Deep Reinforcement Learning-Enabled Resource Allocation for UAV-assisted Communications

Xuli Cai

Thesis submitted to the University of Ottawa
in partial fulfillment of the requirements for the
Master of Applied Science

School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

Declaration of Authorship

I hereby certify that this thesis is entirely my own original work except where otherwise indicated. I am aware of the University's regulations concerning plagiarism, including those concerning consequent disciplinary actions. Any use of the works of any other author, in any form, is properly acknowledged at their point of use.

Abstract

Unmanned Aerial Vehicles (UAVs) are increasingly employed in wireless networks to provide dynamic, on-demand connectivity, particularly in emergency and infrastructure-limited scenarios. This thesis presents a comprehensive AI-enabled framework that integrates user clustering, mobility modeling, and multi-agent reinforcement learning for optimizing UAV-assisted communications. The proposed system leverages a realistic user mobility model (STEP), silhouette-based K-Means clustering for UAV-UE association, and a hybrid reinforcement learning architecture combining Deep Q-Networks (DQN) and Multi-Agent Deep Deterministic Policy Gradient (MADDPG) to jointly optimize UAV placement, bandwidth allocation, and power control.

The research progresses through three stages: (1) joint resource allocation in a single-UAV static-user scenario; (2) power optimization in a multi-UAV static-user environment using user clustering and MADDPG; and (3) adaptive UAV deployment and resource scheduling in a dynamic-user setting. Simulation results demonstrate substantial improvements in data rate, UAV utility, and user coverage, with the hybrid DRL approach outperforming traditional baselines by up to 41%. The findings validate the potential of AI-driven, mobility-aware UAV coordination for scalable and intelligent next-generation wireless communication networks.

Dedication

I would like to express my deepest gratitude to my supervisor, *Professor Burak Kantarci* , for his invaluable guidance, support, and encouragement throughout the course of this thesis. His insightful advice and mentorship have played a crucial role in shaping both the direction and quality of my research.

I am also sincerely thankful to *Dr. Poonam Lohan* for her continuous support, constructive feedback, and helpful discussions that significantly contributed to the success of this work.

My heartfelt thanks go to all my lab mates — *Tony, Xinyu, Lansu, Arda, Parisa, Arild, Ghazal, Didem, Mahsa, Samhita* — whose collaboration, camaraderie, and shared knowledge have made the research environment inspiring and enjoyable.

I would also like to acknowledge all my fellow classmates in the master's program — *Han, Zijian, Xiangyi, Xu* — for their friendship, discussions, and encouragement throughout this academic journey. The experiences we have shared together have been both enriching and memorable.

I am especially grateful to my former bachelor classmate, *Renjie*, who is currently pursuing his Ph.D. at McMaster University. Working on a similar topic in the field of IoT, he has provided me with valuable insights, technical support, and constant encouragement, for which I am truly thankful.

Finally, I extend my deepest appreciation to my beloved parents for their unwavering love, patience, and support throughout my academic journey. Their belief in me has been a constant source of strength and motivation.

This thesis stands as a testament to the collective support and inspiration I have received from all of you.

Publications

- **X. Cai**, P. Lohan, and B. Kantarci, "A Novel Joint DRL-Based Utility Optimization for UAV Data Services," in 2024 IEEE 10th World Forum on Internet of Things(WF-IoT), 2024, Ottawa, Canada, pp. 930–935.

  (Published in 2024.11)

- **X. Cai**, P Lohan, B Kantarci "Multi-Agent Deep Reinforcement Learning for Optimized Multi-UAV Coverage and Power-Efficient UE Connectivity," in 2025 IEEE International Symposium on Personal, Indoor and Mobile Radio Communications(PIMRC), 2025, Istanbul, Türkiye.

  (Published in 2025.9)

- **X. Cai**, P Lohan, B Kantarci "FLARE: Flying Learning Agents for Resource Efficiency in Next-Gen UAV Networks," IEEE Networking Letters.

  (Published in 2025.9)

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

The explosive growth in mobile devices and the emergence of data-intensive applications have significantly increased the demand for high-capacity, low-latency, and adaptive wireless communication networks. Traditional terrestrial infrastructure often struggles to meet such requirements in dynamic environments, including disaster zones, rural areas, or high-density temporary events.

Unmanned Aerial Vehicles (UAVs) [1], due to their mobility, flexibility, and ease of deployment, have become a promising solution to support next-generation wireless communication. UAVs can serve as aerial base stations, mobile relays, or edge computing platforms, complementing terrestrial networks by providing rapid and adaptive coverage with improved line-of-sight (LoS) connectivity. These advantages make UAV-assisted wireless communication highly suitable for scenarios requiring on-demand, dynamic service delivery.

In particular, disaster response has emerged as one of the most impactful domains for UAV deployment. When conventional infrastructure is damaged by earthquakes, hurricanes, floods, or wildfires, UAVs can be quickly deployed to establish emergency communication links, deliver medical supplies, perform real-time surveillance, and support rescue coordination. For instance, UAV-mounted LTE/5G base stations can re-establish wireless connectivity in disconnected zones, while drones equipped with thermal imaging cameras aid in locating survivors trapped in rubble. In flood-stricken or isolated areas, UAVs reduce the delivery time of essential medicines or food packages. These capabilities have been increasingly adopted by both government and humanitarian organizations for real-time, scalable disaster management.

A detailed summary of UAV roles in disaster scenarios is provided in Table 1.1, highlighting their operational diversity across communication, logistics, sensing, and coordination domains.

Reinforcement learning (RL) [2], especially deep RL and multi-agent approaches, further enhances UAV capabilities by enabling adaptive decision-making under uncertainty. RL algorithms allow UAVs to learn optimal policies for trajectory planning, user association, and resource allocation by interacting with dynamic environments. These intelligent policies are particularly vital in time-sensitive and resource-constrained disaster contexts, where real-time adaptability and coordination are essential.

Table 1.1: UAV Applications in Disaster Scenarios

Application	**Description**
Emergency Communication	Deploy aerial base stations to restore wireless coverage.
Search and Rescue	Locate survivors using thermal/visual sensors.
Medical Delivery	Transport medical kits to inaccessible areas.
Damage Assessment	Capture aerial images for infrastructure analysis.
Environmental Sensing	Detect hazards (gas, radiation) via onboard sensors.
Network Relay	Extend signal range in blocked or rural zones.
Situational Monitoring	Provide real-time aerial views for coordination.

1.2 Motivation

While UAV-assisted wireless networks offer substantial flexibility and rapid deployment capabilities, they still face significant limitations when tasked with serving a large number of mobile users, especially in environments characterized by dynamic mobility, uneven user distributions, and limited spectrum resources. Traditional static deployment strategies or heuristic rule-based control mechanisms often result in inefficient coverage, resource underutilization, and degraded quality of service (QoS). Moreover, as the scale and complexity of the system grow, these conventional approaches suffer from poor scalability and high computational overhead.

A key challenge lies in the real-time adaptation of UAV positions and resource allocations in response to unpredictable user mobility patterns. Without dynamic repositioning and intelligent decision-making, UAVs may hover over low-density areas, cause inter-UAV

interference, or fail to maintain consistent data rates for fast-moving users. Furthermore, the heterogeneous and stochastic nature of user demand in practical scenarios (e.g., disaster recovery, event coverage, or urban congestion) further exacerbates the control complexity.

Motivated by these challenges, this thesis aims to explore the application of Artificial Intelligence (AI)—specifically, reinforcement learning (RL) and clustering-based spatial partitioning—to enhance the autonomy and adaptability of UAV-assisted communication systems. The key motivation is to enable UAVs to learn environment-aware, data-driven policies that govern user association, flight trajectory, power control, and bandwidth scheduling, without relying on predefined rule sets.

Clustering serves as a critical first step to simplify the control problem. By grouping users based on spatial proximity and mobility trends, UAVs can be more efficiently assigned to localized regions, reducing system-wide interference and balancing load distribution. Coupled with this, reinforcement learning—particularly deep multi-agent variants such as MADDPG [3] and DQN—empowers UAVs to continuously adapt their strategies based on long-term performance feedback, mobility dynamics, and interference states.

The integration of user clustering and multi-agent reinforcement learning offers a promising path toward achieving scalable, robust, and context-aware UAV coordination. This hybrid learning framework is expected to yield substantial improvements in spectral efficiency, user fairness, and energy consumption, while maintaining real-time responsiveness in highly dynamic wireless environments.

Ultimately, the motivation behind this work is to bridge the gap between theoretical UAV control models and practical deployment challenges in next-generation wireless networks by harnessing the power of learning-based intelligence.

1.3 Problem Statement

The overarching challenge of this thesis is to develop and evaluate an AI-driven framework whereby unmanned aerial vehicles (UAVs), acting as aerial base stations, can effectively serve mobile ground users under varying scenarios. In particular, we address three progressive research stages:

1. **Single-UAV, static users:** A fixed-position UAV serves a set of fixed ground users. The key task is to jointly optimize power allocation and bandwidth assignment via a hybrid DDPG–DQN approach.

2. **Multi-UAV, static users:** Multiple UAVs cooperate to serve fixed ground users. We first partition users into spatial clusters using k-means, then employ MADDPG for inter-UAV power coordination across clusters.

3. **Multi-UAV, dynamic users:** Ground users exhibit mobility. UAVs must adapt their positions frame by frame to track user clusters, while a hybrid MADDPG–DQN scheme determines both UAV trajectories and real-time resource allocations (power levels and bandwidth).

1.4 Objectives

The specific objectives of this thesis are:

- **I:** Model a static-user, single-UAV environment and implement a DDPG+DQN algorithm for joint power and bandwidth allocation.
- **II:** Extend to a multi-UAV scenario by applying k-means clustering for spatial user grouping and MADDPG for coordinated UAV power control.
- **III:** Incorporate user mobility to generate per-frame optimal UAV placements, and design a hybrid MADDPG+DQN agent for simultaneous trajectory planning, power level selection, and bandwidth allocation.

1.5 Contributions

This thesis makes the following contributions:

- *I:* A hybrid DDPG–DQN framework for single-UAV, static-user scenarios, achieving optimized power and bandwidth allocation.
- *II:* A multi-UAV coordination method combining k-means user clustering with MADDPG-based power control, demonstrating scalable gains in coverage and throughput.
- *III:* A dynamic-user-aware system that generates frame-wise UAV trajectories and resource allocations via MADDPG+DQN, showing enhanced adaptability in data rate compared to baselines.

1.6 Thesis Organization

This thesis is organized into six chapters, each addressing a critical aspect of AI-enabled resource allocation in UAV-assisted wireless communications:

- **Chapter 1 – Introduction:** This chapter introduces the background, motivation, problem statement, objectives, and key contributions of the thesis. It outlines the challenges of UAV-enabled communication in dynamic environments and motivates the use of AI-based solutions.
- **Chapter 2 – Literature Review:** A comprehensive survey of recent advances in UAV-assisted wireless networks is presented. The chapter covers AI-driven communication architectures, reinforcement learning algorithms, and identifies key research gaps that this thesis aims to address.
- **Chapter 3 – Joint DRL-Based Utility Optimization:** This chapter proposes a hybrid Deep Reinforcement Learning (DRL) framework that integrates Deep Q-Network (DQN) and Deep Deterministic Policy Gradient (DDPG) to jointly optimize bandwidth allocation and power control for a single UAV serving static users. *This chapter is based on the work published in the 2024 IEEE WF-IoT.*
- **Chapter 4 – Multi-UAV Deployment and Power Allocation Optimization:** This chapter extends the framework to a multi-UAV setting with static users. K-Means clustering is employed for user association, and a Multi-Agent DDPG (MADDPG) algorithm is used to coordinate power allocation. *This chapter is based on the work accepted by the 2025 IEEE PIMRC.*
- **Chapter 5 – Mobility-Aware User Data Service Optimization:** This chapter considers dynamic user mobility and introduces a realistic mobility model (STEP). A combination of clustering and hybrid DRL (DQN + MADDPG) techniques is applied for UAV positioning, bandwidth allocation, and power control in dynamic environments. *This chapter contributes to a journal article currently under submission.*
- **Chapter 6 – Conclusion and Future Work:** This chapter summarizes the key findings, discusses the limitations of the proposed methods, and outlines promising directions for future research in intelligent, mobility-aware UAV communication systems.

Chapter 2

Literature Review

2.1 UAV-Assisted Wireless Networks

Recent literature emphasizes the critical role of architecture and deployment strategies in the design and performance of UAV-assisted wireless communication systems. According to Owaid et al. [1], UAVs can be deployed in various modes including peer-to-peer formations, swarm-based structures, or as mobile relays, depending on mission objectives and operational constraints. The authors detail three major deployment strategies: point-to-point connections among UAVs for coordinated missions, area coverage using Complete Coverage Path Planning (CCPP) algorithms for trajectory optimization, and swarm deployment for efficient collaborative operations in dynamic or large-scale environments. Each strategy offers distinct trade-offs between coverage efficiency, energy consumption, and coordination complexity.

The paper also categorizes communication architectures into three types: centralized (infrastructure-based), ad-hoc (structure-free), and hybrid. Centralized architectures rely on a ground control station (GCS) for decision-making and data aggregation but face vulnerability and range limitations. Ad-hoc architectures support decentralized operation, increasing robustness and adaptability at the expense of requiring more intelligent onboard processing and routing. Hybrid designs aim to balance these approaches by leveraging cellular networks and distributed intelligence to achieve scalable, long-range connectivity—particularly suitable for 5G/6G UAV networks. These architectural insights provide foundational knowledge for designing robust, adaptable UAV communication frameworks.

As UAV-assisted wireless networks grow more complex and dynamic, interference has become a critical challenge, particularly in dense deployments and multi-UAV environ-

ments. Murtadha et al. [4] provide a comprehensive classification of interference mitigation techniques across multiple layers of the communication stack. The review categorizes these techniques into five major domains: antenna-based methods, spectrum access coordination, power control, trajectory optimization, and artificial intelligence (AI)-enhanced methods.

Antenna-based methods such as beamforming and directional antennas help to spatially isolate signals, while dynamic spectrum access protocols reduce co-channel interference by coordinating frequency reuse. Power control strategies aim to minimize unnecessary signal leakage, especially in altitude-variable UAV networks. Trajectory optimization algorithms, often powered by reinforcement learning, play a dual role by reducing both physical path loss and interference footprints. AI-enhanced methods combine these approaches, using contextual information to dynamically predict and mitigate interference in real-time.

Importantly, the authors highlight that while many interference mitigation strategies exist, few are explicitly designed for the high-mobility, 3D operational environments of UAVs. Future research is encouraged to explore cross-layer and context-aware solutions that incorporate trajectory planning, adaptive transmission, and multi-agent cooperation simultaneously.

In time-sensitive UAV-assisted wireless sensor networks (UAV-WSNs), minimizing the age of information (AoI) is essential to ensuring data freshness and real-time responsiveness. Sun et al. [5] propose a two-stage optimization framework for reducing AoI in UAV-enabled data collection systems. The first stage uses an enhanced possibilistic fuzzy c-means (PFCM) clustering algorithm, optimized by the honey badger algorithm (HBA), to determine UAV hovering points and associated sensor nodes. This approach improves spatial efficiency in data acquisition and energy utilization.

In the second stage, the authors adopt the Proximal Policy Optimization 2 (PPO2) reinforcement learning algorithm to dynamically plan UAV trajectories across multiple data collection rounds. The integration of PPO2 allows the UAV to adapt its path in response to environmental changes, energy constraints, and real-time communication metrics. Simulation results show that the combined HBA-PFCM and PPO2 framework significantly outperforms conventional approaches such as ADC and generic AI-based planners, particularly in large-scale sensor deployments. The proposed model is especially effective in jointly minimizing AoI while respecting the UAV's limited energy budget, demonstrating its potential for scalable, intelligent aerial monitoring solutions.

Accurate channel modeling is fundamental to the design and reliability of UAV-assisted wireless sensor networks (WSNs). Xia et al. [6] propose a novel methodology for bounding path loss under various UAV flight trajectories by incorporating the concept of radio ir-

regularity—a key factor often neglected in traditional models. Unlike deterministic models which rely on fixed trajectories and assumptions, the proposed approach estimates path loss bounds based on directional irregularity and dynamic UAV orientation changes.

Their framework introduces the Degree of Irregularity (DOI) to model anisotropic variations in received signal strength (RSS) caused by propagation media and antenna orientation shifts. By employing a probabilistic Weibull-based compensation factor, the model successfully estimates upper and lower bounds for path loss rather than relying on single-point estimates. Experimental validation in diverse real-world environments shows that this method covers 94.7% of all measured data points across multiple scenarios, significantly outperforming traditional free space, two-ray, and log-normal models. This work provides a generalizable and robust path loss estimation tool suitable for dynamic UAV applications in WSNs, especially under non-line-of-sight (NLoS) and complex urban conditions.

Security threats such as reactive jamming are a growing concern in UAV-assisted wireless communication systems, particularly in adversarial environments. Zhang et al. [7] address this challenge by proposing a Collaborative Multi-Agent Jamming Deceiving (CMJD) method, designed to counter sophisticated multi-tone reactive jamming without relying on prior channel knowledge. The interaction between legitimate users (LUs) and a malicious jammer (MU) is modeled as a Stackelberg game, with LUs as leaders and the MU as the follower. To approximate the game equilibrium under uncertainty, the authors adopt a Multi-Agent Advantage Actor-Critic (MAA2C) reinforcement learning algorithm within a centralized training and decentralized execution (CTDE) framework.

The CMJD method allows each LU to independently select both communication and deception bands, encouraging the MU to misallocate jamming power and improving the signal-to-interference ratio (SIR). The UAV evaluates system-wide communication quality and updates LU policies accordingly. Simulation results demonstrate that CMJD outperforms both optimal Stackelberg equilibria (with full information) and independent multi-agent reinforcement learning (IMARL) baselines, achieving superior SIR and convergence speed. This approach is particularly valuable for securing dynamic UAV networks where centralized control and real-time channel feedback are limited.

Sheshashayee et al. [8] present an in-depth evaluation of UAV-assisted data collection strategies in wireless sensor networks (WSNs) employing wake-up radio (WuR) technology. The authors compare a naïve collection strategy—with fixed UAV paths—to an adaptive strategy that optimizes flight paths using metadata-based node localization. Physical and simulation experiments reveal that WuR-based systems drastically outperform duty-cycled WSNs across key metrics such as awake time, latency, and network lifetime. Specifically, WuR reduces node awake time by over 98%, enabling network lifetimes several orders of

magnitude longer than duty-cycled alternatives. Notably, even the naïve collection strategy achieves substantial energy savings when WuR is employed, suggesting that complex route optimizations may be unnecessary in WuR-dominated deployments. These findings validate the viability of lightweight and energy-efficient UAV-WSN architectures for long-term environmental sensing, agriculture, and IoT applications where node replacement is impractical.

Unmanned aerial vehicles (UAVs) have emerged as a versatile component in modern wireless communication networks, offering flexible deployment, high mobility, and favorable line-of-sight (LoS) conditions. Their ability to act as aerial base stations or mobile relays enables rapid response in challenging environments such as natural disasters, rural regions, and high-density events. A critical consideration in UAV-assisted wireless sensor networks (WSNs) is maintaining the freshness of collected data, especially in dynamic settings where sensor node (SN) positions may shift due to external factors like earthquakes or landslides. Traditional deep reinforcement learning (DRL) approaches often struggle to adapt quickly when such environmental changes occur. To overcome this limitation, a meta-learning DRL framework has been proposed that enables UAVs to optimize their flight trajectories by minimizing both the average and maximum age of information (AoI). By modeling the UAV trajectory planning problem as a Markov decision process with nonuniform time steps, the framework leverages experience from past tasks to achieve rapid convergence and superior adaptability in new, unseen scenarios. Simulation results show that this approach significantly outperforms standard DRL baselines in terms of responsiveness and data timeliness, making it highly suitable for real-time data collection in disaster-prone WSNs [2].

The integration of federated learning (FL) into UAV-assisted wireless networks has emerged as a solution for enabling distributed machine learning tasks while addressing challenges such as communication overhead, device heterogeneity, and data privacy. A recent study introduces a joint client selection and model compression framework (csmcFL) that enhances the efficiency of FL in UAV-based systems by optimizing UAV deployment and selectively compressing local models based on client channel quality [9]. The scheme applies singular value decomposition (SVD) to compress fully connected layers in CNN models and dynamically chooses which ground terminals (GTs) transmit compressed or full models. Simulation results demonstrate significant reductions in communication time and training latency, especially under both IID and Non-IID data distributions, while maintaining high model accuracy. This highlights the feasibility of deploying FL in UAV networks where resource constraints and dynamic channel conditions are prevalent.

Energy consumption is a critical constraint in UAV-assisted wireless networks due to limited onboard power resources. To address this, recent research proposes reinforcement

learning (RL) algorithms—Q-learning and deep Q-networks (DQN)—for optimizing energy efficiency over hybrid communication protocols, including BLE, LTE, Wi-Fi, and LoRa [10]. The approach dynamically assigns the most energy-efficient communication link between UAVs and ground stations (GS), considering both free space (FS) and free space multi-path (FSMP) path loss models. The RL agents learn to minimize the total network energy consumption by selecting communication technologies based on link distance and energy models, outperforming traditional rule-based and random hybrid schemes. Analytical models and simulations demonstrate that the RL-based hybrid networks achieve significantly lower energy usage and latency, particularly when employing the FSMP model, highlighting their suitability for adaptive and sustainable UAV communication infrastructures.

With the emergence of quantum computing, quantum edge computing devices (QECDs) can be integrated into UAV-assisted wireless networks to significantly enhance distributed intelligence and computational capacity. Quantum Federated Learning (QFL), which combines federated learning with quantum computing, enables decentralized model training while preserving data privacy and increasing computational efficiency. A recent study introduces a trust-enhanced game-theoretic framework to secure QFL in UAV-assisted networks by mitigating issues like malicious behaviors and selfish participation of QECDs [11]. The scheme features a Bayesian trust assessment mechanism to filter out untrustworthy devices and a Stackelberg game-based incentive model, solved via deep Q-learning, to optimize training participation and payment strategies. Simulation results show that this framework improves model accuracy, convergence speed, and resource utilization, demonstrating the practical viability of QFL for highly dynamic and sensitive UAV-assisted environments.

To systematically identify the limitations and research gaps in current UAV-assisted wireless communication literature, we conduct a focused comparative analysis of recent state-of-the-art works discussed. Table 2.1 summarizes each study based on four critical dimensions: the extent of UAV application, the incorporation of adaptive or AI-driven mechanisms, attention to energy efficiency, and whether the proposed models have been validated through simulations or real-world testing. This structured gap analysis highlights that while UAVs are widely adopted across various scenarios, there is still a significant lack of comprehensive solutions that jointly optimize adaptability, energy constraints, and deployment realism. The table reveals opportunities for integrated research efforts that bridge these isolated advancements into robust, energy-aware, and practically validated UAV communication systems.

Table 2.1: Gap Analysis of UAV-Assisted Wireless Networks Literature

Paper	Deployment Strategy	Network Architecture	Interference Management	Energy Efficiency
[1]	✓	✓	✗	✓
[4]	✓	✗	✓	✓
[5]	✓	✓	✗	✓
[6]	✗	✗	✓	✗
[7]	✗	✓	✓	✗
[8]	✓	✓	✗	✓
[2]	✗	✓	✗	✗
[9]	✓	✓	✗	✓
[10]	✗	✓	✗	✓
[11]	✓	✓	✓	✗
Our work in Chap.5	✓	✓	✓	✓

2.2 AI in Wireless Communication

Recent advancements in optical wireless communication (OWC) also demonstrate promising integration potential with UAV platforms, particularly in scenarios requiring high-sensitivity and low-power signal reception. He and Chen [12] compare artificial neural networks (ANN) and radial basis function neural networks (RBFNN) for signal demodulation in photon-counting-based OWC systems using silicon photomultiplier (SiPM) sensors. These sensors, known for their high sensitivity and nonlinear behavior due to microcell recovery time, pose challenges such as intersymbol interference (ISI) under high data rates. The authors show that both ANN and RBFNN architectures significantly outperform traditional demodulation schemes in such nonlinear regimes, reducing the bit error rate (BER) across a wide range of irradiance levels. While not UAV-specific, this research underscores the broader applicability of AI-enhanced demodulation techniques in non-linear, high-noise environments—conditions often encountered in aerial communication platforms. Integrating such robust signal processing models on UAVs equipped with OWC systems could further enhance their communication reliability and data throughput, especially under adverse conditions like urban clutter or disaster recovery operations. Another critical perspective on UAV-assisted wireless networks is provided by the broader landscape of wireless communication advancements. As discussed in Dandekar et al. [13], next-generation wireless systems—including 5G and beyond—are poised to transform connectivity through ultra-reliable low latency communication (URLLC) and enhanced mobile broadband (eMBB). These capabilities are particularly relevant for UAV applications that require low-latency, high-throughput links for mission-critical operations such as autonomous navigation and real-time video surveillance. Furthermore, the review emphasizes the significance of in-

tegrating artificial intelligence (AI) and edge computing into wireless networks, enabling intelligent control and reduced latency at the network's edge—capabilities that align well with UAV operation requirements. The paper also highlights the importance of mesh and ad hoc networking topologies, which are often employed in UAV swarms and emergency deployments where centralized infrastructure is absent. Collectively, these insights underscore how UAV-assisted wireless systems can leverage cutting-edge wireless technologies to improve responsiveness, reliability, and scalability in both civilian and tactical contexts.

The evolution of UAV-assisted wireless networks is further enriched by the paradigm shift toward semantic-aware communication, as discussed in the work by Shi et al. [14]. Traditional wireless systems operate at the bit level, optimizing transmission and decoding accuracy. However, for UAVs operating under dynamic, resource-constrained environments, transmitting semantic information—such as mission-relevant content or event-specific alerts—can be significantly more efficient. The concept of semantic communication prioritizes the meaning behind data rather than exact bit reconstruction, allowing UAVs to communicate more purposefully and adaptively. AI-driven techniques, particularly deep learning models, enable semantic extraction and prioritization from raw data streams like video, sensor data, or telemetry. This is especially advantageous in UAV applications requiring real-time decision-making with limited bandwidth. Semantic-aware UAV communication can reduce latency, improve resilience, and optimize energy usage, representing a major advancement in the fusion of AI with aerial network intelligence.

In UAV-assisted wireless networks, accurate and low-latency channel estimation is critical for ensuring reliable communication, especially under highly dynamic conditions. The study by Kashyap et al. [15] introduces an innovative AI-driven channel estimation framework that integrates geo-spatial data and modified Demodulation Reference Signals (DMRS). Their approach employs a deep learning model—specifically, a hybrid CNN-RNN architecture—that adapts to various urban, rural, and semi-urban scenarios using location-specific environmental features. This design allows for robust prediction of channel characteristics even in the presence of severe multipath fading and Doppler shifts, issues frequently encountered in UAV-based deployments. By treating DMRS configurations as learnable parameters and incorporating environmental context through geo-spatial inputs, their model significantly reduces channel estimation error compared to traditional methods. These capabilities are particularly relevant for UAV communication systems operating in heterogeneous landscapes, where conventional estimation techniques struggle to maintain accuracy. The framework also shows compatibility with 6G use cases, suggesting its potential for future UAV network architectures.

To support the high-capacity demands of UAV-assisted communication, especially in next-generation wireless networks like 5G and 6G, AI-based optimization techniques have

gained prominence. Gurupandi and Premkumar [16] investigate the role of artificial neural networks (ANNs) in estimating and enhancing the channel capacity of massive MIMO systems—an architecture highly relevant to UAV communication platforms. Their model uses AI to learn the relationship between signal-to-noise ratio (SNR), bandwidth, and channel coefficients, producing an intelligent estimation of communication capacity. The ANN-based model, trained using backpropagation and Levenberg–Marquardt algorithms, shows a strong correlation between predicted and simulated capacity values, confirming the potential of AI in supporting UAV networks operating under dynamic SNR and mobility conditions. This AI-driven approach not only offers improved spectral efficiency but also paves the way for adaptive UAV communication systems that can optimize throughput in real time.

As UAV networks increasingly rely on AI-driven models, the quality of wireless datasets becomes paramount to achieving robust and generalizable performance. Tang et al. [17] introduce a data quality assessment (DQA) framework that emphasizes the importance of dataset similarity and diversity in AI-enabled wireless communication systems. Their work is particularly relevant to UAV-assisted networks, where machine learning models trained on air-interface data (e.g., CSI, SINR, RSRP) must generalize across heterogeneous, dynamic environments. The framework proposes methods for quantifying similarity between synthetic and real-world datasets, as well as measuring diversity to ensure the trained models can perform across unseen scenarios. Using the CsiNet model as a case study, the authors show that selecting training datasets based on similarity to the target scenario significantly improves performance, even with reduced data volumes. This insight can be pivotal for UAV systems where on-board learning and data collection are resource-constrained, and targeted training using diverse yet statistically similar datasets can maximize model efficacy and deployment efficiency.

The emergence of generative AI in the context of 6G wireless networks presents transformative opportunities for UAV-assisted systems. As discussed by Zhang et al. [18], generative AI models—including large language models (LLMs) and diffusion models—are set to redefine the landscape of wireless intelligence by enabling data generation, protocol synthesis, and autonomous control. In UAV networks, these models can be used for predictive channel modeling, real-time decision-making, and cross-layer optimization. Moreover, generative AI fosters semantic communication, where the transmission focuses on relevant content rather than raw bits, aligning well with the data and energy constraints typical in UAV operations. The paper further highlights challenges such as computational complexity, training-data scarcity, and model security, which are critical when deploying generative models on resource-limited UAV platforms. Integrating generative AI into UAV-assisted networks could therefore enable smarter, context-aware communications and collaborative

behaviors, paving the way for truly intelligent aerial systems in future 6G infrastructures.

Future UAV-assisted wireless systems will benefit greatly from advances in AI-driven multi-band and multi-connectivity frameworks, especially as sub-THz communications are considered for 6G deployments. Choi et al. [19] investigate the challenges of achieving Tbps wireless communication by leveraging AI for managing the complexity of sub-THz band impairments, channel prediction, and seamless radio access. Their study emphasizes the importance of ultra-massive MIMO, beamforming, and distributed multi-point transmission to overcome the limitations of sub-THz propagation such as high path loss and narrow coverage. Particularly for UAVs operating at high altitudes and with frequent mobility-induced blockages, such architectures—enhanced with AI-based prediction of radio link failure (RLF)—can ensure reliable connectivity and low-latency control. By integrating deep learning models to anticipate and mitigate RLF events, UAV networks can dynamically switch between links in a multi-connectivity setup, thereby maintaining uninterrupted service in diverse operating environments. This capability is essential for real-time applications such as aerial surveillance, disaster response, and industrial automation in the 6G era.

Recent developments in reconfigurable intelligent surfaces (RIS) and holographic beamforming offer new possibilities for optimizing UAV-assisted wireless communication. Gomathi et al. [20] propose an integrated 6G architecture combining Edge AI, AI-controlled RIS, and Low Earth Orbit (LEO) satellites to deliver high-speed, low-latency, and energy-efficient wireless connectivity. For UAVs operating in diverse and unpredictable environments, AI-enhanced RIS provides dynamic control over signal propagation, improving spectral and energy efficiency while minimizing interference. Moreover, the use of Edge AI facilitates real-time processing of UAV data, enabling predictive resource allocation and adaptive beamforming strategies. The incorporation of holographic beamforming allows precise wavefront shaping, which enhances signal robustness and directionality—critical for maintaining reliable UAV-to-ground links during high-mobility operations. This architecture's ability to integrate LEO satellites also ensures global UAV connectivity, especially in remote or infrastructure-sparse regions. These innovations signal a significant leap toward resilient, intelligent, and flexible aerial communication systems in 6G networks.

