to Channel Variations: Requires dynamic resource allocation to cope with interference fluctuations in industrial settings.

As shown in Table 1, CS-ASIC [2] optimizes image semantic compression, achieving a 3.6x compression ratio, improving mAP by 1.15% in object detection tasks while reducing FLOPs by 99%, making it suitable for edge device deployment. L-DeepSC [3] employs structured pruning and quantization to achieve a 2.5x compression ratio, maintaining near-lossless BLEU score in text transmission while significantly reducing computational load and bandwidth demand. Dynamic resource allocation techniques (e.g., SBQ-DRL [15]) use deep reinforcement learning to adaptively adjust bandwidth, improving semantic similarity by 50%, thereby strongly supporting real-time control scenarios like predictive maintenance. These technologies achieve substantial resource savings (computation, bandwidth, storage) with minimal performance degradation. Combined with dynamic optimization, they ensure the real-time and reliable execution of critical tasks in complex industrial environments, significantly enhancing system energy efficiency.

(2) Smart Healthcare

In Smart Healthcare scenarios, transmitted content often consists of high-resolution medical images and real-time physiological data, demanding extremely high task performance (e.g., diagnostic accuracy) while simultaneously satisfying dual constraints of data privacy and low energy consumption [11, 18]. The primary demand drivers are:

High Accuracy & Stability: Errors directly impact diagnostic conclusions.

Privacy Protection: Requires local inference or encrypted transmission mechanisms.

Low Energy Consumption: Portable medical devices require prolonged operation.

Table 1 indicates that ECG-SC-DARTS [13], targeting ECG classification tasks, elevates the F1-score to 92.1% even under harsh channel conditions (SNR = -15 dB), while reducing bandwidth usage by 50% and memory consumption by 25%. This meets the resource constraints of wearable devices and helps reduce misdiagnosis risks. MSS+ESC [16] achieves ultra-high image compression (10x), reducing client-side computational load by 40%, balancing medical image reconstruction quality with terminal energy efficiency. This makes it suitable for remote diagnosis and mobile monitoring. These technologies are crucial for resource-constrained mobile medical terminals, reducing communication resource consumption and energy usage while guaranteeing diagnostic accuracy.

(3) Smart Home

Smart Home scenarios involve numerous terminal devices handling diverse data types (images, voice, sensor signals, etc.), experience fluctuating communication conditions, and emphasize low power consumption and long-term stable operation [8, 9]. The main demand for drivers includes:

Device Heterogeneity: Significant variations in computational power across different terminals.

Event-Driven Sampling: Data acquisition and transmission occur only upon event triggers to minimize power consumption.

Adaptive Transmission: Requires dynamic adjustment of compression rates and transmission strategies under fluctuating network conditions.

As seen in Table 1, Q-GESCO [6] optimizes image semantic transmission, achieving a 4x compression ratio. While FID slightly increases by 6.6%, FLOPs are reduced by 79%, balancing visual quality with energy consumption and supporting real-time response in scenarios like security monitoring. LSSC [14] reduces bandwidth demand by 55% while sacrificing only 3% mIoU accuracy (compared to UNet), enabling deployment across devices ranging from low-power sensors to mid-range cameras. In home environments, these methods skillfully balance accuracy and resource consumption, trading minimal perceived quality degradation for significant bandwidth savings and computational load reduction, thereby enhancing user experience.

In summary, IIoT emphasizes low latency and high reliability, favoring a combination of high compression ratios and dynamic resource allocation. Smart Healthcare prioritizes high accuracy and privacy protection, best served by high-fidelity codec optimization and energy-efficient techniques. Smart Home focuses on low power consumption and environmental adaptability, benefiting from event-driven approaches and bandwidth compression techniques. This scenario-specific technology mapping not only provides guidance for practical engineering deployment but also lays the groundwork for designing future device-edge-cloud semantic communication systems capable of multi-scenario collaboration.

4. Technical Challenges and Future Directions

Despite demonstrating promising application prospects across multiple typical scenarios, lightweight semantic communication still faces several critical challenges during practical deployment and large-scale adoption, which require resolution through future research.

