Leader-Follower Coordination in UAV Swarms for Autonomous 3D Exploration via Reinforcement Learning

Robert Kathrein[1,2*], Julian Bialas[1,2*], Mohammad Reza Mohebbi[1,2*], Simone Walch[1], Mario Döller[1] and Kenneth Hakr[3]

[1]*University of Applied Sciences Kufstein Tirol, Kufstein, Austria*

[2]*University of Passau, Passau, Germany*

[3]*Twins GmbH, Austria*

Keywords: Unmanned Aerial Vehicle, Leader-Follower Architecture, Swarm Algorithms, Reinforcement Learning, Path Planning, 3D Exploration.

Abstract: Autonomous volumetric scanning in three-dimensional environments is critical for environmental monitoring, infrastructure inspection, and search and rescue applications. Efficient coordination of multiple Unmanned Aerial Vehicles (UAVs) is essential to achieving complete and energy-aware coverage of complex spaces. In this work, a Reinforcement Learning (RL)-based framework is proposed for the coordination of a leader-follower UAV system performing volumetric scanning. The system consists of two heterogeneous UAVs with directional sensors and constant mutual orientation during the mission. A centralized control policy is learned based on Proximal Policy Optimization (PPO) to control the leader UAV, which produces trajectory commands for the follower to achieve synchronized movement and effective space coverage. The observation space includes a local 3D occupancy map of the leader and both UAVs' battery levels, enabling energy-aware decision-making. The reward function is carefully designed to favor exploration and visiting new regions without penalizing collision and boundary crossing. The proposed method is verified using both simulation experiments and real-world experiments on the ArduPilot platform, showing the applicability of RL to scalable autonomous multi-UAV scanning operations.

1 INTRODUCTION

Unmanned Aerial Vehicles (UAVs) are finding core platforms in environmental monitoring, disaster response, infrastructure inspection, and precision agriculture applications due to their autonomous operation and ability to acquire high-resolution data (Mohebbi et al., 2024). However, a vast majority of existing UAV missions remain founded upon single-drone missions, which restrain efficiency, scalability, and precision (Mohsan et al., 2023). To address these challenges, coordinated multi-UAV systems have been conceptualized, with the ability to allow simultaneous collaboration among a number of drones to undertake large-scale and time-critical missions. In such a scenario, parallel flight has been a foundational capability, as it ensures simultaneous control among a number of UAVs and provides a basis for higher-level cooperative tasks, including environmental scanning over large areas and taking a quick situation assessment (Chen and Deng, 2023).

One significant use of parallel flight coordination involves the observation of emissions from gases, particularly methane (CH_4), which is highly potent greenhouse gas causing global warming (Shaw et al., 2021; Filonchyk et al., 2024). Existing UAV-based monitoring techniques are typically two-dimensional scans, which neglect vertical dispersion within gases to provide incomplete surveys. Because gases are diffused within a three-dimensional (3D) volume, volumetric scans are required to achieve complete and accurate mapping of concentration levels within gases (Tosato et al., 2019). This necessity suggests a need to develop intelligent UAV systems to achieve optimized volumetric scans by adaptive controls.

To address these challenges, a volumetric scanning system with parallel coordination of UAVs inspired by reinforcement learning is proposed here. The system integrates concurrent maneuvering of

*Robert Kathrein, Julian Bialas, and Mohammad Reza Mohebbi contributed equally and are corresponding authors.

Kathrein, R., Bialas, J., Mohebbi, M. R., Walch, S., Döller, M. and Hakr, K.
Leader-Follower Coordination in UAV Swarms for Autonomous 3D Exploration via Reinforcement Learning.
DOI: 10.5220/0013673400003982

In *Proceedings of the 22nd International Conference on Informatics in Control, Automation and Robotics (ICINCO 2025) - Volume 1*, pages 176-183
ISBN: 978-989-758-770-2; ISSN: 2184-2809

multiple UAVs with a Proximal Policy Optimization (PPO) algorithm to, in real time, dynamically recalculate flight paths so as to attain maximum coverage of methane concentrations with minimized energy expenditures. This developed system has the capability to operate adaptively within uncertain environments and can be scaled up to larger UAV swarms to attain scalable environmental monitoring.

