IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 December 2022, accepted 10 January 2023, date of publication 13 January 2023, date of current version 2 March 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3236801*

SURVEY

# Reinforcement Learning in the Sky: A Survey on Enabling Intelligence in NTN-Based Communications

**TAREK NAOUS[1,4], MAY ITANI[2], MARIETTE AWAD[1], (Member, IEEE), AND SANAA SHARAFEDDINE[3], (Senior Member, IEEE)**

[1]Electrical and Computer Engineering Department, American University of Beirut, Beirut 1107-2020, Lebanon
[2]Mathematics and Computer Science Department, Beirut Arab University, Beirut 1107-2809, Lebanon
[3]Department of Computer Science, American University of Beirut, Beirut 1107-2020, Lebanon
[4]Georgia Institute of Technology, Atlanta, GA 30332, USA

Corresponding author: Mariette Awad (mariette.awad@aub.edu.lb)

This work was supported in part by the Maroun Semaan Faculty of Engineering and Architecture.

**ABSTRACT** Non terrestrial networks (NTN) involving 'in the sky' objects such as low-earth orbit satellites, high altitude platform systems (HAPs) and Unmanned Aerial Vehicles (UAVs) are expected to be integral components of next generation cellular systems. With the deployment of 5G services and beyond, NTNs are leveraged to assist as aerial base stations in providing ubiquitous network connectivity and service to ground users or be deployed as aerial users connected to the cellular network. NTN-aided wireless communication offers multiple benefits such as mobility, flexibility, resistance to ground physical attacks and wide coverage. However, due to their limited resources and the current design of terrestrial cellular systems that do not account for aerial users, and other restrictions such as service requirements, limited available power and storage resources on high-throughput satellites, resource allocation, location of the high altitude platform base station and the flight trajectory of the UAVs need to be intelligently controlled to satisfy various objectives both from an aerial base station and overall network perspectives. To achieve this, many works have explored Reinforcement Learning (RL) techniques to allow aerial platforms in non-terrestrial networks to learn from past observations and achieve some optimal control policy. In this paper and differently from prior surveys, we contribute a comprehensive review of the control objectives required by non-terrestrial platforms that have been solved using RL formulations. We provide an up-to-date overview of the latest applications of RL techniques for different NTN-aided wireless communication aspects. The survey focuses on Markov Decision Process (MDP) formulations in terms of states, actions, and rewards. We synthesize a taxonomy from the surveyed literature and provide a comprehensive representation of the current usages of RL in NTN-aided wireless communications. A qualitative analysis of the level of realism achieved in the works presented in the literature is provided based on several factors that pertain to the simulation environment, station deployment setting, wireless channel assumption, and energy considerations. We also curate a list of challenges that remain to be considered by the research community in order to achieve more efficient deployments and close the simulation-to-reality gap.

**INDEX TERMS** Reinforcement learning, non terrestrial networks, satellite communication, high altitude platforms, NTN, NTN-aided communication, AI-enabled communications.

## I. INTRODUCTION

NTNs have witnessed an increased interest over the last few years and are expected to become a key part of next-generation wireless communication. With the rapid growth of wireless communication systems, terrestrial base stations are challenged to provide connectivity and performance requirements including throughput, latency and energy efficiency especially in rural areas, deserts and oceans, harsh and remote environments [1].

The associate editor coordinating the review of this manuscript and approving it for publication was Olutayo O. Oyerinde.

Many recent work has been published on the integration of space and terrestrial networks that involve flying objects including satellites, high altitude platforms, and UAVs [2]. Research has been mainly focused on the usage of NTNs in general and UAVs in specific and their integration into cellular networks as either *NTN-aided wireless communications* [3] or *cellular-connected NTNs* [4], [5]. Fig. 1 captures a comparative illustration of these two integration scenarios into cellular networks. In NTN-aided wireless communications, NTNs are deployed as aerial base stations to assist the cellular network infrastructure that is required to keep up with the exploding service demand for a higher quality of wireless services. Cellular-connected NTNs, on the other hand, are deployed as user equipment in the air, enabling an unlimited operation range and world-wide accessibility through the cellular network. It is worth noting that the majority of existing literature explores UAVs as part of the latter two scenarios, while research interest is being shifted towards aerial platforms in general.

Mobility of non-terrestrial base stations results in a dynamic unstable environment imposing challenges in the coverage optimization. Moreover, flexibility of aerial platforms has led to carrying out more research to explore the potential of NTNs in optimizing various performance metrics in wireless communication such as SNR, data rate, power and time consumption.

UAVs can be equipped with light-weight base station equipment and act as aerial base stations in challenging scenarios including hard-to-reach areas and emergencies when terrestrial base stations are damaged. This is also applicable to scenarios where the terrestrial network infrastructure becomes incapable of meeting the stringent demand for wider coverage, higher capacity, and better service quality such as large crowd gatherings and hotspot areas. Hence, such non-terrestrial platforms in general can be useful for the on-demand assistance of cellular communication networks and the mitigation of the unexpected surge in cellular traffic and its implication on the network performance [6]. Meanwhile, cellular-connected NTNs make use of the already available cellular network infrastructure for various purposes such as package delivery, search and rescue operations, building inspections, security surveillance, live streaming of events, and many others [7]. Thus, cellular-connected non-terrestrial base stations can be controlled in a very wide operation range without the need to build a new infrastructure dedicated to a given service. This type of non-terrestrial platform integration has therefore become a very attractive technology for the industry due to the possibility of enabling a wide range of applications [8].

In both of the aforementioned integration scenarios, non-terrestrial platforms including satellites, HAPs and UAVs face challenging objectives that need to be satisfied. These objectives are, among others, maximizing quality of service, minimizing energy consumption, guaranteeing connectivity between the core network and ground users, and avoiding interference. In this respect, UAVs, for example, are required

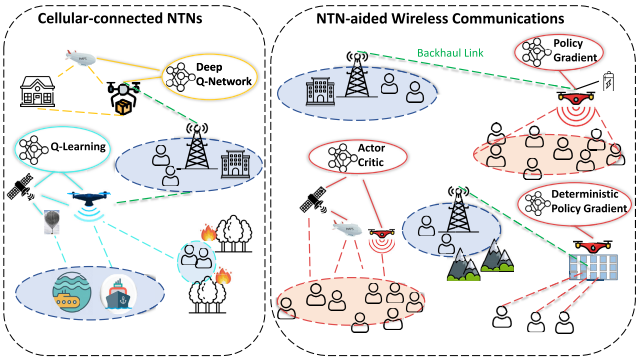



**FIGURE 1. A comparative illustration of cellular-connected NTNs and NTN-aided wireless communications.**

to optimize their flying trajectory to deliver the desired service and meet the performance criteria, while being cognizant of the system constraints. The need for sophisticated algorithms to assist in the decision-making and achieving various goals is therefore inevitable. However, the efficient control of the non-terrestrial platform resources and mobility is a complex problem, especially in highly uncertain scenarios where user information cannot be predicted reliably due to the unavailability of dedicated control channels for information exchange, or simply due to the unavailability of information. Conventional mathematical optimization approaches may not converge within the desired time range to the optimal solution of these problems that are in most cases non-convex, and hence sub-optimal approaches are usually applied to obtain results. Nonetheless, the latter approach may not be feasible or practical after all due to the unavailability of its input data in uncertain environments. Recently, reinforcement learning (RL) algorithms have found their way into various applications in both NTN-aided wireless communications and cellular-connected NTNs. Most of the design problems of NTNs in general can be formulated as a Markov Decision Process (MDP). To solve this MDP, many works in the literature have used a variety of RL techniques for different objectives in NTN communications. This is shown in Fig. 1 where each non-terrestrial platform acts as an RL agent that leverages past observations and rewards to reach an optimal control policy.

We note that several surveys in the literature have addressed different aspects of the integration of NTN-platforms in cellular networks. Comprehensive tutorials on non-terrestrial networks including space and air-borne platforms in general were presented in [2], [9], [10], [11], [12], [13], [14], and [15] and they illustrated how space-air-ground networks can be integrated in 5G/6G systems yielding a heterogeneous network architecture that involves non-terrestrial stations (satellites, HAPs, UAVs) assisting terrestrial ones. In [16], [17], and [18], the convergence of satellite and terrestrial networks was surveyed and different architectures were presented, while in [19], satellite communication applications were explored. Related surveys in [20] and [21] present a recent review of wireless

communications involving High-Altitude Platforms (HAPs) in rural areas exploiting cellular radio spectrum. In [22], authors present services that could be provided by considering cloud-enabled HAPs as flying data centers. All of the afore-mentioned surveys have no focus on RL. Few surveys have considered machine learning (ML) techniques, those of which include RL in the context of wireless IoT [23] or 5G network slicing [24]. Others included subsections related to RL as open issues and challenges [25]. In [26], authors reviewed artificial intelligence techniques in general as applied to satellite communications. By all means, the literature is rich in surveys that study UAV-assisted communications as compared to surveys addressing other forms of non-terrestrial platforms, thanks to the agility and practicality of deploying UAVs to assist terrestrial networks in critical situations and in enabling novel services. Surveys dedicated to UAV communications and their applications were presented in [27], [28], [29], [30], [31], [32], and [33], where they highlighted how UAVs are expected to be integrated in fifth-generation (5G) wireless networks and beyond. In [34], the challenges in UAVs standardization were discussed, and a set of regulations were proposed for their integration into society. An extensive overview of software-defined networking and network function virtualization in UAV-assisted systems is presented in [35]. The routing demands and protocols required for UAVs are detailed in [36], along with the associated challenges. In [3], an overview of the networking architecture of UAV-aided wireless communications is provided, along with key design considerations. Surveys on trajectory design techniques for UAVs are provided in [37] and [38]. However, these latter surveys have limited focus on RL-based approaches and the challenges associated with them. In [39], a survey on UAV-aided Internet of Things (IoT) networks is presented. Game-theoretic formulations for objectives in UAV communications are reviewed in [40] while machine learning techniques for UAV-based communications are presented in [41], [42], [43], [44], [45], and [46], also with little focus on RL techniques in specific. The scope of the existing surveys in terms of their focus on RL-based problem formulations is shown in Fig. 2. Surveys labeled with 'NTN', 'S,' and 'H' respectively represent surveys related to non-terrestrial platforms in general, Satellite in specific, HAPs in specific, and UAVs in specific.

Additionally, surveys on the applications of RL in communications and networking are provided in [47], [48], [49], [50], [51], and [52], but they have a limited focus on applications for non-terrestrial platforms.

While these many surveys have discussed the current state-of-the-art of different non-terrestrial platforms and UAVs in specific, no survey has previously addressed the applications of RL for intelligent NTN communications. Specifically, no survey has already provided a comprehensive review of the control objectives required by satellites, HAPs and/or UAVs in NTN-assisted communication problems that have been addressed using RL formulations. In this regard, our survey is the first to bridge that gap and present an up-to-date discussion on RL for NTN-aided wireless communications as well as cellular-connected NTNs. We cluster the literature around different integration categories that constitute (i) improving network key performance indicators (KPIs), (ii) maintaining reliable integrated access and backhaul links, (iii) improving data integrity and security, and (iv) minimizing the age of information (AoI) in information dissemination and data collection applications under NTN-aided wireless communications. In the context of cellular-connected NTNs, three main categories are defined constituting (i) enhanced connectivity, (ii) interference management, and (iii) spectral management. We then synthesize a taxonomy from the literature based on what control objective is considered in each RL problem formulation. The developed taxonomy gives a complete representation of what the current applications of RL are in NTN communications. We, then, discuss challenges for adopting RL for different objectives in NTN communications and aim to set a basis for future directions and insights to potentially further improve effective real-world deployment.

The rest of this survey is organized as follows: A brief overview on RL is provided in Section II, covering some basic fundamental concepts. In Section III, we briefly introduce the control challenges in NTN-assisted networks and present a taxonomy of RL objectives in NTN communications. Section IV surveys the literature that employs RL techniques for NTN-assisted wireless networks. Section V surveys the works that propose RL-based solutions for various challenges in cellular-connected NTNs. A qualitative analysis on the level of achieved realism in the surveyed literature is provided in Section VI. A discussion on remaining challenges and insights for future research directions is presented in Section VII with a focus on bridging the gap between simulation and real-world environments. Finally, concluding remarks follow in Section VIII.

## II. AN OVERVIEW ON REINFORCEMENT LEARNING

Poised to be the next stage in the evolution of machine learning algorithms that learn how to learn, RL is a subfield within artificial intelligence where the learner, referred to as an agent, learns how to map situations to actions in a way that maximizes a numerically-defined reward function. In an RL setting, the agent is not given any prior knowledge on what actions it should take. Instead, it interacts with the environment and explores different actions in different situations, called states, to discover which decisions will yield the most reward. Moreover, the concepts of delayed reward and trial-and-error search are important as they allow RL discount meaningless reward in anticipation of a longer term gain and explore solutions without being fixated on the exploitation of the knowledge it accumulated. Based on its interactions with the environment, the agent learns from its past actions and experiences and becomes better in future decision making [53].

RL distinguishes itself from other learning paradigms, such as supervised learning approaches, that rely on instructive

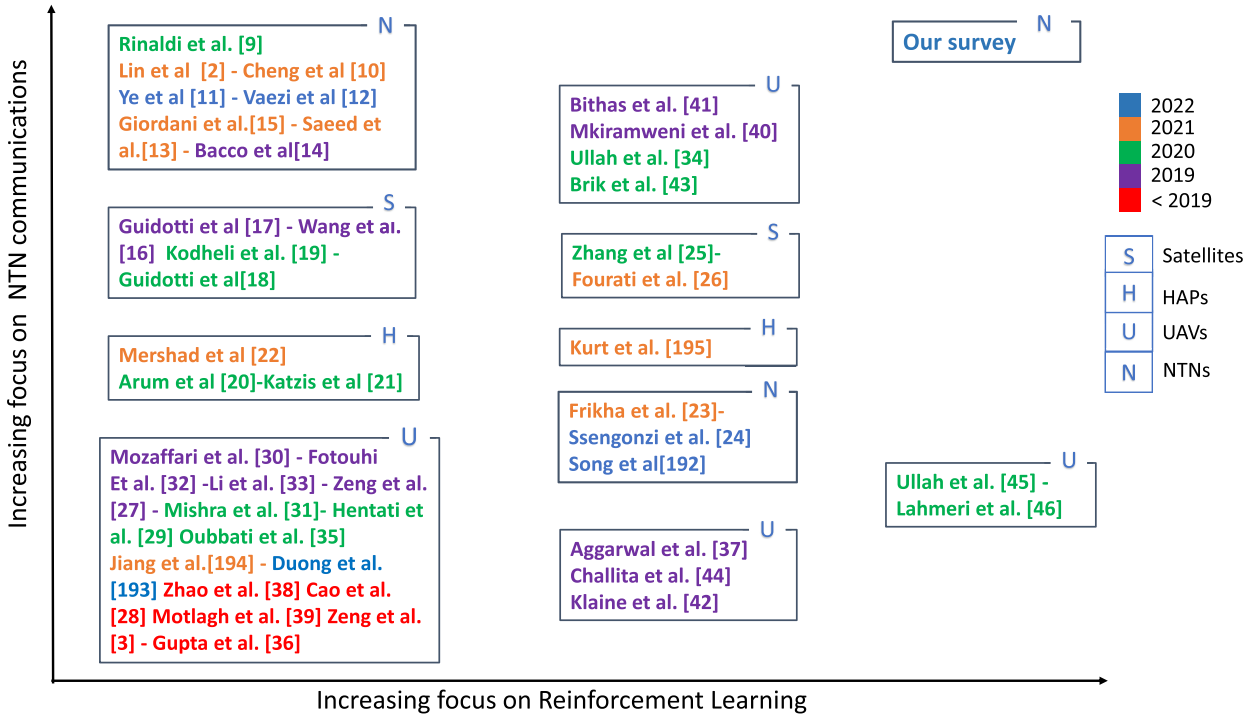



**FIGURE 2.** Visualisation of the scope of related surveys in terms of increasing focus on NTN communications and RL formulations.

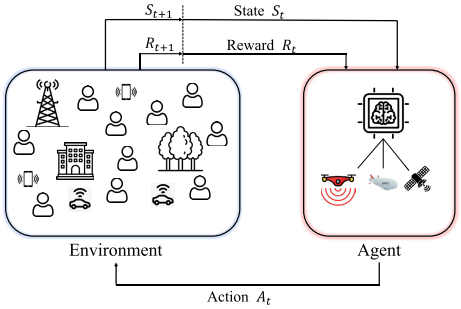



**FIGURE 3.** The interactions of the agent and the environment in an MDP.

feedback instead of evaluative feedback. Instructive feedback indicates what action is correct to take, independently from the action that has been taken, while purely evaluative feedback gives insights on how good the action performed by the agent was. Table 1 compares RL to other learning approaches. RL is distinct from other machine learning paradigms in that the lack of supervision for the optimal solution is substituted by a choice and a feedback in a dynamic environment which makes RL an active learning process [54].

A mathematical idealization of the RL problem is the Markov Decision Process (MDP), a discrete-time stochastic control process that is generally used as a framework for sequential decision-making algorithms. It satisfies the Markov property which states that a future state relies only on the present state and is independent of the past states. An MDP is represented by a five-element tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ where:

- $\mathcal{S}$ represents the set of states $s$ of the environment
- $\mathcal{A}$ represents the set of actions $a$ that the agent can take
- $\mathcal{P}$ represents the transition probability function. Specifically, at a time step $t$, $\mathcal{P}$ determines the probability of going from state $S_t$ to state $S_{t+1}$ when action $A_t$ is performed
- $\mathcal{R}$ represents the reward function that gives the agent a reward when transitioning from state $S_t$ to state $S_{t+1}$ by performing action $A_t$
- $\gamma$ is the discount factor that can take on a value between 0 and 1

The agent-environment interactions in an MDP are shown in Fig. 3, where at each time step $t$, the agent receives a representation of the environment's state $S_t \in \mathcal{S}$. Based on this state, the agent performs an action $A_t \in \mathcal{A}$. At the subsequent time step, the agent receives a numerical reward $R_{t+1} \in \mathcal{R}$ and transitions into a new state $S_{t+1}$.

The solution of an MDP is a policy function $\pi$ that maps states to actions ($\pi : s \rightarrow a$). The goal of the agent is to find the optimal policy $\pi^*$ by maximizing the total reward it receives, that is the cumulative reward and not the immediate reward. The cumulative reward is represented by the discounted expected return, denoted $G_t$, that is computed using:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+2} + \dots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \quad (1)$$

**TABLE 1. RL Compared to Other Learning Approaches.**

| Key Feature | AI Planning | Supervised Learning | Unsupervised Learning | Reinforcement Learning | Imitation Learning |
|---|---|---|---|---|---|
| Optimization | ✓ | | | ✓ | ✓ |
| Learns from experience | | ✓ | ✓ | ✓ | ✓ |
| Generalization | ✓ | ✓ | ✓ | ✓ | ✓ |
| Delayed Consequences | ✓ | | | ✓ | ✓ |
| Exploration | | | | ✓ | |

where $\gamma$ is referred to as the discount rate. The concept of discounting is essential to make the agent select actions that maximize the expected return $\mathbb{E}[G_t|s, \pi]$ by assigning weights to the cumulative set of rewards. If $\gamma$ is selected to be 0, the agent is said to be myopic or short-sighted and only focuses at immediate rewards. If $\gamma$ becomes closer to 1, the agent is said to be more far-sighted and weighs future rewards more strongly in its decision making.