As UAVs integrate more AI functionality for real-time decision-making, the need for efficient onboard AI computation becomes paramount. Cai et al. [21] propose a novel AI accelerator architecture tailored for next-generation wireless systems, addressing the computation bottlenecks in data-intensive AI tasks. Their design separates computing and control using a dataflow-driven paradigm and a specialized compiler-emulator stack, achieving ultra-low-latency inference suitable for UAVs with tight energy and time constraints. This hardware architecture supports tensor and vector operations optimized for common deep

learning tasks such as channel estimation, feedback compression, and end-to-end communication—all essential for intelligent UAV network operations. Simulation results show sub-millisecond latency even for complex transformer-based models like SwinCFNet, making this accelerator architecture viable for real-time deployment in UAVs. The development of such domain-specific accelerators ensures that UAV-assisted wireless networks can scale their AI capabilities without compromising on efficiency, latency, or power consumption.

Effective resource allocation is critical in UAV-assisted wireless networks, especially when operating in dynamic and heterogeneous environments. Zhang et al. [22] propose a comprehensive AI-based resource allocation and optimization model that leverages deep reinforcement learning (DRL), Q-learning, Proximal Policy Optimization (PPO), and Long Short-Term Memory (LSTM) networks. Their framework employs a hierarchical architecture with real-time monitoring, predictive analytics, and closed-loop feedback control, making it particularly well-suited to the adaptive needs of UAV communication systems. The model dynamically allocates wireless resources such as spectrum and transmission power by analyzing network state and user behavior, leading to significant improvements in throughput (by over 20%), reduced latency (by up to 25%), and lower energy consumption. These capabilities align with the operational requirements of UAV networks, where energy efficiency, low-latency, and context-aware control are essential. By applying DRL and LSTM in tandem, the framework demonstrates robust performance in scenarios with time-varying traffic and mobility—conditions commonly encountered in UAV deployments.

The fusion of Tensor Space-Time (TST) coding with artificial intelligence presents a promising avenue for enhancing UAV-assisted wireless communication systems. Kapula et al. [23] review how AI-enhanced TST coding can improve spectral efficiency, reliability, and adaptability by exploiting space-time diversity across multiple-input multiple-output (MIMO) channels. By modeling transmitted signals as multidimensional tensors and applying deep learning techniques such as CNNs, LSTMs, and reinforcement learning, wireless systems can dynamically optimize code design, beamforming, and power allocation in response to fluctuating channel conditions. This has direct applications in UAV networks, where mobility introduces spatiotemporal channel variability. The paper also highlights the use of AI to reduce TST decoding complexity and improve real-time responsiveness, which is vital for delay-sensitive UAV tasks such as swarm coordination and emergency response. The integration of TST coding and AI not only boosts transmission robustness but also provides a flexible framework adaptable to various UAV mission profiles in 5G and beyond networks.

The integration of generative AI and semantic communication (SemCom) is emerging as a powerful paradigm for enhancing bandwidth efficiency and intelligence in UAV-assisted wireless networks. Zhou et al. [24] propose a framework that combines SemCom and gener-

ative AI (GAI) to optimize the construction and distribution of radio maps—a critical tool for spectrum awareness and resource allocation in dynamic wireless environments. Their approach leverages multi-modal semantic information, extracted from textual and visual inputs, to compress and transmit only the most essential features over the wireless channel. This significantly reduces communication overhead while maintaining high fidelity in map reconstruction. Although originally targeted at smart city infrastructure, this method is highly applicable to UAV networks, where low-latency, energy-efficient semantic transmission can support real-time environmental awareness, node coordination, and adaptive communication planning. The use of diffusion models, particularly denoising diffusion implicit models (DDIM), enables high-quality radio map generation at the user equipment (UE) level despite channel constraints. This semantic-aware radio mapping approach enhances the scalability, responsiveness, and reliability of UAV-based communication systems in next-generation wireless networks.

Table 2.2 provides a comparative overview of the surveyed literature across key criteria relevant to UAV-assisted wireless communication systems. While a wide range of contributions have been made in areas such as AI-enhanced signal processing, semantic communication, channel estimation, and resource allocation, the table highlights several consistent research gaps. Notably, most works do not explicitly target UAV-specific scenarios, lack real-time adaptive capabilities, or fail to address the hardware and deployment constraints faced in aerial networks. Furthermore, while hybrid AI models (e.g., combining deep learning and reinforcement learning) are beginning to emerge, experimental validation and end-to-end system integration remain underexplored. This comparative analysis underlines the necessity for a unified, real-time, and mobility-aware AI framework tailored to the unique operational environment of UAV-assisted networks, which this thesis aims to address.

2.3 Reinforcement Learning for UAVs

Reinforcement Learning (RL) has demonstrated substantial promise in enhancing UAV autonomy, particularly in dynamic environments requiring real-time decision-making and adaptive control. In [25], an adaptive access control algorithm based on deep RL is proposed for UAV front-end sensing systems. The authors design a Deep Q-Network (DQN) framework that integrates UAV state parameters—such as position, velocity, and environmental context—into a fully connected neural network to generate optimal control actions. By leveraging a reward function that balances sensing accuracy and control performance, the UAV dynamically adjusts throttle, pitch, and yaw to complete complex flight tasks.

Table 2.2: Metric-Based Gap Analysis of AI-Enabled UAV Wireless Communication

Paper	**Real-Time Adaptive**	**Hybrid AI Models**
[12]	✗	✗
[13]	✗	✗
[14]	✗	✗
[15]	✗	✓
[16]	✗	✗
[17]	✗	✗
[18]	✗	✓
[19]	✓	✗
[20]	✓	✗
[21]	✗	✗
[22]	✓	✗
[23]	✗	✓
[24]	✓	✗
Our work in Chap.5	✓	✓

Experimental validation shows that the RL-based controller significantly outperforms traditional algorithms in trajectory tracking accuracy, flight stability, and task completion time, demonstrating superior adaptability in varied flight scenarios.

An enhanced deep reinforcement learning (DRL) approach for autonomous UAV visual navigation is proposed in [26], addressing limitations of traditional DQN-based methods in dynamic environments. The method introduces a two-stage learning process: a reinforced training stage based on Deep Q-Networks (DQN) and a self-supervised fine-tuning stage driven by contrastive loss. This hybrid strategy significantly accelerates convergence by improving scene encoding from UAV-captured depth images. Furthermore, a ResNet50-based obstacle detection model is integrated to mitigate UAV collisions with both static and moving obstacles such as pedestrians. Experimental results in a simulated outdoor environment demonstrate that the proposed self-supervised DQN outperforms baseline DQN, Double DQN, and Dueling DQN models in terms of navigation distance and collision avoidance, highlighting its efficacy for UAV deployment in dynamic and cluttered environments.

In the context of multi-UAV collaborative missions, reinforcement learning has shown potential to address coordination and formation control challenges. Jiang et al. [3] propose an ATT-MATD3 algorithm that integrates a multi-attention mechanism with a dual-delay deterministic policy gradient framework to enhance collaborative encirclement strategies among UAV swarms. The proposed architecture leverages attention-augmented critic networks to improve decision-making by focusing on relevant teammate and environmental features. In adversarial simulations involving an evasive target, UAVs trained with ATT-MATD3 demonstrated superior convergence and coordination compared to standard MATD3 and MADDPG baselines. This attention-enhanced multi-agent framework not only ensures stable learning but also improves encirclement formation maintenance and capture success rate, showcasing its applicability in high-stakes autonomous multi-UAV systems.

Multi-agent reinforcement learning (MARL) has been successfully applied to cooperative UAV decision-making tasks, particularly in adversarial environments where reward signals are sparse. Huo and Li [27] proposed a cooperative maneuver decision-making strategy where two UAVs must collaborate to overcome a stronger opponent. To address suboptimal policy learning due to reward sparsity and sample imbalance, they introduced a novel reward augmentation method that maximizes the *weighted entropy of reward* (MWRE). This approach encourages exploration by penalizing frequently encountered rewards, thus incentivizing agents to discover more diverse and effective strategies. Experimental results in a continuous action space confirm that MAPPO with MWRE significantly outperforms standard MAPPO in both learning speed and final performance, enabling UAV agents to develop coordinated tactics such as flanking and luring maneuvers.

In UAV-assisted mobile edge computing (MEC) scenarios, decentralized and privacy-preserving navigation policies are essential to cope with dynamic environments and heterogeneous mission demands. To address this, Wang et al. [28] propose a heterogeneous federated reinforcement learning (HFRL) framework for decentralized UAV navigation. The method enables individual UAVs to train local navigation policies using proximal policy optimization (PPO), while periodically aggregating high-level policy parameters via a cloud server without sharing raw data. To handle heterogeneity in computation and communication capabilities among UAVs, the authors introduce a dynamic weighting mechanism during policy aggregation. Simulation results demonstrate that HFRL not only accelerates convergence but also improves task completion rates and energy efficiency compared to traditional centralized and homogeneous federated RL baselines. This approach presents a scalable and secure solution for collaborative UAV navigation in MEC systems.

Improving generalization in deep reinforcement learning (DRL) for multi-UAV collision avoidance remains a key challenge due to spurious correlations in training environments. Han et al. [29] propose a novel approach that integrates Causal Representation Learning (CRL) with DRL to address this limitation. Their method uses causal intervention to extract invariant features from depth images by manipulating obstacle shapes while preserving semantic content. These learned causal representations, combined with speed and goal data, are used as input to a Soft Actor-Critic (SAC) policy network. By introducing invariance and decorrelation losses, the CRL framework effectively disentangles task-relevant features, enhancing policy robustness across unseen environments. Experimental results in the AirSim simulator show substantial improvements in success rate, SPL, and velocity compared to state-of-the-art methods such as SAC+RAE, AutoAugment, and DrAC. This highlights CRL's potential to significantly improve DRL generalization in safety-critical UAV applications.

Deep reinforcement learning (DRL) methods have shown promise in UAV applications, yet transferring learned policies from simulation to real-world environments remains challenging. Xu et al. [30] address this issue by developing a 3DQN (dueling-double-deep Q-network) algorithm for autonomous UAV landing, and crucially validate it through real-world flight tests. Their approach integrates a discrete action space in the X-Y plane and uses downward-looking RGB imagery as input to the DRL model. To bridge the sim-to-real gap, they introduce a database-driven pretraining phase using real flight data, accelerating convergence and enhancing generalization. The UAV executes landing using DRL-generated control outputs in "Offboard" mode with safety fallback mechanisms. Experiments with varying initial conditions show consistent and accurate landings, with oscillatory adjustments near touchdown due to limited visual field. This work exemplifies a successful physical deployment of DRL in safety-critical UAV operations.

Deep reinforcement learning (DRL) has been explored for optimizing UAV scheduling in mobile IoT networks, where energy efficiency and timely data collection from mobile IoT devices (IoTDs) are critical. Singh and Hegde [31] propose a DRL-based UAV scheduling framework using the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm to maximize Global Energy Efficiency (GEE) in UAV-assisted IoMT networks. The model captures dynamic IoTD data levels and mobility patterns, reformulating the scheduling as a Markov Decision Process. The TD3-based approach mitigates Q-value overestimation and efficiently learns optimal hovering locations and durations for UAVs under constraints such as AoI, onboard storage, and mobility. Simulation results confirm superior cumulative rewards, improved GEE, and reduced AoI compared to baseline methods like DDPG and ISAC, thereby demonstrating the robustness and applicability of the proposed method for autonomous and energy-aware UAV operations in dynamic IoT environments.

To address the complexity of heterogeneous UAV swarm confrontation, Su et al. [32] propose a novel reinforcement learning framework that incorporates a heterogeneous policy network with local parameter sharing. Within the Multi-Agent Proximal Policy Optimization (MAPPO) paradigm, the method enables different types of UAVs—reconnaissance and attack—to use specialized policy networks tailored to their distinct state and action spaces. The approach also introduces a collaborative state space reconstruction mechanism that integrates relative positional and directional information to promote coordinated strategies such as encirclement, decoying, and maneuvering. Simulation results in adversarial swarm environments show that the proposed method significantly outperforms baseline algorithms like IPPO, MAPPO, and MASAC in terms of reward and win rate, particularly in imbalanced scenarios (e.g., 6 vs. 7 agents). This highlights its robustness and strategic effectiveness in real-world multi-agent confrontation tasks.

Ensuring timely and adaptive data collection in dynamic wireless sensor networks (WSNs) is a critical challenge for UAV-assisted systems, particularly in disaster scenarios where sensor node (SN) positions may shift. Xiao et al. [2] propose a Meta-Learning Deep Reinforcement Learning (MLDRL) framework to optimize UAV trajectories for minimizing both average and maximum Age of Information (AoI). Their approach segments the environment into square sub-areas, enabling the UAV to visit only sub-area centers to collect data efficiently. The core idea leverages model-agnostic meta-learning (MAML) to enable fast policy adaptation when the SN distribution changes. Simulation results demonstrate that MLDRL achieves faster convergence and better data freshness compared to baseline algorithms such as PPO, A2C, and DQN (with and without transfer learning). This highlights MLDRL's effectiveness for real-time mission planning under uncertain and evolving environmental conditions.

To solve the NP-hard problem of multi-UAV trajectory planning in unknown and

obstacle-rich environments, Xing et al. [33] propose a deep reinforcement learning framework called PF-LSTM-MATD3. This method combines potential field-based dense rewards with a long short-term memory (LSTM) enhanced multi-agent twin delayed deep deterministic policy gradient (MATD3) network. The proposed hierarchical architecture integrates three layers—adaptive formation, trajectory planning, and action execution—and supports dynamic formation transformations to safely traverse narrow passages. The LSTM module enables UAV agents to retain historical state-action sequences, improving learning stability and convergence speed. Simulation results demonstrate that PF-LSTM-MATD3 outperforms both MADDPG and standard MATD3 in trajectory optimality, policy learning efficiency, and adaptability, especially in constrained environments where formation transitions are critical for obstacle avoidance.

In the realm of UAV-enabled secure communication, Liu et al. [34] propose a multi-agent deep reinforcement learning (MADDPG) framework to jointly optimize trajectory planning, transmission power control, and user scheduling for multiple UAVs serving as aerial base stations. The goal is to maximize the sum secrecy rate while accounting for interference, eavesdropping threats, and UAV collision constraints. The system is modeled as a Markov decision process, and each UAV agent learns a deterministic policy using actor-critic networks with centralized training and decentralized execution. The MADDPG-based solution significantly outperforms DDPG in both convergence and secrecy rate across multiple scenarios, demonstrating effective coordination among UAVs for secure multi-user communication in dynamic environments.

In emergency communication scenarios, organizing UAVs as mobile access points is essential for network recovery. Xu et al. [35] propose a novel multi-agent reinforcement learning (MARL) framework enhanced by large language model (LLM) guidance to optimize UAV-based relay networks. The approach introduces a grouping strategy and selective parameter sharing mechanism to improve scalability and collaboration among UAVs in large-scale environments. A dual reward system is employed, combining communication performance and intrinsic motivation derived from LLM-inferred action guidance. UAV agents receive intrinsic rewards based on the similarity between LLM predictions and actual outcomes, effectively accelerating policy optimization. Experimental results show that this method surpasses MAPPO and existing graph-based alternatives in terms of device connectivity and average data rate, demonstrating its promise for intelligent, scalable, and context-aware UAV networking in disaster recovery.

Kulkarni and Patil [36] provide a comprehensive review of reinforcement learning (RL) methods applied to autonomous systems, with a focus on UAV navigation. The paper discusses the progression from traditional Deep Q-Networks (DQN) to advanced techniques like Double DQN, Dueling DQN, Actor-Critic, and Twin Delayed DDPG (TD3). Each

technique addresses specific limitations in dynamic and high-dimensional environments, such as Q-value overestimation, unstable training, or poor generalization. Notably, the integration of self-supervised learning into DQN frameworks helps reduce collisions and improve path optimization by extracting better feature representations from UAV camera inputs. These enhancements collectively contribute to improved convergence rates and robustness in tasks such as path planning, obstacle avoidance, and target tracking. The review underscores the potential of combining reinforcement learning with auxiliary learning strategies to meet the challenges of real-world UAV autonomy.

Kim et al. [37] survey reinforcement learning-based handover algorithms for UAVs operating in cellular networks, addressing challenges such as frequent disconnections and ping-pong effects due to UAV altitude and velocity. The paper compares three RL-based methods—Proximal Policy Optimization (PPO), Dueling Double Deep Q-Network (D3QN), and Q-learning—against traditional handover techniques. PPO employs a value network that considers UAV trajectory, altitude, and signal strength to optimize handover timing. D3QN decouples state evaluation from action selection to stabilize learning and reduce Q-value overestimation, leading to improved uplink interference and delay metrics. The enhanced Q-learning model incorporates SINR and BS-UE distance to refine action spaces. Across various simulations, these methods demonstrate significant reductions in handover ratio (up to 95%), uplink interference, and delay, showcasing the effectiveness of RL in maintaining stable UAV connectivity in cellular systems.

Task offloading in UAV-assisted edge computing environments demands efficient strategies to minimize energy consumption while maintaining quality of service. Dai et al. [38] present a multi-agent reinforcement learning framework based on Double Deep Q-Networks (DDQN) to optimize dynamic task offloading decisions. In this approach, UAVs equipped with edge servers provide computational resources to ground users who offload delay-sensitive tasks. The system is modeled as a Markov Decision Process (MDP), where agents select offloading targets to balance energy usage and task deadlines. The reward function integrates energy efficiency with task completion penalties to guide learning. Experimental results demonstrate that the proposed method reduces energy consumption and achieves stable convergence compared to local computing and conventional schemes, validating its potential for intelligent and adaptive task scheduling in UAV-enabled mobile edge computing systems.

Ghomri et al. [39] investigate the use of deep reinforcement learning (DRL) algorithms for optimizing the 3D placement of UAVs in NOMA-enabled wireless networks. The study compares three DRL methods—Deep Q-Network (DQN), Advantage Actor-Critic (A2C), and Proximal Policy Optimization (PPO)—to improve UAV positioning for maximum sum-rate performance. The placement problem is modeled as a Markov Decision Process, where

the agent's state includes UAV position and user distances, and actions involve discrete 3D movements. Simulation results reveal that PPO consistently achieves the highest average sum-rate and most stable convergence, outperforming both fixed-placement baselines and other DRL algorithms. The findings highlight the effectiveness of policy-gradient-based methods and underscore the importance of hyperparameter tuning, neural architecture design, and reward shaping for UAV 3D deployment in dynamic communication environments.

To better understand the distribution and capabilities of reinforcement learning algorithms across UAV-related applications, we conducted a comparative analysis of recent literature. Table 2.3 presents an algorithm-oriented gap analysis, highlighting the presence or absence of core DRL methods—such as DQN, PPO, A2C, and DDPG-based approaches—alongside whether each work proposes a novel framework. The categorization reflects the dominant algorithmic strategies and research focuses, ranging from sensing and navigation to trajectory optimization and secure communications. This analysis reveals a growing preference for actor-critic and hybrid models, while also uncovering underexplored areas such as PPO integration in edge computing and hardware-tested deployments for secure UAV communication frameworks.

2.4 Limitations of Existing Works

Despite the growing body of research leveraging deep reinforcement learning (DRL) for UAV-enabled wireless communication systems, several limitations persist in the current literature. These gaps hinder the full realization of intelligent, adaptive, and scalable UAV networks, especially under dynamic and real-world deployment scenarios.

First, **limited real-world deployment and validation** remains a prominent shortcoming. A majority of studies conduct evaluations solely in simulated environments such as AirSim, Unity 3D, or custom MATLAB platforms. While these simulations provide control and reproducibility, they often fail to capture the uncertainty, noise, and latency of real-world conditions. Only a few works, such as [30], have validated DRL-based control policies through physical UAV testbeds, highlighting a pressing need for hardware-in-the-loop evaluations.

Second, most existing works rely on **homogeneous agent models** and **simplified state/action representations**, which restrict the applicability of learned policies in diverse mission settings. For instance, many multi-agent systems assume identical UAV capabilities, ignoring hardware-level heterogeneity in computation, energy capacity, and

Table 2.3: Algorithm-Oriented Gap Analysis of DRL in UAV Applications

Paper	**UAV Function**	**DQN**	**PPO**	**A2C**	**DDPG**	**MADDPG**
[25]	Front-End Sensing	✓	✗	✗	✗	✗
[26]	Visual Navigation	✓	✗	✗	✗	✗
[3]	Encirclement	✗	✗	✗	✗	✓
[27]	Maneuvering	✗	✓	✗	✗	✗
[29]	Collision Avoidance	✓	✗	✗	✗	✗
[30]	Landing	✓	✗	✗	✗	✗
[31]	GEE Optimization	✓	✗	✗	✓	✗
[32]	Swarm Confrontation	✗	✓	✓	✗	✗
[2]	AoI-Driven Sensing	✓	✓	✓	✗	✗
[33]	Formation Trajectory	✓	✗	✗	✓	✓
[34]	Secure Comms	✓	✗	✗	✗	✓
[37]	Handover	✓	✗	✗	✗	✗
[38]	Edge Offloading	✓	✗	✗	✗	✗
[39]	3D Placement	✓	✓	✓	✗	✗
Our work in Chap.3 4 5	Data Service	✓	✗	✗	✓	✓

communication modules. This leads to reduced performance in cooperative tasks like swarm confrontation [32] and decentralized navigation [28].

Third, the majority of algorithms operate under **fixed reward functions** without incorporating multi-objective optimization or domain-specific metrics such as Age of Information (AoI), secrecy rate, or energy fairness. This is evident in trajectory planning and formation control applications [33], where static reward formulations may lead to brittle policies that fail under changing objectives.

Furthermore, **semantic-awareness and context-driven decision-making** are generally absent. Except for limited studies incorporating contrastive or causal representation learning [29], most methods depend on raw or low-level sensory data, without leveraging high-level semantics, scene understanding, or language-based priors that could enhance adaptability in dynamic environments.

Lastly, **algorithmic redundancy and isolated experimentation** are common. Several studies independently implement variations of DQN or PPO without benchmarking against newer algorithms (e.g., SAC, TRPO, or federated DRL approaches). This redundancy leads to limited generalizability and comparability, obstructing the establishment of standardized benchmarks for UAV autonomy.

These limitations collectively underscore the need for unified frameworks that integrate advanced DRL architectures, real-world deployment feedback, semantic modeling, and multi-objective optimization tailored to UAV-specific constraints and missions.

2.5 Summary

This chapter explored the application of deep reinforcement learning (DRL) in UAV-enabled wireless communication systems, with a particular focus on recent advancements, algorithmic strategies, and deployment scenarios. A comprehensive survey of the literature revealed a growing trend toward leveraging DRL for various UAV tasks, including trajectory optimization, collision avoidance, cooperative control, secure communication, and edge computing. Notably, actor-critic methods such as PPO and MADDPG, as well as hybrid frameworks combining continuous and discrete action spaces, have demonstrated significant potential for enabling autonomous and adaptive UAV behavior.

An algorithm-oriented gap analysis table was constructed to identify the extent to which key DRL algorithms have been adopted across different UAV functions. The results show that while DQN and its variants remain popular, many studies are now transitioning

toward more robust and scalable algorithms such as PPO, A2C, and multi-agent extensions like MATD3. Furthermore, the review highlighted that only a limited number of works incorporate real-world validation, semantic awareness, or system-level integration, which are critical for bridging the gap between simulation and deployment.

Despite promising results, existing works are constrained by simulation-only evaluations, rigid reward formulations, homogeneous agent assumptions, and lack of semantic or hardware-aware learning. These limitations point toward future research directions that prioritize generalization, interpretability, and real-time adaptability in complex and uncertain UAV communication environments.

Overall, this chapter lays the foundation for proposing a more holistic, intelligent, and deployable reinforcement learning framework tailored to UAV-assisted networks, which will be further detailed in the subsequent methodology chapter.

Chapter 3

Joint DRL-Based Utility Optimization

This chapter is published in X. Cai, P. Lohan, B. Kantarci, "A Novel Joint DRL-Based Utility Optimization for UAV Data Services," IEEE 10th World Forum on Internet of Things, 10–13 November 2024, Ottawa, Canada

3.1 Abstract

In this paper, we propose a novel joint deep reinforcement learning (DRL)-based solution to optimize the utility of an uncrewed aerial vehicle (UAV)-assisted communication network. To maximize the number of users served within the constraints of the UAV's limited bandwidth and power resources, we employ deep Q-Networks (DQN) and deep deterministic policy gradient (DDPG) algorithms for optimal resource allocation to ground users with heterogeneous data rate demands. The DQN algorithm dynamically allocates multiple bandwidth resource blocks to different users based on current demand and available resource states. Simultaneously, the DDPG algorithm manages power allocation, continuously adjusting power levels to adapt to varying distances and fading conditions, including Rayleigh fading for non-line-of-sight (NLoS) links and Rician fading for line-of-sight (LoS) links. Our joint DRL-based solution demonstrates an increase of up to 41% in the number of users served compared to scenarios with equal bandwidth and power allocation.

3.2 Introduction

Uncrewed Aerial Vehicles (UAVs), commonly known as drones, have emerged as a promising solution for enhancing communication networks, especially in areas with limited infrastructure or during emergency situations [40]. The agility, flexibility, and cost-effectiveness of UAVs make them ideal for temporary or supplementary network coverage. However, integrating UAVs into communication networks poses challenges in efficient power and bandwidth (BW) management due to their constrained energy budgets and dynamic deployment conditions.

Power allocation in UAV-assisted networks is crucial for ensuring prolonged operation and coverage. UAVs have limited energy resources, and optimizing their power usage is essential for maximizing their utility and communication capabilities. Moreover, efficient bandwidth assignment is key to supporting heterogeneous data services across diverse users in varying channel conditions. Addressing these challenges requires advanced techniques capable of adaptive decision-making under uncertainty.

Recent advances in Deep Reinforcement Learning (DRL) provide promising solutions for resource allocation in such dynamic environments. These algorithms learn optimal policies through environmental interactions, making them ideal for complex scenarios. For example, leveraging the discrete action capabilities of Deep Q-Networks (DQN) [41] and the continuous action handling of Deep Deterministic Policy Gradient (DDPG) [42], resource allocation to users can be efficiently optimized to maximize UAV utility.

In this paper, we propose a hybrid DRL-based framework for UAV resource optimization under realistic wireless channel conditions. The primary contributions of our work are:

1. A novel joint DRL-based algorithm for optimal resource allocation to users, maximizing UAV utility for heterogeneous data services. DQN manages discrete bandwidth resource block allocation, while DDPG continuously adjusts power levels.

2. Consideration of practical air–ground channel models with Line-of-Sight (LoS) links experiencing Rician fading and Non-Line-of-Sight (NLoS) links experiencing Rayleigh fading.

3. A hybrid architecture where a trained DQN model optimizes bandwidth allocation based on power levels determined by a DDPG agent, achieving a 41% increase in served users compared to equal BW and power allocation baselines.

The rest of the paper is structured as follows: Section II delineates the system model and problem formulation. Section III elaborates on the proposed joint DRL-based solution. Section IV presents the numerical findings, while Section V concludes the paper.

3.3 Related Work

Many existing works [43–46] have proposed various UAV applications in emergency scenarios. The research in [44] presents a system for emergency control using UAVs, offering a method to determine the optimal UAV fleet size based on reliability and task requirements. This enhances UAV deployment in emergencies by considering redundancy and aircraft quality. Similarly, [45] examines the deployment of UAVs as aerial base stations (ABSs) to restore communication networks following severe disasters. In [46], an architecture is presented for integrating UAVs with Non-Terrestrial Networks (NTNs) for sustainable, quality-aware data collection from the ground.

For power optimization, the authors in [47] aim to maximize the downlink rate for a D2D pair in a UAV-aided wireless network, considering coexisting D2D users. The study in [48] addresses UAV utility maximization while analyzing rate coverage probabilities but does not consider detailed air-to-ground fading models. In contrast, [49] provides outage probability expressions for UAV relaying under Rician fading, while [50] optimizes uplink power to minimize total consumption subject to rate requirements. Navigation and swarm control strategies are explored in [51], which lies outside the resource allocation focus of this work.

On the DRL front, [52] maximizes UAV service time and throughput using DDPG. The work in [53] introduces an actor–critic framework for continuous UAV deployment to maximize user coverage and data rate. A comparative study in [54] evaluates DQN and DDPG for autonomous UAV control, demonstrating their suitability for discrete and continuous resource optimization, respectively.

While these studies make significant strides in UAV control, few jointly consider both bandwidth and power optimization under realistic channel conditions using a hybrid DRL approach—an important gap this work aims to fill.

3.4 System Model and Problem Formulation

3.4.1 System and Channel Models

We investigate a UAV-assisted communication system, where N ground users (GUs) are uniformly distributed across a two-dimensional circular field Ψ with a radius of R meters. In this setup, illustrated in Fig. 1, these users communicate with a UAV that remains static and hovers at an altitude of h meters above the center of Ψ. Each ground user GU_i is identified by $i \in I \triangleq 1, 2, \ldots, N$, and their respective positions are defined by coordinates $(x_i, y_i, 0)$ relative to the center of Ψ, $(0, 0, 0)$. The UAV's location is represented by coordinates $(0, 0, h)$ in three-dimensional space.

Consider a designated user GU_i depicted in Fig. 1, situated at a distance $r_i \triangleq \sqrt{x_i^2 + y_i^2}$ from the center of Ψ and the elevation angle of the UAV to that user is θ_i rad. For simplicity, we utilize Euclidean distance metrics in our analysis. Given that the UAV maintains an altitude of h meters above the center of Ψ, the distance between user GU_i and the UAV can be calculated as $d_i \triangleq \sqrt{r_i^2 + h^2} = \frac{h}{sin(\theta_i)}$. Both the UAV and all users are assumed to be equipped with a single antenna.

