

Review

Reinforcement Learning for UAV Control: From Algorithms to Deployment Readiness

Georgios Memlikai  and Konstantinos A. Tsintotas * 

Department of Information and Electronic Engineering, International Hellenic University,
GR-574 00 Thessaloniki, Greece; iee2019206@ihu.gr

* Correspondence: tsintotas@ihu.gr

Abstract

The rapid expansion of unmanned aerial vehicles (UAVs) across diverse application domains has underscored the need for reliable autonomy in complex and dynamic environments. To advance toward this goal, in recent years, learning-based control strategies have emerged as a promising alternative, offering adaptability and decision-making capabilities beyond those of conventional model-based ones. Bearing this in mind, the proposed article examines reinforcement learning methodologies for controlling UAVs, with particular emphasis on commonly used virtual environments, benchmark tasks, and the challenges of bridging the gap between simulation and real-world deployment. Therefore, key limitations, including high computational demands, reliance on extensive training data, and reduced robustness under environmental variability, are critically analyzed from a practical implementation perspective. Rather than adopting an algorithm-centric viewpoint, this work aggregates existing knowledge and categorizes learning-based approaches by their level of control abstraction and their treatment of safety and stability, thereby identifying the key factors limiting large-scale real-world deployment and the trends shaping intelligent controllers.

Keywords: UAV control; reinforcement learning; sim-to-real transfer; deployment readiness

1. Introduction

Unmanned aerial vehicles (UAVs) have evolved from specialized technological platforms into widely deployed systems across civil [1], commercial [2], and industrial sectors [3], emerging as one of the most dynamic areas in contemporary aeronautics and autonomous systems [4]. Such growth is underpinned by developments in wireless communications [5], microelectronics [6], and, in particular, artificial intelligence [7]. Applications such as environmental monitoring [8], precision agriculture [9], infrastructure inspection [10], emergency response [11], and autonomous surveillance increasingly rely on UAVs that can operate safely and efficiently in complex, dynamic environments [12]. Hence, as operational demands continue to expand, ensuring reliable UAV control has become a primary research topic [13].

Conventional approaches, such as proportional–integral–derivative (PID) controllers [14], linear quadratic regulators [15], and model predictive control (MPC) [16], have exhibited strong performance under structured and predictable conditions [17]. However, reliance on accurate system modeling and limited adaptability constrain their effectiveness in the presence of nonlinear dynamics, environmental uncertainty, load variations, or tasks that demand high-level decision-making [18]. These limitations have, in recent years, driven the development of intelligent control techniques [19], namely methodologies that integrate



Academic Editors: Haibin Duan,
Maolong Lv and Junkai Ren

Received: 21 December 2025

Revised: 29 January 2026

Accepted: 30 January 2026

Published: 3 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

machine learning [20], advanced machine learning [21], and deep learning (DL) [22], thereby enabling aerial platforms to make decisions and adapt their behavior autonomously. These advancements marked a shift in control frameworks from manual and semi-automatic operation toward fully autonomous systems driven by onboard decision-making.

In contrast to traditional model-based methods, reinforcement learning (RL) allows an agent to learn control policies through direct interaction with the environment, optimizing performance based on cumulative reward signals rather than predefined rules. More specifically, via a trial-and-error mechanism, an intelligent platform incrementally refines its behavior by optimizing a function aligned with mission objectives. Furthermore, when integrated with DL approaches [23], such methods significantly enhance a UAV's capacity to process high-dimensional sensory data from visual cameras, light detection and ranging (LiDAR), and inertial measurement units, thereby enabling perception capabilities comparable to those of human operators [24]. Therefore, through these mechanisms, autonomous navigation can dynamically adapt to environmental changes, thus highlighting the important role of intelligent control in complex operational scenarios [25]. Nevertheless, despite their advantages, several crucial challenges remain that impede the widespread deployment of DRL controllers [26]. These encompass the substantial computational cost associated with training large-scale models, the extensive data requirements induced by repeated trial-and-error interactions, the limited generalization across varying environmental conditions, and the difficulty of transferring policies learned in simulation to physical platforms, commonly referred to as the sim-to-real gap [27].

In this context, prior studies have extensively investigated UAV control pipelines, largely concentrating on specific tasks, such as multi-agent coordination [28] or autonomous navigation [29], offering detailed algorithmic analyses while largely overlooking real-world deployment. On the contrary, our article directly addresses the aforementioned gap by adopting a structured analytical framework that organizes learning-based approaches along three practically relevant dimensions: (i) control abstraction, (ii) sim-to-real readiness, and (iii) the incorporation of safety and stability considerations. Viewed through this lens, the proposed study moves beyond a purely algorithm-centric perspective, i.e., study-by-study comparison or quantitative meta-analysis of learning-based approaches for UAV control, allowing for a clearer differentiation between methods predominantly confined to simulation and those exhibiting partial readiness for real-world deployment.

The remainder of the paper is organized as follows. Section 2 presents the problem formulation and outlines the elements of the Markov decision process (MDP) typically used in RL approaches [30]. Section 3 reviews applicable methods, with emphasis on value-based techniques, policy-based approaches, and continuous-action actor–critic algorithms. Next, Section 4 compares learning- and traditional-based pipelines while highlighting the detected key challenges. Finally, Section 5 summarizes the main conclusions and provides future research directions.

2. Problem Formulation

UAVs constitute an inherently challenging control problem, stemming from their non-linear dynamics [31], underactuated structure [32], and heightened sensitivity to external disturbances, including wind, turbulence, and sensor noise [33]. While maintaining stable flight under such uncertainties is nontrivial, the challenge becomes significantly more pronounced in autonomous missions [34], where the platform is required to simultaneously ensure stability [35], perform obstacle avoidance [36], and achieve mission objectives without human intervention [37]. These requirements can be broadly classified into four task categories [38]: attitude stabilization [39], position and trajectory tracking [40], collision-free navigation [41], and high-level mission planning [32].

2.1. Control Framework

This task is further complicated by the tightly coupled dynamics of a six-degree-of-freedom system [42]. Each control objective entails distinct requirements in terms of perception [43], control accuracy [44], and decision-making [45]. Meeting these requirements necessitates a comprehensive understanding of both low-level physical control mechanisms and high-level autonomous decision-making processes, each introducing its own set of challenges.

2.2. Markovian Decision Process for UAV Control

A rigorous formulation of state and action representations is fundamental to learning-based controllers. RL methodologies are commonly formalized as MDPs, characterized by the tuple (S, A, P, r, γ) . In this framework, S denotes the state space, encompassing all potential UAV configurations, and A denotes the action space, e.g., motor thrusts. The term $P(s_{t+1}|s_t, a_t)$ defines the state transition probability, which controls the system's dynamics using physics equations. The reward function $R(s_t, a_t)$ offers a scalar feedback signal that indicates the quality of an action, and the discount factor $\gamma \in [0, 1)$ establishes the significance of future benefits in relation to immediate ones.

However, in practical scenarios, the agent seldom has direct access to the actual system state s_t [46], as perception is mediated by noisy, incomplete sensor measurements [47]. As a result, the control process is more accurately described within the framework of partially observable MDP [48], defined by $(S, A, P, R, \Omega, O, \gamma)$. Under this formulation, the state S corresponds to an observation-based representation derived from sequences of sensor readings, typically processed by temporal feature-extraction architectures to recover latent state information [42]. Here, Ω represents the observation space (the raw sensor data), and $O(o_t|s_{t+1}, a_t)$ is the observation function, which models the probability of perceiving observation o_t given the true state. This way, the agent can select the corresponding action based on the perceived environment [29].

It is further noted that existing control architectures are commonly categorized into low- and high-level controllers [23]. The former ones are typically responsible for generating thrust commands or attitude setpoints that directly influence motor actuation, whereas the latter ones produce waypoints or velocity references [47]. Within this hierarchy, the agent's role is to map control objectives, such as desired roll, pitch, yaw rate, or thrust, into appropriate motor commands that maintain stable flight [48]. Toward this goal, the learning process is driven by a reward signal, defined as a scalar feedback received when the agent executes an action a_t that induces a transition from state s_t to state s_{t+1} . This reward function is designed to encode task-specific objectives and guide policy optimization [49]. Accordingly, the control objective is commonly expressed as the maximization of the expected cumulative return [50], a formulation that provides a concise and widely adopted abstraction for sequential decision-making [51].

2.3. Constraints in Controlling UAV

Controlling autonomous aerial agents is subject to a range of physical [2], operational [18], and safety constraints [23] that must be satisfied throughout both the training and deployment phases [17]. Physical constraints arise from inherent system characteristics, including underactuation [31], actuator saturation [52], and limited onboard energy resources [23], which collectively bound the magnitude and temporal characteristics of feasible control inputs [52]. Safety constraints are equally fundamental, encompassing collision avoidance requirements [36], adherence to stable flight envelopes [35], and the prevention of unsafe state transitions during exploration [53]. Within DRL-based frameworks, these directly impact the definition of admissible action spaces [54], the formulation of reward functions [26],

and the adoption of conservative policies or explicit safety-filtering mechanisms [53]. Collectively, they establish the operational limits within which UAVs must operate, driving the development of learning-based controllers that achieve stability [32], efficiency [55], and robustness under sim-to-real transfer [27].

