

Autonomous Unmanned Aerial Vehicle Navigation using Reinforcement Learning: A Systematic Review

Fadi AlMahamid, Katarina Grolinger*

Department of Electrical and Computer Engineering, Western University, London, Ontario, Canada

ARTICLE INFO

Keywords:
Reinforcement Learning
Autonomous UAV Navigation
UAV
Systematic Review

ABSTRACT

There is an increasing demand for using Unmanned Aerial Vehicle (UAV), known as drones, in different applications such as packages delivery, traffic monitoring, search and rescue operations, and military combat engagements. In all of these applications, the UAV is used to navigate the environment autonomously - without human interaction, perform specific tasks and avoid obstacles. Autonomous UAV navigation is commonly accomplished using Reinforcement Learning (RL), where agents act as experts in a domain to navigate the environment while avoiding obstacles. Understanding the navigation environment and algorithmic limitations plays an essential role in choosing the appropriate RL algorithm to solve the navigation problem effectively. Consequently, this study first identifies the main UAV navigation tasks and discusses navigation frameworks and simulation software. Next, RL algorithms are classified and discussed based on the environment, algorithm characteristics, abilities, and applications in different UAV navigation problems, which will help the practitioners and researchers select the appropriate RL algorithms for their UAV navigation use cases. Moreover, identified gaps and opportunities will drive UAV navigation research.

Symbol	Description
$s \in S$	State s belongs to all possible states S
$a \in A$	Action a belongs to the set of all possible Actions A
$r \in R$	Reward r belongs to the set of all generated Rewards R
γ	Discounted factor γ decreases the contribution of the future rewards, where $0 < \gamma < 1$
G_t	The Expected Summation of the Discounted Rewards; $G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$
$P(s', r s, a)$	The probability of the transition to state s' with reward r from taking action a in state s at time t
τ	A trajectory τ consists of a sequence of actions and states pairs, where the actions influence the states, also called an episode. Each trajectory has a start state and ends in a final state that terminates the trajectory
$Q(s, a)$	Action-value function expresses the expected return of the state-action pairs (s, a) ; $Q^w(s, a)$ is $Q(s, a)$ parameterized by w

Symbol	Description
$V(s)$	State-value function is similar to $Q(s, a)$ except it measures how good to be in a state s ; $V^w(s)$ is a State-value function parameterized by w
$A(s, a)$	Advantage-Value function $A(s, a)$ measures how good an action is in comparison to alternative actions at a given state; $A(s, a) = Q(s, a) - V(s)$
$\pi(a s)$	Stochastic Policy π is a function that maps the probability of selecting an action a from the state s . It describes agent behavior
$\mu(s)$	Deterministic Policy μ is similar to Stochastic Policy π , except μ symbol is used to distinguish it from Stochastic Policy π
$Q_\pi(s, a)$	Action-value function $Q(s, a)$, when following a policy π
$V_\pi(s)$	State-value function $V(s)$, when following a policy π
ρ	State visitation probability
\sim	Sampled from. For example, $s \sim \rho$ means s sampled from state visitation probability ρ
$\eta(\pi)$	The expected discounted reward following a policy π , similar to G_t

*Corresponding author.

✉ fadalmahamid@uwo.ca (F. AlMahamid); kgrolinger@uwo.ca (K. Grolinger)

ORCID(s): 0000-0002-6907-7626 (F. AlMahamid); 0000-0003-0062-8212 (K. Grolinger)

<https://doi.org/10.1016/j.engappai.2022.105321>

The article submitted 5 February 2022 to *Engineering Applications of Artificial Intelligence Journal*; Received in revised form 12 July 2022; Accepted 4 August 2022.

0952-1976/© 2022 Elsevier Ltd. All rights reserved.

1. Introduction

Autonomous Systems (AS) are systems that can perform desired tasks without human interference, such as robots performing tasks without human involvement, self-driving cars, and delivery drones. AS are invading different domains to make operations more efficient and reduce the cost and risk incurred from the human factor.

An Unmanned Aerial Vehicle (UAV) is an aircraft without a human pilot, mainly known as a drone. Autonomous

UAVs have been receiving an increasing interest due to their diverse applications, such as delivering packages to customers, responding to traffic collisions to attain injured with medical needs, tracking military targets, assisting with search and rescue operations, and many other applications.

Typically, UAVs are equipped with cameras, among other sensors, that collect information from the surrounding environment, enabling UAVs to navigate that environment autonomously. UAV navigation training is typically conducted in a virtual 3D environment because UAVs have limited computation resources and power supply, and replacing UAV parts due to crashes can be expensive.

Different Reinforcement Learning (RL) algorithms are used to train UAVs to navigate the environment autonomously. RL can solve various problems where the agent acts as a human expert in the domain. The agent interacts with the environment by processing the environment's state, responding with an action, and receiving a reward. UAV cameras and sensors capture information from the environment for state representation. The agent processes the captured state and outputs an action that determines the UAV movement's direction or controls the propellers' thrust, as illustrated in Figure 1.

The research community provided a review of different UAV navigation problems, such as Visual UAV navigation [1, 2], UAV Flocking [3] and Path Planning [4]. Nevertheless, to the best of the authors' knowledge there is no survey related to applications of RL in UAV navigation. Hence, this paper aims to provide a comprehensive and systematic review on the application of various RL algorithms to different autonomous UAV navigation problems. This survey has the following contributions:

- Help the practitioners and researchers to select the right algorithm to solve the problem on hand based on the application area and environment type.
- Explain primary principles and characteristics of various RL algorithms, identify relationships among them, and classify them according to the environment type.

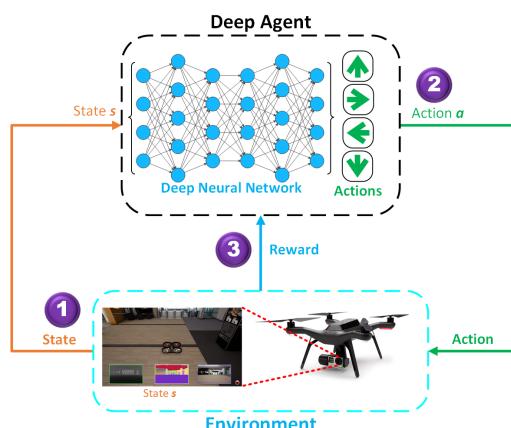


Figure 1: UAV training using deep reinforcement agent

- Discuss and classify different RL UAV navigation frameworks according to the problem domain.
- Recognize the various techniques used to solve different UAV autonomous navigation problems and the different simulation tools used to perform UAV navigation tasks.

The remainder of the paper is organized as follows: Section 2 presents the systematic review process, Section 3 introduces RL, Section 4 provides a comprehensive review of the application of various RL algorithms and techniques in autonomous UAV navigation, Section 5 discusses the UAV Navigation Frameworks and simulation software, Section 6 classifies RL algorithm and discusses the most prominent algorithms, Section 7 explains RL algorithms selection process, and Section 8 identifies challenges and research opportunities. Finally, Section 9 concludes the paper.

2. Review Process

This section described the inclusion criteria, paper identification process, and threats to validity.

2.1. Inclusion Criteria and Identification of Papers

The study's main objective is to analyze the application of Reinforcement Learning in UAV navigation and provide insights into RL algorithms. Therefore, the survey considered all papers in the past five years (2016-2021) written in the English language that include the following terms combined, alongside with their variations: *Reinforcement Learning*, *Navigation*, and *UAV*.

In contrast, RL algorithms are listed based on the authors' domain knowledge of the most prominent algorithms and by going through the related work of the identified algorithms with no restriction to the publication time to include a large number of algorithms.

The identification process of the papers went through the following stages:

- First stage: The authors identified all studies that strictly applied RL to UAV Navigation and acknowledged that model-free RL is typically utilized to tackle UAV navigation challenges, except for a single article [5] that employs model-based RL. Therefore, the authors choose to concentrate on model-free RL and exclude research irrelevant to UAV Navigation, such as UAV networks and traditional optimization tools and techniques [6–10].
- Second stage: The authors listed all RL algorithms based on authors' knowledge of the most prominent algorithms, the references of recognized algorithms, then identified the corresponding paper of each algorithm.
- Third stage: the authors identified how RL is used to solve different UAV navigation problems, classified the work, and then recognized more related papers using exiting work references.

IEEE Xplore and Scopus were the primary sources of papers' identification between 2016 and 2021. The search query was applied using different terminologies that are used

to describe the UAV alternatively, such as *UNMANNED AERIAL VEHICLE, DRONE, QUADCOPTER, or QUADTOR*, and these terms are cross-checked with *REINFORCEMENT LEARNING*, and *NAVIGATION*, which resulted in a total of 104 papers. After removing 15 duplicate papers and 5 unrelated papers, the count became 84.

The authors identified another 75 papers that mainly describe the RL algorithms based on the authors' experience and the references list of the recognized work, using Google Scholar as the primary search engine. While RL for UAV navigation studies were restricted to five years, all RL algorithms are included as many are still extensively used regardless of their age. The search was completed in November 2021, with a total of 159 papers after all exclusions.

2.2. Threats to Validity

Despite the authors' effort to include all relevant papers, the study might be subject to the following main threats to the validity:

- Location bias: The search for papers was performed using two primary digital libraries (databases), IEEE Xplore and Scopus, which might limit the retrieved papers based on the published journals, conferences, and workshops in the database.
- Language bias: Only papers published in English are included.
- Time Bias: The search query is only limited to retrieving papers between 2016 and 2021, which results in excluding relevant papers published before 2016.
- Knowledge reporting bias: The research papers of RL algorithms are identified using authors' knowledge of variant algorithms and the related work in the recognized algorithms. It is hard to pinpoint all algorithms utilizing a search query, which could result in missing some RL algorithms.

3. Reinforcement Learning

RL can be explained using the Markov Decision Process (MDP), where a RL agent learns through experience by taking actions in the environment, causing a change in the environment's state, and receiving a reward for the action taken to measure the success or failure of the action. Equation 1 defines the transition probability from state s by taking the action a to the new state s' with a reward r , for all $s' \in S$, $s \in S$, $r \in R$, $a \in A(s)$ [11].

$$P(s', r|s, a) = Pr\{S_t = s', R_t = r | S_{t-1} = s, A_{t-1} = a\} \quad (1)$$

The reward R is generated using a reward function, which can be expressed as a function of the action $R(a)$, or as a function of action-state pairs $R(a, s)$. The reward helps the agent learn good actions from bad actions, and as the agent accumulates experience, it starts taking more successful actions and avoiding bad ones [11].

All actions the agent takes from a start state to a final (terminal) state make an *episode* (trajectory). The goal of

MDP is to maximize the expected summation of the discounted rewards by adding all the rewards generated from an episode. However, sometimes the environment has an infinite horizon, where the actions cannot be divided into episodes. Therefore, using a discounted factor (multiplier) γ to the power k , where $\gamma \in [0, 1]$ as expressed in Equation 2 helps the agent to emphasize the reward at the current time step and reduce the reward value granted at future time steps, and, moreover, helps the expected summation of discounted rewards to converge if the horizon is infinite [11].

$$G_t = E \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \right] \quad (2)$$

The following subsections introduce important reinforcement learning concepts.

3.1. Policy and Value Function

A policy π defines the agent's behavior by defining the probability of taking action a while being in a state s , which is expressed as $\pi(a|s)$. The agent evaluates its behavior (action) using a value function, which can be either *state-value function*, which estimates how good it is to be in state s after executing an action a , or using a *action-value function* that measures how good it is to select action a while being in a state s . The value produced by the action-value function in Equation 3 is known as the Q-value and is expressed in terms of the expected summation of the discounted rewards [11].

$$Q_\pi(s, a) = E_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a \right] \quad (3)$$

Since the objective is to maximize the expected summation of discounted rewards under the optimal policy π , the agent tries to find the optimal Q-value $Q_*(s, a)$ as defined in Equation 4. This optimal Q-value must satisfy the *Bellman Optimality Equation* 5 which is defined as the sum of the expected reward received from executing the current action R_{t+1} , and sum of all future rewards (discounted) received from any possible future state-action pairs (s', a') . In other words, the agent tries to select the actions that grant the highest rewards in an episode. In general, selecting the optimal value means selecting the action with the highest Q-value; however, the action with the highest Q-value sometimes might not lead to better rewarding actions in the future [11].

$$Q_*(s, a) = \max_{\pi} Q_{\pi}(s, a) \quad (4)$$

$$Q_*(s, a) = E \left[R_{t+1} + \gamma \max_{a'} Q_*(s', a') \right] \quad (5)$$

3.2. Exploration vs Exploitation

Exploration vs. Exploitation may be demonstrated using the multi-armed bandit dilemma, which accurately portrays the behavior of a person experiencing their first slot machine experience. The money (reward) player receives early in the game is unrelated to any previously selected choices, and as the player develops a comprehension of the reward, he/she begins selecting choices that contribute to earning a greater reward. The choices made randomly by the player to acquire knowledge might be defined as the player *Exploring* the environment. In contrast, the player's *Exploiting* the environment is described as the options selected based on his/her experience.

The RL agent needs to find the right balance between exploration and exploitation to maximize the expected return of rewards. Constantly exploiting the environment and selecting the action with the highest reward does not guarantee that the agent performs the optimal action because the agent may miss out on a higher reward provided by future actions taking alternative sets of actions in the future. Finding the ratio between exploration and exploitation can be defined through different strategies such as ϵ -greedy strategy, Upper Confidence Bound (UCB), and Gradient Bandits [12].

3.3. Experience Replay

RL agent does not need data to learn; rather, it learns from experiences by interacting with the environment. The agent experience e can be formulated as tuple $e(s, a, s', r)$, which describes the agent taking an action a at a given state s and receiving a reward r for the performed action and causing a new state s' . *Experience Replay (ER)* [13] is a technique that suggests storing experiences in a replay memory (buffer) D and using a batch of uniformly sampled experiences for RL agent training.

On the other hand, *Prioritized Experience Replay (PER)* [14] prioritizes experiences according to their significance using *Temporal Difference error (TD-error)* and replays experiences with lower TD-error to repeatedly train the agent, which improves the convergence.

3.4. On-Policy vs Off-Policy

In order to interact with the environment, the RL agent attempts to learn two policies: the first one is referred to as the target policy $\theta(a|s)$, which the agent learns through the value function, and the second one is referred to as the behavior policy $\beta(a|s)$, which the agent uses for action selection when interacting with the environment.

A RL algorithm is referred to as *on-policy algorithm* when the same target policy $\theta(a|s)$ is employed to collect the training sample and to determine the expected return. In contrast, *off-policy algorithms* are those where the training sample is collected in accordance to the behavior policy $\beta(a|s)$, and the expected reward is generated using the target policy $\theta(a|s)$ [15]. Another main difference is that Off-policy algorithms can reuse past experiences and do not require all the experiences within an episode (full episode) to generate training samples, and the experiences can be collected from different episodes.

3.5. Deep Reinforcement Learning

Deep Reinforcement Learning (DRL) uses deep agents to learn the optimal policy where it combines artificial Neural Networks (NN) with Reinforcement Learning (RL). The NN type used in DRL varies from one application to another depending on the problem being solved, inputs type (state), and the number of inputs passed to the NN. For example, the RL framework can be integrated with Convolutional Neural Network (CNN) to process images representing the environment's state or combined with Recurrent Neural Network (RNN) to process inputs over different time steps.

