International Journal of Robotics and Control Systems
Vol. 5, No. 5, 2025, pp. 2522-2545
ISSN 2775-2658
http://pubs2.ascee.org/index.php/ijrcs

Dual-Memory Architecture for Robust UAV: Navigation Integrating LSTM and Transformer within a PPO Framework

Maryam Allawi Haddad [a,1,*], Dhayaa Raissan Khudher [b,2]

[a,b] Department of Computer Engineering, University of Basrah, Basrah 61004, Iraq
[1] pgs.maryam.allawi@uobasrah.edu.iq; [2] dhayaa.khudher@uobasrah.edu.iq
* Corresponding Author

ARTICLE INFO

Article history
Received September 21, 2025
Revised November 13, 2025
Accepted November 20, 2025

Keywords
Autonomous Drone Navigation;
Partial Observability;
POMDP;
Memory-Augmented RL;
Trajectory Smoothness

ABSTRACT

Autonomous UAV navigation typically suffers from partial observability (POMDP), where noisy and limited sensing degrades the reliability of decisions. We introduce a dual-memory PPO that augments an LSTM for short-horizon responsiveness with a Transformer for long-horizon context, fused by a learnable gate that adaptively weights both streams end-to-end. Unlike Dual-Transformer PPO and other attention-only variants our model retains recurrent memory and learns the fusion rather than prespecifying it (e.g., concatenation or sum). The observation vector merges normalized proprioceptive and range data the reward balances progress collision penalties and trajectory smoothness with tuned coefficients to avoid dominance. In simulated corridor worlds (with a dynamic variant) the hybrid policy completes 96.5% of episodes 9.7 pp over PPO-LSTM and 17.1 pp over PPO-Transformer while reducing final collisions to 2 per episode, reductions of 37.5% vs PPO-LSTM 64.3% vs PPO-Transformer: 85.7% vs PPO. It converges in 20k episodes (vs 25–29k for baselines), with shorter episodes (150 steps), and greater path efficiency (0.85) than either baseline. Findings are presented as the average plus or minus the standard deviation for all five seeds when $p < 0.05$. Limitations include a simulation-only study and limited environment diversity further, larger-scale environments and fusion and reward design ablations are pending. Overall learnable gating of complementary short- and long-term memories improves reliability under partial observability without compromising on practical training efficiency.

1. Introduction

Recently, the use of drones has become widespread in many applications [1], including search, flood detection [2], rescue [3], surveillance [4], and infrastructure inspection [2], [5]. Transitions from human-in-the-loop piloting to autonomous flight require policies that react under partial observability in a POMDP environment in which the agent observes noisy, incomplete information from sensor constraints and occlusions [6], [7]. This transition can be achieved by two prominent approaches: the adoption of classical control methods (system modelling) [8] or utilizing Deep Reinforcement Learning (DRL) [9], [10], [11]. There are three main aspects for selecting an appropriate DRL algorithm: training stability, convergence speed, and sampling efficiency [12], Further, the ease of algorithm implementation for a specific application and its capacity to generate continuous action