Research on Semantic Communication and AI-Driven Cloud-Edge Resource Collaborative Transmission Strategies under the 6G Vision

Xiaolu Wang[1], Peiying Zhang[2]
[1]Department of Intelligent Engineering, Jiangsu Vocational Institute Of Commerce, Nanjing 211168, Jiangsu, China
[2]China University of Petroleum (East China), College of Computer Science and Technology, Qingdao 266580, Shandong, China

Corresponding author: Xiaolu Wang (e-mail: xiaolu_w2023@163.com)

ABSTRACT To address the extreme requirements of low latency, high energy efficiency, and resource coordination for massive heterogeneous services under the 6G vision, existing cloud-edge transmission mechanisms lack deep coupling between semantic information representation and dynamic resource scheduling, making it difficult to simultaneously meet the semantic fidelity and real-time requirements of multimodal services. Therefore, this paper proposes a cloud-edge resource collaborative transmission strategy driven by semantic communication and AI. First, a hierarchical abstract model of multimodal semantic information is constructed, dividing semantics into signal, object, and target layers to provide differentiated semantic fidelity descriptions for different services. Second, a two-layer cloud-edge collaborative decisionmaking architecture based on deep reinforcement learning is designed: the upper-layer cloud center agent predeploys semantic models and plans baseline resource quotas on a minute-level cycle; the lower-layer edge node agent completes task offloading, joint scheduling of computing and communication resources within milliseconds. Through an adaptive optimization mechanism of immediate feedback in the inner loop and periodic aggregation in the outer loop, the strategy achieves continuous closed-loop evolution. Simulation results show that under balanced mixed load, the proposed strategy achieves an effective throughput of 215.7 tasks/second, an average processing latency of 36.5 milliseconds, an average edge computing resource utilization of 88.7%, and a total system energy efficiency of 24.1 tasks/kJ, verifying its comprehensive superiority in terms of throughput, latency, resource utilization, and energy efficiency. As the work targets semantic communication and cloud-edge transmission, it directly concerns electromagnetic wave propagation and resource scheduling in 6G wireless networks.

INDEX TERMS semantic communication,6g network,cloud-edge collaboration,deep reinforcement learning,electromagnetic propagation

I. INTRODUCTION

THE sixth-generation mobile communication system (6G) will support the presence of all air, land, sea, and air coverage supporting new service types, including immersive cloud XR, holographic communication, digital twins, and autonomous driving. These services present new challenges to communication networks: peak data rates must be in the Tbps range, end-to-end latency must be down to sub-milliseconds, connection density must be tens of millions of devices per square kilometer, and energy efficiency and spectral efficiency must be enhanced tens of times. The traditional communication paradigms founded on bit-precise transmission are guided by the information theory of Shannon, which approximates the capacity of the channel by the enhancement of signal-to-noise ratio, augmenting bandwidth, or enhancing the coding and modulation techniques, but not the redundancy of semantics of the information source, or the divergent demands of various applications to information content. Against the backdrop of the exponential growth of massive data in 6G and the dwindling amount of spectrum available, just the enhancement of the physical layer has already been reducing to the theoretical threshold, and novel leaps at a new level of information processing are urgently desired.

As a new communication paradigm, semantic communication is based on the principle that the correctness of the tasks is central, rather than the correctness of bits. This includes the extraction and transmission of the semantic information which is most likely to be relevant to target task in the sending end and reconstruction and reasoning on the same based on a common semantic knowledge base in the receiving end. This paradigm can dramatically minimize

the data that is sent across the channel and decreases the needs of channel quality which represents a new avenue of the extreme performance needs of 6G. However, the success of semantic communication depends significantly on two major aspects: the effective and faithful semantic abstraction of multimodal data (text, images, video, and sensor data), and the efficient representation of this data during processing; and second, the rational utilization of semantic processing operations (semantic encoding, decoding, and reasoning) in a cloud-edge-device collaborative computing infrastructure to address the latency and resource constraints of various services. Although the technology of cloud-edge collaborative has already advanced to some degree, most studies remain centered around the offloading of conventional computing jobs and resource distribution without involving in-depth modeling of the layer of semantic information and also not incorporating the semantic fidelity requirements as the guiding principle of resource allocation as the core decision-making principle of resource allocation. Consequently, integrating the unique advantages of semantic communication with AI-driven cloud-edge resource collaboration has emerged as a critical research priority for 6G networks.