Two main strategies are commonly adopted to incorporate constraints into DRL formulations. A first approach relies on soft restrictions implemented through reward shaping, where penalty terms are added to the original reward function to discourage constraint violations [56]. The resulting composite objective balances task performance and constraint satisfaction but does not guarantee safety [56]. A second class of methods enforces complex restrictions through safety layers or constrained MDPs, often leveraging Lagrangian relaxation to maximize expected return while bounding expected cost below a predefined threshold [43].

2.4. Simulation Environments

Simulation environments constitute a fundamental component in the development and evaluation of learning-based UAV control systems, as they directly impact sensor realism, communication interfaces, and experimental scalability [28]. Commonly adopted platforms, including Gazebo [19], AirSim [55], CoppeliaSim [43], and FlightGear [48], exhibit substantial differences in licensing models, sensor fidelity, middleware integration, and support for multi-UAV scenarios [23]. Open-source simulators with native robot operating system (ROS) support are frequently favored for research prototyping, while commercial or hybrid solutions often provide superior visualization and high-fidelity modeling capabilities [23]. As a result, simulator selection plays a pivotal role not only in training efficiency but also in determining the viability of sim-to-real transfer [28].

Table 1 provides a qualitative overview of simulation environments commonly reported in UAV control studies [45]. Due to its extensive sensor library and native integration with ROS, Gazebo is commonly regarded as a standard platform for hardware-in-the-loop (HITL) testing [48]. AirSim, although computationally demanding and less suited to large-scale multi-agent training, offers high-fidelity photorealistic rendering that is particularly advantageous for perception-driven DRL policies [29]. In parallel, specialized simulators, such as Flightmare, emphasize high-speed rendering to enable the simultaneous training of UAV swarms [28], whereas CoppeliaSim combines a flexible dual-license model with mature physics engines [23].

Table 1. Overview of commonly used simulation environments for learning-based UAV control, detailing licensing, sensor capabilities, and integration features.

Simulator	License	Sensor Support	ROS Integration	Multi-UAV
Gazebo [23]	Open-source	Yes (Extensive)	Native	Yes
AirSim [55]	Open-source	Yes (Vision-based)	Yes	Limited
CoppeliaSim [23]	Free/Commercial	Yes (All types)	Yes	Yes
FlightGear [48]	Open-source	Limited	Partial	Limited
Flightmare [28]	Open-source	RGB/Depth	Yes	Yes (Swarm)

3. Learning-Based Approaches for UAV Control

RL algorithms are broadly categorized into model-based and model-free approaches [30]. The latter are frequently preferred in UAV control due to the complexity of aerodynamic interactions, even when the former employ learned dynamics models for planning [2]. Within the model-free paradigm, pipelines are commonly further classified into policy-based strategies, which directly optimize the control policy, and value-based methods, which learn a value function from which a policy is derived. Modern continuous-control applications predominantly rely on actor–critic architectures, a hybrid policy optimization

framework in which a critic (value function) is used to reduce the variance of the actor's policy gradient updates [31]. To provide a structured overview of the algorithmic landscape, Table 2 summarizes the key characteristics, typical applications, and trade-offs of the following value- and policy-based techniques, as well as actor–critic algorithms [7,26,57].

Table 2. Summary of representative reinforcement learning algorithms for UAV control, their typical applications, advantages, and main limitations derived from a synthesis of empirical findings reported in prior UAV control studies [7,26,57].

Algorithm	Category	Typical UAV Tasks	Main Advantages/Limitations
SARSA	Value-based	Basic navigation, simple way-point reaching	Adv.: Stable on-policy updates; integrates exploration into value estimation. Lim.: Discrete actions; unsuitable for continuous flight dynamics.
Q-Learning	Value-based	Grid-based path planning, simplified obstacle avoidance	Adv.: Off-policy; converges to optimal value. Lim.: Requires discretization; poor scalability to high-dimensional state spaces.
REINFORCE	Policy-based	Continuous maneuvering, simple trajectory tracking	Adv.: Naturally supports continuous actions. Lim.: High variance; sample inefficient.
DDPG	Actor–Critic (off-policy)	Attitude control, continuous thrust/torque regulation, precision maneuvering	Adv.: Handles continuous actions; replay buffer improves data efficiency. Lim.: Sensitive to hyperparameters; unstable without careful tuning.
PPO	Actor–Critic (on-policy)	Navigation, collision avoidance, take-off/landing, visually guided tasks	Adv.: Very stable updates; robust in long-horizon control. Lim.: Requires many samples (on-policy); slower training.
SAC	Actor–Critic (off-policy)	Robust control under uncertainty, complex dynamic environments	Adv.: Entropy maximization improves exploration and robustness. Lim.: Computationally heavier; tuning entropy temperature is nontrivial.
MADDPG	Multi-agent RL	Formation control, cooperative tracking, swarm coordination	Adv.: CTDE reduces non-stationarity; supports continuous multi-agent actions. Lim.: Scalability challenges; requires joint state/action information during training.
MA-PPO	Multi-agent RL	Distributed planning, mapless cooperative navigation	Adv.: Stable joint training; decentralized execution possible. Lim.: On-policy→slower data collection; communication assumptions may be required.

3.1. Value-Based Methods

Early learning-based approaches predominantly relied on value-based methods, in which an agent learns a state–action value function that estimates the expected return of executing a given action in a given state. Classical algorithms such as SARSA and Q-learning established the theoretical foundations by evaluating the utility of state–action pairs using tabular representations [23]. However, these methods inherently assume discrete state and action spaces, which renders them impractical for high-dimensional problems [58]. Most UAV applications require smooth and continuous inputs, including thrust modulation, attitude regulation, and fine-grained motor-level actuation, which cannot be effectively captured by discrete-action representations. Consequently, while value-based approaches remain suitable for high-level decision-making and simplified planning tasks, their applicability to low-level flight control and complex autonomous navigation is severely limited [15]. These constraints have driven the transition toward

policy-gradient and actor–critic methods, which natively support continuous control and enable end-to-end learning of UAV behaviors [48].

The Deep Q-Network (DQN) represents a key advancement [19], as it enables agents to process high-dimensional sensory inputs that were previously intractable for conventional value-based methods by approximating the Q-value function using deep neural networks [34]. To enhance learning stability, DQN introduces two fundamental mechanisms: target networks, which provide slowly varying update targets, and experience replay, which stores and randomly samples past interactions to mitigate correlations in the training data and smooth changes in the data distribution [19]. Despite being restricted to discrete action spaces, standard DQN has demonstrated effectiveness in high-level UAV decision-making tasks, particularly in complex environments where discrete planning and action selection are sufficient.

As noted above, value-based techniques have primarily been applied to high-level tasks with discrete decision spaces, but they exhibit significant limitations in low-level flight control. Discretization of inherently continuous action spaces can lead to unstable flight trajectories, actuator saturation, and mechanical chattering [47]. More generally, such approaches are ill-suited for precise maneuvering, as the resulting lack of smooth control signals increases energy consumption and can induce aerodynamic instability.

3.2. Policy-Based Methods and Actor–Critic Architectures

In contrast to value-based strategies, policy-based methods directly optimize the policy parameters by ascending the gradient of the expected return [48]. These approaches use a parameterized policy $\pi_{\theta}(a | s)$ that maps system states to probability distributions over actions. This formulation naturally accommodates continuous action spaces and adaptive behavior, making policy-gradient methods particularly well-suited for UAV control tasks that require smooth signals under dynamic, uncertain conditions. This section reviews the evolution of policy-based techniques, distinguishing between pure policy-gradient frameworks, which provide unbiased but high-variance updates, and actor–critic architectures, which incorporate value function estimates to stabilize learning and improve data efficiency.

3.2.1. Policy Gradient

A representative example is REINFORCE [59], which updates its parameters using Monte Carlo estimates of the cumulative return [30]. Despite its conceptual simplicity, this approach exhibits high gradient variance, leading to unstable learning dynamics and slow convergence. These limitations are further amplified in UAV control applications, where tasks are typically long-horizon, system dynamics are highly nonlinear, and exploration is safety-critical. As a result, practical implementations often rely on extensive reward shaping and variance-reduction techniques to achieve stable, effective training performance.

A key advantage of policy-based methods lies in their inherent support for continuous control, enabling end-to-end learning of commands such as thrust modulation, attitude regulation, and velocity setpoints [51]. Furthermore, the stochastic nature of policy-gradient formulations provides a principled mechanism for exploration, which is essential in environments affected by sensor noise, external disturbances [50], or partial observability [48]. Despite these advantages, pure policy-gradient strategies remain sample-inefficient and highly sensitive to hyperparameter selection [60]. Their reliance on Monte Carlo return estimates can lead to unstable or inconsistent performance, particularly in high-dimensional observation spaces encountered in vision-based UAV control.

