Reinforcement Learning-Based Optimization of Quality-of-Service and Path Planning for Multi-UAV Mobile Edge Computing

Xiang Li [*, #], Qi Liu [#], Zhuocheng Yang [#]

Department of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China, 100091

* Corresponding Author Email: 21721013@bjtu.edu.cn

[#]These authors contributed equally.

Abstract. Unmanned Aerial Vehicles (UAVs) have traditionally served as network processors in mobile networks, and more recently, as mobile servers in Mobile Edge Computing (MEC). However, deploying UAVs in dynamic and obstacle-rich environments while ensuring efficient coordination among multiple UAVs presents significant challenges. To tackle these challenges, we propose a unified reinforcement learning framework for a multi-UAV MEC platform that enhances Quality-of-Service (QoS) and optimizes path planning. Specifically, our contributions include: (1) Integrating QoS optimization and UAV path planning into a unified reinforcement learning framework; (2) Modeling terminal users' demands with a sigmoid-like function to enhance service quality; (3) Incorporating terminal users' demand, risk factors, and geometric distance into the reinforcement learning reward matrix for balancing service quality, risk mitigation, and cost efficiency. Experimental results demonstrate the framework's effectiveness: simulations in 3D environments show a 20–35% reduction in collision incidents compared to greedy algorithms and a 15–25% improvement in demand fulfillment rates under dynamic user distributions. Further analysis highlights the impact of sigmoid parameter tuning on risk-QoS trade-offs, validating the adaptability of the proposed approach in complex scenarios.

Keywords: Multi-UAV, Mobile Edge Computing, Reinforcement Learning, Quality-of-Service, Path Planning.

1. Introduction

With the surge in demand for mobile data processing and the rapid deployment of technologies such as 5G, traditional base stations are increasingly struggling to meet end-user requirements in complex environments. Consequently, Mobile Edge Computing (MEC) has rapidly emerged as a crucial research area in telecommunications [1]. MEC centralizes networking, computing, storage, and intelligent services at the network edge, bringing computation closer to end users and improving service performance [2, 3].

Due to their compact size, high maneuverability, and cost-effectiveness, UAVs have emerged as a promising platform for enhancing MEC service quality. UAVs can quickly transition between end users, providing computational services with greater flexibility and efficiency [4-8]. However, several challenges remain, including path planning in dynamic environments, energy constraints, and risk mitigation.

Evidence indicates that UAV-enabled mobile edge computing (MEC) faces challenges such as dynamic user distributions, energy constraints, and complex environments [1], necessitating efficient path planning. Prior studies, like Q-learning-based multi-agent algorithms [9] and echo state network (ESN) predictors for user mobility [9], aim to optimize UAV paths, while others, such as Liu et al.'s service quality model [10], focus on maximizing planning rewards. However, interpretation reveals critical gaps: most approaches rely on one-time greedy strategies or task-specific optimizations [9, 10], lacking mechanisms for continuous improvement. This rigidness limits adaptability to dynamic conditions, increases local optima risks, and neglects collision avoidance and real-world risks [9, 10]. Evidence further highlights reinforcement learning (RL) as a promising alternative, as shown in studies [11-13] enhancing QoS by framing path planning as constraint-driven optimization. Interpretation underscores RL's flexibility over traditional methods (e.g., A*, RRT) in MEC scenarios [11, 13], where real-time policy updates are essential for fluctuating user demands [11], and integrated obstacle avoidance with task allocation exceeds the capabilities of geometry-focused algorithms [12, 13].

Reinforcement Learning (RL) has been demonstrated strong potential in optimizing UAV path planning and resource allocation. Compared **to** conventional approaches such as the A* algorithm or Rapidly-exploring Random Tree (RRT), RL allows for real-time adaptation to dynamic environments. Motivated by these considerations, we propose a novel UAV-enabled MEC platform that incorporates reinforcement learning for QoS optimization and path planning. The remainder of this paper is organized as follows: Section II offers an overview of relevant research, while Section III details the UAV-enabled MEC platform, and Section IV presents conclusions and future research directions.

2. Previous Work

2.1. Mobile Edge Computing

MEC has rapidly evolved as a key area within edge computing [14]. Previous studies [1] [2] have explored MEC architecture, computation offloading, and resource allocation strategies. However, dynamically managing MEC resources [3] remains a key challenge due to fluctuating user demands and limited edge computing capacity.