One common approach for air-to-ground channel modeling between the UAV and users is to consider the LoS and NLoS links separately along with their different occurrence probabilities [55]. Note that for NLoS link, the path loss exponent factor α_{NLoS} is higher than that in the LoS link α_{LoS} due to the shadowing effect and reflection from obstacles. Also to incorporate the effect of small-scale fading, we are considering Rician fading in LoS links and Rayleigh fading in NLoS links. Consequently, the random channel power gains, g_i, for LoS link are noncentral-χ^2 distributed with mean μ and rice factor K [56], and the random channel power gains, k_i, for NLOS link are exponentially distributed with mean μ. Here, μ is the average channel power gain parameter that depends on antenna characteristics and average channel attenuation. With this consideration, the received signal-to-noise ratio (SNR) for Los and NLos links at GU_i can be written as:

$$\mathrm{SNR}_{i,\mathrm{LoS}} = P_i g_i d_i^{-\alpha_{LoS}} / (B_i \sigma^2), \ \forall i \in I, \quad \text{LoS link}, \tag{3.1}$$

$$\mathrm{SNR}_{i,\mathrm{NLoS}} = P_i k_i d_i^{-\alpha_{NLoS}} / (B_i \sigma^2), \ \forall i \in I, \text{NLoS link}, \tag{3.2}$$

where P_i and B_i, respectively, are the transmission power and BW allocated to user GU_i. σ^2 denotes AWGN (additive white Gaussian noise) power density. The probability of LoS link between GU_i and UAV depends upon the elevation angle $\theta_i = \sin^{-1}(\frac{h}{d_i})$, density and height of buildings, and environment. The LoS probability p_{LoS} is written as [55]:

$$p_{LoS} = 1/(1 + C \exp(-B[(180/\pi)\theta_i - C])), \tag{3.3}$$

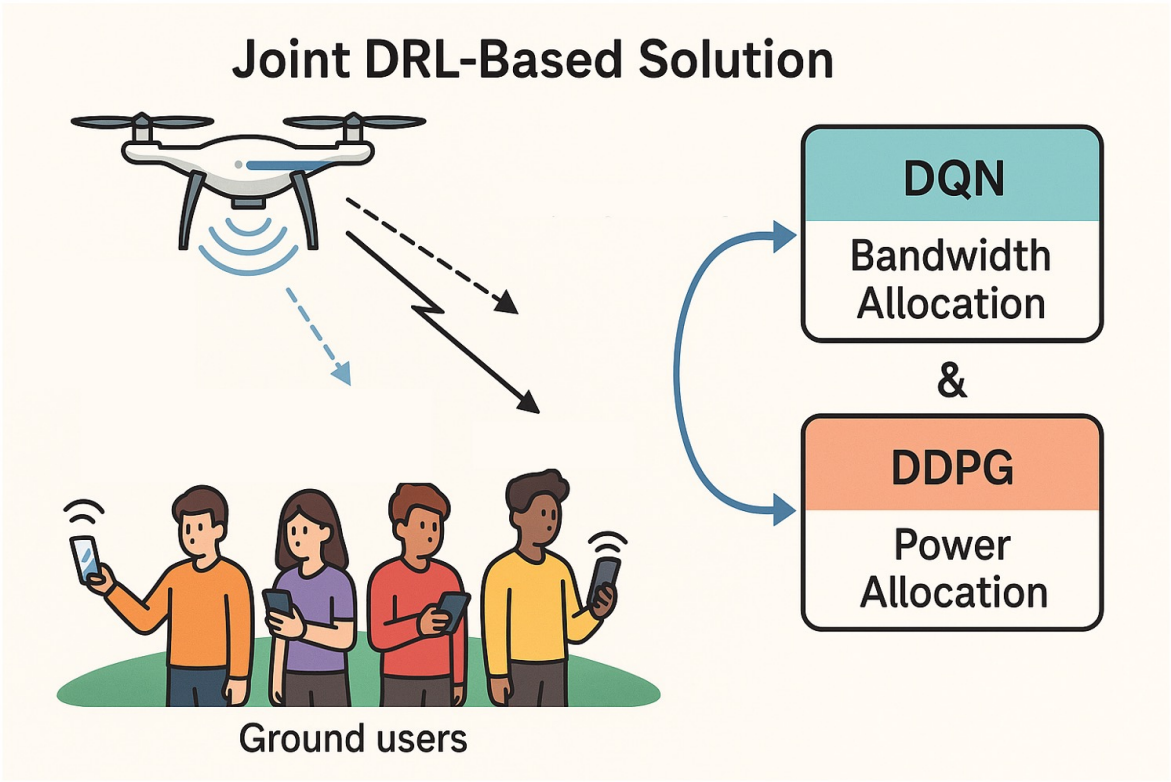

Figure 3.1: Problem Introduction

where C and B are constants that depend on the environment (rural, urban, dense urban). The probability of NLoS link is $p_{NLoS} = 1 - p_{LoS}$. Thus the effective SNR received by user GU_i is expressed as:

$$\mathrm{SNR}_{\mathrm{eff}_i} = P_{\mathrm{LoS}_i} \cdot \mathrm{SNR}_{\mathrm{LoS}_i} + P_{\mathrm{NLoS}_i} \cdot \mathrm{SNR}_{\mathrm{NLoS}_i} \tag{3.4}$$

Using Shannon's capacity formula and SNR representation from equation (3.5), the data rate, η_i bits per sec (bps) for user GU_i through UAV communication link can be expressed as:

$$\eta_i \triangleq B_i \log_2(1 + \mathrm{SNR}_{\mathrm{eff}_i}) \; \forall i \in I. \tag{3.5}$$

Note that a user $GU_i, \forall i \in I$, is considered under UAV rate-coverage or served by the UAV, if the data rate for that user is greater than its desired rate threshold η_{th_i}, i.e., $\eta_i \geq \eta_{th_i}$.

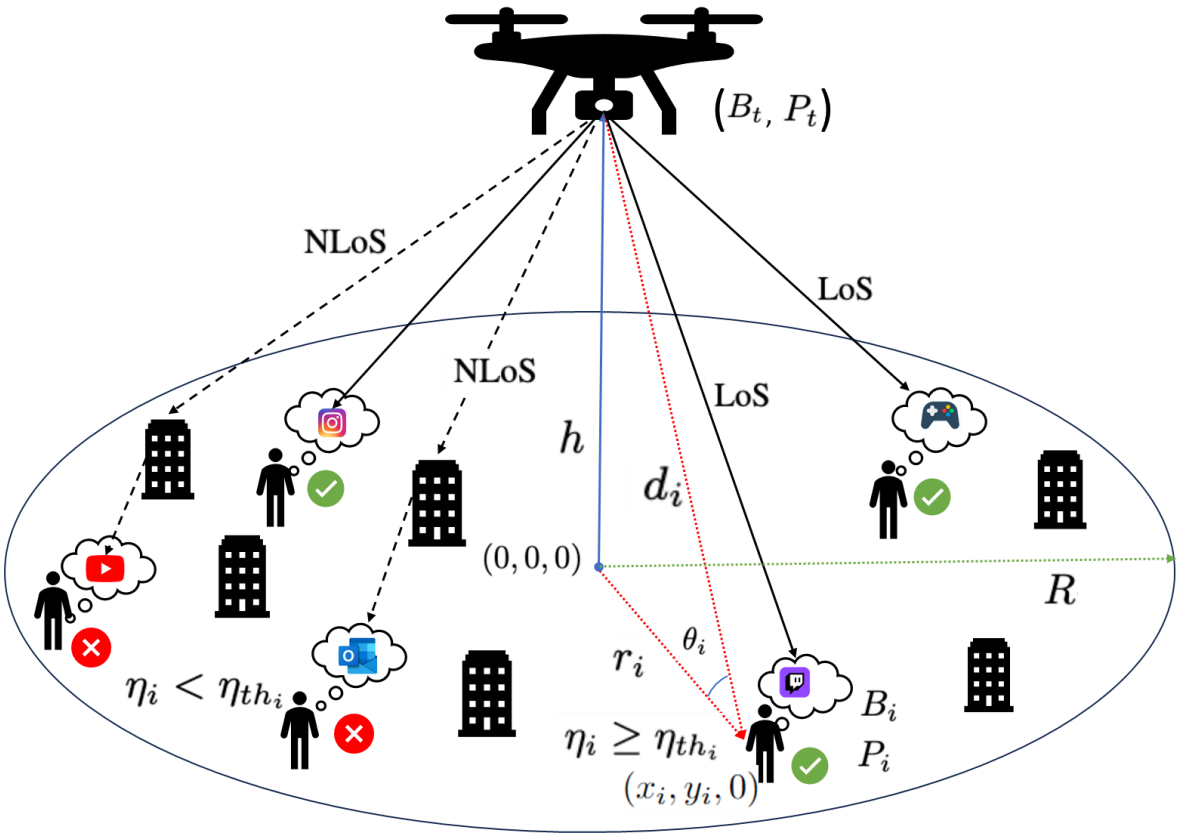

Figure 3.2: System model with ground users in circular field served by a UAV

3.4.2 Problem Formulation

Following the system model, our objective is to enhance the UAV utility by maximizing the number of served users, N_s. This objective can be achieved by optimally allocating the BW and power resources to users, while considering the heterogeneous data rate demand, η_{th_i}, for each user GU_i, $\forall i \in I \triangleq \{1, 2, \ldots, N\}$, limited power budget P_t, and BW resources B_t constraints of the UAV. So, the proposed design framework is mathematically expressed

as follows:

$$
\begin{aligned}
&(\mathcal{P}): \underset{N_s,\mathbf{P},\mathbf{B}}{\text{maximize}}\ N_s, \\
&\text{s.t.: } (C1): \eta_i(P_i, B_i) \geq \eta_{th_i},\ \forall i \in (1, 2, ...N_s), \\
&(C2): \sum_{i=1}^{N_s} P_i \leq P_t,\ (C3): \sum_{i=1}^{N_s} B_i \leq B_t,\ (C4): N_s \in \mathrm{I}, \\
&(C5): P_i \geq 0,\ \forall i \in (1, 2, ...N_s), \\
&(C6): B_i \geq 0,\ \forall i \in (1, 2, ...N_s).
\end{aligned} \tag{3.6}
$$

where, constraint $(C1)$ represents the different data rate requirements for each user, $(C2)$ ensures that the sum of power allocated to the served users should not be more than the total power budget, $(C3)$ is the BW resource constraint, $(C4)$ Specifies that the maximum number of served users must be an integer, while $(C5)$ and $(C6)$ are the boundary conditions for power and BW allocation to each user. The power and BW allocation vectors are represented by $\mathbf{P} = \{P_1, P_2, ..., P_{N_s}\}$ and $\mathbf{B} = \{B_1, B_2, ..., B_{N_s}\}$, respectively. The objective, N_s, of this optimization problem $\mathcal{P}$ is an integer variable. All constraints and sizes of vectors $\mathbf{P}$ and $\mathbf{B}$ are dependent on N_s, which itself is unknown. Therefore, $\mathcal{P}$ is a combinatorial, non-convex and NP-hard problem. To solve this problem, we present a novel joint DRL-based algorithm in the next section.

3.5 Proposed Joint Model

Here, we introduce a joint DRL-based algorithm devised for optimizing resource allocation to users and maximizing UAV utility. In this algorithm, the DQN model is employed to allocate optimal BW to each user, considering its allocated power value, channel condition, and path loss effects. Subsequently, based on the allocated BW and channel conditions, DDPG allocates power to users. This iterative process continues until achieving optimal resource allocation, maximizing the number of users served within the given BW and power resources of the UAV. The proposed joint DRL-based algorithm is provided in Algorithm 1. Now the implementation details of DQN and DDPG models are discussed as follows:

3.5.1 DQN Part Implementation

DQN is a robust DRL method that integrates the traditional Q-learning algorithm with deep neural networks. DQN is particularly effective in environments with discrete action

spaces, making it ideal for applications such as discrete BW resource block allocation to the users. It is worth noting that the available BW is partitioned into resource blocks, each consisting of 1.6KHz BW. This approach enables efficient exploration and exploitation of the action space and provides the means to handle complex decision-making processes in dynamically changing environments.

Model Objective

The objective of the DQN model is to allocate optimal BW resource blocks to users for various scenarios in the shortest possible time.

Initialization

At the beginning of the training process, uniformly distributed ground users' locations are generated within a defined circular field. Then, the data requirement for each user is randomly generated to simulate different data rate requirements of different applications.

State and Observation Space

In this DQN implementation, we initialize the state and observation space with three key elements which are random power levels P_i within a specified limit, random (GU_i) location information $(x_i, y_i, 0)$, and initial BW B_i for each user. Note that, every training episode considers a user with a different initial state.

Action Space

The action space in our model consists of two discrete actions: first, addition of one BW resource block and second, subtraction of one BW resource block. These actions allow DQN model to incrementally adjust BW allocation to each ground user in discrete manner.

State Updates

In the step updation, we perform the following steps: First, the selected BW change is applied to the current BW and the resulting data rate η_i is calculated for the user based on the updated BW. Then, the user's data rate is compared to the required data rate η_{th_i}.

Reward Function

The reward is calculated as the ratio of the achieved data rate to the required data rate. Specifically: Reward $= \eta_i / \eta_{th_i}$.

Penalty Function

If the reward is over 1, a penalty, $(\eta_i / \eta_{th_i} - 1)^2$ would be charged for the reward's deviation from 1. This means that rewards closer to 1 incur smaller penalties, while rewards further away from 1 incur larger penalties. The goal is to minimize resource waste by encouraging rewards that are close to the target value of 1.

Training Termination

One training episode is considered completed and is stopped if the reward for a user is between 1 and $1 + \epsilon$. Here, ϵ depends on the data rate gap above the required threshold with an additional resource block while without this additional resource block, η_i is less than η_{th_i}. DQN model training is terminated when the number of steps required to get optimal BW allocation in each episode is converged. This indicates that the model has learned to allocate the optimal BW to meet the user's data rate requirements with the minimum number of time steps.

Integration with DDPG Training

While the DQN model provides an efficient method for determining the best BW allocation, continuous power levels pose a challenge for subsequent DDPG training. To address this, we adopt a comparison strategy where power levels are considered identical if they match up to four decimal places. This approximation ensures that the discrete experiences from the DQN model are compatible with the continuous action space of power levels resulting from DDPG model.

3.5.2 DDPG Part Implemention

DDPG is a DRL algorithm that combines Deep Learning and Policy Gradient methods. It works well with continuous action spaces, making it suitable for tasks like UAV network optimization where actions, such as power allocation to users, are not discrete. DDPG

uses a deterministic policy to choose actions and a Q-function (critic) to evaluate the expected return of those actions. This combination allows for efficient exploration and stable learning of optimal policies.

Model Objective

The objective of the DDPG model is to optimize the power levels for all ground users in a UAV network to maximize the number of users served within the available power budget. This model operates in a continuous action space, making it suitable for fine-tuning power levels.

State and Observation Space

The state and observation space are defined as follows: power levels $\mathbf{P} = (P_1, P_2, ...P_N$ and BW allocations $\mathbf{B} = (B_1, B_2, ...B_N)$ for all ground users. These elements capture the current configuration of the network, providing the necessary context for decision-making.

Action Space

The action space in the DDPG model consists of continuous adjustments to the power levels for all ground users. This allows for precise control over the power allocation, enabling more efficient resource optimization.

State Updates

In the state updation, the following steps are performed: First, generate a continuous change in the power levels for all users based on the current policy. Second, apply this change to the current state to update the power levels. Third, use the trained DQN model to determine the optimal BW allocation for each user one by one, given the updated power levels. Fourth, If the total used BW exceeds the total BW B_t, count the number of served users N_s as the reward. Fifth, Update the state to reflect the BW allocation by the DQN model for served users. Sixth, Set the BW for unserved users equal to 0.

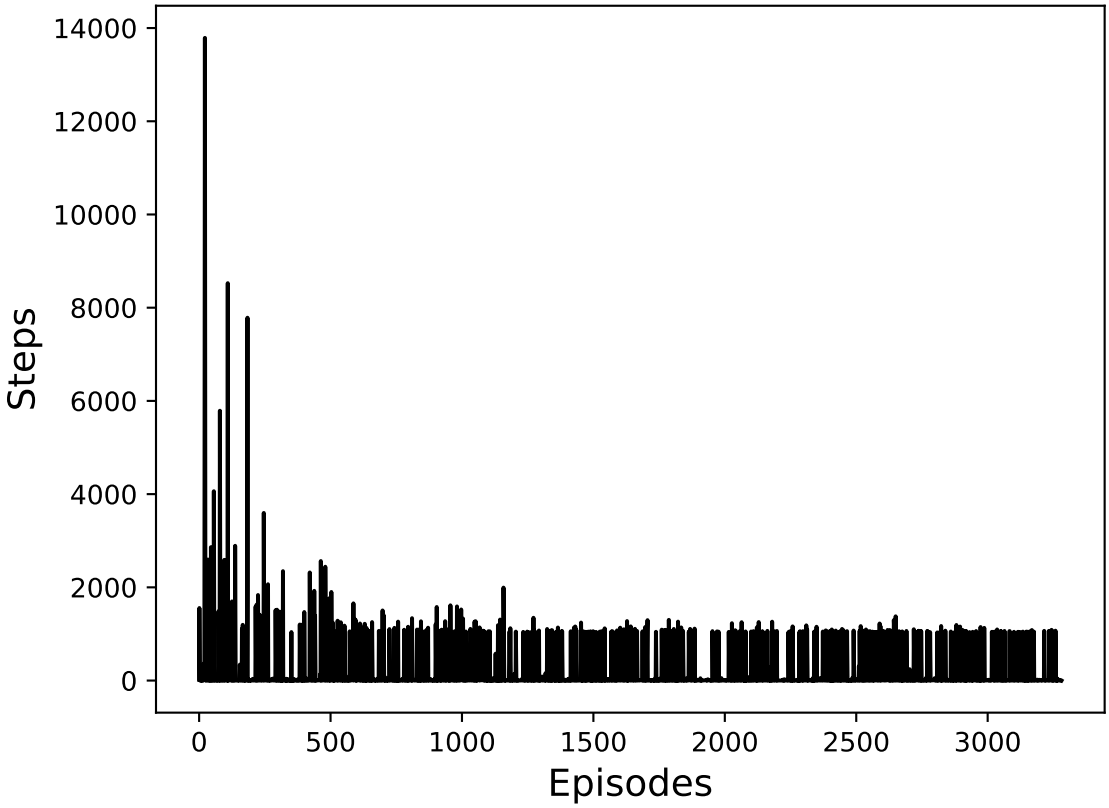

Figure 3.3: DQN training for BW allocation
with random power allocation and user positions

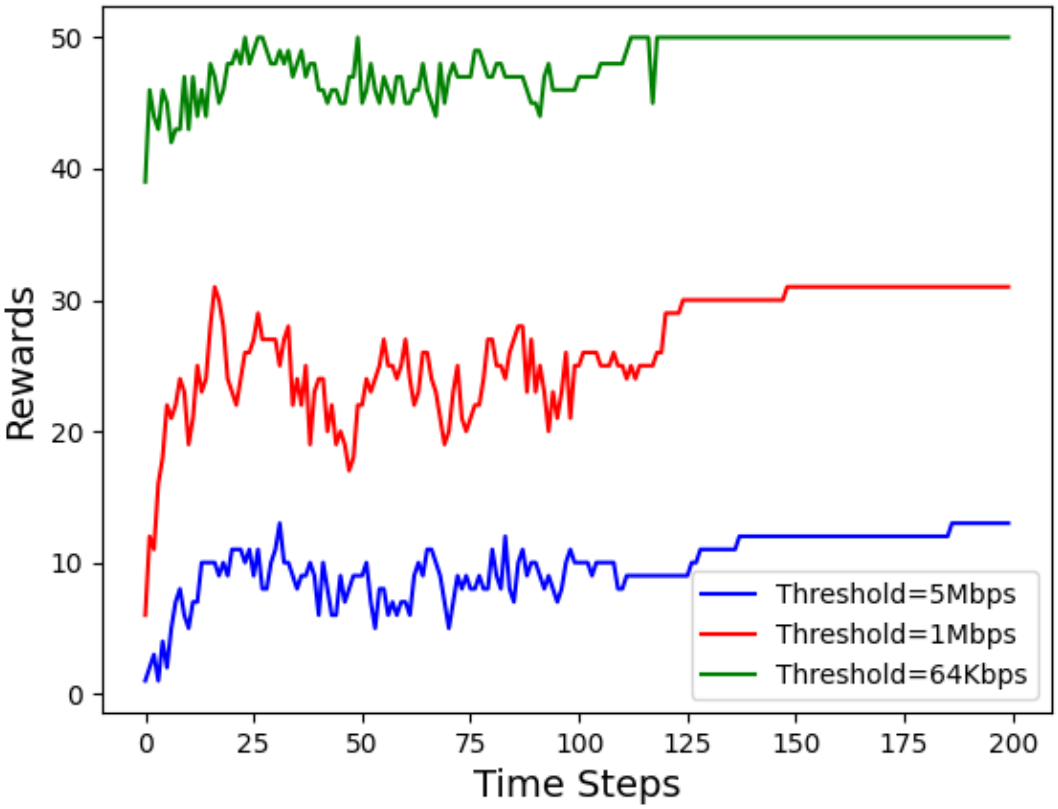

Figure 3.4: Joint DRL-based model's training
with different user thresholds

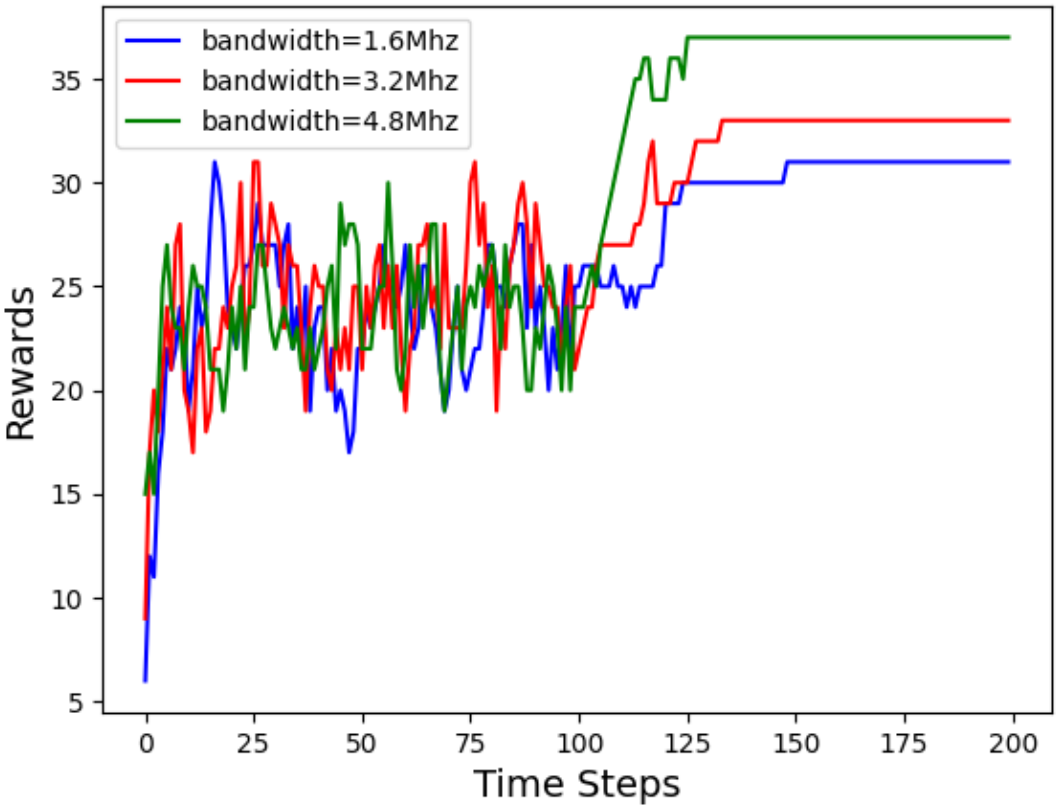

Figure 3.5: Joint DRL-based model's training with different total BW

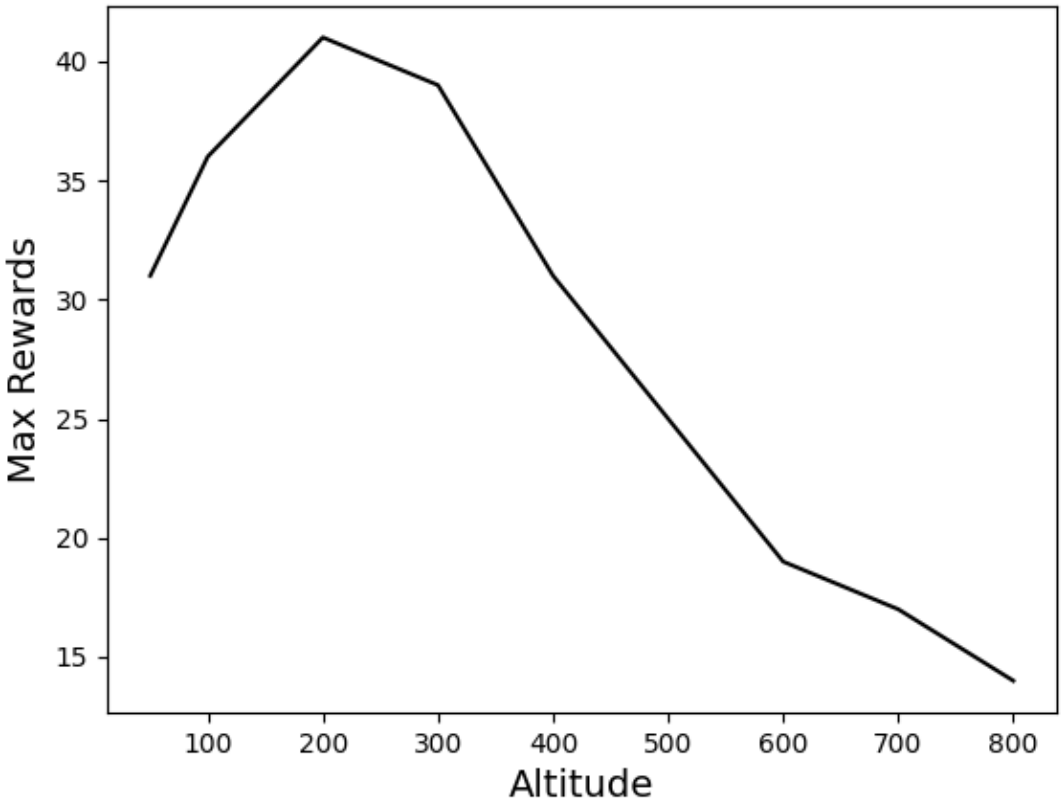

Figure 3.6: Rewards at different altitudes

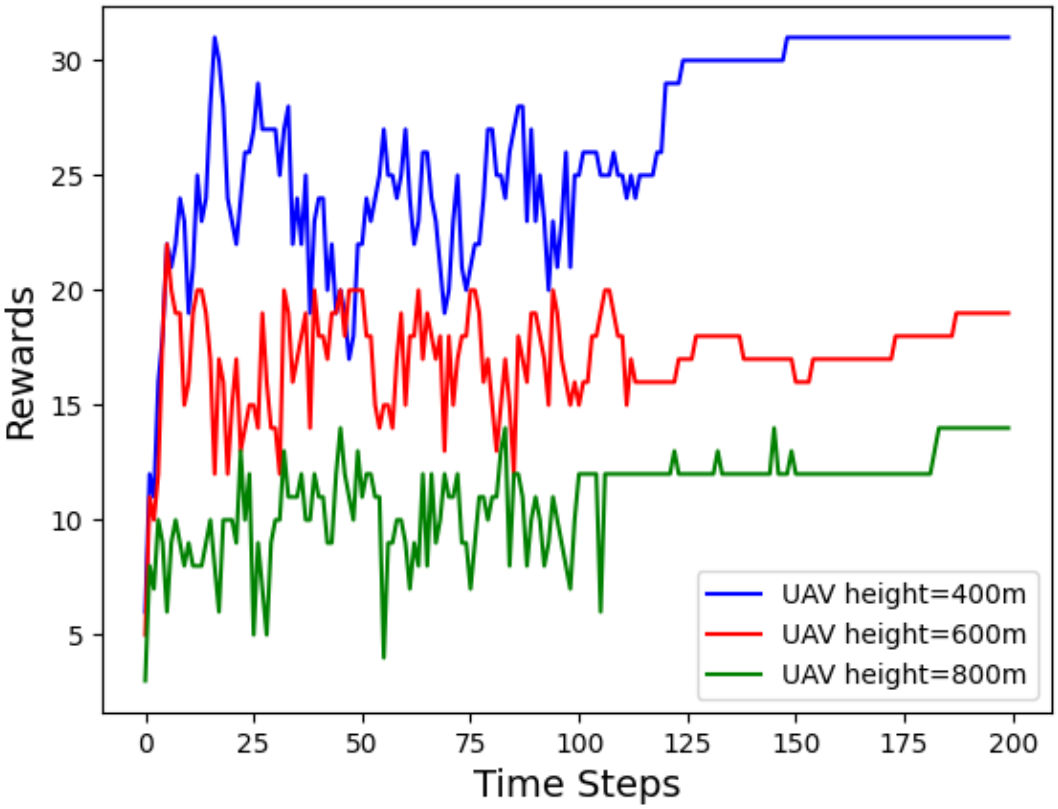

Figure 3.7: Joint DRL-based model's training with different UAV heights

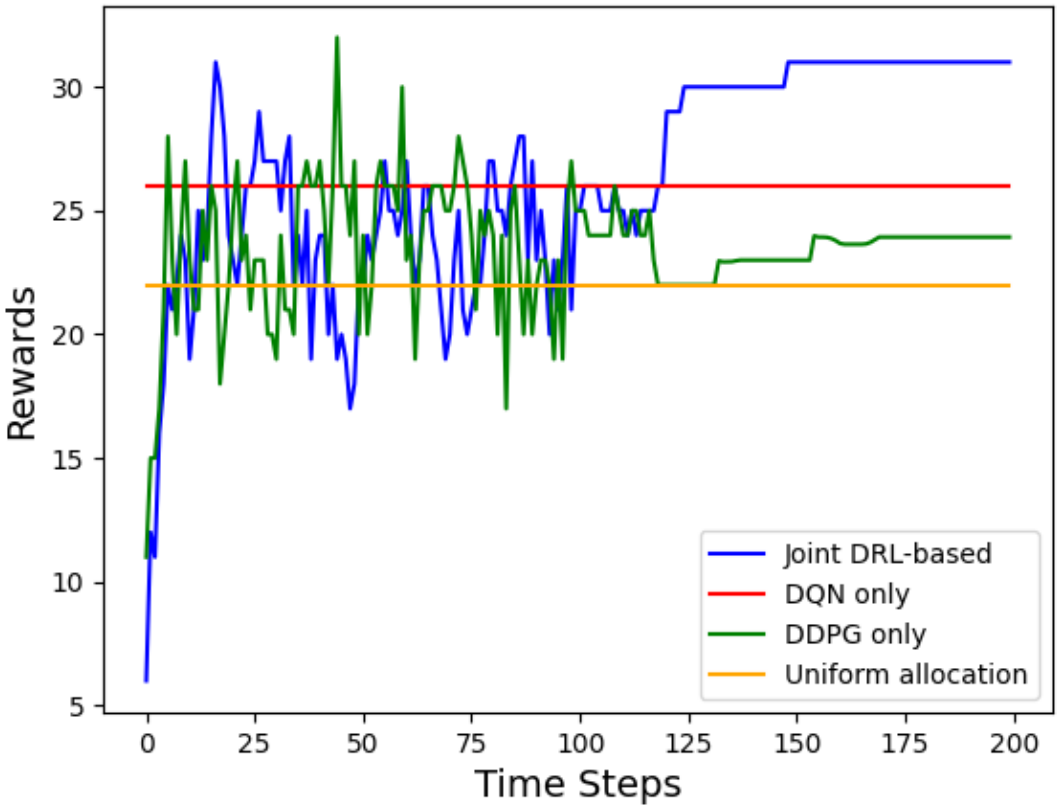

Figure 3.8: Comparative analysis of joint DRL-based algorithm performance

Table 3.1: Environmental Parameters Used in the Simulation

Symbol	Description	Value
R	Radius of the Circular Field	200m
h	Height of the UAV	400m
N	Total Number of Ground Users	50
B_t	UAV Total BW	1.6Mhz=1000 blocks
P_t	UAV Total Power for Transmission	1
B	Environmental Constant (Dense Urban)	0.136
C	Environmental Constant (Dense Urban)	11.95
α_{LoS}	Path Loss Exponent for LoS	2.5
α_{NLoS}	Path Loss exponent for NLoS	3.5
σ^2	Noise Power Spectral Density	10e-17
η_{th_i}	Data Rate Requirement for user GU_i	1Mbps
K	Fading Factor	10
μ	Mean Power	0.5

Reward Function

The reward is defined as the number of served users within the total BW and power constraints, i.e. Reward = N_s. This metric directly incentivizes the model to maximize the efficiency of power allocation.

Penalty Function

If the sum of power allocated to users exceeds the total power budget, P_t, a high penalty is charged to the model that is equal to $(\sum_{i=1}^{n} P_i - P_t) * 10$. It ensures that the model will learn to avoid using more power than budget.

Training Termination

Once the reward stabilizes and converges, we consider the resulting number of users served as the maximum served user number.

3.6 Simulation Results and Discussion

3.6.1 Simulation parameters

Unless otherwise stated all the environmental parameters used in the simulations are provided in Table 3.1. Regarding hyperparameter tuning of DRL models, the learning rate for DQN is set at 0.0001 to ensure stable convergence by fine-tuning the Q-values accurately, whereas DDPG uses a higher learning rate of 0.001 to learn faster in the continuous action space. Both algorithms share a buffer size of 1,000,000 to store experiences for training, aiding in data decorrelation for stability. The batch size for DQN is 32, which suffices for its simpler Q-value updates, while DDPG uses a larger batch size of 256 for more stable gradient updates in its actor-critic framework. Both use a discount factor γ of 0.99 to balance between immediate and future rewards. DQN updates its model every 4 steps to manage computational load and allow Q-values to propagate, whereas DDPG updates at every step for continuous policy and value improvement. Learning starts after 100 steps for both, ensuring enough data in the replay buffer for meaningful training batches. The τ value for DQN is set to 1, indicating hard updates, while for DDPG, τ is 0.005, allowing soft updates for gradual and stable target network changes. Both algorithms employ the MultiInputPolicy type for structured input processing, tailored to the problem domain's specific nature. The values of all hyperparameters are summarized in Table 3.2. In our simulations, the path loss exponent is set to 2.5 for LoS links and 3.5 for NLoS links. These values are adopted in the literature [57], where LoS propagation typically exhibits an exponent between 2 and 3, while NLoS conditions experience more severe attenuation with exponents ranging from 3 to 4.

