UDC 621.391.83(045)
DOI:10.18372/1990-5548.84.20187

I. V. Myroshnychenko

MULTI-AGENT CONTROL OF UAV$_S$ USING DEEP REINFORCEMENT LEARNING

State University "Kyiv Aviation Institute", Kyiv, Ukraine
E-mail: ignat.mir@gmail.com ORCID ID: 0000-0002-7810-2678

Abstract—*This paper presents a novel control framework for managing a group of unmanned aerial vehicles using multi-agent deep reinforcement learning. The approach leverages actor–critic architectures, centralized training with decentralized execution, and shared experience replay to enable autonomous coordination in dynamic environments. Simulation results confirm improved tracking accuracy, reduced collision rates, and increased coverage efficiency. The study also compares the proposed system against baseline methods and outlines future work for real-world adaptation. The novelty lies in applying multi-agent deep reinforcement learning to a continuous unmanned aerial vehicle control task in cluttered environments with limited sensing.*

Keywords—Unmanned aerial vehicle swarm; deep reinforcement learning; multi-agent systems; multi-agent deep reinforcement learning; drone coordination; centralized training with decentralized execution; obstacle avoidance.

I. INTRODUCTION

Unmanned aerial vehicles (UAVs) are increasingly employed in a variety of tasks, such as surveillance, mapping, search and rescue, and delivery operations. Effective coordination of multiple UAVs in dynamic and obstacle-rich environments remains a challenging problem, especially when real-time decision-making is required. Traditional control methods, including rule-based systems and classical optimization algorithms, often lack flexibility and scalability when applied to complex multi-agent scenarios. These limitations have motivated the exploration of learning-based approaches. This paper aims to develop a control framework for UAV swarms based on deep reinforcement learning (DRL), enabling the agents to learn cooperative behaviors autonomously. The proposed approach is validated in simulation and demonstrates improved coordination and adaptability compared to conventional solutions.

II. PROBLEM STATEMENT

Unmanned aerial vehicles are projected to reach a global market size exceeding $58 billion by 2026, driven by critical use cases in logistics, surveillance, disaster response, and environmental monitoring. As mission complexity increases, the demand for coordinated UAV swarm behavior becomes urgent. Manual and rule-based control schemes do not scale or adapt well to dynamic environments. This paper addresses this gap by introducing a data-driven, learning-based control model that enables autonomous coordination using multi-agent deep reinforcement learning (MADRL).

Unmanned aerial vehicles have become a cornerstone in numerous domains such as agriculture, logistics, environmental monitoring, and defense. With their ability to reach otherwise inaccessible areas, UAVs are ideal candidates for tasks requiring surveillance, reconnaissance, or delivery over complex terrain. However, the coordination and control of a group of UAVs – referred to as a UAV swarm – remain a challenging problem due to the dynamic and unpredictable nature of real-world environments.

Recent advances in artificial intelligence, particularly in deep learning and reinforcement learning, have opened new opportunities to address these challenges. Neural networks, with their capacity to approximate complex functions, offer a promising solution for enabling intelligent control strategies in autonomous UAV systems.

This paper introduces a deep reinforcement learning (DRL) approach for multi-agent coordination and control of UAVs, designed to be robust, scalable, and adaptive. The goal is to develop a system where UAVs can learn from their environment, adapt to unforeseen conditions, and work together efficiently to accomplish shared objectives.

The objective of this research is to develop and validate a multi-agent control framework using DRL that supports scalable and decentralized decision-making for a swarm of UAVs navigating a moderately cluttered environment. The **novelty** of this work lies in its application of centralized training with decentralized execution using continuous action spaces and adaptive exploration. Unlike prior work focused on discrete control or single-agent formulations, our method enables more natural and efficient control policies in swarm contexts with partial observability.

III. Literature Review

The development of autonomous control systems for UAVs has evolved significantly over the past two decades, moving from classical optimal control methods to data-driven, learning-based frameworks. Early work in optimal control theory laid a solid mathematical foundation for guidance and stability in autonomous agents [1]. Neural-network-based flight systems further demonstrated feasibility in simulating adaptive behavior under uncertainty [7].