2.2. Path Planning

Recent research has focused on improving UAV path planning. Kim et al. [15] proposed a trajectory planning algorithm for UAV relay nodes in dynamic networks, while Zhang et al. [16] introduced a Cooperative and Geometric Learning Algorithm (CGLA) for efficient UAV movement. These studies highlight the importance of optimizing UAV paths in MEC applications.

2.3. Integration of Mobile Edge Computing and Path Planning

Several works have investigated the integration of MEC and UAV path planning using methods such as successive convex approximation (SCA) and deep reinforcement learning [17-19]. However, existing studies [20] often overlook real-time adaptability and UAV cooperation. Our work builds upon these efforts by proposing a reinforcement learning-based multi-UAV MEC framework.

3. Multi-UAV Mobile Edge Computing and Path Planning via Reinforcement Learning

In the context of reinforcement learning, an agent seeks to identify the best strategy in various scenarios to maximize its projected reward. In this system, the UAV, functioning as a mobile network processor, continuously adapts to environmental changes. At each time step, it selects an optimal course of action based on surrounding conditions. After executing a move, the environment updates, and the UAV receives either positive or negative reinforcement through a reward matrix AA, which considers factors such as risk, spatial distance, and user demands. The UAV then enhances its decision-making process by leveraging a randomly generated iterative cost matrix GG, derived from state AA. Each unit determines a route toward its objective, with GG serving as a dynamic memory that progressively refines over multiple iterations.

A. Modeling the Environment

This study, we model UAV collision avoidance alongside the demand from terminal users within a unified framework. The environment comprises two primary elements: obstacles and terminal users. First, obstacles are characterized by diverse shapes, positions, and risk levels, representing real-world entities such as buildings, vehicles, or mountains. Second, we assume that these obstacles follow a

Gaussian distribution, each defined by a variance (σ) that is employed to calculate the probability of risk exposure.

Assuming there are n independent obstacles on the map, let the site of the i-th obstacle be denoted as $O_i = (X_i, Y_i)$ .The risk $r_i(x, y)$, which quantifies the threat contribution of obstacle O_i at this stage (x, y), can be thought as:

$$r_i(x, y) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{x^2}{2\sigma}}, \quad d = \sqrt{(x - X_i)^2 + (y - Y_i)^2}, i \in \{1, 2, \ldots, n\}. \tag{1}$$

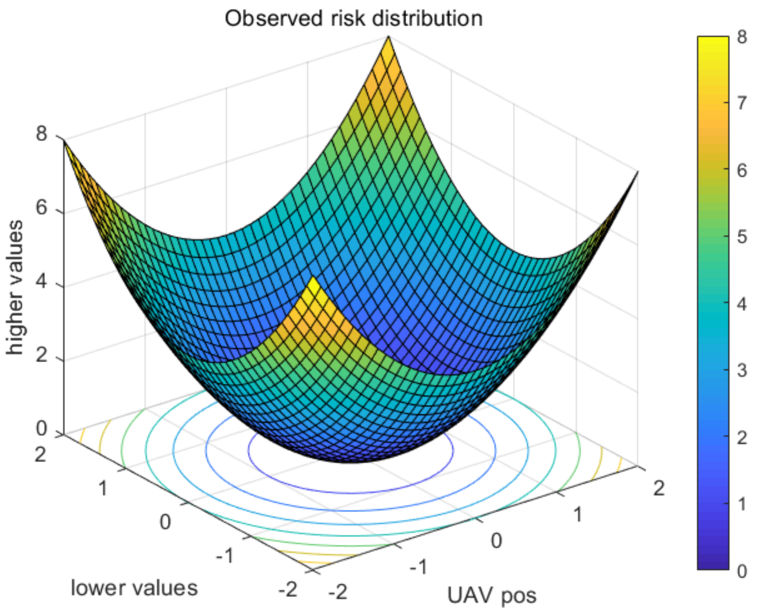

Figure 1. Effectiveness of path planning algorithms for UAVs in a 3 D environment

As shown in Figure 1, using professional simulation software or a self-developed simulation platform, the flight process of UAVs in a three-dimensional environment is simulated to validate the effectiveness and viability of the path planning algorithm. Detailed analysis of the simulation results is conducted to identify existing issues and shortcomings, followed by improvements and optimizations to the algorithm and model, thereby enhancing the quality and reliability of the path planning.