RL agents are categorized as (i) value-based that have a value function and implicit policy, (ii) policy-based that maintain a data structure of every state without storing value function, or (iii) actor-critic that combine both policy and value functions. As for RL algorithms they can be categorized as (i) model-free where the agent learns directly by collecting rewards from the environment then updating their value function estimation thus figuring out the policy or (ii) model-based where an RL agent is involved and no need for direct environment interaction since the agent learns the model which consists of state transitions and reward function. Policy is then figured out with simple information about state values. Note that in model-based scenario solution may fail if the state space is too large [23], [54]. Meta-RL algorithms including model-agnostic meta-learning (MAML), Simple Neural AttentIve Learner (SNAIL) and Proximal Meta-Policy Search (ProMP) algorithms are more recent RL algorithms that emerged in years 2017 and 2018 where the agent is trained over a variety of distributed tasks and tries to solve new related unseen tasks from the knowledge it learns [55], [56], [57]. Fig. 4 shows selected RL algorithms from model-free and model-based categories where the lower taxonomies in a branch are the most recent ones [58], [59], [60], [61], [62], [63], [64], [65], [66], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76], [77].

Below we provide a brief overview on policy gradient and Q-based learning which constitute the basics of RL algorithms as indicated in Fig. 4

- ***Policy Gradient:*** In policy gradient based methods, the policy is directly tuned after being parameterized with respect to the expected long term cumulative reward by gradient descent. By adopting a stochastic policy, various actions that yield different trajectories are sampled to check those that yield the best rewards and update the policy direction parameters. Policy gradient methods do not suffer from the lack of guarantees of a value function, the intractability problem that results from uncertain state information and the complexity arising from continuous states-actions [78].
  Policy Gradient was first introduced in 2014, and its variants Asynchronous Advantage Actor-Critic, Proximal Policy Gradient (PPO) and Maximum a Posteriori Policy Optimization (MPO) in 2016, 2017 and 2018 respectively. MPO combines the sample efficiency of off-policy methods with the scalability and robustness of on-policy methods. It achieves state of the art results on continuous control tasks while using fewer order of magnitude samples than PPO [61].
- ***Q-Based Learning*** is an off-policy temporal difference (TD) learning algorithm. TD learning is a combination of Monte Carlo ideas and dynamic programming (DP) concepts. Q-learning is widely used for model free problems. The learned action-value function, Q, directly approximates the optimal action-value function, independent of the policy being followed thus simplifying the algorithm and enabling early convergence. However, with large Q-tables or infinite spaces, the algorithm will take long to converge and becomes impractical [53].
  Deep Q- Networks which consist of Q-Learning with deep neural networks as state-action value estimators and use replay buffers to sample experiences from previous trajectories were first introduced in 2013 [65]. Categorical 51 with Hindsight Experience Replay (HER) were introduced in 2017. HER provides efficient learning without the need for complicated reward engineering [67]. Recurrent Replay Distributed DQN (R2D2) was introduced in 2019 and was the first agent to exceed human-level performance in 52 out of 57 Atari games as demonstrated in [68].

Integration of non-terrestrial base stations (NT-BSs) with terrestrial networks implies a heterogeneous dynamic environment (due to NT-BSs mobility) imposing new challenges different from terrestrial wireless communication requirements. Many recent work to solve NTN wireless control and management problems such as channel estimation, joint beam forming, resource allocation, multi-user access control, trajectory and power optimization are being motivated by and based on ''RL Techniques'' since the latter techniques rely on systematic trial and error. Application of RL methods have a showed an increased potential in building low latency, ultra-reliable, and scalable systems for future wireless generations including IoT networks [23], [50], [52], [79]. In [80] the RL approach outperformed benchmark learning approach by 33.85% in terms of improving the network throughput, and by 95% in terms of enhancing the energy efficiency. Compared to a non learning-based approach,

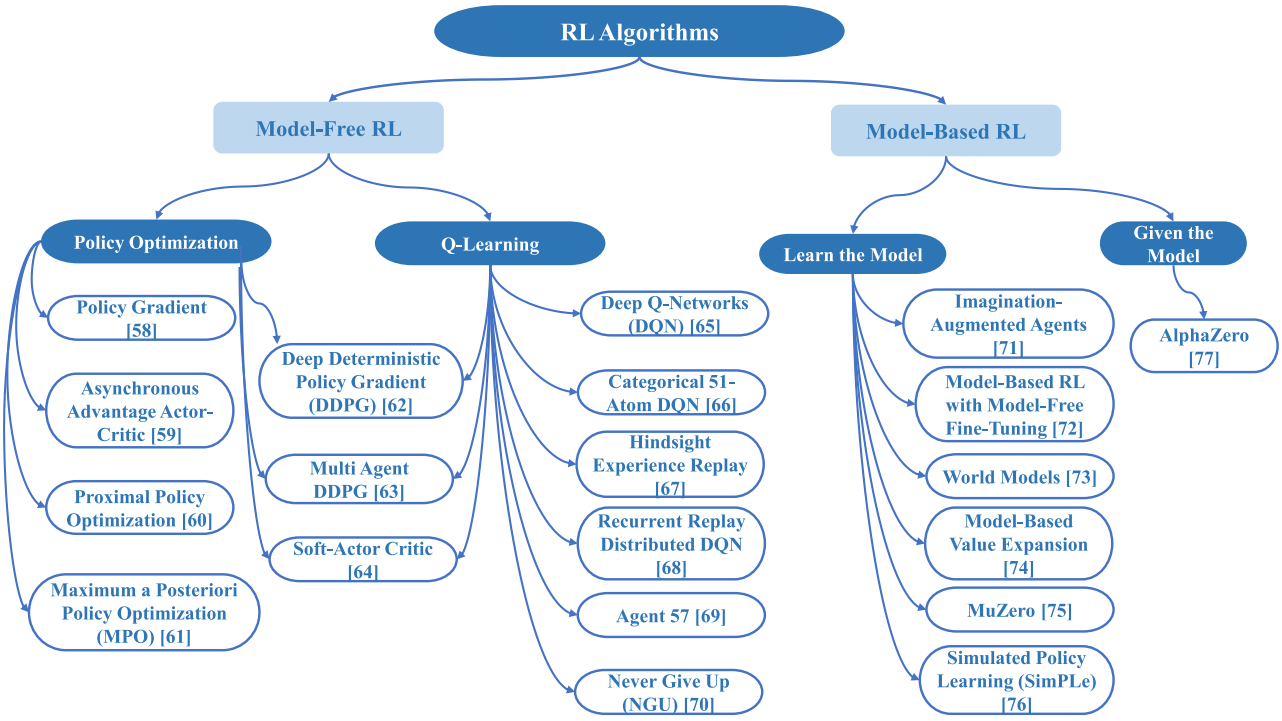



**FIGURE 4.** Selected RL Taxonomies.

RL improves the throughput and energy efficiency by 46.61% to approximately 110%.

## III. CONTROL CHALLENGES & TAXONOMY

As stated earlier, the main non-terrestrial platforms are classified as satellites, HAPs and UAVs. To support various applications in the 6G era, a wide area network integrating non-terrestrial and terrestrial networks is needed to deliver the desired control objective(s). Each NTN platform, however, has its own significant role. In fact, to support orbit or space Internet services and provide wireless coverage for flight applications, low-Earth-orbit, medium-Earth-orbit, and geostationary-Earth-orbit satellites are to be deployed. Satellites with mm-wave communications are also utilized for high-capacity satellite–ground transmission. Floating and flying base stations known as HAPs and UAVs respectively are, however, installed to provide coverage and reliability in rural and hard-to-reach areas. Floating base stations (HAPs) usually assist space networks and reachable UAVs [81]. Flying base stations (UAVs), on the other hand, are the most significant NTN component platforms and are considered a promising technology to assist future wireless communications due to their flexibility, swiftness and low-cost features. UAVs have been regarded as a solution of aerial networking and a complement of terrestrial communication infrastructure by $3^{rd}$ Generation Partnership Project (3GPP) Long Term Evolution-Advanced (LTE-A). They have a stronger line of sight connection with ground users, a better mobility that provides real-time and on demand services in critical situations as in floods and hurricanes, flexibility and lower cost to enhance specific terrestrial links such as cellular network links in sport stadiums, and others [82].

### A. NTNs TO ASSIST WIRELESS NETWORKS

Non terrestrial platforms can be extremely useful in assisting the wireless network as aerial base stations or relays, given their ability to establish a dominant LoS feature to ground users and their high agility and mobility features. However, flying platforms with flexible routes including UAVs may need to adjust their positions and trajectories to optimize their intended service, such as ensuring that several areas are receiving coverage and service for a specified duration. The challenge of determining an adaptive trajectory becomes more compound when the environment is stochastic, such as having mobile users in vehicles, or users with dynamic access demands. Additionally, when multiple non terrestrial platforms are deployed, cooperative coordination needs to be ensured among them to reach the desired objective. In this regard, autonomous non terrestrial base station deployment and trajectory optimization specifically of UAVs used as aerial base stations or relays is extremely important for the full exploitation of the potential of NTNs in assisting cellular networks. Given their mobility, non terrestrial platforms can adjust their locations to achieve various control objectives. For instance, the HAP and/or UAV could be required to optimize their/its trajectory to provide a target coverage to ground users or adjust its location to maintain favorable channel conditions and provide a better service and experience to users. Another critical factor that needs to be considered is the limited energy and storage resources of

high throughput satellites (HTSs) where resource allocation got to be optimized to enhance the performance of the HTS based communication system. Nonetheless, battery-powered UAVs cannot keep flying around for a long duration before they need to move to a charging station. The trajectory of the UAV also needs to be optimized to maximize its utility while preserving energy resources and prolonging network lifetime. It is worth noting that major investments are dedicated to improving the endurance of aerial platforms including extending their lifetime. Solar energy harvesting and laser beaming are example techniques to provide non terrestrial platforms with sustainable energy sources.

### B. NTNs AS AERIAL USERS

Nonterrestrial communications have emerged to support high data rate communications among aerial platforms (satellites, HAPS and UAVs) and cellular networks, achieving anywhere and anytime connections. Cellular-connected NTNs leverage the ubiquitous accessibility of cellular communication networks to enable various new NTN applications. These applications include search and rescue operations, package delivery, streaming of live events, security surveillance, edge computing, and many others [11], [31]. Indeed, cellular-connected NTNs are a promising technology that offers many potential benefits. However, many practical implementation limitations exist including dynamic propagation environment, overload energy consumption issues and high probability of blockage [11] and other challenges [83] when integrated into the existing cellular network infrastructure. At present, terrestrial base stations are designed to provide reliable connectivity to ground users, without considerations for the aerial user equipment. The antennas of current terrestrial base stations are down-tilted to maximize the coverage probability for users on the ground level or within buildings. Aerial platforms specifically UAVs are required to cleverly optimize their navigation to coordinate with HAPs and satellites and to take advantage of the existing infrastructure to maintain reliable connectivity to the network [84], which is critical for their command, control, and data communications with terrestrial base stations.

Since UAVs enjoy more favorable propagation conditions as their altitude increases, their link to the serving base station becomes stronger with the increase in altitude [85], [86]. However, this fact is also a limiting factor for cellular-connected UAVs. As UAVs hover at a higher altitude, they also start receiving signals from an increasing number of base stations that they have dominant LoS links to. This leaves them prone to aggregate interference which can dominate over the increased received signal power from the serving base station [87], [88]. The LoS-dominated links of cellular-connected UAVs also cause another issue, that is the increased number of unnecessary handovers [89]. This effect could be mitigated by optimizing the altitude of the UAVs. Hence, cellular-connected UAVs require intelligent navigation and height optimization policies to assist them in achieving their objective efficiently.

### C. TAXONOMY OF RL FOR NTN COMMUNICATIONS

To address the aforementioned control challenges in both non-terrestrial platform integration scenarios, many works in the literature have leveraged RL techniques as effective strategies in reaching optimal control policies. The richness of the literature in RL formulations inspires the synthesis of a taxonomy of RL objectives in NTN communications. Our proposed taxonomy is presented in Fig. 5 and clusters the objectives of RL under the two broad categories of NTN-aided wireless communications and cellular-connected NTNs.

The objectives of RL for designing the trajectory of non terrestrial platforms specifically UAVs deployed to assist wireless communication networks can be classified into five main categories. These categories are namely improving network key performance indicators (KPIs), maintaining reliable integrated access and backhaul links, improving data integrity and security, and minimizing the age of information (AoI) in information dissemination and data collection applications. The approaches for improving network KPIs can be clustered into two sub-categories, which are enhancing coverage for ground users or improving the quality of service (QoS) or quality of experience (QoE) of the ground users. Likewise, two sub-categories fall under the data integrity and security category. These sub-categories separate the works that focus on combating scenarios where terrestrial base stations are jammed or aim to combat ground eavesdroppers within the network.

In the context of cellular-connected NTNs, the objectives of RL techniques can be classified into three main categories, namely enhanced connectivity, interference management, and spectral management. The works that focus on enhancing the connectivity of cellular-connected NTNs to the wireless network are also separated into two classes which are coverage hole avoidance, and hand-over rate reduction.

## IV. RL FOR NTN-AIDED WIRELESS COMMUNICATIONS

Various control objectives exist that require optimization of the trajectory of the high and low altitude platforms acting as base stations. To achieve the optimal trajectory design, RL approaches have been explored in several works in the literature and have shown great promise in reaching high performance. In what follows, we present the literature that leveraged RL algorithms for control objectives in NTN-assisted cellular communications. We focus on the MDP formulations proposed in terms of states, actions, and reward functions. A summary of these formulations is provided in Table 2 where we highlight which works addressed finite or infinite state and action spaces.

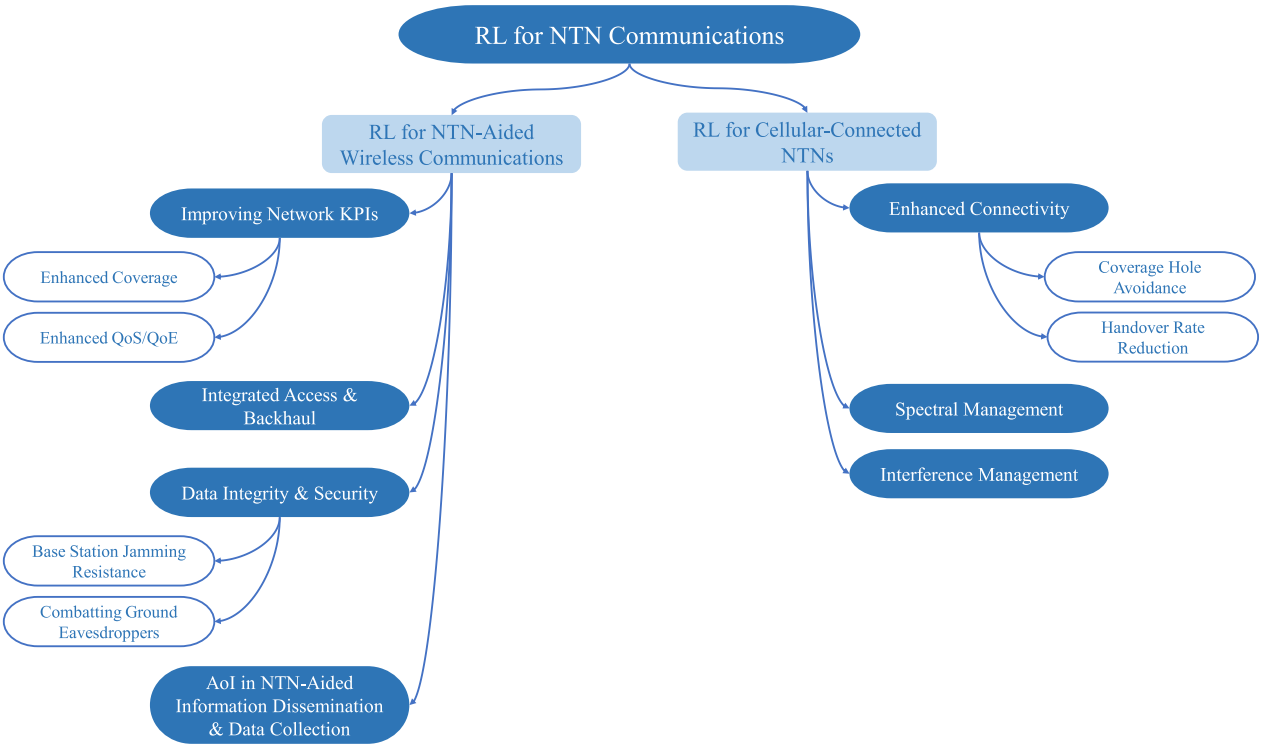



**FIGURE 5.** Our proposed taxonomy for the applications of RL in NTN Communications.

### A. IMPROVING NETWORK KEY PERFORMANCE INDICATORS

#### 1) ENHANCED COVERAGE

Mobility of non terrestrial base stations (NT-BSs) and non terrestrial user equipment (NT-UE) leads to a dynamic non-stationary environment, and creates unique challenges in the coverage optimization specifically in deployment of multiple non terrestrial base stations. In this regards, Lien et al. [90] proposed a reinforcement learning (RL) scheme where multiple NT-BSs autonomously determine deployment trajectories to maximize the number of NT-UEs that can access NTBSs. Anicho et al. [91] work analyzes the performance of Reinforcement Learning (RL) versus Swarm Intelligence (SI) for coordinating multiple unmanned High Altitude Platform Stations (HAPS) for communications area coverage. It builds upon previous work which looked at various elements of both algorithms. The main aim of this paper is to address the continuous state-space challenge within this work by using partitioning to manage the high dimensionality problem. This enabled comparing the performance of the classical cases of both RL and SI establishing a baseline for future comparisons of improved versions. From previous work, SI was observed to perform better across various key performance indicators. However, after tuning parameters and empirically choosing suitable partitioning ratio for the RL state space, it was observed that the SI algorithm still maintained superior coordination capability by achieving higher mean overall user coverage (about 20% better than the RL algorithm), in addition to faster convergence rates. Though the RL technique showed better average peak user coverage, the unpredictable coverage dip was a key weakness, making SI a more suitable algorithm within the context of this work. Another setting constitutes hybrid satellite networks where UAVs serve as relay mobile base stations to enhance satellite terrestrial communication. Moreover, lightweight base station equipment can be mounted on UAVs to provide coverage in areas of the cellular network where coverage is poor, or when the terrestrial base station is down or non-existent. Given the diverse distribution of users, the challenge of the non terrestrial platform is to maximize the number of users covered. Huang et al. [92] proposed a Deep Q Network (DQN) model to optimize the navigation of 32 UAVs acting as aerial base stations. The state space was represented by the received signal strengths, while the reward was determined by the Signal to Interference and Noise Ration (SINR) of the UAVs. The SINR was chosen to determine the reward since it varies with the change in location of the UAVs. Hence, the UAVs will vary their locations in a way to maximize the long-term expected reward. A three-dimensional user space was considered by Liu et al. [93] where user equipment can have various altitudes. Such a simulation environment is important to model the real-world scenarios where users may be on the ground level or in high buildings and skyscrapers. The double Q-learning algorithm was used to maximize the total number of served users and was selected over standard Q-learning to overcome its drawback of overestimation. The state of the UAV was represented by several vectors that describe the situation of each user in terms of receiving service, the

maximum time the user can wait for a service, and the time a UAV needs to fly to a user. The action of the UAV is specified as the provision of service to a user, while the reward is the total number of served users. A Double Deep Q-Network (DDQN) with Prioritized Experience Replay (PER) was proposed by Qiu et al. [94] to find the optimal locations of UAV base stations that maximize the coverage rate, defined as the number of ground users covered to the total number of users, given the constraint of possible blockage in the air-ground channel. The state of the UAV was defined by a coverage bitmap that represents the spatial correlation between the UAVs and the users and provides information on the total coverage. The UAV changes its moving direction to maximize the long-term expected reward. To prevent the UAV from flying beyond the borders of the area considered in the simulation, the authors defined a negative error function in the reward to penalize the agent for such behavior.

