Check for updates

Comprehensive Review of Reinforcement Learning for Autonomous Drone Systems: Algorithms, Simulation, and Deployment Challenges

Yogesh S. Dighe[1,*], R.A. Kapgate[1], P. M. Patare[1] and Vishant Kumar[2]

[1] Sanjivani College of Engineering, Kopargaon, Maharashtra, India
[2] Parul Institute of Technology, Parul University, Vadodara, Gujarat, India
yogeshin2009@gmail.com

Abstract. Reinforcement learning (RL) is proving to be a significant enabler of intelligent autonomy in UAVs, such that perception, decision-making, and control policies can be learned within drones through their interactions with complex and uncertain environments. Overall, while traditional model-based control methods rely on accurate system models, RL offers an alternative means of realising adaptive behaviour in the context of experience optimisation. However, recent advances in Deep Reinforcement Learning (DRL) and multi-agent reinforcement learning (MARL) have significantly expanded the use of RL for various complex UAV tasks such as autonomous navigation, obstacle avoidance, trajectory planning, target tracking, agile flight and collaborative swarm.

This work presents a comprehensive systematic review of the reinforcement learning approaches applied to autonomous drone systems. The survey systematically investigates RL fundamentals, common DRL algorithms, MARL paradigms, simulators, simulation-to-real transfer techniques, application domains and deployed issues. Comparative tables are also added to sum up the algorithmic trade-offs, simulator features and application–algorithm relations. We then discuss the critical issues on safety, sample efficiency, computational constraints and evaluation procedures, followed by the potential research directions. This paper is intended to offer a well-organized and easy entry reference for those who are interested in the RL-based UAV autonomy area such as researchers or practitioners.

Keywords: Reinforcement learning; deep reinforcement learning; unmanned aerial vehicles; autonomous drones; multi-agent reinforcement learning; simulation-to-real transfer.

1 Introduction

Unmanned Aerial Vehicles (UAVs), also referred to as drones, have moved from being remotely controlled platforms into systems with higher autonomy that can be used for a wide range of missions like surveillance, infrastructure inspection, precision

© The Author(s) 2026
J. Saxena et al. (eds.), *Proceedings of the International Conference on Sustainable Micro-Nano Materials & Innovative Technology (ICSUMMIT 2026)*, Atlantis Highlights in Engineering 44,
https://doi.org/10.2991/978-94-6239-727-9_13

agriculture, disaster response, delivery services and cooperative swarm missions. This rapid increase in the use of UAVs has significantly increased the demand for intelligent autonomy capable of operating robustly in uncertain, nonlinear and dynamic environments.

Due to the complexity of flight dynamics, uncertain aerodynamic disturbances, the variation of payload weight and state estimation errors (e.g., sensor noise, lag compensation) in UAV systems, the robust control problem still remains a challenge [1]. One of the fundamentally difficult problems in conventional model-based control and optimisation approaches seems to be that of how such an accurate mathematical model of the UAV and its environment could ever be achieved or maintained in practice [2]. As a result, classic control methodologies may lack the versatility required when operating in open or variable environments.

Learning approaches such as reinforcement learning (RL) enable to learn an optimal policy that maximises cumulative reward through interacting with the environment [3]. Reinforcement learning allows for model-free, adaptive decision making instead of rule-based or optimisation-based [4]. The integration of deep neural networks with reinforcement learning (RL), referred to as Deep Reinforcement Learning (DRL), has considerably increased the applicability of various RL techniques to high-dimensional sensory input and continuous action space [5].

Recent studies show encouraging results with DRL-enabled UAV navigation and flight control [6]. Reinforcement learning methods have already been implemented for low-level attitude stability [7], trajectory tracking and auto landing [8], collision-free in cluttered environments [9] and energy aware flying, ensuring an infinite duration flight time horizon [10]. It is simulation-based training that makes such advances possible, it enables safe, scalable learning to precede real-world evaluation on physical air assets. [11].

However, the practical implementation of RL-based UAV systems is still constrained by safety issues, lack of sample efficiency, computation constraints and the simulation/reality gap. To this end, the present work focuses on Strategies and Techniques for reinforcement learning in autonomous drones, elaborating these issues through a structured analysis of RL methods for drone systems with applications to algorithms, simulators, platforms, application domains, as well as sim-to-real transfer solutions shared in recent Scopus-indexed ICs

1.1 Fundamentals of Reinforcement Learning for UAV Systems

Reinforcement learning structures the problem of autonomous control as one of online, sequential decision making, where an agent interacts with a system to maximise cumulative reward in the long run [3]. The majority of UAV tasking problems in reinforcement learning are formulated as MDPs, which are composed of state space, action space, transition dynamics (i.e., the system model), reward function and discount factor.

