Intelligent Spectrum Sharing in Integrated TN-NTNs: A Hierarchical Deep Reinforcement Learning Approach

Muhammad Umer, Muhammad Ahmed Mohsin, Ali Arshad Nasir, Hatem Abou-Zeid and Syed Ali Hassan

Abstract—**Integrating non-terrestrial networks (NTNs) with terrestrial networks (TNs) is key to enhancing coverage, capacity, and reliability in future wireless communications. However, the multi-tier, heterogeneous architecture of these integrated TN-NTNs introduces complex challenges in spectrum sharing and interference management. Conventional optimization approaches struggle to handle the high-dimensional decision space and dynamic nature of these networks. This paper proposes a novel hierarchical deep reinforcement learning (HDRL) framework to address these challenges and enable intelligent spectrum sharing. The proposed framework leverages the inherent hierarchy of the network, with separate policies for each tier, to learn and optimize spectrum allocation decisions at different timescales and levels of abstraction. By decomposing the complex spectrum sharing problem into manageable sub-tasks and allowing for efficient coordination among the tiers, the HDRL approach offers a scalable and adaptive solution for spectrum management in future TN-NTNs. Simulation results demonstrate the superior performance of the proposed framework compared to traditional approaches, highlighting its potential to enhance spectral efficiency and network capacity in dynamic, multi-tier environments.**

I. Introduction

The exponential growth of wireless data traffic and the emergence of new use cases have driven the need for integrating non-terrestrial networks (NTNs) with terrestrial networks (TNs) to meet the demanding requirements of future communications. NTNs, such as satellites, high-altitude platforms (HAPs), and unmanned aerial vehicles (UAVs), offer unique advantages in terms of coverage extension, service continuity, and rapid deployment. However, the increasing demand for wireless services has led to a scarcity of available spectrum resources. Spectrum sharing between NTNs and TNs has emerged as a promising solution to address this issue, enabling the efficient utilization of limited spectrum while accommodating the diverse requirements of future wireless networks.

The Third Generation Partnership Project (3GPP) has been actively working on incorporating NTN components into the 5G New Radio (NR) architecture, with ongoing studies and specifications in Release 17 and beyond [1], [2]. The integration of NTNs with TNs introduces significant challenges in spectrum sharing and interference management due to the multi-tier, heterogeneous nature of the resulting network. The coexistence of diverse network elements with different characteristics, such as altitude, mobility, and transmission power, leads to a complex and dynamic interference landscape. Moreover, the high-dimensional decision space arising from the joint allocation of resources across multiple tiers renders conventional optimization approaches impractical for real-time spectrum management [3].

Recent work has explored various aspects of spectrum sharing in integrated TN-NTNs, including cognitive radio techniques [4], dynamic spectrum access [5], and interference mitigation schemes. However, these approaches often rely on simplifying assumptions and struggle to adapt to rapidly changing network conditions. Machine learning, particularly deep reinforcement learning (DRL), has emerged as a promising tool for complex resource management problems in wireless networks.

Several studies have applied DRL to spectrum sharing and resource allocation in NTNs, demonstrating its effectiveness in optimizing spectrum access and power control in cognitive satellite-terrestrial networks [6], coordinating multiple HAPs for enhanced coverage [7], and managing interference between UAVs and ground users [8]. However, most existing DRL-based solutions focus on a single network tier or limited resource types, failing to capture the full complexity of the integrated TN-NTN environment. Cao et al. [1] proposed a multi-tier DRL approach for spectrum sharing in NTN networks, using separate policies for different network tiers. While this improves upon single-agent DRL by considering each tier's unique characteristics, it lacks an explicit hierarchical structure and the ability to handle multiple agents within each tier, which is crucial for efficient coordination and scalability in large-scale networks.

To address these limitations, we propose a novel hierarchical DRL (HDRL) framework for intelligent spectrum sharing in integrated TN-NTNs. Our approach mirrors the network's hierarchy to decompose the complex spectrum sharing problem into manageable sub-tasks, with meta-controllers guiding the learning of sub-controllers at each tier. By incorporating multiple policies within each tier and enabling coordination across tiers, the proposed framework offers a scalable and adaptive solution for dynamic spectrum management in integrated TN-NTNs.

The remainder of this paper is organized as follows. The following section provides an overview of spectrum sharing challenges and opportunities in integrated TN-NTNs. We then introduce hierarchical DRL and its application to wireless

Muhammad Umer and Syed Ali Hassan are with the National University of Sciences and Technology (NUST), Pakistan.

Muhammad Ahmed Mohsin is with Stanford University, USA

Hatem Abou-Zeid is with University of Calgary, Canada.

Ali Arshad Nasir is with King Fahd University of Petroleum and Minerals (KFUPM), Saudia Arabia.

networks. Next, we present the proposed HDRL framework for intelligent spectrum sharing, detailing the system model, learning architecture, and optimization objectives. Subsequently, we present simulation results and performance analysis, comparing our approach with existing benchmarks. Finally, we conclude and discuss future research directions.

II. Spectrum Sharing in Integrated TN-NTNs

In this section, we provide an overview of established strategies for spectrum sharing in integrated TN-NTNs. We also discuss recent advances in these strategies and highlight the trend towards AI-driven spectrum sharing. Finally, we identify some common limitations of existing approaches to motivate the need for more intelligent and adaptive frameworks.

