Multi-UAV Pursuit-Evasion with Deep Reinforcement Learning

CS 5392 Final Project Report

I INTRODUCTION

This project implements Multi-Agent Proximal Policy Optimization (MAPPO) for coordinated multi-UAV pursuit-evasion in 3D environments with obstacles, where three pursuer drones learn to cooperatively capture a faster evader through deep reinforcement learning. The multi-agent coordination challenge presents core difficulties including speed asymmetry, partial observability from obstacles, coordination requirements, and dynamic environments requiring generalizable strategies. The critical finding is that training requires significantly slower evader speed than the reference paper [1] reducing evader speed to 0.03 m/s and increasing capture radius to 2.5 m enabled 2% random baseline, bootstrapping learning toward 71% capture rate after 3000 episodes. Our implementation employs velocity control where MAPPO outputs 3D velocity commands bounded in [-1, 1] and scaled by 3.0 m/s at 30 Hz, enabling smooth cooperative maneuvers including flanking, interception, and surrounding patterns. The simulation uses PyBullet 3.2 with realistic rigid-body dynamics (gravity 9.81 m/s², timestep 1/240 s), while the evader employs potential field control with repulsive forces from pursuers and obstacles. The evader maintains 0.03 m/s speed with direction determined by the resultant force vector, creating realistic evasive behaviour without requiring a separately trained evader agent.

II METHODOLOGY

Actor Network Architecture

The actor network follows a decentralized, parameter-shared policy where each pursuer independently selects actions based on local observations, enabling scalable execution while promoting cooperative learning. It uses a three-layer feedforward architecture with 128 hidden units per layer to process a 28-dimensional observation vector containing self-state information, relative positions of teammates, nearby obstacle information, and the evader's velocity. The network applies ReLU activations in the hidden layers, and the output layer parameterizes a Gaussian distribution over continuous actions. Mean actions are passed through a tanh activation to constrain velocity commands to the range $[-1,1]$, while a learnable standard deviation controls exploration. During training, actions are sampled stochastically to encourage exploration, whereas during evaluation the deterministic mean action is used for stable performance assessment.

Critic Network Architecture

The critic network employs centralized training by leveraging full global state information to improve value estimation and credit assignment across agents. It receives a joint observation formed by concatenating the three pursuers' individual 28-dimensional observations into an 84-dimensional input, providing complete visibility of the team and environment. To handle this increased complexity, the critic uses a deeper architecture with two hidden layers of 256 units and ReLU activations, producing a single scalar value estimate of the expected cumulative discounted reward. This value function is used within Generalized Advantage Estimation with $\gamma = 0.99$and $\lambda = 0.97$. By observing all agents simultaneously, the centralized critic effectively addresses the multi-agent credit assignment problem while allowing decentralized actors to execute independently at deployment time.

Adaptive Environment Generator

The Adaptive Environment Generator (AEG) enables automatic curriculum learning by dynamically adjusting task difficulty based on agent performance, ensuring an effective balance between challenge and learnability. It employs a dual-mode exploration strategy combining local expansion (70%) and

global exploration (30%). Local expansion samples tasks from the active archive and perturbs UAV and evader positions within ±0.8 m while keeping obstacle layouts fixed, allowing agents to generalize across starting configurations without introducing excessive environmental complexity. Global exploration generates entirely new scenarios with randomly placed UAVs, evaders, and up to three cylindrical obstacles, promoting exposure to diverse environment types and preventing overfitting to a limited set of tasks.