The NN loss function, also known as the Temporal Difference (TD), is generically computed by finding the difference between the output of the NN $Q(s, a)$ and the optimal Q-value $Q_*(s, a)$ obtained from the Bellman equation as shown in Equation 6 [11]:

$$\underbrace{E \left[R_{t+1} + \gamma \max_{a'} Q_*(s', a') \right]}_{\text{Target}} - \underbrace{E \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \right]}_{\text{Predicted}} \quad (6)$$

The architecture of the deep agent can be simple or complex based on the problem at hand, where a complex architecture combines multiple NN. But what all deep agents have in common is that they receive the state as an input, then they output the optimal action and maximize the discounted return of rewards.

The application of Deep NN to the RL framework enabled the research community to solve more complex problems in autonomous systems that were hard to solve before and achieve better performance than previous state-of-the-art, such as drone navigation and avoiding obstacles using images received from the drone's monocular camera.

4. Autonomous UAV Navigation using DRL

Different DRL algorithms and techniques were used to solve various problems in autonomous UAV navigation, such as UAV control, obstacle avoidance, path planning, and flocking. The DRL agent acted as an expert in all of these problems, selecting the best action that maximizes the reward to achieve the desired objective. The input and the output of the DRL algorithm are generally determined based on the desired objective and the implemented technique.

RL agent design for UAV navigation depicted in Figure 2 shows different UAV input devices used to capture the state processed by the RL agent. The agent produces action values that can be either the movement values of the UAV or the waypoint values where the UAV needs to relocate. Once the agent executes the action in the environment, it receives the new state and the generated reward based on the performed action. The reward function is designed to generate the reward subject to the intended objective while using various information from the environment. The agent design ('Agent' box in the figure) is influenced by the RL algorithms discussed in Section 6 where the agent components and inner working varies from one algorithm to another.

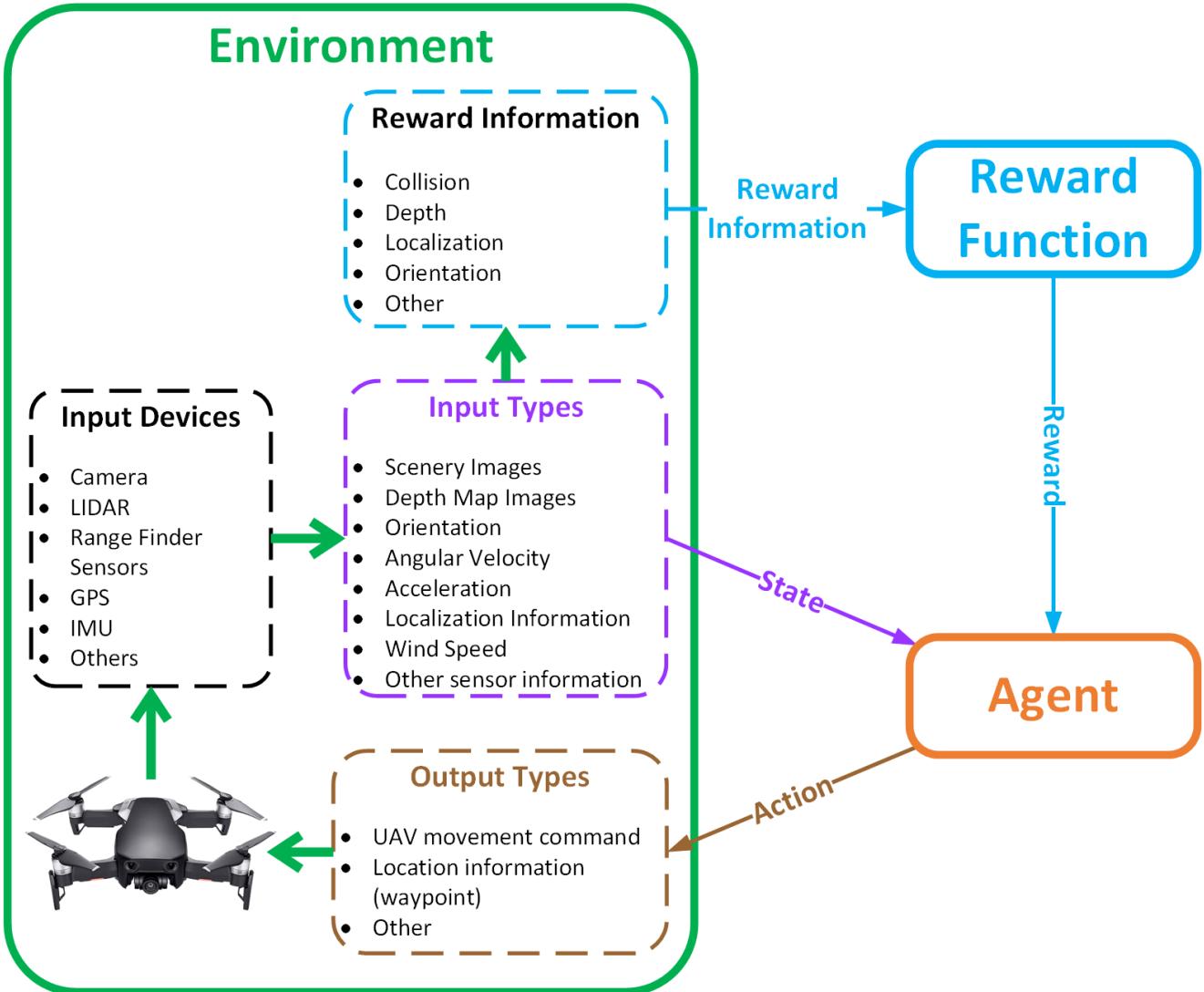


Figure 2: RL agent design for UAV navigation task

Table 1 summarizes the application of RL to different UAV navigation tasks (objectives), and the following subsections discuss the UAV navigation tasks in more detail. As seen from this table, most of the research focused on two UAV navigation objectives: 1) Obstacle avoidance using various UAV sensor devices such as cameras and LIDARs and 2) Path planning to find the optimal or shortest route.

4.1. UAV Control

RL is used to control the movement of the UAV in the environment by applying changes to the flight mechanics of the UAV, which varies based on the UAV type. In general UAVs can be classified based on the flight mechanics into 1) Multirotor, 2) Fixed-Wing, and 3) single-rotor, and 4) fixed-wing hybrid Vertical Take-Off and Landing (VTOL) [92].

Multirotor, also known as multicopter or drone, uses more than two rotors to control the flight mechanics by applying different amounts of thrust to the rotors causing changes in principal axes leading to four UAV movements

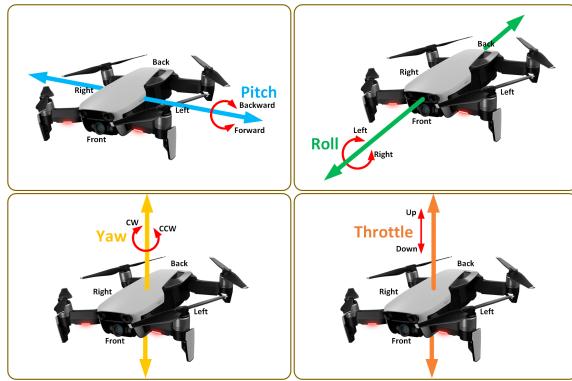
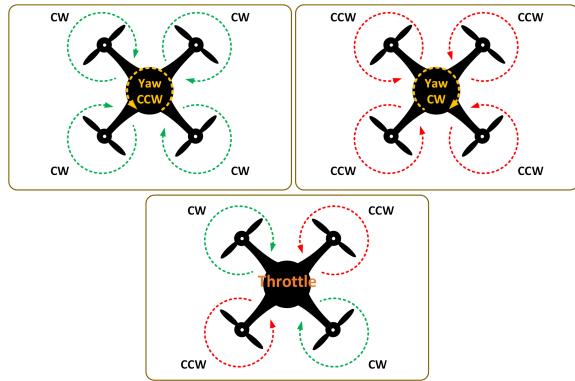
1) pitch, 2) roll, 3) yaw, and 4) throttle as explained in Figure 3. Similarly, single-rotor and fixed-wing hybrid VTOL apply changes to different rotors to generate the desired movement, except they both use tilt-rotor(s) and wings in fixed-wing hybrid VTOL. On the other hand, fixed-wing can only achieve three actions pitch, roll, and yaw, where they take off by generating enough speed that causes the air-dynamics to lift-up the UAV.

Quad-rotors has four propellers: two diagonal propellers rotate clockwise and the other two propellers rotate counter-clockwise causing the throttle action. When the propellers generate a thrust more significant than the UAV weight they cause elevation, and when the thrust power equals the UAV weight, the UAV stops elevation and starts hovering in place. In contrast, if all propellers rotate in the same direction, they cause a yaw action in the opposite direction, as shown in Figure 4.

Table 1

DRL application to different UAV Navigation tasks

Objective	Sub-Objective	Paper
UAV Control	Controlling UAV flying behavior (attitude control)	[5, 16–22]
Obstacle Avoidance	Obstacle avoidance using images and sensor information	[21, 23–58]
	Obstacle avoidance while considering the battery level	[59]
Path Planning	Local and global path planning (finding the shortest/optimal route)	[24, 26, 44, 51, 55, 58, 60–70]
	Path planning while considering the battery level	[24, 71, 72]
	Find fixed or moving targets (points of interest)	[58, 67, 73–77]
	Landing the UAV on a selected point	[78–80]
Flocking	Maintain speed and orientation with other UAVs (formation)	[39, 81–83]
	Obstacle avoidance	[83, 84]
	Target tracking	[85–89]
	Flocking while considering the battery level	[86]
	Covering geographical region	[86, 90]
	Path planning and finding the safest route	[83, 91]

**Figure 3:** Multirotor Flight Mechanics**Figure 4:** Yaw vs Throttle Mechanics

The steps described in Figure 5 depicts the RL process used to control the UAV, which depends on the used RL algorithm, but the most important takeaway is that RL uses the UAV state to produce actions. These actions are responsible for moving the UAV in the environment and can be either direct changes in the value of pitch, roll, yaw, and throttle values or indirect changes that require transformation to commands understood by the UAV.

4.2. Obstacle Avoidance

Avoiding obstacles is an essential task required by the UAV to navigate any environment, which can be achieved by estimating the distance to the objects in the environment using different devices such as front-facing cameras or distance sensors. The output generated by these different devices provides input to the RL algorithm and plays a significant role in the NN architecture.

Lu *et al.* [1] described different front-facing cameras such as monocular cameras, stereo cameras, and RGB-D cameras that a UAV can use. Each camera type produces a different image type used as raw input to the RL agent. However, regardless of the camera type, these images can be preprocessed using computer vision to produce specific image types as described below:

- **RGB Images:** are renowned colored images where each pixel is represented in three values (Red, Green, Blue) ranging between (0, 255).
- **Depth-Map Images:** contains information related to the distance of the objects from the Field Of View (FOV).
- **Event-Based Images:** are special images that output the changes in brightness intensity instead of standard images. Event-based images are produced by an event camera, also known as Dynamic Vision Sensor (DVS).

RGB images lack depth information, and, therefore, the agent cannot estimate how far or close the UAV is to the object leading to unexpected flying behavior. On the other hand, depth information is essential for building a successful reward function that penalizes moving closer to the objects. Some techniques used RGB images and depth-map simultaneously as input to the agent to provide more information about the environment. In contrast, event-based images data are represented as one-dimensional sequences of events over time, which is used to capture quickly changing information in the scene [23].

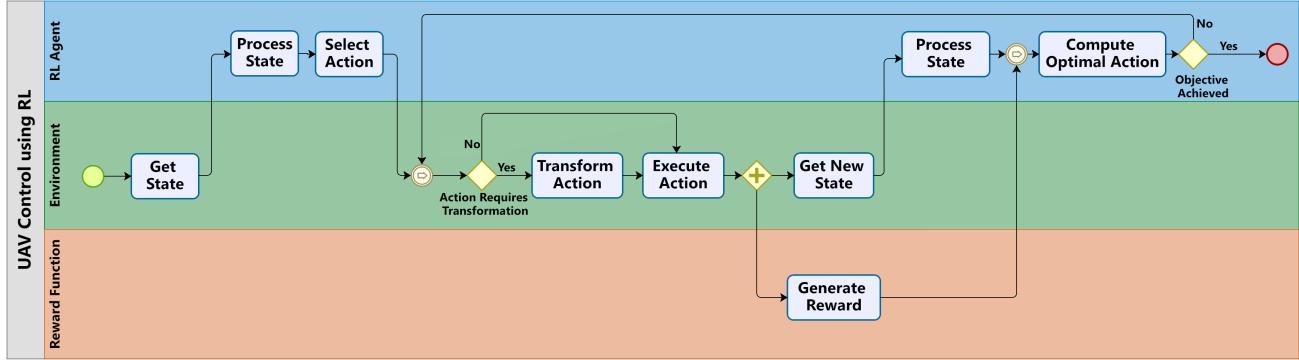


Figure 5: UAV Control using RL

Similar to cameras, distance sensors have different types, such as LiDAR, RADAR, and acoustic sensors: they estimate the distance of the surrounding objects to the UAV but require less storage size than 2D images since they do not use RGB channels.

The output generated by these devices reflects the different states that the UAV has over time, used as an input to the RL agent to make actions causing the UAV to move in different directions to avoid obstacles. The NN architecture of the RL agent is based on: 1) input type, 2) the number of inputs, and 3) the used algorithm. For example, processing RGB images or depth-map images using the DQN algorithm requires Convolutional Neural Network (CNN) followed by fully-connect layers since CNN is known for its power in processing images. In contrast, processing event-based images is performed using Spiking Neural Networks (SNN), which is designed to handle spatio-temporal data and identify spatio-temporal patterns [23].

4.3. Path Planning

Autonomous UAVs must have a well-defined objective before executing a flying mission. Typically, the goal is to fly from a start to a destination point, such as in delivery drones. But, the goal can also be more sophisticated, such as performing surveillance by hovering over a geographical area or participating in search and rescue operations to find a missing person.

Autonomous UAV navigation requires path planning to find the best UAV path to achieve the flying objective while avoiding obstacles. The optimal path does not always mean the shortest path or a straight line between two points; instead, the UAV aims to find a safe path while considering UAV's limited power and flying mission.

Path planning can be divided into two main types:

- Global Path Planning: concerned with planning the path from the start point to destination point in attempt to select the optimal path.
- Local Path Planning: concerned with planning the local optimal waypoints in an attempt to avoid static and dynamic obstacles while considering the final destination.

Path planning can be solved using different techniques and algorithms; in this work, we focus on RL techniques used to solve global and local path planning, where the RL agent receives information from the environment and outputs the optimal waypoints according to the reward function. RL techniques can be classified according to the usage of the environment's local information 1) map-based navigation and 2) mapless navigation.

4.3.1. Map-Based Navigation

A UAV that adopts map-based navigation uses a representation of the environment either in 3D or 2D format. The representation might include one or more of the following about the environment: 1) the different terrains, 2) fixed-obstacles locations, and 3) charging/ground stations.

Some maps oversimplify the environment representation: the map is divided into a grid with equally-sized smaller cells that store information about the environment [68, 73, 93]. Others oversimplify the environment's structure by simplifying objects representation or by using 1D/2D to represent the environment [34, 36, 39, 41, 47, 55, 65, 67, 74, 86, 87]. The UAV has to plan a safe and optimal path over the cells to avoid cells containing obstacles until it reaches its destination and has to plan its stopover at the charging stations based on the battery level and path length.

In a more realistic scenario, the UAV calculates a route using the map information and the GPS signal to track the UAV's current location, starting point, and destination point. The RL agent evaluates the change in the distance and the angle between the UAV's current GPS location and target GPS location, and penalizes the reward if the difference increases or if the path is unsafe depending on the reward function (objective).

4.3.2. Mapless Navigation

Mapless navigation does not rely on maps; instead, it applies computer vision techniques to extract features from the environment and learn the different patterns to reach the destination, which requires computation resources that might be overwhelming for some UAVs.