A. Existing Strategies and Advances

Spectrum sharing in integrated TN-NTNs aims to efficiently utilize the limited spectrum resources by allowing both TN and NTN components to access the same spectrum, either simultaneously or opportunistically. Traditional spectrum sharing techniques can be broadly classified into static and dynamic approaches. Static spectrum sharing involves pre-allocating fixed portions of the spectrum to different networks or users, often based on geographical separation or orthogonal frequency bands. In legacy systems, for example, geostationary orbit (GEO) satellites and TNs operated in separate frequency bands. However, as spectrum demands grow, such fixed allocation becomes inefficient and leaves spectrum resources largely underutilized [9].

Dynamic spectrum access (DSA), on the other hand, allows for more flexible and adaptive spectrum usage. One common approach is opportunistic spectrum access, where secondary users (SUs) can access licensed spectrum bands when they are not being used by primary users (PUs). In this scenario, SUs perform spectrum sensing to detect the presence or absence of PUs and dynamically adjust their transmission parameters accordingly. Various spectrum sensing techniques, such as energy detection, matched filtering, and cyclostationary feature detection, have been employed in cognitive radio (CR) [10]. Another approach for dynamic spectrum sharing is to allow co-channel or adjacent-channel coexistence between different networks provided that the interference levels are constrained. This can be achieved through different interference management techniques, including power control, beamforming, and interference cancellation [11].

Recent advanced have focused on making DSA more efficient, scalable, and adaptive to dynamic network conditions. For instance, database-driven spectrum sharing uses a central database to store information about spectrum availability and usage rights, thereby enabling more efficient coordination between different users and networks [12]. Advancements in CR systems have also contributed to the development of more intelligent DSA techniques; for example, cooperative spectrum sensing allows SUs to collaborate and share sensing results to improve detection performance. Similarly, spectrum mobility techniques enable SUs to switch between different frequency bands or networks dynamically, based on changing channel conditions or traffic demands.

B. AI-Driven Spectrum Sharing

While traditional DSA techniques offer improvements over static spectrum allocation, they often face challenges in dealing with the complex and dynamic nature of integrated TN-NTNs. Conventional optimization and game theory-based approaches, commonly used in DSA, often require complete system information and rely on simplifying assumptions, making them impractical in real-world scenarios with limited information and heterogeneous network components. The emergence of artificial intelligence (AI)-based solutions has opened new avenues for developing more intelligent spectrum sharing solutions. AI-driven techniques excel at learning from experiences, adapting to changing network conditions, and making decisions in complex environments with incomplete information, resulting in efficient spectrum utilization and improved network performance.

Several AI techniques have been utilized for spectrum sharing in wireless networks, including deep learning (DL), deep reinforcement learning (DRL), and federated learning (FL). DL models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), can be used for tasks like spectrum sensing, channel prediction, and interference classification. When availability of data is limited, DRL algorithms enable autonomous agents to learn optimal spectrum access policies by interacting with their deployed environment and receiving rewards for desirable actions and penalties for undesirable actions [6]. FL emerges as a promising solution for addressing privacy concerns and enabling efficient learning. It allows multiple devices to collaboratively train a shared ML model without sharing their raw data and has been used for cooperative spectrum sensing in CR [13].

C. Challenges and Limitations

Despite advances in spectrum sharing strategies, several limitations still hinder their full potential in integrated TN-NTNs. Traditional methods often rely on simplified assumptions about network characteristics and user behavior, which may not hold in real-world scenarios with heterogeneous network components and dynamic operating environments. Many recent methods assume perfect or near-perfect spectrum sensing capabilities, which is challenging to achieve in practice, especially in satellite communication where weak received signal strength and atmospheric effects can significantly affect sensing accuracy. Additionally, centralized approaches may face scalability issues, especially in large, geographically dispersed networks with many TN and NTN components.

AI-driven approaches also encounter challenges related to training data requirements, computational constraints, and explainability. The complexity of DL models and DRL agents can limit their deployment on resource-constrained devices, such as UAVs and satellites, which often have limited processing capabilities and energy budgets. Furthermore, ensuring the explainability and interpretability of AI-based spectrum sharing decisions is crucial for gaining trust and facilitating regulatory approval. These limitations motivate the need for more intelligent, robust, and adaptable spectrum sharing frameworks for integrated TN-NTNs that can handle the challenges of