Table 3.2: Hyperparameters for DQN and DDPG

Hyperparameter	**DQN**	**DDPG**
Learning Rate	0.0001	0.001
Buffer Size	1000000	1000000
Batch Size	32	256
Gamma (γ)	0.99	0.99
Train Frequency	4	1
Learning Starts	100	100
τ for DQN and DDPG	1	0.005
Policy	MultiInputPolicy	MultiInputPolicy

3.6.2 Results and discussion

Table. 3.3 and Table. 3.4 show the power levels and bandwidth blocks optimized result under 700m altitude and 1Mbps data rate threshold. We can serve up to 17 users under this senario.

Fig. 3.3 illustrates the convergence time of the DQN algorithm as it searches for optimal BW allocation strategy. The results indicate that DQN quickly adapts to different user positions and power settings after 500 training episodes, demonstrating its robustness and efficiency in dynamic environments.

Furthermore, DDPG's performance in the joint DRL-based model utilizing the trained DQN model is extensively tested under various network conditions to assess its adaptability and learning efficiency. In Fig. 3.4, the joint DRL-based algorithm's response to different user data rate thresholds is depicted. The training curves suggest that the proposed solution can effectively adjust its policy to meet varying data demands, optimizing BW and power allocation to users to enhance the overall number of users served and fulfill the users' requirements. As discussed earlier, the model begins learning after 100 steps to ensure sufficient data in the replay buffer for meaningful mini-batch sampling and leads to convergence.

Fig. 3.5 shows how the proposed joint DRL-based algorithm manages different amounts of total BW. The results indicate a direct correlation between available BW and the network performance and higher BW allows for a higher number of users served. Fig. 3.8 compares the performance of the proposed joint DRL-based algorithm utilizing DQN for BW allocation and DDPG for power allocation under scenarios where either one of DDPG and DQN is activated or both are deactivated. In the case when both DRL models are deactivated, uniform/ equal resource allocation is considered in that scenario. It is demonstrated that the proposed solution boosts the served users' number by up to 41% compared to the equal resource allocation strategy. In the case When only DDPG is active, it represents optimal power allocation alongside equal BW allocation, and when only DQN is active, it represents optimal BW allocation alongside equal power allocation. Compared to these scenarios, our proposed joint DRL-based solution shows 29%, and 19% improvements, respectively, in number of served users.

Fig. 3.7 explores the impact of UAV altitude adjustments on joint-DRL model training performance. Higher altitudes generally improve LoS link probability but may introduce higher path loss, which is why with an increase in height, the number of served users decreases. Additionally, Fig.3.6 illustrates that lowering the UAV height reduces path loss but diminishes LoS link probability. Consequently, there exists an optimal altitude of approximately 200 meters that balances these effects and yields the highest performance.

Fig. 3.9 shows the number of served users with different K and μ. With better LoS link conditions number of served users is high. Fig. 3.10 displays the performance of the proposed joint DRL-based algorithm with the variation in total power budget and users' data rate requirements. As demonstrated with a higher power budget and lower data rate demands, a larger number of users can be served.

Table 3.3: Power Levels of 50 Users (700m/1Mbps/17 Users Served)

User ID	Power	User ID	Power	User ID	Power	User ID	Power	User ID	Power
1	0.00857309	2	0.00100000	3	0.00818038	4	0.00100000	5	0.02718233
6	0.00374094	7	0.00972122	8	0.00815676	9	0.01180006	10	0.01213871
11	0.02099606	12	0.03000000	13	0.03000000	14	0.03000000	15	0.00493765
16	0.00850547	17	0.02943767	18	0.02448420	19	0.00358069	20	0.01601649
21	0.02115150	22	0.01519261	23	0.03000000	24	0.00100000	25	0.00100000
26	0.00100000	27	0.01589931	28	0.00100000	29	0.03000000	30	0.00641263
31	0.00100000	32	0.00100000	33	0.01672291	34	0.01789321	35	0.00854530
36	0.00190162	37	0.03000000	38	0.02638846	39	0.01524097	40	0.03000000
41	0.02959371	42	0.02787926	43	0.01412626	44	0.00689800	45	0.03000000
46	0.01054110	47	0.01761402	48	0.01783395	49	0.03000000	50	0.01400508

Table 3.4: Bandwidth Blocks of 50 Users (700m/1Mbps/17 Users Served)

User ID	Blocks	User ID	Blocks	User ID	Blocks	User ID	Blocks	User ID	Blocks
1	0	2	0	3	0	4	0	5	0
6	0	7	0	8	0	9	412	10	0
11	4	12	4	13	2	14	0	15	0
16	0	17	29	18	0	19	0	20	0
21	0	22	3	23	6	24	0	25	0
26	0	27	3	28	0	29	0	30	0
31	0	32	0	33	0	34	0	35	0
36	0	37	2	38	3	39	0	40	5
41	0	42	7	43	25	44	0	45	0
46	0	47	4	48	0	49	5	50	4

3.7 Conclusion

This study has proposed the merger of two advanced reinforcement learning algorithms, DQN and DDPG, in the resource management of UAV communication networks. A practical air-to-ground channel modeling has been considered while integrating the small-scale fading effects into it. Through a series of simulations, it has been demonstrated how the

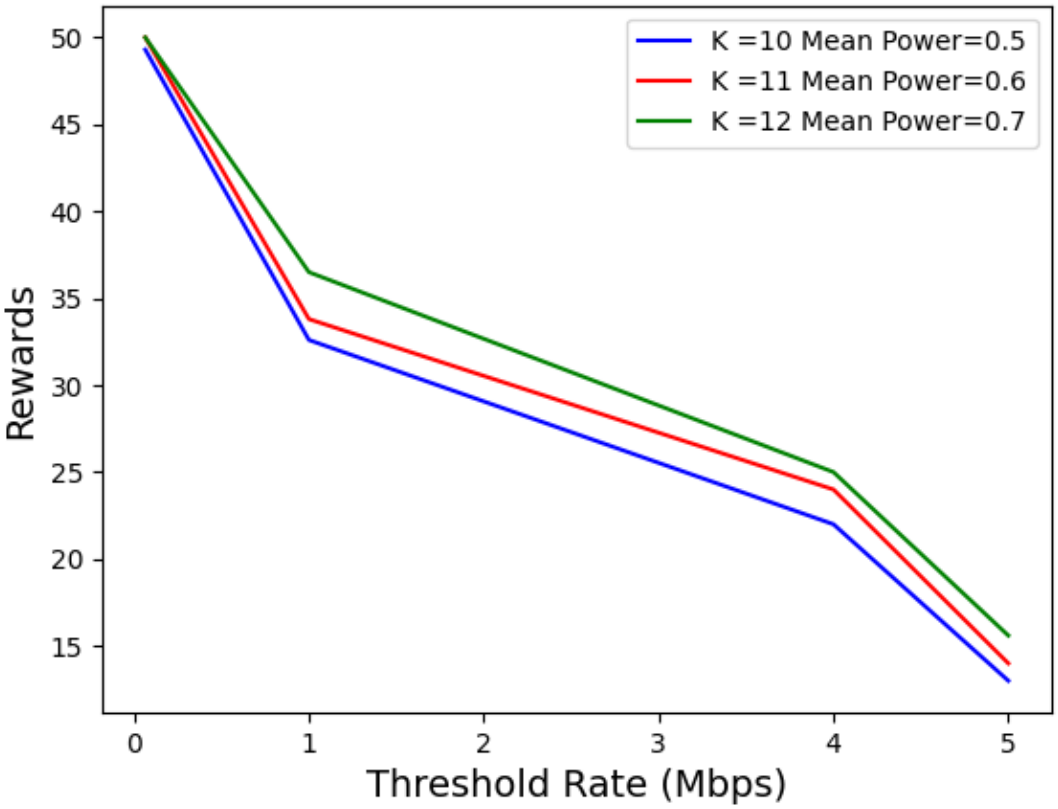

Figure 3.9: Rewards under different total power and thresholds

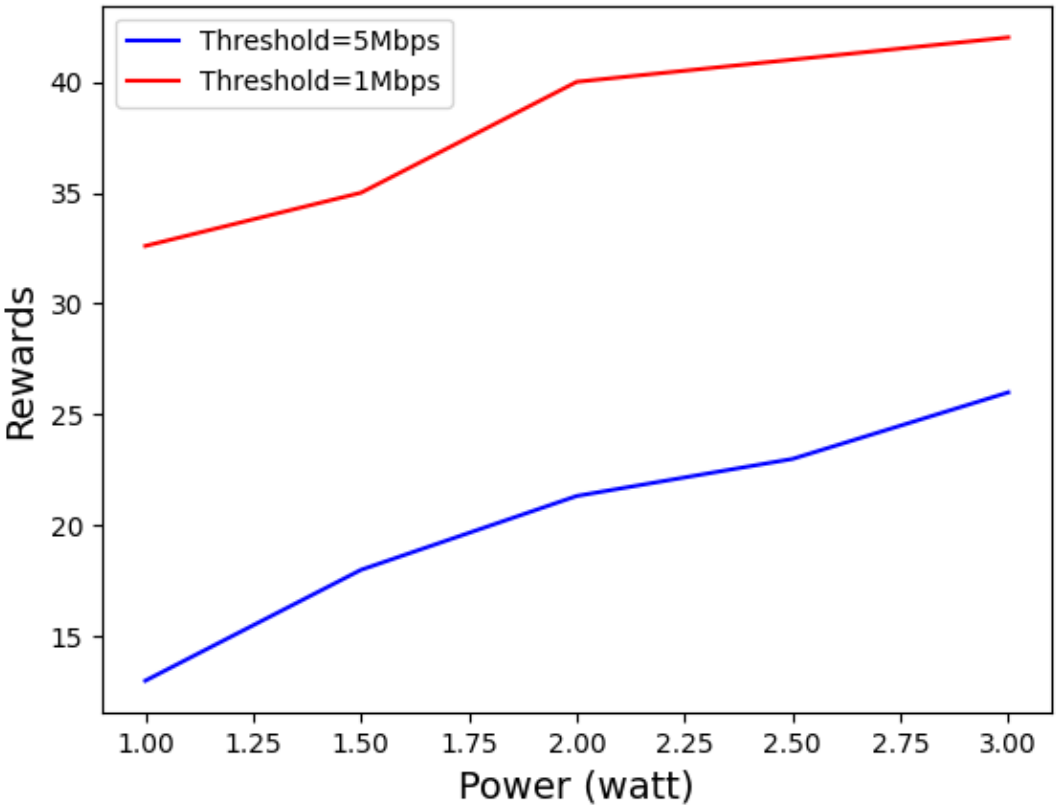

Figure 3.10: Rewards under different total power and thresholds

proposed joint DRL-based algorithm can be effectively utilized to optimize BW allocation and power management in dynamically changing environments and enhance the number of served users by 41% in comparison to the benchmark scheme of uniform resource allocation. Our ongoing research agenda includes exploration of scenarios involving multiple UAVs, taking into account interfering signals.

Algorithm 3.1 Proposed Joint DRL-based Algorithm

Input: Ground users' positions $(x_i, y_i, 0)$ and data rate requirements η_{th_i} for all $i = 1, 2, \ldots, N$
Output: Optimized power and bandwidth allocation via DQN + DDPG

Initialize DQN model with action-value function Q with random weights θ Initialize replay buffer D
for *episode = 1* **to** J **do**
  Observe initial state $s_1 = (P_i, B_i, (x_i, y_i))$ **for** $t = 1$ **to** T **do**
    With probability ϵ select a random action a_t from adding one BW block or decreasing one BW block Otherwise select $a_t = \arg\max_a Q(s_t, a; \theta)$ Execute action a_t in the UAV network and observe reward $r_t = \eta_i/\eta_{\text{th}_i}$ and update new state s_{t+1} **if** $r_t > 1$ **then**
      Calculate the squared deviation of the ratio $(r_t - 1)^2$ as the penalty to avoid too much resource wastage
    Store transition (s_t, a_t, r_t, s_{t+1}) in D Sample a random minibatch of transitions (s_j, a_j, r_j, s_{j+1}) from D Set $y_j = r_j + \gamma \max_{a'} Q(s_{j+1}, a'; \theta)$ Perform a gradient descent step on $(y_j - Q(s_j, a_j; \theta))^2$ with respect to θ
  Gradually reduce ϵ (exploration rate)
Initialize DDPG model with critic network $Q(s, a|\theta^Q)$ and actor $\mu(s|\theta^\mu)$ with weights θ^Q and θ^μ Initialize target network Q' and μ' with weights $\theta^{Q'} \leftarrow \theta^Q$, $\theta^{\mu'} \leftarrow \theta^\mu$ Initialize replay buffer D'
for *episode = 1* **to** K **do**
  Initialize a random process $\mathcal{N}$ for action exploration Receive initial observation state $s_1 = P_1, P_2, \ldots, P_N, B_1, B_2, \ldots, B_N$
  for $t = 1$ **to** T **do**
    Select action $a_t = \mu(s_t|\theta^\mu) + \mathcal{N}_t$ according to the current policy and exploration noise Execute action a_t (continuous power change for all GUs) and use the well-trained DQN model by entering the (P_i, B_i) state, then get the optimal BW B_i, at last observe reward $r_t = N_s$ and new state s_{t+1} Calculate the exceed power $(\sum_{i=1}^{n} P_i - P_t) \cdot 10$ as the penalty to avoid overusing Store transition (s_t, a_t, r_t, s_{t+1}) in D' Sample a random minibatch of W transitions (s_i, a_i, r_i, s_{i+1}) from D' Set $y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'})$ Update critic by minimizing the loss:

$$L = \frac{1}{W} \sum_i (y_i - Q(s_i, a_i|\theta^Q))^2$$

Update the actor policy using the sampled policy gradient:

$$\nabla_{\theta^\mu} J \approx \frac{1}{W} \sum_i \nabla_a Q(s, a|\theta^Q)\big|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^\mu} \mu(s|\theta^\mu)\big|_{s_i}$$

Update the target networks:

$$\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}, \quad \theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}$$

Chapter 4

Multi UAV Deployment and Power Allocation Optimization

This chapter is published in X. Cai, P. Lohan, B. Kantarci, "Multi-Agent Deep Reinforcement Learning for Optimized Multi-UAV Coverage and Power-Efficient UE Connectivity, " IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 1–4 September 2025, Istanbul, Türkiye

4.1 Abstract

In critical situations such as natural disasters, network outages, battlefield communication, or large-scale public events, Unmanned Aerial Vehicles (UAVs) offer a promising approach to maximize wireless coverage for affected users in the shortest possible time. In this paper, we propose a novel framework where multiple UAVs are deployed with the objective to maximize the number of served user equipment (UEs) while ensuring a predefined data rate threshold. UEs are initially clustered using a K-means algorithm, and UAVs are optimally positioned based on the UEs' spatial distribution. To optimize power allocation and mitigate inter-cluster interference, we employ the Multi-Agent Deep Deterministic Policy Gradient (MADDPG) algorithm, considering both LOS and NLOS fading. Simulation results demonstrate that our method significantly enhances UEs coverage and outperforms Deep Q-Network (DQN) and equal power distribution methods, improving their UE coverage by up to 2.07 times and 8.84 times, respectively.

4.2 Introduction

Wireless communication networks are evolving to support the increasing demand for high data rates and low-latency services [58]. Traditional terrestrial infrastructure faces challenges in providing seamless coverage, particularly in remote, disaster-struck, or high-density urban environments. As a result, UAV-assisted communication has emerged as a viable solution to complement existing networks and provide on-demand connectivity. UAVs, functioning as aerial base stations, offer flexibility in deployment, mobility for coverage optimization, and the ability to adapt to dynamic network conditions [53, 59, 60]. However, effective UAV placement and power allocation remain critical challenges due to interference and fading conditions [61]. To address these challenges, we introduce a novel framework for UAV deployment and resource management to maximize the number of UE served at a predefined data rate threshold. Much previous work has applied DRL approaches [62–64], such as DQN [65], DDPG [66, 67], SAC [68], and PPO [69]. We believe that adopting a multi-agent DRL framework could provide a highly effective solution for multi-UAV scenarios.

Our approach begins with the uniform distribution of UE on a grid, followed by K-means clustering to form UE groups. Each cluster is assigned a UAV, which is optimally positioned based on spatial distribution of UEs. To manage power allocation efficiently and mitigate interference, we employ the MADDPG algorithm, a reinforcement learning (RL)-based technique that enables cooperative decision-making among multiple UAVs. Furthermore, our model incorporates both LOS and NLOS fading effects to ensure realistic channel modeling. The contributions of this work are as follows:

- Propose a K-means clustering-based approach for UEs grouping, UAVs' allocation, and determine the optimal UAV positions.
- Integrate MADDPG with dynamic power allocation, improving multi-UAVs coverage efficiency and interference management.

Compared to centralized DQN and equal power allocation, the proposed decentralized MADDPG strategy improves UE coverage efficiency by serving up to 2.07 times and 8.84 times more users, respectively.

The rest of the paper is organized as follows: Section II discusses related works. Section III presents the system model and problem formulation. Section IV details the proposed solution methodology, and Section V provides simulation results and performance analysis. Finally, Section VI concludes the paper and outlines future research directions.

4.3 Related Work

The optimization of power allocation and deployment strategies in UAV-assisted systems has garnered significant attention, particularly in mobile edge computing (MEC) scenarios and energy-efficient UAV operations. Several studies have employed RL techniques to address these challenges, focusing on trajectory design, task offloading, and energy management. [58] proposed various MEC frameworks utilizing MADDPG algorithms to optimize task scheduling, trajectory planning, and resource allocation. These studies demonstrated significant improvements in energy efficiency, reduced task processing delays, and fairness in resource distribution across UAVs. Additionally, [70] extended these approaches to ensure geographical and load fairness while optimizing energy consumption for UAV-assisted MEC networks. For energy-efficient UAV path planning, [61, 71] introduced MADDPG-based algorithms that minimized energy usage through techniques like pruning and optimization of neuron layers, as well as by addressing eavesdropping threats in MEC systems with ground-based jamming. The study presented in [72] focuses on the joint optimization of content caching probability, resource allocation, and UAV flight trajectory. To achieve this, the authors propose a MADDPG-based Resource allocation and UAV trajectory Optimization (MRFO) algorithm to maximize the overall system energy efficiency. In air-ground collaborative networks, [73] proposed multi-UAV systems leveraging Lyapunov optimization and MADDPG for adaptive task offloading, service instance management, and resource allocation. These approaches minimized energy consumption and economic expenditure, demonstrating fast convergence and superior cost efficiency compared to baseline methods. Dynamic and adaptive UAV operations were further explored in [60, 74], where advanced algorithms like MADDPG-LC, Multi-Agent Proximal Policy Optimization (MAPPO), and PPO2-based DRL were employed for dynamic trajectory control, cooperative UAV swarm management, and 3-D trajectory design. These studies highlighted improved energy efficiency, faster convergence, and robustness in addressing flight dynamics and disaster recovery scenarios. [62] employs DQN and DDPG to address bandwidth and power allocation for a single UAV operating in a static, interference-free environment. However, cooperation among multiple UAVs is essential for more complex scenarios.

In contrast to previous works that primarily focus on energy efficiency, task offloading, and trajectory planning using RL techniques such as MADDPG, MAPPO, and DDPG, our contributions differentiate our work from the related literature by addressing the challenges of optimal spatial UAV deployment and dynamic interference management by optimally allocating power to UEs in multi-UAV-assisted networks to enhance coverage efficiency by maximizing the number of users served, rather than prioritizing energy efficiency. While the

existing literature provides a robust foundation for UAV-assisted wireless communication, several key areas remain for further investigation, such as extending current approaches to address dense urban environments and large-scale UAV deployments and enhanced coverage efficiency by optimizing power usage without compromising network performance.

4.4 System Model and Problem Formulation

4.4.1 System Model

UAVs enhance wireless coverage by providing flexible deployment and connectivity in challenging environments. We consider a multi-UAV-assisted communication system, where N UEs are uniformly distributed across a two-dimensional square field Ψ with sides of L meters. The square field consists of 100×100 grids with each grid cell being a square of side $l = L/100$ meters. In this setup, illustrated in Fig. 1, these users are grouped in different clusters and communicate with a dedicated UAV for each cluster. The number of UAVs deployed are equal to the number of UEs' clusters. Note that the user clustering and UAVs deployment approach is discussed in the next section. Each UE UE_i is identified by $i \in I \triangleq 1, 2, \ldots, N$, and their respective positions are defined by coordinates $(x_i, y_i, 0)$ relative to the left lower vertex of Ψ, $(0, 0, 0)$. The location of each UAV UAV_j is represented by coordinates (x_j, y_j, h_j) in three-dimensional space, where h represents the hovering height of all UAVs.

Consider a designated UE i depicted in Fig. 1, situated at a horizontal distance $d_{i,j} \triangleq \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}$ from the associated UAV UAV_j, and the elevation angle of the UAV UAV_j to that user is $\theta_{i,j}$ rad. For simplicity, we utilize Euclidean distance metrics in our analysis. Given that UAVs maintain an altitude of h meters above the field Ψ, the distance between UE UE_i and the UAV UAV_j can be calculated as $r_{i,j} \triangleq \sqrt{d_{i,j}^2 + h^2} = \frac{h}{sin(\theta_i)}$. All the UAVs and users are assumed to be equipped with a single antenna.

One common approach for air-to-ground channel modeling between the UAV and users is to consider the LoS and NLoS links separately along with their different occurrence probabilities [55]. Note that for NLoS link, the path loss exponent factor α_{NLoS} is higher than that in the LoS link α_{LoS} due to the shadowing effect and reflection from obstacles. Also, to incorporate the effect of small-scale fading, we are considering Rician fading in LoS links and Rayleigh fading in NLoS links. Consequently, the random channel power gains, g_i, for LoS link are noncentral-χ^2 distributed with mean μ and rice factor $\mathcal{K}$ [56], and the random channel power gains, k_i, for NLOS link are exponentially distributed with

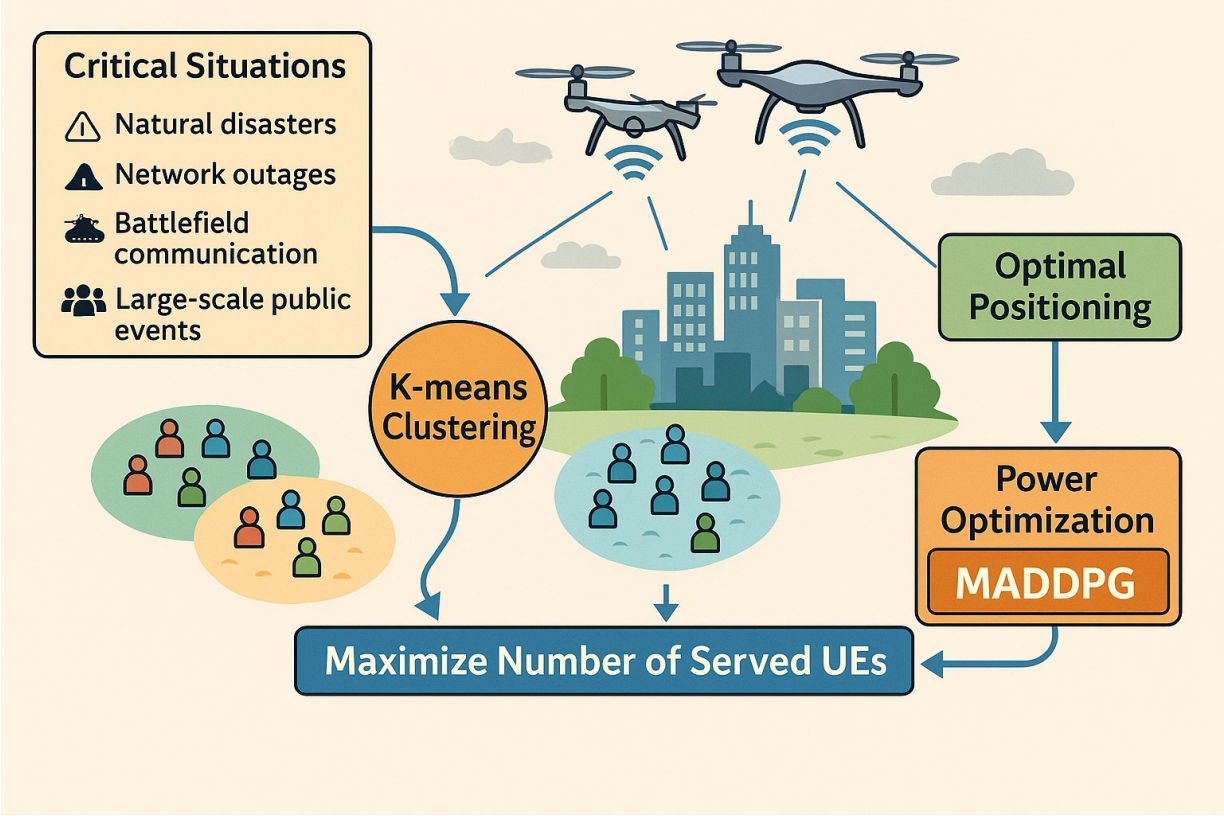

Figure 4.1: Problem Introduction

mean μ. Here, μ is the average channel power gain parameter that depends on antenna characteristics and average channel attenuation. With this consideration, the received power for LoS and NLoS links at UE_i associated with UAV_j can be written as:

$$\mathrm{P}^{r}_{\mathrm{LoS}_{i,j}} = P_{i,j} g_{i,j} r_{i,j}^{-\alpha_{LoS}}, \ \forall i \in I, \tag{4.1}$$

$$\mathrm{P}^{r}_{\mathrm{NLoS}_{i,j}} = P_{i,j} k_{i,j} r_{i,j}^{-\alpha_{NLoS}}, \ \forall i \in I. \tag{4.2}$$

where $P_{i,j}$ is the transmission power allocated to UE_i by UAV_j. The probability of LoS link between UE_i and UAV_j depends upon the elevation angle $\theta_{i,j} = \sin^{-1}(\frac{h}{r_{i,j}})$, density and height of buildings, and environment. The LoS probability $P_{\mathrm{LoS}_{i,j}}$ is written as [55]:

$$P_{\mathrm{LoS}_{i,j}} = 1/(1 + c\exp(-b[(180/\pi)\theta_{i,j} - c])), \tag{4.3}$$

where c and b are constants that depend on the environment (rural, urban, dense urban). The probability of NLoS link is $P_{NLoS} = 1 - P_{LoS}$. Thus the effective power received by

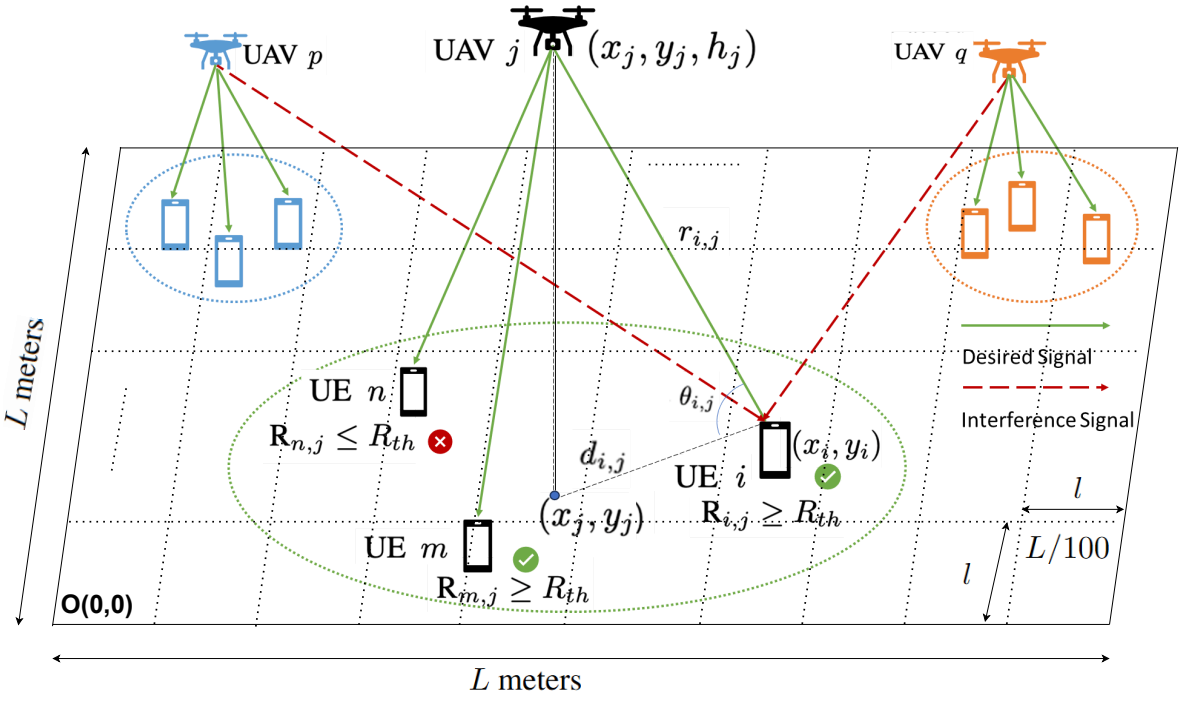

Figure 4.2: System Model

UE_i associated with UAV_j is expressed as:

$$\mathrm{P}^r_{\mathrm{eff}_{i,j}} = P_{\mathrm{LoS}_{i,j}} \cdot \mathrm{P}^r_{\mathrm{LoS}_{i,j}} + P_{\mathrm{NLoS}_{i,j}} \cdot \mathrm{P}^r_{\mathrm{NLoS}_{i,j}} \tag{4.4}$$

Since we are considering the multi-UAV scenario, inter-cluster interference arises from adjacent UAVs transmitting on overlapping frequency bands. Using Shannon's capacity formula, the data rate, $R_{i,j}$ bits per sec (bps) for UE_i through UAV_j communication link can be expressed as:

$$R_{i,j} \triangleq (B/N_j) \log_2 \left(1 + \frac{\mathrm{P}^r_{\mathrm{eff}_{i,j}}}{I_{i,j} + N_o}\right) \ \forall i \in I. \tag{4.5}$$

Here N_o denotes AWGN (additive white Gaussian noise) power, and the inter-cluster interference experienced by UE_i is defined as $I_{i,j} = \sum_{s \neq j} P_{s,avg} k_{i,s} r_{i,s}^{-\alpha_{NLoS}}$. In this expression, only NLoS links are considered for interference calculation, as the probability of a LoS link from interfering UAVs is very low. Since each UAV distributes its entire bandwidth among its associated users, the average transmit power is used as the interfering power. Note that

a user $UE_i, \forall i \in I$, is considered under coverage or served by the UAV, if its data rate meets or exceeds the desired rate threshold R_{th}, i.e., $R_{i,j} \geq R_{th}$.

