Hierarchical Deep Reinforcement Learning for Adaptive Resource Management in Integrated Terrestrial and Non-Terrestrial Networks

Muhammad Ahmed Mohsin[1], Hassan Rizwan[2], Muhammad Umer[3], Sagnik Bhattacharya[1], Ahsan Bilal[4], John M. Cioffi[1]

[1]Stanford University, 450 Jane Stanford Way, Stanford, CA 94305-2004, USA
[2]University of California, Riverside, 900 University Ave, Riverside, CA 92521, USA
[3]National University of Sciences and Technology, Sector H-12, Islamabad, Pakistan
[4]University of Oklahoma, 660 Parrington Oval, Norman, OK 73019, USA
muahmed@stanford.edu, hrizwan@email.com, mumer.bee20seecs@seecs.edu.pk, sagnikb@stanford.edu, ahsan.bilal-1@ou.edu, cioffi@stanford.edu

Abstract

Efficient spectrum allocation has become crucial as the surge in wireless-connected devices demands seamless support for more users and applications, a trend expected to grow with 6G. Innovations in satellite technologies such as SpaceX's Starlink have enabled non-terrestrial networks (NTNs) to work alongside terrestrial networks (TNs) and allocate spectrum based on regional demands. Existing spectrum sharing approaches in TNs use machine learning for interference minimization through power allocation and spectrum sensing, but the unique characteristics of NTNs like varying orbital dynamics and coverage patterns require more sophisticated coordination mechanisms. The proposed work uses a hierarchical deep reinforcement learning (HDRL) approach for efficient spectrum allocation across TN-NTN networks. DRL agents are present at each TN-NTN hierarchy that dynamically learn and allocate spectrum based on regional trends. This framework is 50x faster than the exhaustive search algorithm while achieving 95% of optimum spectral efficiency. Moreover, it is 3.75x faster than multi-agent DRL, which is commonly used for spectrum sharing, and has a 12% higher overall average throughput.

1 Introduction

The increasingly complex ecosystem of wireless networks—from cellular systems to satellite constellations, Internet of things (IoT) devices to vehicular networks—operating within a constrained spectrum band necessitates efficient spectrum sharing. Spectrum sharing refers to the dynamic allocation and reuse of radio frequency bands among multiple users, particularly in complex interference channel (IC) scenarios where multiple transmitters and receivers operate simultaneously. Efficient spectrum sharing reduces co-channel interference and thereby maintains a higher signal-to-interference-plus-noise ratio (SINR) and preserves reliable network throughput for communication. Technological breakthroughs in satellite constellation deployments, led by innovators such as SpaceX's Starlink, Amazon's Project Kuiper, and OneWeb, transform wireless communication landscapes. These non-terrestrial networks (NTNs) now seamlessly coexist with terrestrial networks (TNs), establishing intricate, multi-tiered network architectures that operate across diverse altitudinal ranges. Globally, the number of operational satellites is expected to increase from 13,000 to 33,000 by 2030, enabling 20-30% of mobile users to shift to satellite internet services (Wilson 2024). This market is projected to grow from $9 billion in 2023 to $37 billion by 2034 (Zoting 2024). Therefore, developing efficient spectrum sharing mechanisms for integrated TN-NTN has become crucial to ensure resource optimization and seamless coexistence between these diverse network architectures.

Deep reinforcement learning (DRL) emerges as a particularly compelling solution for optimal spectrum sharing due to its ability to adapt to complex wireless environments and learn optimal policies through continuous interaction (Si et al. 2024). Unlike traditional machine learning approaches that rely on static training datasets and struggle to generalize beyond their training distributions, DRL agents can dynamically adjust their spectrum allocation strategies based on real-time network conditions, interference patterns, and quality of service (QoS) (Zhang, Li, and Mu 2024).

Existing spectrum sharing strategies often oversimplify spectrum allocation as they fail to account for the nested hierarchy present in modern network architectures (Patil et al. 2023). These approaches demonstrate efficacy in controlled environments but, their centralized nature introduces a significant overhead in execution time and leads to suboptimal outcomes due to computational bottlenecks and delays in gathering state information (Zhang and Luo 2023). In large-scale deployments with heterogeneous architectures, spectrum sharing faces severe scalability limitations as the complexity of interference management and resource allocation in ICs grows exponentially with expanding network size and user density. Traditional solutions have primarily focused on TNs, employing conventional techniques like power control (Nasir and Guo 2021), interference management (Oyedare et al. 2022), and spectrum sensing (Li et al. 2020) for efficient spectrum usage. However, these approaches fail to address the unique challenges of TN-NTN which require coordinated decision-making across multiple network tiers.

This paper proposes a novel hierarchical deep reinforcement learning (HDRL)-based spectrum allocation scheme for integrated TN-NTN networks. The network architecture is hierarchically decomposed into three distinct sub-networks: satellite, high altitude platforms (HAPs), and a combined layer of unmanned aerial vehicle (UAVs) and terrestrial base stations (TBSs). Each sub-network's agent operates on a different temporal scale with interconnected policies, where higher-level agents guide the behavior of lower-level agents through metacontrol signals. The hierarchical structure ensures that each subsequent network layer operates within the spectrum constraints imposed by its preceding layer, creating a cascaded decision-making framework that maximizes spectrum utilization across the entire network.