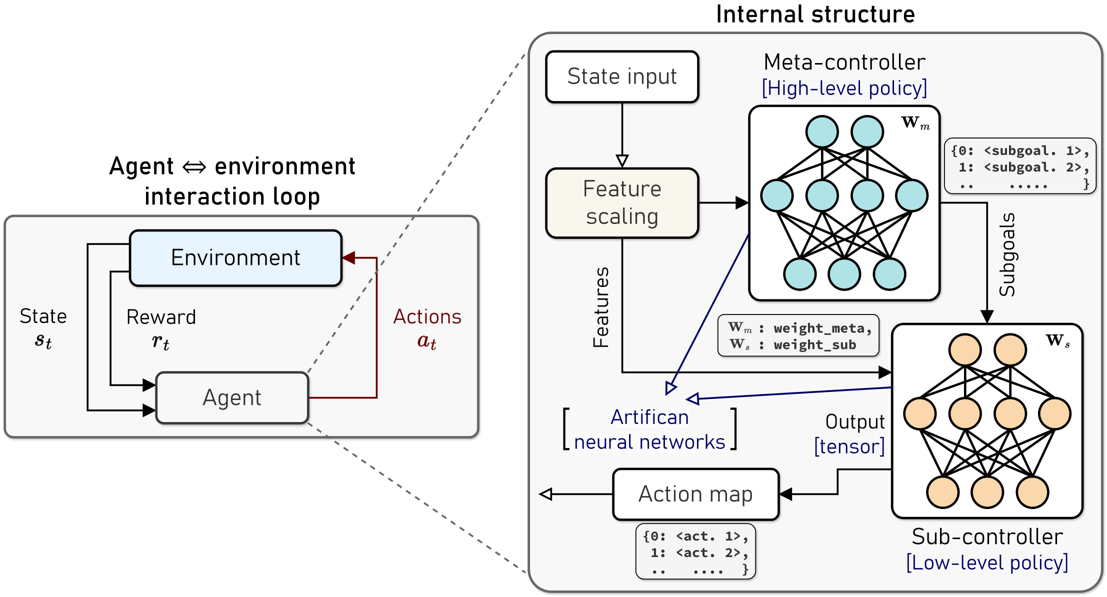

Fig. 1. Illustration of the hierarchical policy structure and agent-environment interaction loop.

complex, dynamic environments while considering practical constraints and addressing explainability concerns.

III. Fundamentals of Hierarchical Deep Reinforcement Learning

Before detailing the proposed framework for intelligent spectrum sharing in integrated TN-NTNs, we provide a general overview of DRL and its hierarchical extension, which forms the basis of our approach.

A. *Deep Reinforcement Learning*

DRL combines the function approximation capabilities of deep neural networks (DNNs) with the learning and decision-making framework of RL. This allows agents to learn complex policies directly from high-dimensional sensory inputs, such as images, sensor readings, or feedback signals, eliminating the need for manual feature engineering, though techniques like normalization and preprocessing remain beneficial.

1) Markov Decision Process: The mathematical framework underlying most RL algorithms is the Markov decision process (MDP). An MDP can be represented as a tuple $\langle S, A, P, R, \gamma \rangle$, where S is the set of possible states, A is the set of actions, P is the state transition probability function, R is the reward function, and γ is the discount factor. $P(s'|s,a)$ defines the probability of transitioning to state s' from state s when action a is taken. The reward function $R(s,a)$ specifies the immediate reward for taking action a in state s and plays a crucial role in shaping the agent's policy. The discount factor $\gamma \in [0,1]$ determines the trade-off between immediate and future rewards. The goal is to learn a policy $\pi(a|s)$ that maximizes the expected cumulative discounted reward, known as the return [8].

2) DNNs as Function Approximators: In many practical RL problems, the state and action spaces can be extremely large or even continuous, making it infeasible to store and update Q-values [1] or policy probabilities for every state-action pair in a Q-table. DNNs, a class of artificial neural networks (ANNs) with multiple layers, serve as function approximators in DRL. According to the universal approximation theorem, ANNs can approximate any continuous function given sufficient capacity. DNNs allow agents to learn complex mappings from states to actions or Q-values. The network takes the state as input and outputs either Q-values (in Q-learning) or the probability distribution over actions (in policy gradient methods). By training the DNN using appropriate algorithms and loss functions, the agent learns an effective policy to maximize its cumulative reward and optimize system control [6].

B. *Hierarchical Deep Reinforcement Learning*

HDRL extends DRL by introducing hierarchies to the decision-making process to address the challenges posed by complex real-world problems, such as spectrum sharing in integrated TN-NTNs. These problems often involve large action spaces that make learning slow and challenging. Instead of learning a single policy to map states to actions, HDRL agents learn multiple policies at different levels of abstraction, decomposing the overall task into a hierarchy of sub-tasks. Each level of the hierarchy manages a different sub-task contributing to the core goal while maintaining scalability [14]. Key principles of HDRL are as follows.

- *Hierarchical Policy:* High-level policies define abstract goals or sub-tasks, while low-level policies determine the specific actions to achieve them. This allows the agent to learn complex strategies by focusing on different levels of abstraction.
- *Temporal Abstraction:* Decisions are made at different timescales; high-level policies operate slower, setting

[1] Q-values represent the expected cumulative reward of taking an action a in state s and following a specific policy thereafter.

long-term goals, while low-level policies act faster, refining actions based on immediate observations and higher-level directives. Essentially, an agent can learn efficient strategies by avoiding constant decision-making for every aspect of the task.
- *Sample Efficiency:* HDRL improves sample efficiency through its hierarchical structure and agents can learn from fewer environmental interactions. This is particularly valuable in scenarios where data collection is expensive or time-consuming.