The contributions of this work are summarized as follows:

- **Parallel Flight Coordination:** Running a single-input control scheme to obtain coordinated operation among UAVs, increasing scalability and operational efficiency.
- **Volumetric Scanning Method:** Application of a 3D environmental monitoring method to accurately chart gases, beyond 2D scan limitations.
- **Optimization of Reinforcement Learning:** Developing a PPO policy to enable autonomous path planning with better coverage and energy optimization.
- **Scalability with Multi-UAV Systems:** Development of a system that can be deployed within large UAV swarms to be applicable to adaptive monitoring of extensive regions.

2 LITERATURE REVIEW

Various methods have been proposed to achieve coordination between pairs of UAVs, generally categorized into two strategies: physical interconnections and wireless communications. Both approaches have been shown to enhance stability, improve coordination, and increase operational efficiency, particularly in scenarios requiring precise synchrony under diverse environmental conditions.

Physical interconnecting approaches have focused primarily on improving stability during cooperative maneuvers. In this regard, Sato et al. (Sato and et al., 2020) employed interconnected rods or strings to reduce the influence of rotor downwash during gas measurements to improve reliability in detecting. In similar endeavors, Six et al. (Six et al., 2018) envisioned a hard-bodied articulated body with a higher payload capacity but reduced aerodynamic interferences, resulting in higher precision and stable maneuvers. Spurny et al. (Spurny et al., 2019) showed a cooperative transportation system with a suspended payload from a two-cable system by adopting a Rapidly-Exploring Random Tree (RRT)-based path planning algorithm to improve safety during external missions. Bulka et al. (Bulka et al., 2022) further showed the potential advantages of mechanical linkages during payload delivery missions but emphasized robust communication protocols to achieve proper coordination. Despite these advantages, physical interconnecting approaches are inherently constrained by limited scalability and adaptability to dynamic environments, leading to the consideration of alternative solutions (Liu et al., 2021).

Wireless communication-based approaches were generally popular because of their ability to easily coordinate multiple UAVs. Rafifandi et al. (Rafifandi et al., 2019) employed a Leader–Follower (L-F) control structure involving visual positioning and Proportional-Derivative (PD) controllers to obtain collision-free denser formations. Walter et al. (Walter et al., 2019) created Ultraviolet Direction and Ranging (UVDAR), an ultraviolet positioning system that provides stable relative localization irrespective of ambient conditions. Chen et al. (Chen et al., 2018) exceeded the classical approaches of L-F methods by synergistically combining Ultra-Wide Band (UWB) distance measurements with GPS navigation, resulting in higher precision coordination. Similarly, Zhang et al. (Zhang et al., 2022) improved formation control with encircling potentials with consensus-based observers, while Cross et al. (Cross, 2023) created a macroscale UAV system with real-time communications and adaptive response, supporting real small-scale implementations. Individually, these research articles show robust communicative mechanisms and adaptive algorithms to attain reliable multi-UAV collaboration.

Although both approaches are effective, hard-wired interconnections restrict operational flexibility, and wireless methods are hampered by issues such as latency and drift in dynamic environments (Li et al., 2021). Previous work using L-F control has been dedicated to formation control strategies, with no additional work given to adaptive strategies with Reinforcement Learning (RL) as a foundation. To our knowledge, no study has previously applied any combination of RL with L-F strategies to achieve coordination with parallel flight. The neglect inspires our present study, which provides an RL-based approach that integrates synchronized UAV maneuvering with adaptive optimization to achieve enhanced stability, scalability, and performance with multi-UAV systems.

3 METHODOLOGY

To enable efficient, collision-free volumetric scanning, an RL framework coordinates a swarm of UAVs

in a L-F setup. Each agent consists of two UAVs: the leader handles navigation and issues high-level commands, while the follower mirrors movements. Both maintain continuous sensor alignment since the methane sensor is split between them, measuring concentration via a laser scanning the air volume between leader and follower.

This paper focuses on navigation and coordination of the UAV pair. Training uses PPO to maximize volumetric coverage, minimize collisions, and prevent excursions beyond scan boundaries. Built in ArduPilot, the simulation supports strong sim-to-real transfer with tests in both simulated and physical environments. The policy is hierarchical: the leader runs the RL policy, commanding the follower to maintain sensor alignment and avoid terrain collisions through controlled positioning and rotation.