Liu et al. [95] adopted a deep RL approach to enable energy-efficient control of UAVs while providing fair coverage and connectivity to ground users. The Deep Deterministic Policy Gradient (DDPG) actor-critic method was chosen to handle the continuous control problem with an unlimited action space. A network of multiple UAVs was controlled via a deep RL agent that sends command signals to orchestrate the UAVs based on the observations it receives. The state of the agent was defined by the coverage score and coverage state of each cell in the network, which are metrics defined by the authors to represent whether a cell is receiving fair coverage or not. Energy efficiency was ensured in this formulation by considering the energy consumption of each UAV as a part of its state. The authors assumed that the UAV only consumes energy as it hovers from one location to another. The action was defined as the angle or flying direction of each UAV and its flying distance. The reward was an energy efficiency equation that the agent needs to maximize. This multi-UAV setting was extended by Liu et al. [96], where each UAV not only acts as an aerial base station to serve ground users but also as a hotspot for the other UAVs. The state of the agent was modified to include the positions of all UAVs, and their flying directions. Additionally, the authors ensured the UAVs remain connected to each other by including the UAV's distance to the other agents in its state and penalizing the agent, via the reward it receives, when these distances fall behind a pre-defined threshold.

Anicho et al. [91] a reinforcement learning method to solve the coordination problem of multiple unmanned high altitude platform stations (HAPs) is compared to swarm intelligence where reinforcement learning showed better average peak user coverage. The authors implement a classical Q-learning method where HAPs are considered as agents and user mobility is considered a part of the environment and states are mapped to predefined fixed coordinates. HAPs adjust their positions to achieve higher mean overall user coverage. Lien et al. [90] authors propose k-step SR QD-learning scheme where each NT-BS constituting either HAP or UAV in a multiple NT-BSs scenario autonomously determines the deployment trajectory to maximize the number of NT-UEs that can access the non terrestrial base station. In [97], Chen et al. first allowed optimal link selection via a designed graph neural network (GNN), and then adjusted the UAV locations by using model-free reinforcement learning (RL). The state of the UAV is composed of its location, embedding features, and energy consumption and the action consists of its direction and moving distance. Whereas the instantaneous reward received by a specific UAV is defined as the coverage at time t.

#### 2) ENHANCED QoS/QoE

Several RL problem formulations have been proposed to improve the QoS and QoE of users. Yin et al. [98] considered the maximization of the uplink sum rate using the Deterministic Policy Gradient (DPG) with no access of the UAV to user-side information such as transmit power or location. The state of the UAV was represented by the time difference between received signal strengths at each time slot. The UAV changes its movement represented by spherical coordinates (step size, elevation, and azimuth angles) to maximize the long-term reward defined as the uplink sum rate in each time slot. A similar Q-learning-based approach was proposed by Bayerlein et al. [99] where the state of the UAV is represented by its current position and time. In this formulation, the agent moves its location in four possible directions to maximize the reward signal defined as the sum rate between the UAV and the users. Dai et al. [100] used deep reinforcement learning to solve dynamic resource allocation problem caused by the limited buffer of the GEO satellite and the time varying parameter channel in the NTN scenario to enhance long-term average throughput performance. In [101] a deep deterministic policy gradient (DDPG)-based algorithm was used to optimize the overall uplink throughput and energy consumption where the state constituted an HAP equipped with MEC server & multiple UAVs. Cui et al. [102] also used deep deterministic policy gradient (DDPG) algorithm for UAV trajectory design and power allocation to maximize the downlink throughput & service time considering UAVs as aerial base stations.

The remaining battery of the UAV was considered in the agent's state by Guo [103], in addition to several QoS and QoE measures. Based on its state the UAV can choose to continue serving in one area, move to serve in another area, or move to recharge its battery at a charging station. The reward included a penalty that relies on battery capacity. In a different setting, Cui et al. [104] defined an energy-efficiency constrained reward function for Q-learning based multi-agent UAV resource allocation. To ensure the agent will learn to optimize its trajectory while optimizing for throughput maximization and energy efficiency, the authors defined the reward as the difference between achieved throughput and the power consumed. Hence, the agent would be rewarded when this difference is increased, that is when

throughput increases and energy consumption decreases. Authors in [105] and [106] also worked on energy efficiency optimization. Zhan et al. [105] modeled a joint design problem of mission completion time, UAV trajectory, as well as communication BS associations and solved it using multi-step DDQN RL algorithm to minimize the energy consumption of the UAV. In [106] a deep reinforcement learning based online channel allocation and power control algorithm in a Satellite-IoT uplink scenario was proposed. The transmission channel and the power are determined by the intelligent agent based on contextual information. A reward to balance increased resource efficiency and met QoS requirements was used.

A QoE-driven formulation was proposed by Liu et al. [107], [108] where Q-learning was used to optimize the trajectory of a UAV in three-dimensional space to maximize the Mean Opinion Score (MOS) of users. The convergence of the agent to an optimal policy was ensured by defining a reward function that rewards the agent for increased MOS at each time step and penalizes the agent when the MOS decreases.

Another line of research focuses on leveraging the usage of Intelligent Reflecting Surfaces (IRS) [109], [110] to assist UAVs in their objectives. IRS have been receiving significant research interest and are viewed as a promising energy-efficient technology for 6G communication networks, as they are capable of enhancing the transmission quality between a sender and a receiver by the intelligent configuration of the wireless environment [111]. In this regard, Zhang et al. [112] considered the mitigation of attenuation in millimeter-wave networks by deploying a UAV that carries an IRS. This approach helps to compensate the N-LoS link by several connected LoS links such as a base station to UAV-IRS link and UAV-IRS to ground user link. Hence, the authors proposed a formulation to optimize the UAV location and the reflection parameters of the IRS using a deep Q-Learning approach. The usage of RL in this context showed effectiveness in reaching a higher average data rate compared with a non-learning approach. Another line of work proposes the placement of IRS on the facade of several buildings to enhance the communication quality between ground users and UAVs. For instance, the joint optimization of both the UAV trajectory and the phase shifts of an IRS was considered by Wang et al. [113] to maximize the overall weighted data rate of all users in the network. A DQN approach was used where the state of the UAV consists of its current coordinates and energy level. The UAV can change its flying direction and distance to maximize the weighted data rate and fairness of all users.

### B. INTEGRATED ACCESS AND BACKHAUL

Instead of acting as an independent aerial base station, non terrestrial platforms can be equipped with wireless transceivers for usage as aerial relays. In this setting, wireless backhauling is employed in NTNs to act as nodes for Integrated Access and Backhaul (IAB) operations. IAB has been justified for usage over 5G infrastructure by the 3GPP [114] and is deemed as useful in enhancing capacity, coverage, as well as connectivity. However, additional challenges are imposed on the UAV that needs to guarantee stable backhaul and access links [115]. Cao et al. [116] proposed a UE-driven deep reinforcement learning (DRL) based scheme, in which a centralized agent deployed at the backhaul side of NT-BSs is responsible for training the parameter of a deep Q-network (DQN), and each UE is able to access a proper NT-BS intelligently to enhance the long-term system throughput and avoid frequent handovers among NT-BSs. A local reward related to the transmission rate and handover cost is collected autonomously by the UE. Integrating LEO satellite and UAV relaying in [117]to maximize the end-to-end data rate, satellite association and HAP location were optimized using deep reinforcement learning where correlation between system utility and achievable rate was modeled by a sigmoid function to calculate the reward. The problem considered the scenario of having a single satellite - HAP link that could be extended in future research to consider a multi-link scenario. Moreover, this same problem can be tackled using a distributed deep learning architecture such as actor-critic or multi-agent reinforcement learning (MARL) to minimize complexity arising from additional communication overhead.

Fotouhi et al. [118] proposed an RL method, based on the brute force search, to optimize the heading direction of the UAV given the locations of neighboring macro base stations and ground users. The reward was defined as the average user performance and was estimated through the received signal power of associated users, interference signal power of neighboring UAV IAB nodes, and the backhaul link performance. A dynamic environment was considered by Tafintsev et al. [119] where a UAV can switch to another association node that can provide better performance as it is moving from one location to another. The association nodes could be ground base stations or other UAVs acting as IAB nodes. In [120], the authors considered the problem where low-Earth Orbits provide backhaul connectivity to UAVs. The authors formulated the problem of maximizing user fairness and minimizing of all terrestrial base stations as a multi-armed bandit problem that can be solved using Q-Learning.

### C. DATA INTEGRITY & SECURITY

#### 1) BASE STATION JAMMING RESISTANCE

Among non terrestrial platforms, UAVs have been proposed as a strategy to resist jamming which cellular systems are vulnerable to. Specifically, jamming occurs when replayed signals are sent to the serving base station to block ongoing communications. Smart jammers have made the problem even worse, where the defense policy of the cellular system is learned through machine learning techniques and smart radio devices [121]. Given their LoS channels to the user

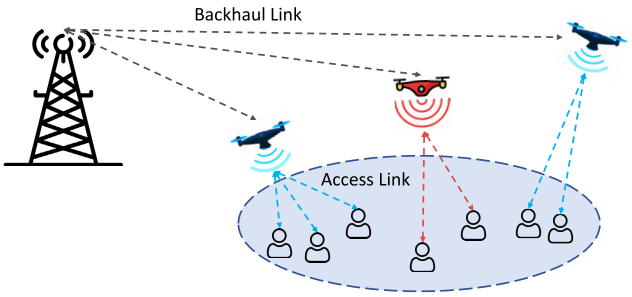



**FIGURE 6.** Illustration of the usage of UAVs as Integrated Access and Backhaul nodes.

equipment, in addition to their high altitude and mobility, UAVs can help mitigate jamming effects by acting as relays when a serving base station is heavily jammed. In this regard, a UAV can be used to relay the traffic of users to a neighboring backup base station. This relay solution is effective since the UAV-to-user and UAV-to-backup base station links will have better channel states than the link between users and the jammed base station. Lu et al. [122] proposed a DQN approach where the UAV is required to find an optimal power relay policy in a way to reduce jamming while maximizing its utility. The learned policy, therefore, allows the UAV to adjust its relay power depending on its current state, which was defined as the bit-error-rate (BER) values of messages received by the jammed base station, and the ground users. Zhou et al. [123] proposed a multi-agent double deep Q-network (MADDQN) to solve channel selection problem and a multi-agent twin delayed deep deterministic policy gradient (MATD3PG) to jointly optimize trajectory design and power control. The study considered an unmanned aerial vehicle (UAV)-assisted downlink transmission and solved the joint optimization problem to maximize the average achievable channel capacity among the ground users. It should be noted that the computational complexity of the algorithm is higher than the general multi-agent deep (MADRL) scheme but this comes at the expense of having dynamic rather than static resource allocation.

### 2) COMBATING GROUND EAVESDROPPERS

Despite the benefits of the LoS-dominated channel links of UAVs, they make it easier for ground eavesdroppers to wiretap the UAV acting as an aerial base station [124], [125], [126]. This fact threatens the security of UAV-aided wireless networks. To solve this issue, UAVs have been proposed as aerial jammers that send artificial noise to the ground eavesdroppers, thus helping the serving UAV. Zhang et al. [127] considered the scenario where the number of UAVs is larger than the number of ground eavesdroppers, requiring the UAV to optimize its flying trajectory in a way to improve the secure rate. To achieve this, a cooperative multi-agent deep deterministic policy gradient (MADDPG) approach was proposed, where the agent could be a serving UAV or a jammer UAV. The state of each UAV was defined as the locations of the other agents, the transmission or jamming

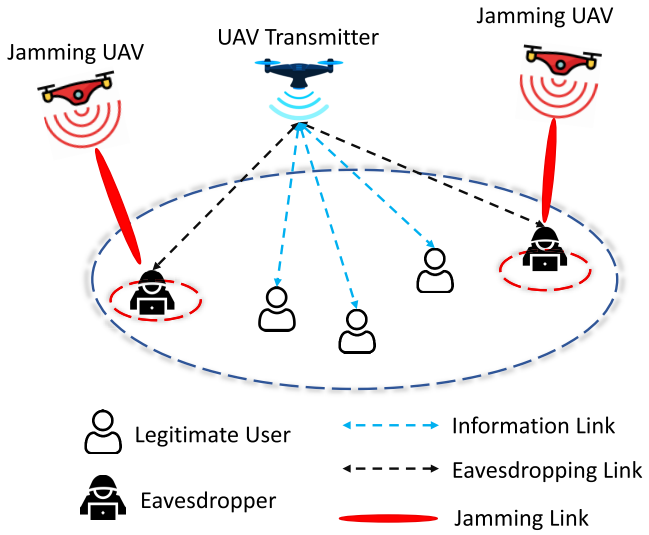



**FIGURE 7.** Illustration of the usage of UAVs for as aerial jammers for combating ground eavesdroppers in the network.

power, and the secure rate of users. Based on this state, each UAV adjusts its location and power level to maximize the reward function, defined as the difference between the secure rate and the jamming power penalty. Further adjustments to this problem formulation were provided by Zhang et al. [128]. The reward function was modified to penalize the UAV when it changes its location beyond the specified map. The agent is also rewarded when it minimizes its distance with the ground users or ground eavesdroppers, depending on whether the agent is a serving or jamming UAV respectively. The authors also reduced the exploration space by the introduction of an attention layer [129], [130] in the neural network architecture of the MADDPG algorithm. Hence, the UAV agent learns to pay attention to the location of ground users and eavesdroppers, resulting in improved learning efficiency.

The same problem was proposed in another setting where information security of UAV-to-vehicle (U2V) communications was considered. Authors in [131] proposed a U2V communications subject to multi-eavesdroppers on the ground in urban scenarios. The study aimed to maximize the secrecy rates in physical layer security perspective while considering both the energy consumption and flight zone limitation, by jointly optimizing the UAV's trajectory, the transmission power of the UAV, and the jamming power sent by the roadside unit (RSU). After modeling the problem as an MDP problem, a curiosity-driven deep reinforcement learning (DRL) algorithm was implemented to solve the problem in which the agent is reinforced by an extrinsic reward supplied by the environment and an intrinsic reward defined as the prediction error of the consequence after executing its actions. However, this study imposes limitations on the number of UAVs & vehicles in the system. Future work may consider multiple UAVs and vehicles deployed.

### D. AGE OF INFORMATION IN NTN-AIDED INFORMATION DISSEMINATION AND DATA COLLECTION

While many works focus on maximizing coverage and enhancing various QoS measures, it is important to ensure the freshness of information received when dealing with time-sensitive applications. This is specifically needed when UAVs are deployed to collect information from IoT devices and sensors in the wireless network. Recently, the AoI was introduced as a time-related metric that measures the time elapsed since the generation of the last received update packet by the destination node from a transmission source [132]. In real-time sensing applications, UAVs can be employed as access points to collect and relay information from ground nodes in IoT networks or wireless sensor networks. However, due to their limited communication range, UAVs will have to fly closer to their targets for better data collection. This could result in lower throughput as the UAV moves farther from the terrestrial base station to which it relays information. In such settings, UAVs are therefore required to optimize their flight trajectory in a way to minimize the AoI [133], [134].

Abd-Elmagid et al. [135] proposed a deep RL approach to minimize the weighted-sum AoI of update packets collected from ground nodes while jointly optimizing the scheduling of packet transmissions. A DQN with Experience Replay (ER) was used where the state of the UAV was represented by its location during a time slot, in addition to the difference between the time left before its battery runs out and the time needed to reach the recharging location. Accordingly, the UAV can choose to move to an adjacent cell in the next time slot or remain in its current position. This work was extended in [136] where a neural combinatorial-based deep RL algorithm was proposed using a DQN. To handle a very large number of nodes, a Long Short-Term Memory (LSTM) auto-encoder was used to reduce the dimensions of the state space to a fixed-length vector. The reward was defined as the reduction in the normalized weighted sum AoI. A similar study was presented in [137] where UAVs were deployed as virtual queues between base stations and low-resource IoT devices to relay recent information. Aiming to minimize the expected weighted sum AoI, a proximal policy optimization approach was used to control the UAV's altitude and scheduling behavior. IRS were also made use of in the context of AoI minimization by Samir et al. [138] where the phase shifts of the IRS were optimized along with the altitude of the UAV.

Yi et al. [139] tackled the AoI minimization problem with UAV energy constraints. The state was represented by the UAV's location, the AoI value for each sensor node in the network, the difference between the UAV's remaining time and energy, and the time, and energy needed to reach its final destination. The UAV's actions consist of its movement and scheduling of a sensor node. A custom reward function was defined to reward the UAV when the weighted sum AoI is reduced, and penalize the UAV when several defined energy, location, and scheduling constraints are violated. Another energy-efficient trajectory optimization of a UAV with considerations for data freshness was proposed by *Abedin et al.* [140]. A DQN with ER approach was adopted where the agent is required to minimize the AoI while maximizing its reward that was defined as the instantaneous energy efficiency function.

A multi-UAV approach for cooperative sensing and AoI minimization was introduced by Hu et al. [141], where a distributed sense-and-send protocol was presented. The protocol defines several cycles that the UAV goes through to complete its tasks of sensing and transmission of its results to a base station. A set of UAVs was considered, where each UAV acts as an RL agent. The state of the UAV is represented by the number of considered cycles, the amount of sensing data it will transmit to the base station, its selected task, and its target sensing location. At every state, the UAV takes the actions of selecting a task and a sensing location. The reward was defined as the negative average AoI of all tasks. However, due to the nature of this formulation where the action space contains discrete variables (task selection) and continuous variables (sensing location), a compound-action actor-critic (CA2C) algorithm was proposed to deal with this problem since traditional deep RL methods can either deal with purely discrete or continuous action spaces [142]. This formulation was improved in [143] where the the reward function was altered to become the reduction in AoI when transitioning from one state to another.

To investigate the benefits of integrating unmanned aerial vehicles (UAVs) with reconfigurable intelligent surface (RIS) elements to passively relay information sampled by Internet of Things devices (IoTDs) to the base station (BS), an optimization problem was proposed in [144] with the objective of minimizing the expected sum Age-of-Information (AoI). Proximal policy optimization algorithm was adopted to solve the problem and optimize the altitude of the UAV, the communication schedule, and phases-shift of RIS elements. Simulation results showed that the proposed algorithm outperforms all others in terms of AoI. It is observed that if the number of reflecting elements per RIS increase, the quality of the communication link between the IoTD and the BS will be enhanced thus improving SNR and expected sum of AoI. A variant of this work maybe to consider multiple antennas in source/destination nodes in the future and study overall system performance.

## V. RL FOR CELLULAR-CONNECTED NTNs

Multiple works in the literature have leveraged RL techniques to aid cellular-connected non terrestrial platforms specifically UAVs in optimizing their trajectory for various objectives. In what follows, we present the applications of RL in cellular-connected NTNs with a focus on the proposed MDP formulations. A summary of these formulations is provided in Table 3 where we highlight which works addressed finite or infinite state and action spaces.

