# An improved Q-learning path planning algorithm based on IAPF

Jiahao Zhao [1, *], Shizhuo Chen [2]

[1] College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin, China, 150001

[2] College of Energy and Power Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China, 211106

* Corresponding Author Email: heuzgy1112@163.com

**Abstract.** Unmanned Aerial Vehicles (UAVs), particularly quadrotor UAVs, are widely recognized for their cost-effectiveness, operational flexibility, and vertical takeoff and landing capabilities, making them ideal for specific airspace operations and emergency response scenarios. Efficient path planning is critical for UAV mission execution, requiring optimal flight paths that ensure safe obstacle avoidance while addressing challenges such as dynamic environments, energy optimization, and multi-parameter management. Despite advancements in path planning techniques, including the artificial potential field (APF) method and reinforcement learning (RL), issues like local optima and parameter tuning complexity persist, limiting adaptability in dynamic environments. This study proposes a hybrid approach integrating Markov Decision Process (MDP) theory with an improved artificial potential field (IAPF) method to enhance quadrotor UAV path planning in three-dimensional environments. By generating global waypoints based on known obstacle, the method minimizes flight path deviations and improves navigation performance. The results demonstrate significant advancements in trajectory accuracy and adaptability, offering a robust solution for UAV path optimization in complex scenarios.

**Keywords:** Q-learning, APF, Reinforce learning, Path planning.

## 1. Introduction

UAVs are pilot-less flight platforms that operate based on fluid dynamics principles. They can perform missions autonomously or via remote control and are often reusable, with some models capable of carrying equipment or payloads [1].

Structurally, UAVs are classified into three main types: fixed-wing, flapping-wing, and rotary-wing [2]. Among these, quadrotor UAVs are particularly notable for their cost-effectiveness, ease of operation, vertical takeoff and landing capabilities, and ability to hover. These advantages make them well-suited for specific airspace operations and emergency response scenarios.

Efficient path planning is crucial for quadrotor UAVs to execute missions effectively. An optimal flight path must be determined while ensuring safe obstacle avoidance. However, path planning faces challenges such as adapting to dynamic environments, optimizing time and energy consumption, and managing multiple parameters, including altitude, velocity, and acceleration [3]. Consequently, developing advanced technologies for adaptive and intelligent path planning remains a key research focus.

Extensive research has been conducted on quadrotor UAV path planning, leading to significant advancements. Lu [4] employed the APF method for low-altitude obstacle avoidance and proposed a "selective crossing method" to mitigate local minima. Si [5] introduced RL to improve convergence in multi-UAV path planning, refining the Q-learning algorithm. Kan [6] applied MDP theory with a negative reward mechanism in Q-learning to visualize obstacle threat zones and enhance trajectory planning. Guo [7] integrated deep RL with the APF method, optimizing its parameters through DDPG-based training for increased robustness in complex scenarios. Yao [8] combined an optimized black hole potential field model with RL, leveraging curriculum learning to enhance environmental

adaptability. Despite these advancements, the APF method still struggles with local optima and parameter tuning complexity, limiting its ability to dynamically adjust to environmental changes.

To address these limitations, this study investigates the path optimization of quadrotor UAVs in three-dimensional environments. By integrating MDP theory with an improved APF method, global waypoints are generated based on known obstacle and threat constraints. This hybrid approach enhances trajectory accuracy by minimizing flight path deviations, ultimately improving the navigation performance of UAVs.

## 2. Artificial potential field

### 2.1. Traditional APF

The APF method, introduced by Khatib in 1986 [9], constructs a virtual force field within the mission space of an UAV. This field consists of two components: an attractive potential field generated by the target and a repulsive potential field induced by obstacles. The attractive field exerts a force that guides the UAV toward the target, while the repulsive field generates a force that pushes the UAV away from obstacles.