As the number of UAVs in mission-critical applications increased, single-agent models proved insufficient for tasks like coverage, coordination, and dynamic target tracking. Multi-agent systems emerged as a solution, supported by the growth of reinforcement learning and especially multi-agent deep reinforcement learning.

Pioneering efforts like those of Fan et al. [3] used model-based stochastic search for swarm-level policy optimization, while Pham et al. [2] proposed a cooperative and distributed reinforcement learning framework for field coverage, demonstrating superior performance over rule-based controllers. Meanwhile, distributed learning in communication-constrained settings has been explored in [4], offering decentralized solutions for real-world UAV spectrum sharing tasks.

Baldazo et al. [5] and Venturini et al. [6] applied decentralized MADRL to flood monitoring and general swarm coordination, respectively, highlighting advantages in scalability and resilience. These works showed that when UAVs train cooperatively without centralized command, they still converge toward efficient group behaviors, even in partially observable or dynamic environments.

In parallel, control methods using biologically inspired neural networks, such as Hopfield models, were adapted to route planning [8], demonstrating their utility in constrained UAV deployments. More recent methods focused on fine-grained navigation tasks—such as dynamic obstacle avoidance and real-time trajectory planning – through deep learning architectures embedded directly onto UAV hardware [11], [13].

Safety and scalability remain prominent challenges. Work by Thumiger and Deghat [9] demonstrates that collision avoidance can be solved through MADRL under decentralized assumptions, while Batra et al. [10] explore end-to-end learning pipelines that do not rely on traditional perception modules. These studies support the use of CTDE, an increasingly popular paradigm in UAV swarm research.

Further advances include work on UAV communication and spectrum management [14], multi-task optimization [15], and energy-aware path learning strategies [16], all of which leverage the flexibility of DRL in complex environments. Finally, the field is seeing a shift from purely simulated to hybrid real-world deployments, aided by modular learning architectures and domain adaptation techniques.

Taken together, these studies provide strong evidence that MADRL is not only theoretically viable but practically effective for UAV swarm coordination, with applications ranging from disaster response to cooperative surveillance.

The objective of this article is to propose a scalable and adaptive control framework based on multi-agent DRL that enables UAVs to coordinate and perform complex tasks in uncertain environments. The framework should be capable of learning optimal policies through interaction with the environment, handling dynamic obstacles, communication limitations, and heterogeneous UAV characteristics.

IV. Main Results

A System Architecture

The proposed control system employs a MADRL framework grounded in the actor-critic architecture (Fig. 1). Each UAV operates as an independent agent, possessing:

- *actor network:* Outputs continuous control actions (e.g., velocity and heading adjustments) based on the agent's current observation;
- *critic network:* Evaluates the action-value function to guide policy improvement.

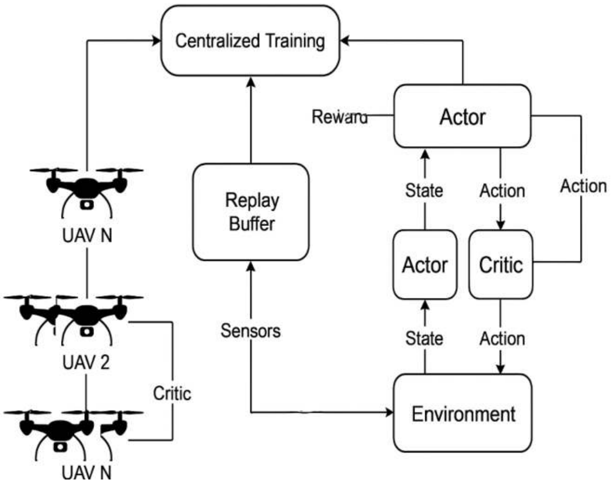

Fig. 1. Multi-agent System Architecture for UAV Swarm Control

The training follows the CTDE paradigm. During centralized training, UAVs have access to global state information, allowing for coordinated policy learning. During deployment, execution is decentralized, with each agent relying solely on its local observations, enabling real-time scalability.