Localization information of the UAV obtained by different means such as Global Positioning System (GPS)

or Inertial Measurement Unit (IMU) is used in mapless navigation to plan the optimal path. DRL agent receives the scene image, the destination target, and the localization information as input and outputs the change in the UAV movements.

For example, Zhou *et al.* [50] calculated and tracked the angle between the UAV and destination point, then encoded it with the depth image extracted from the scene and used both as a state representation for the DRL agent. Although localization information seems essential to plan the path, some techniques achieved navigation with high speed using monocular visual reactive navigation system without a GPS [94].

4.4. Flocking

Although UAVs are known for performing individual tasks, they can flock to perform tasks efficiently and quickly, which requires maintaining flight formation. UAV flocking has many applications, such as search and rescue operations to cover a wide geographical area.

UAV flocking is considered a more sophisticated task than a single UAV flying mission because UAVs need to orchestrate their flight to maintain flight formation while performing other tasks such as UAV control and obstacle avoidance. Flocking can be executed using different topologies:

- Flock Centering: maintaining flight formation as suggested by Reynolds [95] involves three concepts: 1) flock centering, 2) avoiding obstacles, and 3) velocity matching. This topology was applied in several research papers [82, 96–99].
- Leader-Follower Flocking: the flock leader has its mission of reaching destination, while the followers (other UAVs) flock with the leader with a mission of maintaining distance and relative position to the leader [100, 101].
- Neighbors Flocking: close neighbors coordinate with each other, where each UAV communicates with two or more nearest neighbors to maintain flight formation by maintaining relative distance and angle to the neighbors [81, 102, 103].

Maintaining flight formation using RL requires communication between UAVs to learn the best policy to maintain the formation while avoiding obstacles. These RL systems can be trained using a single agent or multi-agents in centralized or distributed settings.

4.4.1. Centralized Training

A centralized RL agent trains a shared flocking policy to maintain the flock formation using the experience collected from all UAVs, while each UAV acts individually according to its local environment information such as obstacles. The reward function of the centralized agent can be customized to serve the flocking topology, such as flock centering or leader-follower flocking.

Yan *et al.* [39] used Proximal Policy Optimization (PPO) algorithm to train a centralized shared flocking control policy, where each UAV flocks as close as possible to the center

and decentralized execution for obstacle avoidance according to each UAV local environment information. Similarly, Hung and Givigi [101] trained a leader UAV to reach a destination while avoiding obstacles and trained a shared policy for followers to flock with the leader considering the relative dynamics between the leader and the followers.

Zhao *et al.* [57] used a Multi-Agent Reinforcement Learning (MARL) to train a centralized flock control policy shared by all UAVs with decentralized execution. MARL received position, speed, and flight path angle from all UAVs at each time step and tried to find the optimal flocking control policy.

The centralized training would not produce a good generalization in neighbors flocking topology since the learned policy for one neighbor is different from other neighbors' policies due to the differences in neighbors' dynamics.

4.4.2. Distributed Training

UAV flocking can be trained using a distributed (decentralized) approach, where each UAV has its designated RL agent responsible for finding the optimal flock policy for the UAV. The reward function is defined to maintain distance and flying direction with other UAVs and can be customized to include further information depending on the objective.

Flight information such as location and heading angle should be communicated to other UAVs since the RL agents are distributed, and the state representation must include information of other UAVs. Any UAV that fails to receive the information from other UAVs will cause the UAV to be isolated from the flock.

Liu *et al.* [86] proposed a decentralized DRL framework to control each UAV in a distributed setting to maximize average coverage score, geographical fairness, and minimize UAVs' energy consumption.

5. UAV Navigation Frameworks and Simulation Software

Subsection 5.1 discusses and classifies the UAV navigation frameworks based on the UAV navigation objectives/sub-objectives explained in Section 4, and identifies categories such as Path Planning Frameworks, Flocking Frameworks, Energy-Aware UAV Navigation Frameworks, and others. On the other hand, Subsection 5.2 explains the simulation software's components and the most common simulation software utilized to perform the experiments.

5.1. UAV Navigation Frameworks

In general, a software framework is a conceptual structure analogous to a blueprint used to guide the comprehending construction of the software by defining different functions and their interrelationships. By definition, RL can be considered a framework by itself. Therefore, we considered only UAV navigation frameworks that add to traditional navigation using sensors or camera data for navigation. As a result, Table 2 classifies UAV frameworks based on the framework objective. The subsequent sections discuss the frameworks in more detail.

Table 2
UAV Navigation Frameworks

Framework Objective	Papers
Energy-aware UAV Navigation	[59, 71]
Path Planning	[30, 32, 47, 60, 62, 66, 70]
Flocking	[61, 91]
Vision-Based Frameworks	[42, 43, 73, 77]
Transfer Learning	[40]

5.1.1. Energy-Aware UAV Navigation Frameworks

UAVs have limited flight time, hence operate mainly using batteries. Therefore, planning flight route and recharge stopover are crucial to reach destinations. Energy-aware UAV navigation frameworks aim to provide obstacles avoidance navigation while considering the UAV battery capacity.

Bouhamed *et al.* [59] developed a framework based on Deep Deterministic Policy Gradient (DDPG) algorithm to guide the UAV to a target position while communicating with ground stations, allowing the UAV to recharge its battery if it drops below a specific threshold. Similarly, Imanberdiyev *et al.* [71] monitor battery level, rotors' condition, and sensor readings to plan the route and apply necessary route changes for required battery charging.

5.1.2. Path Planning Frameworks

Path planning is the process of determining the most efficient route that meets the flight objective, such as finding the shortest, fastest, or safest route. Different frameworks [47, 60] implemented a modular path planning scheme, where each module has a specialized function to achieve while exchanging data with other modules to train action selection policies and discover the optimal path.

Similarly, Li *et al.* [32] developed a four-layer framework in which each layer generates a set of objective and constraint functions. The functions are intended to serve the lower layer and consider the upper layer's objectives and constraints, with their primary goal generating trajectories.

Other frameworks suggested stage-based learning to choose actions from the desired stage depending on current environment encounters. For example, Camci and Kayacan [66] proposed learning a set of motion primitives offline, then using them online to design quick maneuvers to enable switching seamlessly between two modes: *near-hover motions*, which is responsible for generating motion plans allowing a stable completion of maneuvers and *swift maneuvers* to deal smoothly with abrupt inputs.

In a collaborative setting, Zhang *et al.* [62] suggested a coalition between Unmanned Ground Vehicle (UGV) and UAV complementing each other to reach the destination, where UAV cannot get to far locations alone due to limited battery power, and UGV cannot reach high altitude due to limited abilities.

5.1.3. Flocking Frameworks

UAV flocking frameworks have functionality beyond UAV navigation while maintaining flight formation. For

example, Bouhamed *et al.* [61] presented a RL-based spatiotemporal scheduling system for autonomous UAVs. The system enables UAVs to autonomously arrange their schedules to cover the most significant number of pre-scheduled events physically and temporally spread throughout a specified geographical region and time horizon. On the other hand, Majd *et al.* [91] predicted the movement of drones and dynamic obstacles in the flying zone to generate efficient and safe routes.

5.1.4. Vision-Based Frameworks

Vision-Based Framework depends on UAV camera for navigation, where the images produced by the camera are used to draw on additional functionality for improved navigation. It is possible to employ frameworks that augment the agent's CNN architecture to fuse data from several sensors, use Long-Short Term Memory cells (LSTM) to maintain navigation choices, use RNN to capture the UAV states over different time steps, or pre-process images to provide more information about the environment [42, 42, 43].

5.1.5. Transfer Learning Frameworks

UAVs are trained on target environments before executing the flight mission; the training is carried either in a virtual or real-world environment. The UAV requires retraining when introduced to new environments or moving from virtual training as the environments have different terrains and obstacle structures or textures. Besides, UAV training requires a long time and it is hardware resource intensive while actual UAVs have limited hardware resources. Therefore, when UAV is introduced to new environments, transfer learning frameworks reduce the training time by reusing the NN weights trained from the previous environment and retraining only parts of the agent's NN.

Yoon *et al.* [40] proposed algorithm-hardware co-design, where the UAV is trained in a virtual environment, and after the UAV is deployed to a real-world environment; the agent loads the weights stored in embedded Non-Volatile Memory (eNVM), and then evaluates new actions and only trains the last few layers of CNN whose weights are stored in the on-die SRAM (Static Random Access Memory).

5.2. Simulation Software

The research community used different evaluation methods for autonomous UAV navigation using RL. Simulation software is used widely over actual UAVs to execute the evaluation due to the cost of the hardware (drone) in addition to the cost of replacement parts required due to UAV crashes. Comparison between simulation software is not the intended purpose, rather than making the research community aware of the most commonly used tools for evaluation as illustrated in Figure 6. 3D UAV navigation simulation requires mainly three components as illustrated in Figure 7:

- **RL Agent:** represents the RL algorithm used with all computations required to generate the reward, process the states, and compute the optimal action. RL Agent interacts directly with the *UAV Flight simulator* to send/receive UAV actions/states.

- **UAV Flight Simulator:** responsible for simulating the UAV movements and interactions with the 3D environment, such as obtaining images from the UAV camera or reporting UAV crashes with different obstacles. Examples of UAV flight simulators are Robot Operating systems (ROS) [104] and Microsoft AirSim [105].
- **3D Graphics Engine:** provides a 3D graphics environment with the physics engine, which is responsible for simulating the gravity and dynamics similar to the real world. Examples of 3D graphics engines are Gazebo [106] and Unreal Engine [107].

Due to compatibility/support issues, ROS is used in conjunction with Gazebo, where AirSim uses Unreal Engine to run the simulations. However, the three components might not always be present, especially if the simulation software has internal modules or plugins that provide the required functionality, such as MATLAB.

6. Reinforcement Learning Algorithms Classification

The previous sections discussed the UAV navigation tasks and frameworks without elaborating on RL algorithms. However, to choose a suitable algorithm for the application environment and the navigation task, the comprehension of RL algorithms and their characteristics is necessary. For example, the DQN algorithm and its variations can be used for UAV navigation tasks that use simple movement actions (UP, DOWN, LEFT, RIGHT, FORWARD) since they are discrete. Therefore, this section examines RL algorithms and their characteristics.

AlMahamid and Grolinger [11] categorized RL algorithms into three main categories according to the number of states and the type of actions: 1) limited number of states and discrete actions, 2) unlimited number of states

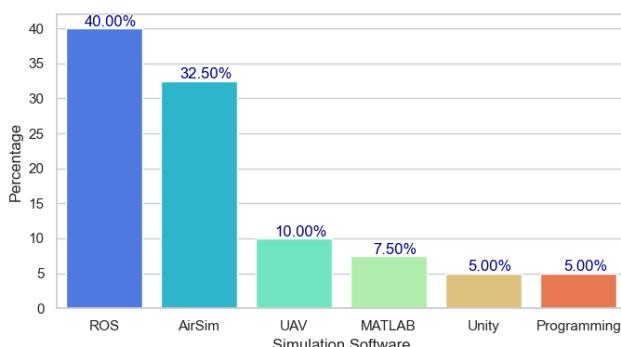


Figure 6: UAV Simulation Software Usage



Figure 7: UAV Simulation Software Components

and discrete actions, and 3) unlimited number of states and continuous actions. We extend this with sub-classes, analyze more than 50 RL algorithms, and examine their use in UAV navigation. Table 3 classifies all RL algorithms found in UAV Navigation studies and includes other prominent RL algorithms to show the intersection between RL and UAV navigation. Furthermore, this section discusses algorithms characteristics and highlights RL applications in different UAV navigation studies. Note that Table 3 includes many algorithms, but only the most prominent ones are discussed in the following subsections.

6.1. Limited States and Discrete Actions

Generally, simple environments have a limited number of states and the agent transitions between states by executing discrete (limited number) actions. For example, in a tic-tac-toe game, the agent has a predefined set of two actions X or O that are used to update the nine boxes constituting the predefined set of known states. Q-Learning [108] and State–Action–Reward–State–Action (SARSA) [110] algorithms can be applied to environments with a limited number of states and discrete actions, where they maintain a Q-Table with all possible states and actions while iteratively updating the Q-values for each state-action pair to find the optimal policy.

SARSA is similar to Q-Learning except to update the current $Q(s, a)$ value it computes the next state-action $Q(s', a')$ by executing the next action a' [126]. In contrast, Q-learning updates the current $Q(s, a)$ value by computing the next state-action $Q(s', a')$ using the Bellman equation since the next action is unknown, and takes a greedy action by selecting the action that maximizes the reward [126].

6.2. Unlimited States and Discrete Actions

An RL agent uses Deep Neural Network (DNN) - usually a CNN, in complex environments such as the pong game, where the states are unlimited and the actions are discrete (UP, DOWN). The deep agent/DNN processes the environment's state as an input and outputs the Q-values of the available actions. The following subsections discuss the different algorithms that can be used in this type of the environment, such as DQN, Deep SARSA, and their variations [11].

6.2.1. Deep Q-Networks Variations

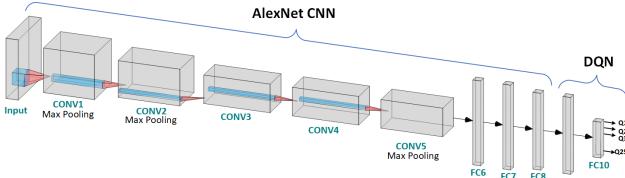
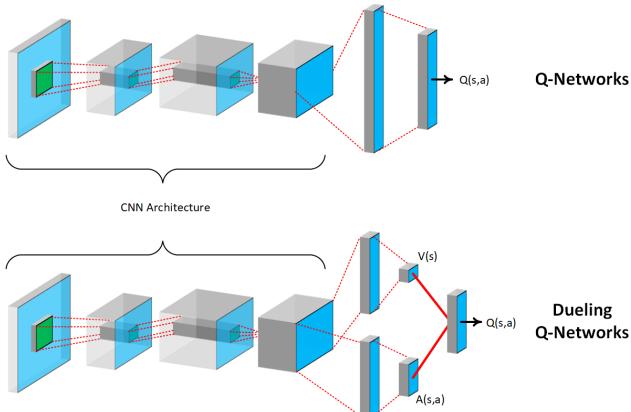
Deep Q-Learning, also known as Deep Q-Network (DQN)[111], is a primary method used in settings with an unlimited number of states and discrete actions, and it serves as an inspiration for other algorithms used for the same goal. As illustrated in Figure 8 [29], DQN architecture frequently employs convolutional and pooling layers, followed by fully connected layers that provide Q-values corresponding to the number of actions. A significant disadvantage of the DQN algorithm is that it overestimates the action-value (Q-value), with the agent selecting the actions with the highest Q-value, which may not be the optimal action [151].