**TABLE 2. A Summary of research that uses RL algorithms for various objectives in NTN-aided wireless communications. Blue background coloring reflects finite state or action spaces. Red background coloring reflects infinite state or action spaces.**

| Ref. | RL Algorithm | State | Action | Reward |
|---|---|---|---|---|
| **Enhanced Coverage** | | | | |
| [93] | Double Q-Learning | • User receiving service<br>• User waiting time<br>• UAV-to-user flying time | • Provide service to user | • Total users served |
| [92] | DQN | • Received signal strength | • Change location | • SINR of UAV |
| [94] | DDQN with PER | • Coverage bitmap | • Change location | • Negative error function |
| [96] | DDPG | • Coverage score of each cell<br>• Energy consumption of UAV<br>• Position & direction of UAVs<br>• UAV-to-UAV distances | • Change flying direction<br>• Change flying distance | • Energy efficiency function<br>• Penalize UAV when it gets disconnected from the network |
| [91] | Classical Q-learning | • Coordinates of HAP | • Change HAP position | • Maximize user coverage |
| [90] | k-step SR QD Learning | • Location of NT-BS<br>• Location of backhaul node on ground | • Determine location to move at next time step | • Maximize mean of accumulated rewards to maximize user coverage |
| [97] | Q-Learning (Model free) | • UAV location<br>• UAV embedding features<br>• UAV energy consumption | • Change UAV direction and moving distance<br>• User link selection | • Maximize mean of accumulated rewards to maximize user coverage |
| **Enhanced QoS/QoE** | | | | |
| [99] | Q-Learning | • UAV location and time | • Change moving direction | • UAV-users sum rate |
| [107] | Q-Learning | • UAV position & altitude | • Change direction or altitude | • Instantaneous MOS of users |
| [98] | DPG | • Temporal difference between received signal strength at each timeslot | • Change movement (in terms of spherical coordinates) | • Uplink sum rate |
| [113] | DQN | • UAV coordinates<br>• UAV energy level | • Change flying direction<br>• Change flying distance | • Weighted data rate and fairness |
| [103] | Dueling DDQN with PER | • Remaining battery of UAV<br>• Expected battery left<br>• QoS & QoE information | • Continue serving in area<br>• Move to another area<br>• Move to recharging location | • Difference between met and unmet load with a battery capacity penalty |
| [104] | Q-Learning | • SINR of UAV | • Selected user & subchannel<br>• Selected power level | • Throughput & power consumption difference |
| [105] | Multi-step DDQN | • UAV waypoint | • Selecting UAV flying direction | • Energy consumption of UAV along the n-th line segment<br>• SINR measurement at waypoint |
| [145] | Double deep Q-learning | • Local channel condition<br>• Transmission data rate<br>• Total cost of computation and communication overhead<br>• Number of data bits transmitted<br>• Computational complexity ofthe task | • Offloading decision<br>• Sub-band allocation<br>• Transmission power of vehicle<br>• Computing resources allocation to vehicle | • Minimization function of the total communication and computation overhead incurred by each vehicle |
| [146] | Deep Q-learning | • Starting and ending time of user request<br>• Duration and priority of request | • Select next request according to probability distribution for deciding the next action | • Selection and scheduling of user tasks where failure rate and timeliness of requests are optimized |
| [100] | Deep Q-learning | • Real-time CSI of the up-link and down-link of the GEO satellite<br>• Data information stored in satellite buffer | • Change the power used by the receive gain<br>• Change the transmitted power allocated by the satellite for each beam | • Actual total power of down-link<br>•Long term average throughput of the system |
| [102] | DDPG | • Position of UAV<br>• Remaining energy | • Choose most appropriate next position and power allocation during the flight of the UAV | • Average down-link throughput of the random flight<br>•A binary number that represents whether the UAV has reached its destination |
| [101] | DDPG | • Dynamic aerial vehicle | • Offloading rate of AVU<br>• Offloading decision of AVU | • Negative value of total latency and the energy consumption for all AVUs processing tasks |
| [147] | Multi-tier cooperative deep RL | • Group of channel gains in sub-channels<br>• Group of interferences suffered by vehicle-to-satellite links<br>• Group of achievable uplink rates from vehicles | • Allocate sub-channels to vehicles | • Maximize the overall up-link throughput of vehicles |
| [148] | Parallel Adaptive DRL | • UAV-BS flight height, horizontal angle, elevation and speed<br>• Users state (for ten users) | • Adjust UAV-BS speed, flight horizontal angle and elevation angle | • MSE of prediction error |
| [82] | MAT3DPG | • Coordinates of all UAVs in served area | • Adjust UAV flying speed and angle | • Weighted function of average channel capacity, transmission power and collision avoidance |
| [106] | GNN-QL | • Channel quality information<br>• Set of terminals with allocated resources | • Allocate resources to terminal | • Maximize terminal's long term energy efficiency<br>• Penalty of unsatisfied QoS requirement on transmission rate, outage probability, and latency |

**TABLE 2.** *(Continued.)* **A Summary of research that uses RL algorithms for various objectives in NTN-aided wireless communications. Blue background coloring reflects finite state or action spaces. Red background coloring reflects infinite state or action spaces.**

| | | | | |
|---|---|---|---|---|
| **Integrated Access & Backhaul** | | | | |
| [118] | Brute force search | • Location of UAV<br>• Location of other UAVs | • Change moving direction | • Average user performance |
| [140] | DQN with ER | • Position and height of UAV<br>• Target position of UAV<br>• Energy efficiency<br>• Age of UAV navigation | • Change position | • Energy efficiency function |
| [149] | Q-Learning | • HAP moving direction<br>• HAP speed | • Re-position HAP | • Improve communication performance via HAP when direct link is blocked |
| [116] | Deep Q-Learning | • Connected NT-BS of user equipment (UE) at time t<br>• Received signal strength at specific UE from each NT-BS<br>• Number of connected UE of each NT-BS at specific time slot<br>• Transmission rate of specific UE at specific time slot | • User taking action to access NT-BS in specific time slot | • Sum-rate of all user equipment connecting the same NT-BS |
| [150] | Multi-agent soft actor critic | • Current location of UAV<br>• Current energy level of UAV<br>• Harvested energy estimate | • Task partitioning<br>• Power allocation | • Cost of task execution delay and energy consumption |
| [151] | 2D Multi-armed Bandit problem | • UAV state constituting UAV movement direction and channel | • Allocate Channel<br>• Change direction of UAV | • Reward Function for each BS capturing the fairness and the load of the BS |
| [152] | MCMOPSO-RL | • UAV position | • Adjust UAV position | • Length of UAV path |
| [153] | Multi-armed Bandit | • Power level and channel for BSs<br>• Trajectory of UAVs | • BS action is composed of its transmission power and channel<br>• UAV action is composed of its transmission power, channel, and movement direction | • Sum of all maximum achievable data rates provided by BSs |
| [80] | Multi-armed Bandit | • UAV position, speed and movement direction | • Change UAV speed | • Captures two performance metrics, including throughput and energy consumption blocked |
| [154] | Federated PPO RL | • Relative locations of neighboring UAVs<br>• Traffic Information | • Move UAV in one of four directions or keep in position | • Minimize sum of remaining traffic |
| [155] | Deep Q-Network | • Communication link<br>• Transmission power<br>• Antenna tilt<br>• Number of transmission antennas in link | • Transmit or Re-transmit | • Energy efficiency<br>• Spectral efficiency<br>• No fading |
| [117] | DRL | • Position and velocity of HAP<br>• Position of SAT in orbital plane<br>• Link distance for each link<br>• Achievable rates for each link from source-to-destination | • Selecting the serving satellite in specific orbital plane | • Sigmoid function to describe correlation between system utility and achievable rate |
| **Base Station Jamming Resistance** | | | | |
| [122] | DQN | • BER of received message from jammed base station | • Adjust relay power | • UAV utility |
| [156] | Q-Learning | • Channel status of current device | • Allocate free channel with reflecting service level information | • Function based on channel load and blocking condition |
| [123] | MA-LB-DRL | • Coordinates of UAVs<br>• Coordinates of ground users<br>• Current channel occupancy | • Select channel<br>• Adjust flying speed and angle<br>• Adjust transmission power | • Mean achievable channel capacity |
| **Combating Ground Eavesdroppers** | | | | |
| [79] | MA-DRL | • Speed and position of the ground node<br>• Power state controlled by UAV<br>• Nodes scheduling state | • Control transmission power | • Maximize sum secracy rate<br>• Set a penalty based on speed and distance of UAVs |
| [131] | C-DQN | • UAV's coordinates and speed<br>• Vehicle's coordinates, velocity and acceleration<br>• RSU's coordinates<br>• Eves-dropper's coordinates | • Adjust UAV's speed<br>• Adjust transmission power<br>• Adjust jamming power | • Minimizing loss function |
| **Age of Information in NTN-aided Information Dissemination and Data Collection** | | | | |
| [136] | LSTM DQN | • Energy levels of nodes | • Schedule node for sending an update packet | • Reduction in the normalized weighted sum AoI |
| [139] | DQN | • UAV location<br>• AoI value for each sensor node<br>• Time and energy difference to reach destination | • Change movement<br>• Schedule a sensor node | • Rewards reduction in weighted sum AoI<br>• Penalizes violation of energy, location, and scheduling constraints |
| [137] | Proximal Policy Optimization | • AoI of all IoT devices<br>• Status of the virtual queue<br>• Time elapsed of the queue<br>• Achievable rate<br>• Status-update size | • Adjust altitude<br>• Decide which IoT device to transmit a status update<br>• Schedule transmission from virtual queue to base station | • Penalize the agent when collecting status updates from device with high AoI<br>• Penalize the agent when it flies beyond the specific altitude constratints |

**TABLE 2.** *(Continued.)* **A Summary of research that uses RL algorithms for various objectives in NTN-aided wireless communications. Blue background coloring reflects finite state or action spaces. Red background coloring reflects infinite state or action spaces.**

| | | | | |
|---|---|---|---|---|
| [143] | CA2C | • Number of considered cycles<br>• Amount of sensing data<br>• Selected task by UAV<br>• Sensing location | • Select task (sensing or transmission to base station)<br>• Select sensing location | • Reduction in AoI |
| [157] | Deep Q-learning | • Back-scatter device channel information | • Move UAV one step of nine directions horizontally | • Energy Efficiency Reward of the UAV-aided BC data collection system |
| [158] | Q-learning | • Decoding Window Width where repair packets are collected<br>• Expected decoding delay given the current in-flight packets | • Update the current system state<br>•Update the FEC code rate corresponding to the in-flight packets | • Design reward signal to maximize the expected in-order good-put while controlling the delay |
| [144] | Proximal Policy Optimization | • SNR at the BS<br>• Altitude of UAV | • Adjust deployed UAV altitude<br>•Schedule an IoTD depending on the phases of RIS elements and the activation pattern | • Negative summation of AoI |

## A. ENHANCED CONNECTIVITY

### 1) COVERAGE HOLE AVOIDANCE

An important challenge for cellular-connected UAVs is guaranteeing connectivity to the cellular network as they hover to a specific destination [159]. This challenge is imposed by the fact that currently, the existing terrestrial base stations are designed to serve terrestrial user equipment. Thus, the antennas of these base stations are typically down-tilted [160]. A ubiquitous coverage in the sky for UAVs is therefore not available by current cellular networks such as Long-Term Evolution (LTE) networks [161]. This challenge can be addressed by leveraging the UAV's controllable mobility feature to design a communication-aware trajectory that can enhance connectivity to the cellular network. Zeng et al. [162] proposed a model-free RL approach, based on Temporal Difference (TD) learning, to avoid coverage holes by minimizing the UAV's disconnection duration from the network. The state was represented by the location of the UAV. At every state, the UAV can choose to change its flying direction. The UAV is rewarded if it is in a location that is connected to the cellular network and is penalized otherwise. This problem was extended to a deep RL setting in [163], where the dueling DDQN was used. To enable the UAV to learn how to avoid being disconnected from the network, the authors modified their reward function to penalize the UAV when it is in a location with a certain outage probability. In the context of the internet of connected vehicles a cooperative approach for content caching and delivery is presented in [164]. A RSU with a limited communication coverage collaborates with a UAV to deliver contents to vehicles on a road segment. An MDP problem is modeled with the goal of maximizing the number of served vehicles and solved using a dual task reinforcement learning method. The problem was modeled as a single-cell scenario in which one RIS-aided air-to-ground uplink is deployed. A more realistic and interesting problem might be the case of having a multi-cell scenario, where the RISs provide both signal enhancement and inter-cell interference mitigation.

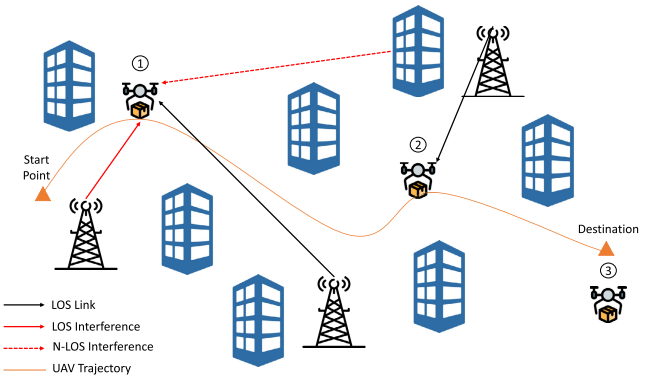



**FIGURE 8.** **Illustration of the trajectory design of cellular-connected UAVs for enhanced connectivity. LOS stands for Line-of-Sight. NLOS stands for No-Line-of-Sight.**

### 2) HANDOVER RATE REDUCTION

Another line of research focuses on reducing the potential number of handovers which can lead to radio link failure and signaling overhead [165]. By adopting an efficient handover mechanism, the robustness of the connection between the aerial platform and the cellular network can be improved. A Q-learning approach was presented by Chen et al. [166] to design the UAV's trajectory in a way that optimizes the number of handovers. In baseline handover schemes the UAV connects to the cell that provides the strongest received signal strength. In this formulation, this is not always the case since the UAV may connect to a cell with lower received signal strength but would go through fewer handovers while maintaining reliable connectivity. The state of the UAV was represented by its position, movement direction, and the cell it is connected to. At every state, the UAV can take the action of choosing what next cell to connect to. The reward function was defined as a weighted combination of the received signal power of the cell at the next state and the handover rate. This work was extended in [167] to a deep RL setting based on DQN that can handle real-world scenarios where the state space becomes too large, making it more appealing to approximate Q-values

rather than relying on tabular Q-learning. Azari et al. [168] formulated the handover reduction problem as a multi-armed bandit problem, where the agent changes its movement speed to reduce the disconnectivity time given additional energy and link reliability constraints.

### B. SPECTRAL MANAGEMENT

The rapidly increasing number of communication devices that a network needs to handle has made the communication environment highly complex. This problem is augmented when limited spectral resources are available. Additional burdens are imposed on this environment when cooperative UAVs are deployed as aerial users in these networks [169]. Under limited available channels to serve these UAVs, a robust dynamic channel allocation strategy is required to maximize spectral efficiency [170]. Given the time-varying and complex environment that UAVs need to operate within, RL methods have been found useful in achieving an optimal action strategy for spectral management. Zhou et al. [171] proposed a DQN approach that incorporates an LSTM neural network for dynamic channel allocation. In this approach, several UAVs are deployed for various tasks and need to send information to receiving nodes, but the number of channels available is smaller than the total number of UAVs. Each UAV was represented as an agent. The state was defined as the channel occupancy status, residual channel capacity, and collision of UAV access. The authors defined a reward function that penalizes the UAV when a collision occurs, that is when it tries to access a channel that is already occupied by another UAV. Otherwise, the UAV receives a reward that depends on its distance from the receiving node. This work was extended in [172] to consider information sharing among UAVs, where one UAV would broadcast information to the rest of the UAVs in the network, allowing the better accomplishment and survivability of the tasks. In this setting, a strategy for dynamic allocation of time slots is required since only one UAV needs to be in the transmission state while the rest of the UAVs need to be in the information reception state. The agent can decide at every state whether to share information with the rest of the UAVs or not, depending on the reward it receives which was adjusted to be the MOS, which was defined to consider the sending bit rate, frame rate, and total packet error rate. In the afore-mentioned studies, authors were simulating the channel using dominant/probabilistic empirical models since channel state information (CSI) is unavailable due to UAVs mobility. More realistic CSI estimations to get more accurate channel models are performed through learning-based approaches. Luong et al. [173] proposed a novel algorithm that employs a deep Q-learning approach to tackle the issue of CSI unavailability for determining UAVs' positions in a multi-cooperative UAV network. Numerical results demonstrated that the approach was efficient with a network performance gain of up to 70%. In [174] the authors presented a machine learning based channel estimation technique to help reduce the CSI feedback delay as the UAV feeds the CSI information only to the primary base stations. Simulation results showed that both the bit error rate (BER) and the sum rate performance are enhanced when appropriate CSI estimation results are utilized.

### C. INTERFERENCE MANAGEMENT

Despite the benefits UAVs get from being connected to the cellular network such as high-speed data access, this comes at the cost of increased inter-cell interference to ground users and among the UAVs. It is therefore important to optimize the trajectory of the UAV to overcome the interference challenge in cellular networks that serves users in the ground and the air [175], [176]. Hence, the UAV should be able to adapt its movements depending on the requirements of the ground and aerial user equipment. A non-cooperative game-theoretic formulation for interference management was proposed by Challita et al. in [177], [178] and was solved using a deep RL algorithm based on echo state networks. The approach aims at mitigating the interference caused by the UAV on the ground users while minimizing the time required to reach the destination location as well as the transmission delay. It was shown that a vital role is played by the UAV's altitude when aiming to minimize interference levels on ground users. The challenge of UAV height optimization was tackled in [179] using a DQN with ER, where the UAV agent adapts its height in a way to increase throughput under interference constraints. A similar study with energy constraints was presented in [180].

## VI. QUALITATIVE ANALYSIS: SIMULATION REALISM

To investigate how well the surveyed works of the literature emulated a realistic simulation environment, we provide a comparative illustration in Tables 4 and 5 that classify the literature according to several factors we define as important to achieving realism in simulation. In this regard, we consider four main factors: the simulation environment, the nature of the aerial platform mainly UAV in Table 4, the wireless channel, and the energy of the UAV. Additionally in Table 5 we consider non terrestrial platforms in general specifying the platform type. Under the simulation environment, we classify the works on whether their simulated environment was static or dynamic, and whether it was 2-Dimensional (2D) or 3-Dimensional (3D). The nature of the NT-BS proposed in the problem formulation is classified as single, multiple independent, or multiple platforms that coordinate cooperatively to achieve a certain goal. In terms of the wireless channel considered in the proposed system model, we classify the works according to four levels: a simple path loss model that considered the presence of a dominant LoS link, a path loss model with shadowing and/or fading consideration, a probabilistic path loss model that considers probabilities of having LoS or N-LoS links, or the case of where the UAV performs estimation of the channel state information (CSI). Finally, we also classify the works on

**TABLE 3. A Summary of research that uses RL algorithms for various objectives in cellular-connected UAVs. Blue background coloring reflects finite state or action spaces. Red background coloring reflects infinite state or action spaces.**

| Ref. | RL Algorithm | State | Action | Reward |
|---|---|---|---|---|
| **Enhanced Connectivity** | | | | |
| [162] | TD Learning | • UAV location | • Change flying direction | • Negative reward when the UAV is connected to the cellular network<br>• Penalized when disconnected from the network |
| [166] | Q-Learning | • UAV position<br>• Movement direction<br>• Cell connected to | • Choose cell to connect to | • Weighted combination of the handover rate and received signal power at the next state |
| [167] | DQN | • UAV position<br>• Movement direction<br>• Cell connected to | • Choose cell to connect to | • Weighted combination of the handover rate and received signal power at the next state |
| [163] | Dueling DDQN | • UAV location | • Change flying direction | • Penalized when entering a zone with a certain outage probability |
| [181] | Dual-Task DRL | • Position of UAV<br>• Position of vehicles at time t<br>• Downloaded content | • Sub-action 1 is for RSU and it specifies whether to continue or cut the service<br>• Sub-action 2 is for UAV and it selects one velocity out of a list of different velocities | • Sum of positive rewards due to serving vehicles sufficiently |
| [182] | DDPG and REL-DDPG | • Environment information obtained by UAV | • Generate required load factor and UAV acceleration | • Distances between UAV and target at two consecutive time instants<br>• Yaw and pitch angles between UAV and target |
| [164] | Distributionally Robust Soft Actor-Critic | • Location of UAV<br>• Distance from UAV to center of obstacles<br>• Sum rate of UAV and GU | • UAV's maneuver direction<br>• Phase shift of each RIS sub-surface<br>• Power control for UAV and ground user (GU) | • Instantaneous sum rate of UAV and GU<br>• Penalty if any constraint is not satisfied |
| **Spectral Management** | | | | |
| [171] | LSTM DQN | • Channel occupancy status<br>• Collision of UAV access<br>• Residual channel capacity | • Decide to access a channel or not | • Rewarded when distance to destination node is minimized<br>• Penalized when accessing the same channel of another UAV |
| [172] | LSTM DQN | • Throughput<br>• Collision of UAV access<br>• Residual channel capacity | • Decide to share information with other UAVs or not | • MOS |
| **Interference Management** | | | | |
| [179] | DQN with ER | • Density of buildings<br>• Density of base station<br>• SINR<br>• Distance to closest base station<br>• Height of UAV | • Change height (up, down, or no change) | • Spectral Efficiency |
| [180] | DQN | • Distance to other UAVs<br>• Path-loss to nearby stations<br>• Interference Measurements<br>• Current serving base station<br>• Radio Resource Bocks available<br>• UAV Speed | • Adjust transmit power<br>• Adjust altitude<br>• Select base station | • Custom reward function that include energy and spectral efficiency, as well as interference incurred |

whether they considered the UAV's limited energy resources in their proposed RL formulation.