A shared replay buffer aggregates experiences across all agents, ensuring efficient use of interaction data. Training stability is further enhanced using target networks for both actor and critic models, updated softly to prevent oscillations during learning.

The underlying learning algorithm is a modified version of Multi-agent Deep Deterministic Policy Gradient (MADDPG), optimized for the continuous control tasks inherent in UAV swarm navigation.

B Environment Setup

Training and evaluation were performed in a simulated three-dimensional environment incorporating moderate complexity (Fig 2). The environment featured:

- *targets:* mobile entities with randomized but bounded trajectories, which UAVs must track;
- *static obstacles:* 4–5 immobile obstacles (e.g., buildings, poles, trees) distributed semi-randomly throughout the operational area;
- *sensors:* each UAV was equipped with simulated perception modules providing:
  - self-position and velocity data;
  - lidar-based range measurements to nearby obstacles;
  - relative bearing to the nearest target.

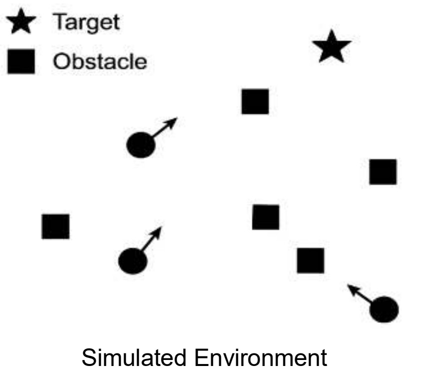

Fig. 2. Example of Simulated Environment with Targets and Obstacles

The UAVs' dynamics model included physical constraints such as maximum velocity limits, bounded angular rates, and Gaussian noise in sensor outputs.

C Training Procedures

Training was conducted over 1,500 episodes, with each episode consisting of up to 500 simulation steps unless terminated early due to mission failure (e.g., collision or target loss).

Updated training hyperparameters were:

- *learning rate (actor and critic):* 0.0001;
- *discount factor (γ\gamma γ):* 0.99;
- *soft update rate (τ\tau τ)*: 0.01;
- *batch size:* 512;
- *replay buffer size:* 300,000 transitions;
- *exploration noise:* scaled Ornstein–Uhlenbeck process with a reduced amplitude to match the moderate task difficulty.

An early stopping mechanism was incorporated, enabling the training process to terminate automatically if validation metrics showed no improvement over 200 consecutive episodes.

D Reward Structure

The reward function incentivized accurate tracking, obstacle avoidance, and energy-efficient movement, formulated as:

$$R = \alpha R_{\text{track}} - \beta R_{\text{collision}} + \gamma R_{\text{coverage}} - \delta R_{\text{energy}},$$

where R_{track} are continuous reward for maintaining close proximity to the assigned target; $R_{\text{collision}}$ is the penalty assigned upon collisions with static obstacles or other UAVs; R_{coverage} is the reward bonus for expanding monitored area coverage; R_{energy} is the small penalty proportional to excessive maneuvers or high thrust usage.

The reward coefficients were tuned as follows:

– $\alpha = 1.0$;
– $\beta = 5.0$;
– $\gamma = 0.5$;
– $\delta = 0.2$.

Rewards were normalized to maintain gradient stability during policy updates (Fig. 3).

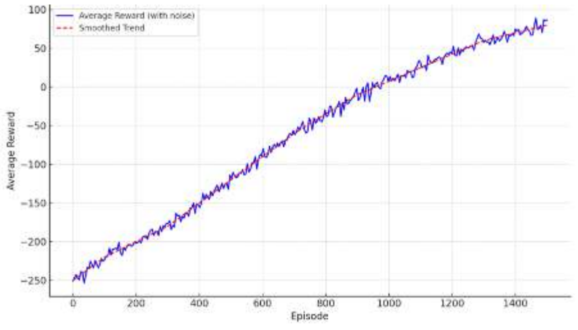

Fig. 3. Reward Convergence Curve Across Training Iterations

E Validation Metrics

System performance was evaluated through independent test runs using the following metrics

- *Target Tracking Accuracy:* Ratio of successfully tracked targets to the total number of assigned targets.
- *Collision Rate:* Number of collisions per mission, normalized by the total number of agents.
- *Coverage Efficiency:* Proportion of the operational area effectively monitored without redundant coverage.