Double DQN solves the overestimation issue in DQN by using two networks. The first network, known as the Policy

Table 3

RL Algorithms usage and classification

State	Action	Class	Algorithm	On/Off Policy	Actor-Critic	Multi-Thread	Distributed	Multi-Agent	Usage
Limited	Discrete	Simple RL	Q-Learning [108]	-	No	No	No	No	[17, 21, 24, 61, 63–65, 67, 68, 74, 75, 109]
			SARSA [110]	-	No	No	No	No	-
	Discrete	DQN Variations	DQN [111]	Off	No	No	No	No	[19, 23, 25, 26, 33, 35, 37, 41, 44, 46, 47, 49, 55, 66, 70, 72, 77, 83, 88, 89, 109, 112]
			Double DQN [113]	Off	No	No	No	No	[16, 26, 29, 31, 37, 40, 45, 52, 78, 79, 109]
			Dueling DQN [114]	Off	No	No	No	No	[26, 37]
			DRQN [115]	Off	No	No	No	No	[43, 58, 73, 76]
			DD-DQN [114]	Off	No	No	No	No	[26, 48]
			DD-DRQN	Off	No	No	No	No	-
Unlimited	Discrete	Distributional DQN	Noisy DQN [116]	Off	No	No	No	No	-
			C51-DQN [117]	Off	No	No	No	No	-
			QR-DQN [118]	Off	No	No	No	No	-
			IQN [119]	Off	No	No	No	No	-
			Rainbow DQN [120]	Off	No	No	No	No	-
			FQF [121]	Off	No	No	No	No	-
			R2D2 [122]	Off	No	No	Yes	No	-
			Ape-X DQN [123]	Off	No	No	Yes	No	-
			NGU [124]	Off	No	No	Yes	No	-
			Agent57 [125]	Off	No	No	Yes	No	-
	Continuous	Deep SARSA Variations	Deep SARSA [126]	On	No	No	No	No	-
			Double SARSA	On	No	No	No	No	-
			Dueling SARSA	On	No	No	No	No	-
			DR-SARSA	On	No	No	No	No	-
			DD-SARSA	On	No	No	No	No	-
			DD-DR-SARSA	On	No	No	No	No	-
Unlimited	Actor-Critic	Policy Based	REINFORCE [127]	On	No	No	No	No	-
			TPRO [128]	On	No	No	No	No	[22]
			PPO [129]	On	No	No	No	No	[18, 22, 38, 39, 51, 53, 56, 62, 69]
			PPG [130]	Off	No	No	No	No	-
			SVPG [131]	Off	No	No	No	No	-
			SLAC [132]	Off	Yes	No	No	No	-
			ACE [133]	Off	Yes	Yes	No	No	-
			DAC [134]	Off	Yes	No	No	No	-
			DPG [15]	Off	Yes	No	No	No	[20]
			RDPG [135]	Off	Yes	No	No	No	-
	Continuous	Actor-Critic	DDPG [136]	Off	Yes	No	No	No	[22, 24, 30, 32, 34, 36, 42, 50, 54, 59, 80, 81, 84, 86]
			TD3 [137]	Off	Yes	No	No	No	[87]
			SAC [138]	Off	Yes	No	No	No	[34]
			Ape-X DPG [123]	Off	Yes	No	Yes	No	-
			D4PG [139]	Off	Yes	No	Yes	Yes	-
			A2C [140]	On	Yes	Yes	Yes	No	[76, 82]
			DPPO [141]	On	No	No	Yes	Yes	-
			A3C [140]	On	Yes	Yes	No	Yes	[36]
Unlimited	Multi-Agent and Distributed Actor-Critic		PAAC [142]	On	Yes	Yes	No	No	-
			ACER [143]	Off	Yes	Yes	No	No	-
			Reactor [144]	Off	Yes	Yes	No	No	-
			ACKTR [145]	On	Yes	Yes	No	No	-
			MADDPG [146]	Off	Yes	No	No	Yes	-
			MATD3 [147]	Off	Yes	No	No	Yes	-
			MAAC [148]	Off	Yes	No	No	Yes	[90]
			IMPALA [149]	Off	Yes	Yes	Yes	Yes	-
			SEED [150]	Off	Yes	Yes	Yes	Yes	-

**Figure 8:** DQN using AlexNet CNN**Figure 9:** DQN vs. Dueling DQN

Network, optimizes the Q-value, while the second network, known as the Target Network, is a clone of the Policy Network and is used to generate the estimated Q-value [113]. After a specified number of time steps, the parameters of the target network network are updated by copying the policy network parameters instead of performing backpropagation.

Dueling DQN, as depicted in Figure 9 [114], is a further enhancement to DQN. To improve Q-value evaluation, Dueling DQN employs the following functions in place the Q-value function:

- The State-Value function $V(s)$ quantifies how desirable it is for an agent to be in a state s .
- The Advantage-Value function $A(s, a)$ assesses the superiority of the selected action in a given state s over other actions.

The two functions depicted in Figure 9 are integrated using a custom aggregation layer to generate an estimate of the state-action value function [114]. The aggregation layer has the same value as the sum of the two values produced by the two functions:

$$Q(s, a) = V(s) + (A(s, a) - \frac{1}{|\mathcal{A}|} \sum_{a'} A(s, a)) \quad (7)$$

The term $\frac{1}{|\mathcal{A}|} \sum_{a'} A(s, a)$ denotes the mean, whereas $|\mathcal{A}|$ denotes vector A length. This assists the *identifiability* problem while having no effect on the relative rank of the A (and thus Q) values. This also improves the optimization because the advantage function only needs to change as fast as the mean [114].

As explained by Wang *et al.* [114], Double Dueling DQN (DD-DQN) extends DQN by combining Dueling DQN and Double DQN to determine the optimal Q-value, with the output of Dueling DQN passed to Double DQN.

The Deep Recurrent Q-Network (DRQN) [115] algorithm is a DQN variation, using a recurrent LSTM layer in place of the first fully connected layer. This changes the input from a single environment state to a group of states as a single input, which aids in the integration of information over time [115]. The techniques of doubling and dueling can be utilized independently or in combination with a recurrent neural network.

6.2.2. Distributional DQN

The goal of distributional Q-learning is to obtain a more accurate representation of the distribution of observed rewards. Fortunato *et al.* [116] introduced NoisyNet, a deep reinforcement learning agent that uses gradient descent to learn parametric noise added to the network weights, and demonstrated how the agent's policy's induced stochasticity can be used to aid efficient exploration [116].

Categorical Deep Q-Networks (C51-DQN) [117] applied a distributional perspective using Wasserstein metric to the random return received by Bellman's equation to approximate value distributions instead of the value function. The algorithm first performs a heuristic projection step and then minimizes the Kullback-Leibler (KL) divergence between the projected Bellman update and the prediction [117].

Quantile Regression Deep Q-Networks (QR-DQN) [118] performs a distributional reinforcement learning over the Wasserstein metric in a stochastic approximation setting. Using Wasserstein distance, the target distribution is minimized by stochastically adjusting the distributions' locations using quantile regression [118]. QR-DQN assigns fixed, uniform probabilities to N adjustable locations and minimizes the quantile Huber loss between the Bellman updated distribution and current return distribution [121], whereas C51-DQN uses N fixed locations ($N = 51$) for distribution approximation and adjusts the locations probabilities [118].

Implicit Quantile Networks (IQN) [119] incorporates QR-DQN [118] to learn full quantile function controlled by the size of the network and the amount of training, in contrast to QR-DQN quantile function that learns a discrete set of quantiles dependent on the number of quantiles output [119]. IQN distribution function assumes the base distribution to be non-uniform and reparameterizes samples from a base distribution to the respective quantile values of a target distribution.

Rainbow DQN [120] combines several improvements of the traditional DQN algorithm into a single algorithm, such as **1)** addressing the overestimation bias, **2)** using Prioritized Experience Replay (PER) [14], **3)** using Dueling DQN [114], **4)** shifting the bias-variance trade-off and propagating newly observed rewards faster to earlier visited states as implemented in A3C [140], **5)** learning a distributional reinforcement learning instead of the expected return similar

to C51-DQN [117], and **6**) implementing stochastic network layers using Noisy DQN [116].

Yang *et al.* [121] proposed **Fully parameterized Quantile Function (FQF)** for distributional RL providing full parameterization for both quantile fractions and corresponding quantile values. In contrast, QR-DQN [118] and IQN [119] only parameterize the corresponding quantile values, while quantile fractions are either fixed or sampled [121].

FQF for distributional RL uses two networks: 1) quantile value network that maps quantile fractions to corresponding quantile values, and 2) fraction proposal network that generates quantile fractions for each state-action pair with the goal of distribution approximation while minimizing the 1-Wasserstein distance between the approximated and actual distribution [121].

6.2.3. Distributed DQN

Distributed DRL architecture used by different RL algorithms as depicted in Figure 10 [125] aims to decouple acting from learning in distributed settings relying on prioritized experience replay to focus on the significant experiences generated by actors. The actors share the same NN and replay experience buffer, where they interact with the environment and store their experiences in the shared replay experience buffer. On the other hand, the learner replays prioritized experiences from the shared experience buffer and updates the learner NN accordingly [123]. In theory, both acting and learning can be distributed across multiple workers or running on the same machine [123].

Ape-X DQN [123], based on the Ape-X framework, was the first algorithm to suggest distributed DRL, which was later extended by Recurrent Replay Distributed DQN (R2D2) [122] with two main differences: 1) R2D2 adds an LSTM layer after the convolutional stack to overcome partial observability, and 2) it trains a recurrent neural network from randomly sampled replay sequences using the “burn-in” strategy, which produces a start state through using a portion

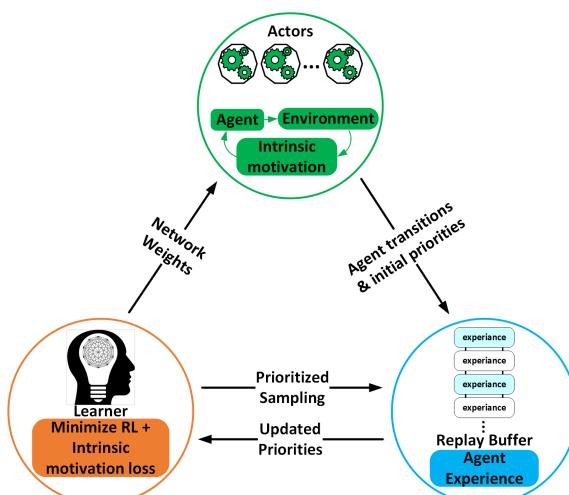


Figure 10: Distributed DRL agent scheme

of the replay sequence and updates the network only on the remaining part of the sequence [122].

Never Give Up (NGA) [124] is another algorithm that combines R2D2 architecture with a novel approach that encourages the agent to learn exploratory strategies throughout the training process using a compound intrinsic reward consisting of two modules:

- Life-long novelty module uses Random Network Distillation (RND) [152], which consists of two networks used to generate an intrinsic reward: 1) target network, and 2) prediction network. This mechanism is known as curiosity because it motivates the agent to explore the environment by going to novel or unfamiliar states.
- Episodic novelty module uses dynamically-sized episodic memory M that stores the controllable states in an online fashion, then turns state-action counts into a bonus reward, where the count is computed using the k -nearest neighbors.

While NGA uses intrinsic reward to promote exploration, it promotes exploitation by generating extrinsic reward using the Universal Value Function Approximator (UVFA). NGA uses conditional architecture with shared weights to learn a family of policies that separate exploration and exploitation [124].

Agent57 [125] is the first RL algorithm that outperforms the human benchmark on all 57 games of Atari 2600. Agent57 implements NGA algorithms with the main difference of applying an adaptive mechanism for exploration-exploitation trade-off and utilizes parameterization of the architecture that allows for more consistent and stable learning [125].

6.2.4. Deep SARSA

SARSA is based on Q-learning and is designed for situations with limited states and discrete actions, as explained in subsection 6.1. Deep SARSA [126] uses a deep neural network similar to DQN and has the same extensions: Double SARSA, Dueling SARSA, Double Dueling SARSA (DD-SARSA), Deep Recurrent SARSA (DR-SARSA), and Double Dueling Deep Recurrent (DD-DR-SARSA). The main difference compared to DQN is that Deep SARSA computes $Q(s', a')$ by taking the next action a' , which is necessary to determine the current state-action $Q(s, a)$ rather than taking a greedy action that maximizes the reward.

6.3. Unlimited States and Continuous Actions

While discrete actions are adequate to drive a car or unmanned aerial vehicle in a simulated environment, they do not enable realistic movements in real-world scenarios. Continuous actions specify the quantity of movement in various directions, and the agent does not select from a predetermined set of actions. For instance, a realistic UAV movement defines the amount of roll, pitch, yaw, and throttle changes necessary to navigate the environment while avoiding obstacles, as opposed to flying the UAV in preset directions: forward, left, right, up, and down [11].

Continuous action space demands learning a parameterized policy π_θ that maximizes the expected summation of

discounted rewards since it is not feasible to determine the action-value $Q(s, a)$ for all continuous actions in all distinct states. Learning a parameterized policy π_θ is considered a maximization problem, which can be handled using gradient descent methods to get the optimal θ in the following manner [11]:

$$\theta_{t+1} = \theta_t + \alpha \nabla J(\theta_t) \quad (8)$$

Here, ∇ is the gradient and α is the learning rate.

The goal of the reward function J is to maximize the expected reward applying the following parameterized policy π_θ [12]:

$$\begin{aligned} J(\pi_\theta) &= \sum_{s \in S} \rho_{\pi_\theta}(s) V^{\pi_\theta}(s) \\ &= \sum_{s \in S} \rho_{\pi_\theta}(s) \sum_{a \in A} Q^{\pi_\theta}(s, a) \pi_\theta(a|s) \end{aligned} \quad (9)$$

where $\rho_{\pi_\theta}(s)$ denotes the stationary probability π_θ starting from state s_0 and transitioning to future states according to the policy π_θ . To determine the best θ that maximizes the function $J(\pi_\theta)$, the gradient $\nabla_\theta J(\theta)$ is calculated as follows:

$$\begin{aligned} \nabla_\theta J(\theta) &= \nabla_\theta \left(\sum_{s \in S} \rho_{\pi_\theta}(s) \sum_{a \in A} Q^{\pi_\theta}(s, a) \pi_\theta(a|s) \right) \\ &\propto \sum_{s \in S} \mu(s) \sum_{a \in A} Q^{\pi_\theta}(s, a) \nabla_\theta \ln \pi_\theta(a|s) \end{aligned} \quad (10)$$

Due to the fact that $\sum_{s \in S} \eta(s) = 1$ and the action space is continuous, Equation 10 can be rewritten as:

$$\nabla_\theta J(\theta) = \mathbb{E}_{s \sim \rho^{\pi_\theta}, a \sim \pi_\theta} \left[Q^{\pi_\theta}(s, a) \nabla_\theta \ln \pi_\theta(a|s) \right] \quad (11)$$

Equation 12 [15], referred to as the off-policy gradient theorem, defines the policy change in relation to the ratio of target policy $\pi_\theta(a|s)$ to behavior policy $\beta(a|s)$. Take note that the training sample is selected according to the target policy $s \sim \rho^{\pi_\theta}$, and the expected return is calculated for the same policy π_θ , where the training sample adheres to the behavior policy $\beta(a|s)$.

$$\nabla_\theta J(\theta) = \mathbb{E}_{s \sim \rho^\beta, a \sim \beta} \left[\frac{\pi_\theta(a|s)}{\beta_\theta(a|s)} Q^{\pi_\theta}(s, a) \nabla_\theta \ln \pi_\theta(a|s) \right] \quad (12)$$

The policy gradient theorem depicted in Equation 9 [153] served as the foundation for a variety of other Policy Gradients (PG) algorithms, including REINFORCE, Actor-Critic algorithms, and various multi-agent and distributed actor-critic algorithms.

6.3.1. Policy-Based Algorithms

Policy-based algorithms are devoted to improving the gradient descent performance by means of applying different methods such as REINFORCE [127], Trust Region Policy Optimization (TRPO) [128], Proximal Policy Optimization (PPO) [129], Phasic Policy Gradient (PPG) [130], and Stein Variational Policy Gradient (SVPG) [131].

REINFORCE

REINFORCE is a Monte-Carlo policy gradient approach that creates a sample by selecting from an entire episode proportionally to the gradient and updates the policy parameter θ with the step size α . Given that $\mathbb{E}_\pi[G_t | S_t, A_t] = Q^\pi(s, a)$, REINFORCE may be defined as follows [12]:

$$\nabla_\theta J(\theta) = \mathbb{E}_\pi \left[G_t \nabla_\theta \ln \pi_\theta(A_t | S_t) \right] \quad (13)$$

The Monte Carlo method has a high variance and, hence, a slow pace of learning. By subtracting the baseline value from the expected return G_t , REINFORCE decreases variance and accelerates learning while maintaining the bias [12].