4.4.2 Problem Formulation

Following the system model, our objective is to maximize coverage by serving the maximum possible number of users with a given set of multiple UAVs. This objective can be achieved by optimally positioning the multiple UAVs and optimally allocating power resources to their associated users while considering the limited power budget P_t constraints of each UAV and reducing the inter-cluster interference. Let I be the set of UEs and $\mathcal{J}$ be the set of UAVs. Define the binary variable in (4.6).

$$a_{i,j} = \begin{cases} 1, & \text{if UE } i \text{ is served by UAV } j, \\ 0, & \text{otherwise,} \end{cases} \tag{4.6}$$

and let $N_j = \sum_{i \in I} a_{i,j}$ denote the number of UEs served by UAV j. The optimization problem is formulated as follows:

$$\begin{aligned} (\mathcal{P}): & \max_{(a_{i,j}, P_j, \mathbf{p}_j)} \sum_{j \in \mathcal{J}} \sum_{i \in I} a_{i,j}, \\ \text{s.t.: } & (C1): R_{i,j} \mathbb{1}(a_{i,j} = 1) \geq R_{th}, \quad \forall\, i \in I, \forall\, j \in \mathcal{J}; \\ & (C2): \sum_{j \in \mathcal{J}} a_{i,j} \leq 1, \quad \forall\, i \in I; \\ & (C3): \ a_{i,j} \in \{0, 1\}, \quad \forall\, i \in I, \forall\, j \in \mathcal{J}; \\ & (C4): \sum_{l=1}^{N_j} P_{lj} \leq P_t; \quad \forall\, j \in \mathcal{J}; \\ & (C5): \mathbf{p}_j = (x_j, y_j) \in \Psi, \quad \forall\, j \in \mathcal{J}. \end{aligned} \tag{4.7}$$

Constraint (C1) ensures that for UE i to be considered as a served user, its achieved data rate $R_{i,j}$ should be at least the minimum required rate threshold R_{th}. Constraint (C2) guarantees that each UE is served by at most one UAV. The third constraint (C3) enforces the binary nature of the variable $a_{i,j}$, meaning a UE is either served by a UAV or not. Constraint (C4) restricts the total transmit power of each UAV to not exceed its maximum allowable power P_t. Constraint (C5) ensures that the horizontal position of each UAV, given by $\mathbf{p}_j = (x_j, y_j)$, lies within the feasible region Ψ. To solve this combinatorial and non-convex problem, we present a novel MADDPG-based solution in the next section.

4.5 Proposed Methodology

The proposed methodology comprises two main parts: 1) clustering UEs' locations and determining UAVs' positions using K-Means and 2) power allocation using MADDPG.

4.5.1 Clustering

We have a set of UE positions, where each UE $i \in \{1, 2, \ldots, N\}$ is observed once. The UE position for UE i is denoted as $\mathbf{u}_i = [x_i, y_i] \in \Psi$. Thus, the set of all UEs' positions is given by $\mathcal{U} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N\}$.

For a given number of clusters $K \in \{1, 2, \ldots, K_{\max}\}$, the K-Means clustering problem is formulated as:

$$\min_{\{c_j\}_{j=1}^{K}, \{s(i)\}_{i=1}^{N}} \sum_{i=1}^{N} \left\| \mathbf{u}_i - c_{s(i)} \right\|^2, \tag{4.8}$$

where $c_j \in \Psi$ denotes the centroid of cluster j, representing a candidate UAV horizontal position and $s(i) \in \{1, 2, \ldots, K\}$ is the cluster assignment for UE i. The solution to the above problem provides the set of centroids $\{c_1, c_2, \ldots, c_K\}$, which serve as the positions for UAVs and the corresponding UE clusters $\mathcal{C}_j = \{\mathbf{u}_i \mid s(i) = j\}$ for $j = 1, \ldots, K$.

4.5.2 MADDPG For Power Allocation

The UAV power allocation problem is modeled as a multi-agent system, where each UAV is an independent agent interacting with the environment. The objective of each UAV is to maximize the number of served UE within its cluster while minimizing penalties due to data rate oversupply and excessive power usage. The MADDPG algorithm is employed for solving this problem by leveraging a centralized training and decentralized execution framework.

Reinforcement Learning Problem Formulation

The problem is defined as a Markov Decision Process (MDP) for K UAV agents, with the following components:

State Space (s) The state space for each agent j at timestep t is defined in (4.9) where $\mathbf{P}_j = \{P_{1j}, P_{2j}, \ldots, P_{N_{s,j}j}\}$ stands for the power allocation to UE within the cluster of UAV j, $\mathbf{R}_j = \{R_{1j}, R_{2j}, \ldots, R_{N_{s,j}j}\}$ denotes the achieved data rates for UE associated with UAV j, and $N_{s,j}$ represents the number of UEs in cluster $\mathcal{C}_j$ associated with UAV j.

$$\mathbf{s}_j^t = \{\mathbf{P}_j, \mathbf{R}_j, N_{s,j}\} \tag{4.9}$$

Action Space ($\mathbf{a}_j$) The action space for each agent j corresponds to the adjustment of power allocated to the UE in its cluster as formulated in (4.10):

$$\mathbf{a}_j^t = \{\Delta P_{j1}, \Delta P_{j2}, \ldots, \Delta P_{jN_{s,j}}\}, \tag{4.10}$$

where ΔP_{ji} represents change in power allocated by UAV j to UE i.

Reward Function (r) The reward in (4.11) is computed as the sum of two components: the total number of served UEs and the effective total data rate (i.e., the aggregate data rate across all UEs after subtracting the wasted data rate). Let total served users be defined as the sum of the number of served UEs (U_{C_j}) in each cluster $C_j; \forall j = 1, ..., k$ (i.e., those all UEs whose effective data rate R_{ij} meets or exceeds the threshold R_{th}) as formulated in the first component of the reward function. The total data rate is formulated in the second summation component in the reward function where W_d denotes the unutilized data rate (i.e., the excess data rate above the required threshold that is not effectively utilized).

$$r = \sum_{j=1}^{K} U_{C_j} + \left(\sum_{j=1}^{K} \sum_{i=1}^{U_{C_j}} R_{ij} - W_d \right) \tag{4.11}$$

Transition Dynamics The environment transitions from state $\mathbf{s}_j^t$ to $\mathbf{s}_j^{t+1}$ based on the UAV's action $\mathbf{a}_j^t$. The updated power allocation affects the SINR, data rate, and the resulting reward.

Centralized Training and Decentralized Execution

In the MADDPG framework, centralized training is employed using a global critic network, while execution is decentralized using individual actor networks.

Critic Network (Q_j) The centralized critic evaluates the joint action-value function in (4.12) where $\mathbf{s}$ is the global state, $\mathbf{a} = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_K\}$ is the joint action of all agents, and γ is the discount factor.

$$Q_j(\mathbf{s}, \mathbf{a}) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r_j^t \mid \mathbf{s}, \mathbf{a}\right], \tag{4.12}$$

Actor Network (μ_j) Each agent j uses an actor network to determine its action as shown in (4.13) where θ_{μ_j} are the parameters of the actor network for agent j.

$$\mathbf{a}_j^t = \mu_j(\mathbf{s}_j^t | \theta_{\mu_j}), \tag{4.13}$$

MADDPG Algorithm

This proposed MADDPG solution outlined as Algorithm 4.1, enables UAVs to learn cooperative policies that maximize the number of served UEs while minimizing penalties for inefficient resource utilization.

4.6 Numerical Results

4.6.1 Environment and MADDPG Training Parameters

Unless otherwise stated all the environmental parameters used in the simulation setup are provided in Table 4.1. The hyperparameters in MADDPG training are selected to ensure stable and efficient learning. A replay buffer of 100,000 experiences supports off-policy learning, while a batch size of 64 stabilizes updates. A learning rate of 0.0001 is chosen to prevent drastic weight updates and enhance convergence stability, and a discount factor $\gamma = 0.95$ balances short and long-term rewards. A soft update rate $\tau = 0.01$ ensures smooth policy updates, while an exploration noise of $\sigma_{\text{noise}} = 0.2$ promotes exploration, preventing premature convergence to suboptimal policies. The selected values of all hyper-parameters are summarized in Table 4.2. In our setup, training typically requires over 500 episodes where each episode includes up to 500 time steps, resulting in several hours of GPU computation time. The simulations have been conducted over 10 different seeds, and all figures present the average results across these 10 runs, with 95% confidence interval (CI) bars indicating variability. In our simulations, the path loss exponent is set to 3 for LoS links and 4 for NLoS links. These values are adopted in the literature [57], where

LoS propagation typically exhibits an exponent between 2 and 3, while NLoS conditions experience more severe attenuation with exponents ranging from 3 to 4.

Table 4.1: Default Environmental Parameters Used in the Simulations

Symbol	Description	Value
L	Side of square field Ψ	10000 meters
$l \times l$	Default cell size	100×100 meters
h	Height of UAVs	500 meters
N	Total Number of UE	30
P_t	Total Power of each UAV	1 W
B	Total bandwidth of each UAV	10 MHz
R_{th}	Data Rate Requirement per UE	30 Mbps
N_o	Noise Power	4×10^{-15} W
α_{LOS}	Path Loss Exponent for LoS	3
α_{NLOS}	Path Loss Exponent for NLoS	4
c	Environmental Constant (Dense Urban)	11.95
b	Environmental Constant (Dense Urban)	0.136
$\mathcal{K}$	Fading Factor	10
μ	Mean Power	0.5

4.6.2 Results and Discussion

Training Convergence

Fig. 4.3 shows the convergence behavior of MADDPG training over 500 time-steps for 3, 5, and 7 clusters. The reward starts low and steadily increases, stabilizing after 100 time-steps. Minor drops likely result from policy updates or exploration-exploitation trade-offs. Training with 7 clusters yields the most stable policy with minimal variance and the highest reward, suggesting that more clusters enhance learning through richer interactions and exploration.

Served UEs Comparison

Fig. 4.4 compare the performance of the proposed MADDPG solution against DQN and equal power allocation. To ensure a fair comparison, three different R_{th} values are consid-

Table 4.2: MADDPG Hyperparameters

Symbol	Description	Value
B_s	Replay Buffer Size	$100,000$ samples
W	Batch Size for Training	64 samples
α_a	Actor Network Learning Rate	0.0001
α_c	Critic Network Learning Rate	0.0001
γ	Discount Factor for Future Rewards	0.95
τ	Target Network Update Rate	0.01
σ_{noise}	Exploration Noise Level	0.2
H	Hidden Layer Sizes for Actor and Critic	$[128, 128]$

Table 4.3: Mean Transmission Power Usage

Clusters	MADDPG	DQN	Equal Power
5	98.81%	99.00%	100%
10	40.53%	46.00%	100%
15	7.19%	10.27%	100%
20	1.73%	3.85%	100%
25	0.89%	2.80%	100%

ered. The results indicate that as R_{th} increases, the performance gap between MADDPG and DQN widens, particularly for a smaller number of clusters, leading to a higher number of served users with the proposed approach. Also, with less R_{th} requirements, fewer clusters/UAVs are sufficient to cover all the users. Furthermore, Fig. 4.7 shows the comparison results with higher user density (with N=60 UEs) indicating better performance of MADDPG over DQN and equal power allocation. With an increasing number of clusters and UAVs in the air, the system is able to provide greater transmission power. Moreover, when each cluster contains fewer users, the available resources can be distributed more generously, allowing each user to receive a higher share. This, in turn, enables the network to support a larger number of users overall. However, we are still exploring the optimal balance point between efficiency and scalability. Our objective is to design the system such that each UAV can effectively serve more users.

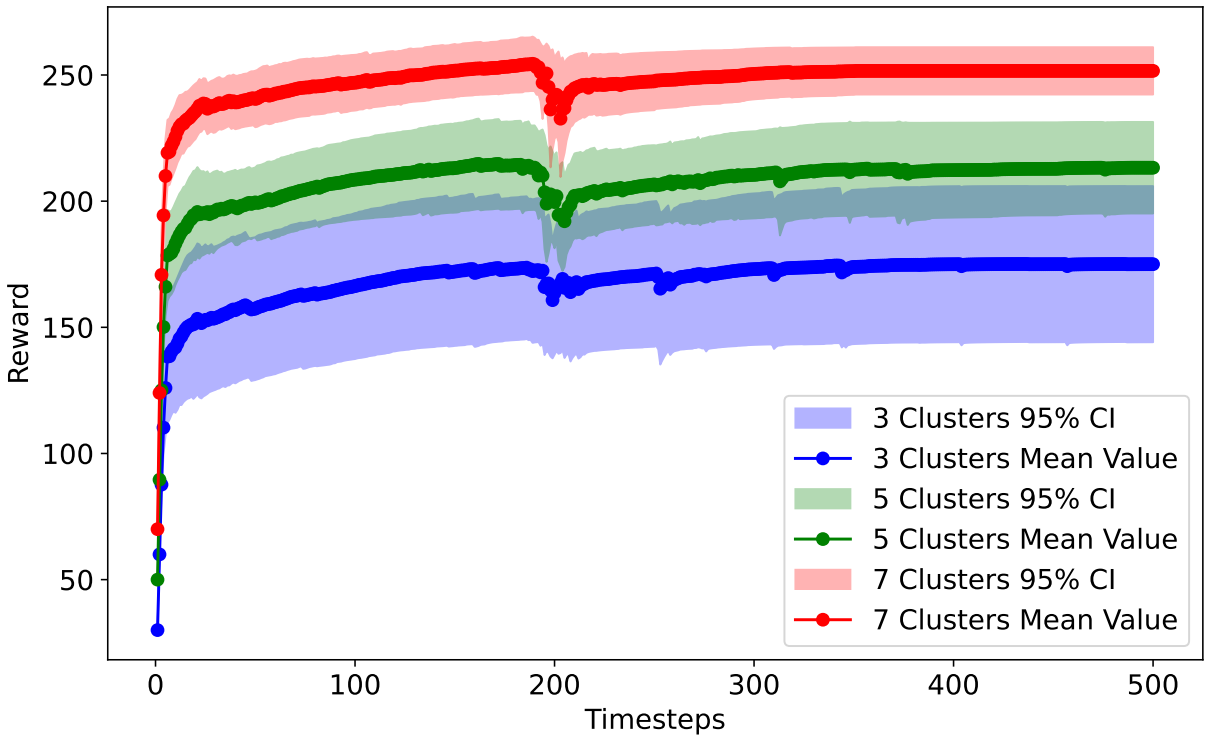

Figure 4.3: MADDPG Training Convergence ($N = 30$ UEs, $R_{th} = 30$ Mbps)

Power Usage Comparison

Table 4.3 presents the average transmission power consumption of UAVs for MADDPG, DQN, and equal power allocation across varying cluster counts. With fewer clusters, UAVs must cover larger distances to serve users, resulting in higher transmission power requirements. As the number of clusters increases, power consumption decreases since users are distributed more evenly, reducing the distance between UAVs and their associated users. Among the evaluated approaches, MADDPG demonstrates better performance with lower transmission power usage.

Cell Scale Impact

Fig. 4.8 shows MADDPG performance across different cell scales/coverage areas with L=10000 m, L=30000 m, and L=50000 m. Larger coverage areas/cell scales result in greater distances between UEs and their associated UAV clusters, leading to a lower number

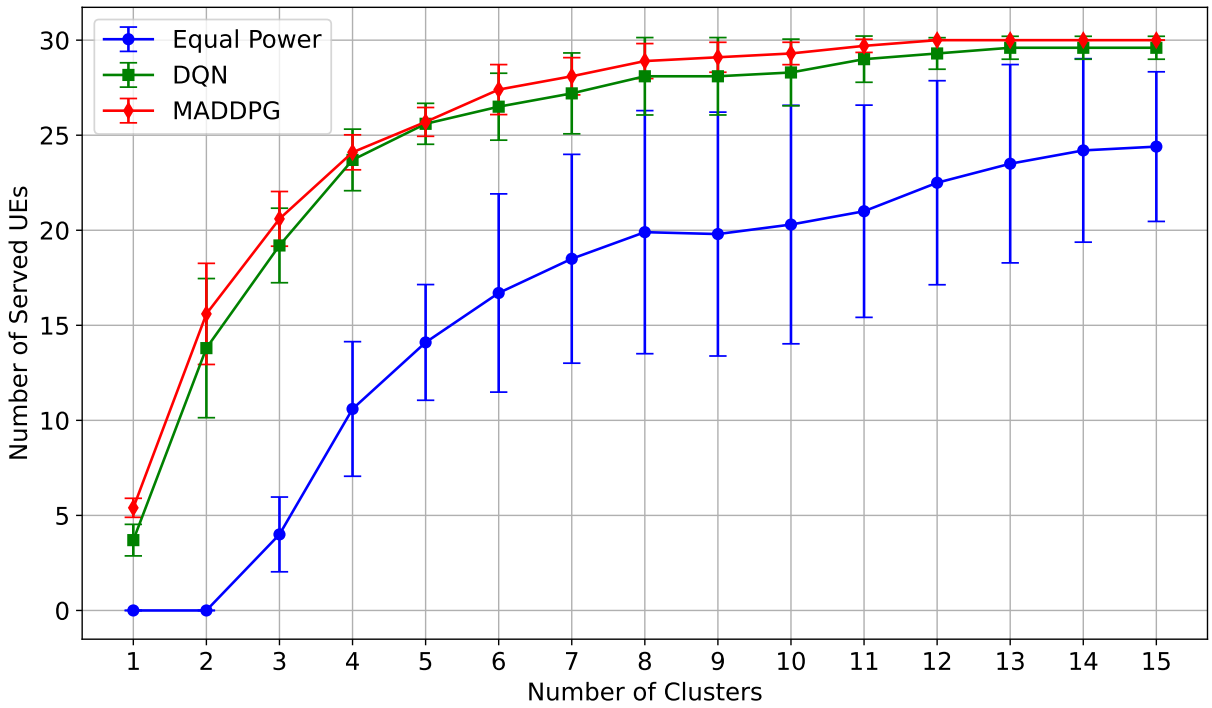

Figure 4.4: Performance comparison for $N = 30$ UEs $R_{th} = 10$ Mbps

of served users. However, as the number of clusters/UAVs increases, this performance gap narrows since UAVs are positioned closer to UEs, improving coverage efficiency.

4.7 Conclusion

This paper has presented a multi-UAV-assisted wireless network proposing K-means clustering and MADDPG-based solution for optimal positioning of UAVs and optimal power allocation, respectively. Compared to centralized DQN and equal power distribution, our decentralized MADDPG approach improves UE coverage efficiency maximum of 2.07 times and 8.84 times, respectively. The framework incorporates realistic LoS/NLoS fading and interference modeling, accurately capturing wireless dynamics. By leveraging MADDPG, UAVs autonomously learn optimal strategies and enhance UE coverage, as well as data rates while maximizing network performance. Our ongoing research aims to further enhance the system's adaptability by incorporating dynamic UE mobility models and tackling energy-efficient trajectory optimization.

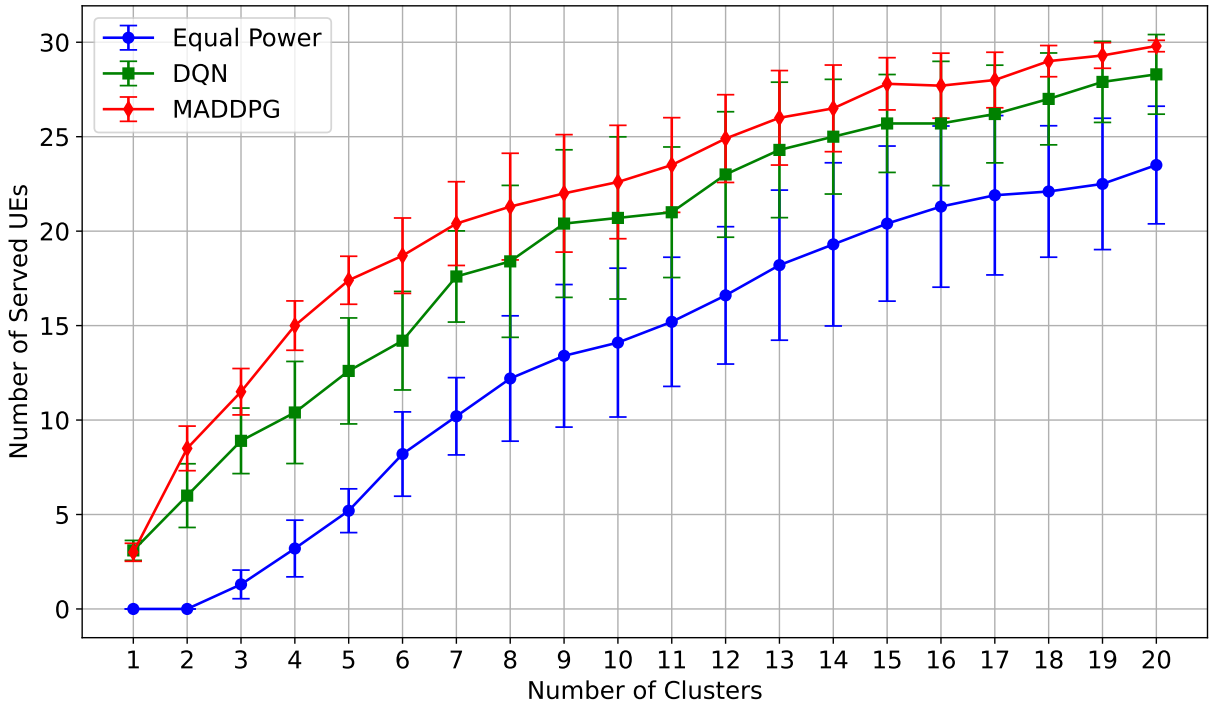

Figure 4.5: Performance comparison for $N = 30$ UEs $R_{th} = 20$ Mbps

Algorithm 4.1 Proposed MADDPG Solution

Input: Number of agents k, replay buffer $\mathcal{D}$, batch size W, discount factor γ, target network update rate τ

foreach *agent* j **do**
  Initialize actor network μ_{θ_j} and critic network Q_{ϕ_j} with random parameters θ_j and ϕ_j Initialize target networks $\mu_{\theta'_j}$ and $Q_{\phi'_j}$ with weights $\theta'_j \leftarrow \theta_j$, $\phi'_j \leftarrow \phi_j$
Initialize replay buffer $\mathcal{D}$ (shared by all agents)
for *episode* $= 1$ **to** M **do**
  Initialize a random process $\mathcal{N}$ for action exploration Receive initial global state $\mathbf{s}_0$
  for $t = 0$ **to** $T - 1$ **do**
    foreach *agent* $j \in \{1, \ldots, K\}$ **do**
      Select action $a_j^t = \mu_{\theta_j}(\mathbf{o}_j^t) + \epsilon$, where $\epsilon \sim \mathcal{N}$, and $\mathbf{o}_j^t$ is agent j's local observation
    Execute joint action $\mathbf{a}^t = (a_1^t, \ldots, a_K^t)$ in the environment Check the power limit and prioritize to nearby UE Collect the power matrix of all UEs, calculate the interference and finalize the data rate matrix Observe next global state $\mathbf{s}_{t+1}$ and immediate rewards $r_1^t, \ldots, r_K^t$ Store transition $(\mathbf{s}_t, \mathbf{a}^t, r^t, \mathbf{s}_{t+1})$ in $\mathcal{D}$ $\mathbf{s}_t \leftarrow \mathbf{s}_{t+1}$
    if *replay buffer* $\mathcal{D}$ *has enough samples* **then**
      foreach *agent* j **do**
        Sample a mini-batch of W transitions from $\mathcal{D}$ Compute target $y_j = r_j + \gamma\, Q_{\phi'_j}(\mathbf{s}_{t+1}, \mathbf{a}_1^{t+1}, \ldots, \mathbf{a}_K^{t+1})\big|_{a_j^{t+1} = \mu_{\theta'_j}(\mathbf{o}_j^{t+1})}$ Update critic by minimizing loss:

$$\mathcal{L}(\phi_j) = \frac{1}{W} \sum \big(y_j - Q_{\phi_j}(\mathbf{s}_t, \mathbf{a}_1^t, \ldots, \mathbf{a}_K^t)\big)^2$$

        Update actor using policy gradient: $\nabla_{\theta_j} J$

$$\approx \frac{1}{W} \sum \nabla_{a_j} Q_{\phi_j}(\mathbf{s}_t, \mathbf{a}_t)\big|_{a_j = \mu_{\theta_j}(\mathbf{o}_j^t)} \nabla_{\theta_j} \mu_{\theta_j}(\mathbf{o}_j^t)$$

        Update target networks:

$$\theta'_j \leftarrow \tau\theta_j + (1-\tau)\theta'_j, \quad \phi'_j \leftarrow \tau\phi_j + (1-\tau)\phi'_j$$

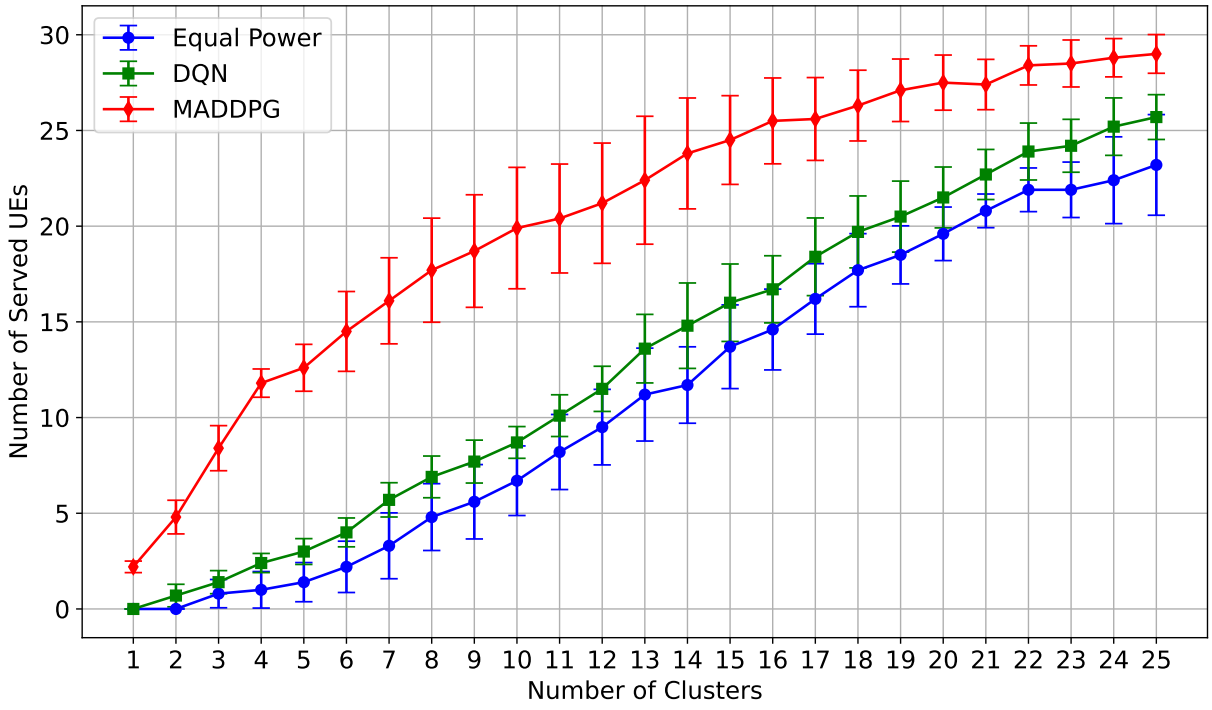

Figure 4.6: Performance comparison for $N = 30$ UEs $R_{th} = 30$ Mbps

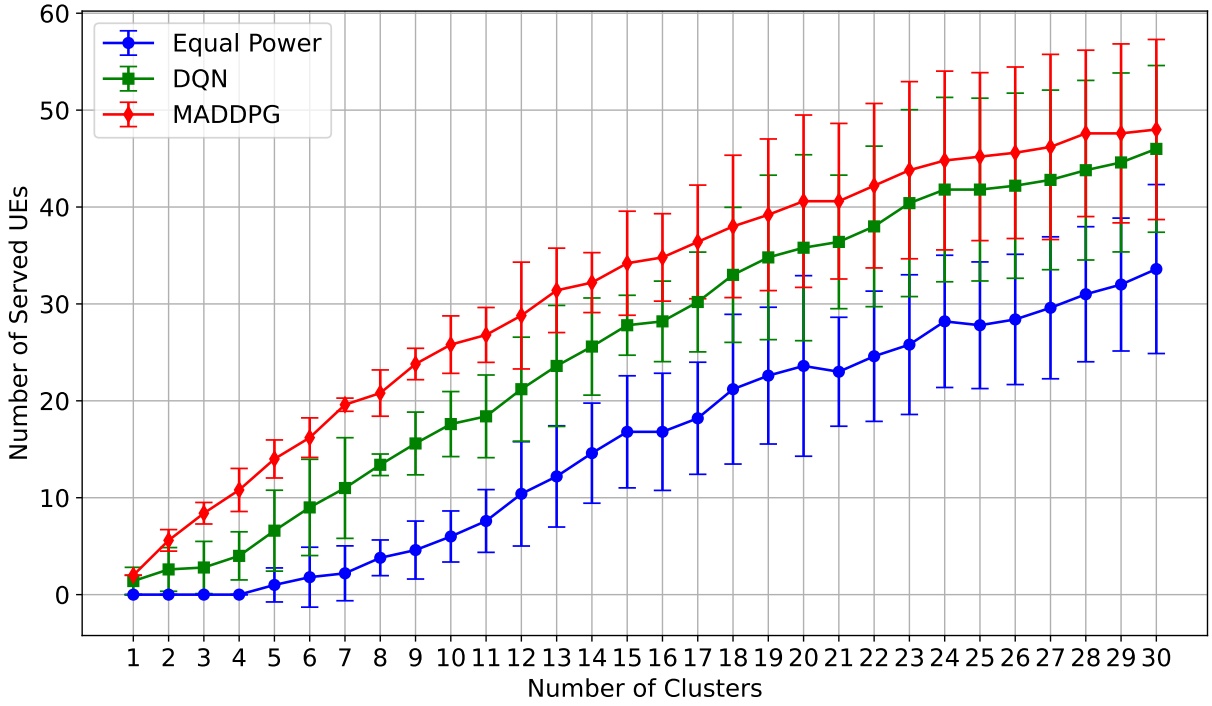

Figure 4.7: Performance comaprison for $N = 60$ UEs ($R_{th} = 30$ Mbps)

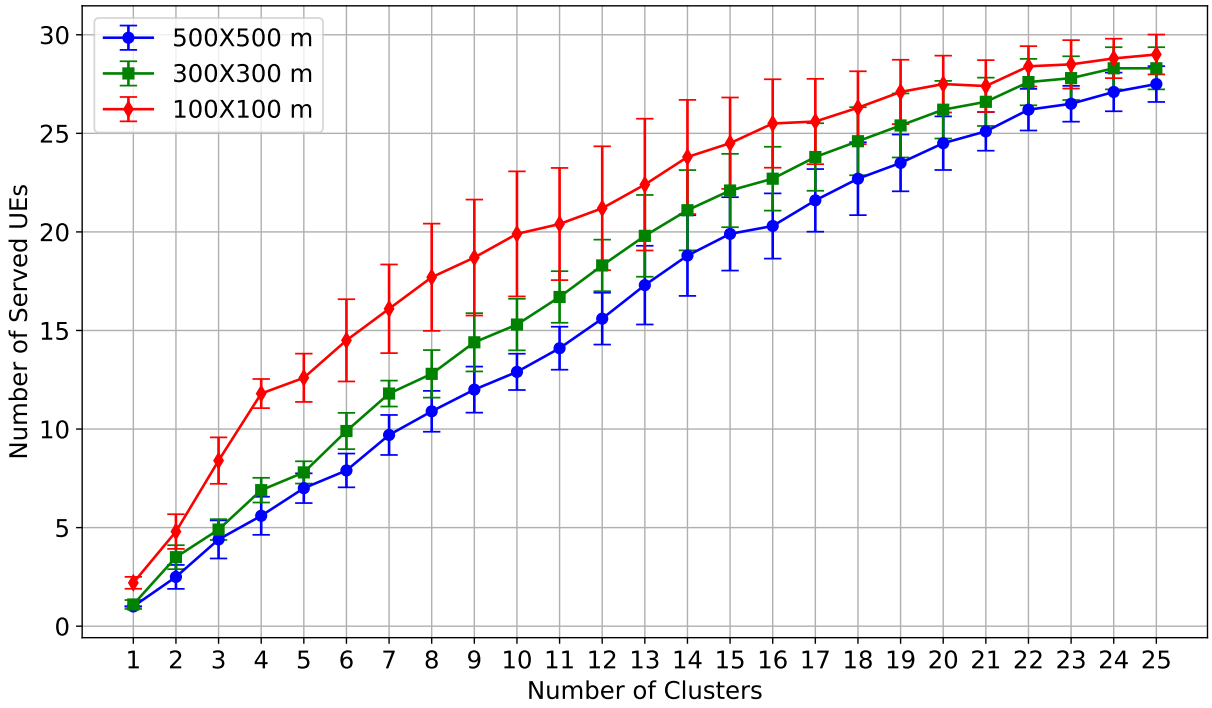

Figure 4.8: MADDPG performance for different cell scales (N=30, R_{th}=30 Mbps)

Chapter 5

Mobility-Aware Users Data Service Optimization

This chapter is published in X. Cai, P. Lohan, B. Kantarci, "FLARE: Flying Learning Agents for Resource Efficiency in Next-Gen UAV Networks" IEEE Networking Letters

5.1 Abstract

This chapter addresses a critical challenge in the context of 6G and beyond wireless networks, the joint optimization of power and bandwidth resource allocation for aerial intelligent platforms, specifically unmanned aerial vehicles (UAVs), operating in highly dynamic environments with mobile ground user equipment (UEs). We introduce FLARE (Flying Learning Agents for Resource Efficiency), a learning-enabled aerial intelligence framework that jointly optimizes UAV positioning, altitude, transmit power, and bandwidth allocation in real-time. To adapt to UE mobility, we employ Silhouette-based K-Means clustering, enabling dynamic grouping of users and UAVs' deployment at cluster centroids for efficient service delivery. The problem is modeled as a multi-agent control task, with bandwidth discretized into resource blocks and power treated as a continuous variable. To solve this, our proposed framework, FLARE, employs a hybrid reinforcement learning strategy that combines Multi-Agent Deep Deterministic Policy Gradient (MADDPG) and Deep Q-Network (DQN) to enhance learning efficiency. Simulation results demonstrate that our method significantly enhances user coverage, achieving a 73.45% improvement in the number of served users under a 5 Mbps data rate constraint, outperforming MADDPG baseline.