The three primary contributions of the paper are as follows: First, it suggests a hierarchical abstract model of multimodal semantic information, breaking semantics down into layer signal, layer object, and layer target, and offers a level of semantic fidelity requirement per business, thereby offering a finer frame of semantic description to differentiated business requirements and a quantitative parameter on latter resource scheduling decisions. Second, it implements a two-layer deep reinforcement learning cooperative decision making architectures: the upper-layer cloud center agent performs long-cycle (minute-level) semantic model pre-deployment and baseline resource quota planning, and the lower-layer edge node agent performs short-cycle (milliseconds-level) real-time task offloading and multi-dimensional resource joint scheduling, effectively resolving the coupling problem between decisions at various time scales. Third, it builds a closed-loop adaptive optimization system: the inner loop uses real-time feedback to steer online fine-tuning of the lower-layer strategy, the outer loop periodically summarizes data to steer regular updates of the upper-layer model, and joint retraining in case of business model drift to achieve a continuous evolution and strong adaptation of the system strategy.

II. RELATED WORK

Under the 6G vision, the demand for massive IoT devices, ultra-low latency communication and high spectral efficiency transmission is driving cloud-edge computing toward deep collaboration. Ebrahimi Mood et al. [1] used a balanced optimization algorithm to perform evolutionary training on the weights and structure of a recurrent neural network to improve the ability to predict and schedule cloud-edge resource status. Souri et al. [2] constructed a dynamic trust model by evaluating the historical behavior of devices and identity credentials, and integrated it into the cloud-edge resource management framework to achieve secure and reliable resource scheduling and sharing. Zhang et al. [3] used the idea of ensemble learning to combine multiple evolutionary strategies to solve the high-dimensional multi-objective optimization problem in edge cloud resource scheduling. Lyu et al. [4] designed a heterogeneous cloud-edge collaborative computing architecture and introduced an affinity-aware workflow scheduling strategy. Kumaran and Sasikala et al. [5] used a dynamic arithmetic optimization algorithm to dynamically adjust the hyperparameters and exploration strategy of DDQN to enhance the joint learning ability of task offloading and resource allocation. Mansouri et al. [6] proposed a resource utilization optimization framework for edge cloud distributed databases. Kaftantzis et al. [7] designed a variety of representative resource management strategies and conducted comparative experiments under different load and network conditions. Qadeer and Lee [8] adopted a hierarchical reinforcement learning structure, with the upper layer performing coarse-grained allocation of multiple resources and the lower layer combining heuristic rules for fine-grained adjustment, to achieve efficient joint allocation of multiple resources. Kumar and Agrawal [9] dynamically pre-allocated resources by predicting device movement trajectories and event triggering times, and entered a low-power mode during inactive periods. Wang et al. [10] designed a low-complexity iterative algorithm to alternately optimize offloading decisions and computing/communication resource allocation. Existing research lacks explicit modeling of the semantic hierarchical structure of information. This paper proposes a cloud-edge collaborative transmission strategy driven by semantic communication and AI, and establishes an intrinsic mapping relationship between information value and resource supply from two dimensions: semantic hierarchical abstraction and two-layer deep reinforcement learning decision-making.

III. METHODS

A. HIERARCHICAL ABSTRACTION AND REPRESENTATION MODEL OF MULTIMODAL SEMANTIC INFORMATION

The effectiveness of semantic communication depends first and foremost on the proper definition and extraction of the "semantics" of information. Considering the highly heterogeneous nature of 6G services (such as text, image, video, and sensor data streams), this study proposes a hierarchical abstraction model for multimodal semantic information. This model divides semantics into three layers: the signal layer, the object layer, and the target layer. The signal layer focuses on low-level features of the data, utilizing traditional Auto-Encoder (AE) structures to compress raw pixel or waveform data into latent space representations; the object layer employs Convolutional Neural Networks (CNNs) and Graph Neural Networks (GNNs) to identify and extract structured entities (e.g., bounding boxes and class

labels) from the decoded signal features; and the target layer utilizes Transformer-based reasoning engines to understand the logical relationships between entities to generate task-specific commands, such as trajectory predictions. The degree of semantic abstraction varies across different layers, and the required computational complexity and the amount of semantic representation data generated also differ by orders of magnitude. For example, for autonomous driving video streams, transmitting signal layer semantics (compressed video) requires Mbps of bandwidth, transmitting object layer semantics (structured target list) requires Kbps of bandwidth, while transmitting target layer semantics (a warning command) may only require a few hundred bits. This model defines a "semantic fidelity requirement level" for each network access service. This level determines the highest level of abstraction that can be performed for semantic extraction under the premise of meeting task performance, providing a differentiated decision basis for subsequent resource scheduling [11,12]. Specifically, for the target layer (e.g., warning commands), semantic fidelity is mathematically defined as the "Task-Success Probability" (P_{ts}), calculated as the inverse of the semantic loss between the inferred action and the ground-truth objective. For a warning command, fidelity is binary-weighted based on the critical latency threshold; if the command arrives after the safety-critical window, the fidelity is recorded as zero, regardless of bit-level accuracy.