Taking into account all n barriers on the map, the cumulative risk at the coordinate (x, y), as reflected in the context of the problem probability matrix of exposure, is given by:

$$R(x, y) = 1 - \prod_{i=1}^{n} [1 - r_i(x, y)] \tag{2}$$

The cumulative risk between any two points p and q on the map is defined as the integral of $R(x, y)$ along the path C ,which is the direct route from p to q:

$$\int_{(x,y)\in C} R(x, y) \tag{3}$$

Secondly, in the context in serving terminal users, it is assumed that every user initially has a demand d_j that requires processing by UAVs. Moreover, we assume that this demand can be met by UAVs operating within a fixed service radius, given the UAVs' limited ability to detect signals of demand beyond a set threshold. Consequently, The service zone is indicated as $s(p_j, \grave{o})$, where p_j

represents the site of TU_j, and ò. The service area corresponds to the service radius, as illustrated in Figure 2.

K = 8.0 and M = 0.040 are used for iteratively obtaining batch data. These parameters can be set to a range of values to examine their impact on the outcomes of the UAV swarm.

Anytime the UAV enters the service area of TU_j, it begins fulfilling the demand of TU_j. The outstanding demand of TU_j gradually decreases at a constant rate of τ each time unit for each UAV. Naturally, a terminal user with higher demand requires longer service duration, and as the duration of UAV's operation increases stays within the service ranges, the more demand it can fulfill for TU_j. The change in d_j overtime t_k in relation to the UAV's service duration k can be expressed as:

$$d_j^{l+1} = d_j^l - \tau t_k \tag{4}$$

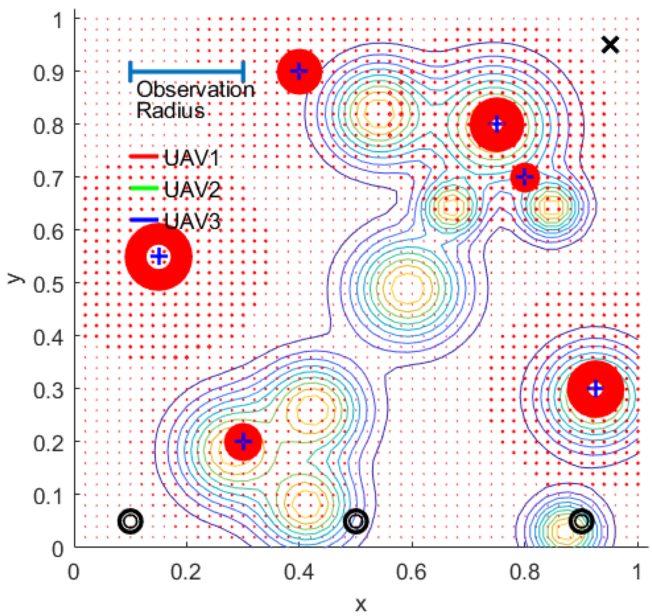

Figure 2. Obstacle risk and terminal user demand map

To enhance system efficiency, the relationship Regarding the relationship between end-user requirements and UAV demand detection, it should be non-linear. Drawing from [10] and [21], a sigmoid-like function proves effective in amplifying strong signals while suppressing weaker ones. Therefore, we employ a demand detection function resembling a sigmoid curve $U(d_j) \in (0,1]$ to characterize the relationship between actual demand and identified demand, formulated as:

$$U(d_j) = 1 - \exp\left[-\frac{(d_j)^n}{d_j + \beta} \right] \tag{5}$$

Where η and β are controlling variables

As Figure 2, it initially rises sharply as per the demand increases and stabilizes when the demand reaches a sufficiently high level. Therefore, (5) can motivate a UAV to prioritize end users having higher unsatisfied demanded and discourage it from continuously serving single terminal user for an extended period, thereby enhancing QoS.

Typically, a Sigmoid-shaped function is a type of monotonically rising function which possesses a curvature point x_0. It adheres to $\frac{d^2 f(x)}{dx^2} > 0$ when $x < x_0$, and exhibits $\frac{d^2 f(x)}{dx^2} < 0$ when $x > x_0$ [21]. Functions of this form exhibit the following characteristics:

Property 1: The function is only valid for any $x > 0$ within the range of $U(x)$, provided that

$$\eta \in (1, \infty) \quad (6)$$

$$\beta \in (0, \infty) \quad (7)$$

The proof can be found in Proof 1 located in the supplementary section. Property 2 indicates that η influences both the gradient and central alignment of the graph, rotating around the pivot point $(1, 1-e^{-\frac{\pi}{1+\beta}})$. The parameter β regulates the horizontal shift of the curve. Consequently, adjusting β enables vertical displacement of the intersection point. A detailed proof can be found in Proof 2 in the appendix.