- The proposed framework was benchmarked against different algorithms, including exhaustive search, random access, single-agent DRL (SADRL), and multi-agent DRL (MADRL), across three network hierarchies. The framework achieved 95% of the spectral efficiency of exhaustive search while being 50x faster. It also demonstrated 3.75x faster execution than MADRL and yielded 10-18% performance improvements over SADRL and random access methods across all scenarios.
- In a single 500-step episode, the framework achieved 5%, 11%, and 25% higher average throughput compared to MADRL, SADRL, and random access, respectively. The framework maintained superior stability with minimal throughput fluctuations across all steps relative to both MADRL and SADRL.
- The framework's learning progression and convergence behavior were evaluated across different network hierarchies to assess adaptability. Training results over 1000 episodes showed consistent learning progress, with the space-air-ground (SAG) network achieving the highest cumulative reward, followed by the air-ground and UAV-aided networks.

2 Related Work

2.1 Spectrum Sharing

Several previous works focus on interference management for spectrum sharing in satellite-terrestrial networks: (Lee et al. 2021) introduces reverse spectrum pairing for TN-NTN systems, while (Lee et al. 2024) optimizes TN-NTN grouping with earth-fixed satellite beamforming. To accommodate different network architectures, (Wang et al. 2020) develops a cognitive control system for air ground integrated networks (AGIN) spectrum sharing, whereas, (Wang, Ding, and Zhang 2020) and (Zhang et al. 2019) focus on satellite spectrum sharing frameworks—the former for geostationary earth satellites (GEO) and low-orbit earth satellites (LEO) networks using overlay/underlay modes, and the latter for satellite-terrestrial mmWave networks using protection areas. These works allow spectrum sharing via interference mitigation through spectrum sensing and power control mechanisms, however, they overlook the challenge of spectrum allocation across integrated TN-NTN networks where spectrum resources need to be dynamically distributed based on regional demand patterns.

2.2 DRL for Spectrum Management

There are several works exploring DRL approaches for spectrum management: (Song et al. 2021) and (Chen et al. 2022) utilize deep Q-network (DQN) variants for dynamic spectrum access (DSA) coordination, with the latter implementing dueling DQNs and prioritized experience replay for balanced primary user and secondary user performance. (Cui and Yu 2021) introduces a scalable single-agent reinforcement learning (RL) method for joint routing and spectrum optimization in wireless ad-hoc networks. For distributed approaches, (Tan et al. 2022) employs cooperative multi-agent RL with recurrent DQNs for DSA, while (Jo et al. 2022) proposes multi-agent DQL for HAPs power control with interference management. (Han et al. 2022) addresses urban air mobility using DRL-based spectrum allocation between aerial and terrestrial users.

While these approaches demonstrate the effectiveness of utilizing DRL algorithms for spectrum sharing, they primarily focus on TNs and single-layer optimization, neglecting the spectrum management required for NTNs in conjunction with TNs, and the need for hierarchical decision-making across multiple network layers.

3 HDRL Based Intelligent Spectrum Allocation

3.1 System Overview

The considered system comprises a LEO satellite, HAPs, UAVs, TBSs, and users. The LEO satellite covers designated geographical regions using fixed multi-beam technology, where each beam cell serves as a dedicated coverage area for a HAP. Each HAP acts as a regional hub, relaying data and control signals between the satellite and lower-tier network nodes. Within each HAP's coverage area, multiple TBSs and UAVs are deployed. TBSs provide fixed, high-capacity connectivity for ground-based users, while UAVs act as aerial base stations, offering flexible, on-demand coverage. Users dynamically associate with either a TBS or UAV based on factors such as signal strength, network load, and QoS. A satellite gateway facilitates communication between the TN and the LEO satellite while a central control and compute unit manages network operations, acting as the coordinator for spectrum allocation and interference management. The LEO satellite allocates portions of its spectrum $A_{\text{satellite}}$ to each beam cell. Each beam's spectrum is then further divided by the HAPs into subbands A_{HAP} and is shared by both UAVs and TBSs in that coverage region.

3.2 Problem Formulation

The hierarchical spectrum sharing framework is modeled as a Markov decision process (MDP), captures the multi-tiered decision-making process essential for efficient spectrum allocation in a TN-NTN system. The MDP consists of a state space S, an action space A, a state transition function $s' = f(s, a)$, and a reward function $r(s, a)$. DRL agent interacts with the environment over a sequence