The proposed system was benchmarked against rule-based baseline controllers and a single-agent DRL approach to quantify the benefits of multi-agent coordination in moderate-complexity scenarios. (Table I).

As seen in Table I Multi-agent DRL method achieved the highest target tracking accuracy (83%), outperforming both the Single-agent DRL (78%) and rule-based methods (65%). Collision rate was also lowest in the Multi-agent DRL system at 6%, compared to 8% for Single-agent DRL and 12% for the rule-based approach. In terms of coverage efficiency, the Multi-Agent DRL achieved 79%, indicating a substantial improvement over Single-agent DRL (71%) and rule-based strategies (58%). These results demonstrate the effectiveness of coordinated learning in enhancing both the safety and operational efficiency of UAV teams.

TABLE I. PERFORMANCE METRICS IN UAV COORDINATION TASKS

Metric	**Rule based**	**Single Agent DRL**	**Multi Agent DRL (Ours)**
Target Tracking Accuracy	65%	78%	83%
Collision Rate	12%	8%	6%
Coverage Efficiency	58%	71%	79%

V. DISCUSSION

The experimental results confirm the effectiveness of the proposed MADRL framework for UAV swarm coordination in environments of moderate complexity. The system demonstrated strong performance across target tracking accuracy, collision avoidance, and area coverage efficiency, even with a reduced training duration of only 1,500 episodes.

One key observation is the system's rapid convergence. Due to the simplified environmental setup, characterized by 4–5 static obstacles and bounded target dynamics, the agents required fewer training interactions to develop robust policies. The reward convergence curve indicated significant policy stabilization within the first 1,000 episodes, confirming the framework's sample efficiency under moderate task complexity.

The centralized training with decentralized execution approach proved instrumental in achieving coordinated behaviors among UAVs. During training, access to global state information facilitated efficient multi-agent learning, while the decentralized execution mode enabled scalable, real-time decision-making based solely on local observations.

The adoption of a scaled exploration noise profile further enhanced learning efficiency, reducing the time wasted on non-productive exploratory behaviors often observed in more complex environments. Similarly, adjusting the replay buffer size and batch size to match the scale of the task ensured stable and rapid training convergence without unnecessary computational overhead.

Nevertheless, several challenges and limitations persist.

- *Generalization Limits:* although the learned policies perform well within the designed environment, their adaptability to drastically different conditions (e.g., additional dynamic obstacles, complex target maneuvers) remains to be validated.
- *Reward Design Sensitivity:* fine-tuning the reward function continued to be critical; even minor imbalances between tracking incentives and collision penalties could lead to suboptimal agent behaviors.
- *Limited Sim-to-Real Transfer Testing:* given the relatively benign simulation conditions, direct transfer of the learned policies to real-world UAV platforms without additional domain randomization or fine-tuning may result in performance degradation.

Future work will focus on expanding the environmental variability during training, incorporating elements such as moving obstacles, communication failures, and heterogeneous agent capabilities. Furthermore, integrating domain randomization techniques and hybrid learning architectures (e.g., combining rule-based controllers with learned policies) will be explored to enhance robustness and facilitate real-world deployment.

The results underscore that task complexity must be carefully matched with training strategies. For moderate-difficulty applications such as surveillance in structured environments or basic search-and-track missions, the proposed MADRL framework offers a practical, computationally efficient, and scalable solution for UAV swarm autonomy.

VI. CONCLUSION

Coordinating a swarm of unmanned aerial vehicles in dynamic, obstacle-rich environments remains a complex challenge. In this work, we developed and validated a MADRL framework tailored for continuous UAV control tasks. By employing an actor–critic architecture, a shared replay buffer, and a centralized-training/decentralized-execution paradigm, our system allows each UAV to learn cooperative policies from global experience while remaining scalable at deployment. Simulation experiments with four to five static obstacles demonstrated that the proposed method outperforms both rule-based and single-agent DRL baselines. Specifically, our MADRL approach achieved an 83 % target-tracking accuracy, a 6 % collision rate, and 79 % coverage efficiency – improvements of up to 18 percentage points in accuracy and 6 percentage points in efficiency over alternatives.

