BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2026
http://excel.fit.vutbr.cz

Optimization of Multi-Agent Unmanned Aerial Systems Behavior

Filip Jahn

Abstract
Wildfires require rapid, autonomous, and coordinated intervention in dangerous environments. We propose a heterogeneous Unmanned Aerial Vehicle (UAV) swarm simulated using our environment and PyBullet for real physics, utilizing a novel Multi-Agent Proximal Policy Optimization (MAPPO) framework with Cross-Attention communication. By introducing an auxiliary prediction task and curriculum learning, the commander agent successfully learns to navigate and extinguish fires based entirely on scout data. This architecture bridges the gap between simulated multi-agent reinforcement learning and real-world safety-critical deployment.

*xjahnf00@vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

[Motivation] Autonomous aerial firefighting is a highly promising, safety-critical application. Traditional drone operations rely on manual piloting or hardcoded waypoints, which fall short in dynamic, unpredictable environments like wildfires. We need autonomous swarms that can cooperate and adapt on the fly.

[Problem definition] The core challenge is coordinating a *heterogeneous swarm*. Agile quadcopters (Scouts) act as visual sensors, while a heavy fixed-wing aircraft (Commander) carries the payload. The Commander is completely blind to the fire and must rely entirely on a learned communication protocol from the Scouts to navigate.

[Our solution] We tackle this using Centralized Training with Decentralized Execution (CTDE) driven by the MAPPO algorithm. We developed a custom environment that integrates real-world GIS data, complex aerodynamics (thermal updrafts), and a stochastic fire spread model and uses PyBullet environment for physics.

[Contributions] The main contributions are the design of a robust Cross-Attention communication channel, the introduction of a specialized Auxiliary Loss to prevent shortcut learning, and a 4-phase Curriculum Learning pipeline that enables the Commander to achieve pure autonomy.

2. Simulation Environment & The Swarm

As shown in **Section 1** of the poster, the environment bridges the gap between abstract reinforcement learning and real-world physics.

Instead of flat, featureless planes, we parse OpenStreetMap Geographic Information System (GIS) data to generate realistic 2D semantic grids (Figure 3). The fire itself is not static; it spreads dynamically through stacked cellular automata layers representing fuel and moisture (Figure 2). Furthermore, the simulation models dangerous thermal updrafts (Figure 1), which physically impact the drones' aerodynamics, forcing the neural network to learn active stabilization alongside navigation.

Because the swarm is heterogeneous, the agents operate on different temporal scales. The quadcopter Scouts evaluate their environment at a high frequency, while the fixed-wing Commander operates on "macro-actions" (waypoints) to maintain strategic stability.

3. CTDE Methodology & Network Architecture

Section 2 of the poster illustrates the training loop of the swarm's neural networks. As shown in Figure 4 , the system heavily relies on the CTDE paradigm.

During execution, the drones (Actors) rely solely on local observations and swarm messages. However, during the training phase, the value estimation is handled

by centralized **Critics** that are fed the exact, noise-free Privileged Global State. Importantly, our architecture employs independent critics for the Scouts and the Commander. Because the agents operate on vastly different temporal scales (as explained in Section 2), using a single shared critic would mathematically collapse the temporal discounting and advantage estimations (Adv_t).

The diagram also highlights the addition of an **Auxiliary Loss** ($\mathcal{L}_{AUX}$) during the Backpropagation Through Time (BPTT) phase. The Commander must learn to decode the stream of messages from the Scouts to locate the fire. However, training this end-to-end often results in "shortcut learning"—the agent ignores the messages simply to avoid crashing. By explicitly forcing the Commander to predict the fire's relative position, we inject a supervised gradient that teaches the network to actively "listen" to the swarm.

Because learning flight dynamics, water management, and communication decoding simultaneously is difficult problem, we implemented a 4-phase **Curriculum Training** pipeline. We start with heavy assists (explicit compass, auto-drop) and gradually remove them until the agent reaches *Pure Autonomy*, where it relies purely on the learned communication protocol.

4. Training Results & Action Demo

The effectiveness of our protocol is best seen in the swarm's emergent coordination. As shown in **Section 3** of the poster, the Commander agent successfully learns to execute the full operational cycle required for fire suppression without manual intervention or synthetic navigational aids.

Figure 5 illustrates this sequence: (A) the Commander localizes the Scout and itself, adjusting its heading to approach the target; (B) it aligns its trajectory with the Scout's position to prepare for the drop; (C) it executes a precise water drop. Remarkably, the Commander often completes the mission even after the Scout is terminated. Its GRU layer maintains an internal memory of the fire's coordinates, allowing the agent to successfully extinguish the fire while being completely "blind" and without active communication.

To see the system in motion and observe how the Commander dynamically reacts to real-time data from the Scout swarm, please scan the QR Code on the poster, which links to a full video demonstration of the trained policies in the PyBullet simulation.

5. Conclusions

This thesis represents a comprehensive development effort, bridging the gap between theoretical multi-agent research and complex physics-based simulation. We have successfully built a high-fidelity environment that integrates real-world GIS terrain data with a stochastic fire spread model and realistic thermal updraft physics.

Our work demonstrates that combining hierarchical temporal control with a specialized Cross-Attention communication protocol enables heterogeneous swarms to solve high-dimensional coordination tasks. By introducing a critical auxiliary prediction loss and a rigorous 4-phase curriculum learning pipeline, we moved beyond simple navigation to a state of pure autonomy. The resulting system successfully transitions from raw, noisy environmental data to a fully coordinated, strategic firefighting mission. This robust framework proves that multi-agent reinforcement learning, when supported by physics-aware architectures, is a viable solution for complex, safety-critical disaster response.

Future work will scale the swarm for urban protection. We aim to leverage terrain features like natural firebreaks to autonomously optimize resource allocation and firefighting efficiency.

Acknowledgements

I would like to thank my supervisor Ing. Jiří Novák, Ph.D. for his invaluable guidance and technical support.

References

[1] YU, C., VELU, A., VINITSKY, E., GAO, J., WANG, Y. et al. The surprising effectiveness of ppo in cooperative multi-agent games. Advances in neural information processing systems, 2022, vol. 35, p. 24611-24624.

[2] GRIFFITH, J. D., KOCHENDERFER, M. J., MOSS, R. J., MIŠIC, V. V., GUPTA, V. et al. Automated dynamic resource allocation for wildfire suppression. Lincoln Laboratory Journal, 2017, vol. 22, no. 2, p. 38-59.

[3] LOWE, R., WU, Y. I., TAMAR, A., HARB, J., PIETER ABBEEL, O. et al. Multi-agent actor-critic for mixed cooperative-competitive environments. Advances in neural information processing systems, 2017, vol. 30.