3.1 Problem Formulation

The environment is represented as a 3D occupancy grid, discretized into a finite set of volumetric elements (voxels). It is encoded as a Boolean tensor: $O \in \{0,1\}^{X\times Y\times Z}$ where $X,Y,Z \in \mathbb{N}$ denote the spatial resolution along each axis. Each voxel $O(x,y,z)$ is assigned a value of 0 if it corresponds to free space, or 1 if it is occupied by an obstacle.

A UAV agent is modeled as a coordinated pair of UAVs, indexed by $i \in \{1,2\}$, referred to as the *leader* and the *follower*, respectively. At each discrete time step $t \in \mathcal{T} = \{t_0,\ldots,t_{\text{terminal}}\}$, the state of the UAV i is given by: $a_{i,t} = (p_{i,t}, b_{i,t})$ where $p_{i,t} \in \mathbb{N}^3$ denotes the position of the UAV in the voxel grid, and $b_{i,t} \in [0,1]$ is its normalized battery level, interpreted as the remaining movement budget.

At each timestep, a volumetric measurement is obtained via a laser-based sensor system that scans the air volume between the two UAVs. This measurement is valid only if the line of sight between them is unobstructed. The set of measured voxels is defined as:

$$\mathcal{M}_t = \begin{cases} \text{line}(p_{1,t}, p_{2,t}), & \text{if } \forall v \in \text{line}(p_{1,t}, p_{2,t}) : O(v) = 0 \\ \emptyset, & \text{otherwise} \end{cases} \tag{1}$$

Here, $\text{line}(p_{1,t}, p_{2,t})$ are the discrete set of voxels along the straight-line segment joining $p_{1,t}$ and $p_{2,t}$, computed using a 3D extension of Bresenham's line algorithm.

Let $\mathcal{U}_t$ represent the set of all voxels that have been successfully measured up to time t. The mission's objective is to maximize the number of previously unmeasured voxels over the entire time horizon: $\max \sum_{t=0}^{t_{\text{terminal}}} |\mathcal{M}_t \setminus \mathcal{U}_t|$ subject to the update rule: $\mathcal{U}_{t+1} = \mathcal{U}_t \cup \mathcal{M}_t$. Accordingly, the UAVs must coordinate their trajectories to continuously reposition themselves so as to yield new, non-redundant measurements at each timestep. Simultaneously, battery constraints must be respected to ensure operational feasibility throughout the mission.

3.2 Framework Overview

The L-F coordination approach addresses the problem by running the dynamic coverage planner only on the leader UAV. The method depends on maintaining a direct line-of-sight between leader and follower, converting sensing into volumetric line-based measurements, and extending the single-agent planner from earlier work (Bialas et al., 2023).

The deep reinforcement learning-based path planner trains the leader to maximize exploration of unseen voxels. Unlike prior single-agent methods, the framework leverages spatial separation of two coordinated UAVs for broader visibility. The follower receives high-frequency position updates and maintains a constant distance to ensure line-of-sight sensing.

To improve navigation safety in cluttered environments, a control mechanism lets the leader command follower rotations for lateral corrections. Clockwise or counterclockwise turns are issued to preserve line of sight and avoid potential collisions.

3.3 RL Coverage Path Planner

The leader UAV's coverage path planner is designed as a partially observable Markov decision process (POMDP) (Littman, 2009) using the same architecture as in previous work (Bialas et al., 2023). A POMDP is defined as (S,A,Ω,T_a,O,R), where: S is the state space, A is the action space, Ω is the observation space, T_a is the transition function under action a, O is the observation function, and R is the reward function.

The objective of the RL agent is to learn an optimal policy to maximize the overall reward by moving through the environment and executing volumetric coverage.

3.3.1 State Space

The state space contains both environmental and agent-specific information required for decision-making. It is described as:

$$S = \underbrace{\{0,1\}^{X\times Y\times Z}}_{O} \times \underbrace{\{0,1\}^{X\times Y\times Z}}_{U} \times \underbrace{\mathbb{N}^{2\times 3}}_{p_{1,t},p_{2,t}} \times \underbrace{[0,1]^2}_{b_{1,t},b_{2,t}}. \tag{2}$$

Where O is the 3D occupancy representation of the environment in which each voxel is either empty

or blocked. U is the accumulated set of observed voxels by the UAVs as a record of previous measurements. The third factor represents the current locations of both the leader and follower UAVs at time t. The last component represents their respective battery levels normalized in the range $[0,1]$.