Trust Region Policy Optimization (TRPO)

Trust Region Policy Optimization (TRPO) [128] belongs to a category of PG methods: it enhances gradient descent by performing protracted steps inside trust zones specified by a KL-Divergence constraint and updates the policy after each trajectory instead of after each state [11]. Proximal Policy Optimization (PPO) [129] may be thought of as an extension of TRPO, where the KL-Divergence constraint is applied as a penalty and the objective is clipped to guarantee that the optimization occurs within a predetermined range [154].

Phasic Policy Gradient (PPG) [130] is an extension of PPO [129]: it incorporates a recurring auxiliary phase that distills information from the value function into the policy network to enhance the training while maintaining decoupling.

Stein Variational Policy Gradient (SVPG) Stein Variational Policy Gradient (SVPG) [131] algorithm updates the policy π_θ using Stein variational gradient descent (SVGD) [155], therefore reducing variance and improving convergence. When used in conjunction with REINFORCE and the advantage actor-critic algorithms, SVPG enhances average return and data efficiency [155].

6.3.2. Actor-Critic

The term "Actor-Critic algorithms" refers to a collection of algorithms based on the policy gradients theorem. They are composed of two components:

1. The Actor who is liable of finding the optimal policy π_θ .
2. The Critic who estimates the value function $Q^w(s_t, a_t) \approx Q^\pi(s_t, a_t)$ utilizing a parameterized vector w and a policy assessment technique such as temporal-difference learning [15].

The actor can be thought of as a network that is attempting to discover the probability of all possible actions and perform the one with the largest probability, whereas the critic can be thought of as a network that is evaluating the chosen action by assessing the quality of the new state created by the performed action. Numerous algorithms can be classified under the actor-critic category including Deterministic policy gradients (DPG) [15], Deep Deterministic Policy Gradient (DDPG) [136], Twin Delayed Deep Deterministic (TD3) [137], and many others.

Deterministic Policy Gradients (DPG)

Deterministic policy gradients (DPG) algorithms implement a deterministic policy $\mu(s)$ instead of a stochastic policy $\pi(s, a)$. The deterministic policy is a subset of a stochastic policy in which the target policy objective function is averaged over the state distribution of the behavior policy, as depicted in 14 [15].

$$\begin{aligned} J_\beta(\mu_\theta) &= \int_S \rho^\beta(s) V^\mu(s) ds \\ &= \int_S \rho^\beta(s) Q^\mu(s, \mu_\theta(s)) ds \end{aligned} \quad (14)$$

Importance sampling is frequently used in off-policy techniques with a stochastic policy to account for mismatches between behavior and target policies. The deterministic policy gradient eliminates the integral over actions; therefore, the importance sampling can be skipped, resulting in the following gradient [11]:

$$\begin{aligned} \nabla_\theta J_\beta(\mu_\theta) &\approx \int_S \rho^\beta(s) \nabla_\theta \mu_\theta(a|s) Q^\mu(s, \mu_\theta(s)) ds \\ &= \mathbb{E}_{s \sim \rho^\beta} \left[\nabla_\theta \mu_\theta(s) \nabla_a Q^\mu(s, a) \Big|_{a=\mu_\theta(s)} \right] \end{aligned} \quad (15)$$

Numerous strategies are employed to enhance DPG; for example, Experience Replay (ER) can be used in conjunction with DPG to increase the stability and efficiency of data [135]. Deep Deterministic Policy Gradient (DDPG) [136], on the other hand, expands DPG by leveraging DQN to operate in continuous action space whereas Twin Delayed Deep Deterministic (TD3) [137] expands on DDPG by utilizing Double DQN to prevent the overestimation of the value function by taking the minimum value between the two critics [137].

Recurrent Deterministic Policy Gradients (RDPG)

Wierstra *et al.* [156] applied RNN to Policy Gradient (PG) to build a model-free RL - namely Recurrent Policy Gradient (RPG), for Partially Observable Markov Decision Problem (POMDP), which does not require the agent to have a complete assumption about the environment [156]. RPG applies a method for backpropagating return-weighted characteristic eligibilities through time to approximate a policy gradient for a recurrent neural network [156].

Recurrent Deterministic Policy Gradient (RDPG) [135] implements DPG using RNN and extends the work of RGP

[156] to partially observed domains. The RNN with LSTM cells preserves information about past observations over many time steps.

Soft Actor-Critic (SAC)

The objective of Soft Actor-Critic (SAC) is to maximize anticipated reward and the entropy [138]. By adding the anticipated entropy of the policy across $\rho_\pi(s_t)$, SAC improves the maximum sum of rewards established by adding the rewards over states transitions $J(\pi) = \sum_{t=1}^T \mathbb{E}_{s \sim \rho^\pi, a \sim \pi} [r(s_t, a_t)]$ [138]. Equation 16 illustrates an extended entropy goal, in which the temperature parameter α influences the stochasticity of the optimum policy by specifying the importance of the entropy $\mathcal{H}(\pi(.|s_t))$ term to the reward [138].

$$J(\pi) = \sum_{t=1}^T \mathbb{E}_{s \sim \rho^\pi, a \sim \pi} [r(s_t, a_t) + \alpha \mathcal{H}(\pi(.|s_t))] \quad (16)$$

By using function approximators and two independent NNs for the actor and critic, SAC estimates a soft Q-function $Q_\theta(s_t, a_t)$ parameterized by θ , a state value function $V_\psi(s_t)$ parameterized by ψ , and an adjustable policy $\pi_\phi(a_t|s_t)$ parameterized by ϕ [11].

6.3.3. Multi-Agent and Distributed Actor-Critic

This group of algorithms includes multi-agent and distributed actor-critic algorithms. They are grouped together as multi-agents can be deployed across several nodes making it a distributed system.

Advantage Actor-Critic

Asynchronous Advantage Actor-Critic (A3C) [140] is a policy gradient algorithm that parallelizes training by using multi-threads, commonly known as workers or agents. Each agent has a local policy $\pi_\theta(a_t|s_t)$ and a value function estimate $V_\theta(s_t)$. The agent and the same-structured global network asynchronously exchange the parameters in both directions, from agent to the global network and vice-versa. After t_{max} actions or when a final state is reached, the policy and the value function are modified [140].

Advantage Actor-Critic (A2C) [140] is a policy gradient method identical to A3C, except that it includes a coordinator for synchronizing all agents. After all agents complete their work, either by arriving at a final state or by completing t_{max} actions, the coordinator updates the policy and value function in both directions between the agents and the global network and vice versa.

Another variant of A3C is **Actor-Critic with Kronecker-Factored Trust Region (ACKTR)** [145] which uses Kronecker-factored approximation curvature (K-FAC) [157] to optimize the actor and critic. It improves the computation of natural gradients by efficiently inverting the gradient covariance matrix.

Actor-Critic with Experience Replay (ACER)

Actor-Critic with Experience Replay (ACER) [143] is an off-policy actor-critic algorithm with experience replay that

estimates the policy $\pi_\theta(a_t|s_t)$ and the value function $V_{\theta_v}^\pi(s_t)$ using a single deep neural network [143]. In comparison to A3C, ACER employs a stochastic dueling network and a novel *trust region policy optimization* [143], while improving importance sampling with a bias correction [140].

ACER applies an improved Retrace algorithm [158] by using a truncated importance sampling with bias correction and the value Q^{ret} as the target value to train the critic [143]. The gradient \hat{g}_t^{acer} is determined by truncating the importance weights by a constant c , and subtracting $V_{\theta_v}(s_t)$, which reduces variance.

Retrace-Actor (Reactor) [144] increases sampling and time efficiency by combining contributions from different techniques. It employs Distributional Retrace [158] to provide multi-step off-policy distributional RL updates while prioritizing replay on transitions [144]. Additionally, by taking advantage of action values as a baseline, Reactor improves the trade-off between variance and bias via β -leave-one-out (β -LOO) resulting in an improvement of the policy gradient [144].

Multi-Agent Reinforcement Learning (MARL)

Distributed Distributional DDPG (D4PG) [139] adds features such as N-step returns and prioritized experience replay to the distributed settings of DDPG [139]. On the other hand, Multi-Agent DDPG (MADDPG) [146] expands DDPG to coordinate between multiple agents and learn policies while considering each agent's policy [146]. In comparison, Multi-Agent TD3 (MATD3) expands TD3 [147] to work with multi-agents using centralized training and decentralized execution while, similarly to TD3, controlling the overestimation bias by employing two centralized critics for each agent.

Importance Weighted Actor-Learner Architecture (IMPALA) [149] is an off-policy algorithm that separates action execution and policy learning. It can be applied using two distinct configurations: 1) a single learner and multiple actors, or 2) multiple synchronous learners and multiple actors.

Using a single learner and several actors, the trajectories generated by the actors are transferred to the learner. Before initiating a new trajectory, the actors are waiting for the learner to update the policy, while the learner simultaneously queues the received trajectories from the actors and constructs the updated policy. Nonetheless, actors may acquire an older version due to their lack of awareness of one another and the lag between the actors and the learner. To address this challenge, IMPALA employs a unique v-trace correction approach that takes into account a truncated importance sampling (IS), defined as the ratio of the learner's policy π to the actor's present policy μ [11]. Likewise, with multiple synchronous learners, policy parameters are spread across numerous learners who communicate synchronously via a master learner [149].

Scalable, Efficient Deep-RL (SEED RL) [150] provides a scalable architecture that combines IMPALA with R2D2 and can train on millions of frames per second with a lower cost of experiments compared to IMPALA [150]. SEED

moves the inference to the learner while the environments run remotely, introducing a latency issue due to the increased number of remote calls, which is mitigated using a fast communication layer using gRPC.

7. Problem Formulation and Algorithm Selection

The previous section categorized RL algorithms based on state and action types and reviewed the most prominent algorithms. With such a large number of algorithms, it is challenging to select the RL algorithms suitable to tackle the task at hand. Consequently, Figure 11 depicts the process of selecting a suitable RL algorithm or a group of RL algorithms through six steps/questions that are answered to guide an informed selection.

The selection process places a greater emphasis on how the environment and RL objective are formulated than on the RL problem type because the algorithm selection is dependent on the environment and objective formulation. For instance, UAV navigation tasks can employ several sets of algorithms dependent on the desired action type. The six steps, indicated in Figure 11, guide the selection of algorithms: the selected option at each step limits the choices available in the next step based on the available algorithms' characteristics. The steps are as follows:

- **Step 1 - Define State Type:** When assessing an RL task, it is essential to comprehend the state that can be obtained from the surrounding environment. For instance, some navigation tasks simplify the environment's states using grid-cell representations [68, 73, 93], where the agent has a limited and predetermined set of states, whereas in other tasks, the environment can have unlimited states [34, 38, 40]. Therefore, this step involves a decision between limited vs. unlimited states.

- **Step 2 - Define Action Type:** Choosing between discrete and continuous action types limits the number of applicable algorithms. For instance, discrete actions can be used to move the UAV in pre-specified directions (UP, DOWN, RIGHT, LEFT, etc.), whereas continuous actions, such as the change in pitch, roll, and yaw angles, specify the quantity of the movement using a real number $r \in \mathbb{R}$.

- **Step 3 - Define Policy Type:** As addressed and explained in Subsection 3.4, RL algorithms can be either off-policy or on-policy algorithms. The policy type selected restricts the alternatives accessible in the subsequent stage. On-policy algorithms converge faster than off-policy algorithms and find a sub-optimal policy, making them a good fit for environments requiring much exploration. Moreover, on-policy algorithms provide stable training since one policy uses learning and data sampling. On the other hand, off-policy algorithms provide an optimal policy and require a good exploration strategy.

The off-policy algorithms' convergence can be improved using techniques such as *prioritized experience replay* and *importance sampling*, making them a good fit for navigation tasks that require finding the optimal path.

- **Step 4 - Define Processing Type:** While some RL algorithms run in a single thread, others support multi-threading and distributed processing. This step selects the processing type that suits the application needs and the available computational power.
- **Step 5 - Define Number of Agents:** This step specifies the number of agents the application should have. This is needed as some RL algorithms enable MARL, which accelerates learning but requires more computational resources, while other techniques only employ a single agent.
- **Step 6 - Select the Algorithms:** The last phase of the process results in a collection of algorithms that may be applied to the RL problem at hand. However, the performance of the algorithms is affected by a number of factors and may vary depending on variables such as hyper-parameter settings, reward engineering, and the agent's NN architecture. Consequently, the procedure seeks to reduce the algorithm selection to a group of algorithms rather than a single algorithm.

While this section presented the process of narrowing down the algorithm for use case, Section 6 provided a description and references to many algorithms to assist in comprehending the distinctions between the algorithms and making an informed selection.

8. Challenges and Opportunities

Previous sections demonstrated the diversity of UAV navigation tasks besides the diversity of RL algorithms. Due to such a high number of algorithms, selecting the appropriate algorithms for the task at hand is challenging. Table 3 and discussion in Section 6 provide an overview of the RL algorithms and assist in selecting the RL algorithm for navigation task. Nevertheless, there are still numerous challenges and opportunities in RL for UAV navigation, including:

Evaluation and benchmarking: Atari 2600 is a home video game console with 57 built-in games that laid the foundation to establish a benchmarking for RL algorithms. The benchmark was established using 57 different games to compare various RL algorithms and set a benchmark baseline against human performance playing the same games. The agent's performance is evaluated and compared to other algorithms using the same benchmark (Atari 2600), or evaluated using various non-navigation environments other than Atari 2600 [159]. The performance of the algorithms using the benchmark might differ when applied to the UAV navigation simulated on a 3D environment or the real world because the games in Atari 2600 can provide a full state of the environment, which means the agent does not need to make assumptions about the state, and MDP can be applied to these problems. Whereas in UAV navigation simulation, the agent knows partial states of the environment (observation), these observations are used to train the agent using POMDP, which results in changing behavior for some of the algorithms. Furthermore, images processed from the games are 2D images, where the agent in most algorithms tries to learn

an optimal policy based on the pattern of the pixels in the image. The same cannot be inferred for images received from the 3D simulators or the real-world images because objects' depth plays a vital role in learning the optimal policy avoiding nearby objects. Therefore, there is a need for new evaluation and benchmarking techniques for RL driven navigation.

Environment complexity: The tendency to oversimplify the environment and the absence of a standardized benchmarking tools makes it impossible to compare and conclude performances obtained using different algorithms and simulated using various tools and environments. Nevertheless, the UAV needs to perform tasks in different environments and is subject to various conditions, for example:

- Navigating in various environment types such as indoor vs. outdoor.
- Considering the changing environment conditions such as wind speed, lighting conditions, and moving objects.

Some of the simulation tools discussed in Section 5, such as AirSim combined with Unreal Engine, provide different environment types out-of-the-box and are capable of simulating several environmental effects such as changing wind speed and lighting conditions. Still, these complex environments remain to be combined with new benchmarking techniques for improved comparison of RL algorithms for UAV navigation.

Knowledge transfer: Knowledge transfer imposes another challenge, where the RL agent training in a selected environment does not guarantee similar performance in another environment due to the difference in environments' nature such as different object/obstacles types, background texture, lighting density, and added noise. Most of the existing research focused on applying transfer learning to reduce the training time for the agent in the new environment [40]. However, generalized training methods or other techniques are needed to guarantee a similar performance of the agent in different environments and under various conditions.