B. A TWO-LAYER CLOUD-EDGE RESOURCE COLLABORATIVE DECISION-MAKING MODEL BASED ON DEEP REINFORCEMENT LEARNING

Resource collaborative decision-making serves as the centralized decision engine of the proposed framework. This study designed a two-layer deep reinforcement learning model to deal with decision problems of different time scales and spatial ranges [13] .

1) Upper Layer (Cloud Center - Macro-level Collaboration):

The decision-making timescale of this layer is between minutes and hours. The agent is located in cloud center, its state space consists of: the past request distribution and forecasted traffic of services of varying semantic fidelity in the coverage region of each edge node in the whole network; the average utilization rate of computing resource (CPU/GPU) and the condition of storage resource of each edge node; and the version and scale of the global semantic model library in the cloud center. Its action space is the choices about the set of semantic model types and versions that have been pre-deployed to each edge node and the quota of the computing resources that each edge node is entitled to. Its reward functionality tries to maximize the long-term benefits and can be defined as: the aggregate of the total utility of semantic tasks processed divided by the amount of energy consumption in the entire network. The utility of the task is in connection with the level of semantic fidelity and processing latency contentment. The DRL model in this layer periodically monitors the state of the globe and provides pre-deployment strategies, so that edge nodes can be able to respond quickly to the majority of routine semantic tasks, without having to relay them back to the cloud and collaboratively process them.

2) Lower Layer (Edge Node - Micro-Scheduling):

This layer's decision-making timescale is from milliseconds to seconds. Every edge node runs a local DRL agent. Its state space consists of: the set of attributes of all semantic tasks in the node at the current time (e.g., data type, semantic fidelity requirement level, data size, maximum tolerable latency); the real-time computing resources available to the node, wireless channel status information, and network connection status with adjacent nodes. It operates in a multi-dimensional joint space that includes:

1) Task offloading decision: whether to process locally, or to offload tasks to a nearby edge node (cooperatively) or to offload tasks to the cloud;

2) Resource allocation decision: how to allocate particular computing resources (CPU cores/frequency) and communication resources (transmission power, sub-channels) to tasks being processed locally.

Its reward functionality is concerned with immediate performance and is defined as: utility of all tasks that the node processes less the resource usage cost (computing and communication energy consumption). This layer of DRL is acquired to decide the best joint offloading and resource allocation decisions in the dynamic and uncertain network conditions by interacting with the environment in real time to handle the sudden traffic bursts and local resource bottlenecks.

C. STRATEGY EXECUTION, FEEDBACK, AND ADAPTIVE OPTIMIZATION MECHANISM

The output of the decision model requires an efficient execution and feedback mechanism. This study designs a policy execution engine that receives the joint actions of the lower-level DRL agents and converts them into specific system instructions: invoking the corresponding local or remote semantic processing models, allocating resources in the specified computing containers, and scheduling data transmission between the physical layer and the MAC layer. Simultaneously, the system incorporates a comprehensive monitoring and feedback module that continuously collects key performance indicators (KPIs), including the actual end-to-end latency, processing accuracy, energy consumption, and resource utilization of each node for each semantic task. The feedback mechanism is categorized into instantaneous telemetry and periodic aggregation. For the inner loop, raw KPIs (e.g., transmission success and queue length) are directly mapped to the reward function of the lower-layer DRL agent via a hardware-abstracted monitor at the edge node, bypassing complex data cleaning to ensure microsecond-level reward updates. The historical KPI data are then cleaned and aggregated asynchronously to form the

outer loop feedback. The outer loop feedback is periodically (e.g., hourly) aggregated to the cloud center to evaluate the effectiveness of the upper-level macro-cooperative strategy and as a dataset for retraining or updating the upper-level DRL model. In addition, when the business model changes significantly (such as the launch of a new application causing a change in the distribution of semantic fidelity requirements), the monitoring module will trigger an alarm and start the joint retraining process of the upper and lower DRL models, thereby achieving closed-loop adaptive optimization of the strategy and ensuring that the system maintains the best performance in the long term [14,15].