As the UAV navigates within this potential field, it experiences a resultant force determined by the vector sum of the attractive and repulsive forces. The UAV adjusts its trajectory and velocity accordingly, enabling obstacle avoidance while progressing toward the target. Effective path planning is achieved by appropriately tuning the potential function parameters and step size. As illustrated in Fig.1, the quadrotor moves under the influence of the combined attractive force and repulsive forces and $F_{rep2}$, following a trajectory shaped by the resultant force $F_{all}$.

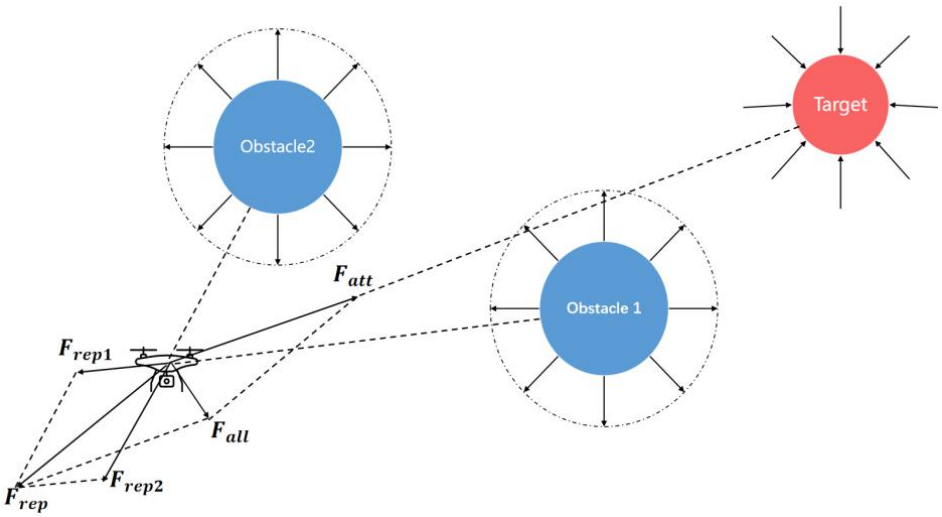



**Figure 1.** APF method

In the traditional APF method, the attractive potential field function is:

$$\begin{cases} U_{att}(\rho_t) = \frac{1}{2} k_{att} \rho_t^2 \\ \rho_t = \|p_t - p_{UAV}\| \end{cases} \tag{1}$$

Where $U_{att}$ represents the magnitude of the attractive potential field, $k_{att}$ is the attractive gain coefficient, $\rho_t$ denotes the Euclidean distance from the UAV to the target point, $p_t$ represents the coordinates of the target point, and represents the coordinates of the UAV.

In the attractive potential field, the magnitude of the attractive force is the negative gradient of the attractive potential field, and its mathematical expression is as follows:

$$F_{att}(\rho_t) = -\nabla U_{att}(\rho_t) = -\eta \rho_t \tag{2}$$

In the traditional APF method, the repulsive potential field function is:

$$\begin{cases} U_{rep}(\rho_g) = \begin{cases} \frac{1}{2}k_{rep}(\frac{1}{\rho_g} - \frac{1}{\rho_0})^2, \rho_g \le \rho_0 \\ 0, \rho_g > \rho_0 \end{cases} \\ \rho_g = \|p_{UAV} - p_{obs}\| \end{cases} \tag{3}$$

Where $U_{att}$ represents the magnitude of the repulsive potential field, $k_{rep}$ is the repulsive gain coefficient, $\rho_g$ denotes the distance between the UAV and the obstacle, represents the maximum influence distance of the obstacle, represents the coordinates of the UAV, and $p_{obs}$ represents the coordinates of the obstacle.