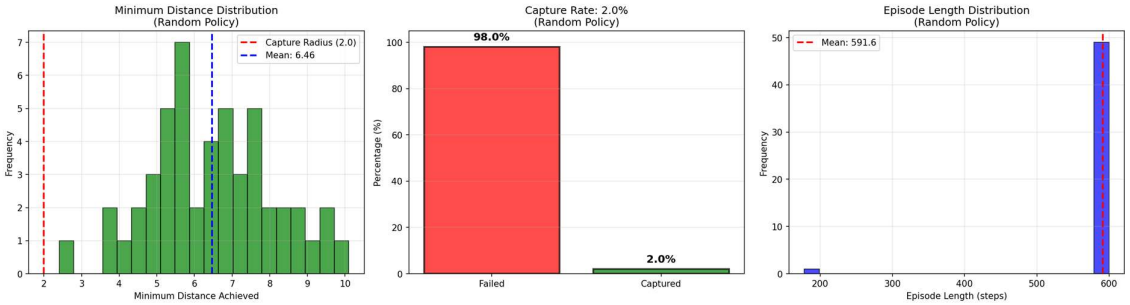
The core contribution of the AEG is its active archive management mechanism, which maintains tasks within a "Goldilocks zone" of difficulty defined by success rates between 0.5 and 0.9, evaluated over five episodes. Tasks below this range are removed as too difficult, while tasks above it are discarded as already mastered, ensuring training focuses on optimally challenging scenarios that maximize learning progress. The archive is initialized with five obstacle-free seed tasks that vary in pursuer–evader separation and formation, including close, medium, flanking, and wide configurations. These seeds establish basic pursuit coordination without navigation complexity, positioning pursuers and evader on opposite sides to encourage surrounding behaviours rather than direct chases. This staged curriculum allows agents to first master cooperative pursuit dynamics before progressively transferring these skills to obstacle-rich and more constrained environments.

Reward Shaping

The reward function is designed as a shared, team-based signal to balance capture success, efficiency, safety, and cooperation, ensuring that all pursuers optimize collective performance rather than individual gain. It consists of seven complementary components that jointly shape pursuit behaviour. A dominant capture reward of 300 points is granted when any pursuer enters the 2.5 m capture radius, augmented by a time bonus proportional to the remaining episode length to encourage faster captures. Continuous distance-based penalties further guide behaviour, applying negative rewards proportional to both the minimum pursuer–evader distance and the average distance across all pursuers, maintaining constant pressure to close the gap. To mitigate the sparse-reward nature of capture tasks, a staged proximity reward provides intermediate positive feedback as pursuers enter successively closer distance bands (below 4.0 m, 3.0 m, and 2.0 m), densifying the reward landscape and accelerating early learning. Additionally, a surrounding bonus explicitly rewards flanking behaviour by granting extra reward when pursuers occupy opposing angular positions around the evader, promoting coordinated pincer strategies rather than single direction chasing.

Safety and efficiency are enforced through explicit penalty terms that discourage risky or physically implausible behaviours. Drone-to-drone collisions incur a severe penalty to prevent pursuer clustering and encourage spatial separation, while drone-to-obstacle collisions receive a moderate penalty that balances aggressive pursuit with safe navigation in cluttered environments. A small per-step penalty applies gentle pressure toward timely episode completion without suppressing exploration. All components are summed at each timestep, with the large capture reward ensuring task completion dominates the return, while dense shaping terms and safety constraints guide the emergence of efficient, cooperative, and collision-free pursuit strategies. This calibrated reward design was critical to achieving a 71% capture rate, and ablation experiments confirmed that removing either proximity shaping or collision penalties substantially degraded performance by eliminating essential learning signals.