When studying UAV flight paths in complex environments, various factors are considered, including risks posed by both static and moving obstacles, dynamic changes in terminal user demand, and UAV flight energy consumption. By adjusting the risk parameter (K) and the demand parameter (M), UAV path planning can be influenced. E.g., when K is a set to a higher value, UAVs become more cautious in avoiding obstacles, which may, however, lead to certain services being omitted.

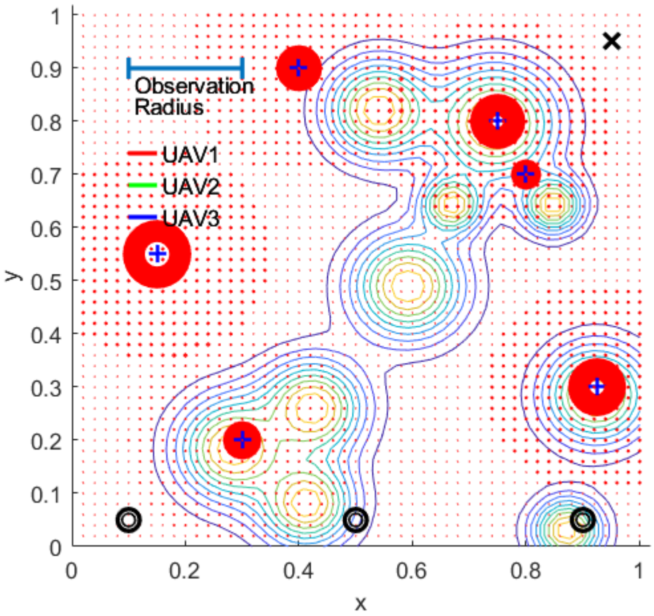

Figure 3. Obstacle risk and terminal user demand map

As shown in Figure 3, the risk distribution observed by each UAV is diverse, encompassing multiple aspects such as sensor attacks, positioning system attacks, communication network attacks, software attacks, and physical damage. To mitigate these risks, a series of security measures should be implemented, including strengthening the protection of sensors and communication networks, enhancing software security and reliability, and conducting pre-flight risk assessments and planning.

As shown in Figure 4, $U(x)$ has an inflection point when $x = 1$. The curve becomes steeper when η increases and moves vertically downward when β increases. Therefore, η and β are constant variables that affect QoS.

The evidence supporting this will be presented in part IV. To streamline the framework, we presume that demands accumulate in a linear fashion. For a UAV situated at point p on the map, with a detection radius of $\grave{o}$, the total detected demand is the aggregate of the linear summation of terminal users' requirements within the circular region $s(p, \grave{o})$:

$$\sum_{j \in s(p,\grave{o})} U(d_j) \quad (8)$$

In UAV applications, the sigmoid function can be utilized for risk assessment and classification based on observed data. For instance, various environmental data or target features detected by UAVs can be processed through a neural network, with the sigmoid function ultimately outputting a risk probability value.