In the repulsive potential field, the magnitude of the repulsive force is the negative gradient of the repulsive potential field, and its mathematical expression is as follows:

$$F_{rep}(\rho_g) = \begin{cases} k_{rep}(\frac{1}{\rho_g} - \frac{1}{\rho_0})\frac{1}{\rho_g^2}\nabla\rho_g, \rho_g \le \rho_0 \\ 0, \rho_g > \rho_0 \end{cases} \tag{4}$$

During the flight, the UAV is influenced by both the attractive potential field and the repulsive potential field. Typically, in path planning problems, there is usually one target point and one or more obstacles. Therefore, the resultant potential field acting on the UAV is:

$$U_{all} = U_{att}(\rho_t) + \sum U_{rep}(\rho_t) \tag{5}$$

The resultant force acting on the UAV in the resultant potential field is:

$$F_{all} = -\nabla U_{all} = F_{att}(\rho_t) + \sum F_{rep}(\rho_t) \tag{6}$$

### 2.2. Improved artificial potential field

Traditional APF have two limitations: the target unreachability problem and the local minimum problem.

The target unreachability problem can generally be explained as follows: As shown in Fig.2, when there are obstacles around the target point, and the UAV is about to reach the target point while being near these obstacles, according to the definitions of the repulsive potential field and the attractive potential field, the UAV will experience an attractive force $F_{att}$ and a repulsive force $F_{rep}$ At this point, the resultant force $F_{all}$ on the UAV will not point towards the target, causing the UAV to hover and rotate around the target point without actually reaching it.

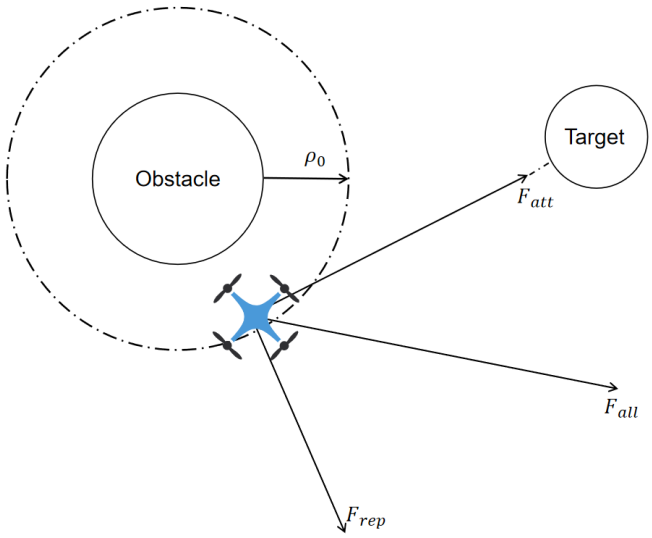



**Figure 2.** Target unreachability problem

The local minimum problem generally manifests in two cases. As shown in Fig.3, case 1 occurs when there is only one obstacle near the target point. If the obstacle and the target point are aligned in

a straight line, the UAV will experience an attractive force $F_{att}$ and a repulsive force $F_{rep}$ in opposite directions along the same line. When the UAV moves to a certain point, it will become trapped in this local minimum, where the attractive force $F_{att}$ and the repulsive force $F_{rep}$ balance each other, resulting in a resultant force $F_{all}$ of zero, causing the UAV to stop moving.

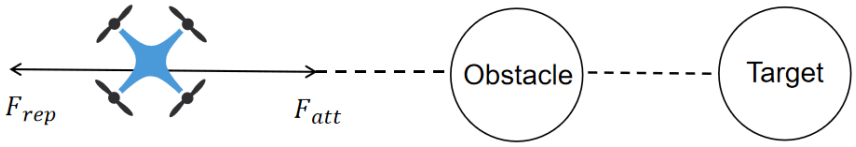



**Figure 3.** Local minimum problem Case 1

As shown in Fig.4, the second case occurs when there are two obstacles near the target point. If these two obstacles are symmetrically distributed about the horizontal central axis of the target point, the UAV will experience symmetrical repulsive forces $F_{rep1}$ and $F_{rep2}$. The attractive force $F_{att}$ and the combined repulsive force $F_{rep}$ will be aligned in the same line but in opposite directions. When the UAV moves to a certain point, it will become trapped in this local minimum, where the attractive force $F_{att}$ and the combined repulsive force $F_{rep}$ balance each other, resulting in a resultant force $F_{all}$ of zero, causing the UAV to stop moving.