5.2 Introduction

Unmanned Aerial Vehicles (UAVs) are increasingly utilized in next-generation wireless communication systems due to their inherent mobility, altitude control, and rapid deployment capabilities. These features make UAVs well-suited to provide on-demand connectivity for mobile and heterogeneous user equipment (UE) with dynamic service requirements [59]. However, effective UAV placement, along with joint power and bandwidth allocation to improve user coverage, remains a critical challenge due to interference and fading effects.

In this letter, we propose a mobility-aware framework that jointly optimizes UAV positioning, altitude, and resource allocation through clustering and a hybrid Multi-Agent Deep Reinforcement Learning (MADRL) approach to maximize the number of served users. The key contributions of this work are as follows:

- Integration of the Spatio-Temporal Parametric Stepping (STEP) model to capture continuous UE mobility across time frames.
- Development of a Silhouette-based K-Means clustering technique for adaptive UAV-UE association under dynamic mobility patterns.
- Design of a hybrid MADDPG+DQN framework with mixed activation functions to enable continuous control of UAV altitude and power allocation, alongside discrete control of bandwidth resource allocation.

5.3 Related Work

To address the complexities of UAV-based communication networks, recent studies have explored the use of Multi-Agent Deep Reinforcement Learning (MADRL), particularly the MADDPG algorithm. For instance, the MADDPG-M&L (MADDPG based on Matching Game and Lagrangian Dual) approach proposed in [75] investigates UAV-assisted user association and slicing resource allocation in heterogeneous networks. In the context of Mobile Edge Computing (MEC), [76] addresses the problem of joint service placement and task offloading in air–ground integrated networks (AGINs), integrating service instance replacement and offload decision coupling.

Other related works include [77] and [78], which leverage MADDPG for UAV dynamic position planning and task offloading. The UAV utility maximization strategy in [62]

utilizes a DQN-DDPG joint bandwidth-power optimization scheme. In emergency scenarios, [79] proposes a greedy reinforcement learning method to improve fairness. Meanwhile, [80] introduces a machine learning-driven vibration analysis framework for UAV condition monitoring, optimizing communication through data aggregation and dimensionality reduction.

UAV placement and backhaul optimization are studied in [81], while [82] presents a joint trajectory and resource allocation framework that includes admission control under fixed user data demands. The work in [83] also tackles joint trajectory and resource allocation for UAVs serving mobile users in vehicular networks via problem decomposition and iterative optimization.

However, these works do not jointly consider UAV and user mobility alongside power and bandwidth resource constraints to enhance user coverage—an integrated perspective that our work aims to address.

5.4 System Model

We consider a multi-UAV-assisted communication system, where N ground UEs are initially distributed uniformly across a two-dimensional square area Ψ, partitioned into a 100×100 grid with each cell measuring l meters. Let I denote the set of UEs and we model UE mobility with the STEP model over F time frames. At each time frame, UEs' positions are updated, K-means clustering (with K chosen via the silhouette score) defines clusters, and UAVs are positioned at the centroids of clusters. The clustering method and UAV deployment strategy are detailed in the subsequent section. As depicted in Fig. 5.2, the scenario involves multiple UAVs, each serving a dynamically formed cluster of UEs.

The optimization is carried out frame-by-frame over F frames, with updated UE locations and clustering at each frame $t \in \{1, 2, \ldots, F\}$. Each UE is identified by $i \in I \triangleq 1, 2, \ldots, N$, with coordinates $(x_i^t, y_i^t, 0)$ representing their positions at frame t. Let $\mathcal{J}^t$ denote the set of active UAVs at frame t. UAV, $j \in \mathcal{J}^t$, is located at (x_j^t, y_j^t, h_j^t) in 3D space. Considering a UE i at a horizontal distance $d_{i,j}^t \triangleq \sqrt{(x_j^t - x_i^t)^2 + (y_j^t - y_i^t)^2}$ from its associated UAV j, the Euclidean distance between them is $r_{i,j}^t = \sqrt{{d_{i,j}^t}^2 + {h_j^t}^2}$, and the elevation angle is $\theta_{i,j}^t = \sin^{-1}(h_j^t / r_{i,j}^t)$. All UEs and UAVs use a single antenna.

We adopt an air-to-ground channel model with LoS and NLoS link components [55]. LoS and NLoS links experience Rician and Rayleigh fading, respectively. In NLoS links,

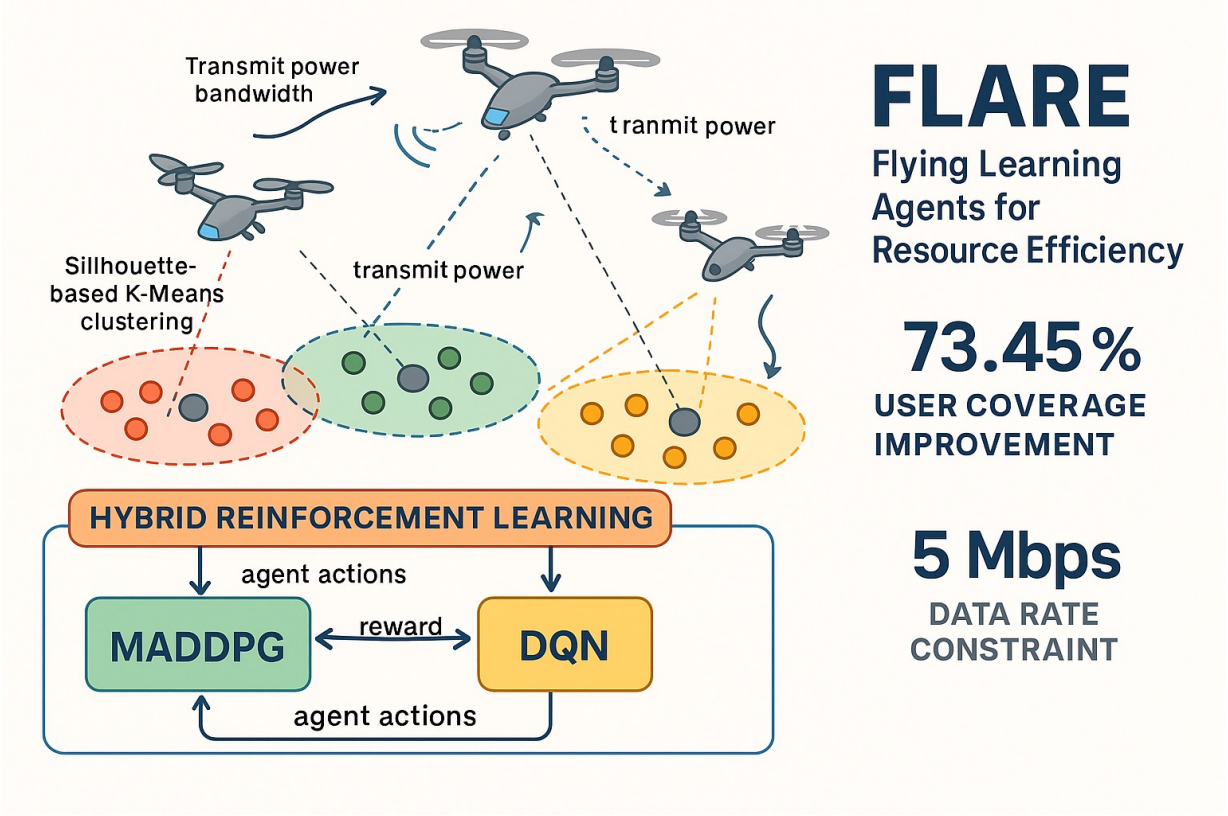

Figure 5.1: Problem Introduction

path loss exponent α_{NLoS} is higher than α_{LoS} due to increased attenuation from shadowing and reflections. The received power at UE i for LoS and NLoS links from UAV j is:

$$\mathrm{P}^{t}_{\mathrm{LoS}_{i,j}} = P^{t}_{i,j} g_{i,j} {r^{t}_{i,j}}^{-\alpha_{LoS}}, \tag{5.1}$$

$$\mathrm{P}^{t}_{\mathrm{NLoS}_{i,j}} = P^{t}_{i,j} k_{i,j} {r^{t}_{i,j}}^{-\alpha_{NLoS}}, \tag{5.2}$$

where $P^{t}_{i,j}$ is the transmit power allocated from UAV j to UE i, $g_{i,j}$ is the dynamic rician fading factor and $k_{i,j}$ is the dynamic rayleigh fading factor. The LoS link probability between UE i and UAV j is given as [55]:

$$p^{t}_{\mathrm{LoS}_{i,j}} = \frac{1}{1 + c\exp(-b[(180/\pi)\theta^{t}_{i,j} - c])}, \tag{5.3}$$

with b and c being environment-dependent constants. The effective received power be-

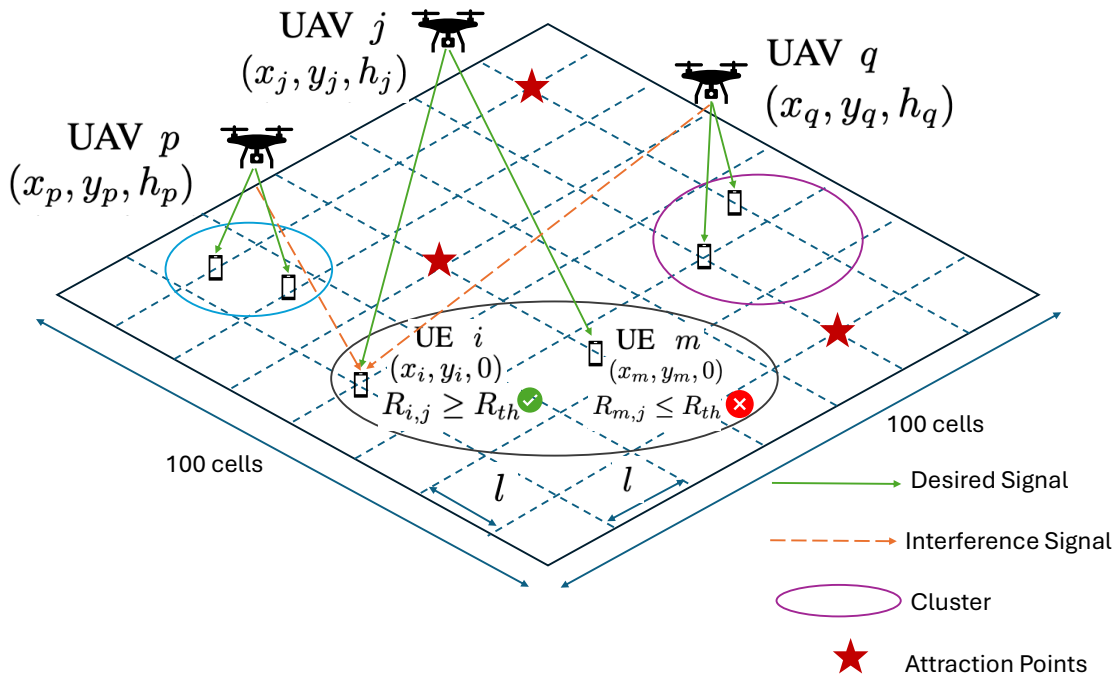

Figure 5.2: System Diagram

comes:

$$\mathrm{P}^t_{\mathrm{eff}_{i,j}} = p^t_{\mathrm{LoS}_{i,j}} \cdot \mathrm{P}^t_{\mathrm{LoS}_{i,j}} + (1 - p^t_{\mathrm{LoS}_{i,j}}) \cdot \mathrm{P}^t_{\mathrm{NLoS}_{i,j}}. \tag{5.4}$$

Considering inter-UAV interference and using Shannon's capacity formula, the achievable data rate of UE i from UAV j is given by:

$$R^t_{i,j} = B^t_{i,j} \log_2 \left(1 + \frac{\mathrm{P}^t_{\mathrm{eff}_{i,j}}}{I^t_{i,j} + N_o} \right), \tag{5.5}$$

where $B^t_{i,j}$ denotes the bandwidth allocated to UE i, N_o is the additive white Gaussian noise (AWGN) power and $I^t_{i,j} = \sum_{s \neq j} P^t_{s,\mathrm{avg}} k_{i,s} {r^t_{i,s}}^{-\alpha_{NLoS}}$ is the interference from other UAVs, assuming only NLoS links contribute to interference. UE i is considered served if $R^t_{i,j} \geq R_{th}$, where R_{th} denotes the target data rate threshold. To quantify the number of served users, we define a binary variable $a^t_{i,j}$ as follows:

$$a^t_{i,j} = \begin{cases} 1, & \text{if UE } i \text{ is served by UAV } j \text{ at frame } t, \\ 0, & \text{otherwise.} \end{cases} \tag{5.6}$$

5.5 Problem Formulation

Our goal is to maximize the number of served UEs over the entire mobility duration by dynamically adjusting UAV positions, altitudes, and resource allocations (power and bandwidth). Let $N_j^t = \sum_{i \in I} a_{i,j}^t$ denote the number of UEs served by UAV j at frame t. The optimization problem for frame t is formulated as:

$$
\begin{aligned}
(\mathcal{P}): & \max_{(P_j^t, B_j^t, h_j^t, \mathbf{p}_j^t)} \sum_{j \in \mathcal{J}^t} N_j^t, \\
\text{s.t.: } & (C1): R_{i,j}^t \mathbb{1}(a_{i,j}^t = 1) \geq R_{th}, \quad \forall\, i \in I, \forall\, j \in \mathcal{J}^t \\
& (C2): \sum_{j \in \mathcal{J}^t} a_{i,j}^t \leq 1, \quad \forall\, i \in I \\
& (C3): \ a_{i,j}^t \in \{0,1\}, \quad \forall\, i \in I, \forall\, j \in \mathcal{J}^t \\
& (C4): \sum_{l=1}^{N_j^t} P_{l,j}^t \leq P_{max}; \quad \forall\, j \in \mathcal{J}^t \\
& (C5): \sum_{l=1}^{N_j^t} B_{l,j}^t \leq B_{max}; \quad \forall\, j \in \mathcal{J}^t \\
& (C6): \mathbf{p}_j^t = (x_j^t, y_j^t) \in \Psi, \quad \forall\, j \in \mathcal{J}^t \\
& (C7): h_{min} \leq h_j^t \leq h_{max}, \quad \forall\, j \in \mathcal{J}^t.
\end{aligned}
\tag{5.7}
$$

Constraint (C1) ensures that each served UE meets the required rate. (C2) ensures each UE is served by at most one UAV. (C3) enforces binary assignment. (C4) and (C5) restrict UAV transmission power and bandwidth under certain power and bandwidth budgets. (C6) and (C7) ensure UAVs stay within the field and height range. Due to the dynamic, non-convex nature of this problem and UE mobility, we propose a joint MADRL approach and silhouette-based K-means clustering to adapt UAV positioning, altitude, and resource allocation over time.

5.6 Spatio-Temporal Parametric Stepping (STEPS): A Mobility Model for UEs

5.6.1 Introduction

Modeling user equipment (UE) mobility is a critical component in the design and optimization of next-generation wireless networks, particularly in scenarios involving dynamic resource allocation, UAV coordination, and real-time service delivery. Accurate modeling of Spatio-Temporal Parametric Stepping (STEPS) enables improved forecasting of user locations, which is essential for efficient coverage planning, handover management, and predictive task allocation in UAV-assisted networks.

Traditional mobility models, such as random walk or Markovian processes, often fall short in capturing the structured, environment-aware behavior exhibited by users in real-world networks. To address these limitations, data-driven spatio-temporal models have emerged as powerful tools that leverage historical movement data and contextual cues to predict future mobility trends.

In this work, we adopt and extend STEPS framework to simulate and analyze UE mobility over a discretized spatial grid. By incorporating attraction points and probabilistic decision-making mechanisms, our model reflects both goal-oriented behavior and stochastic variability observed in mobile users. This hybrid structure ensures that the generated mobility traces are not only diverse but also semantically grounded, making them well-suited for downstream tasks such as UAV clustering, trajectory prediction, and reinforcement learning-based coordination.

The proposed simulator generates rich datasets of user trajectories under varying conditions, supporting robust benchmarking and reproducible experimentation. Ultimately, this mobility modeling framework lays the foundation for developing intelligent, context-aware wireless network management strategies.

5.6.2 Related Work

To improve the modeling of high-dimensional temporal sequences in structured prediction tasks, Zand et al. [84] propose MotionFlow, a conditional normalizing flows (CNF) framework tailored for spatio-temporal structured prediction. Unlike conventional generative models based on VAEs or GANs that suffer from training instability and mode collapse, MotionFlow leverages the tractable, invertible transformations of normalizing

flows to model complex trajectory and motion dynamics. By integrating masked convolutions for autoregressive conditioning and factorized latent variables for temporal modeling, their method accurately captures spatial and temporal dependencies in data such as human motion, time series, and multi-agent trajectories. Notably, it achieves state-of-the-art performance on diverse tasks—including motion prediction on CMU Mocap and trajectory forecasting on the NBA dataset—demonstrating its generalizability and robustness in structured spatio-temporal learning, which makes it highly relevant to UAV trajectory optimization and future mobility prediction frameworks.

To enable fine-grained spatio-temporal forecasting of user traffic in cellular networks, Yu et al. [85] proposed STEP, a deep learning-based prediction framework that leverages Graph Convolutional Gated Recurrent Networks (GCGRN) to jointly model users' mobility and traffic patterns. By capturing base station handoffs and temporal traffic correlations, STEP achieves high-accuracy traffic predictions with low device-side overhead, making it a valuable reference for mobility-aware resource allocation strategies in UAV-assisted wireless networks.

Yang et al. [86] introduce a probabilistic reasoning framework for identifying unique roles in dynamic environments by fusing semantic-interaction and spatio-temporal features. Their method leverages two observation models—Object Existence Model (OEM) and Human Action Model (HAM)—to infer an individual's role based on object proximity and motion behavior, without relying on pre-defined semantic relationships. The hierarchical graphical model they propose enables robust role recognition even under clutter, occlusion, or uncertainty. This work informs our approach to fusing environmental cues and user behavior patterns for UAV task prioritization and adaptive mission planning.

An et al. [87] introduce STMPNet, a novel spatio-temporal multivariate probabilistic model for traffic prediction that effectively captures complex dependencies across space, time, and multiple traffic attributes. By integrating a spatio-temporal fusion graph block and a copula-based joint distribution estimator, their model enables both forecasting and interpolation tasks under conditions of missing or non-uniformly sampled data. This probabilistic approach provides not only accurate predictions but also quantifies uncertainty, which is essential for robust decision-making in intelligent transportation systems.

Zheng et al. [88] propose **STDiff**, a novel conditional denoising diffusion framework for probabilistic spatio-temporal traffic forecasting. Instead of relying on static spatial graphs or complex adaptive structures, their model employs spatio-temporal attention mechanisms and discrete wavelet decomposition to separately model low-frequency trends and high-frequency fluctuations. By leveraging a simple MLP-based conditional network and diffusion processes, STDiff achieves state-of-the-art performance across multiple real-

world traffic datasets. This framework highlights the effectiveness of diffusion-based models in capturing uncertainty and temporal dynamics, which could be beneficial for UAV-based predictive mobility and resource planning tasks.

5.6.3 Implementation

We adopt the STEPS model [89] to govern the mobility behaviour of N UEs, which are initially distributed uniformly over a bounded 100×100 grid Ψ. At each time frame $t = 1, \ldots, F$, the position of UE i is represented as

$$p_i^t = (p_{x_i}^t, p_{y_i}^t) \in \Psi, \quad 0 \leq p_{x_i}^t, p_{y_i}^t \leq 99, \tag{5.8}$$

where UEs are constrained to move only to one of the four adjacent grid cells, which means $p_i^{t+1} - p_i^t \in \{(\pm 1, 0), (0, \pm 1)\}$. Movement decisions are influenced by a predefined set of K attraction points $\mathcal{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_K\} \subset \Psi$. At each time frame, UE selects its nearest attraction point $\mathbf{a}_K \in \mathcal{A}$, and then identifies the valid move $\mathcal{M}$, minimizing $\|\mathbf{a}_k - p_i^{t+1}\|^2$. With a fixed probability $p \in [0, 1]$, UE selects the move $\mathcal{M}$; otherwise, it randomly chooses one of the four adjacent directions from $\{(\pm 1, 0), (0, \pm 1)\}$ at uniform probability. To prevent collisions, if the selected cell is already occupied during the same frame, the move is resampled until an unoccupied cell is found. The resulting position is then recorded as p_i^{t+1}. Note that physical coordinates are obtained by scaling the grid position via $(x_i^t, y_i^t) = (p_{x_i}^t \times l,\ p_{y_i}^t \times l)$, where l denotes the cell size of one grid unit.

5.7 Clustering and UAV Positioning

5.7.1 Introduction

To ensure effective UAV-user association and maximize coverage performance, we design a clustering framework that adaptively determines the optimal number of clusters at each simulation frame. The output of this clustering stage directly informs UAV deployment by assigning each UAV to the centroid of a user group. Our objective is to dynamically minimize intra-cluster user dispersion while avoiding coverage voids caused by under-clustering and resource inefficiencies induced by over-clustering.

K-Means is a well-established clustering algorithm widely used for its simplicity, scalability, and geometric interpretability. It partitions a set of observations into k clusters by minimizing the sum of squared distances between each point and its nearest cluster center.

Given a set of user locations $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^2$, K-Means solves the following optimization problem:

$$\min_{\{\mu_i\}_{i=1}^k} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in C_i} \|\mathbf{x}_j - \mu_i\|^2, \tag{5.9}$$

where μ_i is the centroid of cluster C_i. The algorithm iteratively assigns each point to the nearest centroid and updates the centroids until convergence.

In our system, K-Means is employed to group users spatially within each frame. The resulting centroids serve as candidate UAV deployment points. Compared to density-based (e.g., DBSCAN) or probabilistic (e.g., GMM) alternatives, K-Means is favored for its deterministic behavior, efficiency, and suitability for frame-wise, real-time UAV control. Moreover, we adaptively determine k using silhouette-based validation to ensure that the number of UAVs deployed is both sufficient and efficient, minimizing user dispersion while avoiding overprovisioning.

5.7.2 Related Work

To address the limitations of traditional K-Means clustering in handling non-convex and high-dimensional datasets, Zhu et al. [90] proposed DH-KMeans, which enhances the K-Means framework by dynamically determining initial cluster centers based on local density and inter-cluster distances. This dual-stage approach, integrating Agglomerative Hierarchical Clustering (AHC), improves clustering precision and is well suited for adaptive UAV-user clustering under dynamic conditions.

Yu et al. [91] introduced Trust Cop-KMeans, a trust-aware variant that integrates social trust relationships into clustering via must-link and cannot-link constraints. This is particularly relevant in decentralized UAV networks where interaction history can influence coordination, making trust-aware mechanisms beneficial for spectrum sharing and collaboration.

Li et al. [92] apply K-Means to electricity consumption data to evaluate rural socio-economic development. Their work builds a multi-factor clustering model informed by key energy indicators, offering insights into region classification and adaptive UAV resource planning using context-aware clustering strategies.

Sun et al. [93] propose E-C-KMeans, a hybrid algorithm integrating entropy-based weighting and density peaks into K-Means to enhance scenario reduction in power distribution systems. Their method supports reliable scheduling in the presence of renewable

energy uncertainties, which motivates its potential adaptation to spatio-temporal UAV operations under uncertainty.

In image processing, El Rube' [94] combines Progressive Histogram Quantization (PHQ) with K-Means for color reduction. PHQ reduces dimensionality by merging histogram bins prior to clustering, enabling faster execution with comparable quality. This hybrid approach inspires latency-aware K-Means deployment for UAV visual systems.

A comprehensive evaluation by Raghu et al. [95] compares clustering techniques—K-Means, NMF, Spectral Clustering, and GMM—on high-dimensional gene expression data. K-Means shows robust performance across several validation metrics, reaffirming its utility in high-dimensional feature spaces relevant to UAV-user association tasks.

Applications Across Domains

K-Means and its extensions have found widespread application across multiple disciplines:

- In power systems, it assists in clustering rural energy usage for development planning [92].
- In hybrid AC/DC grids, energy-constrained variants such as E-C-KMeans support multi-time-scale optimization [93].
- In image processing, K-Means aids in color quantization and histogram binning [94].
- In bioinformatics, hybrid methods integrating K-Means with spectral and NMF-based clustering enhance robustness in gene expression data analysis [95].

These use cases showcase K-Means' flexibility in probabilistic, spatio-temporal, and constraint-based contexts, aligning well with the needs of dynamic UAV-assisted networks.

5.7.3 Silhouette-Based Cluster Selection

At each frame t, are grouped using the K-means clustering algorithm. The clustering quality is evaluated using the silhouette score:

$$s_i^t = (b_i^t - a_i^t)/\max\{a_i^t, b_i^t\}, \tag{5.10}$$

where a_i^t and b_i^t denote the average intra-cluster and nearest inter-cluster distances for UE i at frame t, respectively. The optimal number of clusters is determined as:

$$k^{t*} = \arg \max_{k \in \{2,\ldots,k_{\max}\}} S_k^t, \tag{5.11}$$

where S_k^t is the average silhouette score across all UEs for k clusters. Once the optimal cluster count k^{t*} is identified, UEs are partitioned into k^{t*} clusters using the K-means algorithm. Each UAV j is then positioned at the centroid of its assigned cluster C_j^t, with the horizontal coordinates (x_j^t, y_j^t). The corresponding altitude h_j^t is optimized by Algo. 5.3, aiming to minimize the average UE–UAV distance while adapting to dynamic spatial distributions. UAVs dynamically track the centroids of their assigned clusters, while those not associated with any cluster enter sleep mode to conserve energy. Let $|^t$ denote the set of active UAVs at time frame t.The number of active UAVs, $|\mathcal{J}^t| = k^{t*}$, adapts at each time frame t according to the spatial distribution of UEs.

5.8 Joint Multi-Agent Resource Allocation Strategy

5.8.1 Introduction

The growing demand for flexible, intelligent, and rapid-deployment wireless communication infrastructure has made unmanned aerial vehicles (UAVs) a promising solution. Owing to their high mobility, adaptive altitude, and on-demand coverage capability, UAVs are being increasingly adopted as aerial base stations in next-generation wireless networks, particularly in scenarios involving emergency communication, rural access, or dense urban hotspots.

However, managing UAV-based wireless networks introduces several challenges. UAVs must dynamically adjust their 3D positions, transmission power, and bandwidth allocations in real-time to accommodate mobile ground users while minimizing interference and energy consumption. Traditional rule-based or optimization-based methods often fail to address these issues effectively due to the non-convexity of the system, the stochastic nature of wireless channels, and the high dimensionality introduced by multiple agents and user mobility.

In this context, reinforcement learning (RL) has emerged as a data-driven, model-free approach capable of learning adaptive control policies through interaction with the environment. Specifically, deep reinforcement learning (DRL) extends RL to high-dimensional

problems by leveraging deep neural networks for function approximation. DRL has demonstrated success in UAV applications such as trajectory planning, coverage control, and spectrum allocation.

Nonetheless, the unique structure of the UAV control problem—where both continuous (e.g., altitude, power) and discrete (e.g., bandwidth allocation) actions are involved—calls for a hybrid learning framework. Moreover, the multi-agent nature of UAV networks further complicates the learning process, as coordination among agents must be maintained in the face of partial observability and dynamic user behavior.

To address these challenges, we propose a hybrid learning architecture that combines Multi-Agent Deep Deterministic Policy Gradient (MADDPG) with Deep Q-Network (DQN). MADDPG enables fine-grained control over continuous action spaces such as UAV transmission power and flying altitude, while DQN handles discrete decisions like bandwidth allocation to users. By integrating these components in a unified framework, we enable each UAV to jointly optimize its physical placement and communication resource distribution in real time.

The remainder of this chapter details the structure, training process, and design of the proposed hybrid framework, including specific algorithmic formulations, reward engineering, and integration logic that supports convergence and stability in large-scale UAV-assisted communication systems.

5.8.2 Overview of Reinforcement Learning

Reinforcement Learning (RL) is a framework in which agents interact with an environment to learn optimal actions through trial and error. The agent receives feedback in the form of scalar rewards and uses this signal to refine its policy $\pi(a \mid s)$, which maps observed states s to actions a. The objective is to maximize the expected cumulative reward:

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right], \tag{5.12}$$

where $\gamma \in (0, 1]$ is the discount factor and r_t is the reward at time step t.

In UAV communication systems, RL is highly suited for sequential decision-making tasks such as mobility control, power allocation, and bandwidth scheduling. These decisions are often affected by dynamic and partially observable states—like user mobility, interference, and variable channel gains—which makes model-free RL particularly powerful.

5.8.3 Deep Q-Network

DQN is a model-free RL algorithm used for environments with discrete action spaces. It approximates the Q-value function $Q(s, a)$ with a deep neural network, updating weights based on the temporal-difference error:

$$L(\theta) = \mathbb{E}_{(s,a,r,s')} \left[\left(r + \gamma \max_{a'} Q(s', a'; \theta^-) - Q(s, a; \theta) \right)^2 \right], \tag{5.13}$$

where θ^- are target network parameters.

In our framework, DQN is employed to select discrete bandwidth allocation strategies. Each UAV agent chooses the number of resource blocks to assign to each user, ensuring the bandwidth is optimally distributed under discrete constraints.

5.8.4 Multi-Agent Deep Deterministic Policy Gradient

The Multi-Agent Deep Deterministic Policy Gradient (MADDPG) algorithm is an extension of the Deep Deterministic Policy Gradient (DDPG) method tailored for multi-agent systems operating in continuous action spaces. It is particularly well-suited for environments where agents must learn both cooperative and competitive behaviors under partial observability and non-stationary dynamics caused by the presence of other learning agents.

In MADDPG, each agent i maintains two core components: an actor network $\mu_i(s_i)$ that determines the agent's action based on its local observation s_i, and a critic network $Q_i(s, a_1, \ldots, a_N)$ that estimates the expected return given the global state and the joint actions of all agents. This design enables each agent to benefit from centralized training—where full environmental information and other agents' actions are available—while still allowing for decentralized execution during inference.