Fig. 1 illustrates a simple hierarchical policy structure of an agent interacting with an environment. Although only two levels of hierarchy are shown, HDRL can be easily extended to multiple levels, each with its own policy and corresponding DNN. The agent interacts with the environment in a typical RL loop: it observes the state, selects an action based on the current policy, receives a reward, and updates the policy parameters. At the top level, a meta-controller outputs subgoals to the sub-controller, which in turn outputs actions to the action mapper. This hierarchical structure allows the agent to make decisions at different levels of abstraction and timescales, leading to more efficient and adaptive learning. Hierarchical control is particularly beneficial for integrated TN-NTNs, where network tiers with limited computational capabilities can use higher-level directives to simplify their decision-making processes [1]. HDRL agents can adapt to both long-term global changes and short-term local fluctuations, crucial for adaptive spectrum sharing in dynamic integrated TN-NTN environments.

IV. HDRL-Based Intelligent Spectrum Sharing

This section presents our proposed HDRL-based framework for intelligent spectrum sharing in integrated TN-NTNs. We describe the system model and its components, detail the proposed framework and how it addresses the spectrum sharing problem, and analyze its complexity.

A. System Model

Fig. 2 illustrates our considered system comprising a low Earth orbit (LEO) satellite, HAPs, UAVs, TBSs, and users. The LEO satellite serves designated geographical areas using fixed multi-beam technology. Each beam cell is served by a HAP acting as a regional hub that relays data and control signals between the satellite and lower tiers. Multiple TBSs and UAVs are deployed within each HAP's coverage area. TBSs provide fixed high-capacity connectivity while UAVs operate as aerial BSs offering flexible on-demand coverage. This setup addresses coverage gaps, enhances network capacity, and accommodates temporary hotspots or high-traffic events. Users dynamically associate with either a TBS or a UAV based on factors such as signal strength, network load, and QoS requirements. A satellite gateway facilitates communication between the terrestrial network and the LEO satellite. A control and compute unit manages network operations, acting as a central coordinator.

The LEO satellite allocates portions of its available spectrum to each beam cell, which is further divided by the HAPs and jointly accessed by the UAVs and TBSs. To enhance spectral efficiency and user capacity, both UAVs and TBSs employ downlink non-orthogonal multiple access (NOMA). NOMA allows multiple users simultaneous access to the same frequency resource by exploiting power-domain multiplexing and successive interference cancellation (SIC) at the receiver [15]. Furthermore, to improve coverage, capacity, and cell-edge performance, UAVs and TBSs can utilize non-coherent joint transmission coordinated multi-point (JT-CoMP) in combination with NOMA (CoMP-NOMA), allowing multiple transmission points to cooperatively serve users.

This spectrum sharing scenario presents several key challenges: (1) maximizing overall network throughput through efficient inter-tier spectrum allocation; (2) managing inter-tier and intra-tier interference among nodes; and (3) adapting to the dynamic network environment including user mobility, varying channel conditions, and time-varying traffic demands. To address these challenges, we propose an HDRL framework that enables intelligent and adaptive spectrum sharing across TN and NTN.

B. Proposed Framework

The proposed framework mirrors the network's hierarchical structure with three levels: global, regional, and local, as shown in Figure 2. Each level corresponds to a specific network tier and employs a DNN as a policy to learn and optimize spectrum sharing via proximal policy optimization (PPO).

Unlike single-agent PPO, which optimizes a single policy, or multi-agent PPO (MAPPO), where each agent optimizes its own policy independently, our framework utilizes a hierarchical approach. Higher-level policies provide a context or subgoal for lower-level policies, constraining their action spaces. This allows for more efficient learning and decision-making in the complex spectrum sharing environment. The specific roles of each tier are as follows.

- *Global Tier:* The global policy π_g oversees the overall spectrum allocation for the entire network. It determines the optimal distribution of total spectrum chunks across beam cells based on the global network state S_g, which includes aggregated information about spectrum demand, user distribution, and channel conditions. This high-level allocation constrains the regional tier policies, ensuring fair and efficient spectrum distribution across different geographical areas.
- *Regional Tier:* Each HAP employs a regional policy π_r^i to manage spectrum resources within its designated coverage area. Given the spectrum allocation from the global tier, the regional policy further divides and assigns spectrum sub-bands to UAVs and TBSs under its control. This assignment is based on the regional network state S_r^i, which incorporates information about user distribution, traffic demands, and channel conditions within the HAP's coverage region. The regional policy adapts its sub-band assignments to optimize regional performance while adhering to the global tier's allocation.
- *Local Tier:* UAVs and TBSs utilize local policies to control real-time spectrum access and power allocation for their associated users. For each node j, the local policy π_l^j