IV. RESULTS AND DISCUSSION

The proposed strategy is evaluated using an OMNeT++ based simulation platform integrated with Python-based DRL libraries. The simulated network topology consists of 1 central cloud node and 10 distributed edge nodes within a 1km x 1km urban area. Wireless channels are modeled with Rayleigh fading and path loss following the 3GPP TR 38.901 standard. For the DRL agents, both upper and lower layers utilize the Deep Q-Network (DQN) architecture with an Adam optimizer (learning rate = 0.001) and are trained over 2,000 episodes until the reward curves converge. Computation capacities are set to 50 GHz for the cloud and 5-10 GHz for edge nodes.

A. SYSTEM THROUGHPUT AND TASK LATENCY PERFORMANCE ANALYSIS

The system's effective throughput is defined as the total "value" of semantic tasks successfully processed per second that meet latency requirements. In this study, a "task" is defined as a complete end-to-end multimodal service request (e.g., an image-based target recognition or a sensor-based anomaly detection) from generation to final decision execution. A "standard task" is operationally defined as the processing of a 1MB data unit with moderate semantic complexity (Object Layer) to provide a normalized benchmark for throughput calculations. Average task processing latency includes computation latency, transmission latency, and potential queuing latency. Table 1 shows performance comparison data under four strategies.

As shown in Table 1, the proposed strategy significantly outperforms other comparative strategies in all metrics. Compared to Strategy A, the proposed strategy improves effective throughput by 72% and reduces average latency by 57.2 % , directly demonstrating the enormous potential of semantic communication in removing redundancy and improving efficiency. Compared to Strategy B, the proposed strategy still improves in both throughput and latency, highlighting the superiority of AI-based intelligent dynamic collaboration over fixed rules. Although Strategy C also employs DRL, its single-layer structure makes it difficult to accurately coordinate the relationship between long-cycle model deployment and short-cycle real-time scheduling. The proposed two-layer structure, through the pre-configuration of semantic models and baseline resources in the upper layer, creates a more stable and resource-ready local environment for the lower layer, allowing the lower-layer DRL to focus more on handling real-time dynamic changes. The latency satisfaction rate reaches 96.1%, indicating that the proposed strategy can reliably support latency-critical 6G applications.

B. CLOUD-EDGE RESOURCE UTILIZATION AND ENERGY EFFICIENCY ANALYSIS

Resource utilization primarily examines the average utilization rate of computing resources at edge nodes. System energy efficiency is defined as the amount of effective semantic tasks completed per joule of energy (tasks/joule). The results are shown in Table 2.

As shown in Table 2, the strategy presented in this paper achieves the highest edge computing resource utilization (88.7%) and a more balanced cloud load. Strategy B, due to its fixed threshold triggering of offloading, is prone to the "ping-pong effect" and uneven resource utilization, resulting in a higher peak cloud utilization. Although the single-layer DRL of Strategy C can learn scheduling, it lacks upper-layer macro-level pre-configuration, which sometimes forces edge nodes to directly offload a large number of tasks to the cloud due to model missingness, increasing unnecessary data transmission and cloud load. The upper-layer DRL in this paper, through intelligent model pre-deployment, enables more than 95% of routine semantic tasks to find matching models for processing on the edge side, thereby significantly improving edge resource utilization and smoothing cloud load. In terms of energy efficiency, the strategy presented in this paper is about 17.08 % higher than Strategy A and 31% higher than Strategy C. This is mainly due to two aspects: first, semantic communication significantly reduces the amount of data that needs to be transmitted, reducing communication energy consumption; second, intelligent collaborative scheduling reduces the frequency of long-distance backhaul and cross-node collaboration, and more rationally allocates computing tasks to edge nodes with better energy efficiency, thereby achieving joint optimization of computing and communication energy consumption.

C. ADAPTABILITY ANALYSIS OF THE STRATEGY TO DIFFERENT BUSINESS LOADS

To test the robustness of the strategy, this study varied the mixing ratio of the three semantic service traffic types in the simulation, setting up three typical load scenarios: "video stream-dominated," "sensor data-dominated," and "balanced mixture," and observed the changes in system performance. The results are shown in Table 3.