In the context of UAVs, the state space typically consists of the drone pose and velocity, attitude, rate of change of attitude and inertial sensor FDI. [7] In a perception-based task, the state may also contain visual observation from onboard camera or depth sensors. The action space generally includes continuous thrust or torque commands to the motors, although higher-level commands such as velocity or waypoints are also employed.

The mission objectives and operational constraints are encoded into the reward function. In a control task, for example, the reward function penalises instability, tracking error and energy consumption while encouraging smooth flight. For the second navigation and exploration task, their reward consists of distance to goal terms combined with collision penalties and energy consumption components [9]. Reward engineering is a well-known central problem, since reward functions that are specified poorly can cause dangerous behaviours or unintentional convergence to poor policies.

Reinforcement learning (RL) algorithms can be categorized into two major groups, namely model-free and model-based methods. Model-free methods: these algorithms learn policies directly from interaction data; their flexibility and simplicity result in widespread popularity [4]. But they have become dependent on vast amounts of training data. These model-based RL algorithms take advantage of learned or known information with regards to the system dynamics, which could improve sample efficiency, and being data-scarce ultimately is one major consideration in real-world UAV applications. [12] To preserve the robustness, efficiency and real-time feasibility, a hybrid solution combining planning based on models and non–model based learning is increasingly pursued.

2 Methodologies and Adapted Workflow

2.1 Deep Reinforcement Learning Algorithms for Drones

Deep reinforcement learning generalises classical RL by using deep neural networks to value function or policy approximation, leveraging their capability of learning from high-dimensional sensory data. Such capability is a requirement for UAVs operating in visually cluttered and non-standard conditions.

Value-based methods like DQN approximate the action-value function and have been employed towards discrete decision-making problems for UAVs, e.g., way-point selection, simple obstacle avoidance [13]. Nevertheless, the action discretisation will not be applicable for practice in the real-world implementation of UAV flying control since continuous and stable control signals are needed.

Policy-gradient-based method can optimise the policy in continuous action spaces and is more applicable to UAV control [14]. Actor–critic algorithms integrate policy-gradient learning with value estimation to enhance the stability of training and convergence. Low-level UG control has been achieved using DDPG [7], and TD3 makes an improvement in stability by relieving overestimation bias [15]. One of the most widely used techniques is Proximal Policy Optimization (PPO), as it has conservative policy updates and stable training behavior [12]. Even though the Soft Actor-Critic (SAC) has been enhanced through minimising/discriminating an entropy term, maximal-entropy is one of the central mechanisms for improved exploration and robustness in stochastic worlds [16].

The pattern of development for reinforcement learning algorithms in the context of UAV systems is clearly seen to move from discrete value-based methods, into more continuous control (Table 1) — appropriate actor–critic techniques. PPO and SAC are also common for safety-critical and dynamic environments; TD3 is with stability boost for deep continuous low-level fly-control. Thus, performance requirements are to be based on the nature of tasks and safety bounds, along with computational resources must help select appropriate algorithms.

Incorporating multi-agent reinforcement learning (MARL), which extends RL into the realm of collaborative UAV systems. potentially getting us coordination and scalability [17] by centralised training + decentralised execution. The non-stationarity and communication inefficiency problems are still an active area of research in multi-agent reinforcement learning (MARL) problems, e.g comments on environment: formation flying, cooperative navigation, etc.

Table 1. Comparison of Reinforcement Learning Algorithms for UAV Applications.

Algorithm	Learning Paradigm	Action Space	Training Stability	Sample Efficiency	Typical UAV Applications
Q-Learning	Value-based	Discrete	Low	Low	Simple navigation
DQN	Deep value-based	Discrete	Moderate	Moderate	Waypoint selection
DDPG	Actor–Critic	Continuous	Moderate	Moderate	Low-level flight control
TD3	Actor–Critic	Continuous	High	High	Robust UAV control
PPO	Actor–Critic	Continuous / Discrete	Very High	Moderate	Navigation and planning
SAC	Actor–Critic	Continuous	Very High	Very High	Dynamic environments
MARL	Multi-agent RL	Continuous / Discrete	Moderate	Moderate	Swarm coordination

2.2 Simulation Platforms and Training Environments

Given the high-risk and cost of conducting real-time exploration at learning phase, simulation platforms are vital in training reinforcement learning (RL)-based UAV systems [11]. X000D X Simulator-based training is essential as the models of reinforcement learning need to interact millions of times.

Wind is a significant factor in physics-based simulators, which model UAV dynamics and control together with actuator/sensor dynamics [18]. If we want to train our policies and later deploy them on real UAVs, inertia model, thrust limitations, and sensor noise are important tell-tales for our state dynamics.