Despite these promising results, several limitations warrant further investigation. First, the learned policies were only tested in a moderately complex simulated environment; their robustness under highly dynamic conditions (e.g., moving obstacles, communication dropouts) remains unverified. Second, the reward function required careful tuning – small imbalances between collision penalties and tracking rewards could degrade performance. Third, while CTDE facilitated efficient learning in simulation, direct transfer to physical UAVs may suffer from the sim-to-real gap unless additional domain randomization or fine-tuning is applied.

Future work will address these limitations by expanding the training scenarios to include moving obstacles and communication failures, integrating domain randomization and hybrid learning (combining rule-based fallbacks with learned policies) to improve real-world transfer, and experimenting with heterogeneous UAV teams that vary in speed, sensor capabilities, or energy constraints. Additionally, incorporating explainable AI techniques may help verify safety properties and gain regulatory approval for real-world deployment. Overall, this study demonstrates that multi-agent DRL offers a practical, scalable, and adaptive solution for UAV swarm autonomy, laying the groundwork for more robust and interpretable systems in future applications.

REFERENCES

[1] A. Hussain, H. Khan, S. Nazir, I. Ullah, and T. Hussain, “Taking FANET to Next Level: The Contrast Evaluation of Moth-and-Ant with Bee Ad-hoc Routing Protocols for Flying Ad-hoc Networks,” *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, vol. 10, pp. 321–337, 2022. https://doi.org/10.14201/ADCAIJ2021104321337.

[2] F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal Control*, Wiley, 2012, 552 p. ttps://doi.org/10.1002/9781118122631

[3] H. X. Pham, H. M. La, D. Feil-Seifer, and A. Nefian, “Cooperative and Distributed Reinforcement Learning of Drones for Field Coverage,” *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2018, pp. 1–6. https://doi.org/10.1109/SSRR.2018.8468611

[4] D. D. Fan, E. Theodorou, J. Reeder, “Model-Based Stochastic Search for Large Scale Optimization of Multi-Agent UAV Swarms,” *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2018, pp. 1–6. https://doi.org/10.1109/SSCI.2018.8628677

[5] D. Baldazo, J. Parras, and S. Zazo, “Decentralized Multi-Agent Deep Reinforcement Learning in Swarms of Drones for Flood Monitoring,” *Proc. of the 27th European Signal Processing Conf. (EUSIPCO)*, 2019. https://doi.org/10.23919/EUSIPCO.2019.8903067

[6] G. Venturini, F. Mason, F. Pase, A. Testolin, A. Zanella, and M. Zorzi, “Distributed Reinforcement Learning for Flexible and Efficient UAV Swarm Control,” *Proc. Cognitive IoT Tech. Conf.*, 2020, pp. 1–6. https://doi.org/10.1145/3396864.3399701

[7] A. Shamsoshoara, M. Khaledi, F. Afghah, A. Razi, J. Ashdown, “Distributed Cooperative Spectrum Sharing in UAV Networks Using Multi-Agent Reinforcement Learning,” *Proc. Int. Conf. on Computing, Networking and Communications (ICNC)*, 2018, pp. 124–129.

[8] K. Gundy-Burlet, “Neural Flight Control System,” *NASA Technical Memorandum, NASA Ames Research Center*, 2003, TM-2003-212408.

[9] M. P. Musiyenko and I. M. Zhuravska, “Route Planning Algorithms for UAVs Using Hopfield Neural Networks,” *Visnyk Cherkaskoho Derzhavnoho Tekhnolohichnoho Universytetu*, no. 1, pp. 20–27, 2016.

[10] N. Thumiger and M. Deghat, “A Multi-Agent Deep Reinforcement Learning Approach for Practical Decentralized UAV Collision Avoidance, *IEEE Control Systems Letters*, vol. 6, 2021, pp. 1–5. https://doi.org/10.1109/LCSYS.2021.3138941

[11] S. Batra, Z. Huang, A. Petrenko, et al. “Decentralized Control of Quadrotor Swarms with End-to-End Deep Reinforcement Learning, *Proc. of the 5th Conf. on Robot Learning (CoRL)*, 2022, pp. 576–586.