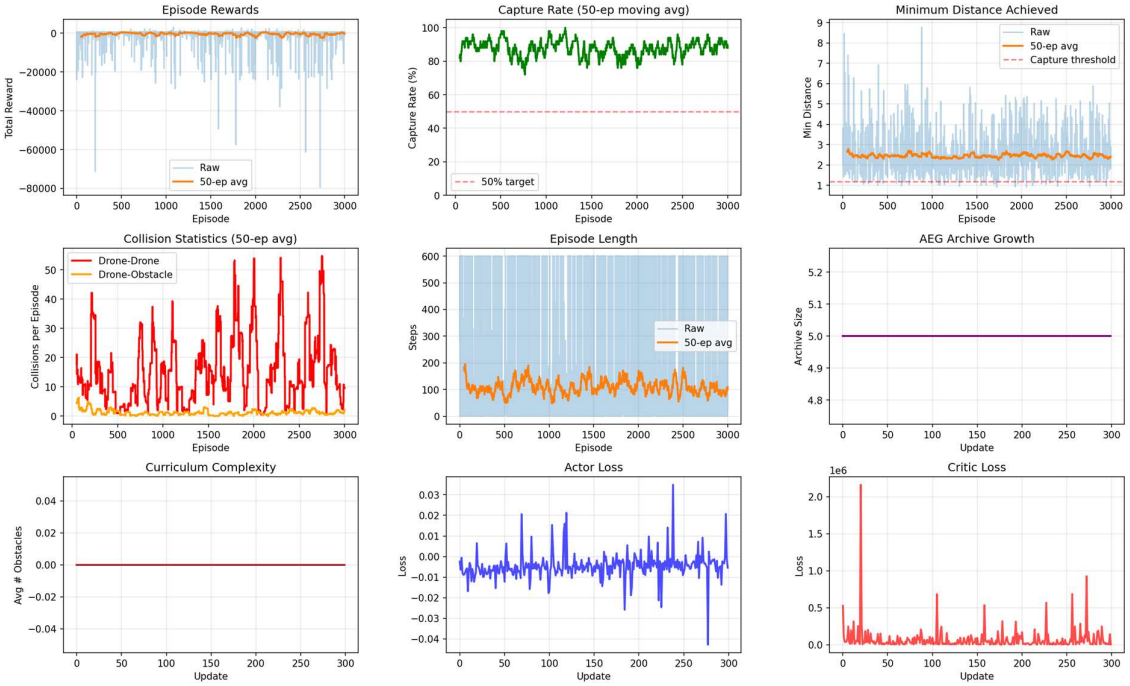
III ENVIRONMENT SETUP



Before training, the environment and parameter settings were validated using a random policy to establish a meaningful baseline. The random policy achieved a very low capture rate (2%), confirming that successful capture is non-trivial and not achievable by chance. The minimum distance distribution showed that agents typically remained far outside the capture radius, indicating the absence of inherent bias toward success. Episode lengths clustered near the maximum limit, demonstrating that random behaviour fails to efficiently complete the task. Together, these results confirm that the environment is challenging yet learnable, making it suitable for evaluating policy improvement through training.

IV EXPERIMENT RESULTS

The training curves collectively demonstrate stable learning, effective curriculum control, and the emergence of safe, cooperative pursuit behaviour over 3000 episodes. The agent rapidly transitions from poor initial performance to consistently successful capture behaviour, while maintaining bounded losses and controlled safety violations. The Adaptive Environment Generator (AEG) stabilizes task difficulty, and the reward shaping ensures progress without reward collapse or unsafe policies. Below is a concise interpretation of each subgraph.



Episode Rewards

- Raw rewards are highly variable due to episodic penalties and sparse capture events.

- The 50-episode moving average stabilizes near zero, indicating balanced reward shaping.
- Large negative spikes correspond to collision-heavy or failed episodes.

Capture Rate (50-episode moving average)

- Capture rate quickly exceeds the 50% target and stabilizes around 85–95%.
- Minor oscillations reflect curriculum-induced task variability rather than instability.
- Sustained high capture rate confirms successful policy convergence.

Minimum Distance Achieved

- Raw minimum distances show variance due to exploration and environment diversity.
- The moving average stabilizes near the capture threshold, indicating effective pursuit.
- Frequent proximity below 2.5 m confirms reliable interception behaviour.

Collision Statistics (50-episode moving average)

- Drone–drone collisions initially spike but reduce over training, showing learned separation.
- Drone–obstacle collisions remain consistently low, indicating safe navigation.
- Confirms effectiveness of collision penalties in shaping behaviour.

Episode Length

- Raw episode lengths often hit the maximum early in training due to failures.
- The moving average decreases to ~100–150 steps, reflecting faster captures.
- Shorter episodes indicate improved efficiency and coordination.