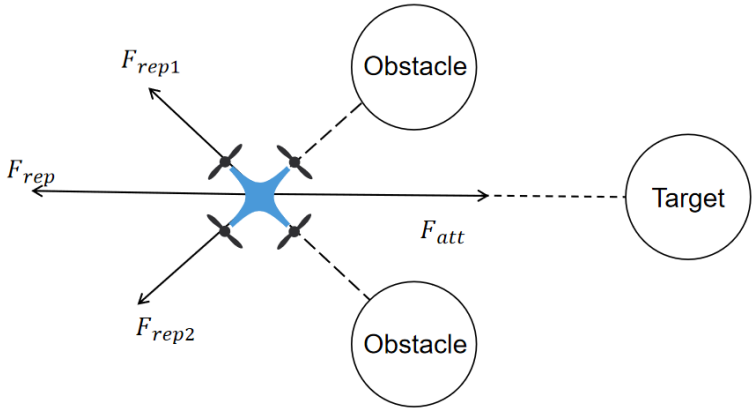



**Figure 4.** Local minimum problem Case 2

To address this issue, introducing the relative distance between the UAV and the target into the repulsive potential field of the obstacles. By modifying the repulsive force, the problem of target unreachability is resolved. Additionally, an auxiliary potential field is introduced to solve the local minimum problem.

### 2.2.1. Target unreachability problem

Introducing the distance $\rho_t$ between the UAV and the target point:

$$U_{rep}(\rho_g, \rho_t) = \begin{cases} \frac{1}{2} k_{rep} \left(\frac{1}{\rho_g} - \frac{1}{\rho_0}\right)^2 {\rho_t}^2, \rho_g \le \rho_0 \\ 0, \rho_g > \rho_0 \end{cases} \tag{7}$$

The repulsive force is then modified as:

$$F_{rep}(\rho_g, \rho_t) = \begin{cases} F_{rep1} + F_{rep2}, \rho_g \le \rho_0 \\ 0, \rho_g > \rho_0 \end{cases} \tag{8}$$

$F_{rep1}$ is the force from the obstacle pointing to the UAV:

$$F_{rep1}(\rho_g, \rho_t) = k_{rep} \left(\frac{1}{\rho_g} - \frac{1}{\rho_0}\right) \frac{{\rho_t}^2}{\rho_g^2} \tag{9}$$

$F_{rep2}$ is the force from the UAV pointing to the target point:

$$F_{rep2}(\rho_g, \rho_t) = k_{rep}\left(\frac{1}{\rho_g} - \frac{1}{\rho_0}\right)^2 \rho_t \tag{10}$$

After modification, as shown in Fig.5, by introducing the distance $\rho_t$ , the repulsive force $F_{rep}$ not only has a component pointing from the obstacle to the UAV but also has a component pointing from the UAV to the target.