Table 3 shows that the proposed strategy maintains stable high performance under different workloads. In the scenario of "sensor data as the main source," the system performs best in terms of throughput, latency, and energy efficiency because the object layer semantic extraction task has lower computational complexity and smaller data packets com-

TABLE 1. Comparison of system throughput and task latency performance under different strategies

Performance metrics	Strategy A (Traditional Static)	Strategy B (Semantics + Fixed Rules)	Strategy C (Semantic + Single-layer DRL)	This article's strategy (semantic + two-layer DRL)
System effective throughput (tasks/second)	125.4	163.8	178.6	215.7
Average task processing latency (ms)	85.2	52.3	41.7	36.5
Latency satisfaction rate (latency < 50ms)	71.5%	88.2%	92.8%	96.1%

TABLE 2. Comparison of resource utilization and system energy efficiency under different strategies

Performance metrics	Strategy A	Strategy B	Strategy C	This article's strategy
Edge average computing resource utilization	63.2%	78.5%	82.1%	88.7%
Peak utilization of cloud computing resources	22.4%	35.6%	30.2%	28.8%
Total Energy Efficiency of the System (per Task/kJ)	8.9	15.6	18.4	24.1

TABLE 3. Performance of the proposed strategy under different business load scenarios

Load scenarios	Effective throughput (tasks/second)	Average latency (ms)	marginal resource utilization rate	System energy efficiency (per task/kJ)
Scenario 1: Primarily video streaming (70%)	198.5	39.2	86.3%	21.8
Scenario 2: Primarily sensor data (70%)	245.6	28.1	91.5%	28.9
Scenario 3: Balanced mixing (each ~33%)	215.7	36.5	88.7%	24.1

pared to video target layer semantic extraction. In the high-load, complex task scenario of "video stream as the main source," although the performance indicators decrease slightly, they are still significantly better than the performance of other strategies in Table 1 under balanced load. This proves that the upper-layer DRL can adaptively adjust the pre-deployed model type (such as deploying more video understanding models) based on the business distribution prediction (part of the state input), while the lower-layer DRL can also learn to allocate more resources to computationally intensive video semantic tasks or make offloading decisions earlier. The adaptive optimization mechanism of the strategy ensures that it can adjust the decision logic with changes in business patterns, demonstrating good robustness and generalization ability.

V. CONCLUSION

This paper explores an innovative strategy for the integration of semantic communication and AI-driven cloud-edge resource collaborative transmission, aiming to achieve a systematic breakthrough from both information theory and resource regulation dimensions. The research first constructs a multimodal semantic information hierarchical abstract model, providing a refined semantic description framework for differentiated business needs. Based on this, an innovative two-layer intelligent collaborative decision-making architecture based on deep reinforcement learning is proposed. This architecture performs macro-level semantic model pre-deployment and resource planning at the cloud center, and micro-level real-time task offloading and multidimensional resource joint scheduling at the edge nodes, achieving continuous adaptive optimization of the strategy through a dual feedback loop. Through large-scale simulation experiments and multi-dimensional comparative analysis, this study verifies the comprehensive superiority of the proposed strategy. Further research directions exist. First, current work mainly focuses on cloud-edge collaboration within a single management domain; future research should investigate how to achieve secure sharing of semantic knowledge and reliable resource collaboration across multiple operators or management domains. Second, the semantic hierarchical model in this paper is relatively static; future research could explore dynamic semantic abstraction hierarchical decision-making mechanisms based on online learning. Finally, a deeper integration of semantic communication with wireless physical layer technologies (such as smart reflectors and reconfigurable smart surfaces) to achieve end-to-end joint optimization of "semantics-resources-channels" is an essential path towards 6G full-domain intelligent communication. This research provides valuable theoretical references and feasible technical solutions for building efficient, green, and adaptive future 6G networks.

AUTHOR CONTRIBUTIONS

Conceptualization –Xiaolu Wang and Peiying Zhang; methodology – Xiaolu Wang and Peiying Zhang; investigation – Xiaolu Wang and Peiying Zhang; writing-original draft preparation – Xiaolu Wang and Peiying Zhang. All

authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This work was sponsored in part by National Natural Science Foundation of China (No.62471493);University Research Project, Jiangsu Vocational Institute of Commerce, No. JSJMZH24001;the school-level Team: Multi-dimensional Heterogeneous Elastic Network Engineering Technology and Application Team

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] S. Ebrahimi Mood, A. Rouhbakhsh, A. Souri, et al., "Evolutionary recurrent neural network based on equilibrium optimization method for cloud-edge resource management in internet of things," Neural Computing and Applications, vol. 37, no. 6, pp. 4957-4969, 2025, doi: 10.1007/s00521-024-10929-1.