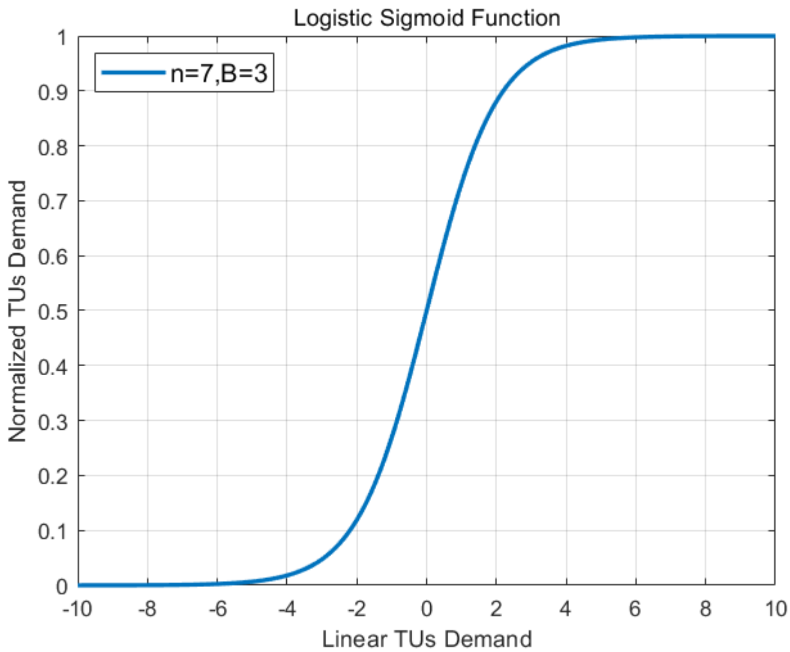

Figure 4. The sigmoid function evaluates the risk

Suppose a UAV needs to assess obstacle risks in its forward path while executing a mission. In this case, a neural network model processes information such as distance, velocity, and object shape obtained from sensors. The sigmoid function then generates a probability value ranging between 0 and 1, indicating the likelihood of an obstacle risk in that area.

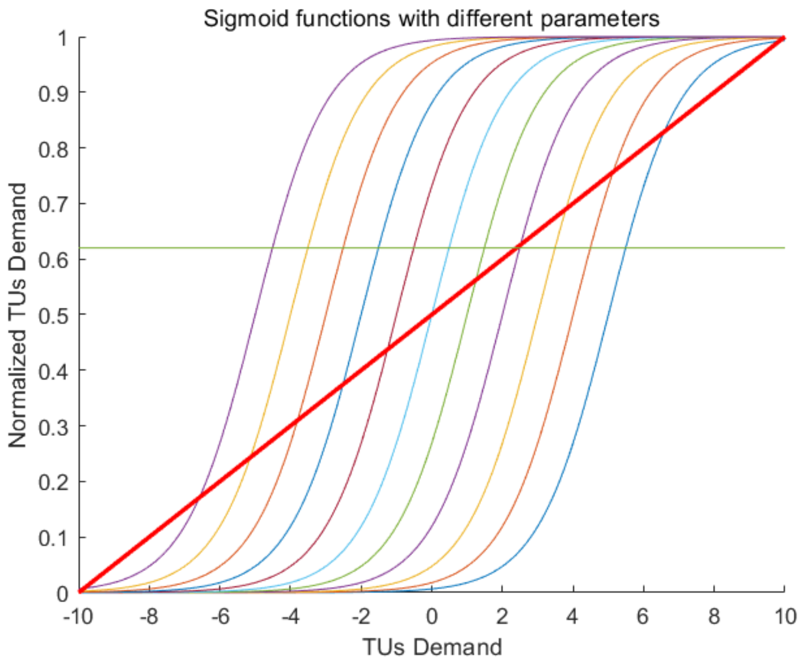

Figure 5. Different parameters for the shape and position of the risk distribution

As shown in Figure 5, the sigmoid function includes two key parameters: the position parameter (μ) and the form parameter (γ). The position parameter determines the central position of the function, while the shape parameter controls the slope and curvature of the function.

In the risk distribution observed by UAVs, different parameter settings can influence the shape and position of the risk distribution curve. For instance, a larger shape parameter results in a steeper curve, indicating more abrupt changes in risk levels. Conversely, a smaller shape parameter leads to a smoother curve, representing a more gradual variation in risk.

A reward matrix is incorporated to facilitate UAVs' learning and adaptation process in finding the most optimized route. This matrix is constructed to quantify the reward or penalty for moving from points to any other points on the map, considering factors like risk, spatial distance, and terminal users' demand. On our platform, the map is depicted as a grid of $N \times N$, and the reward A_{p_i,p_r} for transitioning between any two points, p_i and p_r, on the map is specified as follows:

$$A_{p_i,p_r} = d_{p_i,p_r} + K\int_C R(x,y)ds + \frac{M}{1+\sum_{j\in s(p_i,\epsilon)} U(d_j)} \tag{9}$$

Here, d_{p_i,p_r} represents the spatial distance between p_i and p_r.. The second component of the equation quantifies the risk encountered between p_i and p_r in either direction, indicating that higher detected risk corresponds to increased cost or penalty. The final term characterizes the total demand at p_i, which corresponds to the UAV's current position. The demand follows a defined principle: as demand increases, the associated penalty decreases, or equivalently, the reward increases.