Vision-based simulators are TPMs that offer photorealistic environments necessary for perception-driven reinforcement learning [19]. These simulators are intended to facilitate end-to-end learning, where the perception and actuation are learned in a joint fashion. Simulation systems with wide bandwidth and high-speed capability also speed up training by allowing for a large volume of data collection [20].

Table 2 summarizes the trade-offs among physics fidelity, visual realism and computational efficiency in widely adopted simulators. Although the more realistic simulator (high-fidelity) is better, simulators with faster run-times allow for quicker policy iteration. For sim-to-real transfer, noise and variation in sensing has a greater impact than increased graphical fidelity.

Table 2. Comparison of Simulation Platforms for RL-Based UAV Training.

Simulator	Type	Physics Fidelity	Sensor Support	Vision Realism	Training Speed
Gazebo	Physics-based	High	IMU, GPS, LiDAR	Low	Moderate
AirSim	Hybrid	High	Camera, IMU	High	Moderate
MuJoCo	Physics-based	Very High	Limited	Low	High
PyBullet	Physics-based	Moderate	Limited	Low	High
Unity ML-Agents	Vision-based	Moderate	Camera	Very High	Moderate

2.3 Mapping Reinforcement Learning Algorithms for UAVs

RL has been used in various UAV applications. The autonomous navigation and obstacle avoidance is one of the well studied area, especially in indoor and GPS-denied scenarios [9]. RL from camera input directly to drone navigation is made possible by vision-based RL [19].

These two skills are similar to those in agile flight [18] and drone racing [13], which both show that RL can generalise to aggressive manoeuvres under tight dynamic constraints [21]. Positioning: In both surveillance and target tracking, reinforcement learning techniques enable adaptive decision-making under uncertainty, which can help maintain optimal relative positioning of drones [22]. MARL helps enable cooperative behaviours like formation flying and distributed sensing systems for drone multi-systems [17]. Energy-aware RL-based methods also exist that optimise the trajectories and control policies to prolong a system's lifetime. [10].

Table 3. Mapping of UAV Applications to Reinforcement Learning Algorithms

UAV Application	Primary Algorithm	Input Modality	Control Level
Autonomous navigation	PPO	State / Vision	High-level
Obstacle avoidance	DQN / PPO	Vision / LiDAR	Mid-level
Agile flight	SAC	State vectors	Low-level
Drone racing	PPO / SAC	Vision	Low-level
Target tracking	DDPG	Vision + State	Mid-level
Formation flying	MARL	Local states	Distributed
Swarm coordination	MARL	Communication	Distributed
Energy optimization	PPO / SAC	Battery + State	High-level

Table 3 illustrates a systematic mapping between the class of UAV applications and the standard reinforcement learning algorithms. Stable policy-gradient methods will help with near-optimal decisions for high-level control tasks, while strong continuous-control at the lowest levels is best aided by robust continuous control algorithms. MARL will indeed be a key component for working with swarming applications, as it can fully encapsulate the requirements of coordination and communication-aware schemes.

2.4 Simulation to Real Transfer

Despite its good performance in the simulation, transferring the trained policy to real UAVs is still a challenging task due to the learning-to-real gap. Domain randomisation overcomes this limitation by randomising visual and environmental properties during training [23]. Dynamics randomisation also increases robustness through distilling away physical parameters, like mass, inertia and motor delays [21].

Hybrid control architectures incorporate reinforcement learning within classical controllers to guarantee stability and safety [24]. In practice, extensive real-world fine-tuning from flight data is still necessary to correct the residual modelling errors and further enhance the performance [25]. It has been found that successful sim-to-real transfer usually involves multiple such techniques rather than relying on a single method.

3 Methodologies and Adapted Workflow

3.1 Challenges and Future Directions

Key challenges were ensuring safety during learning and deployment[11], improving sample efficiency[12], addressing the onboard computing[26], and standardizing benchmarks of these methods. We encourage such work to move forward in the community, and suggest a few promising directions and possible tasks for sharing:

1) Safe and constraint-weighed Reinforcement Learning (RL),

2) Explainable Markov Decision Process,

3) Rapid adaptation optimised meta-learning, and

4) Standardised benchmark environments/effects.

3.2 Conclusion

In this paper, a wide review of reinforcement learning in the context of an autonomous drone system. The paper systematically introduces the architectures and methods, including algorithms, simulation platform, application domain, sim-to-real transfer strategy and deployment obstacle with a layered perspective of international conference aspects. Reinforcement learning is a promising way of achieving adaptive UAV autonomy, and more work still needs to be done to prove safe, reliable and scalable in real-world scenarios.