3.2.2. Actor–Critic Architectures

To address the aforementioned limitations, actor–critic frameworks combine elements of both value- and policy-based RL by jointly training two coupled networks: an actor that generates control actions, and a critic that evaluates them through value function estimation to guide learning updates. This hybrid structure reduces the variance inherent in pure policy-gradient updates while improving sample efficiency, thereby making actor–critic algorithms among the most widely adopted approaches in continuous, high-dimensional settings.

DDPG (deep deterministic policy gradient) is a deterministic actor–critic algorithm in which the latter approximates a state–action value function, while the former produces continuous outputs, such as thrust commands or attitude adjustments [17]. The approach integrates experience replay, target networks, and temporally delayed updates to enhance learning stability. Exploration is facilitated by injecting stochastic noise, commonly modeled as Ornstein–Uhlenbeck processes [61], enabling compelling exploration in continuous-action spaces. Owing to these characteristics, DDPG is well-suited to fine-grained UAV tasks, including attitude stabilization [17] and regulation [48], as well as precision maneuvering [42].

Similarly, proximal policy optimization (PPO) is an on-policy actor–critic approach that employs a clipped surrogate objective to restrict update magnitudes and limit excessive deviations between successive iterations [48]. This mechanism substantially enhances training stability and robustness, particularly in long-horizon UAV control tasks [62]. Within this framework, the critic provides state-value estimates, while parameter updates are guided by advantage signals that incorporate temporal-difference information [48]. PPO has demonstrated strong performance in navigation, obstacle avoidance, and take-off and landing scenarios, as well as in hybrid operational settings where UAVs must operate reliably under dynamic disturbances or partial observability.

In contrast, soft actor–critic (SAC) extends this type of technique by introducing an entropy-regularized objective that explicitly promotes stochastic action selection [63], thereby improving exploration efficiency and training stability [19]. By jointly optimizing expected return and policy entropy, SAC produces smoother control behavior and enhanced robustness to model uncertainty. Its off-policy formulation and improved sample efficiency make it well-suited for training in photorealistic simulation environments and for applications subject to strict safety constraints or limited energy budgets.

In cooperative UAV missions [22,24,28], actor–critic methods have been extended to multi-agent settings through frameworks such as multi-agent DDPG (MADDPG) and variants of PPO designed for centralized training with decentralized execution [24,28]. These approaches address the non-stationarity inherent in multi-agent learning by conditioning the critic on joint state or action information during training, while enabling each agent to operate independently at execution time. Such architectures support effective swarm coordination, collision avoidance, and efficient area coverage with limited communication requirements.

3.3. Deep Reinforcement Learning for Continuous Control

DRL combines function approximation with continuous-action learning, enabling the synthesis of nonlinear control laws directly from high-dimensional state representations. In practice, actor–critic algorithms constitute the predominant approaches for continuous UAV control, as they offer a balanced trade-off between representational expressiveness and training stability [42]. Despite these advantages, it remains highly sensitive to reward formulation, hyperparameter selection, and modeling assumptions, with performance often deteriorating under environmental uncertainty or domain shift [24]. Consequently, although DRL facilitates control behaviors that are difficult to achieve using classical

methods alone, its practical deployment continues to depend strongly on simulation fidelity and the integration of hybrid control architectures [26,48,62].

3.4. Multi-Agent and Cooperative Reinforcement Learning

Many UAV applications require coordinated behavior among multiple platforms [28], including formation control [64], coordinated exploration [65], area coverage [66], and decentralized obstacle avoidance [67]. Such tasks introduce additional complexity, as each agent must learn in the presence of other adaptive agents, leading to a non-stationary learning environment. Multi-agent RL (MARL) addresses these challenges by facilitating coordinated decision-making while explicitly accounting for inter-agent interactions, communication constraints, and shared mission objectives [22].

Within this cooperative setting, centralized training with decentralized execution (CTDE) has emerged as a dominant paradigm in MARL [28]. During training, a centralized evaluator has access to global observations or joint state–action information, thereby stabilizing learning through collective assessment of agent behavior. At deployment, each UAV operates solely on local observations, preserving autonomy and robustness under communication constraints.

Building on this principle, MADDPG extends deterministic actor–critic formulations to multi-agent scenarios by conditioning the value estimator on the joint state–action space [28]. This allows agents to learn policies that respond to their teammates’ and opponents’ behaviors while still producing continuous control commands. In practical UAV applications, MADDPG has been employed for formation maintenance, cooperative target tracking, and distributed swarm control, where smooth coordination is required under dynamically changing environmental conditions [24,68,69]. Related actor–critic variants further improve scalability through shared reward structures, communication graphs, or attention mechanisms. In parallel, on-policy MARL approaches, such as multi-agent PPO, extend the stability properties of PPO to cooperative UAV domains [70]. The clipped surrogate objective limits destabilizing updates during joint training, while decentralized execution allows each UAV to operate independently after deployment. Collectively, these methods have demonstrated strong potential in multi-agent navigation, distributed path planning, and collaborative obstacle avoidance, particularly in environments characterized by uncertainty or partial observability [71].

In practical deployments, the choice of a MARL architecture is closely linked to the available communication infrastructure and latency constraints. Although a detailed analysis of communication protocols lies outside the scope of this survey, addressing system-level factors that are often abstracted away in simulation is essential for bridging the gap between algorithmic development and deployment readiness. Two such factors are particularly critical.

First, communication architecture imposes strict limitations on coordination in real-world environments. Unlike simulated settings that assume instantaneous access to global state information, deployed UAV swarms must operate under severe range and bandwidth constraints and, consequently, limited observability. This has motivated a shift away from transmitting raw sensory data toward the design of efficient communication strategies, in which agents learn what information to share and when to broadcast it.

Second, coordination and team-level objectives extend beyond mere stabilization. Agents must balance global mission goals with individual safety considerations, a requirement that places strong demands on reward design. Poorly specified objectives can lead to undesirable behaviors, such as inactive or self-serving agents that undermine collective performance. Careful reward formulation is therefore necessary to promote cooperative behavior while preventing actions that jeopardize overall mission success.

3.5. Perception-Driven Reinforcement Learning

Perception-driven pipelines enable UAVs to generate control actions directly from high-dimensional sensory inputs, most commonly images, depth maps, or LiDAR measurements, by tightly integrating deep perception models with policy learning [26,36,43]. This paradigm removes the need for explicit mapping or hand-crafted feature extraction and has demonstrated strong performance in visually guided navigation, obstacle avoidance, and autonomous landing tasks, particularly within simulation-based environments [29,34,72]. End-to-end DRL architectures, as illustrated in Figure 1, have been shown to effectively capture nonlinear visual control policies by directly mapping raw sensory observations to continuous flight commands [42,73].

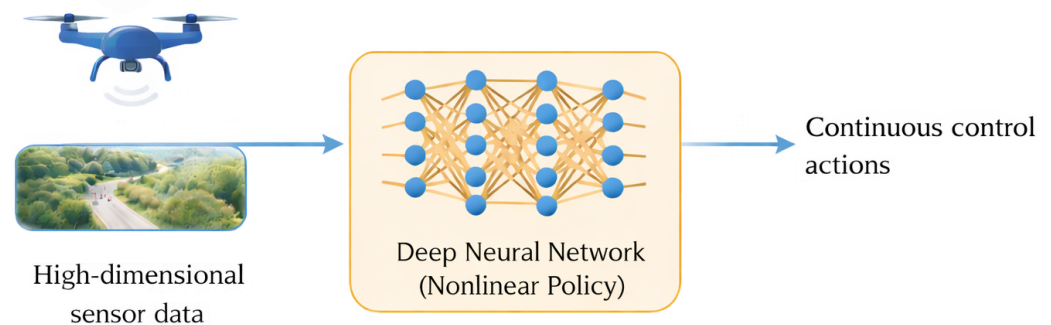


Figure 1. Architecture of an end-to-end deep reinforcement learning agent for UAV control. The system processes high-dimensional raw sensor data (e.g., RGB images) through a deep neural network to directly output continuous control actions.

Certain architectural choices play a critical role in the successful deployment of perception-driven control systems. Owing to their computational efficiency on embedded hardware and strong spatial inductive bias, convolutional neural networks remain the predominant choice for feature extraction in UAV applications [43]. Although more recent architectures, such as Vision Transformers (ViTs), offer improved global context modeling, they typically incur substantially higher computational costs. In safety-critical UAV control, inference latency constitutes a key concern, as instability may arise if the perception module's forward pass exceeds the allowable control period. To mitigate this limitation, recent research has focused on lightweight perception models [74], including optimized variants of YOLO architectures (e.g., TakuNet [75]), which aim to balance representational capacity with real-time execution constraints.

However, perception-driven policies are particularly susceptible to domain shift, as variations in illumination, surface textures, sensor noise, or camera calibration can lead to pronounced performance degradation when transferring from simulation to real-world environments [76]. Empirical studies consistently report substantial drops in the performance of vision-based DRL controllers under environmental conditions not encountered during training, underscoring their limited robustness in safety-critical applications [77].