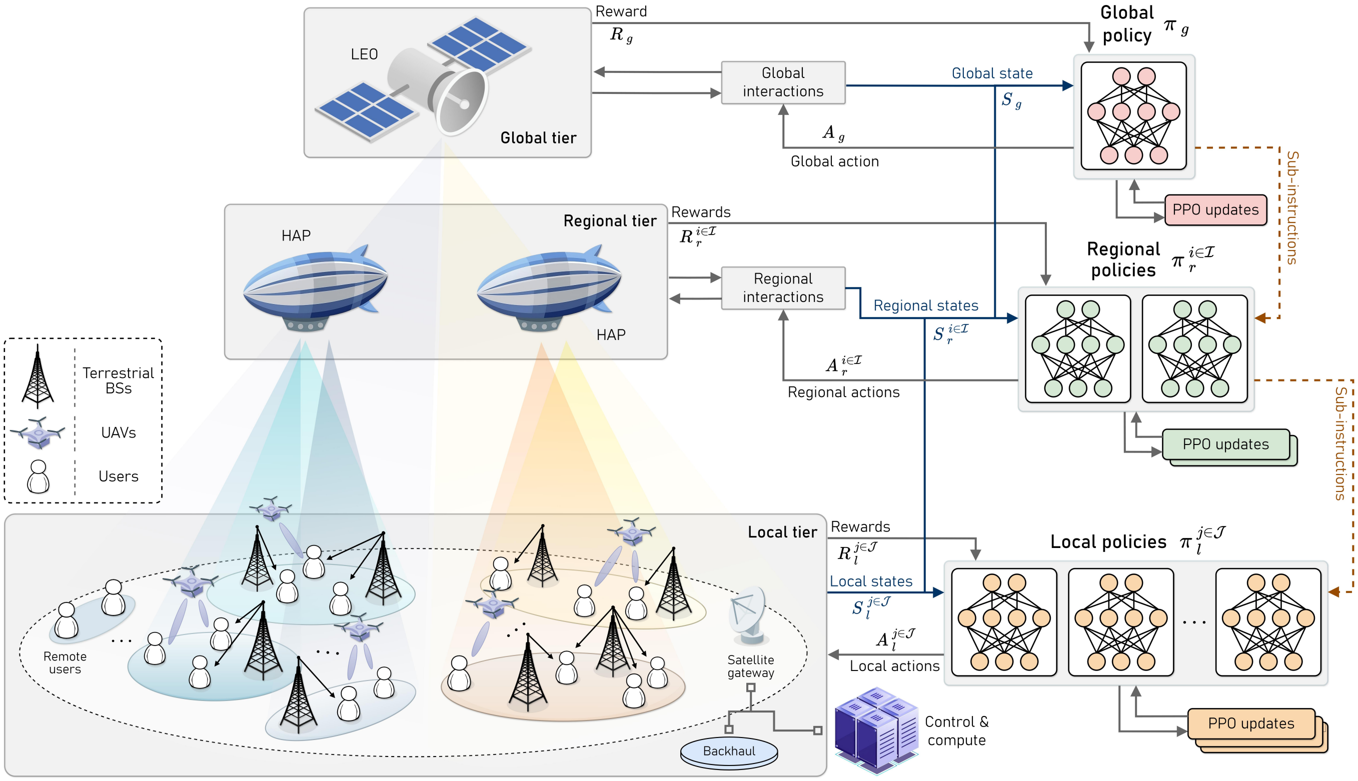

Fig. 2. System model and HDRL framework for spectrum sharing in integrated TN-NTNs.

makes fine-grained decisions on which specific channels and power levels to use for each connected user, based on the local network state S_l^j. This state includes information about individual user requirements, channel gains, and interference levels. The local policies operate within the spectrum sub-bands assigned by the regional tier, ensuring compliance with higher-level decisions while adapting to local network dynamics.

This hierarchical structure allows decision-making at different timescales and abstraction levels, aligning with each tier's unique operational characteristics and requirements. The global tier focuses on long-term, coarse-grained spectrum allocations to optimize overall network performance. The regional tier translates these allocations into more refined assignments tailored to each region's needs. Finally, the local tier makes real-time adjustments to spectrum access and power levels based on rapidly changing local conditions. This multi-tier approach enables efficient coordination and adaptation in the dynamic spectrum sharing environment.

1) Learning Process: The learning process in the proposed HDRL framework involves interactions between policies at different tiers and the environment. Initially, each policy is randomly initialized. The process begins with the local tier sensing their environment and constructing local state representations. These local states are then aggregated at the regional tier to form regional state representations, which incorporate information from the lower tier and the global spectrum allocation. At the highest level, i.e., global tier, aggregated information from all HAPs is received forming the global state.

Based on these state representations, each tier executes actions according to its current policy, with higher-tier actions serving as subgoals or constraints for lower-tier policies. Each policy is updated based on the received reward and the observed state transitions using the PPO algorithm. This iterative process of interaction, reward collection, and policy update continues until the learning converges to a near-optimal solution.

2) Complexity Analysis: The computational complexity of the proposed HDRL framework is influenced by the decision space dimension and the learning complexity at each tier. The decision space dimension, denoted as $|\mathcal{D}|$, represents the number of possible action combinations a policy can choose from.

At the global tier, the satellite allocates spectrum chunks to B beam cells from a pool of F available frequency bands. If each beam cell can receive multiple frequency bands, the decision space dimension is $|\mathcal{D}_g| = \binom{F+B-1}{B}$. If each beam cell can only receive one frequency band, and each frequency band can be allocated to multiple beam cells, the decision space dimension is $|\mathcal{D}_g| = \sum_{k=1}^{\min(F,B)} S(B,k)\binom{F}{k}$, where $S(B,k)$ represents the Stirling numbers of the second kind. At the regional tier, each HAP divides its allocated spectrum into S sub-bands for N_u UAVs and N_t TBSs. The decision space dimension for each HAP is $|\mathcal{D}_r| = S^{(N_u+N_t)}$, assuming each node can be allocated multiple sub-bands. At the local tier, each UAV/TBS controls spectrum access and power allocation for M associated users, selecting from C channels and P power levels. The decision space for each local policy is $|\mathcal{D}_l| = (C \times P)^M$.