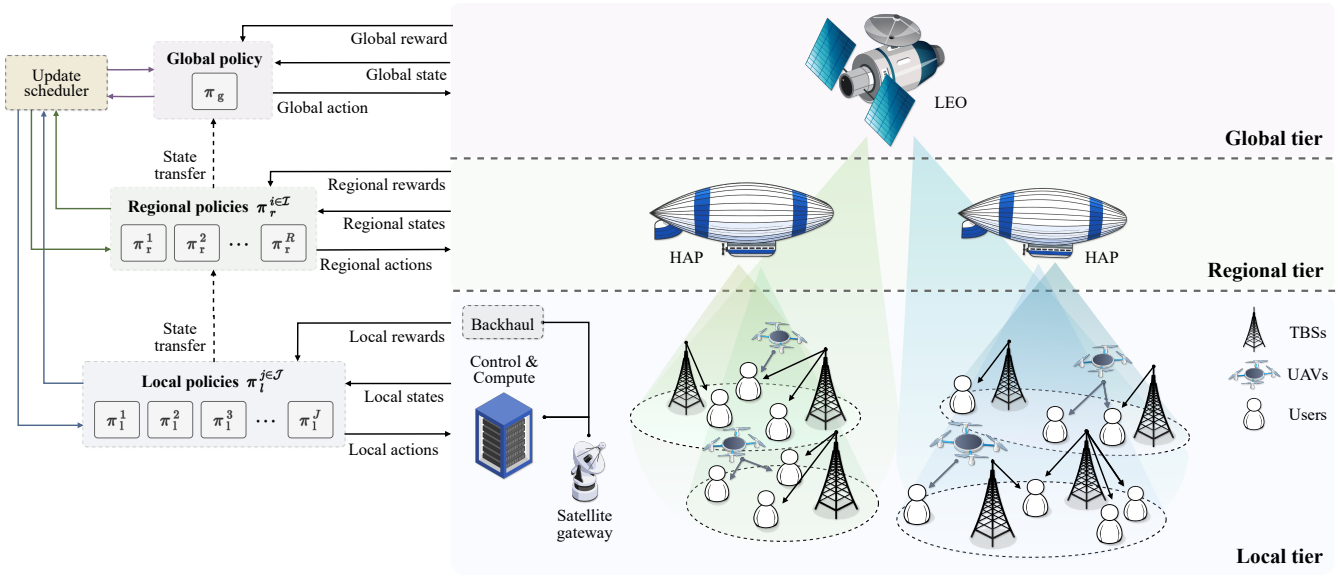


Figure 1: Proposed framework environment for HDRL-based dynamic spectrum sharing in integrated TN-NTN Networks.

of states $s_1, s_2, \ldots, s_t$, actions $a_1, a_2, \ldots, a_t$, and rewards $r_1, r_2, \ldots, r_t$, where s_t, a_t, r_t are state, action and reward at time t and total time steps in the episode are T. The primary objective is to optimize network performance across three hierarchical decision levels—satellite, HAP, and UAV—each responsible for spectrum management within its respective operational scope.

Global Policy. At the top level, the *global policy* π_{g} is managed by the satellite agent, which oversees the entire network spectrum allocation. This agent's goal is to allocate spectrum resources effectively across multiple beam cells, ensuring fair distribution and accommodating varying user demands and channel conditions. The global network state, denoted as

$$S_{\text{g}} = \{A_{\text{spec}}, D_{\text{beam}}, G_{\text{avg}}\},$$

includes aggregated information such as the total available spectrum A_{spec}, the distribution of beams D_{beam} across geographical regions, and the average channel gain G_{avg} across these regions. Given this state S_{g}, the satellite agent determines a spectrum allocation matrix $\mathbf{A}_{\text{g}} \in [0,1]^{B\times N}$, where each element $a_{b,n}$ represents the allocation of subband n to beam b

$$a_{b,n} = \begin{cases} 1 & \text{if subband } n \text{ is allocated to beam } b \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

This allocation must satisfy

$$\sum_{b=1}^{B} a_{b,n} \leq 1, \forall n \in \{1, \ldots, N\}. \tag{2}$$

The selected action $a_{\text{g}} = \mathbf{A}_{\text{g}}$ is constrained by the policy $\pi_{\text{g}}(S_{\text{g}})$, which optimizes the high-level allocation to maximize network performance metrics like spectral efficiency η, system fairness F, and average throughput R_{avg}. This global level allocation subsequently passes down to the regional tier as a constraint on available resources, ensuring that the regional and local policies can operate within these allocations.

Regional Policies. At the intermediate level, each HAP i manages a *regional policy* π_{r}^i, responsible for spectrum allocation within its designated coverage area. Given the global allocation constraints from π_{g}, each HAP agent operates on a regional network state S_{r}^i, which incorporates detailed information relevant to the local context. This state is defined as

$$S_{\text{r}}^i = \{A_{\text{spec}}, D_{\text{region}}, G_{\text{avg}}\},$$

where A_{spec} is the spectrum allocated by the global policy to the HAP's region, D_{region} represents the spatial distribution of users within the HAP's coverage, and G_{avg} indicates average channel conditions in the region.

At this level, each HAP determines a regional spectrum allocation matrix $\mathbf{A}_{\text{r}}^i \in [0,1]^{M\times N}$, where M is the number of subordinate nodes (UAVs and TBSs) in region i. The allocation elements are

$$a_{m,n}^i = \begin{cases} 1 & \text{if subband } n \text{ is allocated to node } m \text{ in region } i, \\ 0 & \text{otherwise,} \end{cases}$$

which are subject to the constraints

$$\sum_{m=1}^{M} a_{m,n}^i \leq 1, \quad \forall n \in \{1, \ldots, N\}, \tag{3}$$

$$a_{m,n}^i \leq a_{b(i),n}, \quad \forall m, n, \tag{4}$$

where $b(i)$ denotes the beam containing region i.