Thus, while perception-driven RL significantly enhances UAV autonomy in unstructured environments, its practical deployment is constrained by challenges of sim-to-real generalization. Addressing these limitations typically requires adopting domain randomization strategies, sensor fusion techniques, or hybrid control architectures to ensure reliable and safe operation in real-world settings [78].

3.6. Training Considerations

Training RL policies for UAV control remains computationally intensive due to sample inefficiency, high-dimensional state representations, and stringent safety constraints. Continuous-control tasks typically require millions of interaction steps to achieve conver-

gence, rendering direct real-world training impractical and necessitating extensive reliance on simulation environments [54,79]. Off-policy algorithms such as DDPG and SAC partially mitigate this challenge through experience replay, whereas on-policy methods like PPO generally require significantly larger data volumes owing to their reliance on freshly collected trajectories [48].

Effective exploration further complicates the training process, as unstructured exploration can induce unsafe behaviors or unstable flight, particularly in cluttered or highly dynamic environments [29,53]. Consequently, techniques such as reward shaping, constrained optimization objectives, and safety-filtering mechanisms are commonly adopted to limit unsafe actions during training [55]. Sparse reward formulations are often insufficient in this context, making careful reward engineering a critical component of successful learning. Dense reward structures that incorporate weighted penalty terms aligned with soft constraints as discussed in Section 1. However, overly harsh penalties can lead to excessively conservative behavior. To balance exploration and safety, curriculum learning strategies are frequently used, gradually increasing task complexity as training progresses [28]. In addition, safety layers are increasingly integrated directly into the training loop to intercept and suppress hazardous exploratory actions before they are executed in the simulator.

Despite continued improvements in simulation fidelity, discrepancies between simulated and real-world dynamics, sensor characteristics, and environmental conditions remain a significant barrier to practical deployment [27]. Policies trained exclusively in simulation often exhibit significant performance degradation when faced with real-world variations in illumination, sensor noise, or aerodynamic disturbances. Techniques such as domain randomization and the use of high-fidelity simulators partially alleviate this issue by exposing agents to a broader range of operating conditions during training, thereby enhancing robustness and generalization [72,80]. Nevertheless, reliable real-world deployment frequently necessitates additional fine-tuning using limited real-flight data or the adoption of hybrid control architectures that integrate learning-based policies with classical control loops in order to maintain stability and ensure safety [17].

4. Comparison and Discussion

The analysis presented in Section 3 highlights clear contrasts among learning-based control approaches for UAVs. While value-based methods offer favorable sample efficiency for discrete, high-level decision-making tasks, the continuous control requirements imposed by the dynamics of aerial vehicles necessitate the use of actor–critic frameworks. However, strong performance in simulation does not directly translate to deployment readiness. Instead, factors such as safety constraints, data inefficiency, and the persistent sim-to-real gap in sensor and dynamics modeling constitute the primary obstacles to real-world operation.

Within this context, this section provides a critical synthesis of the examined techniques, focusing on their practical maturity across diverse deployment conditions. First, the advantages of learning-based strategies are contrasted with those of traditional controllers. This is followed by an assessment of the fundamental limitations that continue to constrain learning-based solutions. Finally, the applicability of different algorithmic classes across distinct UAV operational domains is discussed. To complement this qualitative analysis, Table 3 summarizes the practical readiness of widely adopted RL approaches, with emphasis on continuous control capability, demonstrated sim-to-real transfer, and safety-related considerations. Collectively, this comparison indicates that only a limited subset of existing methods has exhibited partial readiness for real-world UAV deployment.

Table 3. Summary of representative DRL algorithms in UAV control, highlighting typical application scopes and notable studies demonstrating sim-to-real transfer and constraint handling. Note that performance attributes are study-specific rather than inherent to the algorithm.

Algorithm	Typical Application Scope	Key Advantages	Sim-to-Real Examples	Safety/Constraints Approach
DQN/Double DQN	High-level Path Planning, Discrete Action Spaces	Sample efficiency in discrete domains	[66,81] (Sim-only focus)	Penalty-based reward shaping [58]
DDPG	Low-level Attitude Control, Continuous Thrust	Continuous control, deterministic policy	Validated in flight tests for gain tuning [17]	Actuator limits, Safety layers [47]
PPO	End-to-End Navigation, Attitude Control	Stability (clipped objective), Robustness	Demonstrated in fixed-wing attitude control [51,62]	Reward shaping for obstacle avoidance [48]
SAC	Agile Maneuvering, Energy-Efficient Flight	Max entropy exploration, Sample efficiency	Field experiments for attitude control [82]	Entropy regularization for robustness [82]
MADDPG	Swarm Formation, Cooperative Tracking	Centralized Training, Decentralized Execution	Mostly Simulation (Morphing wings) [68,69]	Collision avoidance rewards [69]
MAPPO	Dynamic Target Search, Large-Scale Swarms	On-policy stability for multi-agent settings	Simulation-based (Target Search) [70,71]	Potential-field fusion rewards [71]

Note: This table categorizes algorithms based on their predominant use cases in the reviewed literature. Sim-to-Real and Safety capabilities refer to specific successful implementations cited and do not guarantee performance in all scenarios.

4.1. Comparison with Traditional Controllers

As discussed in Section 2, traditional model-based control schemes remain the dominant solutions for low-level UAV stabilization. Their computational efficiency, ease of implementation, and, most importantly, the availability of formal stability guarantees under nominal operating conditions underpin their continued prevalence in practical systems [17]. For fundamental tasks such as hovering or trajectory tracking, a well-tuned PID controller often outperforms learning-based alternatives in terms of steady-state accuracy and predictability. Nevertheless, despite their effectiveness in nominal regimes, fixed-gain controllers exhibit limited adaptability in highly dynamic scenarios, such as aggressive maneuvers, abrupt payload changes, or actuator degradations, where nonlinearities and unmodeled effects become pronounced, compromising stability. In such cases, robust and adaptive control strategies can partially address modeled uncertainties, but their performance remains constrained by the fidelity of the underlying system model.

DRL offers a complementary, data-driven alternative by enabling control policies to adapt through interaction with the system, without relying on explicit analytical models. Through this interaction, these types of agents can implicitly realize gain-scheduling behavior, adjusting control actions across a wide range of operating conditions. A key advantage of these techniques over classical PID control is their ability to accommodate unmodeled dynamics and structural variations. For instance, PPO-based controllers have been shown to preserve stable flight under structural changes that would require extensive retuning in conventional control frameworks. Similarly, prior work has demonstrated the use of DRL to optimize PID parameters for autonomous landing on moving platforms, highlighting the potential of learning-based adaptation to surpass manual tuning procedures [47].

Despite these advantages, the flexibility of such controllers comes at a high cost. Learned policies typically operate as black-box function approximators, offering limited interpretability and lacking the formal safety and stability guarantees that underpin classical control theory [19]. Consequently, even though DRL methods exhibit strong performance in complex and uncertain environments, they are generally regarded as complementary components that augment traditional architectures rather than as direct replacements.

4.2. Real-World Learning vs. Sim-to-Real Transfer

A distinct line of research explores direct real-world learning, in which control policies are acquired exclusively through physical interaction with the environment. In contrast, most existing studies adopt a sim-to-real paradigm to mitigate the risks of trial-and-error learning on physical platforms. In principle, direct real-world training eliminates the so-called reality gap introduced by modeling inaccuracies and simulator simplifications [43]. In practice, however, its applicability remains severely constrained by substantial logistical, safety, and operational challenges.

Uncontrolled exploration in real flight conditions can have severe consequences, as even a single erroneous action may result in catastrophic hardware failure. Moreover, the limited battery endurance of aerial platforms poses a fundamental constraint, as DRL algorithms typically require millions of interaction steps to converge, resulting in prohibitively long flight times. As a result, direct real-world learning has so far been confined to a narrow set of experimental conditions.

First, nano-UAV platforms are frequently employed, as their low mass and increased robustness reduce the consequences of crashes and allow repeated physical interactions during exploration with limited hardware damage [47]. Second, safe RL frameworks have been introduced, incorporating mechanisms such as virtual safety cages, shielding strategies, or human-in-the-loop safety pilots that intervene to override control commands when imminent collisions or unsafe states are detected [48]. Third, real-world learning is often restricted to simplified control objectives, where the dimensionality of the state and action spaces is significantly reduced by focusing on low-level primitives, such as attitude stabilization [15]. This reduction in task complexity lowers sample requirements, making training feasible within the tight energy and endurance constraints of onboard UAV hardware.

4.3. Strengths and Weaknesses

The strengths and limitations discussed in this section arise largely independently of specific UAV platforms or task formulations. Figure 2 provides a qualitative comparative overview of representative RL algorithms, illustrating key trade-offs among sample efficiency, training stability, and suitability for continuous UAV control. The comparison is based on a synthesis of reported results from existing studies, rather than on a theoretically grounded or quantitatively normalized evaluation framework [26,58,83]. The balanced assessment highlights that DRL techniques exhibit distinct advantages and shortcomings that must be carefully considered when designing aerial systems, particularly for operation under uncertainty.