The overall decision space complexity can be approximated as the product of the decision space dimensions across all tiers. However, the hierarchical structure significantly reduces this complexity compared to a flat, non-hierarchical approach, as higher-tier actions constrain the decision space of lower tiers [1].

The learning complexity depends on factors such as the size and architecture of the DNNs, the number of training episodes, and the specific exploration-exploitation strategy employed. Larger and more complex DNNs require more data and computational resources, leading to longer training times. The number of training episodes required for convergence depends on the environment's complexity and the learning algorithm. The choice of exploration-exploitation strategy, such as ϵ-greedy, softmax, or Boltzmann exploration, also affects the learning time by determining the balance between exploiting current knowledge and exploring new actions. Due to these factors, providing a precise theoretical estimate of learning time is challenging; however, empirical evaluations through simulations can provide valuable insights into the practical learning performance of the proposed framework.

V. Performance Evaluation

In this section, we evaluate the performance of the proposed HDRL-based spectrum sharing framework for integrated TN-NTNs through extensive simulations. We consider various network hierarchies, including space-air-ground, air-ground, and UAV-aided networks, to demonstrate the framework's adaptability. The proposed approach is compared with several baseline methods, including exhaustive search, random access, and single- and multi-agent reinforcement learning algorithms.

A. Simulation Setup

We simulate a system similar to that shown in Fig. 2. Specifically, the LEO satellite operates at an altitude of 550 km with two fixed beams, using Ka-band at 28 GHz with a total bandwidth of 200 MHz. Within each beam, a HAP is deployed at an altitude of 20 km each cover two non-overlapping regions. These regions accommodate two TBSs and one UAV that jointly serve terrestrial users. TBSs have high-power transmitters, while UAVs operate at an altitude of 100 m with lower transmission power, offering flexible deployment to enhance coverage and capacity.

The number of users per HAP region varies from 10 to 30, following a uniform distribution. Users dynamically associate with either TBSs or UAVs based on signal strength, network load, and QoS requirements. The LEO satellite dynamically allocates portions of its 200 MHz bandwidth to HAPs based on demand and network conditions. HAPs further divide their allocated spectrum into 10 sub-bands for TBSs and UAVs, enabling efficient resource utilization across the network tiers.

To capture the diverse propagation characteristics, we employ a shadowed Rician fading model for satellite links and Rayleigh fading for terrestrial links. The channel models incorporate path loss, atmospheric effects, and small-scale fading appropriate for each link type. The simulation time is divided into discrete slots. The satellite makes decisions every 50 slots, HAPs every 10 slots, and TBSs/UAVs every slot, reflecting the different computational capabilities and control cycles of network entities. The SINR for each user is calculated considering intra-tier and inter-tier interference, as well as noise. Throughput is derived using the Shannon capacity formula: $C = B\log_2(1 + \text{SINR})$, where B is the bandwidth and SINR is the signal-to-interference-plus-noise ratio.

It is worth noting that our proposed framework adapts to different network hierarchies. In air-ground network scenarios without the satellite tier, HAPs assume overall spectrum control. For UAV-aided networks, the framework reduces to a single-agent DRL with distributed policies for TBSs and UAVs, with a central control unit allocating spectrum to different regions.

B. Results and Discussion

We compare the performance of our proposed HDRL framework against several baseline approaches:

- *Exhaustive Search*: This method explores all possible spectrum allocation combinations to find the optimal solution. It guarantees the best performance but is computationally intensive and infeasible for large-scale deployments.
- *Random Access*: This approach randomly allocates spectrum resources, serving as a lower bound for performance.
- *PPO*: A single-agent reinforcement learning algorithm optimizing a single policy for the entire network.
- *MAPPO*: A multi-agent version of PPO where each network node optimizes its own policy independently.

Unless stated otherwise, we consider space-air-ground network as the default scenario. The proposed framework is evaluated in terms of learning and convergence, spectral efficiency, average network throughput, and execution time.

1) Learning and Convergence: Fig. 3 illustrates the learning and convergence behavior of our framework across satellite-air-ground, air-ground, and UAV-aided network hierarchies. The average cumulative reward, normalized for fair comparison, is plotted against the number of training episodes. For each hierarchy, the rewards steadily increase over time, converging to a stable value around 700 to 800 episodes. However, the rewards exhibit fluctuations around the converged value due to the dynamic nature of the channels and evolving network conditions. This result demonstrates the ability of HDRL to learn effective spectrum sharing policies in various network configurations while adapting to the specific challenges and opportunities presented by each hierarchy.