Local Policies. At the lowest level, individual UAVs and TBSs operate under *local policies* π_{l}^j for each node j, making real-time decisions on spectrum access and power allocation for their associated users. The local network state S_{l}^j

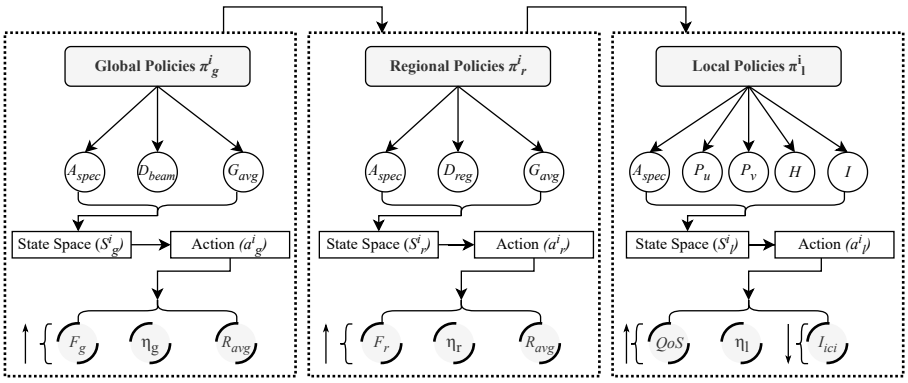


Figure 2: Flow diagram for the proposed hierarchical reinforcement learning for spectrum resource management.

for each UAV or TBS agent j is represented as

$$S_1^j = \{A_{\text{spec}}, P_u, P_v, H, I\},$$

where P_u and P_v indicate the positions of the user equipment (UE) and UAV, respectively, H represents the channel gains, and I denotes interference levels in the immediate vicinity. Based on S_1^j, each local policy π_1^j selects an action a_1^j, which is represented as

$$a_1^j = \{\beta_j, \alpha_j, \Delta p_j\},$$

where the spectrum access vector $\beta_j \in [0,1]^N$, the power allocation vector $\alpha_j \in [0,1]^K$, and the movement vector $\Delta p_j \in [-\Delta p_{\max}, \Delta p_{\max}]^2$. The spectrum access vector β_j indicates which spectrum channels to utilize, allowing the agent to decide the specific channels that are most efficient under current network conditions. The power allocation vector α_j specifies the power levels assigned to each selected channel, ensuring that power resources are optimally distributed to minimize interference while meeting user demands. Lastly, for UAVs, the movement vector Δp_j controls positional adjustments within the UAV's operational range, allowing it to enhance local coverage dynamically by repositioning itself in response to user distribution and channel quality variations. Penalty P_{UAV} is imposed if the UAVs are outside their operational range

$$P_{\text{UAV}} = \frac{\sum_{u=1}^{U} \mathbb{1}(u \notin \mathcal{R})}{U}, \tag{5}$$

where $\mathbb{1}(u \notin \mathcal{R})$ is the indicator function for UAVs outside their designated region, and U is the total number of UAVs per region. This local-level optimization ensures that real-time adjustments to spectrum access and power control are responsive to changes in user demand, channel conditions, and interference levels.

At the regional level, several metrics are evaluated to determine the performance of each region $j \in J$, where J represents the set of all local regions. We utilized appropriate channel distributions for all channel links. Each user's SINR is defined as

$$\gamma_j^u = \frac{H_j^u \alpha_j^u P}{I_j^u + N_0}, \tag{6}$$

where H_j^u represents the channel gain, α_j^u is the power allocation, P is the transmission power, I_j^u denotes interference, and N_0 is the noise power. Using (6), the data rate for each user is

$$R_j^u = \frac{W}{N} \log_2(1 + \gamma_j^u), \tag{7}$$

where W is the total bandwidth available and N is the number of spectrum channels. We compute spectral efficiency in each region as

$$\eta_j = \frac{\sum R_j^u}{W}, \tag{8}$$

where η_j represents the spectral efficiency. To measure fairness, we calculate the Jain's fairness index for each region as

$$F_j = \frac{\left(\sum R_j^u\right)^2}{K \sum (R_j^u)^2}, \tag{9}$$

where K is the number of users in region j. Finally, to monitor QoS adherence, we calculate the QoS violation for each region as

$$V_j = \max(0, R_{\min} - \min(R_j^u)), \tag{10}$$

where $R_{\min}$ is the minimum required rate for QoS compliance. These metrics collectively provide a detailed assessment of each region's performance and are instrumental in formulating the reward function.

3.3 Optimization Objective

The hierarchical structure of the MDP maximizes the cumulative reward over all time steps t formulated as

$$\max_{\pi_g, \pi_r, \pi_l} \mathbb{E}\left[\sum_{t=1}^{T} \left(w_1 R_{\text{avg}} + w_2 \eta + w_3 F + w_4 P_{\text{UAV}}\right)\right], \tag{11}$$

where w_1, w_2, w_3, and w_4 are weights for data rate, spectral efficiency, fairness, and, UAV penalty respectively. This reward structure incentivizes each policy level to contribute towards optimal network performance, balancing local needs and global objectives in a dynamically adaptive manner. Through this hierarchical MDP structure, the HDRL framework facilitates efficient spectrum sharing by decomposing the overall optimization problem into manageable subproblems, each tailored to the operational scope and constraints of the respective agent.