UAVs complexity: Training UAVs is often accomplished in a 3D virtual environment since UAVs have limited computational resources and power supply, with a typical flight time of 10 to 30 minutes. Reducing the computation time will create possibilities for more complex navigation tasks and increase the flight time since it will reduce energy consumption. Figure 6 shows that only 10% of the investigated research used real drones for navigation training. Therefore, more research is required to focus on energy-aware navigation utilizing low-complexity and efficient RL algorithms while simulating using real drones.

Algorithm diversity: As seen from Table 3, many recent and very successful algorithms have not been applied in UAV navigation. As these algorithms have shown great success in other domains outperforming the human benchmark, there is a prodigious potential in their application in UAV navigation. The algorithms are expected to gain better generalization on different environments, speed up the

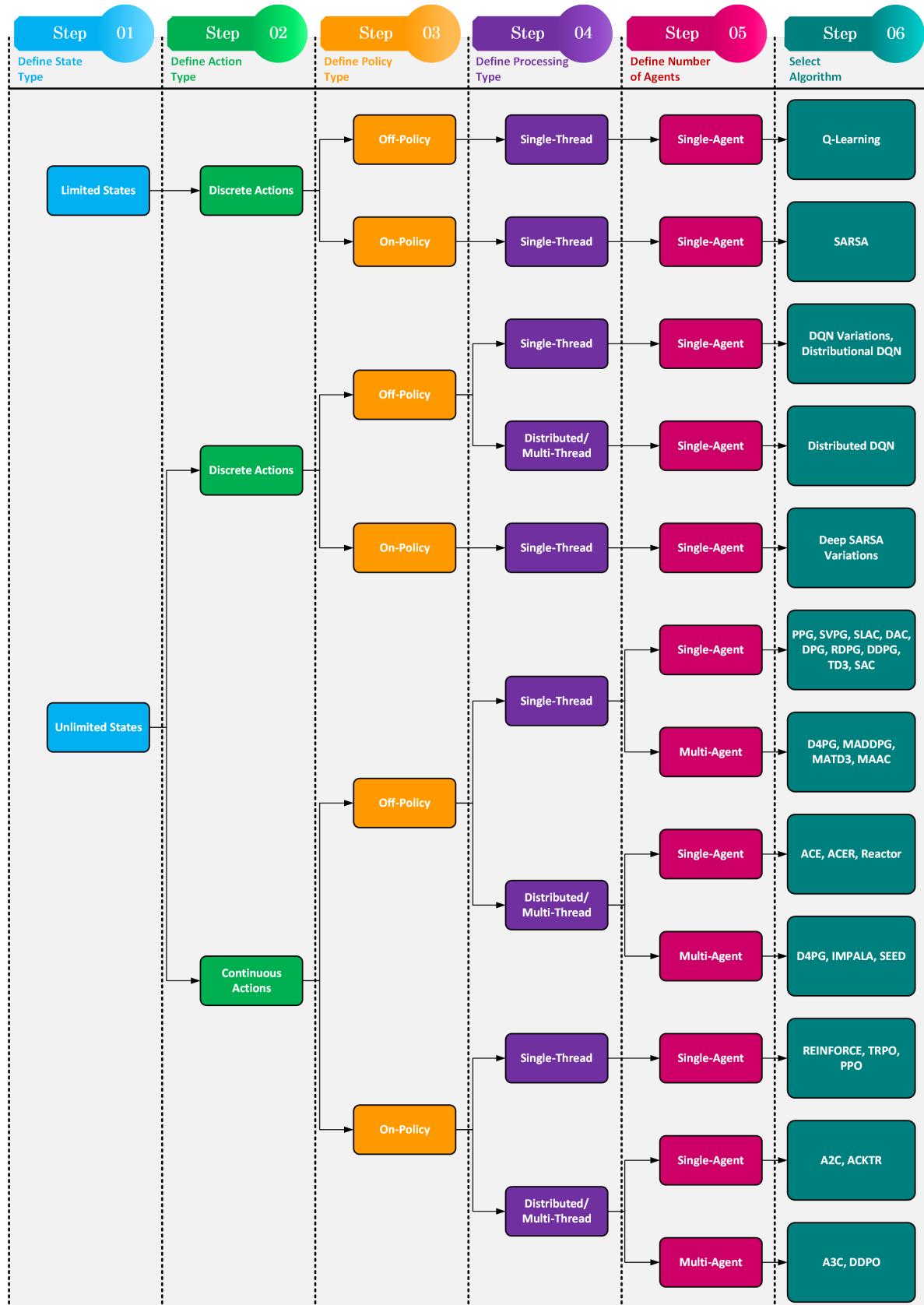


Figure 11: Algorithm Selection Process

training process, and even solve efficiently more complex tasks such as UAVs flocking.

9. Conclusion

This review deliberates on the application of RL for autonomous UAV navigation. RL uses an intelligent agent to control the UAV movement by processing the states from the environment and moving the UAV in desired directions. The data received from the UAV camera or other sensors such as LiDAR are used to estimate the distance from various objects in the environment and avoid colliding with these objects.

RL algorithms and techniques were used to solve navigation problems such as controlling the UAV while avoiding obstacles, path planning, and flocking. For example, RL is used in single UAV path planning and multi-UAVs flocking to plan path waypoints of the UAV(s) while avoiding obstacles or maintaining flight formation (flocking). Furthermore, this study recognizes various navigation frameworks simulation software used to conduct the experiments along with identifying their use within the reviewed papers.

The review discusses over fifty RL algorithms, explains their contributions and relations, and classifies them according to the application environment and their use in UAV navigation. Furthermore, the study highlights other algorithmic traits such as multi-threading, distributed processing, and multi-agents, followed by a systematic process that aims to assist in finding the set of applicable algorithms.

The study observes that the research community tends to experiment with a specific set of algorithms: Q-learning, DQN, Double DQN, DDPG, PPO, although some recent algorithms show more promising results than the mentioned algorithms such as agent57. To the best of the authors' knowledge, this study is the first systematic review identifying a large number of RL algorithms while focusing on their application in autonomous UAV navigation.

Analysis of the current RL algorithms and their use in UAV navigation identified the following challenges and opportunities: the need for navigation-focused evaluation and benchmarking techniques, the necessity to work with more complex environments, the need to examine knowledge transfer, the complexity of UAVs, and the necessity to evaluate state-of-the-art RL algorithms on navigation tasks.

A. Acronyms

- **A2C** : Advantage Actor-Critic
- **A3C** : Asynchronous Advantage Actor-Critic
- **AC** : Actor-Critic
- **ACE** : Actor Ensemble
- **ACER** : Actor-Critic with Experience Replay
- **ACKTR** : Actor-Critic using Kronecker-Factored Trust Region
- **Agent57** : Agent57
- **Ape-X DPG** : Ape-X Deterministic Policy Gradients
- **Ape-X DQN** : Ape-X Deep Q-Networks
- **AS** : Autonomous Systems
- **C51-DQN** : Categorical Deep Q-Networks
- **CNN** : Recurrent Neural Network
- **D4PG** : Distributed Distributional DDPG
- **DAC** : Double Actor-Critic
- **DD-DQN** : Double Dueling Deep Q-Networks
- **DD-DRQN** : Double Dueling Deep Recurrent Q-Networks
- **DDPG** : Deep Deterministic Policy Gradient
- **Double DQN** : Double Deep Q-Networks
- **DPG** : Deterministic Policy Gradients
- **DPPO** : Distributed Proximal Policy Optimization
- **DQN** : Deep Q-Networks
- **DRL** : Deep Reinforcement Learning
- **DRQN** : Deep Recurrent Q-Networks
- **Dueling DQN** : Dueling Deep Q-Networks
- **DVS** : Dynamic Vision Sensor
- **eNVM** : embedded Non-Volatile Memory
- **FOV** : Field Of View
- **FQF** : Fully parameterized Quantile Function
- **GPS** : Global Positioning System
- **IMPALA** : Importance Weighted Actor-Learner Architecture
- **IMU** : Inertial Measurement Unit
- **IQN** : Implicit Quantile Networks
- **K-FAC** : Kronecker-factored approximation curvature
- **KL** : Kullback-Leibler
- **LSTM** : Long-Short Term Memory
- **MAAC** : Multi-Actor-Attention-Critic
- **MADDPG** : Multi-Agent DDPG
- **MARL** : Multi-Agent Reinforcement Learning
- **MATD3** : Multi-Agent Twin Delayed Deep Deterministic
- **MATD3** : Multi-Agent TD3
- **MDP** : Markov Decision Problem
- **NGA** : Never Give Up
- **Noisy DQN** : Noisy Deep Q-Networks
- **PAAC** : Parallel Advantage Actor-Critic
- **PER** : Prioritized Experience Replay
- **PG** : Policy Gradients
- **POMDP** : Partially Observable Markov Decision Problem
- **PPG** : Phasic Policy Gradient
- **PPO** : Proximal Policy Optimization
- **QR-DQN** : Quantile Regression Deep Q-Networks
- **R2D2** : Recurrent Replay Distributed Deep Q-Networks
- **Rainbow DQN** : Rainbow Deep Q-Networks
- **RDPG** : Recurrent Deterministic Policy Gradients
- **Reactor** : Retrace-Actor
- **REINFORCE** : REward Increment = Nonnegative Factor \times Offset Reinforcement \times Characteristic Eligibility
- **RL** : Reinforcement Learning
- **RND** : Random Network Distillation
- **RNN** : Recurrent Neural Network
- **ROS** : Robot Operating System
- **SAC** : Soft Actor-Critic
- **SARSA** : State-Action-Reward-State-Action
- **SEED RL** : Scalable, Efficient Deep-RL

- **SLAC** : Stochastic Latent Actor-Critic
- **SRAM** : Static Random Access Memory
- **SVPG** : Stein Variational Policy Gradient
- **TD** : Temporal Difference
- **TD3** : Twin Delayed Deep Deterministic
- **TRPO** : Trust Region Policy Optimization
- **UAV** : Unmanned Aerial Vehicle
- **UBC** : Upper Confidence Bound
- **UGV** : Unmanned Ground Vehicle
- **UVFA** : Universal Value Function Approximator

CRediT authorship contribution statement

Fadi AlMahamid: Conceptualization, Methodology, Investigation, Formal Analysis, Validation, Writing - Original Draft, Writing - Review & Editing. **Katarina Grolinger:** Supervision, Writing - Review & Editing, Funding Acquisition.

Acknowledgments

This research has been supported by NSERC under grant RGPIN-2018-06222

References

- [1] Y. Lu, Z. Xue, G.-S. Xia, L. Zhang, A survey on vision-based UAV navigation, *Taylor & Francis Geo-spatial information science* 21 (1) (2018) 21–32. doi:[10.1080/10095020.2017.1420509](https://doi.org/10.1080/10095020.2017.1420509).
- [2] F. Zeng, C. Wang, S. S. Ge, A survey on visual navigation for artificial agents with deep reinforcement learning, *IEEE Access* 8 (2020) 135426–135442.
- [3] R. Azoulay, Y. Haddad, S. Reches, Machine Learning Methods for UAV Flocks Management-A Survey, *IEEE Access* 9 (2021) 139146–139175.
- [4] S. Aggarwal, N. Kumar, Path planning techniques for unmanned aerial vehicles: A review, solutions, and challenges, *Computer Communications* 149 (2020) 270–299.
- [5] W. Lou, X. Guo, Adaptive trajectory tracking control using reinforcement learning for quadrotor, *SAGE International Journal of Advanced Robotic Systems* 13 (1) (2016). doi:[10.5772/62128](https://doi.org/10.5772/62128).
- [6] A. Guerra, F. Guidi, D. Dardari, P. M. Djurić, Networks of UAVs of low-complexity for time-critical localization, *arXiv preprint arXiv:2108.13181* (2021).
- [7] A. Guerra, F. Guidi, D. Dardari, P. M. Djurić, Real-Time Learning for THZ Radar Mapping and UAV Control, in: *IEEE International Conference on Autonomous Systems*, 2021, pp. 1–5. doi:[10.1109/ICAS49788.2021.9551141](https://doi.org/10.1109/ICAS49788.2021.9551141).
- [8] A. Guerra, D. Dardari, P. M. Djurić, Dynamic radar network of UAVs: A joint navigation and tracking approach, *IEEE Access* 8 (2020) 116454–116469.
- [9] Y. Liu, Y. Wang, J. Wang, Y. Shen, Distributed 3D relative localization of UAVs, *IEEE Transactions on Vehicular Technology* 69 (2020) 11756–11770.
- [10] S. Zhang, R. Pöhlmann, T. Wiedemann, A. Dammann, H. Wymeersch, P. A. Hoher, Self-aware swarm navigation in autonomous exploration missions, *Proceedings of the IEEE* 108 (7) (2020) 1168–1195.
- [11] F. AlMahamid, K. Grolinger, Reinforcement Learning Algorithms: An Overview and Classification, in: *IEEE Canadian Conference on Electrical and Computer Engineering*, 2021, pp. 1–7.
- [12] R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, 2018.
- [13] L.-J. Lin, Self-improving reactive agents based on reinforcement learning, planning and teaching, *Springer Machine learning* 8 (3–4) (1992) 293–321.
- [14] T. Schaul, J. Quan, I. Antonoglou, D. Silver, Prioritized experience replay, *arXiv:1511.05952* (2015).
- [15] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, M. Riedmiller, Deterministic Policy Gradient Algorithms, in: *PMLR International conference on machine learning*, 2014, pp. 387–395.
- [16] S. Zhou, B. Li, C. Ding, L. Lu, C. Ding, An Efficient Deep Reinforcement Learning Framework for UAVs, in: *IEEE International Symposium on Quality Electronic Design*, 2020, pp. 323–328.
- [17] P. Karthik, K. Kumar, V. Fernandes, K. Arya, Reinforcement Learning for Altitude Hold and Path Planning in a Quadcopter, in: *IEEE International Conference on Control, Automation and Robotics*, 2020, pp. 463–467.
- [18] A. M. Deshpande, R. Kumar, A. A. Minai, M. Kumar, Developmental reinforcement learning of control policy of a quadcopter UAV with thrust vectoring rotors, in: *ASME Dynamic Systems and Control Conference*, Vol. 84287, 2020, p. V002T36A011. doi:[10.1115/DSCC2020-3319](https://doi.org/10.1115/DSCC2020-3319).
- [19] E. Camci, E. Kayacan, Learning motion primitives for planning swift maneuvers of quadrotor, *Springer Autonomous Robots* 43 (7) (2019) 1733–1745.
- [20] S. Li, P. Durdevic, Z. Yang, Optimal Tracking Control Based on Integral Reinforcement Learning for An Underactuated Drone, *Elsevier IFAC-PapersOnLine* 52 (8) (2019) 194–199.
- [21] C. Greatwood, A. G. Richards, Reinforcement learning and model predictive control for robust embedded quadrotor guidance and control, *Springer Autonomous Robots* 43 (7) (2019) 1681–1693.
- [22] W. Koch, R. Mancuso, R. West, A. Bestavros, Reinforcement learning for UAV attitude control, *ACM Transactions on Cyber-Physical Systems* 3 (2) (2019) 1–21.
- [23] N. Salvatore, S. Mian, C. Abidi, A. D. George, A Neuro-Inspired Approach to Intelligent Collision Avoidance and Navigation, in: *IEEE Digital Avionics Systems Conference*, 2020, pp. 1–9.
- [24] O. Bouhamed, H. Ghazzai, H. Besbes, Y. Massoud, A UAV-Assisted Data Collection for Wireless Sensor Networks: Autonomous Navigation and Scheduling, *IEEE Access* 8 (2020) 110446–110460.
- [25] H. Huang, J. Gu, Q. Wang, Y. Zhuang, An Autonomous UAV Navigation System for Unknown Flight Environment, in: *IEEE International Conference on Mobile Ad-Hoc and Sensor Networks*, 2019, pp. 63–68.
- [26] S.-Y. Shin, Y.-W. Kang, Y.-G. Kim, Automatic Drone Navigation in Realistic 3D Landscapes using Deep Reinforcement Learning, in: *IEEE International Conference on Control, Decision and Information Technologies*, 2019, pp. 1072–1077.
- [27] C. Wang, J. Wang, X. Zhang, X. Zhang, Autonomous Navigation of UAV in Large-Scale Unknown Complex Environment with Deep Reinforcement Learning, in: *IEEE Global Conference on Signal and Information Processing*, 2017, pp. 858–862.
- [28] C. Wang, J. Wang, Y. Shen, X. Zhang, Autonomous Navigation of UAVs in Large-Scale Complex Environments: A Deep Reinforcement Learning Approach, *IEEE Transactions on Vehicular Technology* 68 (3) (2019) 2124–2136.
- [29] A. Anwar, A. Raychowdhury, Autonomous Navigation via Deep Reinforcement Learning for Resource Constraint Edge Nodes using Transfer Learning, *IEEE Access* 8 (2020) 26549–26560.
- [30] O. Bouhamed, H. Ghazzai, H. Besbes, Y. Massoud, Autonomous UAV Navigation: A DDPG-Based Deep Reinforcement Learning Approach, in: *IEEE International Symposium on Circuits and Systems*, 2020, pp. 1–5.
- [31] Y. Yang, K. Zhang, D. Liu, H. Song, Autonomous UAV Navigation in Dynamic Environments with Double Deep Q-Networks, in: *IEEE Digital Avionics Systems Conference*, 2020, pp. 1–7.
- [32] Y. Li, M. Li, A. Sanyal, Y. Wang, Q. Qiu, Autonomous UAV with Learned Trajectory Generation and Control, in: *IEEE International Workshop on Signal Processing Systems*, 2019, pp. 115–120.