2) Spectral Efficiency: Fig. 4 compares the spectral efficiency achieved by different spectrum sharing algorithms across the three network hierarchies. Our hierarchical framework achieves near-optimal performance, as compared to exhaustive search, demonstrating its effectiveness in optimizing spectrum utilization. MAPPO also achieves similar performance and even outperforms our framework in air-ground and UAV-aided networks, which can be attributed to simpler network structures and fewer nodes. PPO struggles due to the large action space and lack of coordination among network nodes, resulting in suboptimal performance. The random

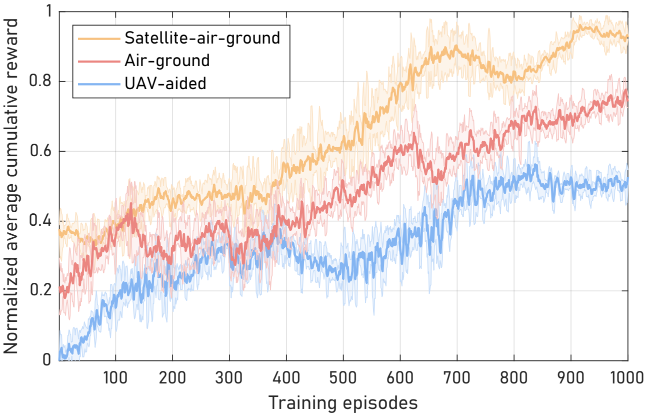

Fig. 3. Normalized average cumulative reward of the proposed HDRL framework for different network hierarchies.

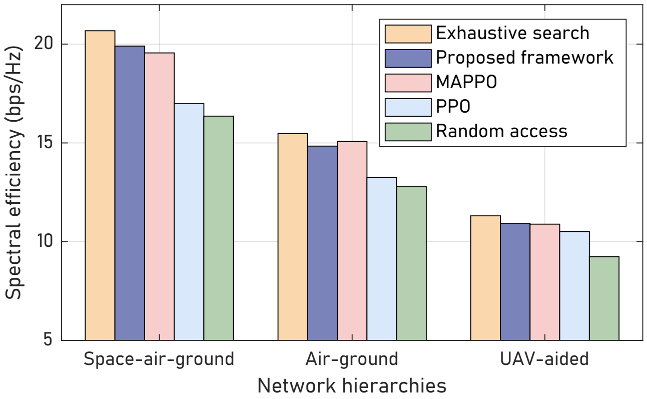

Fig. 4. Spectral efficiency achieved by different algorithms for different network hierarchies.

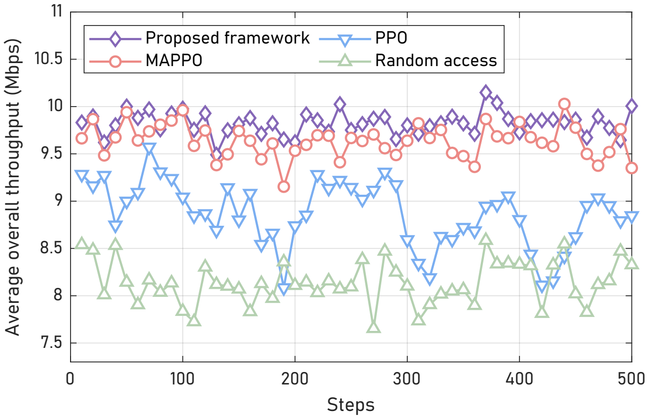

Fig. 5. Average network throughput achieved by different algorithms.

access scheme yields the lowest spectral efficiency since it allocates resources without considering network conditions or user requirements. These results highlight the advantages of our hierarchical approach in efficiently managing spectrum resources while maintaining scalability and adaptability.

3) Average Network Throughput: Fig. 5 presents a comparison of the average network throughput achieved by various algorithms over the course of an episode, with measurements taken every 10 steps. Our hierarchical framework consistently achieves better performance than other algorithms. MAPPO performs similarly to our framework but exhibits higher variance, likely due to autonomous agents making independent decisions without explicit coordination despite being in a collaborative setting. PPO yields the lowest throughput among the learning-based methods, as it struggles to handle the exploded action space and lacks coordination among network nodes. These findings highlight the importance of leveraging the network's hierarchical structure and enabling efficient inter-tier collaboration for maximizing system performance.

4) Execution Time: Table I compares the execution time of different spectrum sharing algorithms. The execution time is calculated by averaging the time taken by each control node to make a decision across all steps in an episode, providing a fair comparison considering the varying control cycles of the global, regional, and local tiers in our hierarchical framework. The random access scheme exhibits the lowest execution time since it does not involve any learning or optimization processes. The exhaustive search method requires the most time due to its loop-based exploration of all possible action combinations. Among the learning-based approaches, PPO achieves a low execution time, despite its suboptimal performance, as it optimizes a single policy for the entire network. Our hierarchical framework outperforms the MAPPO algorithm in terms of execution time, benefiting from the fact that not all policies are executed at each step, unlike the independent agents in the multi-agent setting.

VI. Conclusion and Future Directions

This article proposed an HDRL-based framework for intelligent spectrum sharing in integrated TN-NTNs. The framework leverages the network's inherent hierarchy, with separate policies for each tier, to learn and optimize spectrum allocation decisions at different timescales and levels of abstraction. By decomposing the complex spectrum sharing problem into manageable sub-tasks and enabling efficient inter-tier coordination, the HDRL approach offers a scalable and adaptive solution for spectrum management in future TN-NTNs. Simulation results demonstrate the performance gains of our framework compared to other approaches, such as exhaustive search, random access, and other DRL frameworks. HDRL achieves near-optimal spectral efficiency and network throughput while maintaining low execution times, making it suitable for real-time applications in future 6G networks.