Algorithm 1: HDRL for Spectrum Sharing

Initialize neural networks for policies π_{s}, π_{h}, π_{l}
Initialize replay buffers D_{s}, D_{h}, D_{l}
for *episode* $= 1$ *to* M **do**
  Initialize parameters H, I, P_{u}, P_{v}, α
  for *step* $= 1$ *to* S **do**
    if *step* $\% \ \delta_s = 0$ **then**
      Observe (beam) $s_{\text{s}} = \{A_{\text{spec}}, D_{\text{beam}}, G_{\text{avg}}\}$
      Take action $a_{\text{s}} = \pi_{\text{s}}(s_{\text{s}})$ for $A_{\text{beam spec}}$
    if *step* $\% \ \delta_h = 0$ **then**
      for *each HAP* $i = 1$ *to* $B \times H$ **do**
        Observe $s_{\text{h}}^i = \{A_{\text{spec}}, D_{\text{region}}, G_{\text{avg}}\}$
        Take action $a_{\text{h}}^i = \pi_{\text{h}}(s_{\text{h}}^i)$ for A_{spec}
    for *each region* $j = 1$ *to* $B \times H \times R$ **do**
      Observe $s_{\text{l}}^j = \{A_{\text{spec}}, P_{\text{u}}, P_{\text{v}}, H, I\}$
      Take action $a_{\text{l}}^j = \pi_{\text{l}}(s_{\text{l}}^j)$ for β_j, α_j, Δp_j
      Update UAV positions: $P_{\text{v}} = P_{\text{v}} + \Delta p$
    for *each region* j **do**
      Calculate local metrics $\{\gamma_j^u, R_j^u, \eta_j, F_j, V_j\}$
    Compute rewards: r_{s}, r_{h}, r_{l}
    Store $(s_{\text{s}}, a_{\text{s}}, r_{\text{s}})$ in D_{s}
    Store $(s_{\text{h}}, a_{\text{h}}, r_{\text{h}})$ in D_{h}
    Store $(s_{\text{l}}, a_{\text{l}}, r_{\text{l}})$ in D_{l}
    Update π_{s}, π_{h}, π_{l} using samples from buffers
    if *terminated or truncated* **then**
      break
  Calculate episodic metrics R_{avg}, η, F
return π_{s}, π_{h}, π_{l}, R_{avg}, η, F

4 Numerical Results

4.1 Experimental Settings

The experimental framework utilized a hierarchical reinforcement learning environment with a detailed network topology. It featured 2 beams from LEO, each with 1 HAP, subdivided into 2 regions, each with 2 base stations and 1 UAV, supporting 10 users per region. Spectrum allocation spanned 200 MHz across 10 subbands centered at 28 GHz. Node altitudes included satellites at 550 km, HAPs at 20 km, and UAVs at 100 meters. Transmission power ranged from 33–45 dBm for satellites, 28–36 dBm for HAPs, 16 dBm for base stations, and 8 dBm for UAVs. Each 2×2 km region featured UAVs moving in 10-meter steps, with noise power fixed at -174 dBm/Hz.

The reward mechanism used a weighted multi-objective structure: spectrum efficiency (1.5), fairness (0.5), UAV penalty (-1.0), and QoS violations (-0.5). The core algorithm is based on proximal policy optimization (PPO) as a single-agent RL algorithm with a learning rate of 0.0005, mini-batch size of 512, batch size of 2000, 30 stochastic gradient descent iterations, a discount factor of 0.99, entropy coefficient of 0.01, and value function loss coefficient of 1.0.

Table 1: Simulation parameters.

Parameter	Value
Number of beams (B)	2
HAPs per beam (H)	1
Regions per HAP (R)	2
Time blocks per region (T)	2
UAVs per region (U)	1
Users per region (K)	10
Total bandwidth (W)	200 MHz
Number of subbands (N)	10
Maximum episodes (M)	1000
Steps per episode (S)	500
Decision intervals (δ_s, δ_h, δ_l)	(50, 10, 1)