Key features of MADDPG include:

- **Centralized training and decentralized execution (CTDE):** Each agent is trained using a centralized critic that has access to the full global state and the actions of all agents. During deployment, each agent relies only on its local observation, ensuring scalability and real-world applicability.
- **Deterministic policy updates:** The actor is updated by following the gradient of the Q-value with respect to the action, improving stability and convergence in continuous action domains.

Algorithm 5.1 Deep Q-Network (DQN)

Input: Replay buffer $\mathcal{D}$, discount factor γ, learning rate α, target update rate τ

Output: Trained Q-network parameters θ

Initialize Q-network Q_θ with random weights Initialize target Q-network $Q_{\theta'} \leftarrow Q_\theta$ Initialize replay buffer $\mathcal{D}$

for *each episode* **do**

  Initialize environment and receive initial state s_0

  for *each timestep* t **do**

    Select action a_t using ϵ-greedy policy:

$$a_t = \begin{cases} \text{random action} & \text{with probability } \epsilon \\ \arg\max_a Q_\theta(s_t, a) & \text{otherwise} \end{cases} \tag{5.14}$$

    Execute a_t, observe reward r_t and next state s_{t+1} Store transition (s_t, a_t, r_t, s_{t+1}) in buffer $\mathcal{D}$

    Sample minibatch of transitions (s_j, a_j, r_j, s_{j+1}) from $\mathcal{D}$

    Compute target Q-value for each sample:

$$y_j = r_j + \gamma \max_{a'} Q_{\theta'}(s_{j+1}, a') \tag{5.15}$$

    Perform gradient descent on loss:

$$\mathcal{L} = \frac{1}{M} \sum_j \left(Q_\theta(s_j, a_j) - y_j \right)^2 \tag{5.16}$$

    Update target network:

$$\theta' \leftarrow \tau\theta + (1 - \tau)\theta' \tag{5.17}$$

  end

end

- **Shared experience replay buffer:** Transition tuples of the form (s, a, r, s') are stored in a replay buffer and sampled uniformly to break temporal correlations, which enhances learning stability.
- **Soft target updates:** Target networks for both actor and critic are updated slowly to ensure stable training, using the rule:

$$\theta^{\text{target}} \leftarrow \tau\theta + (1-\tau)\theta^{\text{target}}, \tag{5.18}$$

  where $\tau \ll 1$ is a smoothing coefficient.

In the proposed UAV-assisted communication framework, each UAV is modeled as an independent agent. The actions taken by each UAV include selecting a transmission power level and adjusting its flight altitude—both of which are continuous control variables. The use of MADDPG allows the agents to jointly learn optimal continuous control policies while coordinating implicitly through the centralized critic during training. This facilitates real-time, fine-grained adjustments that improve system performance metrics such as coverage, data rate, and energy efficiency.

5.8.5 Proposed Solution

Following UE clustering and UAV placement at the corresponding cluster centroids, we develop a hybrid MADRL frameowrk, FLARE, detailed in Algo. 5.3, to optimize each UAV's altitude h_j^t, per-user transmit power $\{P_{i,j}^t\}$, and bandwidth allocation $\{B_{i,j}^t\}$ at each time frame t. We propose a hybrid MADRL framework combining MADDPG for continuous control of UAV altitude and transmit power allocation, and DQN for discrete allocation of bandwidth resource blocks to UEs. The framework aims to maximize the number of UEs served while satisfying power and bandwidth constraints at each time frame t.

Hybrid Action with Mixed Activation

To support hybrid actions, the MADDPG actor outputs altitude via `tanh` and per-user power allocation via `softmax`, ensuring bounded and normalized decisions. This separation avoids constraint violations without penalty terms. UAVs observe local states based on users grouped in its associated C_j^t, and training uses Gaussian exploration noise.

Algorithm 5.2 Multi-Agent Deep Deterministic Policy Gradient (MADDPG)

Input: Number of agents N, actor networks $\{\mu_{\theta_i}\}$, critic networks $\{Q_{\phi_i}\}$, experience buffer $\mathcal{D}$

Output: Trained actor and critic parameters $\{\theta_i, \phi_i\}$

Initialize actor μ_{θ_i} and critic Q_{ϕ_i} networks for each agent i Initialize target networks $\mu_{\theta'_i} \leftarrow \mu_{\theta_i}, Q_{\phi'_i} \leftarrow Q_{\phi_i}$ Initialize replay buffer $\mathcal{D}$

for *each episode* **do**

  Initialize environment and obtain initial state $\mathbf{s}_0 = (s_1, ..., s_N)$

  for *each time step* t **do**

    for *each agent* i **do**

      Select action $a_i = \mu_{\theta_i}(s_i) + \mathcal{N}_t$ with exploration noise $\mathcal{N}_t$

    Execute joint action $\mathbf{a} = (a_1, ..., a_N)$, observe reward $r = (r_1, ..., r_N)$ and next state $\mathbf{s}'$

    Store $(\mathbf{s}, \mathbf{a}, \mathbf{r}, \mathbf{s}')$ in buffer $\mathcal{D}$

    if *time to update* **then**

      Sample minibatch of transitions from $\mathcal{D}$

      for *each agent* i **do**

        Compute target actions $a'_j = \mu_{\theta'_j}(s'_j)$ for all j

        Compute target $y_i = r_i + \gamma Q_{\phi'_i}(\mathbf{s}', \mathbf{a}')$

        Update critic by minimizing loss:

$$\mathcal{L}_i = \frac{1}{M} \sum (Q_{\phi_i}(\mathbf{s}, \mathbf{a}) - y_i)^2 \tag{5.19}$$

        Update actor using the policy gradient:

$$\nabla_{\theta_i} J \approx \frac{1}{M} \sum \nabla_{a_i} Q_{\phi_i}(\mathbf{s}, \mathbf{a}) \nabla_{\theta_i} \mu_{\theta_i}(s_i) \tag{5.20}$$

        Soft update target networks:

$$\theta'_i \leftarrow \tau\theta_i + (1-\tau)\theta'_i \quad \phi'_i \leftarrow \tau\phi_i + (1-\tau)\phi'_i \tag{5.21}$$

    $\mathbf{s} \leftarrow \mathbf{s}'$

MADDPG for UAV altitude and transmit power allocation

Each UAV acts as an agent with:

- **State:** Altitude h_j^t, power $\{P_{i,j}^t\}$, bandwidth $\{B_{i,j}^t\}$, and number of served UEs.
- **Action:** Continuous values for $h_j^t \in [h_{\min}, h_{\max}]$, $P_{i,j}^t \in [0, P_{\max}]$.
- **Reward:** Number of UEs achieving target data rate.

DQN for bandwidth resource block allocation

Each UAV integrates a DQN to determine the minimum bandwidth required for each associated UE to satisfy its data rate requirement. Bandwidth is allocated in discrete steps, and once a user is successfully served, its allocation is fixed to enable efficient utilization of the remaining resources.

- **State:** $\{P_{i,j}^t, B_{i,j}^t\}$ for all associated UEs.
- **Action:** Increment or decrement bandwidth resource block.
- **Reward:** 1 if user meets data rate; else 0.

The flow of Algo. 5.3 is explained as follows: Line 1 performs preprocessing; lines 2–3 initialize each UAV's MADDPG; lines 4–5 enter the episode/timestep loops; lines 6–7 compute altitude and power; lines 8–12 run per-UE DQNs to select bandwidth; line 13 execute actions and record transitions; lines 14–19 update MADDPG and DQNs with soft targets; line 20 advance the state.

Figure 5.3 illustrates the overall structure of the proposed Hybrid MADDPG+DQN framework. The framework integrates a multi-agent deterministic policy gradient algorithm (MADDPG) for handling continuous control variables—such as altitude and transmission power—and a Deep Q-Network (DQN) for optimizing discrete actions like bandwidth allocation.

In the top-left subfigure, the architecture adopts a centralized training and decentralized execution (CTDE) paradigm. Each agent observes its local state and executes actions via its own actor policy π_j. These policies are trained using sampled experiences from a shared replay buffer, and the critics Q_j are updated based on global state and joint actions during

training. The agents interact with the environment independently, and their collected experiences are aggregated to optimize the shared reward.

The top-right panel shows the neural architecture of the actor networks. The policy network receives the agent-specific state vector and maps it through multiple hidden layers to generate continuous action outputs. Mixed activation functions (Tanh and Softmax), combined with stochastic noise layers, are used to model both bounded and discrete-continuous hybrid control signals.

The bottom-right section illustrates the step function of the DQN module. It receives environmental observations and uses a neural Q-function to output the optimal bandwidth allocation action. The DQN interacts with both environmental state and MADDPG outputs (e.g., altitude, power) to guide discrete decision-making. The reward signal from the environment is used to update both continuous and discrete learners simultaneously.

This hybrid design ensures effective division of control responsibilities: MADDPG agents manage high-precision continuous control, while the DQN module handles quantized resource allocation, resulting in a scalable and flexible UAV control framework suitable for complex, dynamic wireless environments.

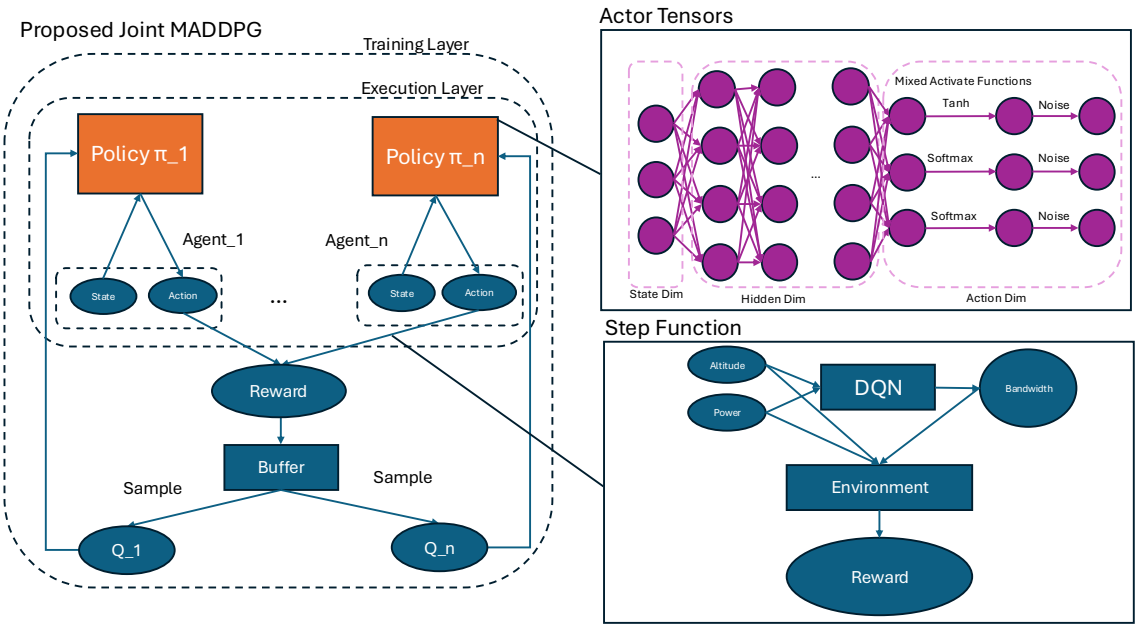

Figure 5.3: Joint Algorithm

5.9 Numerical Results

5.9.1 Simulation Setup

We evaluate the proposed hybrid MADRL framework, via simulations in a realistic urban environment. Table 5.1 summarizes the environmental and system settings for a dense urban scenario, while Table 5.2 lists the MADRL training hyperparameters chosen for stable convergence. UE mobility is generated by the STEP model (Fig. 5.4), with user trajectories and attraction points; Fig. 5.5 plots the mean optimal number of clusters over ten random seeds per frame t.

Users' initial positions are uniformly distributed across the entire field to simulate random initial states without any bias.

In our simulations, the path loss exponent is set to 3 for LoS links and 4 for NLoS links. These values are adopted in the literature [57], where LoS propagation typically exhibits an exponent between 2 and 3, while NLoS conditions experience more severe attenuation with exponents ranging from 3 to 4.

Table 5.1: Environmental Parameters Used in the Simulations

Symbol	Description	Value
$h_{\min}$, $h_{\max}$	Min/Max UAV Altitude	300 m, 1000 m
l	Cell Size	300 m
p	Attraction Probability	0.4
K	Number of Attraction Points	3
N	Total Number of UEs	30
$k_{\max}$	Max Number of Clusters/Agents/UAVs	5
$P_{\max}$	Total Power of each UAV	1 W
$B_{\max}$	Bandwidth of each UAV	3.6 MHz
$Block_{size}$	Bandwidth size of each block	18 kHz
$Block_{max}$	Block Limit of each UAV	200
N_o	Noise Power	4×10^{-15} W
α_{LoS}, α_{NLoS}	Path Loss Exponent for LoS/NLoS	3, 4
c, b	Environmental Constant (Dense Urban)	11.95, 0.136

Table 5.2: Model Training Hyperparameters

Description	Value
Replay Buffer Size	$100,000$
Batch Size	512
Update Steps	2500
Steps per Episode	500
Episodes	100
Actor Network Learning Rate	0.0001
Critic Network Learning Rate	0.0001
Discount Factor for Future Rewards	0.99
Target Network Update Rate	0.01
Hidden Layer Sizes for Actor and Critic	$[64, 64]$

5.9.2 Results and Discussion

Fig. 5.6 illustrates the convergence behavior of the DQN agent during bandwidth block selection. Each episode represents a new attempt to identify the optimal bandwidth allocation, with the objective of minimizing the search time. As training progresses, the agent rapidly converges to optimal solutions, demonstrating improved efficiency over successive episodes. Fig. 5.7 demonstrates the stable training behavior of the MADDPG framework under varying data rate thresholds. Following the initial exploration phase using samples from the replay buffer, the average reward increases consistently, indicating effective learning over time.

Fig.5.8 and Fig.5.9 compare the number of served UEs with target data rates of 5 Mbps and 7.5 Mbps, respectively, under three different methods: our proposed hybrid MADRL framework, a baseline MADDPG method, and a static scheme with equal resource allocation and fixed UAV altitude. Across both scenarios, our solution demonstrates superior performance by consistently serving a higher number of UEs, enhancing user coverage by 73.45% under 5 Mbps and almost 2 times under 7.5 Mbps compared to MADDPG baseline.

5.10 Conclusion

In this letter, we have presented a novel framework, to jointly optimize UAV positioning, altitude, power, and discrete bandwidth allocation in dynamic mobile environments

for user coverage maximization. Numerical Results confirm that our approach, leveraging Silhouette-based K-Mean clustering for adaptive UAV-UE association and hybrid MADDPG+DQN framework, outperforms baseline, MADDPG, enhancing user coverage by 73.45% for 5Mbps data rate threshold. This highlights its promise for real-time UAV resource management. Future work will address security aspects and heterogeneous UE mobility patterns, advancing AI-native aerial access for 6G networks.

Algorithm 5.3 Hybrid MADDPG+DQN framework for UAV altitude control and resource allocation to UEs

Input: Set of UAVs and UEs, $\mathcal{J}$ and I, respectively, No. of episodes E, Timesteps T, $P_{\max}$, $B_{\max}$

Output: Optimized $(h_j^t, \{P_{i,j}^t\}, \{B_{i,j}^t\})$ for each UAV $j \in \mathcal{J}^t$, and each UE $i \in I$

Preprocessing: At each time frame t, update UEs' positions via STEP model, cluster UEs with silhouette-based K-means, and assign UAVs to each cluster, placing UAV at their centroids

foreach *UAV agent* $j \in \mathcal{J}^t$ **do**

  Initialize MADDPG actor μ_j, critic Q_j, targets μ'_j, Q'_j, replay buffer $\mathcal{D}_{\text{MADDPG}}$ and local state s_j **for** *episode* $e = 1$ **to** E **do**

    for *timestep* $t = 1$ **to** T **do**

      $h_j^t \leftarrow \tanh\big(\mu_j^{(h)}(s_j)\big)$ $\{P_{i,j}^t\} \leftarrow \text{softmax}\big(\mu_j^{(p)}(s_j)\big)$

      foreach *UE* $i \in C_j^t$ **do**

        Initialize DQN $Q_{i,j}$, target $Q'_{i,j}$ for all $i \in C_j^t$, replay buffer $\mathcal{D}_{\text{DQN}}$ and state $s_{i,j}^{\text{DQN}}$ $s_{i,j}^{\text{DQN}} \leftarrow \phi\big(s_j,\, h_j^t,\, P_{i,j}^t\big)$ $B_{i,j}^t \leftarrow \epsilon\text{-greedy}\big(Q_{i,j},\, s_{i,j}^{\text{DQN}}\big)$ Apply the action $B_{i,j}^t$ to UE i and store the experience for $\mathcal{D}_{\text{DQN}}$

      Apply joint action $(h_j^t, \{P_{i,j}^t\}, \{B_{i,j}^t\})$ to environment and get observe reward r_j, next MADDPG state s'_j and store $(s_j, h_j^t, \mathbf{P}_j^t, r_j, s'_j)$ in $\mathcal{D}_{\text{MADDPG}}$

      if t mod *update_interval* $= 0$ **then**

        Sample minibatch from $\mathcal{D}_{\text{MADDPG}}$ Update μ_j, Q_j via gradient steps; $\mu'_j \leftarrow \tau\mu_j + (1-\tau)\mu'_j$, $Q'_j \leftarrow \tau Q_j + (1-\tau)Q'_j$

        foreach *user* $i \in C_j$ **do**

          Sample minibatch from $\mathcal{D}_{\text{DQN}}$ Update $Q_{i,j}$ via TD-loss; Soft-update $Q'_{i,j} \leftarrow \tau Q_{i,j} + (1-\tau)Q'_{i,j}$