Upon analyzing Table 4, we can conclude that:

- While a noticeable number of works considered a 3D environment, much less consideration for dynamic and 3D environments was reported, with most of the literature presenting static simulation environments.
- The works in [82], [96], [123], [127], [128], [140], [141], [142], [143] and [172] succeeded in achieving realistic aerial platform deployment scenarios where multiple platforms are expected to perform cooperative decisions instead of independent decisions.
- Most of the works simulated a realistic wireless channel in their system model using the probabilistic path loss model [93], [94], [118], [127], [128], [104], [107], [108], [172], [179], [180], [148], [80], and [123] with only a few works achieving higher realism by considering CSI estimation [79], [112], [131], [137], [158].
- In terms of energy considerations, a fair number of works presented energy-efficient factors and constraints in their formulations such as battery capacity [103], energy harvesting [112], [150], [157], propulsion energy [113], [139], energy quanta [136], [139] and others [95], [96], [140].

Upon analyzing Table 5, we notice that more attention was paid to 3D environments with more realistic deployment scenarios where multiple non terrestrial platforms coordinate together to provide multi-user access control [116] in NTNs, space-air-ground integrated link optimization [151], [153], maximizing end-to-end data rate [117] and others [154], [155].

**TABLE 4.** Classification of the surveyed literature in terms of realism factors pertaining to the simulation environment, UAV, wireless channel, and energy consideration.

| Ref. | Environment | | | | UAV | | | Wireless Channel | | | | UAV Energy Consideration |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Static | Dynamic | 2D | 3D | Single-UAV | Multi-UAV | Multi-UAV Cooperative | Pathloss Model with dominant LoS | Pathloss Model with shadowing and/or fading | Probabilistic Pathloss Model | CSI Estimation | |
| [92] | ✓ | | ✓ | | | ✓ | | | ✓ | | | |
| [93] | ✓ | | | ✓ | ✓ | | | | | ✓ | | |
| [94] | ✓ | | | ✓ | | ✓ | | | | ✓ | | |
| [95] | ✓ | | ✓ | | | ✓ | | ✓ | | | | ✓ |
| [96] | ✓ | | ✓ | | | | ✓ | ✓ | | | | ✓ |
| [98] | ✓ | | | ✓ | ✓ | | | ✓ | | | | |
| [99] | ✓ | | ✓ | | | | | | ✓ | | | |
| [103] | ✓ | | ✓ | | ✓ | | | ✓ | | | | ✓ |
| [104] | | ✓ | ✓ | | | ✓ | | | | ✓ | | ✓ |
| [107] | | ✓ | | ✓ | | ✓ | | | | ✓ | | |
| [108] | | ✓ | | ✓ | | ✓ | | | | ✓ | | |
| [112] | ✓ | | ✓ | | ✓ | | | | | | ✓ | ✓ |
| [113] | | ✓ | ✓ | | ✓ | | | ✓ | | | | ✓ |
| [118] | ✓ | | | ✓ | | ✓ | | | | ✓ | | |
| [119] | | ✓ | | ✓ | | ✓ | | ✓ | | | | |
| [122] | | ✓ | ✓ | | ✓ | | | ✓ | | | | ✓ |
| [127] | ✓ | | | ✓ | | | ✓ | | | ✓ | | |
| [128] | ✓ | | | ✓ | | | ✓ | | | ✓ | | |
| [135] | ✓ | | ✓ | | ✓ | | | ✓ | | | | ✓ |
| [136] | ✓ | | ✓ | | ✓ | | | ✓ | | | | ✓ |
| [137] | | ✓ | | ✓ | ✓ | | | | | | ✓ | |
| [138] | ✓ | | ✓ | | ✓ | | | | ✓ | | | |
| [139] | ✓ | | ✓ | | ✓ | | | ✓ | | | | ✓ |
| [140] | | ✓ | ✓ | | | | ✓ | | | ✓ | | ✓ |
| [141] | ✓ | | ✓ | | | | ✓ | | | ✓ | | |
| [142] | ✓ | | | ✓ | | | ✓ | | | ✓ | | |
| [143] | ✓ | | | ✓ | | | ✓ | | | ✓ | | |
| [163] | ✓ | | | ✓ | ✓ | | | | ✓ | | | |
| [166] | ✓ | | ✓ | | ✓ | | | ✓ | | | | |
| [167] | ✓ | | ✓ | | ✓ | | | ✓ | | | | |
| [168] | ✓ | | ✓ | | ✓ | | | | | ✓ | | |
| [172] | | ✓ | ✓ | | | | ✓ | ✓ | | | | |
| [178] | | ✓ | | ✓ | ✓ | | | ✓ | | | | |
| [179] | | ✓ | | ✓ | ✓ | | | | | ✓ | | |
| [180] | | ✓ | ✓ | | | ✓ | | | | ✓ | | |
| [158] | | ✓ | | ✓ | ✓ | | | | | | ✓ | ✓ |
| [102] | | ✓ | ✓ | | | ✓ | | ✓ | | | | ✓ |
| [148] | | ✓ | | ✓ | ✓ | | | | | ✓ | | ✓ |
| [82] | ✓ | | | ✓ | | | ✓ | | ✓ | | | ✓ |
| [80] | ✓ | | | ✓ | | ✓ | | | | ✓ | | ✓ |
| [123] | ✓ | | | ✓ | | | ✓ | | | ✓ | | ✓ |
| [79] | | ✓ | | ✓ | ✓ | | | ✓ | | | | |
| [131] | | ✓ | | ✓ | ✓ | | | | | | ✓ | ✓ |
| [157] | ✓ | | ✓ | | ✓ | | | | ✓ | | | ✓ |
| [144] | | ✓ | | ✓ | ✓ | | | | ✓ | | | |
| [181] | | ✓ | ✓ | | ✓ | | | | | | ✓ | |
| [182] | | ✓ | | ✓ | ✓ | | | ✓ | | | | |
| [164] | ✓ | | | ✓ | ✓ | | | | ✓ | | | |

**TABLE 5. Classification of the surveyed literature in terms of realism factors pertaining to platform, simulation environment, wireless channel, and energy consideration.**

| Ref. | Platform | Environment | | | | Platform | | | Wireless Channel | | | | Energy Consideration |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Static | Dynamic | 2D | 3D | Single | Multiple | Multi-Cooperative | Pathloss Model with dominant LoS | Pathloss Model with shadowing and/or fading | Probabilistic Pathloss Model | CSI Estimation | |
| [90] | HAPs & LAPs | | ✓ | | ✓ | | | ✓ | | | | ✓ | ✓ |
| [97] | LEOs & UAVs | | ✓ | | ✓ | | | ✓ | | | | ✓ | ✓ |
| [145] | LEOs & HAPs | | ✓ | | ✓ | | ✓ | | | ✓ | | | ✓ |
| [146] | Satellite | | ✓ | | ✓ | ✓ | | | ✓ | | | | ✓ |
| [100] | HTS | | ✓ | | ✓ | ✓ | | | | | | ✓ | ✓ |
| [101] | HAP & UAVs | | ✓ | ✓ | | | ✓ | | ✓ | | | | ✓ |
| [147] | LEOs | ✓ | | ✓ | | | | ✓ | | ✓ | | | |
| [106] | LEOs | ✓ | | ✓ | | | ✓ | | | | | ✓ | |
| [149] | HAP | ✓ | | | ✓ | ✓ | | | ✓ | | | | ✓ |
| [116] | NT-BS | ✓ | | | ✓ | | | ✓ | | ✓ | | | |
| [150] | HAP & UAVs | ✓ | | | ✓ | | | ✓ | | | | | ✓ |
| [151, 153] | LEOs & UAVs | | ✓ | | ✓ | | | ✓ | | | ✓ | | |
| [154] | LEOs & UAVs | | ✓ | | ✓ | | | ✓ | ✓ | | | | |
| [155] | LEOs & HAPs | | ✓ | | ✓ | | | ✓ | | ✓ | | | ✓ |
| [117] | LEO & UAV | ✓ | | | ✓ | ✓ | | | ✓ | | | | |
| [156] | HAPs | ✓ | | | ✓ | | | ✓ | ✓ | | | | ✓ |

## VII. BROAD RESEARCH DIRECTIONS

In this section, we discuss some challenges that arise when adopting RL techniques for NTN communications. Our set of challenges highlight open research that integrates NTN communications and intelligence, and includes some key ideas that should be considered to bridge the gap between simulation-based experimentation and real-field implementation.

### A. EXPERIMENTATION AND ADAPTATION TO REAL ENVIRONMENTS

RL-based solutions proposed for both NTN-aided wireless communications and cellular-connected NTNs have been experimented on in simulation environments. Although simulation-based environments enable the collection of larger data sets for training, it will be difficult for a model trained on data generated by simulated environments to generalize in real-world environments. Dynamic environments need to be further explored in problem formulations to accurately mimic real-world situations that include various uncertainty in terms of user behavior, demand, or mobility. Statistical efficiency is needed in the real world since we can not obtain as many samples as we can during simulations. In this case, a possible solution could be the investigation of domain adaptation techniques for RL [183], [184], [185], [186] since they can allow models trained on data from one domain to generalize in a target domain, which is the real-world environment. Additionally, to validate the usefulness of RL methods for intelligent NTN communications, it is necessary to perform experiments of these approaches in the real-world using wireless testbeds [187], [188]. Such procedures are important as they may uncover challenges that a non terrestrial platform will face in a real deployment, and that are not easily deducible from experiments in simulated environments. By performing experimentation in the real-world and adapting models from simulated to real environments, the simulation-reality gap can be mitigated. One sample consideration is that non terrestrial platforms especially UAVs have to move very close to users mainly in extremely harsh environments to achieve better performance. In order to adapt to such harsh environments, the hardware material used to manufacture the platform itself should be robust to tolerate real situations. Harsh atmospheric conditions, sensor accuracy, equipment size and battery endurance affect the flight time and in turn the performance. This should be taken into consideration so that UAVs will be able to provide an adaptable and reliable communication backbone [189], [190].

Integrating NTN and free space optical (FSO) technologies can provide low cost broadband solutions in extremely harsh environments, and can be the next disruptive technology for 6G remote connectivity. Hybrid RF/FSO Satellite Communication is proposed in [191] where the satellite selects RF or FSO links depending on the weather conditions obtained from sensors knowing that the impact of rain on FSO transmission is less significant compared to fog. In hybrid RF/FSO two configurations are possible. The first one enables RF communication at one hop and FSO communication at the other in a dual-hop or relay-assisted networks. For regions that have high probability of a certain weather condition (mainly cloud, rain, fog), frequencies with tolerable attenuation should be preferred in order to

complement the behaviour of FSO main link by a RF back up link [192]. The hybrid radio frequency/free-space optical (RF/FSO) network can be employed in backhaul-to-relay and relay-to-user communications when considering the limited backhaul communication in HAPs [193]. It resulted in improved power & spectral efficiency in [191] and [193], respectively. Joint optimization problems can be formed to help link aerial and terrestrial terminals by optimizing multiple-HAP deployment, power and spectral efficiency.

### B. METAVERSE REALITY

With the advancement of wireless communication technologies and the creation of a digital twin of the physical world, known as the meta-verse or 3D virtual reality, new open research problems arise. Networks are expected to support super-high-definition (SHD) and extremely high-definition (EHD) videos, with super-high throughput demands and to provide ultra-reliable low-latency communications. To achieve this, bands in the range of 275GHz–3000GHz, which are known as Terahertz (THz) bands and are not yet allocated for specific active services, will be considered. However, these available bands at terahertz (THz) and millimeter-wave (mmWave) frequencies are limited by a short communication range and a high susceptibility to molecular absorption, blockage, and deep fade. Recent proposed work in this area is presented in [194] and [195]. Non-terrestrial platforms will play a crucial role in offering expected Tbps-level throughput and sub-millisecond latencies to assist terrestrial networks via 6G technology since current terrestrial network capabilities do not satisfy 6G requirements. 6G is supposed to be a cell free four-layer architecture network that combines space, air, terrestrial, and underwater (or sea) network tiers where full wireless coverage and ubiquitous connectivity will be provided in an intelligent information society to support support various applications, such as flight in the sky, voyage at sea, or vehicles on land. Low-Earth-orbit, medium-Earth-orbit, and geostationary-Earth-orbit satellites will be deployed to support orbit or space Internet services to serve areas not covered or partially covered by terrestrial networks. Satellites with mm-wave communications will be deployed for high-capacity satellite–ground transmission.As for long-distance inter-satellite transmission in free space, laser communications may be used. Flying and floating base stations such as UAVs and HAPs can be deployed to work in the low-frequency, microwave, and mm-wave bands to provide more flexible and reliable connectivity for urgent events or remote areas [81]. 6G will be an autonomous ecosystem where intelligence and machine learning will be needed to integrate sensing, communication, computing, caching, control, positioning, radar, navigation, and imaging, to support full-vertical applications. [196] implement deep reinforcement learning to enhance communication efficiency and trajectory of THz-empowered NTNs where new constraints are imposed by dynamic THz channel conditions for ground users (GUs) association. Metaverse will also support space communications where users in crewed aircraft will be able to access various kinds of Internet services with the aid of non terrestrial platforms. Other applications include space exploration where NTNs play a vital role in establishing connection to investigate the universe beyond Earth's atmosphere. [197] recently proposed the need of non terrestrial wireless communication and social connection between planets in the virtual world. The paper illustrates a vision of an interplanetary Metaverse that connects Earthian and Martian users in Metaverse.

### C. NTNs ENABLING ZERO-TOUCH NETWORKS

Evolving 6G envisions the deployment of non-terrestrial networks (NTNs) in 3D platforms UAVs, HAPSs and satellites since they provide standalone networking solutions to preserve connectivity in the absence of other already-deployed network infrastructures, or when terrestrial towers are out of service especially in rural areas. In such scenarios, manual configuration of the network will no longer be possible. Network intelligence and automation will be a must, thus the need for computationally intensive algorithms. To achieve this, energy resources will remain a challenge. To illustrate more, specifically when dealing with deep RL models that perform continual learning instead of models that follow a fixed policy, high computational cost will impose additional power consumption due to data processing operations. This will require additional energy demands from the non terrestrial platform that has limited energy resources [198]. In this regard, an important design consideration for real-world deployment is the investigation of accurate RL methods with moderate computational and energy demands to comply with the resources available to the aerial platforms. Other potential solutions are the powering using solar cells [199], [200], [201] and integrating energy harvesting solutions [202], [203], [204], which could lead to extended flight duration and further reduce energy consumption. An additional gap identified in the literature is the lack of consideration for multiple UAV charging stations in problem formulations for UAV-assisted wireless networks. This consideration is important for real-world deployment scenarios and would add a constraint on the RL-based trajectory design where the UAV would not be limited with only one choice of location for recharging its battery. Open research problems related to ambient backscatter communication where transmitters can harvest the surrounding signals and waves radiated by towers, base stations, as well as access points and reflect them towards receivers without the need of external power resources, include spectral efficiency, energy efficiency and protocol design. Regarding spectral efficiency, careful planning of backscatter devices is needed. As for energy efficiency, a large IoT network composed of hundreds or thousands of devices may still need energy efficiency optimization on a system level although individual backscatter communication devices

demonstrate good energy performance [205]. Considering protocol design, since ambient backscatter communication systems are mainly used for dedicated application-specific purposes, compatibility issues with other wireless devices need to be considered where key operation and management aspects of ambient backscatter communications, such as packet size, routing protocols, and others might need to be formalized by specific standardization methods and/or protocol design formalization.

Other open research problems are in the field of medical IoT and autonomous vehicles. The overall aim of zero-touch networks is for devices to learn how to become more autonomous so that we can perform complex tasks on them. NTN platforms will help enhance the availability of rural healthcare solutions via the Internet of Space things. Within the domain of healthcare, NTNs enabling 6G will help in disease diagnosis and treatment by integrating different components (NTN platforms, physician devices, biosensors,..) at heterogeneous levels where remote metric evaluations and treatment plans will be proposed.

Space connectivity will also help enable connected autonomous vehicles where large amount of data related to high-resolution real-time mapping of the terrain, route optimization, and traffic and safety information is exchanged between vehicles and aerial platforms. In autonomous vehicular networks a predictive model based on real-time data would be more accurate than traditional theoretical models due to mobility of vehicular nodes. Reinforcement-learning algorithms for intelligent resource management and network management problems mainly when the orchestrator performs optimal placement of virtual network functions onto the underlying physical substrate prove to be highly applicable and efficient [205]

Wide-area coverage of satellite communications together with hybrid satellite–terrestrial networks complemented high capacity shore-based systems by providing ubiquitous maritime connectivity. By employing solutions for new radio technologies to support non-terrestrial networks, 6G maritime networks can benefit from the 5-layer architecture for 6G setups as proposed in [206] to extend the coverage of terrestrial systems and provide access to maritime services in offshore areas and non-line-of-sight (NLOS) scenarios. Whenever the line-of-sight link is unavailable, reinforcement learning can help in identifying relay nodes to solve the beam misalignment problem. Since reinforcement learning requires no prior knowledge of the environment, it helps in identifying optimal relay nodes in dynamic maritime environments where beam misalignment leads to data rate and energy efficiency deterioration [206]. To tackle such challenges a recent study in [207] proposes a deep reinforcement learning algorithm to solve the alignment issue by obtaining the optimal beam divergence angle to maximize the link availability. Another study proposes an RL-based approach for optimizing positioning and beam width of the light source for underwater wireless communication [208].

### D. NTN-AIDED PERVASIVE COMPUTING

Communication implies computation everywhere. As different devices are performing different heterogeneous in a multiagent stochastic environment attention to the RL algorithm should be considered. Deep RL techniques dominate as the choice of the algorithm in the majority of the surveyed articles where proposed approaches were evaluated in simulated environments with limited considerations of the non terrestrial platform resources especially UAV resources. UAVs are sometimes used as edge servers, so they are expected to carry computational resources [209], [210]. However, these resources would be limited. Hence, if UAVs are to be operated in the real world, and if the computational load is expected to take place at the UAV side and not the base station side, the adopted RL methods need to be computationally efficient for real-time decision making. This would be difficult when using deep RL methods that rely on complex neural network architectures with high computational costs. The problem is augmented in cases where incremental learning is applied, where the agent will be continuously learning from its interactions with the environment while being in operation. Suitable selection mechanisms of the device hardware that is suitable for deep learning tasks [211], [212], [213] and RL technique is needed.