This formulation ensures the agent's decisions are guided by spatial constraints and energy limitations to allow it to safely and efficiently explore and maximize new voxel coverage.

3.3.2 Action Space

The agent operates within a discrete action space composed of nine predefined actions: $A = \{0,1,2,3,4,5,6,7,8\}$.

Six actions control translational movement along the Cartesian axes—north, south, east, west, up, and down—moving both leader and follower synchronously to maintain formation.

Two actions adjust the follower's orientation via orbit maneuvers, rotating it around the leader clockwise or counterclockwise for adaptive positioning in constrained or dynamic spaces. An explicit Return-To-Launch (RTL) command ends the episode and directs both UAVs back to their launch points, enabling safe mission termination under critical conditions like battery depletion or blocked paths.

3.3.3 Observation

At each time step, the agent receives an observation vector that describes the spatial and operational context. Central to this is a local occupancy map centered on the leader UAV's position within a fixed spatial window. The map represents static obstacles and visited areas in a compact form, giving an immediate environment state. The map's resolution and update rate trade off accuracy for computational speed, so the map is suitable for real-time use in simulations and physical tests.

The perception also incorporates L-F levels of the battery, $b_{1,t}$ and $b_{2,t}$, so the agent is able to consider energy constraints, reduce unnecessary motion, and plan energy-optimal routes. The integration of spatial and resource information enables the model to strike a balance among exploration efficiency, safety, and staying power.

3.3.4 Reward Function

The reward function guides the RL agent to maximize volume coverage efficiently while minimizing risk and waste. It combines positive and negative terms to encourage safe, effective exploration in 3D environments (Nguyen et al., 2020).

Positive reward comes from scanning previously unvisited voxels. At each step, the agent is rewarded based on new voxels uncovered via the leader-follower line-of-sight, driving it to seek novel viewpoints and maintain wide coverage. This prevents local saturation and supports area-wide surveillance in 3D scan tasks.

Penalties promote safety and feasibility. Collisions with terrain incur strong negative rewards, and leaving operational limits—such as exceeding altitude or scanning boundaries—triggers critical mission penalties. These act as hard constraints, steering policies away from unsafe behavior.

Energy efficiency is enforced by penalizing low-battery, energy-hungry maneuvers or redundant high-speed motion, pushing the agent toward sustainable flight to extend mission time.

Together, these rewards create a structured signal that fosters high coverage, collision avoidance, and endurance-aware strategies. As a result, cooperative behaviors emerge between leader and follower UAVs for adaptive, redundant exploration of unknown 3D environments.

Positive Reward – Scanning The primary positive reward is obtained from scanning new voxels, incentivizing the agent to explore unvisited areas and maximize the overall volumetric coverage of the environment.

$$R_{\text{scan}}(s_t, a_t) = r_{\text{scan}} \cdot |M_t \setminus U_t|. \tag{3}$$

Negative Rewards – Safety Violations. Penalties are applied for violations of operational constraints, such as collisions, exceeding altitude limits, or moving outside the designated scanning area.

- **Collision Penalty:**

$$R_{\text{collision}}(s_t, a_t) = -r_{\text{collision}} \sum_{i \in \{1,2\}} \mathbf{1}_{\text{collision}}(p_{i,t}), \tag{4}$$

- **Boundary Violation Penalty:**

$$R_{\text{boundary}}(s_t, a_t) = -r_{\text{boundary}} \sum_{i \in \{1,2\}} \mathbf{1}_{\text{out of bounds}}(p_{i,t}), \tag{5}$$

- **Battery Depletion Penalty:**

$$R_{\text{battery}}(s_t, a_t) = -r_{\text{battery}} \sum_{i \in \{1,2\}} \mathbf{1}_{\text{battery depleted}}(b_{i,t}). \tag{6}$$

3.3.5 Policy Optimization with PPO

The agent's objective is to learn a policy π that maps observed states to a distribution over possible actions, thereby maximizing the expected cumulative reward

over time. Formally, the policy is defined as a probabilistic function parameterized by θ:

$$\pi : A_i \times S \to [0,1], \quad (a_i, s) \mapsto \pi(a_i \mid s; \theta), \quad (7)$$

where A_i is the action space for agent i, S is the state space, and $\pi(a_i \mid s; \theta)$ denotes the probability of selecting action a_i in state s under the current policy parameters θ.