[2] A. Souri, Y. Zhao, M. Gao, et al., "A trust-aware and authentication-based collaborative method for resource management of cloud-edge computing in social internet of things," IEEE Transactions on Computational Social Systems, vol. 11, no. 4, pp. 4899-4908, 2023, doi: 10.1109/TCSS.2023.3241020.

[3] J. Zhang, Z. Ning, R. H. Ali, et al., "A many-objective ensemble optimization algorithm for the edge cloud resource scheduling problem," IEEE Transactions on Mobile Computing, vol. 23, no. 2, pp. 1330-1346, 2023, doi: 10.1109/TMC.2023.3235064.

[4] S. Lyu, X. Dai, Z. Ma, et al., "A heterogeneous cloud-edge collaborative computing architecture with affinity-based workflow scheduling and resource allocation for Internet-of-Things applications," Mobile Networks and Applications, vol. 28, no. 4, pp. 1443-1459, 2023, doi: 10.1007/s11036-023-02113-x.

[5] K. Kumaran and E. Sasikala, "An efficient task offloading and resource allocation using dynamic arithmetic optimized double deep Q-network in cloud edge platform," Peer-to-Peer Networking and Applications, vol. 16, no. 2, pp. 958-979, 2023, doi: 10.1007/s12083-022-01440-2.

[6] Y. Mansouri, V. Prokhorenko, F. Ullah, et al., "Resource utilization of distributed databases in edge-cloud environment," IEEE Internet of Things Journal, vol. 10, no. 11, pp. 9423-9437, 2023, doi: 10.1109/JIOT.2023.3235360.

[7] N. Kaftantzis, D. G. Kogias, and C. Z. Patrikakis, "Exploring the Impact of Resource Management Strategies on Simulated Edge Cloud Performance: An Experimental Study," Network, vol. 4, no. 4, pp. 498-522, 2024, doi: 10.3390/network4040025.

[8] A. Qadeer and M. J. Lee, "Hrl-edge-cloud: Multi-resource allocation in edge-cloud based smart-streetscape system using heuristic reinforcement learning," Information Systems Frontiers, vol. 26, no. 4, pp. 1399-1415, 2024, doi: 10.1007/s10796-022-10366-2.

[9] R. Kumar and N. Agrawal, "Edma-rm: an event-driven and mobility-aware resource management framework for green iot-edge-fog-cloud networks," IEEE Sensors Journal, vol. 24, no. 14, pp. 23004-23012, 2024, doi: 10.1109/JSEN.2024.3404470.

[10] S. Wang, X. Li, and Y. Gong, "Energy-efficient task offloading and resource allocation for delay-constrained edge-cloud computing networks," IEEE Transactions on Green Communications and Networking, vol. 8, no. 1, pp. 514-524, 2023, doi: 10.1109/TGCN.2023.3306002.

[11] A. Mani, G. Kavya, and B. R. T. Bapu, "A proficient resource allocation using hybrid optimization algorithm for massive internet of health things devices contemplating privacy fortification in cloud edge computing environment," Wireless Networks, vol. 30, no. 3, pp. 1187-1199, 2024, doi: 10.1007/s11276-023-03549-5.

[12] M. D. Nguyen, L. B. Le, and A. Girard, "Integrated computation offloading, UAV trajectory control, edge-cloud and radio resource allocation in SAGIN," IEEE Transactions on Cloud Computing, vol. 12, no. 1, pp. 100-115, 2023, doi: 10.1109/TCC.2023.3339394.

[13] P. S. S. R. Patchamatla, "Integrating AI for Intelligent Network Resource Management across Edge and Multi-Tenant Cloud Clusters," International Journal of Advanced Research in Computer Science & Technology (IJARCST), vol. 6, no. 6, pp. 9378-9385, 2023, doi: 10.15662/IJARCST.2023.0606003.

[14] H. Lahza, S. BR, and H. F. M. Lahza, "Adaptive multi-objective resource allocation for edge-cloud workflow optimization using deep reinforcement learning," Modelling, vol. 5, no. 3, pp. 1298-1313, 2024, doi: 10.3390/modelling5030067.

[15] Z. Nezami, E. M. Chaniotakis, and E. Pournaras, "When computing follows vehicles: Decentralized mobility-aware resource allocation for edge-to-cloud continuum," IEEE Internet of Things Journal, vol. 12, no. 13, pp. 23324-23340, 2025, doi: 10.1109/JIOT.2025.3552504.