A critical issue is the location of data storage and that used to be in cloud data centers. For devices distributed in a wide geographic area, this introduces significant performance delays. Edge AI pushes operation and management tasks to local devices. This will increase the burden on local devices since they are not equipped with as powerful processing units as the cloud processing center. Research efforts in accelerating the hardware's processing capability, and increasing the coordination between local and central processing units to optimization task distribution are being introduced [205].

Federated learning concept can be implemented where generated raw data is used locally to train a local model and then send the local trained model to the central node for aggregation. This will help in minimizing communication overhead and latency. Moreover less data will be communicated which ensures better privacy preservation. How to use federated learning with integrated space-terrestrial networks woud be another challenge. A critical open research question is how to jointly optimize aerial station locations, resource allocation, and training parameters to boost the learning process [214], [215].

Other challenging problems arise with the introduction of multi-access mobile edge computing and intelligent computation offloading. Non terrestrial aided pervasive computation allows different devices to be involved in the computation process. Due to energy and computation resource constraints of aerial platforms, especially for UAVs, offloading computationally heavy tasks from cellular-connected NTNs to edge nodes will improve the network perseverance. In this regard, joint task offloading, commu-

nication and computation resource allocation problems to minimize the energy consumption of mobile devices and UAVs and/or latency especially in a multi-UAV scenario can be formulated and solved using reinforcement learning methods [216], [217].

In a multi-NTN platform system, action coordination of individual NTN device is required so that mission is complete in the best possible way. In order to adapt to the environment with uncertain changes, the system should decide on where aerial platforms should move and what tasks to perform. Coordination algorithms can be classified based on the actions they need to decide on, data to use for decision making, the decision making algorithms and decentralization degree [218].

### E. SECURITY

Machine learning has recently drawn research attention in terms of security in diverse systems and platforms of satellite-terrestrial communication and more research is needed in this area. One of the main open problems is that traditional terrestrial security approaches are adopted and they are not sufficient for NTNs. Even though same security challenges exist such as DoS and jamming attacks, however these do not apply due to latency and high mobility involved. Key management for cryptographic protocols is considered critical in NTNs. Furthermore, security measures should be applied on ground-based stations including gateways and end user IoT devices since they are prone to be used as launchpads for security attacks. Reinforcement learning can aid in secure computation offloading, as proposed in [219] to meet security challenges arising due to lack of resources on board in satellites. Authors implemented RL methods to dynamically alter the computation offloading policies for different scenarios based on threat levels. Techniques that require high energy and computation resources should only be used in cases of serious security threats. Blockchain-based techniques have been proposed and proved to improve security through distributed computing using ground-based cellular networks [220]. Thus blockchain technologies can be implemented to enhance communication security between terrestrial and non terrestrial stations [221].

## VIII. CONCLUSION

RL has been an attractive choice for researchers aiming to achieve various control objectives in NTN-aided wireless communications and cellular-connected NTNs. RL techniques can reach an optimal control policy that the NTN platform can adopt to satisfy the desired objective. In this paper, we surveyed the literature for the different RL formulations applied to solve control problems in NTN communications, with a focus on MDP formulations. We consider the two integration scenarios where non terrestrial platforms are deployed as aerial base stations or relays to assist wireless networks or connected to the cellular network as aerial user equipment. While many surveys in the literature have addressed different aspects of NTN communications, no survey has comprehensively tackled the applications of RL. In this respect, we synthesize a taxonomy from the surveyed literature that represents the investigated objectives of RL in the context of NTN communications.

Despite the promising results achieved in the literature by using RL, many challenges remain to be addressed before RL techniques can be used in real-world non terrestrial platform deployment. An important design consideration for is the investigation of accurate RL methods with moderate computational and energy demands to comply with the resources available to the aerial platforms. Problem formulations should mimic real world multi-agent stochastic scenarios more accurately. Other aspects that need to be considered are integration with 3D virtual reality where networks are expected to support super-high-definition (SHD) and extremely high-definition (EHD) videos, with super-high throughput demands and to provide ultra-reliable low-latency communications. To achieve this, we need non-terrestrial platforms to assist terrestrial networks. Moreover, space and underwater connectivity, autonomous devices, backscatter communication and energy harvesting are to

**TABLE 6. List of Acronyms.**

| | |
|---|---|
| 5G | Fifth-Generation |
| 6G | Sixth-Generation |
| AoI | Age of Information |
| BER | Bit Error Rate |
| CA2C | Compound-Action Actor-Critic |
| CSI | Channel State Information |
| DDPG | Deep Deterministic Policy Gradient |
| DDQN | Double Deep Q-Network |
| DPG | Deterministic Policy Gradient |
| DQN | Deep Q-Network |
| ER | Experience Replay |
| IAB | Integrated Access and Backhaul |
| IoT | Internet of Things |
| IRS | Intelligent Reflecting Surfaces |
| KPI | Key Performance Indicator |
| LoS | Line of Sight |
| LSTM | Long Short-Term Memory |
| MADDPG | Multi-Agent Deep Deterministic Policy Gradient |
| MDP | Markov Decision Process |
| MOS | Mean Opinion Score |
| N-LoS | No Line of Sight |
| PER | Prioritized Experience Replay |
| QoE | Quality of Experience |
| QoS | Quality of Service |
| RL | Reinforcement Learning |
| SINR | Signal to Interference and Noise Ratio |
| UAV | Unmanned Aerial Vehicle |
| MCMOPSO-RL | Multi-Objective Particle Swarm Optimization algorithm with Multi-mode Collaboration based on Reinforcement Learning |
| MA-DRL | Multi-agent deep reinforcement learning |
| MA-LB-DRL | Multi-Agent Low-Bias Deep Reinforcement Learning |
| GNN-QL | Graph Neural Network Q-learning |
| REL-DDPG | Relevant Experience Learning-DDPG |
| MAT3DPG | Multi-Agent Twin Delayed Deep Deterministic Policy Gradient |
| MASAC | Multi-Agent Soft Actor Critic |
| HTS | High Through-put Satellite |
| AVU | Aerial Vehicle User |

considered in the context of non terrestrial networks as stated in VII. As machine learning has recently been implemented in diverse systems and platforms of satellite-terrestrial communication for secure communication, more research is needed in this area. One of the main open problems is developing security mechanisms that tailor to the design and functionality of NTN platforms rather than utilizing or customizing existing traditional terrestrial security approaches.

## APPENDIX

See Table 6.

## REFERENCES


[1] F. A. Dicandia, N. J. G. Fonseca, M. Bacco, S. Mugnaini, and S. Genovesi, "Space-air-ground integrated 6G wireless communication networks: A review of antenna technologies and application scenarios," *Sensors*, vol. 22, no. 9, p. 3136, Apr. 2022.

[2] X. Lin, S. Rommer, S. Euler, E. A. Yavuz, and R. S. Karlsson, "5G from space: An overview of 3GPP non-terrestrial networks," *IEEE Commun. Standards Mag.*, vol. 5, no. 4, pp. 147–153, Dec. 2021.

[3] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 36–42, May 2016.

[4] X. Lin, V. Yajnanarayana, S. D. Muruganathan, S. Gao, H. Asplund, H.-L. Maattanen, M. Bergstrom, S. Euler, and Y.-P. E. Wang, "The sky is not the limit: LTE for unmanned aerial vehicles," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 204–210, Apr. 2018.

[5] X. Lin, Z. Zou, V. Yajnanarayana, R. Wiren, S. Euler, A. Sadam, H.-L. Maattanen, S. Muruganathan, S. Gao, Y.-P.-E. Wang, and J. Kauppi, "Mobile network-connected drones: Field trials, simulations, and design insights," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 115–125, Sep. 2019.

[6] L. Zhang, H. Zhao, S. Hou, Z. Zhao, H. Xu, X. Wu, Q. Wu, and R. Zhang, "A survey on 5G millimeter wave communications for UAV-assisted wireless networks," *IEEE Access*, vol. 7, pp. 117460–117504, 2019.

[7] H. Wang, J. Wang, J. Chen, Y. Gong, and G. Ding, "Network-connected UAV communications: Potentials and challenges," *China Commun.*, vol. 15, no. 12, pp. 111–121, Dec. 2018.

[8] H. Zhang, L. Song, and Z. Han, *Unmanned Aerial Vehicle Applications Over Cellular Networks for 5G and Beyond*. Cham, Switzerland: Springer, 2020.

[9] F. Rinaldi, H.-L. Maattanen, J. Torsner, S. Pizzi, S. Andreev, A. Iera, Y. Koucheryavy, and G. Araniti, "Non-terrestrial networks in 5G & beyond: A survey," *IEEE Access*, vol. 8, pp. 165178–165200, 2020.

[10] N. Cheng, J. He, Z. Yin, C. Zhou, H. Wu, F. Lyu, H. Zhou, and X. Shen, "6G service-oriented space-air-ground integrated network: A survey," *Chin. J. Aeronaut.*, vol. 35, no. 9, pp. 1–18, Sep. 2022.

[11] J. Ye, J. Qiao, A. Kammoun, and M.-S. Alouini, "Nonterrestrial communications assisted by reconfigurable intelligent surfaces," *Proc. IEEE*, vol. 110, no. 9, pp. 1423–1465, Sep. 2022.

[12] M. Vaezi, A. Azari, S. R. Khosravirad, M. Shirvanimoghaddam, M. M. Azari, D. Chasaki, and P. Popovski, "Cellular, wide-area, and non-terrestrial IoT: A survey on 5G advances and the road toward 6G," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 2, pp. 1117–1174, 2nd Quart., 2022.

[13] N. Saeed, H. Almorad, H. Dahrouj, T. Y. Al-Naffouri, J. S. Shamma, and M.-S. Alouini, "Point-to-point communication in integrated satellite-aerial 6G networks: State-of-the-art and future challenges," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 1505–1525, 2021.

[14] M. Bacco, F. Davoli, G. Giambene, A. Gotta, M. Luglio, M. Marchese, F. Patrone, and C. Roseti, "Networking challenges for non-terrestrial networks exploitation in 5G," in *Proc. IEEE 2nd 5G World Forum (5GWF)*, Sep. 2019, pp. 623–628.

[15] M. Giordani and M. Zorzi, "Non-terrestrial networks in the 6G era: Challenges and opportunities," *IEEE Netw.*, vol. 35, no. 2, pp. 244–251, Mar./Apr. 2021.

[16] P. Wang, J. Zhang, X. Zhang, Z. Yan, B. G. Evans, and W. Wang, "Convergence of satellite and terrestrial networks: A comprehensive survey," *IEEE Access*, vol. 8, pp. 5550–5588, 2020.

[17] A. Guidotti, A. Vanelli-Coralli, M. Conti, S. Andrenacci, S. Chatzinotas, N. Maturo, B. Evans, A. Awoseyila, A. Ugolini, T. Foggi, L. Gaudio, N. Alagha, and S. Cioni, "Architectures and key technical challenges for 5G systems incorporating satellites," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2624–2639, Mar. 2019.

[18] A. Guidotti, S. Cioni, G. Colavolpe, M. Conti, T. Foggi, A. Mengali, G. Montorsi, A. Piemontese, and A. Vanelli-Coralli, "Architectures, standardisation, and procedures for 5G satellite communications: A survey," *Comput. Netw.*, vol. 183, Dec. 2020, Art. no. 107588.

[19] O. Kodheli, "Satellite communications in the new space era: A survey and future challenges," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 1, pp. 70–109, 4th Quart., 2021.

[20] S. C. Arum, D. Grace, and P. D. Mitchell, "A review of wireless communication using high-altitude platforms for extended coverage and capacity," *Comput. Commun.*, vol. 157, pp. 232–256, May 2020.

[21] K. Katzis, L. Mfupe, and H. M. Hussien, "Opportunities and challenges of bridging the digital divide using 5G enabled high altitude platforms and TVWS spectrum," in *Proc. IEEE 8th Int. Conf. Commun. Netw. (ComNet)*, Oct. 2020, pp. 1–7.

[22] K. Mershad, H. Dahrouj, H. Sarieddeen, B. Shihada, T. Al-Naffouri, and M.-S. Alouini, "Cloud-enabled high-altitude platform systems: Challenges and opportunities," 2021, *arXiv:2106.02006*.

[23] M. S. Frikha, S. M. Gammar, A. Lahmadi, and L. Andrey, "Reinforcement and deep reinforcement learning for wireless Internet of Things: A survey," *Comput. Commun.*, vol. 178, pp. 98–113, Oct. 2021.

[24] C. Ssengonzi, O. P. Kogeda, and T. O. Olwal, "A survey of deep reinforcement learning application in 5G and beyond network slicing and virtualization," *Array*, vol. 14, Jul. 2022, Art. no. 100142.

[25] S. Zhang, D. Zhu, and Y. Wang, "A survey on space-aerial-terrestrial integrated 5G networks," *Comput. Netw.*, vol. 174, Jun. 2020, Art. no. 107212.

[26] F. Fourati and M.-S. Alouini, "Artificial intelligence for satellite communication: A review," *Intell. Converged Netw.*, vol. 2, no. 3, pp. 213–243, Sep. 2021.

[27] Y. Zeng, Q. Wu, and R. Zhang, "Accessing from the sky: A tutorial on UAV communications for 5G and beyond," *Proc. IEEE*, vol. 107, no. 12, pp. 2327–2375, Mar. 2019.

[28] X. Cao, P. Yang, M. Alzenad, X. Xi, D. Wu, and H. Yanikomeroglu, "Airborne communication networks: A survey," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 10, pp. 1907–1926, Sep. 2018.

[29] A. I. Hentati and L. C. Fourati, "Comprehensive survey of UAVs communication networks," *Comput. Standards Interface*, vol. 72, Oct. 2020, Art. no. 103451.

[30] M. Mozaffari, W. Saad, M. Bennis, Y.-H. Nam, and M. Debbah, "A tutorial on UAVs for wireless networks: Applications, challenges, and open problems," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2334–2360, 3rd Quart., 2019.

[31] D. Mishra and E. Natalizio, "A survey on cellular-connected UAVs: Design challenges, enabling 5G/B5G innovations, and experimental advancements," *Comput. Netw.*, vol. 182, Dec. 2020, Art. no. 107451.

[32] A. Fotouhi, H. Qiang, M. Ding, M. Hassan, L. G. Giordano, A. Garcia-Rodriguez, and J. Yuan, "Survey on UAV cellular communications: Practical aspects, standardization advancements, regulation, and security challenges," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3417–3442, 4th Quart., 2019.

[33] B. Li, Z. Fei, and Y. Zhang, "UAV communications for 5G and beyond: Recent advances and future trends," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2241–2263, Apr. 2019.

[34] Z. Ullah, F. Al-Turjman, and L. Mostarda, "Cognition in UAV-aided 5G and beyond communications: A survey," *IEEE Trans. Cognit. Commun. Netw.*, vol. 6, no. 3, pp. 872–891, Sep. 2020.

[35] O. S. Oubbati, M. Atiquzzaman, T. A. Ahanger, and A. Ibrahim, "Softwarization of UAV networks: A survey of applications and future trends," *IEEE Access*, vol. 8, pp. 98073–98125, 2020.

[36] L. Gupta, R. Jain, and G. Vaszkun, "Survey of important issues in UAV communication networks," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1123–1152, 2nd Quart., 2016.

[37] S. Aggarwal and N. Kumar, "Path planning techniques for unmanned aerial vehicles: A review, solutions, and challenges," *Comput. Commun.*, vol. 149, pp. 270–299, Jan. 2020.

[38] Y. Zhao, Z. Zheng, and Y. Liu, "Survey on computational-intelligence-based UAV path planning," *Knowl.-Based Syst.*, vol. 158, pp. 54–64, Dec. 2018.

[39] N. H. Motlagh, T. Taleb, and O. Arouk, "Low-altitude unmanned aerial vehicles-based Internet of Things services: Comprehensive survey and future perspectives," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 899–922, Dec. 2016.

[40] M. E. Mkiramweni, C. Yang, J. Li, and W. Zhang, "A survey of game theory in unmanned aerial vehicles communications," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3386–3416, 4th Quart., 2019.

[41] P. S. Bithas, E. T. Michailidis, N. Nomikos, D. Vouyioukas, and A. G. Kanatas, "A survey on machine-learning techniques for UAV-based communications," *Sensors*, vol. 19, no. 23, p. 5170, 2019.

[42] P. V. Klaine, R. D. Souza, L. Zhang, and M. Imran, "An overview of machine learning applied in wireless UAV networks," in *Wiley 5G Ref: The Essential 5G Reference Online*. U.K.: Wiley, 2019, pp. 1–15.

[43] B. Brik, A. Ksentini, and M. Bouaziz, "Federated learning for UAVs-enabled wireless networks: Use cases, challenges, and open problems," *IEEE Access*, vol. 8, pp. 53841–53849, 2020.

[44] U. Challita, A. Ferdowsi, M. Chen, and W. Saad, "Machine learning for wireless connectivity and security of cellular-connected UAVs," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 28–35, Feb. 2019.

[45] Z. Ullah, F. Al-Turjman, U. Moatasim, L. Mostarda, and R. Gagliardi, "UAVs joint optimization problems and machine learning to improve the 5G and beyond communication," *Comput. Netw.*, vol. 182, Dec. 2020, Art. no. 107478.

[46] M.-A. Lahmeri, M. A. Kishk, and M.-S. Alouini, "Artificial intelligence for UAV-enabled wireless networks: A survey," 2020, *arXiv:2009.11522*.

[47] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133–3174, 4th Quart., 2019.

[48] Y. Qian, J. Wu, R. Wang, F. Zhu, and W. Zhang, "Survey on reinforcement learning applications in communication networks," *J. Commun. Inf. Netw.*, vol. 4, no. 2, pp. 30–39, Jun. 2019.

[49] Y. Huang, C. Xu, C. Zhang, M. Hua, and Z. Zhang, "An overview of intelligent wireless communications using deep reinforcement learning," *J. Commun. Inf. Netw.*, vol. 4, no. 2, pp. 15–29, 2019.

[50] A. Feriani and E. Hossain, "Single and multi-agent deep reinforcement learning for AI-enabled wireless networks: A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 2, pp. 1226–1252, 2nd Quart., 2021.

[51] L. Lei, Y. Tan, K. Zheng, S. Liu, K. Zhang, and X. Shen, "Deep reinforcement learning for autonomous Internet of Things: Model, applications and challenges," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1722–1760, 3rd Quart., 2020.

[52] W. Chen, X. Qiu, T. Cai, H.-N. Dai, Z. Zheng, and Y. Zhang, "Deep reinforcement learning for Internet of Things: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 1659–1692, 3rd Quart., 2021.

[53] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[54] E. Brunskill.(2022). *Cs234: Reinforcement Learning Winter 2022*. [Online]. Available: https://web.stanford.edu/class/cs234/

[55] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1126–1135.

[56] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel, "A simple neural attentive meta-learner," 2017, *arXiv:1707.03141*.

[57] J. Rothfuss, D. Lee, I. Clavera, T. Asfour, and P. Abbeel, "ProMP: Proximal meta-policy search," 2018, *arXiv:1810.06784*.

[58] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 387–395.

[59] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1928–1937.

[60] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.

[61] A. Abdolmaleki, J. T. Springenberg, Y. Tassa, R. Munos, N. Heess, and M. Riedmiller, "Maximum *a posteriori* policy optimisation," 2018, *arXiv:1806.06920*.

[62] 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," 2015, *arXiv:1509.02971*.

[63] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Advances in Neural Information Processing Systems*, vol. 30, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/68a9750337a418a86fe06c1991a1d64c-Paper.pdf

[64] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1861–1870.

[65] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," 2013, *arXiv:1312.5602*.

[66] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 449–458.