Value-based methods, such as DQN, are well suited to high-level decision-making problems with discrete action spaces, for example, target selection or simplified planning. However, they are ill-suited to low-level attitude control, where the continuous nature of UAV dynamics demands smooth control signals. In the context of sim-to-real transfer, on-policy approaches such as PPO are commonly preferred when robustness and training stability are prioritized over raw sample efficiency, as their constrained update mechanisms facilitate more predictable behavior and simpler tuning. In contrast, off-policy algorithms including SAC and TD3 leverage experience reuse to achieve higher sample efficiency, making them theoretically more attractive for onboard learning scenarios [34]. Moreover, their capacity to generate smooth, continuous control commands is particularly beneficial for reducing energy consumption and mitigating aerodynamic instabilities during flight.

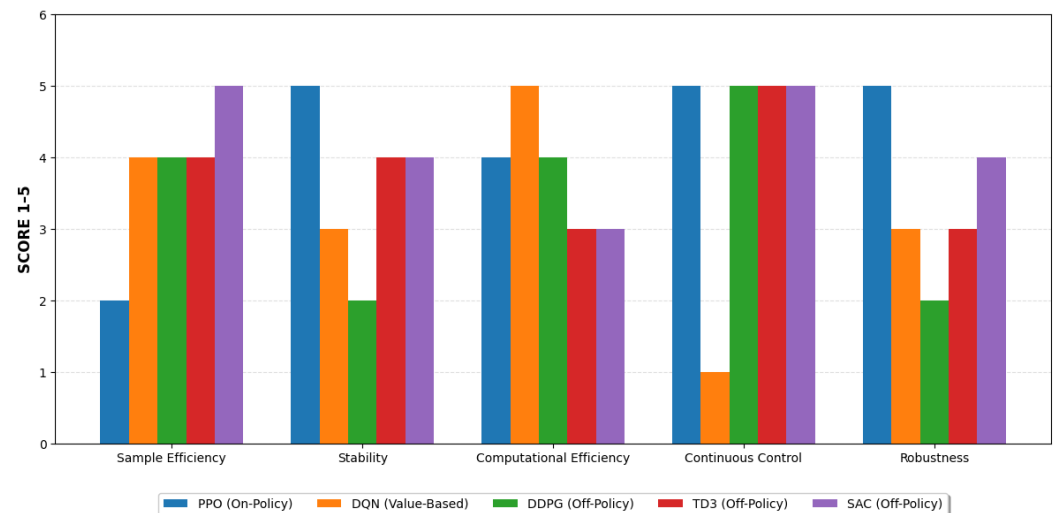


Figure 2. Qualitative comparison of reinforcement learning algorithms (PPO, DQN, DDPG, TD3, and SAC) for UAV control. This chart is based on a synthesis of empirical findings reported in the literature and illustrates the trade-offs among sample efficiency, training stability, and suitability for continuous flight control [26,58,83].

4.3.1. Strengths

End-to-end learning enables the direct mapping of raw sensory inputs to control commands, eliminating the need for explicitly separated state-estimation and path-planning modules and thereby reducing overall system latency [72]. Previous studies have shown that UAVs can acquire stable flight behavior using image-based inputs alone, without relying on explicit state estimation pipelines [73]. Beyond architectural simplicity, RL policies are particularly effective at optimizing long-term performance objectives and handling complex operational scenarios. In contrast, traditional control methods are typically designed around local or short-horizon optimization criteria. By leveraging function approximators, DRL frameworks can automatically extract task-relevant features from high-dimensional sensory data using convolutional neural networks, recurrent neural networks, and transformer-based architectures [43]. This representational capacity enables unified neural controllers to manage heterogeneous system dynamics; for instance, learning-based approaches have demonstrated the ability to handle mixed flight regimes in hybrid UAV platforms without the need for explicit mode-switching logic [84].

4.3.2. Weaknesses

Despite the aforementioned advantages, most reported successes of RL-based controllers are achieved under carefully designed training pipelines and highly controlled experimental conditions. A primary limitation arises from generalization failures, which occur when the distribution of real-world operating conditions deviates from those encountered during training. Policies learned in a specific environment may degrade significantly or fail entirely when exposed to even minor environmental variations, posing serious safety risks in real flight scenarios [27]. This vulnerability is particularly pronounced for vision-based controllers, which are highly sensitive to changes in visual appearance and environmental conditions, potentially leading to catastrophic failures during deployment [77].

Sample inefficiency constitutes a further critical drawback. DRL methods require millions of interaction steps to converge to a stable policy, resulting in substantial computational cost and extended training times [54]. Although algorithmic advances, such as proximal policy optimization, partially alleviate this issue by improving training stability, the overall computational burden remains considerable [59]. In addition, their performance is highly sensitive to hyperparameter selection, including learning rates, reward

weights, and exploration noise characteristics. Inappropriate parameter choices can lead to unstable behavior or complete training failure, particularly under extreme operating conditions [17,61]. Finally, the inherent lack of interpretability of neural-network-based policies presents a major obstacle in safety-critical applications as these controllers act as black-box approximators, making it difficult to predict, verify, or certify their behavior, limiting their adoption in scenarios where transparency and formal safety guarantees are required [31].

4.4. Applicability Across UAV Tasks

At the lowest level, maintaining aircraft stability and tracking angular velocity are the primary goals. The technical issue lies in the control loop's high-frequency characteristics and the requirement for continuous action spaces. By optimizing a substitutive objective function that limits policy updates, PPO enables consistent policy enhancement and avoids harmful updates that could lead to catastrophic failure during flight training [15]. Beyond standard stabilization, DRL offers a framework for managing complex or time-varying dynamics. Moreover, a well-trained policy could sustain fixed-point flights even when physical parameters are changed, demonstrating agents' resilience to physical changes [48]. Alternatively, DRL serves as a high-level tuner, using algorithms such as DDPG to adjust gains rather than completely replacing the PID loop [17].

Subsequently, in navigation tasks that require sequential decision-making under uncertainty, learning-based approaches have shown impressive performance [45]. This layer shifts the emphasis from stability to trajectory generation and spatial awareness. End-to-end methods, which condense the conventional perception and mapping pipeline into a single neural architecture, are the main advantage of DRL in this sector [72]. Moreover, such frameworks are helpful for real-time autonomous aerial navigation, as they directly map raw sensor input to control actions when paired with DL-based perception [78]. In addition, it is vital to maintain a balance between efficiency and safety when navigating in dynamic environments [29]. When obstacles move unpredictably, static path planners frequently fail. On the contrary, DRL agents navigate high-density settings by anticipating possible collisions through dynamic reward functions that learn to adjust their exploration tactics in real-time [29]. Recent architectures combine fluid dynamical systems and long short-term memory networks to address the specific kinematic constraints of UAVs [42]. This integration provides enhanced feasibility while enabling the generation of smooth, 6-DOF collision-free trajectories that respect the airframe's physical limitations.

Lastly, at the highest level of autonomy, swarm intelligence poses a scaling problem that centralized control techniques cannot handle due to computational constraints and communication bandwidth limitations [22]. Swarm coordination and formation flights are examples of cooperative tasks that showcase some of RL's most promising applications. UAV teams can learn decentralized policies that coordinate actions based on shared or partially shared observations, thanks to MARL frameworks. This is especially helpful when coordination needs to adjust to difficult circumstances. On top of this, the swarm can simultaneously optimize each agent's formation geometry by employing algorithms such as MADDPG [24]. Potential-field reward shaping is used by algorithms such as MAPPO in cooperative scenarios to encourage coverage and avoid redundant exploration by multiple agents [71].

4.5. Key Open Challenges

Despite the rapid progress in learning-based UAV control pipelines, several fundamental challenges continue to limit their real-world applicability. A primary bottleneck arises from the high dimensionality of modern policies, which often rely on processing rich

sensory inputs. Such architectures impose substantial computational and energy demands, making real-time onboard inference challenging on resource-constrained flight controllers while at the same time increasing latency during high-frequency control loops [2,75,85]. These constraints limit the deployment of complex DRL policies to platforms with sufficient on-board compute or require aggressive model compression, which may degrade control performance.

Another vital limitation concerns data efficiency and training scalability. Most algorithms require extensive interaction with the environment to converge, particularly in long navigation and cooperative multi-UAV tasks [29,54]. While simulation-based training mitigates safety risks, the reliance on large-scale simulated experience introduces a strong dependency on simulator fidelity. Policies that achieve high performance in such environments often fail to generalize when exposed to real-world disturbances, unmodeled aerodynamics, or sensor imperfections [26,27]. This sim-to-real gap remains one of the most persistent barriers to deployment, especially in safety-critical missions.

Robustness under environmental uncertainty represents an additional failure mode. Vision-based DRL policies are particularly sensitive to changes in illumination, weather conditions, and visual occlusions, leading to severe performance degradation or unsafe behavior when operating outside the training distribution [77]. Experimental studies report significant drops in perception accuracy and control reliability under adverse conditions, underscoring the fragility of purely perception-driven control pipelines [50]. Although domain randomization and sensor fusion improve robustness, they do not yet provide formal guarantees on stability or constraint satisfaction.