4.1. Enhancing Robustness

In real-world channel environments, noise, fading, interference, and packet loss can cause semantic information distortion. For lightweight models, structural simplification may further weaken their fault tolerance under abnormal conditions. Specifically, compression techniques for diffusion models like Q-GESCO [6] show increased sensitivity to channel noise in industrial electromagnetic interference environments after simplification. To address this, future research could investigate multi-modal redundant encoding, adaptive error recovery mechanisms, and meta-learning-based robustness enhancement methods. These approaches would enable models to maintain semantic integrity under varying channel conditions. For example, dynamically adjusting encoding granularity by combining channel state prediction with semantic importance weighting could minimize critical semantic loss.

4.2. Balancing Privacy and Performance

In high-sensitivity scenarios like smart healthcare and industrial IoT, data privacy protection is a mandatory requirement. However, traditional privacy-preserving methods (e.g., homomorphic encryption, differential privacy) introduce additional computational and communication overhead, impacting real-time performance and energy efficiency. While existing technologies effectively prevent data leakage, the trade-off between performance degradation from privacy protection and model capacity compression remains suboptimal. Therefore, future work should explore lightweight privacy-preserving mechanisms for semantic communication, such as:

Feature space anonymization.

Device-edge collaborative encryption.

Low-complexity secure encoding.

Additionally, by drawing inspiration from dynamic resource allocation techniques like MSS+ESC [16], researchers could develop privacy-performance joint optimization frameworks. These would use multi-objective optimization algorithms to achieve dynamic balance between protection levels and transmission efficiency.

4.3. Cross-Device Semantic Consistency

In device-edge-cloud scenarios with multi-device collaboration, differences in semantic encoding models across devices may lead to inconsistent semantic representations, ultimately affecting task performance. Lightweight design could further amplify the impact of model discrepancies. For instance, smart home terminals range from low-power sensors to intelligent gateways with significant heterogeneity. While existing dynamic resource allocation techniques (e.g., E2E semantic policy [17]) can dynamically adjust the number of knowledge vectors and symbols, they still lack sufficient capability for extreme adaptation to ultra-low-computation devices and guaranteeing cross-device semantic consistency. Future research should focus on:

Semantic alignment mechanisms for heterogeneous devices.

Cross-model feature mapping.

Unified embedding space construction.

Online alignment update strategies.

Furthermore, introducing adaptable modules into lightweight codec architectures could facilitate semantic interoperability across different devices.

4.4. Standardization and Reproducibility

The field currently lacks unified evaluation standards and publicly reproducible benchmark systems, making direct comparison between different research outcomes challenging. To address this gap, efforts should be promoted:

Standardized test datasets.

Evaluation of metric systems.

Open-source implementation frameworks.

These should establish unified evaluation benchmarks covering multiple scenarios, modalities, and metrics. Concurrently, industry-academia collaboration should be encouraged to advance protocol standardization, ensuring seamless migration of lightweight semantic communication technologies across different platforms and devices.

5. Conclusion

This paper presents a comprehensive survey of lightweight semantic communication technologies for Resource-Constrained Internet of Things (RC-IoT) systems. We propose a unified taxonomy and evaluation framework, systematically reviewing three major categories of key technologies: model compression, codec optimization and lightweight architecture, and dynamic resource allocation. Practical adaptation recommendations are provided for three representative application scenarios: industrial IoT, smart healthcare, and smart home environments.

This survey demonstrates that lightweight semantic communication has attained practical viability for deployment in diverse resource-constrained settings. Nevertheless, significant challenges persist, including:

Maintaining semantic consistency across heterogeneous devices.

Balancing privacy protection with system performance.

Enhancing robustness under dynamic channel conditions.

Establishing standardized and reproducible evaluation frameworks.

Future research should prioritize cross-modal optimization, adaptive scheduling, and the development of open-source, reproducible platforms. Addressing these directions will accelerate the advancement of lightweight semantic communication toward efficient, secure, and intelligent deployment in broader 6G-enabled applications.

References

[1] X. Yu, D. Li, N. Zhang, and X. Shen, “A novel lightweight joint source-channel coding design in semantic communications,” IEEE Internet of Things J., vol. 12, no. 11, pp. 18447–18450, Jun. 2025.