The goal is to find the optimal policy π^* that maximizes the expected cumulative discounted reward across a trajectory $\{s_0, \ldots, s_{t_{\text{terminal}}}\}$: $\pi^* = \arg\max_\theta \mathbb{E}_{\pi_\theta}[R_0]$, where the cumulative reward from timestep t is given by: $R_t = \sum_{k=0}^{t_{\text{terminal}}-t} \gamma^k r_{t+k}$ with $\gamma \in [0,1)$ representing the discount factor, which prioritizes immediate rewards over distant future ones. The immediate reward at time t is defined as a function of the current state s_t and the actions sampled from the policy:

$$r_t = R(s_t, s_{t+1}) = \sum_{i \in U} R(s_t, \pi(a_{i,t} \mid s_t; \theta)), \quad (8)$$

where U is the set of agents—in this case, the leader and follower UAVs.

In order to obtain the policy, PPO is employed. This is an on-policy RL algorithm that is highly regarded for its balance between training stability and exploration. The policy is optimized through the refinement of a surrogate objective, within which a clipping mechanism is utilized to constrain deviations between the new and old policies. By doing this, training is more consistent and dependable, with the avoidance of sudden or harmful shifts throughout the learning process.

The PPO objective is expressed as:

$$L_{\text{CLIP}}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(r_t(\theta) A_t, \text{clip}(r_t(\theta), 1-\varepsilon, 1+\varepsilon) A_t \right) \right], \quad (9)$$

where:

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}. \quad (10)$$

is the probability ratio between the new and old policies, A_t is the estimated advantage function at time t, quantifying the relative value of action a_t in state s_t compared to the expected value of the current policy, and ε is a small hyperparameter (typically $\varepsilon = 0.1$–0.3) controlling the allowed deviation.

By maximizing this clipped objective, the PPO algorithm ensures that policy updates improve performance while preserving stability and sample efficiency. This makes PPO particularly suitable for real-time UAV coordination tasks in complex and partially observable 3D environments, where large and unstable policy changes could lead to erratic behaviors or unsafe trajectories.

4 EXPERIMENTAL RESULTS

This section details experimental testing of the UAV parallel flight system for volumetric scanning tasks. The system uses an RL-based control policy optimized with PPO to manage UAV maneuvers in real-time 3D environments, aiming to maximize scan coverage while ensuring efficiency, stability, and safety.

Validation involved both simulations and real-world tests in controlled indoor and outdoor settings. Scenarios evaluated the UAVs' ability to perform coordinated parallel flight for full volumetric exploration. Key performance metrics included maneuver stability, trajectory accuracy, scan completeness, and responsiveness to environmental constraints.

The RL agent was first trained in a custom 3D simulator where PPO optimized policy through episodic interaction. The learned policy was then deployed on real UAVs to test robustness and sim-to-real transfer. For benchmarking, RL coordination was compared with a rule-based control scheme to highlight performance gains.

4.1 L-F UAV Coordination

In order to validate the performance of the proposed L-F coordination scheme, extensive testing was conducted in simulation as well as in outdoor real-world environments. The intent was to validate the follower UAV's ability to emulate the flight trajectory of the leader with high accuracy and with minimal response delay. Special emphasis was made on the stability of the formation and the preservation of even inter-UAV separations through dynamic multi-axis maneuvers.

In the simulation stage, a representative set of 3D trajectories was carried out by the leader UAV, such as paths with sharp direction changes as well as vertical transitions. The follower UAV was provided with real-time waypoints and performed corresponding maneuvers with negligible deviation. The recorded trajectories in Figure 1 show high correlation between the leader (blue) and follower (orange), reflecting successful synchronization as well as responsiveness over a variety of path complexities.