Finally, the lack of interpretability and formal safety guarantees remains a major barrier to certification and regulatory acceptance. As already mentioned, DRL policies typically operate as black-box approximators, making it difficult to predict or verify their behavior under rare or extreme conditions [31]. Unlike classical controllers, learning-based approaches generally lack Lyapunov-based stability proofs or explicit constraint enforcement mechanisms, limiting their adoption in certified UAV systems [19]. As a result, current research increasingly favors hybrid architectures, where their policies augment rather than replace traditional control loops, combining adaptive decision-making with provable safety properties [23,50].

5. Conclusions

This article has examined learning-based approaches for UAV control, with an emphasis on their practical applicability and readiness for real-world deployment. Although such controllers exhibit strong potential in handling nonlinear dynamics, continuous control, and high-dimensional sensory inputs, their effectiveness remains highly task-dependent and uneven across control layers. Actor–critic architectures, particularly PPO and SAC, emerge as the most promising candidates for continuous low-level control and navigation, whereas value-based and purely perception-driven methods are confined mainly to simulation-based or simplified settings. Beyond algorithmic performance alone, deployment readiness is primarily determined by factors such as sim-to-real transferability, safety-aware operation, computational feasibility, and robustness under environmental uncertainty.

Future Directions

Raw algorithmic performance is no longer the primary limiting factor for real-world adoption. Instead, the absence of formal safety guarantees, the persistent sim-to-real gap in sensor and dynamics modeling, and the vulnerability of end-to-end policies under unmodeled disturbances continue to hinder widespread deployment. Bridging the gap

between simulation-based success and reliable real-world operation requires a shift in research focus toward the following directions:

- **Safe and Constraints:** Training formulations must explicitly include hard safety constraints to ensure that exploratory actions respect physical flight envelopes and operational limits, enabling deployment beyond controlled laboratory settings.
- **Standardized Sim-to-Real Benchmarks:** The field would benefit from consistent and comparable evaluation frameworks. Future studies should prioritize HITL validation and report deployment-relevant metrics, including energy efficiency, latency, and communication robustness.
- **Implementation of Hybrid and Residual Architectures:** Integrating data-driven approaches with classical control remains a key objective. Residual learning frameworks, in which learning-based components augment rather than replace traditional controllers, offer a promising balance between adaptability and predictability. For instance, learning mechanisms can be used to adaptively tune PID gains to compensate for aerodynamic effects not captured by linear models.

Author Contributions: Conceptualization, G.M. and K.A.T.; methodology, G.M. and K.A.T.; software, G.M.; validation, K.A.T.; formal analysis, K.A.T.; investigation, G.M.; resources, G.M. and K.A.T.; data curation, G.M.; writing—original draft preparation, G.M.; writing—review and editing, K.A.T.; visualization, G.M.; supervision, K.A.T.; project administration, K.A.T.; funding acquisition, K.A.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Shakhathreh, H.; Sawalmeh, A.H.; Al-Fuqaha, A.; Dou, Z.; Almaita, E.; Khalil, I.; Othman, N.S.; Khreishah, A.; Guizani, M. Unmanned Aerial Vehicles (UAVs): A Survey on Civil Applications and Key Research Challenges. *IEEE Access* **2019**, *7*, 48572–48634. [\[CrossRef\]](#)
2. Mohsan, S.A.H.; Khan, M.A.; Noor, F.; Ullah, I.; Alsharif, M.H. Towards the Unmanned Aerial Vehicles (UAVs): A Comprehensive Review. *Drones* **2022**, *6*, 147. [\[CrossRef\]](#)
3. Mourtzis, D.; Angelopoulos, J.; Panopoulos, N. Uavs for industrial applications: Identifying challenges and opportunities from the implementation point of view. *Procedia Manuf.* **2021**, *55*, 183–190. [\[CrossRef\]](#)
4. Mitroudas, T.; Tsintotas, K.A.; Santavas, N.; Psomoulis, A.; Gasteratos, A. Towards 3D printed modular unmanned aerial vehicle development: The landing safety paradigm. In Proceedings of the 2022 IEEE International Conference on Imaging Systems and Techniques (IST), Kaohsiung, Taiwan, 21–23 June 2022; pp. 1–6.
5. Gu, X.; Zhang, G. A survey on UAV-assisted wireless communications: Recent advances and future trends. *Comput. Commun.* **2023**, *208*, 44–78. [\[CrossRef\]](#)
6. Mohsan, S.A.H.; Othman, N.Q.H.; Khan, M.A.; Amjad, H.; Żywiołek, J. A comprehensive review of micro UAV charging techniques. *Micromachines* **2022**, *13*, 977. [\[CrossRef\]](#)
7. Carrio, A.; Sampietro, C.; Rodriguez-Ramos, A.; Campoy, P. A Review of Deep Learning Methods and Applications for Unmanned Aerial Vehicles. *J. Sens.* **2017**, *2017*, 3296874. [\[CrossRef\]](#)
8. Li, X.; Savkin, A.V. Networked unmanned aerial vehicles for surveillance and monitoring: A survey. *Future Internet* **2021**, *13*, 174. [\[CrossRef\]](#)
9. Kim, J.; Kim, S.; Ju, C.; Son, H.I. Unmanned Aerial Vehicles in Agriculture: A Review of Perspective of Platform, Control, and Applications. *IEEE Access* **2019**, *7*, 105100–105115. [\[CrossRef\]](#)
10. Gao, C.; Wang, X.; Chen, X.; Chen, B.M. A hierarchical multi-UAV cooperative framework for infrastructure inspection and reconstruction. *Control Theory Technol.* **2024**, *22*, 394–405. [\[CrossRef\]](#)