Disclosure of Interest: The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Achtelik, M., Doth, K.-M., Gurdan, D., Stumpf, J.: Design of a Multi-Rotor MAV with regard to Efficiency, Dynamics and Redundancy. In: AIAA Guidance, Navigation, and Control Conference 2012. Springer, Heidelberg (2012)
2. Hwangbo, J., et al.: Control of a Quadrotor With Reinforcement Learning. IEEE Robotics and Automation Letters 2(4), 2096–2103 (2017)
3. Sutton, R.S.: Learning to predict by the methods of temporal differences. Machine Learning 3(1), 9–44 (1988)
4. Kober, J., Bagnell, J.A., Peters, J.: Reinforcement learning in robotics: A survey. The International Journal of Robotics Research 32(11), 1238–1274 (2013)
5. Li, Y.: Deep Reinforcement Learning: An Overview. arXiv (Cornell University) (2017)
6. Mnih, V., et al.: Human-level control through deep reinforcement learning. Nature 518(7540), 529–533 (2015)
7. Lillicrap, T., et al.: Continuous control with deep reinforcement learning. arXiv (Cornell University) (2015)
8. Tai, L., Paolo, G., Liu, M.: Virtual-to-real deep reinforcement learning: Continuous control of mobile robots for mapless navigation. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 31–36. IEEE, Piscataway (2017)
9. Sadeghi, F., Levine, S.: CAD2RL: Real Single-Image Flight Without a Single Real Image. In: Robotics: Science and Systems XIII. MIT Press, Cambridge (2017)
10. Dorling, K., Heinrichs, J., Messier, G.G., Magierowski, S.: Vehicle Routing Problems for Drone Delivery. IEEE Transactions on Systems, Man, and Cybernetics: Systems 47(1), 70–85 (2016)
11. Shah, S., Dey, D., Lovett, C., Kapoor, A.: AirSim: High-Fidelity Visual and Physical Simulation for Autonomous Vehicles. In: Springer Proceedings in Advanced Robotics, pp. 621–635. Springer, Cham (2017)
12. Deisenroth, M.P., Rasmussen, C.E.: PILCO: A Model-Based and Data-Efficient Approach to Policy Search. In: Proceedings of the 28th International Conference on Machine Learning, pp. 465–472. Omnipress, Madison (2011)
13. van Hasselt, H., Guez, A., Silver, D.: Deep Reinforcement Learning with Double Q-Learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30, pp. 2094–2100. AAAI Press, Palo Alto (2016)

14. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning 8, 229–256 (1992)
15. Fujimoto, S., van Hoof, H., Meger, D.: Addressing Function Approximation Error in Actor-Critic Methods. arXiv (Cornell University) (2018)
16. Haarnoja, T., Zhou, A., Abbeel, P., Levine, S.: Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. arXiv (Cornell University) (2018)
17. Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., Mordatch, I.: Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. In: Advances in Neural Information Processing Systems 30, pp. 6379–6390. Curran Associates, Red Hook (2017)
18. Koenig, N., Howard, A.: Design and use paradigms for Gazebo, an open-source multi-robot simulator. In: 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems, vol. 3, pp. 2149–2154. IEEE, Piscataway (2004)
19. Dosovitskiy, A., Ros, G., Codevilla, F., López, A.M., Koltun, V.: CARLA: An Open Urban Driving Simulator. In: Proceedings of the 1st Annual Conference on Robot Learning, pp. 1–16. PMLR (2017)
20. Andrychowicz, M., et al.: Hindsight Experience Replay. In: Advances in Neural Information Processing Systems 30, pp. 5048–5058. Curran Associates, Red Hook (2017)
21. Kaufmann, E., Bauersfeld, L., Loquercio, A., Müller, M., Koltun, V., Scaramuzza, D.: Champion-level drone racing using deep reinforcement learning. Nature 620(7976), 982–987 (2023)
22. Sallab, A.E., Abdou, M., Pérot, E., Yogamani, S.: Deep Reinforcement Learning framework for Autonomous Driving. Electronic Imaging 29(19), 70–76 (2017)
23. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 23–30. IEEE, Piscataway (2017)
24. Achiam, J., Held, D., Tamar, A., Abbeel, P.: Constrained Policy Optimization. In: Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 22–31. PMLR (2017)
25. Rajeswaran, A., Ghotra, S., Ravindran, B., Levine, S.: EPOpt: Learning Robust Neural Network Policies Using Model Ensembles. arXiv (Cornell University) (2016)
26. Yousefpour, A., et al.: All one needs to know about fog computing and related edge computing paradigms: A complete survey. Journal of Systems Architecture 98, 289–330 (2019)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.