Experiments in real-world scenarios involved a specially designed UAV platform with onboard sensing, processing, and communications modules. Even in the presence of environmental perturbations like wind, as well as occasional loss of accuracy by the GPS, tracking performance, as well as formation cohesion, was stable. These findings validate robustness and deployability in real-world scenarios of the L-F architecture proposed for cooperative volumetric scanning operations in controlled as well as uncon-

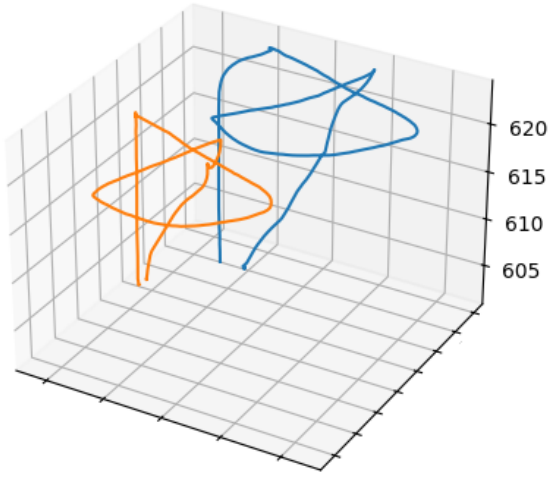

Figure 1: 3D flight trajectories of the leader (blue) and follower (orange) UAVs.

trolled environments.

4.1.1 Latency Estimation in L-F Coordination

Latency in the parallel flight system was analyzed by measuring absolute delay and its components using MAVLink logs of leader and follower UAV positions. The follower's response lag was quantified via mean absolute error (MAE) of speed discrepancies across a range of follower delays.

MAVLink logs were parsed, timestamps converted to UNIX time, and velocity components (VN, VE, VD) derived from the Exogenous Kalman Filter (XKF1). Cubic spline interpolation produced smooth 20ms time-series, enabling precise alignment of leader and follower velocity profiles.

The follower's speed profile was incrementally time-shifted (0–3s in 5ms steps), computing MAE for each shift. As shown in Figure 2, the MAE curve formed a U-shape with a minimum at 1.3s, giving the system's absolute latency.

To validate, total UAV speeds were graphed over the full flight duration. Figure 3 shows the follower's speed mirroring the leader's with a consistent 1.3s lag; peaks and troughs confirm the MAE estimate, visually verifying cooperative flight delay dynamics.

4.1.2 Command Propagation Delay

For further supplement to overall latency evaluation, the time lag from position updates provided by the leader UAV to the follower UAV's resulting action was measured. The time lag, called the command propagation delay, represents the time taken for transmission and translation of movement commands. A specific measurement procedure was used, including

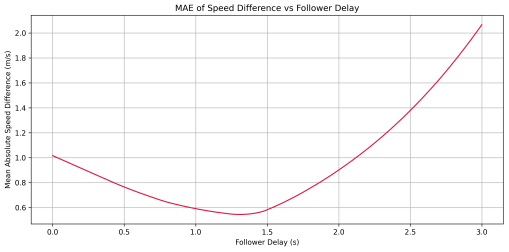

Figure 2: MAE of L-F speed difference across simulated delays.

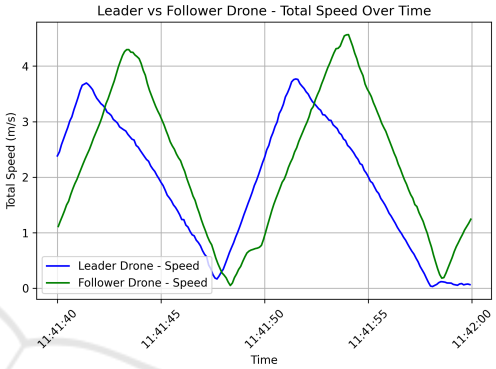

Figure 3: Speed profiles of leader and follower UAVs with observed latency.

noting the time at which the position was requested from the leader and the time at which the follower began its action. The mean delay measured over a number of trials was about 70 milliseconds, indicating a low-order level of latency in receipt of command and its execution.

4.1.3 Position Update Frequency

In addition to latency in execution, the time granularity in the position broadcasting was also examined. Of particular interest was the time interval of adjacent GLOBAL_POSITION_INT messages received by the leader UAV. Inspection of the logging showed that the intervals were consistently 0.07 seconds, equating to an approximately 14 Hz update interval. That's the native capability of the system's motion update mechanism, which is waypoint-based and imposes an upper bound on granularity in motions.

4.1.4 Latency Composition and Remaining Gaps

From the combined analysis, the overall system latency experienced in coordinated UAV flight is about 1.3 seconds. This delay is caused by a variety of factors. Firstly, there is a contribution of about 70 milliseconds from the command propagation delay and a similar contribution of about 70 milliseconds from

the position update interval. These factors account for a total of 140 milliseconds of delay.