11. Wang, Y.; Li, J.; Yang, X.; Peng, Q. UAV–Ground Vehicle Collaborative Delivery in Emergency Response: A Review of Key Technologies and Future Trends. *Appl. Sci.* **2025**, *15*, 9803. [\[CrossRef\]](#)
12. Tsintotas, K.A.; Bampis, L.; Taitzoglou, A.; Kansizoglou, I.; Kaparos, P.; Bliamis, C.; Yakinthos, K.; Gasteratos, A. The MPU RX-4 project: Design, electronics, and software development of a geofence protection system for a fixed-wing vtol uav. *IEEE Trans. Instrum. Meas.* **2022**, *72*, 7000113. [\[CrossRef\]](#)
13. Tsintotas, K.A.; Bampis, L.; Taitzoglou, A.; Kansizoglou, I.; Gasteratos, A. Safe UAV landing: A low-complexity pipeline for surface conditions recognition. In Proceedings of the 2021 IEEE International Conference on Imaging Systems and Techniques (IST), Kaohsiung, Taiwan, 24–26 August 2021; pp. 1–6.
14. Amerttet, S.; Gebresenbet, G.; Alwan, H.M. Modeling of unmanned aerial vehicles for smart agriculture systems using hybrid fuzzy PID controllers. *Appl. Sci.* **2024**, *14*, 3458. [\[CrossRef\]](#)
15. Rinaldi, M.; Primatesta, S.; Guglieri, G. A comparative study for control of quadrotor UAVs. *Appl. Sci.* **2023**, *13*, 3464. [\[CrossRef\]](#)
16. Farid, U.; Khan, B.; Ali, S.M.; Ullah, Z. A Digital Twin Model for UAV Control to Lift Irregular-Shaped Payloads Using Robust Model Predictive Control. *Machines* **2025**, *13*, 1069. [\[CrossRef\]](#)
17. Sönmez, S.; Montecchio, L.; Martini, S.; Rutherford, M.J.; Rizzo, A.; Stefanovic, M.; Valavanis, K.P. Reinforcement Learning-Based PD Controller Gains Prediction for Quadrotor UAVs. *Drones* **2025**, *9*, 581. [\[CrossRef\]](#)
18. Telli, K.; Kraa, O.; Himeur, Y.; Ouamane, A.; Boumehraz, M.; Atalla, S.; Mansoor, W. A Comprehensive Review of Recent Research Trends on Unmanned Aerial Vehicles (UAVs). *Systems* **2023**, *11*, 400. [\[CrossRef\]](#)
19. Azar, A.T.; Koubaa, A.; Ali Mohamed, N.; Ibrahim, H.A.; Ibrahim, Z.F.; Kazim, M.; Ammar, A.; Benjdira, B.; Khamis, A.M.; Hameed, I.A.; et al. Drone Deep Reinforcement Learning: A Review. *Electronics* **2021**, *10*, 999. [\[CrossRef\]](#)
20. Choi, S.Y.; Cha, D. Unmanned aerial vehicles using machine learning for autonomous flight; state-of-the-art. *Adv. Robot.* **2019**, *33*, 265–277. [\[CrossRef\]](#)
21. Kulkarni, S.; Patil, D.D. Reinforcement Learning for Autonomous Systems. In Proceedings of the 2025 4th International Conference on Sentiment Analysis and Deep Learning (ICSADL), Bhimdatta, Nepal, 18–20 February 2025; pp. 816–820. [\[CrossRef\]](#)
22. Javaid, S.; Saeed, N.; Qadir, Z.; Fahim, H.; He, B.; Song, H.; Bilal, M. Communication and Control in Collaborative UAVs: Recent Advances and Future Trends. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 5719–5739. [\[CrossRef\]](#)
23. Elmeseiry, N.; Alshaer, N.; Ismail, T. A Detailed Survey and Future Directions of Unmanned Aerial Vehicles (UAVs) with Potential Applications. *Aerospace* **2021**, *8*, 363. [\[CrossRef\]](#)
24. Zhao, Z.; Wan, Y.; Chen, Y. Deep Reinforcement Learning-Driven Collaborative Rounding-Up for Multiple Unmanned Aerial Vehicles in Obstacle Environments. *Drones* **2024**, *8*, 464. [\[CrossRef\]](#)
25. Tsintotas, K.A.; Sevetlidis, V.; Papapetros, I.T.; Balaska, V.; Psomoulis, A.; Gasteratos, A. BK tree indexing for active vision-based loop-closure detection in autonomous navigation. In Proceedings of the IEEE Mediterranean Conference on Control and Automation (MED), Vouliagmeni, Greece, 28 June–1 July 2022; pp. 532–537.
26. Shen, S.E.; Huang, Y.C. Application of Reinforcement Learning in Controlling Quadrotor UAV Flight Actions. *Drones* **2024**, *8*, 660. [\[CrossRef\]](#)
27. Hu, H.; Zhang, K.; Tan, A.H.; Ruan, M.; Agia, C.; Nejat, G. A Sim-to-Real Pipeline for Deep Reinforcement Learning for Autonomous Robot Navigation in Cluttered Rough Terrain. *IEEE Robot. Autom. Lett.* **2021**, *6*, 6569–6576. [\[CrossRef\]](#)
28. Ekechi, C.C.; Elfouly, T.; Alouani, A.; Khattab, T. A Survey on UAV Control with Multi-Agent Reinforcement Learning. *Drones* **2025**, *9*, 484. [\[CrossRef\]](#)
29. Sheng, Y.; Liu, H.; Li, J.; Han, Q. UAV Autonomous Navigation Based on Deep Reinforcement Learning in Highly Dynamic and High-Density Environments. *Drones* **2024**, *8*, 516. [\[CrossRef\]](#)
30. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2018.
31. Abbas, N.; Abbas, Z.; Zafar, S.; Ahmad, N.; Liu, X.; Khan, S.S.; Foster, E.D.; Larkin, S. Survey of Advanced Nonlinear Control Strategies for UAVs: Integration of Sensors and Hybrid Techniques. *Sensors* **2024**, *24*, 3286. [\[CrossRef\]](#)
32. Zuo, Z.; Liu, C.; Han, Q.L.; Song, J. Unmanned Aerial Vehicles: Control Methods and Future Challenges. *IEEE/CAA J. Autom. Sin.* **2022**, *9*, 601–614. [\[CrossRef\]](#)
33. Thomas, D.G.; Olshanskyi, D.; Krueger, K.; Wongpiromsarn, T.; Jannesari, A. Interpretable UAV Collision Avoidance using Deep Reinforcement Learning. *arXiv* **2021**, arXiv:2105.12254. [\[CrossRef\]](#)
34. Samma, H.; El-Ferik, S. Autonomous UAV Visual Navigation Using an Improved Deep Reinforcement Learning. *IEEE Access* **2024**, *12*, 79967–79977. [\[CrossRef\]](#)
35. Kramar, V.; Kabanov, A.; Dudnikov, S. A mathematical model for a conceptual design and analyses of UAV stabilization systems. *Fluids* **2021**, *6*, 172. [\[CrossRef\]](#)
36. Merei, A.; Mcheick, H.; Ghaddar, A.; Rebaine, D. A survey on obstacle detection and avoidance methods for UAVs. *Drones* **2025**, *9*, 203. [\[CrossRef\]](#)
37. Tsintotas, K.A.; Bampis, L.; Gasteratos, A. *Online Appearance-Based Place Recognition and Mapping: Their Role in Autonomous Navigation*; Springer Nature: Berlin/Heidelberg, Germany, 2022; Volume 133.

38. Rubí, B.; Pérez, R.; Morcego, B. A survey of path following control strategies for UAVs focused on quadrotors. *J. Intell. Robot. Syst.* **2020**, *98*, 241–265. [\[CrossRef\]](#)
39. Godinez-Garrido, G.; Santos-Sanchez, O.J.; Romero-Trejo, H.; Garcia-Perez, O. Discrete integral optimal controller for quadrotor attitude stabilization: Experimental results. *Appl. Sci.* **2023**, *13*, 9293. [\[CrossRef\]](#)
40. Wang, D.; Pan, Q.; Shi, Y.; Hu, J.; Zhao, C. Efficient nonlinear model predictive control for quadrotor trajectory tracking: Algorithms and experiment. *IEEE Trans. Cybern.* **2021**, *51*, 5057–5068. [\[CrossRef\]](#) [\[PubMed\]](#)
41. Verma, S.C.; Li, S.; Savkin, A.V. A hybrid global/reactive algorithm for collision-free UAV navigation in 3D environments with steady and moving obstacles. *Drones* **2023**, *7*, 675. [\[CrossRef\]](#)
42. Wang, Y.; Wang, H.; Liu, Y.; Wu, J.; Lun, Y. 6-DOF UAV Path planning and tracking control for obstacle avoidance: a deep learning-based integrated approach. *Aerosp. Sci. Technol.* **2024**, *151*, 109320. [\[CrossRef\]](#)
43. Rouhi, A.; Arezoomandan, S.; Kapoor, R.; Klohoker, J.; Patal, S.; Shah, P.; Umare, H.; Han, D. An Overview of Deep Learning in UAV Perception. In Proceedings of the 2024 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 6–8 January 2024; pp. 1–6.
44. Cao, H.; Li, Y.; Liu, C.; Zhao, S. ESO-Based Robust and High-Precision Tracking Control for Aerial Manipulation. *IEEE Trans. Autom. Sci. Eng.* **2024**, *21*, 2139–2155. [\[CrossRef\]](#)
45. AlMahamid, F.; Grolinger, K. Autonomous Unmanned Aerial Vehicle navigation using Reinforcement Learning: A systematic review. *Eng. Appl. Artif. Intell.* **2022**, *115*, 105321. [\[CrossRef\]](#)
46. Koutras, D.I.; Kapoutsis, A.C.; Amanatiadis, A.A.; Kosmatopoulos, E.B. Marsexplorer: Exploration of unknown terrains via deep reinforcement learning and procedurally generated environments. *Electronics* **2021**, *10*, 2751. [\[CrossRef\]](#)
47. Wu, L.; Wang, C.; Zhang, P.; Wei, C. Deep Reinforcement Learning with Corrective Feedback for Autonomous UAV Landing on a Mobile Platform. *Drones* **2022**, *6*, 238. [\[CrossRef\]](#)
48. Xue, W.; Wu, H.; Ye, H.; Shao, S. An Improved Proximal Policy Optimization Method for Low-Level Control of a Quadrotor. *Actuators* **2022**, *11*, 105. [\[CrossRef\]](#)
49. Tang, J.; Liang, Y.; Li, K. Dynamic scene path planning of UAVs based on deep reinforcement learning. *Drones* **2024**, *8*, 60. [\[CrossRef\]](#)
50. Ma, Q.; Wu, Y.; Shoukat, M.U.; Yan, Y.; Wang, J.; Yang, L.; Yan, F.; Yan, L. Deep reinforcement learning-based wind disturbance rejection control strategy for uav. *Drones* **2024**, *8*, 632. [\[CrossRef\]](#)
51. Böhn, E.; Coates, E.M.; Moe, S.; Johansen, T.A. Deep Reinforcement Learning Attitude Control of Fixed-Wing UAVs Using Proximal Policy optimization. In Proceedings of the 2019 International Conference on Unmanned Aircraft Systems (ICUAS), Atlanta, GA, USA, 11–14 June 2019; pp. 523–533. [\[CrossRef\]](#)
52. Chen, Y.; Zhang, H.; Li, M.; Wu, H. UAV Path Planning Trends from 2000 to 2024: A Bibliometric Analysis and Visualization. *Drones* **2025**, *9*, 128. [\[CrossRef\]](#)
53. Liu, Z.; Zhou, Z.; Lai, K.; Sim, K.; Ding, D. State-wise Safe Reinforcement Learning: A Survey. In Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI), Macao, China, 19–25 August 2023; pp. 6732–6740. [\[CrossRef\]](#)
54. Feng, S.; Li, X.; Ren, L.; Xu, S. Reinforcement learning with parameterized action space and sparse reward for UAV navigation. *Intell. Robot.* **2023**, *3*, 161–175. [\[CrossRef\]](#)
55. Kim, H.; Choi, J.; Do, H.; Lee, G.T. A Fully Controllable UAV Using Curriculum Learning and Goal-Conditioned Reinforcement Learning: From Straight Forward to Round Trip Missions. *Drones* **2025**, *9*, 26. [\[CrossRef\]](#)
56. Lei, B.; Hu, W.; Ren, Z.; Ji, S. DRL-Based UAV Autonomous Navigation and Obstacle Avoidance with LiDAR and Depth Camera Fusion. *Aerospace* **2025**, *12*, 848. [\[CrossRef\]](#)
57. Zhao, W.; Cui, S.; Qiu, W.; He, Z.; Liu, Z.; Zheng, X.; Mao, B.; Kato, N. A Survey on DRL based UAV Communications and Networking: DRL Fundamentals, Applications and Implementations. *arXiv* **2025**, arXiv:2502.12875. [\[CrossRef\]](#)
58. Jiménez, G.A.; de la Escalera Hueso, A.; Gómez-Silva, M.J. Reinforcement Learning Algorithms for Autonomous Mission Accomplishment by Unmanned Aerial Vehicles: A Comparative View with DQN, SARSA and A2C. *Sensors* **2023**, *23*, 9013. [\[CrossRef\]](#)
59. Zhang, J.; Kim, J.; O'Donoghue, B.; Boyd, S. Sample efficient reinforcement learning with REINFORCE. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 2–9 February 2021; pp. 10887–10895.
60. Zhong, R.; Zhang, D.; Schäfer, L.; Albrecht, S.V.; Hanna, J.P. Robust On-Policy Sampling for Data-Efficient Policy Evaluation in Reinforcement Learning. *arXiv* **2022**, arXiv:2111.14552.
61. Hollenstein, J.; Audy, S.; Saveriano, M.; Renaudo, E.; Piater, J. Action noise in off-policy deep reinforcement learning: Impact on exploration and performance. *arXiv* **2022**, arXiv:2206.03787.
62. Chronis, C.; Anagnostopoulos, G.; Politi, E.; Dimitrakopoulos, G.; Varlamis, I. Dynamic Navigation in Unconstrained Environments Using Reinforcement Learning Algorithms. *IEEE Access* **2023**, *11*, 117984–118001. [\[CrossRef\]](#)
63. Zhou, X.; Wu, P.; Zhang, H.; Guo, W.; Liu, Y. Learn to Navigate: Cooperative Path Planning for Unmanned Surface Vehicles Using Deep Reinforcement Learning. *IEEE Access* **2019**, *7*, 165262–165278. [\[CrossRef\]](#)