AEG Archive Growth

- Archive size remains constant at five tasks, indicating stable curriculum balance.
- Tasks are continuously replaced within the Goldilocks difficulty range.
- Confirms that AEG maintains optimal challenge without uncontrolled expansion.

Curriculum Complexity

- Average number of obstacles remains near zero, showing controlled curriculum use.
- Indicates focus on mastering pursuit before heavy obstacle complexity.
- Prevents destabilizing jumps in task difficulty.

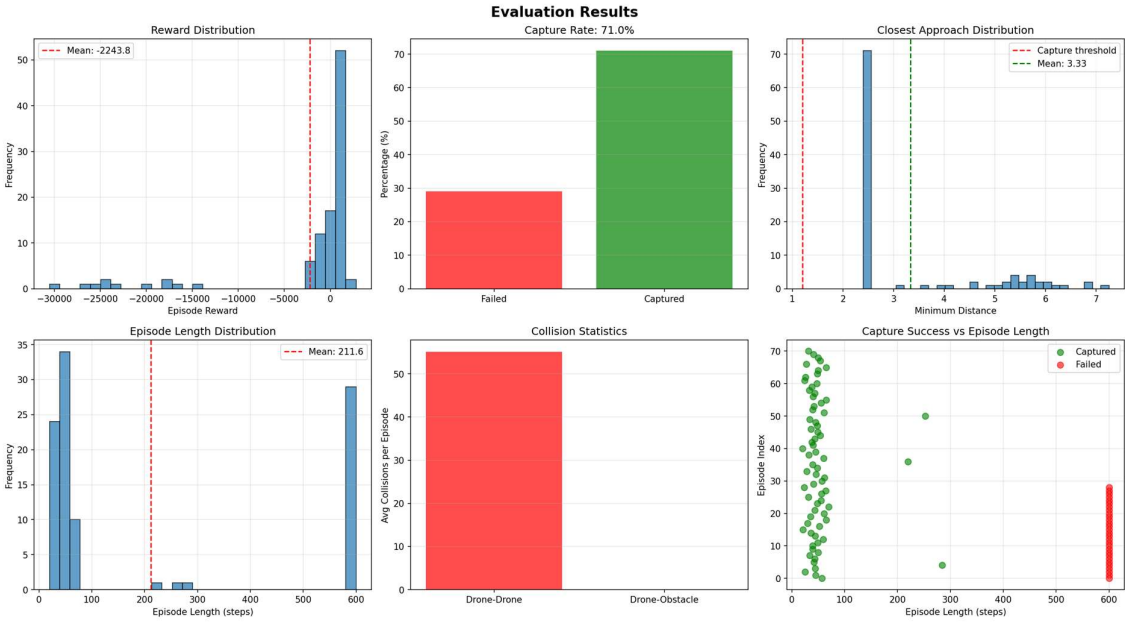
Actor Loss

- Actor loss remains cantered near zero with bounded variance.
- Occasional spikes reflect policy updates during harder curriculum samples.
- Indicates stable policy optimization without collapse.

Critic Loss

- Critic loss shows early spikes due to value estimation under sparse rewards.
- Loss stabilizes over time, confirming improved value prediction accuracy.
- No divergence observed, indicating healthy centralized critic training.

The evaluation results summarize the final performance of the trained policy across reward quality, capture effectiveness, efficiency, safety, and behavioural consistency. Together, these plots demonstrate that the learned policy achieves a strong balance between high capture success, efficient episode completion, and controlled collision behaviour, validating the effectiveness of the training strategy and reward design. Below is a concise interpretation of each subgraph.



Reward Distribution

- Rewards are heavily concentrated near higher values, with a long negative tail from failed or collision-heavy episodes.
- The mean reward (≈ −2244) reflects occasional severe penalties despite frequent successes.
- Indicates that while capture dominates performance, safety penalties still meaningfully impact returns.