      $s_j \leftarrow s'_j$

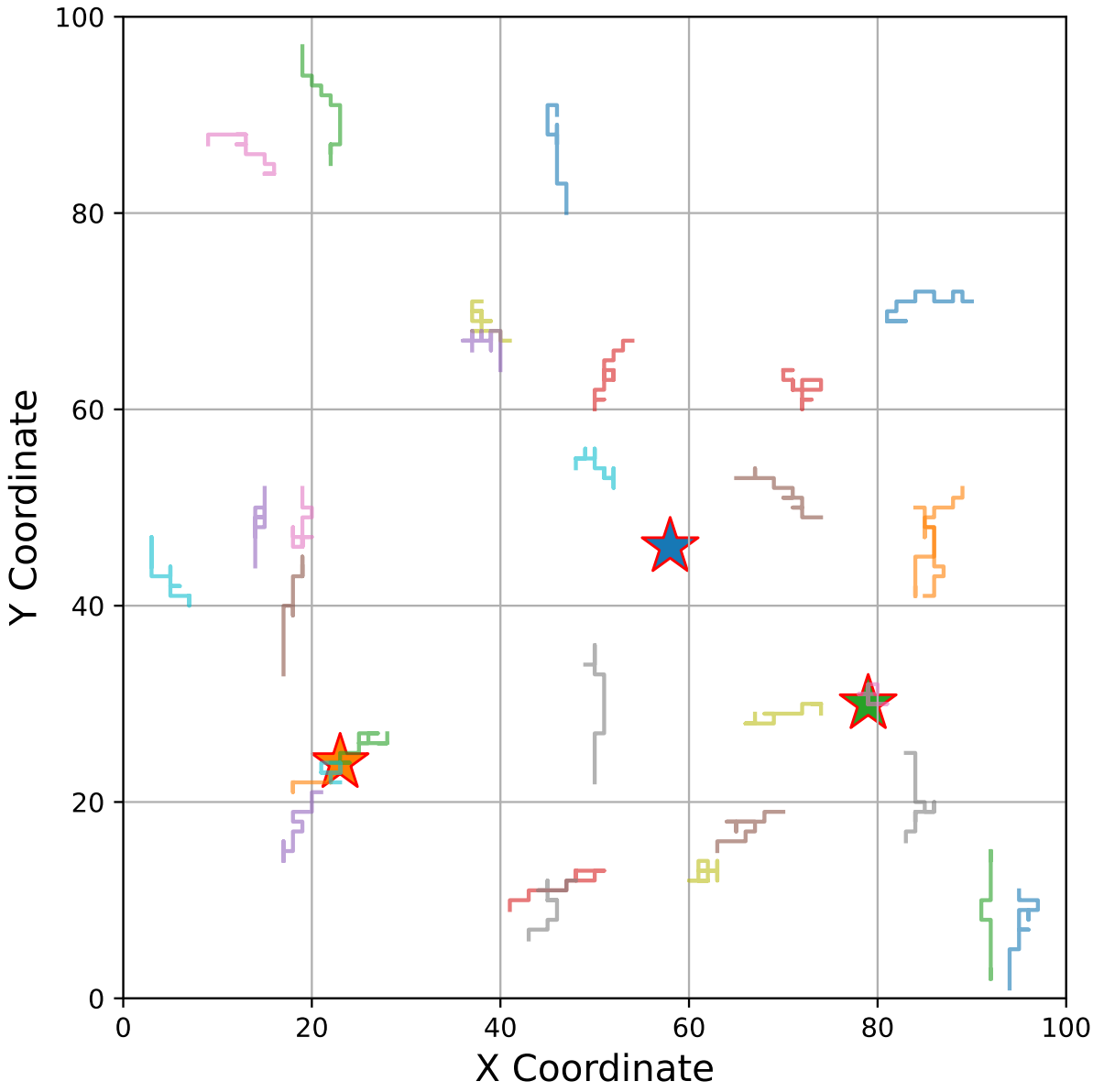

Figure 5.4: Mobility paths

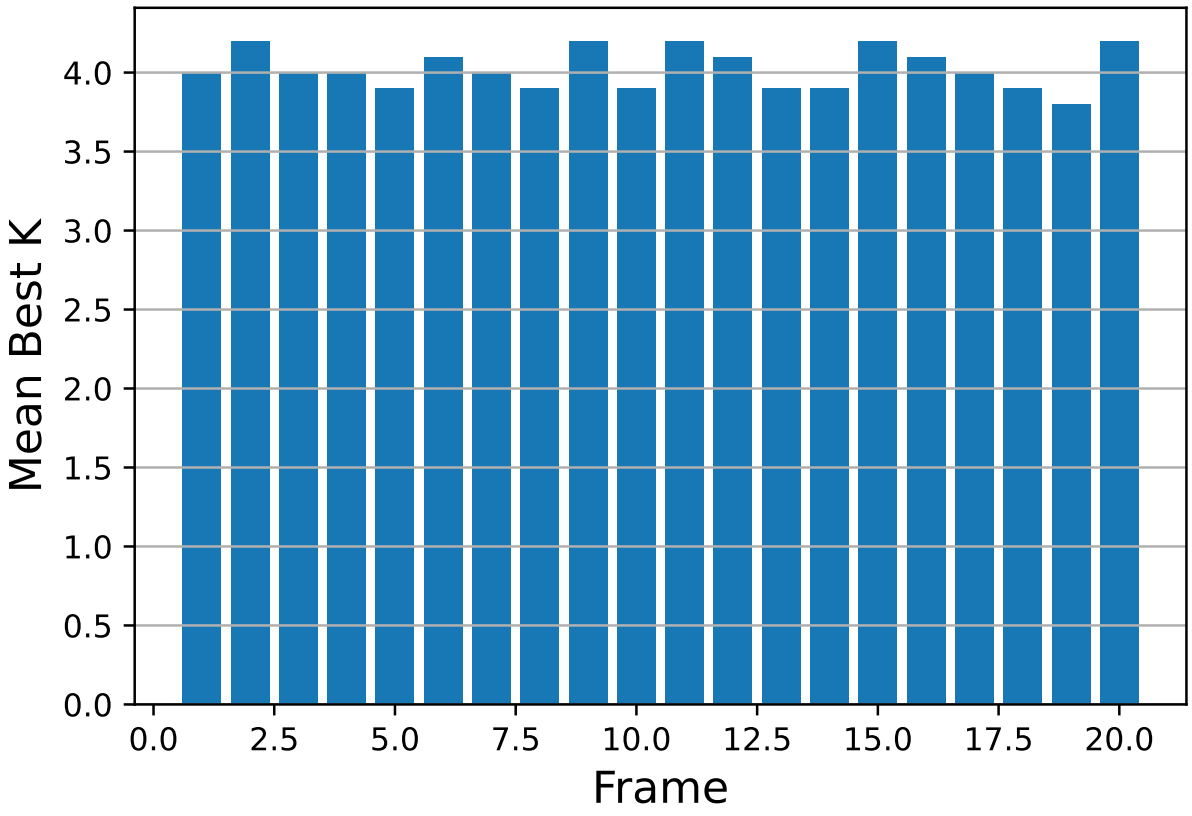

Figure 5.5: Optimal cluster number based on Silhouette score

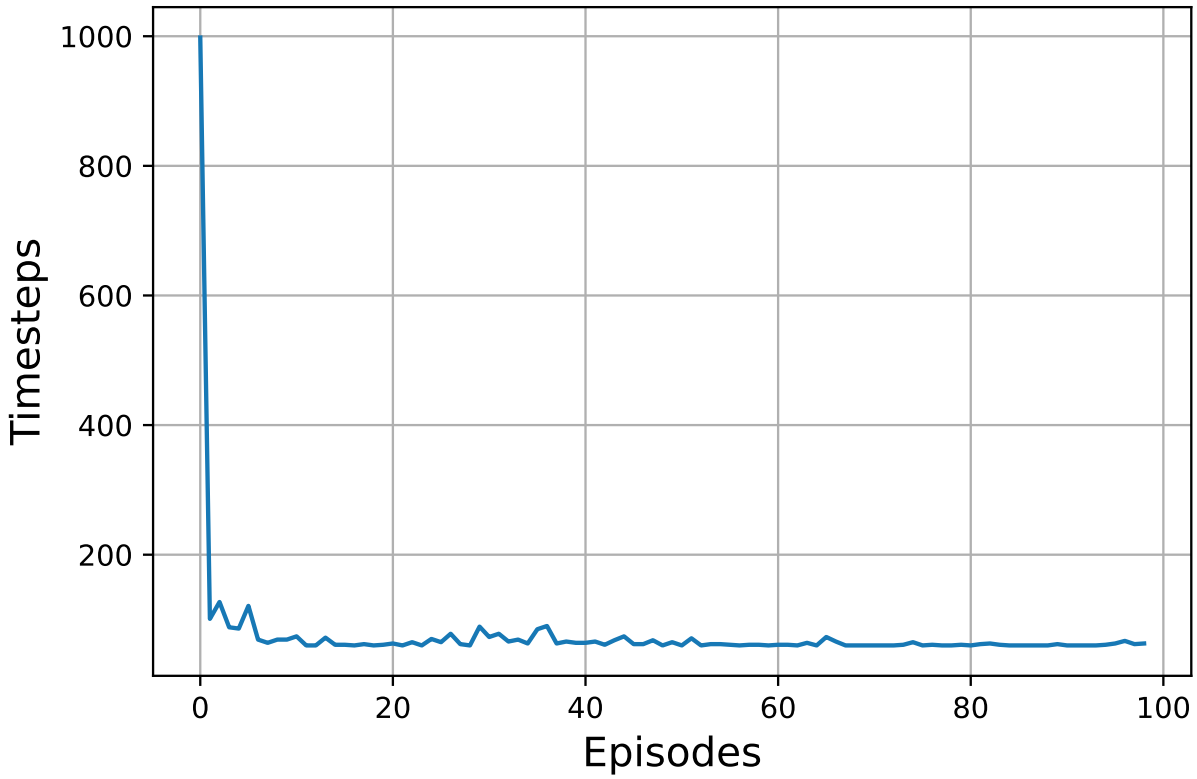

Figure 5.6: DQN optimal bandwidth searching

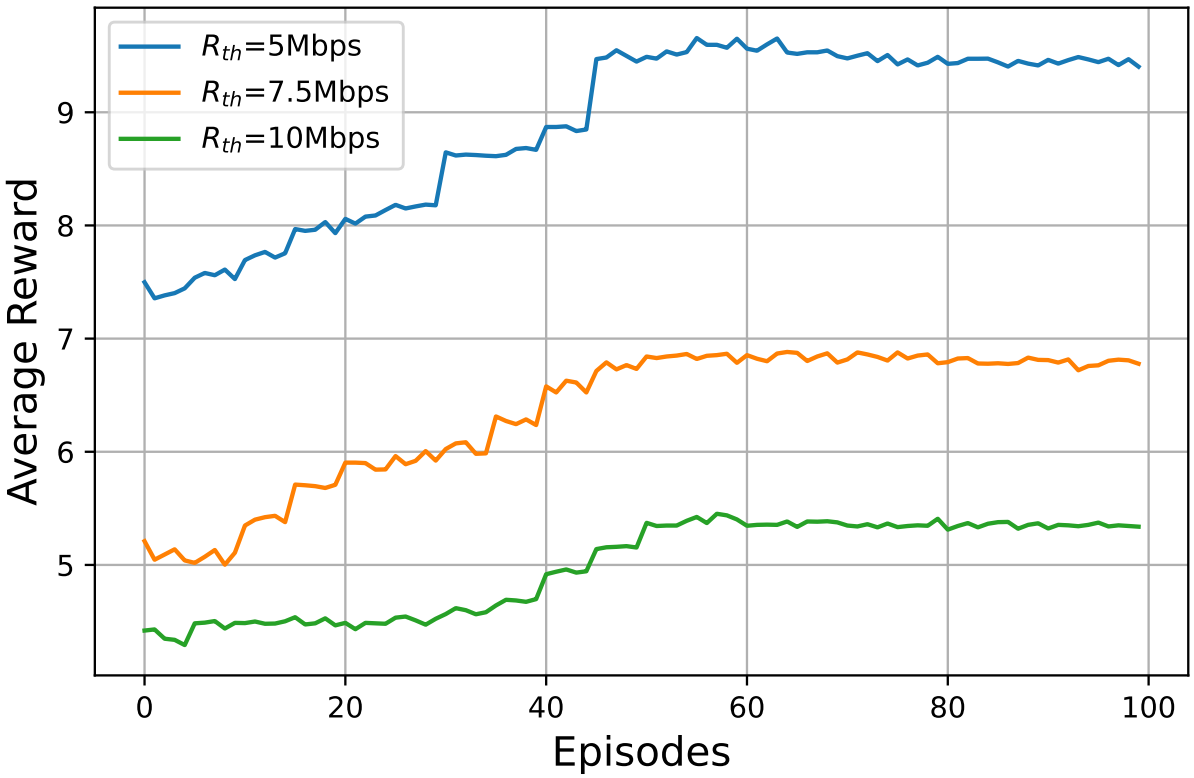

Figure 5.7: Average reward with training epsiodes

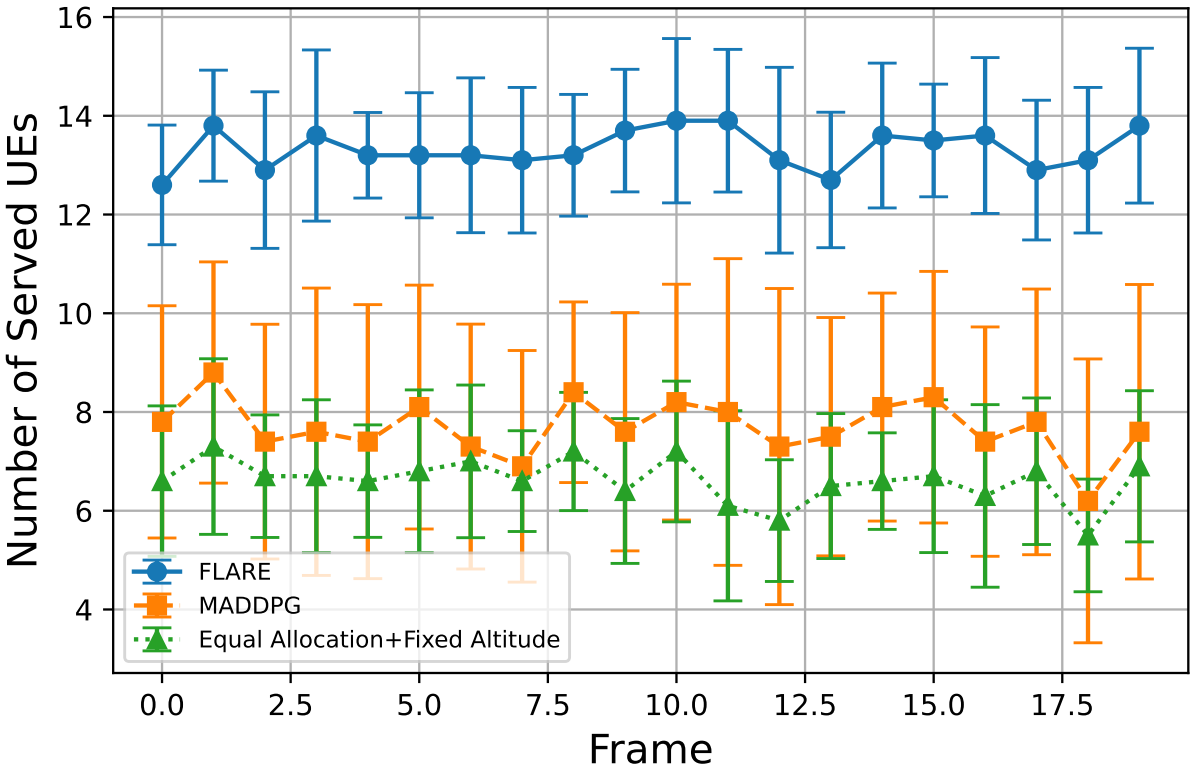

Figure 5.8: Number of served UEs with $R_{th} = 5$Mbps

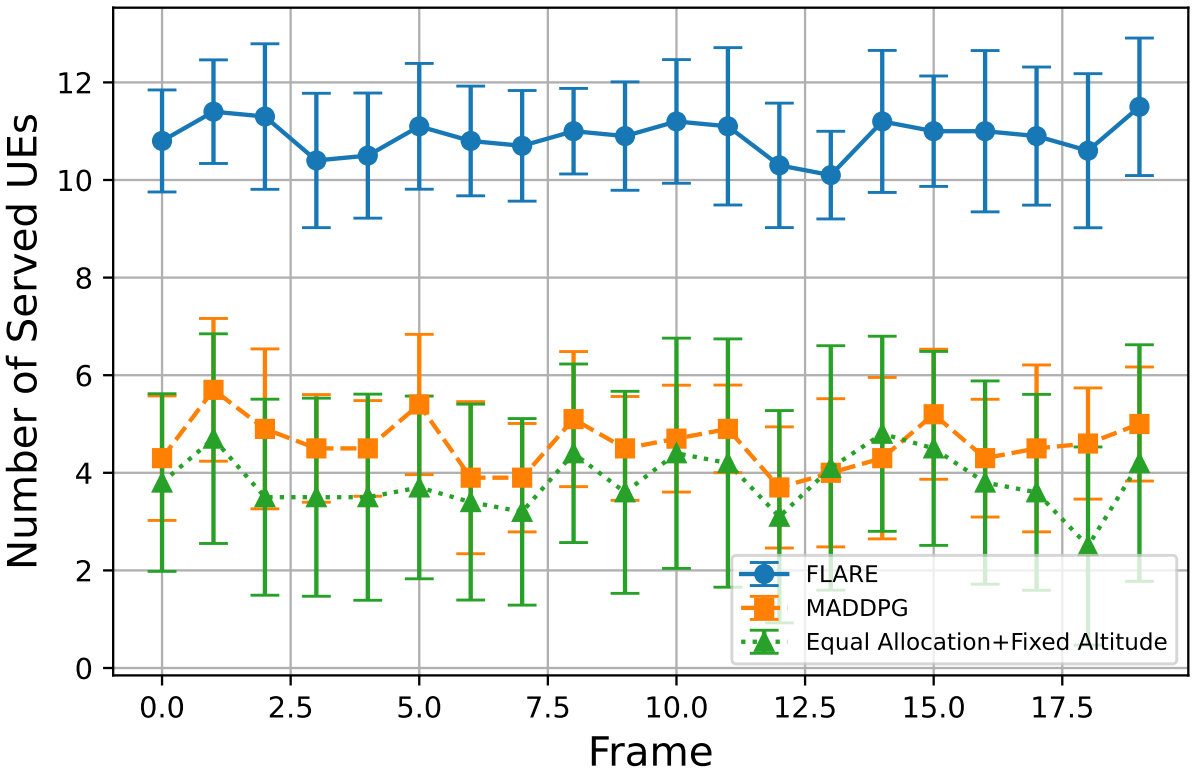

Figure 5.9: Number of served UEs with $R_{th} = 7.5$Mbps

Chapter 6

Conclusion and Future Work

This thesis has investigated the use of AI-enabled control strategies to optimize UAV-assisted wireless communications under increasingly realistic conditions. We have proposed and evaluated a three-stage framework, integrating user clustering, mobility modeling, and hybrid deep reinforcement learning for joint UAV positioning, altitude control, power allocation and bandwidth scheduling. Our main findings and contributions are summarized as follows:

6.1 Contributions and Key Findings

- **I: Single-UAV Joint Utility Optimization (Chapter 3).** We formulated a static users–single UAV scenario and developed a hybrid DQN–DDPG approach for discrete bandwidth and continuous power allocation. Extensive simulations demonstrated a $\approx 20\%$ improvement in served users over benchmark schemes.
- **II: Multi-UAV Coordination via Clustering (Chapter 4).** We introduced a K-means clustering method to partition static users among multiple UAVs and applied MADDPG for cooperative power control. Our decentralized algorithm served up to $8\times$ more users than equal-allocation baselines while respecting inter-UAV interference constraints.
- **III: Mobility-Aware Optimization (Chapter 5).** By integrating the STEP mobility model, silhouette-guided clustering and a hybrid MADDPG+DQN agent, we achieved adaptive, frame-wise UAV trajectories and resource allocations. Results

showed a 73% increase in the number of served mobile users (at 5 Mbps) compared to static-resource schemes.

Here are key findings among this thesis work:

- *Hybrid action spaces* can be effectively handled by separating discrete (DQN) and continuous (DDPG/MADDPG) decision streams, yielding faster convergence and higher stability.
- *Clustering-based UAV assignment* provides substantial gains in load balancing and interference mitigation, compared to fixed or greedy partitions.
- *Multi-agent coordination* via centralized-training/decentralized-execution MADDPG enables UAVs to learn cooperative policies that scale with the number of agents and moving users.
- *Mobility integration* demonstrates that real-time adaptation to user movement preserves coverage and QoS in dynamic environments.

6.2 Limitations

- The proposed algorithms rely on extensive simulation-based training and may incur high computational and sample-complexity costs in real-world deployment.
- Real-world factors—such as UAV flight dynamics, wind, hardware delays and partial state observability—are not fully captured in our simulations.
- Reward design remains manual and can require careful tuning to balance throughput, fairness and energy efficiency.
- The impact of SINR fluctuations on system performance, robustness, and reliability under dynamic channel conditions has not been fully considered.
- Sudden data rate dropouts may affect Quality of Service (QoS), latency, and system stability in HAPS communication scenarios.
- Variations in fading factors can impact link reliability, channel estimation accuracy, and overall system performance.

- In the future, more diverse mobility models need to be considered to better reflect realistic deployment scenarios.
- The framework can be extended to consider more data rate thresholds for finer-grained performance evaluation and adaptation.

6.3 Future Research Directions

- **Field validation:** Implement the hybrid DRL framework on UAV platforms with onboard AI accelerators to assess real-time performance, latency and robustness.
- **Transfer and meta-learning:** Incorporate transfer learning or meta-RL techniques to accelerate policy adaptation to new environments or mobility patterns.
- **Hierarchical and federated RL:** Develop hierarchical control architectures for strategic vs. low-level decisions, and explore federated training to preserve privacy and reduce communication overhead.
- **Multi-objective optimization:** Extend to multi-objective or constrained RL formulations, balancing spectral efficiency, energy consumption, latency and fairness via Pareto-optimal policies.
- **Security and robustness:** Investigate adversarial attacks (e.g., jamming, spoofing) and design risk-aware or robust RL strategies to safeguard aerial networks.
- **Explainable AI:** Integrate interpretability methods to provide insights into agent decisions, thereby improving trust and safety in autonomous UAV operations.

By bridging advanced DRL techniques with practical deployment considerations, this work lays a foundation for truly intelligent and resilient UAV-assisted networks in 6G and beyond.

References

[1] A. H. Owaid, A. Y. Abbas, H. Aldabbas, L. A. Altaee, and W. S. Weli, "A survey on uav-assisted wireless communications: Challenges, technologies, and application," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 1, pp. 216–229, 2024.

[2] X. Xiao, M. Yi, X. Wang, J. Liu, Y. Zhang, and R. Hou, "Meta-learning deep reinforcement learning for fresh data collection in uav-assisted wireless sensor networks," in *2024 IFIP Networking Conference (Networking)*. IEEE, 2024, pp. 118–123.

[3] W. Jiang, T. Cai, C. Zheng, and Y. Wang, "Collaborative encirclement of multiple uavs based on deep reinforcement learning," in *2024 36th Chinese Control and Decision Conference (CCDC)*. IEEE, 2024, pp. 5477–5482.

[4] H. S. Murtadha, A. Badrul, A. Mustapha, S. M. Shamsuddin, and A. Alzahrani, "A systematic review of interference mitigation techniques in current and future uav-assisted wireless networks," *IEEE Access*, vol. 12, pp. 51 456–51 472, 2024.

[5] A. Sun, C. Sun, J. Du, C. Chen, C. Huang, and J. Sui, "Aoi optimization for uav-assisted wireless sensor networks," in *2024 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2024, pp. 1487–1492.

[6] Y. Xia, W. Liu, K. Zhang, C. Xu, and D. Huang, "Bounding the path loss in uav-assisted wireless sensor networks," *IEEE Antennas and Wireless Propagation Letters*, vol. 23, no. 8, pp. 2341–2345, 2024.

[7] C. Zhang, Z. Xiao, L. Zhang, G. Liu, W. Zhang, and X.-G. Xia, "Collaborative multi-agent jamming deceiving for uav-assisted wireless communications," in *2024 International Wireless Communications & Mobile Computing Conference (IWCMC)*. IEEE, 2024, pp. 0549–0554.

[8] A. V. Sheshashayee, M. Bordin, P. Brach del Prever, D. Villa, H. Cheng, C. Petrioli, T. Melodia, and S. Basagni, "Experimental evaluation of the performance of uav-assisted data collection for wake-up radio-enabled wireless networks," in *2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring)*. IEEE, 2024, pp. 1–6.

[9] L. Chen, R. Wang, Y. Cui, P. He, and A. Duan, "Joint client selection and model compression for efficient fl in uav-assisted wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 10, pp. 15 172–15 184, 2024.

[10] W. A. Nelson, S. R. Yeduri, A. Jha, A. Kumar, and L. R. Cenkeramaddi, "Rl-based energy-efficient data transmission over hybrid ble/lte/wi-fi/lora uav-assisted wireless network," *IEEE/ACM Transactions on Networking*, vol. 32, no. 3, pp. 1951–1966, 2024.

[11] Q. Xu, R. Li, Y. Qi, Z. Su, and D. Fang, "Trust-enhanced game incentive for secure quantum federated learning in uav-assisted wireless networks," *IEEE Journal on Selected Areas in Communications*, 2025, early Access.

[12] C. He and C. Chen, "A comparison between ann and rbfnn for signal demodulation in photon counting based optical wireless communications," in *2024 12th International Conference on Intelligent Computing and Wireless Optical Communications (ICWOC)*. IEEE, 2024, pp. 48–53.

[13] P. Dandekar, M. Dandekar, and P. Phutane, "A comprehensive review on wireless communication and networking advances," in *2024 IEEE 3rd International Conference on Electrical Power and Energy Systems (ICEPES)*. IEEE, 2024, pp. 1–4.

[14] Y. Shi, J. Zhang, M. Bennis, and K. B. Letaief, "Ai-empowered wireless communications: From bits to semantics," *IEEE Communications Magazine*, vol. 61, no. 2, pp. 38–44, 2023.

[15] A. Kashyap, A. Munjal, and A. Kaur, "Ai-driven channel estimation using geo-spatial data with modified dmrs," in *2025 14th IEEE International Conference on Communication Systems and Network Technologies (CSNT)*. IEEE, 2025, pp. 636–641.

[16] D. Gurupandi and M. Premkumar, "Analysis on capacity of next generation wireless communication networks using artificial intelligence," in *2024 9th International Conference on Communication and Electronics Systems (ICCES)*. IEEE, 2024, pp. 1747–1750.

[17] H. Tang, L. Yang, R. Zhou, J. Liang, H. Wei, X. Wang, Q. Shi, and Z.-Q. Luo, "Assessing air-interface dataset similarity and diversity for ai-enabled wireless communications," in *2024 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2024, pp. 1623–1628.

[18] J. Zhang, Y. Shi, M. Chen, and K. B. Letaief, "At the dawn of generative ai era: A tutorial-cum-survey on new frontiers in 6g wireless intelligence," *arXiv preprint arXiv:2401.11733*, 2024.

[19] J.-P. Choi, Y. Choi, and H. Chung, "Challenges in ai-powered multi-band multi-connectivity for tbps wireless communications in sub-thz band," in *2024 15th International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2024, pp. 1041–1043.

[20] S. Gomathi, V. Kanakaraja, N. Vasudevan, K. Mahalingam, and K. Rachael, "Integrating edge ai with holographic beamforming using ai-driven reconfigurable intelligent surfaces (ris) in 6g networks," in *2025 International Conference on Computing and Communication Technologies (ICCCT)*. IEEE, 2025, pp. 1–6.

[21] W. Cai, X. Xiong, and C. Wang, "Neural network accelerator architecture designed for next generation communication systems," in *2024 10th International Conference on Computer and Communications (ICCC)*. IEEE, 2024, pp. 1182–1186.

[22] H. Zhang, Z. Wang, and X. Xiang, "Resource allocation and optimization model of wireless communication system based on ai," in *2025 International Conference on Digital Analysis and Processing, Intelligent Computation (DAPIC)*. IEEE, 2025, pp. 607–611.

[23] P. R. Kapula, C. Jayanth, S. Dubey, P. Gahitha, B. R. S. Reddy, and B. Mohit, "Transforming wireless communication systems: A review of ai-based tst coding techniques," in *Proceedings of the 7th International Conference on Inventive Computation Technologies (ICICT)*. IEEE, 2024.

[24] L. Zhou, X. Deng, Z. Ning, H. Zhao, J. Wei, and V. C. M. Leung, "When generative ai meets semantic communication: Optimizing radio map construction and distribution in future mobile networks," *IEEE Network*, vol. 39, no. 3, pp. 47–55, 2025.

[25] Q. Lao, Q. Jiang, and Y. Xing, "Adaptive access control algorithm for uav front-end sensing system based on reinforcement learning," in *2024 3rd International Conference on Artificial Intelligence and Autonomous Robot Systems (AIARS)*. IEEE, 2024, pp. 619–624.

[26] H. Samma and S. El-Ferik, "Autonomous uav visual navigation using an improved deep reinforcement learning," *IEEE Access*, vol. 12, pp. 79 967–79 977, 2024.

[27] W. Huo and J. Li, "Cooperative uav maneuver decision-making based on multi-agent reinforcement learning," in *2024 4th International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*. IEEE, 2024, pp. 475–479.

[28] Y. Wang, K. Wang, K. Yang, Q. Wu, and A. Nallanathan, "Decentralized navigation with heterogeneous federated reinforcement learning for uav-enabled mobile edge computing," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 6, pp. 7304–7319, 2024.

[29] G. Han, Q. Wu, B. Wang, C. Lin, J. Zhuang, W. Li, Z. Hao, and Z. Fan, "Deep reinforcement learning based multi-uav collision avoidance with causal representation learning," in *2024 10th International Conference on Big Data and Information Analytics (BigDIA)*. IEEE, 2024, pp. 833–839.

[30] Y. Xu, Y. Yu, J. Wang, and X. Wang, "Deep reinforcement learning-based physical verification of autonomous landing for uav," in *2024 IEEE International Conference on Unmanned Systems (ICUS)*. IEEE, 2024, pp. 1159–1165.

[31] A. Singh and R. M. Hegde, "Gee maximization in uav-aided mobile iot networks using deep reinforcement learning," in *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–6.

[32] A. Su, F. Hou, and Y. Hong, "Heterogeneous policy network reinforcement learning for uav swarm confrontation," in *2024 China Automation Congress (CAC)*. IEEE, 2024, pp. 722–727.

[33] X. Xing, Z. Zhou, Y. Li, B. Xiao, and Y. Xun, "Multi-uav adaptive cooperative formation trajectory planning based on an improved matd3 algorithm of deep reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 9, pp. 12 484–12 499, 2024.

[34] S. Liu, G. Su, and B. Chen, "Multi-uav collaborative secure communication using multi-agent deep reinforcement learning," in *2024 7th International Conference on Electronics Technology (ICET)*. IEEE, 2024, pp. 772–777.

[35] Y. Xu, Z. Jian, J. Zha, and X. Chen, "Poster abstract: Emergency networking using uavs: A reinforcement learning approach with large language model," in *2024 23rd*

ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN). IEEE, 2024, pp. 281–282.

[36] S. Kulkarni and D. D. Patil, "Reinforcement learning for autonomous systems," in *Proceedings of the Fourth International Conference on Sentiment Analysis and Deep Learning (ICSADL-2025)*. IEEE, 2025, pp. 816–820.

[37] G. Kim, J. Kim, S. Hong, and S. Cho, "Reinforcement learning-based uav handover algorithm in cellular networks: A survey," in *2024 15th International Conference on Ubiquitous and Future Networks (ICUFN)*. IEEE, 2024, pp. 1–6.

[38] Y. Dai, Y. Li, and T. Lyu, "Task offloading based on multi-agent reinforcement learning for uav-assisted edge computing," in *2024 Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*. IEEE, 2024, pp. 426–430.

[39] B. I.-D. Ghomri, M. Y. Bendimerad, and F. T. Bendimerad, "Utilizing deep reinforcement learning for optimal uav 3d placement in noma-uav networks," in *2024 2nd International Conference on Electrical Engineering and Automatic Control (ICEEAC)*. IEEE, 2024, pp. 1–7.

[40] N. Parvaresh, M. Kulhandjian, H. Kulhandjian, C. D'Amours, and B. Kantarci, "A tutorial on ai-powered 3d deployment of drone base stations: State of the art, applications and challenges," *Vehicular Communications*, vol. 36, p. 100474, 2022.

[41] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," 2013.

[42] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," 2019.

[43] O. M. Gul, A. M. Erkmen, and B. Kantarci, "UAV-driven sustainable and quality-aware data collection in robotic wireless sensor networks," *IEEE Internet of Things Journal*, vol. 9, no. 24, pp. 25 150–25 164, 2022.

[44] I. Ruban and V. Lebediev, "Methodology for determining the rational number of uavs taking into account their reliability in emergency situations," in *2022 IEEE 3rd KhPI Week on Advanced Technology (KhPIWeek)*, 2022, pp. 1–5.

[45] G. Peng, Y. Xia, X. Zhang, and L. Bai, "Uav-aided networks for emergency communications in areas with unevenly distributed users," in *2018 IEEE International Conference on Communication Systems (ICCS)*, 2018, pp. 25–29.

[46] O. M. Gul, A. M. Erkmen, and B. Kantarci, "NTN-aided quality and energy-aware data collection in time-critical robotic wireless sensor networks," *IEEE Internet of Things Magazine*, vol. 7, no. 3, pp. 114–120, 2024.

[47] W. Huang, Z. Yang, C. Pan, L. Pei, M. Chen, M. Shikh-Bahaei, M. Elkashlan, and A. Nallanathan, "Joint power, altitude, location and bandwidth optimization for uav with underlaid d2d communications," *IEEE Wireless Communications Letters*, vol. 8, no. 2, pp. 524–527, 2019.

[48] P. Lohan and D. Mishra, "Utility-aware optimal resource allocation protocol for uav-assisted small cells with heterogeneous coverage demands," *IEEE Transactions on Wireless Communications*, vol. 19, no. 2, pp. 1221–1236, 2020.

[49] H. Tu, J. Zhu, and Y. Zou, "Optimal power allocation for minimizing outage probability of uav relay communications," in *2019 11th International Conference on Wireless Communications and Signal Processing (WCSP)*, 2019, pp. 1–6.

[50] Z. Yang, C. Pan, M. Shikh-Bahaei, W. Xu, M. Chen, M. Elkashlan, and A. Nallanathan, "Joint altitude, beamwidth, location, and bandwidth optimization for uav-enabled communications," *IEEE Communications Letters*, vol. 22, no. 8, pp. 1716–1719, 2018.

[51] L. A. Grieco, G. Boggia, G. Piro, Y. Jararweh, and C. Campolo, *Ad-Hoc, Mobile, and Wireless Networks*. Springer, 2020.

[52] Q. T. Do, D. T. Hua, A. T. Tran, and S. Cho, "Energy efficient multi-uav communication using ddpg," in *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*, 2022, pp. 1071–1075.

[53] N. Parvaresh and B. Kantarci, "A continuous actor–critic deep q-learning-enabled deployment of uav base stations: Toward 6G small cells in the skies of smart cities," *IEEE Open Journal of the Communications Society*, vol. 4, pp. 700–712, 2023.

[54] Y. Wang, T. Ren, and Z. Fan, "Autonomous maneuver decision of uav based on deep reinforcement learning: Comparison of dqn and ddpg," in *2022 34th Chinese Control and Decision Conference (CCDC)*, 2022, pp. 4857–4860.

[55] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP altitude for maximum coverage," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 569–572, Dec. 2014.

[56] M. Simon and M.-S. Alouini, *Digital Communication Over Fading Channels*. 2nd ed. New York, NY, USA: Wiley, 2005.

[57] H. Zakeri, P. Khoddami, G. Moradi, M. Alibakhshikenari, R. Abd-Alhameed, S. Koziel, and M. Dalarsson, "Path loss model estimation at indoor environment by using deep neural network and catboost for wireless application," *IEEE Access*, vol. 12, pp. 159 070–159 085, 2024.

[58] Y. He, Y. Gan, H. Cui, and M. Guizani, "Fairness-Based 3-D Multi-UAV Trajectory Optimization in Multi-UAV-Assisted MEC System," *IEEE Internet of Things Journal*, vol. 10, no. 13, pp. 11 383–11 395, 2023.

[59] N. Parvaresh, M. Kulhandjian, H. Kulhandjian, C. D'Amours, and B. Kantarci, "A tutorial on AI-powered 3D deployment of drone base stations: State of the art, applications and challenges," *Vehicular Communications*, vol. 36, p. 100474, 2022.

[60] C. Zhang, Z. Li, C. He, K. Wang, and C. Pan, "Deep Reinforcement Learning Based Trajectory Design and Resource Allocation for UAV-Assisted Communications," *IEEE Commun. Letters*, vol. 27, no. 9, pp. 2398–2402, 2023.

[61] Y. He, K. Xiang, X. Cao, and M. Guizani, "Task Scheduling and Trajectory Optimization Based on Fairness and Commun. Security for Multi-UAV-MEC System," *IEEE Internet of Things Journal*, vol. 11, no. 19, pp. 30 510–30 523, 2024.

[62] X. Cai, P. Lohan, and B. Kantarci, "A Novel Joint DRL-Based Utility Optimization for UAV Data Services," in *2024 IEEE 10th World Forum on Internet of Things (WF-IoT)*, 2024, pp. 930–935.

[63] X. Zhang, Z. Liu, J. Zhang, and B. Ai, "Joint UAV Trajectory and Power Optimization in Cell-Free mMIMO Systems With Deep Q-Network," in *Intl Conf. on Ubiquitous Commun.*, 2024, pp. 395–399.

[64] P. Qin, Y. Fu, J. Zhang, S. Geng, J. Liu, and X. Zhao, "DRL-Based Resource Allocation and Trajectory Planning for NOMA-Enabled Multi-UAV Collaborative Caching 6G Network," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 6, pp. 8750–8764, 2024.

[65] B. Yin, X. Fang, and X. Wang, "Joint Optimization of Trajectory Control, Resource Allocation, and User Association Based on DRL for Multi-Fixed-Wing UAV Networks," *IEEE Transactions on Wireless Commun.*, vol. 23, no. 10, pp. 13 330–13 343, 2024.

[66] L. Wang, H. Liang, Y. Tang, G. Mao, H. Zhang, and D. Zhao, "DRL-Based Joint Resource Allocation and Platoon Control Optimization for UAV-Hosted Platoon Digital Twin," *IEEE Internet of Things Journal*, vol. 11, no. 22, pp. 37 114–37 126, 2024.

[67] M. Zhang, S. Fu, and Q. Fan, "Joint 3D Deployment and Power Allocation for UAV-BS: A Deep Reinforcement Learning Approach," *IEEE Wireless Commun. Letters*, vol. 10, no. 10, pp. 2309–2312, 2021.

[68] F. Xu, Y. Ruan, and Y. Li, "Soft Actor–Critic Based 3-D Deployment and Power Allocation in Cell-Free Unmanned Aerial Vehicle Networks," *IEEE Wireless Commun. Letters*, vol. 12, no. 10, pp. 1692–1696, 2023.

[69] S. Fu, X. Feng, A. Sultana, and L. Zhao, "Joint Power Allocation and 3D Deployment for UAV-BSs: A Game Theory Based Deep Reinforcement Learning Approach," *IEEE Transactions on Wireless Commu.*, vol. 23, no. 1, pp. 736–748, 2024.

[70] L. Wang, K. Wang, C. Pan, W. Xu, N. Aslam, and L. Hanzo, "Multi-Agent Deep Reinforcement Learning-Based Trajectory Planning for Multi-UAV Assisted Mobile Edge Computing," *IEEE Transactions on Cognitive Commun. and Networking*, vol. 7, no. 1, pp. 73–84, 2021.

[71] X. Kong, C. Ni, G. Duan, G. Shen, Y. Yang, and S. K. Das, "Energy Consumption Optimization of UAV-Assisted Traffic Monitoring Scheme With Tiny Reinforcement Learning," *IEEE Internet of Things Journal*, vol. 11, no. 12, pp. 21 135–21 145, 2024.

[72] P. Qin, X. Wu, R. Ding, M. Fu, X. Zhao, Z. Chen, and H. Zhou, "Joint Resource Allocation and UAV Trajectory Design for D2D-Assisted Energy-Efficient Air–Ground Integrated Caching Network," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 11, pp. 17 558–17 571, 2024.

[73] P. Qin, J. Li, J. Zhang, and Y. Fu, "Joint Task Allocation and Trajectory Optimization for Multi-UAV Collaborative Air–Ground Edge Computing," *IEEE Trans. on Network Science and Engineering*, vol. 11, no. 6, pp. 6231–6243, 2024.

[74] Z. Wang, H. Wang, L. Liu, E. Sun, H. Zhang, Z. Li, C. Fang, and M. Li, "Dynamic Trajectory Design for Multi-UAV-Assisted Mobile Edge Computing," *IEEE Trans. on Vehicular Technology*, pp. 1–15, 2024.

[75] G. Chen, F. Sun, H. Liang, Q. Zeng, and Y.-D. Zhang, "MADDPG-M&L: UAV-Assisted Joint User Association and Slicing Resource Allocation in HetNets," *IEEE Transactions on Network Science and Engineering*, pp. 1–16, 2025.

[76] J. Du, Z. Kong, A. Sun, J. Kang, D. Niyato, X. Chu, and F. R. Yu, "MADDPG-Based Joint Service Placement and Task Offloading in MEC Empowered Air–Ground Integrated Networks," *IEEE Internet of Things Journal*, vol. 11, no. 6, pp. 10 600–10 617, 2024.

[77] C. Gu, F. Li, D.-S. Liu, Y.-X. Wu, and H.-X. Wang, "DRL-Based Joint Task Scheduling and Trajectory Planning Method for UAV-Assisted MEC Scenarios," *IEEE Access*, vol. 12, pp. 156 224–156 234, 2024.

[78] H. Sun, Y. Zhou, H. Zhang, L. Ale, H. Dai, and N. Zhang, "Joint Optimization of Caching, Computing, and Trajectory Planning in Aerial Mobile Edge Computing Networks: An MADDPG Approach," *IEEE Internet of Things Journal*, vol. 11, no. 24, pp. 40 996–41 007, 2024.

[79] Z. Kaleem, A. Ahmad, O. Chughtai, and J. J. Rodrigues, "Enhanced Max-Min Rate of Users in UAV-Assisted Emergency Networks Using Reinforcement Learning," *IEEE Networking Letters*, vol. 4, no. 3, pp. 104–107, 2022.

[80] A. Gemayel, D. M. Manias, and A. Shami, "Network Resource Optimization for ML-Based UAV Condition Monitoring with Vibration Analysis," *IEEE Networking Letters*, 2025, to appear.

[81] Q.-V. Pham, N. Iradukunda, N. H. Tran, W.-J. Hwang, and S.-H. Chung, "Joint Placement, Power Control, and Spectrum Allocation for UAV Wireless Backhaul Networks," *IEEE Networking Letters*, vol. 3, no. 2, pp. 56–59, 2021.

[82] M. T. Nguyen and L. B. Le, "Resource Allocation, Trajectory Optimization, and Admission Control in UAV-Based Wireless Networks," *IEEE Networking Letters*, vol. 3, no. 3, pp. 129–132, 2021.

[83] M. Samir, S. Sharafeddine, C. Assi, T. M. Nguyen, and A. Ghrayeb, "Trajectory Planning and Resource Allocation of Multiple UAVs for Data Delivery in Vehicular Networks," *IEEE Networking Letters*, vol. 1, no. 3, pp. 107–110, 2019.

[84] J. Wang, R. Zhou, and J. e. a. Gao, "Flow-based spatio-temporal structured prediction of motion dynamics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[85] L. Yu, M. Li, W. Jin, Y. Guo, Q. Wang, F. Yan, and P. Li, "Step: A spatio-temporal fine-granular user traffic prediction system for cellular networks," *IEEE Transactions on Mobile Computing*, vol. 20, no. 12, pp. 3453–3469, 2021.

[86] Z. Yang, L. He, Z. Wang, and Z. Guo, "Probabilistic reasoning for unique role recognition based on the fusion of semantic-interaction and spatio-temporal features," *IEEE Transactions on Multimedia*, 2024.

[87] Y. An, Z. Li, and X. e. a. Li, "Spatio-temporal multivariate probabilistic modeling for traffic prediction," *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[88] X. Jiang and L. e. a. Zhao, "Probabilistic spatio-temporal traffic flow forecasting based on denoising diffusion framework," *IEEE Transactions on Intelligent Transportation Systems*, 2025.

[89] Y. An, Z. Li, X. Li, W. Liu, X. Yang, H. Sun, M. Chen, Y. Zheng, and Y. Gong, "Spatio-temporal multivariate probabilistic modeling for traffic prediction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 37, no. 5, pp. 2986–3000, 2025.

[90] J. Wang and B. Li, "Dh-kmeans: An improved k-means clustering algorithm based on dynamic initial cluster center determination and hierarchical clustering," *Pattern Recognition Letters*, 2022.

[91] S.-M. e. a. Yu, "Trust cop-kmeans clustering analysis and minimum-cost consensus model considering voluntary trust loss in social network large-scale decision-making," *IEEE Transactions on Fuzzy Systems*, vol. 30, no. 7, 2022.

[92] H. e. a. Liu, "Rural developing level clustering based on kmeans from electricity perspective," *IEEE Access*, 2023.

[93] X. e. a. Zhang, "Multi-time scale stochastic optimization for hybrid ac–dc distribution networks with pet based on e-c-kmeans clustering," *Electric Power Systems Research*, 2023.

[94] Y. e. a. Chen, "Image color reduction using progressive histogram quantization and kmeans clustering," *Multimedia Tools and Applications*, 2023.

[95] Z. e. a. Ali, "Statistical analysis of microarray data: Clustering using nmf, spectral clustering, kmeans and gmm," *Journal of Bioinformatics*, 2023.