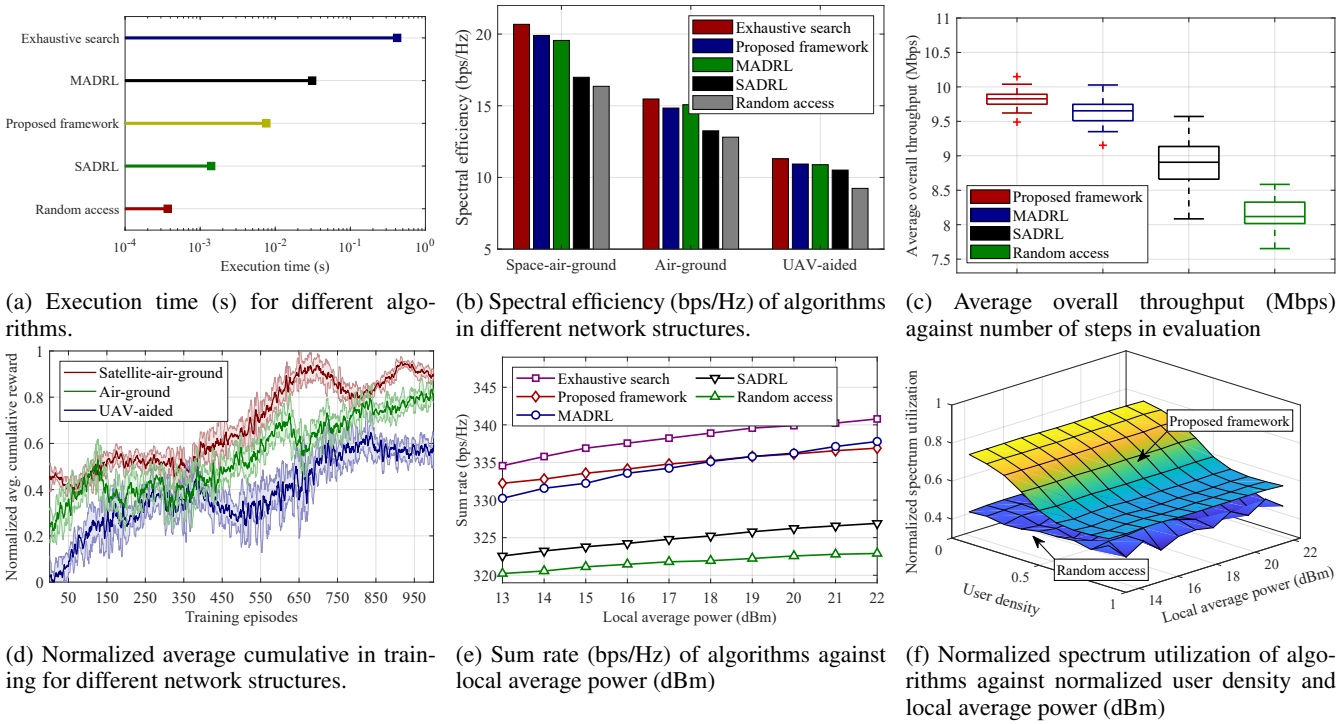
4.2 Baseline Methodologies

Results compare the proposed HDRL framework against SADRL & MADRL algorithms used in spectrum sharing scenarios. A direct comparison with frameworks used in other relevant works, however, is unfeasible as the system model and components are vastly different. Evaluation considers a comprehensive SAG network as the default scenario, unless mentioned otherwise.

1. **Exhaustive Search.** This baseline conducts a full combinatorial search across all possible spectrum allocation configurations to identify the optimal solution. However, the exhaustive search is computationally prohibitive for large-scale networks due to its exponential complexity.
2. **Random Allocation.** Serving as a performance lower bound, this approach assigns spectrum resources randomly without leveraging any intelligent decision-making, offering a baseline for evaluating the gains achieved through reinforcement learning.
3. **SADRL.** PPO is a SADRL algorithm that learns a global policy to optimize network objectives as is commonly used in literature. (Guo et al. 2022) relies on PPO to optimize spectrum sharing in an IRS-aided cognitive radio system, and (Samidi, Radzi, and Aripin 2024) optimizes subcarrier spacing and uplink-downlink allocation with PPO in 5G networks.
4. **MADRL.** Various works formulate MADRL algorithms to enhance spectrum sharing. (Gao et al. 2021) employs MADRL to enable cooperative spectrum sensing among multiple secondary users in cognitive radio networks, enhancing sensing accuracy by sharing detection results and reducing overhead through decentralized execution. (Naderializadeh et al. 2021) proposes a distributed resource management mechanism using MADRL and independent Q-learning to determine user scheduling and power allocation.

4.3 Results

Fig. 3a compares execution times of spectrum allocation algorithms. Exhaustive search has the highest time due to evaluating all possible action-state combinations. MADRL is slower than HDRL due to the exponential growth of



(a) Execution time (s) for different algorithms.

(b) Spectral efficiency (bps/Hz) of algorithms in different network structures.

(c) Average overall throughput (Mbps) against number of steps in evaluation

(d) Normalized average cumulative in training for different network structures.

(e) Sum rate (bps/Hz) of algorithms against local average power (dBm)

(f) Normalized spectrum utilization of algorithms against normalized user density and local average power (dBm)

Figure 3: Results compare the proposed framework with benchmark models for different metrics.

joint state-action spaces and non-stationary learning environments with multiple agents training simultaneously, compared to HDRL's structured task decomposition and transfer learning between hierarchical levels. The proposed HDRL framework achieves a balance, segmenting decision-making hierarchically to improve efficiency without compromising decision quality. The proposed framework is 3.75x faster than MADRL and 50x faster than exhaustive search. SADRL's centralized structure leads to faster execution, while random access, lacking intelligent decision-making, is the fastest. The HDRL framework is 4x and 19x slower than SADRL and random access, respectively.