Looking ahead, several exciting research directions emerge. Extending the HDRL framework to incorporate other resource

TABLE I
Comparison of Execution Time for Different Algorithms.

Algorithm	**Execution Time** (s)
Exhaustive Search	4.2×10^{-1}
Random Access	3.7×10^{-4}
PPO	1.4×10^{-3}
MAPPO	3.1×10^{-2}
Proposed framework	7.6×10^{-3}

management aspects, such as user grouping and mobility management, would create a more comprehensive solution for integrated TN-NTNs. Exploring transfer learning and meta-learning techniques could enable rapid adaptation of the learned policies to new environments or network configurations. Testing the robustness and scalability of framework under various network scenarios, including ultra-dense deployments and heterogeneous traffic demands, would also provide valuable insights for practical implementation. Finally, integrating the HDRL approach with emerging technologies like intelligent reflecting surfaces and edge computing could unlock new possibilities for enhancing the performance and efficiency of future wireless networks.

References

[1] Y. Cao, S.-Y. Lien, Y.-C. Liang, and D. Niyato, "Multi-tier deep reinforcement learning for non-terrestrial networks," *IEEE Wirel. Commun.*, vol. 31, no. 3, pp. 194–201, 2024.

[2] A. Vanelli-Coralli, A. Guidotti, T. Foggi, G. Colavolpe, and G. Montorsi, "5G and beyond 5G non-terrestrial networks: Trends and research challenges," in *IEEE 5G World Forum*, pp. 163–169, 2020.

[3] F. Qamar, M. U. A. Siddiqui, M. N. Hindia, R. Hassan, and Q. N. Nguyen, "Issues, challenges, and research trends in spectrum management: A comprehensive overview and new vision for designing 6G networks," *Electronics*, vol. 9, no. 9, 2020.

[4] H.-W. Lee, A. Medles, V. Jie, D. Lin, X. Zhu, I.-K. Fu, and H.-Y. Wei, "Reverse spectrum allocation for spectrum sharing between TN and NTN," in *Proc. IEEE Conf. Stand. Commun. Netw.*, pp. 1–6, 2021.

[5] H. Martikainen, M. Majamaa, and J. Puttonen, "Coordinated dynamic spectrum sharing between terrestrial and non-terrestrial networks in 5G and beyond," in *Proc. IEEE Int. Symp. World Wireless, Mob. Multimedia Netw.*, pp. 419–424, 2023.

[6] J. Si, R. Huang, Z. Li, H. Hu, Y. Jin, J. Cheng, and N. Al-Dhahir, "When spectrum sharing in cognitive networks meets deep reinforcement learning: Architecture, fundamentals, and challenges," *IEEE Netw.*, vol. 38, no. 1, pp. 187–195, 2024.

[7] S. Jo, W. Yang, H. K. Choi, E. Noh, H.-S. Jo, and J. Park, "Deep Q-learning-based transmission power control of a high altitude platform station with spectrum sharing," *Sensors*, vol. 22, no. 4, 2022.

[8] T. Naous, M. Itani, M. Awad, and S. Sharafeddine, "Reinforcement learning in the sky: A survey on enabling intelligence in NTN-based communications," *IEEE Access*, vol. 11, pp. 19941–19968, 2023.

[9] C. Zhang, C. Jiang, L. Kuang, J. Jin, Y. He, and Z. Han, "Spatial spectrum sharing for satellite and terrestrial communication networks," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 55, no. 3, pp. 1075–1089, 2019.

[10] A. Nasser, H. Al Haj Hassan, J. Abou Chaaya, A. Mansour, and K.-C. Yao, "Spectrum sensing for cognitive radio: Recent advances and future challenge," *Sensors*, vol. 21, no. 7, 2021.

[11] L. Zhang, Z. Wei, L. Wang, X. Yuan, H. Wu, and W. Xu, "Spectrum sharing in the sky and space: A survey," *Sensors*, vol. 23, no. 1, 2023.

[12] Y.-C. Liang, J. Tan, H. Jia, J. Zhang, and L. Zhao, "Realizing intelligent spectrum management for integrated satellite and terrestrial networks," *J. Commun. Inf. Netw.*, vol. 6, no. 1, pp. 32–43, 2021.

[13] Z. Chen, Y.-Q. Xu, H. Wang, and D. Guo, "Federated learning-based cooperative spectrum sensing in cognitive radio," *IEEE Commun. Lett.*, vol. 26, no. 2, pp. 330–334, 2022.

[14] S. Pateria, B. Subagdja, A.-h. Tan, and C. Quek, "Hierarchical reinforcement learning: A comprehensive survey," *ACM Comput. Surv.*, vol. 54, no. 5, 2021.

[15] X. Liu, K.-Y. Lam, F. Li, J. Zhao, L. Wang, and T. S. Durrani, "Spectrum sharing for 6G integrated satellite-terrestrial communication networks based on NOMA and CR," *IEEE Netw.*, vol. 35, no. 4, pp. 28–34, 2021.