This still results in a remaining latency of greater than 1.1 seconds. The unexplained delay could be the result of internal buffering, processing of flight control systems in the UAVs, variable network transmissions, or other conditions such as asynchronous execution of the control loops. The origin of the unexplained latency is left as an open problem for further research, particularly for those applications involving greater coordination or real-time response.

4.2 RL Framework

The RL module enabled autonomous exploration of a bounded 3D world using a high-dimensional voxel state. PPO trained the agent to maximize coverage, avoid collisions, and meet operating constraints in procedurally generated environments. Performance was measured via accumulated reward, reflecting exploration effectiveness and safety.

Efficient policy learning was supported by systematic hyperparameter tuning and large-scale episodic training. The impact of key parameters and overall learning dynamics are detailed in later subsections.

4.2.1 Hyperparameter Optimization

An exhaustive hyperparameter search was conducted to find the optimal policy configuration. The key parameters were tuned across three values each: discount factor $\gamma \in 0.90, 0.95, 0.99$, learning rate $\alpha \in 1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}$, and policy network depth $L \in 2, 3, 4$. Each combination was evaluated over 100 simulation episodes to ensure statistically meaningful comparisons.

The best setup used $\gamma = 0.9$, $\alpha = 1 \times 10^{-4}$, and a two-layer policy network, yielding the highest and most stable rewards. Higher γ values overly favored long-term reward, producing conservative behaviors, while higher learning rates introduced instability and high variance. These results show that for voxel-based, spatially constrained dual-agent search, a low discount factor and conservative learning rate are critical. The strong sensitivity to these parameters highlights the need for careful tuning in complex multi-agent RL systems.

4.2.2 Training Performance and Convergence Behavior

The final policy was trained with PPO over 3,000 episodes in a procedurally generated 3D simulator. Cumulative reward per episode served as the main metric for policy improvement. Despite significant variability from stochastic terrain and randomized spawn points, training revealed a clear upward performance trend.

The logged results trace episodic reward throughout the training horizon. Early episodes tended to show large variability, but when the rewards were smoothed, they exhibited clear increases. For the first 1,000 episodes, average cumulative reward rose from about -60 to greater than 950, indicating that the agent learned basic exploratory behavior. Continued training further refined the policy, and by episode 3,000, the agent was reaching an average reward of 1,300. The slower progressions and additional variability observed in the training were likely due to the environmental randomness inherent in the task, but they further highlighted the robustness of PPO even in adversarial environments. Raw episodic rewards for this set of trials varied greatly, with extreme highs above $3,000$ and extreme lows below $-1,000$. The extreme lows came mainly from unfortunate spawn points in obstructed terrain that limited movement, which also limited exploration. In an open, easily configurable space, this agent would have been able to gather an immense amount of reward.

Despite these fluctuations, the cumulative reward trajectory confirms stable learning and policy convergence. The smoothed graph reflects sustained performance gains and the agent's ability to generalize exploration across diverse conditions, a critical trait for real-world applications with inherent uncertainty and spatial variability.

5 CONCLUSION AND FUTURE OUTLOOK

This research proposed an RL architecture for a leader-follower UAV duo to autonomously and safely explore unfamiliar 3D environments. With PPO and limited action space, complexity was reduced, and efficient training became viable. The supervisory agent was trained to maximize coverage, prevent collisions, and satisfy operational constraints. Training in procedurally generated environments produced diversified, robust policies that transferred to successful real flight tests, verifying policy generalization and the general framework construction. Physical tests also identified that the follower response delay is 1.3s, highlighting synchronization as the major obstacle for the following real-time coordination optimizations. The results also reveal the viability of RL for expansive-scale autonomous surveillance and lay solid ground for intelligent UAV coordination in bounded or unfa-

miliar 3D spaces, including survey, inspection, and search-and-rescue missions. Future work includes scaling to multi-agent scenarios for addressing distributed decision-making and in-real-world tests, such as in agriculture, wildfire, or environmental surveillance.

ACKNOWLEDGEMENTS

This work was supported by FFG as part of the "SpecDrone" project (Application ID: 49003622) and by Interreg programme Bayern-Austria 2021-2027 with co-financing from the European Union as part of the "AI4GREEN: Data Science for Sustainability" (BA0100172) project.