- [33] Y. Chen, N. González-Prelcic, R. W. Heath, Collision-free UAV navigation with a monocular camera using deep reinforcement learning, in: IEEE International Workshop on Machine Learning for Signal Processing, 2020, pp. 1–6.
- [34] R. B. Grando, J. C. de Jesus, P. L. Drews-Jr, Deep Reinforcement Learning for Mapless Navigation of Unmanned Aerial Vehicles, in: IEEE Latin American Robotics Symposium, Brazilian Symposium on Robotics and Workshop on Robotics in Education, 2020, pp. 1–6.
- [35] E. Camci, D. Campolo, E. Kayacan, Deep Reinforcement Learning for Motion Planning of Quadrotors Using Raw Depth Images, Learning (RL) 10 (2020).
- [36] C. Wang, J. Wang, J. Wang, X. Zhang, Deep-Reinforcement-Learning-Based Autonomous UAV Navigation With Sparse Rewards, IEEE Internet of Things Journal 7 (7) (2020) 6180–6190.
- [37] E. Cetin, C. Barrado, G. Muñoz, M. Macias, E. Pastor, Drone navigation and avoidance of obstacles through deep reinforcement learning, in: IEEE Digital Avionics Systems Conference, 2019, pp. 1–7.
- [38] S. D. Morad, R. Mecca, R. P. Poudel, S. Liwicki, R. Cipolla, Embodied Visual Navigation with Automatic Curriculum Learning in Real Environments, IEEE Robotics and Automation Letters 6 (2) (2021) 683–690.
- [39] P. Yan, C. Bai, H. Zheng, J. Guo, Flocking Control of UAV Swarms with Deep Reinforcement Learning Approach, in: IEEE International Conference on Unmanned Systems, 2020, pp. 592–599.
- [40] I. Yoon, M. A. Anwar, R. V. Joshi, T. Rakshit, A. Raychowdhury, Hierarchical memory system with STT-MRAM and SRAM to support transfer and real-time reinforcement learning in autonomous drones, IEEE Journal on Emerging and Selected Topics in Circuits and Systems 9 (3) (2019) 485–497.
- [41] G. Williams, N. Wagener, B. Goldfain, P. Drews, J. M. Rehg, B. Boots, E. A. Theodorou, Information theoretic MPC for model-based reinforcement learning, in: IEEE International Conference on Robotics and Automation, 2017, pp. 1714–1721.
- [42] L. He, N. Aouf, J. F. Whidborne, B. Song, Integrated moment-based LGMD and deep reinforcement learning for UAV obstacle avoidance, in: IEEE International Conference on Robotics and Automation, 2020, pp. 7491–7497.
- [43] A. Singla, S. Padakandla, S. Bhatnagar, Memory-based deep reinforcement learning for obstacle avoidance in UAV with limited environment knowledge, IEEE Transactions on Intelligent Transportation Systems (2019).
- [44] T.-C. Wu, S.-Y. Tseng, C.-F. Lai, C.-Y. Ho, Y.-H. Lai, Navigating assistance system for quadcopter with deep reinforcement learning, in: IEEE International Cognitive Cities Conference, 2018, pp. 16–19.
- [45] M. A. Anwar, A. Raychowdhury, NavRen-RL: Learning to fly in real environment via end-to-end deep reinforcement learning using monocular images, in: IEEE International Conference on Mechatronics and Machine Vision in Practice, 2018, pp. 1–6.
- [46] B. Zhou, W. Wang, Z. Wang, B. Ding, Neural Q-learning algorithm based UAV obstacle avoidance, in: IEEE CSAA Guidance, Navigation and Control Conference, 2018, pp. 1–6.
- [47] Z. Yijing, Z. Zheng, Z. Xiaoyi, L. Yang, Q-learning algorithm based UAV path learning and obstacle avoidance approach, in: IEEE Chinese Control Conference, 2017, pp. 3397–3402.
- [48] A. Villanueva, A. Fajardo, Deep Reinforcement Learning with Noise Injection for UAV Path Planning, in: IEEE International Conference on Engineering Technologies and Applied Sciences, 2019, pp. 1–6.
- [49] A. Walvekar, Y. Goel, A. Jain, S. Chakrabarty, A. Kumar, Vision based autonomous navigation of quadcopter using reinforcement learning, in: IEEE International Conference on Automation, Electronics and Electrical Engineering, 2019, pp. 160–165.
- [50] B. Zhou, W. Wang, Z. Liu, J. Wang, Vision-based Navigation of UAV with Continuous Action Space Using Deep Reinforcement Learning, in: IEEE Chinese Control And Decision Conference, 2019, pp. 5030–5035.
- [51] M. Hasanzade, E. Koyuncu, A Dynamically Feasible Fast Replanning Strategy with Deep Reinforcement Learning, Springer Journal of Intelligent & Robotic Systems 101 (1) (2021) 1–17.
- [52] G. Muñoz, C. Barrado, E. Çetin, E. Salami, Deep reinforcement learning for drone delivery, MDPI Drones 3 (3) (2019). doi:10.3390/drones3030072.
- [53] V. J. Hodge, R. Hawkins, R. Alexander, Deep reinforcement learning for drone navigation using sensor data, Springer Neural Computing and Applications 33 (6) (2021) 2015–2033.
- [54] O. Doukhi, D.-J. Lee, Deep Reinforcement Learning for End-to-End Local Motion Planning of Autonomous Aerial Robots in Unknown Outdoor Environments: Real-Time Flight Experiments, MDPI Sensors 21 (7) (2021) 2534. doi:10.3390/s21072534.
- [55] V. A. Bakale, Y. K. VS, V. C. Roodagi, Y. N. Kulkarni, M. S. Patil, S. Chickerur, Indoor Navigation with Deep Reinforcement Learning, in: IEEE International Conference on Inventive Computation Technologies, 2020, pp. 660–665.
- [56] C. J. Maxey, E. J. Shamwell, Navigation and collision avoidance with human augmented supervisory training and fine tuning via reinforcement learning, in: SPIE Micro-and Nanotechnology Sensors, Systems, and Applications XI, Vol. 10982, 2019, pp. 325 – 334. doi:10.1117/12.2518551.
- [57] Y. Zhao, J. Guo, C. Bai, H. Zheng, Reinforcement Learning-Based Collision Avoidance Guidance Algorithm for Fixed-Wing UAVs, Hindawi Complexity 2021 (2021). doi:10.1155/2021/8818013.
- [58] G. Tong, N. Jiang, L. Biyue, Z. Xi, W. Ya, D. Wenbo, UAV navigation in high dynamic environments: A deep reinforcement learning approach, Elsevier Chinese Journal of Aeronautics 34 (2) (2021) 479–489.
- [59] O. Bouhamed, X. Wan, H. Ghazzai, Y. Massoud, A DDPG-Based Approach for Energy-aware UAV Navigation in Obstacle-constrained Environment, in: IEEE World Forum on Internet of Things, 2020, pp. 1–6.
- [60] O. Walker, F. Vanegas, F. Gonzalez, S. Koenig, A Deep Reinforcement Learning Framework for UAV Navigation in Indoor Environments, in: IEEE Aerospace Conference, 2019, pp. 1–14.
- [61] O. Bouhamed, H. Ghazzai, H. Besbes, Y. Massoud, A Generic Spatiotemporal Scheduling for Autonomous UAVs: A Reinforcement Learning-Based Approach, IEEE Open Journal of Vehicular Technology 1 (2020) 93–106.
- [62] J. Zhang, Z. Yu, S. Mao, S. C. Periaswamy, J. Patton, X. Xia, IADRL: Imitation augmented deep reinforcement learning enabled UGV-UAV coalition for tasking in complex environments, IEEE Access 8 (2020) 102335–102347.
- [63] X. Yu, Y. Wu, X.-M. Sun, A Navigation Scheme for a Random Maze Using Reinforcement Learning with Quadrotor Vision, in: IEEE European Control Conference, 2019, pp. 518–523.
- [64] H. Li, S. Wu, P. Xie, Z. Qin, B. Zhang, A Path Planning for One UAV Based on Geometric Algorithm, in: IEEE CSAA Guidance, Navigation and Control Conference, 2018, pp. 1–5.
- [65] D. Sacharny, T. C. Henderson, Optimal Policies in Complex Large-scale UAS Traffic Management, in: IEEE International Conference on Industrial Cyber Physical Systems, 2019, pp. 352–357.
- [66] E. Camci, E. Kayacan, Planning swift maneuvers of quadcopter using motion primitives explored by reinforcement learning, in: IEEE American Control Conference, 2019, pp. 279–285.
- [67] A. Guerra, F. Guidi, D. Dardari, P. M. Djuric, Reinforcement learning for UAV autonomous navigation, mapping and target detection, in: ION Position, Location and Navigation Symposium, 2020, pp. 1004–1013.
- [68] Z. Cui, Y. Wang, UAV Path Planning Based on Multi-Layer Reinforcement Learning Technique, IEEE Access 9 (2021) 59486–59497.
- [69] Z. Wang, H. Li, Z. Wu, H. Wu, A pretrained proximal policy optimization algorithm with reward shaping for aircraft guidance to a moving destination in three-dimensional continuous space, SAGE International Journal of Advanced Robotic Systems 18 (1) (2021). doi:10.1177/1729881421989546.