[67] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, P. Abbeel, and W. Zaremba, "Hindsight experience replay," in *Advances in Neural Information Processing Systems*, vol. 30, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/453fadbd8a1a3af50a9df4df899537b5-Paper.pdf

[68] S. Kapturowski, G. Ostrovski, W. Dabney, J. Quan, and R. Munos, "Recurrent experience replay in distributed reinforcement learning," in *Proc. Int. Conf. Learn. Represent.*, 2019. [Online]. Available: https://openreview.net/forum?id=r1lyTjAqYX

[69] A. P. Badia, B. Piot, S. Kapturowski, P. Sprechmann, A. Vitvitskyi, Z. D. Guo, and C. Blundell, "Agent57: Outperforming the Atari human benchmark," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 507–517.

[70] A. P. Badia, P. Sprechmann, A. Vitvitskyi, D. Guo, B. Piot, S. Kapturowski, O. Tieleman, M. Arjovsky, A. Pritzel, A. Bolt, and C. Blundell, "Never give up: Learning directed exploration strategies," 2020, *arXiv:2002.06038*.

[71] S. Racanière, T. Weber, D. Reichert, L. Buesing, A. Guez, D. J. Rezende, A. P. Badia, O. Vinyals, N. Heess, Y. Li, R. Pascanu, P. Battaglia, D. Hassabis, D. Silver, and D. Wierstra, "Imagination-augmented agents for deep reinforcement learning," in *Advances in Neural Information Processing Systems*, vol. 30, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/9e82757e9a1c12cb710ad680db11f6f1-Paper.pdf

[72] D. Ha and J. Schmidhuber, "Recurrent world models facilitate policy evolution," in *Advances in Neural Information Processing Systems*, vol. 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran, 2018. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/2de5d16682c3c35007e4e92982f1a2ba-Paper.pdf

[73] A. Nagabandi, G. Kahn, R. S. Fearing, and S. Levine, "Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 7559–7566.

[74] V. Feinberg, A. Wan, I. Stoica, M. I. Jordan, J. E. Gonzalez, and S. Levine, "Model-based value estimation for efficient model-free reinforcement learning," 2018, *arXiv:1803.00101*.

[75] J. Schrittwieser, I Antonoglou, T Hubert, K Simonyan, L Sifre, S Schmitt, A Guez, E Lockhart, D Hassabis, T Graepel, and T. Lillicrap, "Mastering Atari, Go, chess and Shogi by planning with a learned model," *Nature*, vol. 588, no. 7839, pp. 604–609, 2020.

[76] L. Kaiser, M. Babaeizadeh, P. Milos, B. Osinski, R. H Campbell, K. Czechowski, D. Erhan, C. Finn, P. Kozakowski, S. Levine, A. Mohiuddin, R. Sepassi, G. Tucker, and H. Michalewski, "Model-based reinforcement learning for Atari," 2019, *arXiv:1903.00374*.

[77] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis, "Mastering chess and shogi by self-play with a general reinforcement learning algorithm," 2017, *arXiv:1712.01815*.

[78] J. Peters and J. A. Bagnell, "Policy gradient methods," *Scholarpedia*, vol. 5, no. 11, p. 3698, 2010.

[79] C. Wen, Y. Fang, and L. Qiu, "Securing UAV communication based on multi-agent deep reinforcement learning in the presence of smart UAV eavesdropper," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2022, pp. 1164–1169.

[80] A. H. Arani, M. M. Azari, P. Hu, Y. Zhu, H. Yanikomeroglu, and S. Safavi-Naeini, "Reinforcement learning for energy-efficient trajectory design of UAVs," *IEEE Internet Things J.*, vol. 9, no. 11, pp. 9060–9070, Jun. 2022.

[81] Z. Zhang, Y. Xiao, Z. Ma, M. Xiao, Z. Ding, X. Lei, G. K. Karagiannidis, and P. Fan, "6G wireless networks: Vision, requirements, architecture, and key technologies," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 28–41, Sep. 2019.

[82] S. Zhou, Y. Cheng, X. Lei, Q. Peng, J. Wang, and S. Li, "Resource allocation in UAV-assisted networks: A clustering-aided reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 71, no. 11, pp. 12088–12103, Nov. 2022.

[83] Y. Zeng, J. Lyu, and R. Zhang, "Cellular-connected UAV: Potential, challenges, and promising technologies," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 120–127, Feb. 2019.

[84] S. Zhang, Y. Zeng, and R. Zhang, "Cellular-enabled UAV communication: A connectivity-constrained trajectory optimization perspective," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2580–2604, Mar. 2019.

[85] S. Chandrasekharan, K. Gomez, A. Al-Hourani, S. Kandeepan, T. Rasheed, L. Goratti, L. Reynaud, D. Grace, I. Bucaille, and T. Wirth, "Designing and implementing future aerial communication networks," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 26–34, May 2016.

[86] M. M. Azari, H. Sallouha, A. Chiumento, S. Rajendran, E. Vinogradov, and S. Pollin, "Key technologies and system trade-offs for detection and localization of amateur drones," *IEEE Commun. Mag.*, vol. 56, no. 1, pp. 51–57, Jan. 2018.

[87] M. M. Azari, F. Rosas, A. Chiumento, and S. Pollin, "Coexistence of terrestrial and aerial users in cellular networks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2017, pp. 1–6.

[88] M. M. Azari, F. Rosas, and S. Pollin, "Reshaping cellular networks for the sky: Major factors and feasibility," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–7.

[89] A. Azari, F. Ghavimi, M. Ozger, R. Jantti, and C. Cavdar, "Machine learning assisted handover and resource management for cellular connected drones," in *Proc. IEEE 91st Veh. Technol. Conf. (VTC-Spring)*, May 2020, pp. 1–7.

[90] S.-Y. Lien and D.-J. Deng, "Autonomous non-terrestrial base station deployment for non-terrestrial networks: A reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 71, no. 10, pp. 10894–10909, Oct. 2022.

[91] O. Anicho, P. B. Charlesworth, G. S. Baicher, and A. K. Nagar, "Reinforcement learning versus swarm intelligence for autonomous multi-HAPS coordination," *Social Netw. Appl. Sci.*, vol. 3, no. 6, pp. 1–11, Jun. 2021.

[92] H. Huang, Y. Yang, H. Wang, Z. Ding, H. Sari, and F. Adachi, "Deep reinforcement learning for UAV navigation through massive MIMO technique," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 1117–1121, Jan. 2020.

[93] X. Liu, M. Chen, and C. Yin, "Optimized trajectory design in UAV based cellular networks: A double Q-learning approach," in *Proc. IEEE Int. Conf. Commun. Syst. (ICCS)*, Dec. 2018, pp. 13–18.

[94] J. Qiu, J. Lyu, and L. Fu, "Placement optimization of aerial base stations with deep reinforcement learning," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2020, pp. 1–6.

[95] C. H. Liu, Z. Chen, J. Tang, J. Xu, and C. Piao, "Energy-efficient UAV control for effective and fair communication coverage: A deep reinforcement learning approach," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 2059–2070, Sep. 2018.

[96] C. H. Liu, X. Ma, X. Gao, and J. Tang, "Distributed energy-efficient multi-UAV navigation for long-term communication coverage by deep reinforcement learning," *IEEE Trans. Mobile Comput.*, vol. 19, no. 6, pp. 1274–1285, Jun. 2020.

[97] Y.-J. Chen, W. Chen, and M.-L. Ku, "Trajectory design and link selection in UAV-assisted hybrid satellite-terrestrial network," *IEEE Commun. Lett.*, vol. 26, no. 7, pp. 1643–1647, Jul. 2022.

[98] S. Yin, S. Zhao, Y. Zhao, and F. R. Yu, "Intelligent trajectory design in UAV-aided communications with reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8227–8231, Aug. 2019.

[99] H. Bayerlein, P. D. Kerret, and D. Gesbert, "Trajectory optimization for autonomous flying base station via reinforcement learning," in *Proc. IEEE 19th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jun. 2018, pp. 1–5.

[100] N. Dai, D. Zhou, M. Sheng, and J. Li, "Deep reinforcement learning based power allocation for high throughput satellites," in *Proc. IEEE 94th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2021, pp. 1–5.

[101] T. P. Truong, A.-T. Tran, T. M. T. Nguyen, T.-V. Nguyen, A. Masood, and S. Cho, "MEC-enhanced aerial serving networks via HAP: A deep reinforcement learning approach," in *Proc. Int. Conf. Inf. Netw. (ICOIN)*, Jan. 2022, pp. 319–323.

[102] Y. Cui, D. Deng, C. Wang, and W. Wang, "Joint trajectory and power optimization for energy efficient UAV communication using deep reinforcement learning," in *Proc. IEEE INFOCOM Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, May 2021, pp. 1–6.

[103] W. Guo, "Partially explainable big data driven deep reinforcement learning for green 5G UAV," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2020, pp. 1–7.

[104] J. Cui, Y. Liu, and A. Nallanathan, "Multi-agent reinforcement learning-based resource allocation for UAV networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 729–743, Feb. 2020.

[105] C. Zhan and Y. Zeng, "Energy minimization for cellular-connected UAV: From optimization to deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 21, no. 7, pp. 5541–5555, Jul. 2022.

[106] S. Tang, Z. Pan, G. Hu, Y. Wu, and Y. Li, "Deep reinforcement learning-based resource allocation for satellite Internet of Things with diverse QoS guarantee," *Sensors*, vol. 22, no. 8, p. 2979, Apr. 2022.

[107] X. Liu, Y. Liu, and Y. Chen, "Reinforcement learning in multiple-UAV networks: Deployment and movement design," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8036–8049, Aug. 2019.

[108] X. Liu, Y. Liu, and Y. Chen, "Deployment and movement for multiple aerial base stations by reinforcement learning," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2018, pp. 1–6.

[109] S. Gong, X. Lu, D. T. Hoang, D. Niyato, L. Shu, D. I. Kim, and Y.-C. Liang, "Toward smart wireless communications via intelligent reflecting surfaces: A contemporary survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2283–2314, 4th Quart., 2020.

[110] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5394–5409, Nov. 2019.

[111] R. Alghamdi, R. Alhadrami, D. Alhothali, H. Almorad, A. Faisal, S. Helal, R. Shalabi, R. Asfour, N. Hammad, A. Shams, N. Saeed, H. Dahrouj, T. Y. Al-Naffouri, and M.-S. Alouini, "Intelligent surfaces for 6G wireless networks: A survey of optimization and performance analysis techniques," *IEEE Access*, vol. 8, pp. 202795–202818, 2020.

[112] Q. Zhang, W. Saad, and M. Bennis, "Reflections in the sky: Millimeter wave communication with UAV-carried intelligent reflectors," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[113] L. Wang, K. Wang, C. Pan, and N. Aslam, "Joint trajectory and passive beamforming design for intelligent reflecting surface-aided UAV communications: A deep reinforcement learning approach," 2020, *arXiv:2007.08380*.

[114] *Study on Integrated Access and Backhaul (Release 16)*, Standard TR 38.874, 3GPP, 2019.

[115] A. Fotouhi, M. Ding, and M. Hassan, "Deep Q-learning for two-hop communications of drone base stations," *Sensors*, vol. 21, no. 6, p. 1960, Mar. 2021.

[116] Y. Cao, S.-Y. Lien, and Y.-C. Liang, "Deep reinforcement learning for multi-user access control in non-terrestrial networks," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 1605–1619, Mar. 2021.

[117] J.-H. Lee, J. Park, M. Bennis, and Y.-C. Ko, "Integrating LEO satellite and UAV relaying via reinforcement learning for non-terrestrial networks," in *Proc. GLOBECOM IEEE Global Commun. Conf.*, Dec. 2020, pp. 1–6.

[118] A. Fotouhi, M. Ding, L. G. Giordano, M. Hassan, J. Li, and Z. Lin, "Joint optimization of access and backhaul links for UAVs based on reinforcement learning," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2019, pp. 1–6.

[119] N. Tafintsev, D. Moltchanov, M. Simsek, S.-P. Yeh, S. Andreev, Y. Koucheryavy, and M. Valkama, "Reinforcement learning for improved UAV-based integrated access and backhaul operation," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Jun. 2020, pp. 1–7.

[120] A. H. Arani, P. Hu, and Y. Zhu, "Fairness-aware link optimization for space-terrestrial integrated networks: A reinforcement learning framework," *IEEE Access*, vol. 9, pp. 77624–77636, 2021.

[121] Y. Huo, Y. Tian, L. Ma, X. Cheng, and T. Jing, "Jamming strategies for physical layer security," *IEEE Wireless Commun.*, vol. 25, no. 1, pp. 148–153, Feb. 2017.

[122] X. Lu, L. Xiao, C. Dai, and H. Dai, "UAV-aided cellular communications with deep reinforcement learning against jamming," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 48–53, Aug. 2020.

[123] S. Zhou, Y. Cheng, and X. Lei, "Multi-agent low-bias reinforcement learning for resource allocation in UAV-assisted networks," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, May 2022, pp. 1011–1016.

[124] L. Xiao, C. Xie, M. Min, and W. Zhuang, "User-centric view of unmanned aerial vehicle transmission against smart attacks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3420–3430, Apr. 2018.

[125] H. Wang, J. Chen, G. Ding, and J. Sun, "Trajectory planning in UAV communication with jamming," in *Proc. 10th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2018, pp. 1–6.

[126] J. Liu, N. Sha, W. Yang, J. Tu, and L. Yang, "Hierarchical Q-learning based UAV secure communication against multiple UAV adaptive eavesdroppers," *Wireless Commun. Mobile Comput.*, vol. 2020, pp. 1–15, Oct. 2020.

[127] Y. Zhang, Z. Zhuang, F. Gao, J. Wang, and Z. Han, "Multi-agent deep reinforcement learning for secure UAV communications," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, May 2020, pp. 1–5.

[128] Y. Zhang, Z. Mou, F. Gao, J. Jiang, R. Ding, and Z. Han, "UAV-enabled secure communications by multi-agent deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 11599–11611, Oct. 2020.

[129] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[130] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, Y. Bengio and Y. LeCun, Eds. May 2015. [Online]. Available: http://arxiv.org/abs/1409.0473

[131] F. Fu, Q. Jiao, F. R. Yu, Z. Zhang, and J. Du, "Securing UAV-to-vehicle communications: A curiosity-driven deep Q-learning network (C-DQN) approach," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Jun. 2021, pp. 1–6.

[132] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 2731–2735.

[133] J. Liu, X. Wang, B. Bai, and H. Dai, "Age-optimal trajectory planning for UAV-assisted data collection," in *Proc. IEEE INFOCOM Conf. Comput. Commun. Workshops*, vol. 25, no. 1, Apr. 2018, pp. 553–558.

[134] W. Li, L. Wang, and A. Fei, "Minimizing packet expiration loss with path planning in UAV-assisted data sensing," *IEEE Wireless Commun. Lett.*, vol. 8, no. 6, pp. 1520–1523, Dec. 2019.

[135] M. A. Abd-Elmagid, A. Ferdowsi, H. S. Dhillon, and W. Saad, "Deep reinforcement learning for minimizing age-of-information in UAV-assisted networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[136] A. Ferdowsi, M. A. Abd-Elmagid, W. Saad, and H. S. Dhillon, "Neural combinatorial deep reinforcement learning for age-optimal joint trajectory and scheduling design in UAV-assisted networks," 2020, *arXiv:2006.15863*.

[137] M. Samir, C. Assi, S. Sharafeddine, and A. Ghrayeb, "Online altitude control and scheduling policy for minimizing AoI in UAV-assisted IoT wireless networks," *IEEE Trans. Mobile Comput.*, vol. 21, no. 7, pp. 2493–2505, Jul. 2020.

[138] M. Samir, M. Elhattab, C. Assi, S. Sharafeddine, and A. Ghrayeb, "Optimizing age of information through aerial reconfigurable intelligent surfaces: A deep reinforcement learning approach," 2020, *arXiv:2011.04817*.

[139] M. Yi, X. Wang, J. Liu, Y. Zhang, and B. Bai, "Deep reinforcement learning for fresh data collection in UAV-assisted IoT networks," 2020, *arXiv:2003.00391*.

[140] S. F. Abedin, M. S. Munir, N. H. Tran, Z. Han, and C. S. Hong, "Data freshness and energy-efficient UAV navigation optimization: A deep reinforcement learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 9, pp. 5994–6006, Sep. 2021.

[141] J. Hu, H. Zhang, K. Bian, L. Song, and Z. Han, "Distributed trajectory design for cooperative Internet of UAVs using deep reinforcement learning," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[142] F. Wu, H. Zhang, J. Wu, L. Song, Z. Han, and H. Vincent Poor, "UAV-to-device underlay communications: Age of information minimization by multi-agent deep reinforcement learning," 2020, *arXiv:2003.05830*.

[143] J. Hu, H. Zhang, L. Song, R. Schober, and H. V. Poor, "Cooperative internet of UAVs: Distributed trajectory design by multi-agent deep reinforcement learning," *IEEE Trans. Commun.*, vol. 68, no. 11, pp. 6807–6821, Nov. 2020.

[144] M. Samir, M. Elhattab, C. Assi, S. Sharafeddine, and A. Ghrayeb, "Optimizing age of information through aerial reconfigurable intelligent surfaces: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 70, no. 4, pp. 3978–3983, Apr. 2021.

[145] N. Waqar, S. A. Hassan, A. Mahmood, K. Dev, D.-T. Do, and M. Gidlund, "Computation offloading and resource allocation in MEC-enabled integrated aerial-terrestrial vehicular networks: A reinforcement learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 21478–21491, Nov. 2022.

[146] L. Wei, Y. Chen, M. Chen, and Y. Chen, "Deep reinforcement learning and parameter transfer based approach for the multi-objective agile Earth observation satellite scheduling problem," *Appl. Soft Comput.*, vol. 110, Oct. 2021, Art. no. 107607.

[147] Y. Cao, S.-Y. Lien, and Y.-C. Liang, "Multi-tier collaborative deep reinforcement learning for non-terrestrial network empowered vehicular connections," in *Proc. IEEE 29th Int. Conf. Netw. Protocols (ICNP)*, Nov. 2021, pp. 1–6.

[148] S. Yang, Z. Shan, J. Cao, Y. Gao, Y. Guo, P. Wang, X. Wang, J. Wang, T. Zhang, and J. Guo, "Path planning of UAV base station based on deep reinforcement learning," *Proc. Comput. Sci.*, vol. 202, pp. 89–104, Jan. 2022.

[149] N. Gao, S. Jin, X. Li, and M. Matthaiou, "Aerial RIS-assisted high altitude platform communications," *IEEE Wireless Commun. Lett.*, vol. 10, no. 10, pp. 2096–2100, Oct. 2021.

[150] Z. Cheng, M. Liwang, N. Chen, L. Huang, X. Du, and M. Guizani, "Deep reinforcement learning-based joint task and energy offloading in UAV-aided 6G intelligent edge networks," *Comput. Commun.*, vol. 192, pp. 234–244, Aug. 2022.

[151] A. H. Arani, P. Hu, and Y. Zhu, "Fairness-aware link optimization for space-terrestrial integrated networks: A reinforcement learning framework," *IEEE Access*, vol. 9, pp. 77624–77636, 2021.

[152] X. Zhang, S. Xia, X. Li, and T. Zhang, "Multi-objective particle swarm optimization with multi-mode collaboration based on reinforcement learning for path planning of unmanned air vehicles," *Knowl.-Based Syst.*, vol. 250, Aug. 2022, Art. no. 109075.

[153] A. H. Arani, P. Hu, and Y. Zhu, "Re-envisioning space-air-ground integrated networks: Reinforcement learning for link optimization," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2021, pp. 1–7.