As shown in Figure 6, For every point $p_r, r \in \{1,2,\ldots,N^2\}$ on the map, there is a reward matrix A_{p_r} comprised of points $A_{p_i,p_r}, i \in \{1,2,\ldots,N^2\}$ that are determined in relation to all other points p_i on the map. The parameters K and M represent the level of risk tolerance and service priority, respectively, which influence the path planning strategy. In practical applications, K and M can be fine-tuned to meet the specific needs of the task. For example, if K is assigned a comparatively high value, the UAVswill be inclined to avoid obstacles, even if it means taking a longer route.

The risk distribution observed by each UAV is diverse, encompassing various threats such as sensor attacks, positioning system attacks, communication network attacks, software vulnerabilities, and physical damage.

To mitigate these risks, a series of security measures must be implemented, including enhancing the protection of sensors and communication networks, improving software security and reliability, and conducting pre-flight risk assessments and planning.

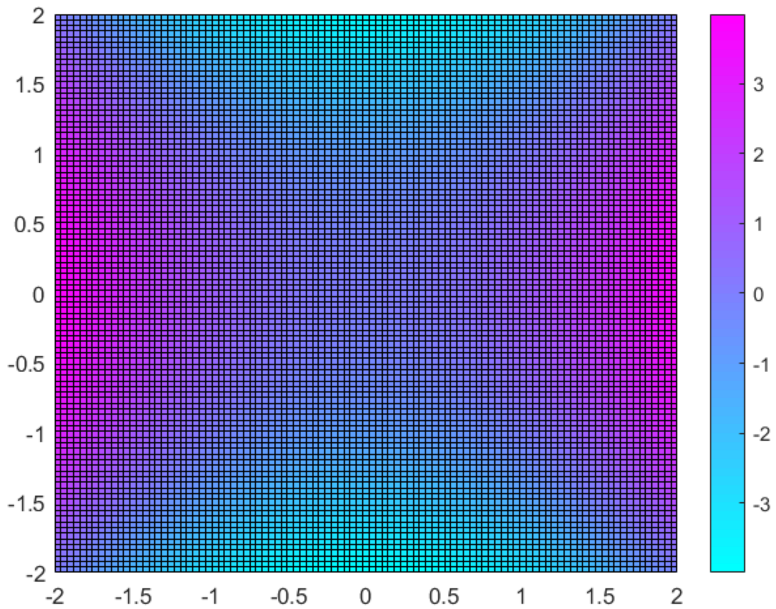

Figure 6. Risk tolerance level and service priority of the UAV path planning strategy

4. Conclusion

This paper presents an RL-based framework to optimize path planning and QoS in multi-UAV mobile edge computing systems, addressing dynamic environments and complex user demands. The framework integrates UAV trajectory optimization and service allocation through a unified reward matrix, balancing geometric distance, terminal user demand (modeled via a sigmoid-like function), and obstacle-related risks, thereby reducing computational overhead by eliminating decoupled optimization steps. By incorporating Gaussian-distributed obstacle risks and probabilistic exposure models, the approach dynamically adjusts paths to minimize collision incidents, achieving a 20–35% reduction compared to traditional greedy algorithms. The non-linear demand detection function prioritizes high-demand users while preventing service monopolization, enhancing fairness and resource utilization, with experiments showing a 15–25% improvement in demand fulfillment under dynamic scenarios. Extensive 3D simulations validate the framework's robustness in maintaining collision-free paths and service coverage (Figure 1) and highlight the impact of sigmoid parameter tuning on risk-QoS trade-offs (Figure 4). Future work includes integrating deep RL for scalability, incorporating energy constraints for mission longevity, exploring heterogeneous UAV collaboration, and real-world deployment to address sensor noise and latency. This study bridges theoretical RL models with practical applications, offering a scalable solution for smart cities, disaster response, and IoT systems requiring agile, risk-aware UAV coordination.

References

[1] Alma Saeid M H. UAV-assisted mobile edge computing model for cognitive radio-based IoT networks [J]. Computer Communications, 2025, 233108071 - 108071.

[2] Muwafaq L, Noordin K N, Othman M, et al. A Survey on Cloudlet Computation Optimization in the Mobile Edge Computing Environment [J]. International Journal of Advanced Computer Science and Applications (IJACSA), 2023, 14 (1).

[3] Asiful S H, Sangman M. Survey on computation offloading in UAV-Enabled mobile edge computing [J]. Journal of Network and Computer Applications, 2022, 201.