- [70] H. Eslamiat, Y. Li, N. Wang, A. K. Sanyal, Q. Qiu, Autonomous waypoint planning, optimal trajectory generation and nonlinear tracking control for multi-rotor UAVs, in: IEEE European Control Conference, 2019, pp. 2695–2700.
- [71] N. Imanberdiyev, C. Fu, E. Kayacan, I.-M. Chen, Autonomous Navigation of UAV by using Real-Time Model-Based Reinforcement Learning, in: IEEE International conference on control, automation, robotics and vision, 2016, pp. 1–6.
- [72] S. F. Abedin, M. S. Munir, N. H. Tran, Z. Han, C. S. Hong, Data freshness and energy-efficient UAV navigation optimization: A deep reinforcement learning approach, *IEEE Transactions on Intelligent Transportation Systems* (2020).
- [73] W. Andrew, C. Greatwood, T. Burghardt, Deep learning for exploration and recovery of uncharted and dynamic targets from UAV-like vision, in: IEEE RSJ International Conference on Intelligent Robots and Systems, 2018, pp. 1124–1131.
- [74] H. X. Pham, H. M. La, D. Feil-Seifer, L. Van Nguyen, Reinforcement learning for autonomous UAV navigation using function approximation, in: IEEE International Symposium on Safety, Security, and Rescue Robotics, 2018, pp. 1–6.
- [75] S. Kulkarni, V. Chaphekar, M. M. U. Chowdhury, F. Erden, I. Guvenç, UAV aided search and rescue operation using reinforcement learning, in: IEEE SoutheastCon, Vol. 2, 2020, pp. 1–8.
- [76] A. Peake, J. McCalmon, Y. Zhang, B. Raiford, S. Alqahtani, Wilderness search and rescue missions using deep reinforcement learning, in: IEEE International Symposium on Safety, Security, and Rescue Robotics, 2020, pp. 102–107.
- [77] M. A. Akhloufi, S. Arola, A. Bonnet, Drones chasing drones: Reinforcement learning and deep search area proposal, *MDPI Drones* 3 (3) (2019). doi:10.3390/drones3030058.
- [78] R. Polvara, M. Patacchiola, S. Sharma, J. Wan, A. Manning, R. Sutton, A. Cangelosi, Toward end-to-end control for UAV autonomous landing via deep reinforcement learning, in: IEEE International conference on unmanned aircraft systems, 2018, pp. 115–123.
- [79] R. Polvara, S. Sharma, J. Wan, A. Manning, R. Sutton, Autonomous Vehicular Landings on the Deck of an Unmanned Surface Vehicle using Deep Reinforcement Learning, *Cambridge Core Robotica* 37 (11) (2019) 1867–1882. doi:10.1017/S0263574719000316.
- [80] S. Lee, T. Shim, S. Kim, J. Park, K. Hong, H. Bang, Vision-based autonomous landing of a multi-copter unmanned aerial vehicle using reinforcement learning, in: IEEE International Conference on Unmanned Aircraft Systems, 2018, pp. 108–114.
- [81] C. Wang, J. Wang, X. Zhang, A Deep Reinforcement Learning Approach to Flocking and Navigation of UAVs in Large-Scale Complex Environments, in: IEEE Global Conference on Signal and Information Processing, 2018, pp. 1228–1232.
- [82] G. T. Lee, C. O. Kim, Autonomous Control of Combat Unmanned Aerial Vehicles to Evade Surface-to-Air Missiles Using Deep Reinforcement Learning, *IEEE Access* 8 (2020) 226724–226736.
- [83] Á. Madridano, A. Al-Kaff, P. Flores, D. Martín, A. de la Escalera, Software Architecture for Autonomous and Coordinated Navigation of UAV Swarms in Forest and Urban Firefighting, *MDPI Applied Sciences* 11 (3) (2021). doi:10.3390/app11031258.
- [84] D. Wang, T. Fan, T. Han, J. Pan, A Two-Stage Reinforcement Learning Approach for Multi-UAV Collision Avoidance Under Imperfect Sensing, *IEEE Robotics and Automation Letters* 5 (2) (2020) 3098–3105.
- [85] J. Moon, S. Papaioannou, C. Laoudias, P. Kolios, S. Kim, Deep Reinforcement Learning Multi-UAV Trajectory Control for Target Tracking, *IEEE Internet of Things Journal* (2021).
- [86] C. H. Liu, X. Ma, X. Gao, J. Tang, Distributed energy-efficient multi-UAV navigation for long-term communication coverage by deep reinforcement learning, *IEEE Transactions on Mobile Computing* 19 (6) (2019) 1274–1285.
- [87] S. Omi, H.-S. Shin, A. Tsourdos, J. Espeland, A. Buchi, Introduction to UAV swarm utilization for communication on the move terminals tracking evaluation with reinforcement learning technique, in: IEEE European Conference on Antennas and Propagation, 2021, pp. 1–5.
- [88] A. Viseras, M. Meissner, J. Marchal, Wildfire Front Monitoring with Multiple UAVs using Deep Q-Learning, *IEEE Access* (2021).
- [89] A. Bonnet, M. A. Akhloufi, UAV pursuit using reinforcement learning, in: SPIE Unmanned Systems Technology XXI, Vol. 11021, International Society for Optics and Photonics, 2019, pp. 51 – 58. doi:10.1117/12.2520310.
- [90] S. Fan, G. Song, B. Yang, X. Jiang, Prioritized Experience Replay in Multi-Actor-Attention-Critic for Reinforcement Learning, *IOP-science Journal of Physics: Conference Series* 1631 (2020). doi:10.1088/1742-6596/1631/1/012040.
- [91] A. Majd, A. Ashraf, E. Troubitsyna, M. Daneshtalab, Integrating learning, optimization, and prediction for efficient navigation of swarms of drones, in: IEEE Euromicro International Conference on Parallel, Distributed and Network-based Processing, 2018, pp. 101–108.
- [92] A. Chapman, Drone Types: Multi-Rotor vs Fixed-Wing vs Single Rotor vs Hybrid VTOL, <https://www.auav.com.au/articles/drone-types/>, (Accessed: 01.11.2021) (2016).
- [93] M. Elnaggar, N. Bezzo, An IRL Approach for Cyber-Physical Attack Intention Prediction and Recovery, in: IEEE American Control Conference, 2018, pp. 222–227.
- [94] H. D. Escobar-Alvarez, N. Johnson, T. Hebble, K. Klingebiel, S. A. Quintero, J. Regenstein, N. A. Browning, R-ADVANCE: Rapid Adaptive Prediction for Vision-based Autonomous Navigation, Control, and Evasion, *WOL Journal of Field Robotics* 35 (1) (2018) 91–100. doi:10.1002/rob.21744.
- [95] C. W. Reynolds, Flocks, herds and schools: A distributed behavioral model, in: Proceedings of the 14th annual conference on Computer graphics and interactive techniques, 1987, pp. 25–34.
- [96] R. Olfati-Saber, Flocking for multi-agent dynamic systems: Algorithms and theory, *IEEE Transactions on automatic control* 51 (3) (2006) 401–420.
- [97] H. M. La, W. Sheng, Flocking control of multiple agents in noisy environments, in: IEEE International Conference on Robotics and Automation, 2010, pp. 4964–4969.
- [98] Y. Jia, J. Du, W. Zhang, L. Wang, Three-dimensional leaderless flocking control of large-scale small unmanned aerial vehicles, *Elsevier IFAC-PapersOnLine* 50 (1) (2017) 6208–6213.
- [99] H. Su, X. Wang, Z. Lin, Flocking of multi-agents with a virtual leader, *IEEE transactions on automatic control* 54 (2) (2009) 293–307.
- [100] S. A. Quintero, G. E. Collins, J. P. Hespanha, Flocking with fixed-wing UAVs for distributed sensing: A stochastic optimal control approach, in: IEEE American Control Conference, 2013, pp. 2025–2031.
- [101] S.-M. Hung, S. N. Givigi, A Q-learning approach to flocking with UAVs in a stochastic environment, *IEEE transactions on cybernetics* 47 (1) (2016) 186–197.
- [102] K. Morihiro, T. Isokawa, H. Nishimura, M. Tomimasu, N. Kamiura, N. Matsui, Reinforcement Learning Scheme for Flocking Behavior Emergence, *Journal of Advanced Computational Intelligence and Intelligent Informatics* 11 (2) (2007) 155–161.
- [103] Z. Xu, Y. Lyu, Q. Pan, J. Hu, C. Zhao, S. Liu, Multi-vehicle flocking control with deep deterministic policy gradient method, in: IEEE International Conference on Control and Automation, 2018, pp. 306–311.
- [104] O. Robotics, ROS Home Page, <https://www.ros.org/>, (Accessed: 01.11.2021) (2021).
- [105] M. Research, Microsoft AirSim Home Page, <https://github.io/AirSim/>, (Accessed: 01.11.2021) (2021).
- [106] O. S. R. Foundation, Gazebo Home Page, <https://gazebosim.org/>, (Accessed: 01.11.2021) (2021).
- [107] E. Games, Epic Games Unreal Engine Home Page, <https://www.unrealengine.com>, (Accessed: 01.11.2021) (2021).
- [108] C. J. Watkins, P. Dayan, Q-learning, *Springer Machine learning* 8 (3–4) (1992) 279–292. doi:10.1007/978-1-4615-3618-5_4.
- [109] A. Fotouhi, M. Ding, M. Hassan, Deep Q-Learning for Two-Hop Communications of Drone Base Stations, *MDPI Sensors* 21 (6)

- (2021). doi:10.3390/s21061960.
- [110] G. A. Rummery, M. Niranjan, On-line Q-learning using connectionist systems, Vol. 37, University of Cambridge, 1994.
- [111] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, M. Riedmiller, Playing Atari with Deep Reinforcement Learning, arXiv:1312.5602 (2013).
- [112] H. Huang, Y. Yang, H. Wang, Z. Ding, H. Sari, F. Adachi, Deep reinforcement learning for UAV navigation through massive MIMO technique, *IEEE Transactions on Vehicular Technology* 69 (1) (2019) 1117–1121.
- [113] H. Van Hasselt, A. Guez, D. Silver, Deep reinforcement learning with double Q-Learning, *AAAI Conference on Artificial Intelligence* (2016) 2094–2100.
- [114] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, N. Freitas, Dueling Network Architectures for Deep Reinforcement Learning, in: PMLR International Conference on Machine Learning, Vol. 48, 2016, pp. 1995–2003.
- [115] M. Hausknecht, P. Stone, Deep Recurrent Q-learning for partially observable MDPS, arXiv:1507.06527 (2015).
- [116] M. Fortunato, M. G. Azar, B. Piot, J. Menick, I. Osband, A. Graves, V. Mnih, R. Munos, D. Hassabis, O. Pietquin, et al., Noisy networks for exploration, arXiv:1706.10295 (2017).
- [117] M. G. Bellemare, W. Dabney, R. Munos, A distributional perspective on reinforcement learning, in: PMLR International Conference on Machine Learning, 2017, pp. 449–458.
- [118] W. Dabney, M. Rowland, M. G. Bellemare, R. Munos, Distributional reinforcement learning with quantile regression, *AAAI Conference on Artificial Intelligence* (2018).
- [119] W. Dabney, G. Ostrovski, D. Silver, R. Munos, Implicit quantile networks for distributional reinforcement learning, in: PMLR International conference on machine learning, 2018, pp. 1096–1105.
- [120] M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, D. Silver, Rainbow: Combining improvements in deep reinforcement learning, *AAAI Conference on Artificial Intelligence* (2018).
- [121] D. Yang, L. Zhao, Z. Lin, T. Qin, J. Bian, T.-Y. Liu, Fully parameterized quantile function for distributional reinforcement learning, *Advances in Neural Information Processing Systems* 32 (2019) 6193–6202.
- [122] S. Kapturowski, G. Ostrovski, J. Quan, R. Munos, W. Dabney, Recurrent experience replay in distributed reinforcement learning, *International Conference on Learning Representations* (2018).
- [123] D. Horgan, J. Quan, D. Budden, G. Barth-Maron, M. Hessel, H. Van Hasselt, D. Silver, Distributed prioritized experience replay, arXiv:1803.00933 (2018).
- [124] A. P. Badia, P. Sprechmann, A. Vitvitskyi, D. Guo, B. Piot, S. Kapturowski, O. Tielemans, M. Arjovsky, A. Pritzel, A. Bolt, et al., Never give up: Learning directed exploration strategies, arXiv:2002.06038 (2020).
- [125] A. P. Badia, B. Piot, S. Kapturowski, P. Sprechmann, A. Vitvitskyi, Z. D. Guo, C. Blundell, Agent57: Outperforming the atari human benchmark, in: PMLR International Conference on Machine Learning, 2020, pp. 507–517.
- [126] D. Zhao, H. Wang, K. Shao, Y. Zhu, Deep reinforcement learning with experience replay based on SARSA, in: IEEE Symposium Series on Computational Intelligence, 2016, pp. 1–6.
- [127] R. J. Williams, Simple statistical gradient-following algorithms for connectionist reinforcement learning, *Springer Machine learning* 8 (3-4) (1992) 229–256.
- [128] J. Schulman, S. Levine, P. Abbeel, M. Jordan, P. Moritz, Trust Region Policy Optimization, in: PMLR International Conference on Machine Learning, Vol. 37, 2015, pp. 1889–1897.
- [129] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal Policy Optimization Algorithms, arXiv:1707.06347 (2017).
- [130] K. Cobbe, J. Hilton, O. Klimov, J. Schulman, Phasic Policy Gradient, arXiv:2009.04416 (2020).
- [131] Y. Liu, P. Ramachandran, Q. Liu, J. Peng, Stein Variational Policy Gradient, arXiv:1704.02399 (2017).
- [132] A. X. Lee, A. Nagabandi, P. Abbeel, S. Levine, Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model, arXiv:1907.00953 (2019).
- [133] S. Zhang, H. Yao, ACE: An Actor Ensemble Algorithm for continuous control with tree search, in: AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 5789–5796. doi:10.1609/aaai.v33i01.33015789.
- [134] S. Zhang, S. Whiteson, DAC: The double actor-critic architecture for learning options, arXiv:1904.12691 (2019).
- [135] N. Heess, J. J. Hunt, T. P. Lillicrap, D. Silver, Memory-Based Control with Recurrent Neural Networks, arXiv:1512.04455 (2015).
- [136] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, D. Wierstra, Continuous control with deep reinforcement learning, arXiv:1509.02971 (2015).
- [137] S. Fujimoto, H. Hoof, D. Meger, Addressing function approximation error in actor-critic methods, in: PMLR International Conference on Machine Learning, 2018, pp. 1587–1596.
- [138] T. Haarnoja, A. Zhou, P. Abbeel, S. Levine, Soft Actor-Critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, in: PMLR International Conference on Machine Learning, 2018, pp. 1861–1870.
- [139] G. Barth-Maron, M. W. Hoffman, D. Budden, W. Dabney, D. Horgan, D. Tb, A. Muldal, N. Heess, T. Lillicrap, Distributed distributional deterministic policy gradients, arXiv:1804.08617 (2018).
- [140] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, K. Kavukcuoglu, Asynchronous methods for deep reinforcement learning, in: PMLR International Conference on Machine Learning, 2016, pp. 1928–1937.
- [141] N. Heess, D. TB, S. Sriram, J. Lemmon, J. Merel, G. Wayne, Y. Tassa, T. Erez, Z. Wang, S. Eslami, et al., Emergence of locomotion behaviours in rich environments, arXiv:1707.02286 (2017).
- [142] C. Alfredo, C. Humberto, C. Arjun, Efficient parallel methods for deep reinforcement learning, in: The Multi-disciplinary Conference on Reinforcement Learning and Decision Making, 2017, pp. 1–6.
- [143] Z. Wang, V. Bapst, N. Heess, V. Mnih, R. Munos, K. Kavukcuoglu, N. de Freitas, Sample efficient actor-critic with experience replay, arXiv:1611.01224 (2016).
- [144] A. Gruslys, W. Dabney, M. G. Azar, B. Piot, M. Bellemare, R. Munos, The reactor: A fast and sample-efficient actor-critic agent for reinforcement learning, arXiv:1704.04651 (2017).
- [145] Y. Wu, E. Mansimov, R. B. Grosse, S. Liao, J. Ba, Scalable Trust-Region method for deep reinforcement learning using Kronecker-Factored Approximation, in: Advances in Neural Information Processing Systems, 2017, pp. 5279–5288.
- [146] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, I. Mordatch, Multi-Agent Actor-Critic for mixed cooperative-competitive environments, arXiv:1706.02275 (2017).
- [147] J. Ackermann, V. Gabler, T. Osa, M. Sugiyama, Reducing overestimation bias in multi-agent domains using double centralized critics, arXiv:1910.01465 (2019).
- [148] S. Iqbal, F. Sha, Actor-Attention-Critic for Multi-Agent Reinforcement Learning, in: PMLR International Conference on Machine Learning, 2019, pp. 2961–2970.
- [149] L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, Y. Doron, V. Firoiu, T. Harley, I. Dunning, et al., IMPALA: Scalable distributed deep-rl with importance weighted actor-learner architectures, in: PMLR International Conference on Machine Learning, 2018, pp. 1407–1416.
- [150] L. Espeholt, R. Marinier, P. Stanczyk, K. Wang, M. Michalski, Seed RL: Scalable and Efficient Deep-RL with accelerated central inference, arXiv:1910.06591 (2019).
- [151] H. Hasselt, Double Q-learning, *Advances in Neural Information Processing Systems* 23 (2010) 2613–2621.
- [152] Y. Burda, H. Edwards, A. Storkey, O. Klimov, Exploration by random network distillation, arXiv:1810.12894 (2018).
- [153] R. S. Sutton, D. A. McAllester, S. P. Singh, Y. Mansour, Policy gradient methods for reinforcement learning with function approximation, in: Advances in Neural Information Processing Systems, 2000, pp.

1057–1063.

- [154] S.-Y. Shin, Y.-W. Kang, Y.-G. Kim, Obstacle Avoidance Drone by Deep Reinforcement Learning and Its Racing with Human Pilot, MDPI Applied Sciences 9 (24) (2019). doi:10.3390/app9245571.
- [155] Q. Liu, D. Wang, Stein variational gradient descent: A general purpose bayesian inference algorithm, arXiv:1608.04471 (2016).
- [156] D. Wierstra, A. Förster, J. Peters, J. Schmidhuber, Recurrent Policy Gradients, Logic Journal of the IGPL 18 (5) (2010) 620–634.
- [157] J. Martens, R. Grosse, Optimizing Neural Networks with Kronecker-Factored Approximate Curvature, in: PMLR International conference on machine learning, 2015, pp. 2408–2417.
- [158] R. Munos, T. Stepleton, A. Harutyunyan, M. G. Bellemare, Safe and efficient off-policy reinforcement learning, arXiv:1606.02647 (2016).
- [159] M. Andrychowicz, A. Raichuk, P. Stanczyk, M. Orsini, S. Girgin, R. Marinier, L. Huszenot, M. Geist, O. Pietquin, M. Michalski, et al., What Matters In On-Policy Reinforcement Learning? A Large-Scale Empirical Study, CoRR abs/2006.05990 (2020).



Fadi AlMahamid received the B.Sc. degree (Hons.) in Computer Science from Princess Sumaya University for Technology (PSUT), Amman, Jordan, in 2001, and M.Sc. degree (Hons.) in Computer Science from New York Institute of Technology (NYIT), Amman, Jordan, in 2003. Also, he obtained another M.Sc. from the University of Western Ontario (UWO), London, Ontario, Canada, in 2019, where he is currently pursuing the Ph.D. degree in Software Engineering with the Department of Electrical and Computer Engineering. He has extensive industry experience of more than 15 years. His current research interests include machine learning, autonomous vehicles focusing on navigation problems, IoT architectures, and sensor data analytics.



Katarina Grolinger received the B.Sc. and M.Sc. degrees in mechanical engineering from the University of Zagreb, Croatia, and the M.Eng. and Ph.D. degrees in software engineering from Western University, London, Canada. She is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Western University. She has been involved in the software engineering area in academia and industry, for over 20 years. Her current research interests include machine learning, sensor data analytics, data management, and the IoT.