[2] Y. Huang, B. Chen, J. Zhang, Q. Han, and S.-T. Xia, “Compressive sensing based asymmetric semantic image compression for resource-constrained IoT system,” in Proc. DAC, San Francisco, CA, USA, Jul. 2022, pp. 877–882.

[3] H. Xie and Z. Qin, “A lite distributed semantic communication system for internet of things,” IEEE J. Sel. Areas Commun., vol. 39, no. 1, pp. 142–153, Nov. 2021.

[4] B. Xie, Y. Wu, Y. Shi, D. W. K. Ng, and W. Zhang, "Communication-efficient framework for distributed image semantic wireless transmission," *IEEE Internet Things J.*, vol. 10, no. 24, pp. 22555–22567, 15 Dec. 2023.

[5] L. Yan, Z. Qin, R. Zhang, Y. Li, and G. Y. Li, "Resource allocation for text semantic communications," IEEE Wireless Commun. Lett., vol. 11, no. 7, pp. 1394-1398, July 2022

[6] E. Grassucci, G. Pignata, G. Cicchetti, and D. Comminiello, “Lightweight Diffusion Models for Resource-Constrained Semantic Communication,” IEEE Wireless Commun. Lett., to be published, DOI: 10.1109/LWC.2025.3578724.

[7] E. Abu-Helalah, J. Serra, and J. Perez-Romero, “On Splitting Lightweight Semantic Image Segmentation for Wireless Communications,” arXiv preprint arXiv:2507.14199v1, Jul. 2025.

[8] B. Chatterjee, N. Cao, A. Raychowdhury, and S. Sen, "Context-Aware Intelligence in Resource-Constrained IoT Nodes: Opportunities and Challenges," IEEE Des. Test, vol. 36, no. 2, pp. 7-40, Mar./Apr. 2019, doi: 10.1109/MDAT.2019.2899334.

[9] Y. Shao, Q. Cao, and D. Gündüz, "A theory of semantic communication," IEEE Trans. Mobile Comput., vol. 23, no. 12, pp. 12211–12234, Dec. 2024.

[10] J. Peng, H. Xing, Y. Li, L. Feng, L. Xu, and X. Lei, “Task-Oriented Multi-User Semantic Communication with Lightweight Semantic Encoder and Fast Training for Resource-Constrained Terminal Devices,” IEEE Wireless Commun. Lett., vol. 13, no. 9, pp. 2427–2431, Sep. 2024.

[11] 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. 2663–2675, 2021.

[12] C. Liu, Y. Zhou, Y. Chen, and S.-H. Yang, ‘‘Knowledge distillation based semantic communications for multiple users,’’ IEEE Trans. Wireless Commun., Dec. 2023.

[13] H. Xing, H. Ma, Z. Xiao, X. Wang, B. Zhao, S. Luo, L. Feng, and L. Xu, “On ECG signal classification: An NAS-empowered semantic communication system,” in Proc. IEEE 22nd Int. Conf. Trust, Secur. Privacy Comput. Commun., Exeter, U.K., 2023, pp. 470–476.

[14] G. Ma, H. Tong, N. Yang, and C. Yin, “Attention-based UNet enabled lightweight image semantic communication system over Internet of Things,” in Proc. IEEE Wireless Communications and Networking Conference (WCNC), 2024, pp. 1–6.

[15] L. Wang, W. Wu, F. Zhou, Z. Yang, Z. Qin, and Q. Wu, "Adaptive resource allocation for semantic communication networks," IEEE Transactions on Communications, vol. 72, no. 11, pp.6900-6916, 2024.

[16] F. Jiang, S. Tu, L. Dong, K. Wang, K. Yang, R. Liu, C. Pan, and J. Wang, “Lightweight vision model-based multi-user semantic communication systems,” arXiv preprint arXiv:2502.16424v1, Feb. 2025.

[17] M. Kountouris and N. Pappas, "Semantics-Empowered Communication for Networked Intelligent Systems,"arXiv preprint arXiv:2007.11579v3, Mar.2021

[18] K. Khatiwada, J. Hopper, J. Cheatham, A. Joshi, and S. Baidya, "Large Language Models in the IoT Ecosystem - A Survey on Security Challenges and Applications,"arXiv preprint arXiv:2502.17586v1, May. 2025.