Fig. 3b compares spectral efficiency across various network scenarios, illustrating the performance of different allocation algorithms. Exhaustive search achieves the highest efficiency possible in all network scenarios and the proposed framework achieves 97% of that optimal spectral efficiency in SAG network, and achieves 2% more than MADRL. This shows that the proposed framework achieves near-optimal performance while maintaining better computational efficiency. In the SAG network scenario, the proposed framework outperforms SADRL and random access by approximately 15% and 18%, respectively. The performance gap between all algorithms narrows in other scenarios due to reduced degrees of freedom and increased interference in simpler architectures. The difference is more pronounced in SAG networks because of a higher system complexity, offering more degrees of freedom to exploit.

Fig. 3c illustrates the throughput performance of different algorithms against varying channel conditions. The proposed framework maintains stable performance across all steps, averaging approximately 9.85 Mbps, while other approaches show more variance in their throughput values. MADRL closely follows with an average throughput of about 9.65 Mbps. The proposed framework achieves achieved 5%, 11% and 25% higher average throughput than MADRL, SADRL, and random access, respectively. SADRL shows lower average throughput due to its limitations in capturing complex network dynamics and has the highest throughput fluctuations at every step. Random access performs worst with the lowest throughput, as expected from its non-intelligent randomized selection of actions.

Fig. 3d demonstrates the learning progression and convergence behavior of the proposed framework across different network architectures over 1000 training episodes. The SAG network achieves the highest normalized average cumulative reward due to its rich multi-layer structure offering more optimization opportunities. The air-ground (AG) network shows moderate performance, with a similar learning pattern but lower overall rewards due to reduced network complexity. The UAV-aided network, being the simplest architecture, exhibits the lowest reward values but shows steady improvement. All scenarios show initial fluctuations during training before stabilizing, with consistent performance gaps.

Fig. 3e illustrates the sum rate of various algorithms relative to the local average power, calculated as the average power across a region. The proposed framework and MADRL perform equally well. Other algorithms display consistent trends, with sum rates gradually increasing as average power rises. Exhaustive search consistently outper-

forms the proposed framework by 5 bps/Hz, equivalent to a 1.5% improvement. Meanwhile, SADRL and random access remain close, within 4 bps/Hz of each other, but lag at least 8 bps/Hz behind the proposed framework.

Fig. 3f is a 3D surface plot illustrating the relationship between normalized spectrum utilization and, user density, and local average power. The color gradient represents utilization, with darker blue indicating lower and yellow-orange denoting higher utilization. The graph demonstrates the adaptability of the proposed framework, revealing that at user density close to 1, spectrum utilization decreases. This reduction is attributed to the larger action space required to accommodate a greater number of users.

5 Ablation Studies

Extensive ablation studies provide deeper insights into HDRL's performance. These evaluate the impact of different reward function formulations, including cumulative reward to capture the total aggregated rewards over the episode trajectory, aggregative reward to normalize the cumulative reward by episode duration and to reduce the influence of varying episode lengths, and difference reward to isolate the unique contribution of individual agents to the overall system performance. Results use rigorous hyperparameter tuning to explore the most impactful parameters such as learning rate, PPO clip parameter, and entropy coefficient. Evaluation tests the robustness of the proposed HDRL solution by examining its performance under varying longitudinal and latitudinal conditions, ensuring its adaptability to diverse network configurations and environments. The ablation studies are extensively covered in Supplementary Materials at the end of this paper.

6 Conclusion

The growing demand for seamless wireless connectivity, further accelerated by 6G adoption, necessitates efficient spectrum allocation to serve diverse user demands. Existing spectrum-sharing solutions predominantly address TNs, overlooking the critical role of NTNs such as satellite constellations from SpaceX, OneWeb, and Amazon. The research introduces an HDRL framework for dynamic spectrum allocation within an integrated TN-NTN infrastructure. By leveraging network nesting, DRL agents at each tier of the network hierarchy coordinate spectrum management based on user demand. This adaptive, multi-layered approach surpasses algorithms like MADRL in efficiency and responsiveness, supporting the diverse connectivity requirements of 6G ecosystems and empowering network operators to manage spectrum resources.

References

Chen, M.; Liu, A.; Liu, W.; Ota, K.; Dong, M.; and N. Xiong, N. 2022. RDRL: A Recurrent Deep Reinforcement Learning Scheme for Dynamic Spectrum Access in Reconfigurable Wireless Networks. *IEEE Transactions on Network Science and Engineering*, 9(2): 364–376.

Cui, W.; and Yu, W. 2021. Scalable Deep Reinforcement Learning for Routing and Spectrum Access in Physical Layer. *IEEE Transactions on Communications*, 69(12): 8200–8213.

Gao, A.; Du, C.; Ng, S. X.; and Liang, W. 2021. A Cooperative Spectrum Sensing With Multi-Agent Reinforcement Learning Approach in Cognitive Radio Networks. *IEEE Communications Letters*, 25(8): 2604–2608.

Guo, J.; Wang, Z.; Li, J.; and Zhang, J. 2022. Deep Reinforcement Learning Based Resource Allocation for Intelligent Reflecting Surface Assisted Dynamic Spectrum Sharing. In *2022 14th International Conference on Wireless Communications and Signal Processing (WCSP)*, 1178–1183.

Han, R.; Li, H.; Knoblock, E. J.; Gasper, M. R.; and Apaza, R. D. 2022. Dynamic Spectrum Sharing in Cellular Based Urban Air Mobility via Deep Reinforcement Learning. In *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, 1332–1337.