REFERENCES

Bialas, J., Doeller, M., and Kathrein, R. (2023). Robust multi-agent coverage path planning for unmanned aerial vehicles (uavs) in complex 3d environments with deep reinforcement learning. In *2023 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1–6. IEEE.

Bulka, E., He, C., Wehbeh, J., and Sharf, I. (2022). Experiments on collaborative transport of cable-suspended payload with quadrotor uavs. In *2022 International Conference on Unmanned Aircraft Systems (ICUAS)*, pages 1465–1473.

Chen, T., Gao, Q., and Guo, M. (2018). An improved multiple uavs cooperative flight algorithm based on leader follower strategy. In *2018 Chinese Control And Decision Conference (CCDC)*, pages 165–169.

Chen, Y. and Deng, T. (2023). Leader-follower uav formation flight control based on feature modelling. *Systems Science & Control Engineering*, 11(1):2268153.

Cross, D. (2023). Uav-to-uav communication establishing a leader-follower formation. In *Wellingt. Fac. Eng. Symp.* Accessed: Nov. 24, 2023.

Filonchyk, M., Peterson, M. P., Zhang, L., Hurynovich, V., and He, Y. (2024). Greenhouse gases emissions and global climate change: Examining the influence of co_2, ch_4, and n_2o. *Science of The Total Environment*, page 173359.

Li, J., Cheng, R., Zhu, J., Tian, Y., and Zhang, Y. (2021). Wireless secure communication involving uav: An overview of physical layer security. In *MATEC Web of Conferences*, volume 336, page 04005. EDP Sciences.

Littman, M. L. (2009). A tutorial on partially observable markov decision processes. *Journal of Mathematical Psychology*, 53(3):119–125.

Liu, Y., Zhang, F., Huang, P., and Zhang, X. (2021). Analysis, planning and control for cooperative transportation of tethered multi-rotor uavs. *Aerospace Science and Technology*, 113:106673.

Mohebbi, M. R., Sena, E. W., Döller, M., and Klinger, J. (2024). Wildfire spread prediction through remote sensing and uav imagery-driven machine learning models. In *2024 18th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pages 827–834. IEEE.

Mohsan, S. A., Othman, N. Q., Li, Y., Alsharif, M. H., and Khan, M. A. (2023). Unmanned aerial vehicles (uavs): Practical aspects, applications, open challenges, security issues, and future trends. *Intelligent Service Robotics*, 16(1):109–137.

Nguyen, T. T., Nguyen, N. D., and Nahavandi, S. (2020). Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. *IEEE Transactions on Cybernetics*, 50(9):3826–3839.

Rafifandi, R., Asri, D. L., Ekawati, E., and Budi, E. M. (2019). Leader–follower formation control of two quadrotor uavs. *SN Appl. Sci.*, 1(6):539.

Sato, R. and et al. (2020). Detection of gas drifting near the ground by drone hovering over: Using airflow generated by two connected quadcopters. *Sensors*, 20(5):Art. no. 5.

Shaw, J. T., Shah, A., Yong, H., and Allen, G. (2021). Methods for quantifying methane emissions using unmanned aerial vehicles: A review. *Philosophical Transactions of the Royal Society A*, 379(2210):20200450.

Six, D., Briot, S., Chriette, A., and Martinet, P. (2018). The kinematics, dynamics and control of a flying parallel robot with three quadrotors. *IEEE Robot. Autom. Lett.*, 3(1):559–566.

Spurny, V., Petrlik, M., Vonasek, V., and Saska, M. (2019). Cooperative transport of large objects by a pair of unmanned aerial systems using sampling-based motion planning. In *2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, pages 955–962.

Tosato, P., Facinelli, D., Prada, M., Gemma, L., Rossi, M., and Brunelli, D. (2019). An autonomous swarm of drones for industrial gas sensing applications. In *2019 IEEE 20th International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, pages 1–6. IEEE.

Walter, V., Staub, N., Franchi, A., and Saska, M. (2019). Uvdar system for visual relative localization with application to leader–follower formations of multirotor uavs. *IEEE Robotics and Automation Letters*, 4(3):2637–2644.

Zhang, D., Duan, H., and Zeng, Z. (2022). Leader–follower interactive potential for target enclosing of perception-limited uav groups. *IEEE Syst. J.*, 16(1):856–867.