64. Gu, J.; Wang, Y. Multi-UAV formation control via deep reinforcement learning and multi-step experience storage in dense obstacle environments. *Trans. Inst. Meas. Control* **2025**, 01423312251387166. [\[CrossRef\]](#)
65. Gui, J.; Yu, T.; Deng, B.; Zhu, X.; Yao, W. Decentralized multi-UAV cooperative exploration using dynamic centroid-based area partition. *Drones* **2023**, *7*, 337. [\[CrossRef\]](#)
66. Ni, J.; Ge, Y.; Zhao, Y.; Gu, Y. An improved multi-UAV area coverage path planning approach based on deep Q-Networks. *Appl. Sci.* **2025**, *15*, 11211. [\[CrossRef\]](#)
67. Thumiger, N.; Deghat, M. A Multi-Agent Deep Reinforcement Learning Approach for Practical Decentralized UAV Collision Avoidance. *IEEE Control Syst. Lett.* **2022**, *6*, 2174–2179. [\[CrossRef\]](#)
68. Xia, Q.; Li, P.; Shi, X.; Li, Q.; Cai, W. Research on Target Capturing of UAV Circumnavigation Formation Based on Deep Reinforcement Learning. In *Proceedings of the 2022 International Conference on Autonomous Unmanned Systems (ICAUS 2022), Xi'an, China, 23–25 September 2022*; Fu, W., Gu, M., Niu, Y., Eds.; Springer: Singapore, 2023; pp. 3751–3762.
69. Cao, Z.; Chen, G. Advanced Cooperative Formation Control in Variable-Sweep Wing UAVs via the MADDPG-VSC Algorithm. *Appl. Sci.* **2024**, *14*, 9048. [\[CrossRef\]](#)
70. Wei, D.; Zhang, L.; Liu, Q.; Chen, H.; Huang, J. UAV Swarm Cooperative Dynamic Target Search: A MAPPO-Based Discrete Optimal Control Method. *Drones* **2024**, *8*, 214. [\[CrossRef\]](#)
71. Hong, X.; Wang, Z.; Wang, Y.; Xue, C.; Gao, Y. Multi-UAV Dynamic Target Search Based on Multi-Potential-Field Fusion Reward Shaping MAPPO. *Drones* **2025**, *9*, 770. [\[CrossRef\]](#)
72. Chen, S.; Zhou, W.; Yang, A.S.; Chen, H.; Li, B.; Wen, C.Y. An End-to-End UAV Simulation Platform for Visual SLAM and Navigation. *Aerospace* **2022**, *9*, 48. [\[CrossRef\]](#)
73. Zhao, J.; Sun, J.; Cai, Z.; Wang, L.; Wang, Y. End-to-end deep reinforcement learning for image-based UAV autonomous control. *Appl. Sci.* **2021**, *11*, 8419. [\[CrossRef\]](#)
74. Dobrzycki, A.D.; Bernardos, A.M.; Casar, J.R. An Analysis of Layer-Freezing Strategies for Enhanced Transfer Learning in YOLO Architectures. *Mathematics* **2025**, *13*, 2539. [\[CrossRef\]](#)
75. Rossi, D.; Borghi, G.; Vezzani, R. TakuNet: An Energy-Efficient CNN for Real-Time Inference on Embedded UAV systems in Emergency Response Scenarios. *arXiv* **2025**, arXiv:2501.05880.
76. Tsintotas, K.A.; Bampis, L.; Gasteratos, A. The revisiting problem in simultaneous localization and mapping: A survey on visual loop closure detection. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 19929–19953. [\[CrossRef\]](#)
77. Munir, A.; Siddiqui, A.J.; Anwar, S.; El-Maleh, A.; Khan, A.H.; Rehman, A. Impact of Adverse Weather and Image Distortions on Vision-Based UAV Detection: A Performance Evaluation of Deep Learning Models. *Drones* **2024**, *8*, 638. [\[CrossRef\]](#)
78. Loquercio, A.; Kaufmann, E.; Ranftl, R.; Dosovitskiy, A.; Koltun, V.; Scaramuzza, D. Deep Drone Racing: From Simulation to Reality With Domain Randomization. *IEEE Trans. Robot.* **2020**, *36*, 1–14. [\[CrossRef\]](#)
79. Chen, Z.; Yan, J.; Ma, B.; Shi, K.; Yu, Q.; Yuan, W. A Survey on Open-Source Simulation Platforms for Multi-Copter UAV Swarms. *Robotics* **2023**, *12*, 53. [\[CrossRef\]](#)
80. Song, Y.; Naji, S.; Kaufmann, E.; Loquercio, A.; Scaramuzza, D. Flightmare: A Flexible Quadrotor Simulator. In *Proceedings of the 2020 Conference on Robot Learning, Virtual, 16–18 November 2020*; Kober, J.; Ramos, F., Tomlin, C., Eds.; Proceedings of Machine Learning Research: Cambridge, MA, USA, 2021; Volume 155, pp. 1147–1157.
81. Chao, Y.; Dillmann, R.; Roennau, A.; Xiong, Z. E-DQN-Based Path Planning Method for Drones in Airsim Simulator under Unknown Environment. *Biomimetics* **2024**, *9*, 238. [\[CrossRef\]](#)
82. Böhn, E.; Coates, E.M.; Reinhardt, D.; Johansen, T.A. Data-Efficient Deep Reinforcement Learning for Attitude Control of Fixed-Wing UAVs: Field Experiments. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 3168–3180. [\[CrossRef\]](#)
83. Jembre, Y.Z.; Nugroho, Y.W.; Khan, M.T.R.; Attique, M.; Paul, R.; Shah, S.H.A.; Kim, B. Evaluation of Reinforcement and Deep Learning Algorithms in Controlling Unmanned Aerial Vehicles. *Appl. Sci.* **2021**, *11*, 7240. [\[CrossRef\]](#)
84. Xu, J.; Du, T.; Foshey, M.; Li, B.; Zhu, B.; Schulz, A.; Matusik, W. Learning to fly: Computational controller design for hybrid UAVs with reinforcement learning. *ACM Trans. Graph.* **2019**, *38*, 42. [\[CrossRef\]](#)
85. Mao, X.; Wu, G.; Fan, M.; Cao, Z.; Pedrycz, W. DL-DRL: A Double-Level Deep Reinforcement Learning Approach for Large-Scale Task Scheduling of Multi-UAV. *IEEE Trans. Autom. Sci. Eng.* **2025**, *22*, 1028–1044. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.