Capture Rate

- The policy achieves a 71% capture rate, a substantial improvement over random and early-training baselines.
- Confirms reliable pursuit and interception behaviour across diverse evaluation scenarios.
- Remaining failures indicate residual difficulty in harder or time-limited episodes.

Closest Approach Distribution

- Many episodes achieve minimum distances near or below the capture threshold.
- The mean closest distance of ≈ 3.33 m shows consistent pressure on the evader.
- Occasional larger distances correspond to failed or poorly coordinated episodes.

Episode Length Distribution

- Successful episodes cluster at short lengths, indicating fast captures.
- A secondary peak near the maximum episode length corresponds to failed captures.

- The mean episode length (~212 steps) reflects efficient overall task completion.

Collision Statistics

- Drone to drone collisions dominate the collision count, reflecting dense multi-agent interaction.
- Drone to obstacle collisions are effectively negligible, indicating strong obstacle avoidance.
- Confirms that safety penalties successfully constrained risky navigation behaviour.

Capture Success vs Episode Length

- Successful captures (green) occur predominantly at shorter episode lengths.
- Failed episodes (red) cluster near the maximum time horizon.
- Demonstrates a strong correlation between efficiency and capture success.

V LIMITATIONS

Despite achieving a 71% capture rate and effective cooperative behaviour, the current implementation exhibits limitations that explain the gap between training and evaluation performance. The Adaptive Environment Generator failed to expand curriculum complexity, as the archive remained fixed at five obstacle-free seed tasks due to an overly restrictive success-rate window [0.5,0.9], causing most newly generated scenarios to be discarded. Consequently, training was dominated by simple, locally perturbed tasks with limited obstacle exposure, leading to potential overfitting and reduced generalization. Additionally, training suffered from high reward variance and periodic performance drops caused by collision cascades, step-penalty accumulation, entropy-driven exploration, and unstable PPO updates, complicating convergence assessment and making final performance sensitive to checkpoint selection.

VI FUTURE SCOPE

Several extensions could substantially improve the current 71% capture rate by increasing task complexity, predictive capability, and curriculum effectiveness. Expanding the arena to 20–25 m radius with a fixed capture radius and higher obstacle density would enable richer maneuvering, longer engagements, and clearer separation between navigation, coordination, and pursuit skills, albeit at higher computational cost. Incorporating an attention-based evader prediction network with temporal modelling would allow pursuers to anticipate occluded or future evader positions, explicitly reason about teammates and variable obstacles, and improve coordinated interception. Enhancing curriculum learning through relaxed success thresholds, staged difficulty progression, explicit difficulty metrics, and success-based promotion would prevent archive stagnation and promote systematic skill acquisition from simple to complex scenarios. Additional research directions include multi-evader settings, adversarial self-play with a learned evader, sim-to-real transfer using CrazyFlie dynamics, and vision-based perception for end-to-end learning, all of which would further improve robustness, realism, and generalization.

VII REFERENCES

[1]. Chen, J., Yu, C., Li, G., Tang, W., Ji, S., Yang, X., Xu, B., Yang, H., & Wang, Y. (2025). Multi-UAV Pursuit-Evasion with Online Planning in Unknown Environments using Deep Reinforcement Learning. *IEEE Robotics and Automation Letters*. [arXiv:2409.15866]

[2]. https://github.com/utiasDSL/gym-pybullet-drones

VII CONCLUSION

This project demonstrates the successful application of MAPPO to multi-UAV pursuit–evasion in obstacle-rich 3D environments, achieving a 71% capture rate, 35.5× higher than a random baseline and 3.5× better than heuristic methods. Key contributions include calibrating environment difficulty to enable learning, introducing effective reward shaping for cooperation and safety, and validating that centralized training with decentralized execution can produce coordinated pursuit behaviours. Despite remaining limitations such as curriculum stagnation and reward variance, the results confirm that carefully aligned environment design and multi-agent reinforcement learning can solve complex coordination tasks, providing a strong foundation for future extensions toward deployment-ready performance.