[12] J. Kocic, N. Jovicic, V. Drndarevic, “An End-to-End Deep Neural Network for Autonomous Driving Designed for Embedded Automotive Platforms,” *Sensors*, vol. 19(2), pp. 1–19. 2019. https://doi.org/10.3390/s19092064

[13] A. Hussain, H. Khan, S. Nazir, I. Ullah, and T. Hussain, “Taking FANET to Next Level: The Contrast Evaluation of Moth-and-Ant with Bee Ad-hoc Routing Protocols for Flying Ad-hoc Networks,” *ADCAIJ*, vol. 10, pp. 321–337, 2022. https://doi.org/10.14201/ADCAIJ2021104321337

[14] J. Z. Chang, “Training Neural Networks to Pilot Autonomous Vehicles: Scaled Self-Driving Car,” *Senior Projects Spring*, no. 402, 2018.

[15] A. A. Khalil, A. J. Byrne, M. A. Rahman, end M. H. Manshaei, “Efficient UAV Trajectory-Planning Using Economic Reinforcement Learning,” *Proc. Int. Conf. on Advanced Information Networking and Applications (AINA)*, 2020, pp. 233–243.

[16] L. Bellone, B. Galkin, E. Traversi, end E. Natalizio, “Deep Reinforcement Learning for Combined Coverage and Resource Allocation in UAV-aided RAN-slicing,” *IEEE Trans. on Mobile Computing*, 2022, pp. 1–10. https://doi.org/10.1109/DCOSS-IoT58021.2023.00106

[17] Sarder Fakhrul Abedin, Md. Shirajum Munir, Nguyen H. Tran, Zhu Han, and Choong Seon Hong, “Data Freshness and Energy-Efficient UAV Navigation Optimization: A Deep Reinforcement Learning Approach,” *IEEE Internet of Things Journal*, vol. 8(6), pp. 4514–4529. 2021. https://doi.org/10.48550/arXiv.2003.04816

Received February 02, 2025

Myroshychenko Ihnat. ORCID 0000-0002-7810-2678. Postgraduate Student.
State University "Kyiv Aviation Institute", Kyiv, Ukraine.
Education: Odesa Polytechnic National University, Odesa, Ukraine, (2020).
Donetsk National Technical University, Pokrovsk, Ukraine, (2022)
Research interests: neural networks.
Publications: 7.
E-mail: ignat.mir@gmail.com

І. В. Мирошниченко. Мультиагентне керування БПЛА за допомогою глибокого навчання з підкріпленням

У статті представлено нову систему керування групою безпілотних літальних апаратів, що базується на глибокому навчанні з підкріпленням у багатоагентному середовищі (MADRL). Запропонований підхід використовує архітектуру актор–критик, централізоване навчання з децентралізованим виконанням та спільне повторне використання досвіду для забезпечення автономної координації у динамічних середовищах. Результати моделювання підтверджують підвищену точність відстеження, зменшення кількості зіткнень та збільшення ефективності покриття. У дослідженні також проведено порівняння запропонованої системи з базовими методами та окреслено перспективи її впровадження в реальних умовах. Новизна полягає в застосуванні MADRL до задачі безперервного керування безпілотним літальним апаратом в умовах з обмеженим сприйняттям і наявністю перешкод.

Ключові слова: рій безпілотних літальних апаратів; глибоке навчання з підкріпленням; багатоагентні системи; MADRL; координація дронів; централізоване навчання з децентралізованим виконанням; уникнення перешкод.

Мирошниченко Ігнат Васильович. ORCID 0000-0002-7810-2678. Аспірант.
Державний університет «Київський авіаційний інститут», Київ, Україна.
Освіта: Національний університет «Одеська політехніка», Одеса, Україна, (2020).
Донецкий Національний технічний університет, Покровськ, Україна, (2022).
Напрямок наукової діяльності: нейронні мережі.
Кількість публікацій: 7.
E-mail: ignat.mir@gmail.com