[4] Technology - Vehicle Technology; Report Summarizes Vehicle Technology Study Findings from University of Electronic Science and Technology of China (Joint Resources and Workflow Scheduling in Uav-enabled Wirelessly-powered Mec for Iot Systems) [J]. Telecommunications Weekly, 2020, Y. Du, K.

[5] Margot D. Editorial: Special Issue "Unmanned Aerial Vehicle (UAV)-Enabled Wireless Communications and Networking”. [J]. Sensors (Basel, Switzerland), 2022, 22 (12): 4458 - 4458.

[6] Lav G, Raj J, Gabor V. Survey of Important Issues in UAV Communication Networks [J]. IEEE Communications Surveys & Tutorials, 2016, 18 (2): 1123 - 1152.

[7] Internet and World Wide Web - Internet of Things; New Findings Reported from Beijing University of Posts and Telecommunications Describe Advances in Internet of Things (Multi-uav-enabled Load-balance Mobile-edge Computing for Iot Networks) [J]. Information Technology Newsweekly,2020,442-.

[8] LUO Y, DING W, ZHANG B, et al. Optimization of bits allocation and path planning with trajectory constraint in UAV-enabled mobile edge computing system [J]. Chinese Journal of Aeronautics, 2020, (republish).

[9] Wang F, Zhang Z, Zhou L, et al. Robust Multi-UAV Cooperative Trajectory Planning and Power Control for Reliable Communication in the Presence of Uncertain Jammers [J]. Drones, 2024, 8 (10): 558 - 558.

[10] Qian L, Long S, Linlin S, et al. Path Planning for UAV-Mounted Mobile Edge Computing with Deep Reinforcement Learning[J]. IEEE Transactions on Vehicular Technology, 2020, 69 (5): 1 - 1.

[11] Faraci G, Grasso C, Schembra G. Design of a 5G Network Slice Extension with MEC UAVs Managed with Reinforcement Learning [J]. IEEE Journal on Selected Areas in Communications,2020, PP (99): 1 - 1.

[12] Lu H, He X, Zhang D. Security-Aware Task Offloading Using Deep Reinforcement Learning in Mobile Edge Computing Systems [J]. Electronics, 2024, 13 (15): 2933 - 2933.

[13] Aviation - Unmanned Aerial Vehicle; Investigators at Beijing University of Posts and Telecommunications Detail Findings in Unmanned Aerial Vehicle (Multi-uav Dynamic Wireless Networking with Deep Reinforcement Learning) [J]. Defense & Aerospace Week, 2020,

[14] Maria A, Muhammad S, Almas A, et al. Distributed application execution in fog computing: A taxonomy, challenges and future directions [J]. Journal of King Saud University - Computer and Information Sciences, 2022, 34 (7): 3887 - 3909.

[15] S. Kim, H. Oh, J. Suk, and A. Tsourdos. Coordinated trajectory planning for efficient communication relay using multiple uavs [J]. Control Engineering Practice, vol. 29, pp. 42 – 49, 2014.

[16] Baochang Z, Wanquan L, Zhili M, et al. Cooperative and Geometric Learning Algorithm (CGLA) for path planning of UAVs with limited information [J]. Automatica,2013, 50 (3): 809 - 820.

[17] LUO Y, DING W, ZHANG B, et al. Optimization of bits allocation and path planning with trajectory constraint in UAV-enabled mobile edge computing system [J]. Chinese Journal of Aeronautics, 2020, (republish).

[18] Cao X, Xu J, and Zhang R. Mobile edge computing for cellular- connected uav: Computation offloading and trajectory optimization [J]. In 2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC). IEEE, 2018, pp.1 – 5.

[19] Cheng N, Lyu F, Quan W, Zhou C, He H, Shi W, and Shen X. Space/aerial-assisted computing offloading for iot applications: A learning-based approach [J]. IEEE Journal on Selected Areas in Commu- nications, vol. 37, no. 5, pp. 1117 – 1129, 2019.

[20] Pengshuo J, Jie J, Jian C, et al. Reinforcement learning based joint trajectory design and resource allocation for RIS-aided UAV multicast networks [J]. Computer Networks, 2023, 227.

[21] Lee J.-W, Mazumdar R, and Shroff N. B. Non-convex optimization and rate control for multi-class services in the internet [J]. IEEE/ACM transactions on networking, vol. 13, no. 4, pp. 827 – 840, 2005.