Jo, S.; Yang, W.; Choi, H. K.; Noh, E.; Jo, H.-S.; and Park, J. 2022. Deep Q-Learning-Based Transmission Power Control of a High Altitude Platform Station with Spectrum Sharing. *Sensors*, 22(4).

Lee, H.-W.; Chen, C.-C.; Liao, C.-I. S.; Medles, A.; Lin, D.; Fu, I.-K.; and Wei, H.-Y. 2024. Interference Mitigation for Reverse Spectrum Sharing in B5G/6G Satellite-Terrestrial Networks. *IEEE Transactions on Vehicular Technology*, 73(3): 4247–4263.

Lee, H.-W.; Medles, A.; Jie, V.; Lin, D.; Zhu, X.; Fu, I.-K.; and Wei, H.-Y. 2021. Reverse Spectrum Allocation for Spectrum Sharing between TN and NTN. In *2021 IEEE Conference on Standards for Communications and Networking (CSCN)*, 1–6.

Li, Y.; Zhang, W.; Wang, C.-X.; Sun, J.; and Liu, Y. 2020. Deep Reinforcement Learning for Dynamic Spectrum Sensing and Aggregation in Multi-Channel Wireless Networks. *IEEE Transactions on Cognitive Communications and Networking*, 6(2): 464–475.

Naderializadeh, N.; Sydir, J. J.; Simsek, M.; and Nikopour, H. 2021. Resource Management in Wireless Networks via Multi-Agent Deep Reinforcement Learning. *IEEE Transactions on Wireless Communications*, 20(6): 3507–3523.

Nasir, Y. S.; and Guo, D. 2021. Deep Reinforcement Learning for Joint Spectrum and Power Allocation in Cellular Networks. In *2021 IEEE Globecom Workshops (GC Wkshps)*, 1–6.

Oyedare, T.; Shah, V. K.; Jakubisin, D. J.; and Reed, J. H. 2022. Interference Suppression Using Deep Learning: Current Approaches and Open Challenges. *IEEE Access*, 10: 66238–66266.

Patil, A.; Iyer, S.; López, O. L. A.; Pandya, R. J.; Pai, K.; Kalla, A.; and Kallimani, R. 2023. A Comprehensive Survey on Spectrum Sharing Techniques for 5G/B5G Intelligent Wireless Networks: Opportunities, Challenges and Future Research Directions. *arXiv preprint arXiv:2308.11716*.

Samidi, F.; Radzi, N.; and Aripin, N. 2024. Reinforcement Learning Model Selection for Resource Allocation and Subcarrier Spacing Optimization in 5G Sliced Spectrum Net-

works. In *2024 IEEE International Conference on Applied Electronics and Engineering (ICAEE)*, 1–6.

Si, J.; Huang, R.; Li, Z.; Hu, H.; Jin, Y.; Cheng, J.; and Al-Dhahir, N. 2024. When Spectrum Sharing in Cognitive Networks Meets Deep Reinforcement Learning: Architecture, Fundamentals, and Challenges. *IEEE Network*, 38(1): 187–195.

Song, H.; Liu, L.; Ashdown, J.; and Yi, Y. 2021. A Deep Reinforcement Learning Framework for Spectrum Management in Dynamic Spectrum Access. *IEEE Internet of Things Journal*, 8(14): 11208–11218.

Tan, X.; Zhou, L.; Wang, H.; Sun, Y.; Zhao, H.; Seet, B.-C.; Wei, J.; and Leung, V. C. M. 2022. Cooperative Multi-Agent Reinforcement-Learning-Based Distributed Dynamic Spectrum Access in Cognitive Radio Networks. *IEEE Internet of Things Journal*, 9(19): 19477–19488.

Wang, H.; Wang, J.; Ding, G.; Xue, Z.; Zhang, L.; and Xu, Y. 2020. Robust Spectrum Sharing in Air-Ground Integrated Networks: Opportunities and Challenges. *IEEE Wireless Communications*, 27(3): 148–155.

Wang, Y.; Ding, X.; and Zhang, G. 2020. A Novel Dynamic Spectrum-Sharing Method for GEO and LEO Satellite Networks. *IEEE Access*, 8: 147895–147906.

Wilson, D. 2024. Revealed: Number of operational satellites in orbit, 2024. *CEOWORLD Magazine*.

Zhang, C.; Jiang, C.; Kuang, L.; Jin, J.; He, Y.; and Han, Z. 2019. Spatial Spectrum Sharing for Satellite and Terrestrial Communication Networks. *IEEE Transactions on Aerospace and Electronic Systems*, 55(3): 1075–1089.

Zhang, Y.; Li, J.; and Mu, G. e. a. 2024. A DRL-based resource allocation for IRS-enhanced semantic spectrum sharing networks. *EURASIP J. Adv. Signal Process.*, 69.

Zhang, Y.; and Luo, Z. 2023. A Review of Research on Spectrum Sensing Based on Deep Learning. *Electronics*, 12(21): 4514.

Zoting, S. 2024. Satellite Internet Market Size to Hit USD 37.64 Bn by 2034. *Precedence Research*.