[154] S. Yoo and W. Lee, "Federated reinforcement learning based AANs with LEO satellites and UAVs," *Sensors*, vol. 21, no. 23, p. 8111, Dec. 2021.

[155] M. Ndong, M. Hayajneh, N. A. Ali, and S. Alkobaisi, "Towards a 3-tiered space-air-ground network with reinforcement learning," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 9, pp. 7001–7013, Oct. 2022.

[156] M. Guan, Z. Wu, Y. Cui, X. Cao, L. Wang, J. Ye, and B. Peng, "An intelligent wireless channel allocation in HAPS 5G communication system based on reinforcement learning," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 1, pp. 1–9, Dec. 2019.

[157] Y. Nie, J. Zhao, J. Liu, J. Jiang, and R. Ding, "Energy-efficient UAV trajectory design for backscatter communication: A deep reinforcement learning approach," *China Commun.*, vol. 17, no. 10, pp. 129–141, 2020.

[158] F. Zhang, Y. Li, J. Wang, and T. Q. S. Quek, "Learning-based FEC for non-terrestrial networks with delayed feedback," *IEEE Commun. Lett.*, vol. 26, no. 2, pp. 306–310, Feb. 2022.

[159] E. Bulut and I. Guevenc, "Trajectory optimization for cellular-connected UAVs with disconnectivity constraint," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, May 2018, pp. 1–6.

[160] M. M. Azari, F. Rosas, and S. Pollin, "Cellular connectivity for UAVs: Network modeling, performance analysis, and design guidelines," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3366–3381, Apr. 2019.

[161] E. Teng, J. D. Falcao, and B. Iannucci, "Holes-in-the-sky: A field study on cellular-connected UAS," in *Proc. Int. Conf. Unmanned Aircr. Syst. (ICUAS)*, Jun. 2017, pp. 1165–1174.

[162] Y. Zeng and X. Xu, "Path design for cellular-connected UAV with reinforcement learning," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[163] Y. Zeng, X. Xu, S. Jin, and R. Zhang, "Simultaneous navigation and radio mapping for cellular-connected UAV with deep reinforcement learning," 2020, *arXiv:2003.07574*.

[164] J. Zhao, L. Yu, K. Cai, Y. Zhu, and Z. Han, "RIS-aided ground-aerial NOMA communications: A distributionally robust DRL approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 4, pp. 1287–1301, Apr. 2022.

[165] A. Fakhreddine, C. Bettstetter, S. Hayat, R. Muzaffar, and D. Emini, "Handover challenges for cellular-connected drones," in *Proc. 5th Workshop Micro Aerial Vehicle Netw., Syst., Appl.*, Jun. 2019, pp. 9–14.

[166] Y. Chen, X. Lin, T. Khan, and M. Mozaffari, "Efficient drone mobility support using reinforcement learning," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, May 2020, pp. 1–6.

[167] Y. Chen, X. Lin, T. A. Khan, and M. Mozaffari, "A deep reinforcement learning approach to efficient drone mobility support," 2020, *arXiv:2005.05229*.

[168] M. M. Azari, A. H. Arani, and F. Rosas, "Mobile cellular-connected UAVs: Reinforcement learning for sky limits," 2020, *arXiv:2009.09815*.

[169] Z. Feng, L. Ji, Q. Zhang, and W. Li, "Spectrum management for mmWave enabled UAV swarm networks: Challenges and opportunities," *IEEE Commun. Mag.*, vol. 57, no. 1, pp. 146–153, Jan. 2018.

[170] E. D. Re, R. Fantacci, and G. Giambene, "Efficient dynamic channel allocation techniques with handover queuing for mobile satellite networks," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 2, pp. 397–405, Feb. 1995.

[171] X. Zhou, Y. Lin, Y. Tu, S. Mao, and Z. Dou, "Dynamic channel allocation for multi-UAVs: A deep reinforcement learning approach," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[172] Y. Lin, M. Wang, X. Zhou, G. Ding, and S. Mao, "Dynamic spectrum interaction of UAV flight formation communication with priority: A deep reinforcement learning approach," *IEEE Trans. Cognit. Commun. Netw.*, vol. 6, no. 3, pp. 892–903, Sep. 2020.

[173] P. Luong, F. Gagnon, L.-N. Tran, and F. Labeau, "Deep reinforcement learning-based resource allocation in cooperative UAV-assisted wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 11, pp. 7610–7625, Nov. 2021.

[174] J. Wang, R. Han, L. Bai, T. Zhang, J. Liu, and J. Choi, "Coordinated beamforming for UAV-aided millimeter-wave communications using GPML-based channel estimation," *IEEE Trans. Cognit. Commun. Netw.*, vol. 7, no. 1, pp. 100–109, Mar. 2021.

[175] B. Van Der Bergh, A. Chiumento, and S. Pollin, "LTE in the sky: Trading off propagation benefits with interference costs for aerial nodes," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 44–50, May 2016.

[176] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Unmanned aerial vehicle with underlaid device-to-device communications: Performance and tradeoffs," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 3949–3963, Jun. 2016.

[177] U. Challita, W. Saad, and C. Bettstetter, "Deep reinforcement learning for interference-aware path planning of cellular-connected UAVs," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–7.

[178] U. Challita, W. Saad, and C. Bettstetter, "Interference management for cellular-connected UAVs: A deep reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2125–2140, Apr. 2019.

[179] E. Fonseca, B. Galkin, R. Amer, L. A. DaSilva, and I. Dusparic, "Adaptive height optimisation for cellular-connected UAVs using reinforcement learning," 2020, *arXiv:2007.13695*.

[180] F. Ghavimi and R. Jantti, "Energy-efficient UAV communications with interference management: Deep learning framework," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshops (WCNCW)*, Apr. 2020, pp. 1–6.

[181] A. Al-Hilo, M. Samir, C. Assi, S. Sharafeddine, and D. Ebrahimi, "A cooperative approach for content caching and delivery in UAV-assisted vehicular networks," *Veh. Commun.*, vol. 32, Dec. 2021, Art. no. 100391.

[182] Z. Hu, X. Gao, K. Wan, Y. Zhai, and Q. Wang, "Relevant experience learning: A deep reinforcement learning method for UAV autonomous motion planning in complex unknown environments," *Chin. J. Aeronaut.*, vol. 34, no. 12, pp. 187–204, Dec. 2021.

[183] K. Arndt, M. Hazara, A. Ghadirzadeh, and V. Kyrki, "Meta reinforcement learning for sim-to-real domain adaptation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 2725–2731.

[184] T. Carr, M. Chli, and G. Vogiatzis, "Domain adaptation for reinforcement learning on the Atari," in *Proc. 18th Int. Conf. Auto. Agents MultiAgent Syst.*, 2019, pp. 1859–1861.

[185] J. Chen, X. Wu, L. Duan, and S. Gao, "Domain adversarial reinforcement learning for partial domain adaptation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 539–553, Feb. 2022.

[186] S. James, P. Wohlhart, M. Kalakrishnan, D. Kalashnikov, A. Irpan, J. Ibarz, S. Levine, R. Hadsell, and K. Bousmalis, "Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12627–12637.

[187] R. Muzaffar, C. Raffelsberger, A. Fakhreddine, J. L. Luque, D. Emini, and C. Bettstetter, "First experiments with a 5G-connected drone," in *Proc. 6th ACM Workshop Micro Aerial Vehicle Netw., Syst., Appl.*, Jun. 2020, pp. 1–5.

[188] S. Hayat, C. Bettstetter, A. Fakhreddine, R. Muzaffar, and D. Emini, "An experimental evaluation of LTE–A throughput for drones," in *Proc. 5th Workshop Micro Aerial Vehicle Netw., Syst., Appl.*, Jun. 2019, pp. 3–8.

[189] S. A. H. Mohsan, M. A. Khan, F. Noor, I. Ullah, and M. H. Alsharif, "Towards the unmanned aerial vehicles (UAVs): A comprehensive review," *Drones*, vol. 6, no. 6, p. 147, Jun. 2022.

[190] S. H. Alsamhi, A. V. Shvetsov, S. Kumar, S. V. Shvetsova, M. A. Alhartomi, A. Hawbani, N. S. Rajput, S. Srivastava, A. Saif, and V. O. Nyangaresi, "UAV computing-assisted search and rescue mission framework for disaster and harsh environment mitigation," *Drones*, vol. 6, no. 7, p. 154, Jun. 2022.

[191] O. B. Yahia, E. Erdogan, G. K. Kurt, I. Altunbas, and H. Yanikomeroglu, "A weather-dependent hybrid RF/FSO satellite communication for improved power efficiency," *IEEE Wireless Commun. Lett.*, vol. 11, no. 3, pp. 573–577, Mar. 2022.

[192] O. B. Yahia, E. Erdogan, G. K. Kurt, I. Altunbas, and H. Yanikomeroglu, "A weather-dependent hybrid RF/FSO satellite communication for improved power efficiency," *IEEE Wireless Commun. Lett.*, vol. 11, no. 3, pp. 573–577, Mar. 2022.

[193] J.-H. Lee, K.-H. Park, Y.-C. Ko, and M.-S. Alouini, "Spectral-efficient network design for high-altitude platform station networks with mixed RF/FSO system," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 7072–7087, Sep. 2022.

[194] Y. Wang, M. Chen, Z. Yang, W. Saad, T. Luo, S. Cui, and H. V. Poor, "Meta-reinforcement learning for reliable communication in THz/VLC wireless VR networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 7778–7793, Sep. 2022.

[195] C. Chaccour, M. N. Soorki, W. Saad, M. Bennis, and P. Popovski, "Can terahertz provide high-rate reliable low-latency communications for wireless VR?" *IEEE Internet Things J.*, vol. 9, no. 12, pp. 9712–9729, Jun. 2022.

[196] S. S. Hassan, Y. M. Park, Y. K. Tun, W. Saad, Z. Han, and C. S. Hong, "3TO: THz-enabled throughput and trajectory optimization of UAVs in 6G networks by proximal policy optimization deep reinforcement learning," 2022, *arXiv:2202.02924*.

[197] L. H. Lee, C. B. Fernandez, A. Alhilal, T. Braud, S. Hosio, P. Hui, and E. H. Henrieke Anne de Haas, "Beyond the blue sky of multimodal interaction: A centennial vision of interplanetary virtual spaces in turn-based metaverse," 2022, *arXiv:2208.05517*.

[198] B. Uragun, "Energy efficiency for unmanned aerial vehicles," in *Proc. Int. Conf. Mach. Learn. Appl. Workshops*, vol. 2, Dec. 2011, pp. 316–320.

[199] S. Morton, R. D'Sa, and N. Papanikolopoulos, "Solar powered UAV: Design and experiments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2015, pp. 2460–2466.

[200] P. Oettershagen, A. Melzer, T. Mantel, K. Rudin, T. Stastny, B. Wawrzacz, T. Hinzmann, K. Alexis, and R. Siegwart, "Perpetual flight with a small solar-powered UAV: Flight results, performance analysis and model validation," in *Proc. IEEE Aerosp. Conf.*, Mar. 2016, pp. 1–8.

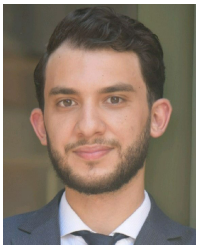
[201] J. Wu, H. Wang, N. Li, P. Yao, Y. Huang, and H. Yang, "Path planning for solar-powered UAV in urban environment," *Neurocomputing*, vol. 275, pp. 2055–2065, Jan. 2018.

[202] B. Ji, Y. Li, B. Zhou, C. Li, K. Song, and H. Wen, "Performance analysis of UAV relay assisted IoT communication network enhanced with energy harvesting," *IEEE Access*, vol. 7, pp. 38738–38747, 2019.

[203] X. Wang and M. C. Gursoy, "Coverage analysis for energy-harvesting UAV-assisted mmWave cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 12, pp. 2832–2850, Dec. 2019.

[204] Z. Yang, W. Xu, and M. Shikh-Bahaei, "Energy efficient UAV communication with energy harvesting," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 1913–1927, Feb. 2020.

[205] I. F. Akyildiz, A. Kak, and S. Nie, "6G and beyond: The future of wireless communications systems," *IEEE Access*, vol. 8, pp. 133995–134030, 2020.
[206] S. Saafi, O. Vikhrova, G. Fodor, J. Hosek, and S. Andreev, "AI-aided integrated terrestrial and non-terrestrial 6G solutions for sustainable maritime networking," 2022, *arXiv:2201.06947*.
[207] N. Bahadori, M. Nabil, and A. Homaifar, "Antenna beamwidth optimization in directional device-to-device communication using multi-agent deep reinforcement learning," *IEEE Access*, vol. 9, pp. 110601–110613, 2021.
[208] I. Romdhane and G. Kaddoum, "A reinforcement-learning-based beam adaptation for underwater optical wireless communications," *IEEE Internet Things J.*, vol. 9, no. 20, pp. 20270–20281, Oct. 2022.
[209] Q. Liu, L. Shi, L. Sun, J. Li, M. Ding, and F. S. Shu, "Path planning for UAV-mounted mobile edge computing with deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 69, no. 5, pp. 5723–5728, May 2020.
[210] A. Asheralieva and D. Niyato, "Hierarchical game-theoretic and reinforcement learning framework for computational offloading in UAV-enabled mobile edge computing networks with multiple service providers," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8753–8769, Oct. 2019.
[211] S. Gu, X. Chen, W. Zeng, and X. Wang, "A deep learning tennis ball collection robot and the implementation on NVIDIA Jetson TX1 board," in *Proc. IEEE/ASME Int. Conf. Adv. Intell. Mechatronics (AIM)*, Jul. 2018, pp. 170–175.
[212] S. Cass, "Nvidia makes it easy to embed AI: The Jetson nano packs a lot of machine-learning power into DIY projects–[hands on]," *IEEE Spectr.*, vol. 57, no. 7, pp. 14–16, Jul. 2020.
[213] J. Hochstetler, R. Padidela, Q. Chen, Q. Yang, and S. Fu, "Embedded deep learning for vehicular edge computing," in *Proc. IEEE/ACM Symp. Edge Comput. (SEC)*, Oct. 2018, pp. 341–343.
[214] B. Brik, A. Ksentini, and M. Bouaziz, "Federated learning for UAVs-enabled wireless networks: Use cases, challenges, and open problems," *IEEE Access*, vol. 8, pp. 53841–53849, 2020.
[215] Y. Qu, C. Dong, J. Zheng, H. Dai, F. Wu, S. Guo, and A. Anpalagan, "Empowering edge intelligence by air-ground integrated federated learning," *IEEE Netw.*, vol. 35, no. 5, pp. 34–41, Sep. 2021.
[216] N. N. Ei, S. W. Kang, M. Alsenwi, Y. K. Tun, and C. S. Hong, "Multi-UAV-assisted MEC system: Joint association and resource management framework," in *Proc. Int. Conf. Inf. Netw. (ICOIN)*, Jan. 2021, pp. 213–218.
[217] N. N. Ei, M. Alsenwi, Y. K. Tun, Z. Han, and C. S. Hong, "Energy-efficient resource allocation in multi-UAV-assisted two-stage edge computing for beyond 5G networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 16421–16432, Sep. 2022.
[218] E. Yanmaz, S. Yahyanejad, B. Rinner, H. Hellwagner, and C. Bettstetter, "Drone networks: Communications, coordination, and sensing," *Ad Hoc Netw.*, vol. 68, pp. 1–15, Jan. 2018.
[219] S. Sthapit, S. Lakshminarayana, L. He, G. Epiphaniou, and C. Maple, "Reinforcement learning for security-aware computation offloading in satellite networks," *IEEE Internet Things J.*, vol. 9, no. 14, pp. 12351–12363, Jul. 2022.
[220] C. Li, X. Sun, and Z. Zhang, "Effective methods and performance analysis of a satellite network security mechanism based on blockchain technology," *IEEE Access*, vol. 9, pp. 113558–113565, 2021.
[221] I. Ahmad, J. Suomalainen, P. Porambage, A. Gurtov, J. Huusko, and M. Höyhtyä, "Security of satellite-terrestrial communications: Challenges and potential solutions," *IEEE Access*, vol. 10, pp. 96038–96052, 2022, doi: 10.1109/ACCESS.2022.3205426.



**TAREK NAOUS** received the B.E. degree in communications and electronics engineering from Beirut Arab University, Lebanon, in 2020, and the M.E. degree in electrical and computer engineering from the American University of Beirut, Lebanon, in 2022. He is currently pursuing the Ph.D. degree in machine learning with the Georgia Institute of Technology, Atlanta, USA. He worked on applied machine learning in wireless communication technology and healthcare at AUB. His research interests include machine learning, natural language processing, multilingual learning, neural text generation and decoding, and clustering algorithms.

**MAY ITANI** received the B.E. degree in electrical engineering and the M.E. degree in computer and communications engineering from the American University of Beirut (AUB), Lebanon, in 1999 and 2004, respectively, and the Ph.D. degree in computer science from Beirut Arab University, in 2017. She worked as a part-time Instructor at Lebanese American University and AUB, from 2012 to 2019. She joined Beirut Arab University as a Full Timer, in September 2020. She is currently an Assistant Professor in computer science at Beirut Arab University. She is also an Assistant Professor at the Mathematics and Computer Science Department, Faculty of Science. She was a recipient of the Charli Korban Award for Outstanding Graduate Student in the Maroun Semaan Faculty of Engineering and Architecture, AUB.

**MARIETTE AWAD** (Member, IEEE) is currently an Associate Professor with the Electrical and Computer Engineering Department, American University of Beirut (AUB) and the Director of the Artificial Intelligence, Data Science and Computing Hub at AUB. She has coauthored a book *Efficient Machine Learning*, in 2015, that is among the top five downloaded open access books according to Springer Nature, in June 2020. She has published in numerous conferences and journals and managed few multimillion grants. Her current research interests include machine learning, data analytics, and the Internet of Things. In 2009, she created the IEEE Women in Engineering Lebanon Chapter and she is a title IX Deputy at AUB. Since 2017, she has been the Organizing Committee for the Stanford Women in Data Science Conference at AUB. She is a reviewer of many conferences and IEEE journals. Prior to her academic position, she was with the IBM System and Technology Group, VT, USA, as a Wireless Product Engineer, where she earned her management recognition, several business awards, and multiple patents. For more information visit the link (mariette.awad@aub.edu.lb).

**SANAA SHARAFEDDINE** (Senior Member, IEEE) received the B.E. and M.E. degrees in computer and communications engineering from the American University of Beirut and the Ph.D. degree in communications engineering from the Munich University of Technology (TUM). She is currently a Professor in computer science at the American University of Beirut. Her Ph.D. research work was conducted in collaboration with the Siemens AG Research Laboratories, Munich, and resulted in two granted patents. She was selected as L'Oreal-Unesco International Rising Talent, in 2015, and received L'Oreal-Unesco Pan Arab Regional Fellowship Award, in 2013. She was a recipient of Siemens Youth and Knowledge Scholarship for Highly Distinguished Students and Charli Korban Award for Outstanding Graduate Student in the Maroun Semaan Faculty of Engineering and Architecture, American University of Beirut. She has served on the Editorial Board of IEEE Networking Letters, *IEEE Internet of Things Magazine*, *Ad Hoc Networks* (Elsevier), IEEE Access, and IEEE Journal on Selected Areas in Communications (JSAC)—Series on Network Softwarization and Enablers. She serves on the Computing Accreditation Commission (CAC) of ABET as a Commissioner and the Team Chair. She was with the Lebanese American University (LAU) as a Professor in computer science, the Vice Chair of the IEEE Computer Society Lebanon Chapter, between 2009 and 2011, the Director of the Software Institute, LAU, between 2014 and 2017, and as a technical program committee member for various international conferences.