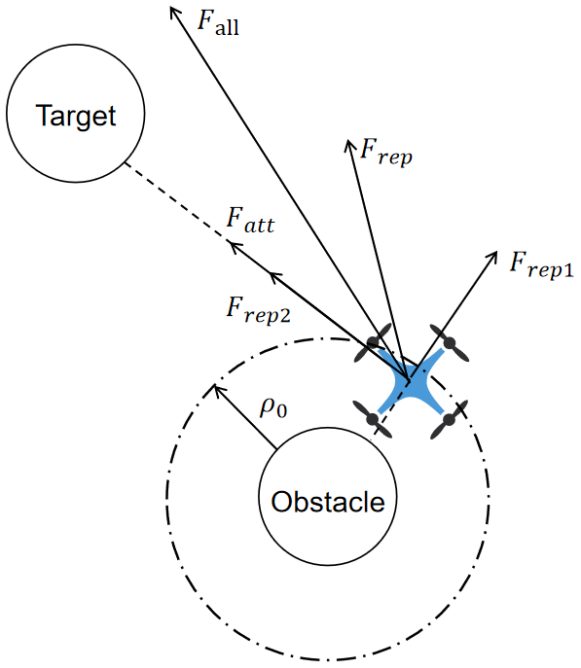



**Figure 5.** Modification of Target unreachability problem

### 2.2.2. Local minimum problem

Introducing an additional potential field to drive the UAV to escape from local minima:

$$\begin{cases} U_{add}(\rho_l) = \begin{cases} \frac{1}{2} k_{add} \frac{1}{\rho_l^2}, \rho_l \le \rho_{amax} \\ 0, \rho_l > \rho_{amax} \end{cases} \\ \rho_l = \|p_{UAV} - p_{local}\| \end{cases} \tag{11}$$

Where$k_{add}$is the additional gain coefficient, $\rho_l$ is the distance between the UAV and the local minimum, and $\rho_{amax}$ is the maximum influence range of the additional potential field.

After modification, as shown in Fig.6, it can be observed that once the UAV approaches the local minimum, the additional potential field force $F_{add}$ will disrupt the previous equilibrium state, and the local minimum will then be bypassed.

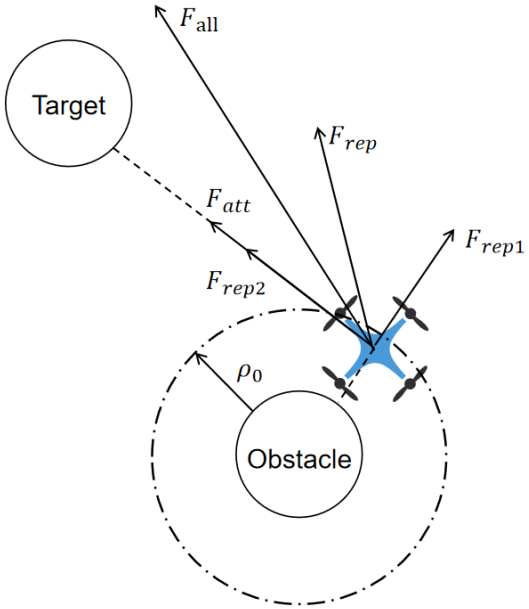



**Figure 6.** Modification of Local minimum problem

## 3. Reinforce learning algorithm

### 3.1. Markov Decision Process

The MDP plays a significant role in handling the interaction between artificial intelligence systems and their surrounding environment. In an MDP, state transitions adhere to the Markov property: future outcomes are influenced solely by the current action and the current state, rather than any historical information from the past. An important concept introduced in the MDP is the reward function (also known as the return function), which is used to evaluate the quality of actions taken by the agent.

A complete MDP can be described by a quintuple$(S, A, P, R, \gamma)$, with the following meanings:

$S = \{s_1, s_2, \cdots, s_i\}$ represents the state space, indicating all possible states.

$A = \{a_1, a_2, \cdots, a_i\}$represents the action space, indicating all possible actions that the agent can take.

$P$ is the state transition probability matrix, describing the probability of transitioning to other states after taking a certain action in the current state.

$\mathrm{R} = \{\mathrm{r}_1, \mathrm{r}_2, \cdots, \mathrm{r}_\mathrm{i}\}$is the reward function, describing the reward obtained by the agent when transitioning from the current state to the next state after taking a certain action.

$\gamma$is the discount factor, used to control the discount rate of future rewards, determining the importance of future rewards relative to immediate rewards. Generally, $0 < \gamma < 1$, when $\gamma$ is close to 1, it indicates that delayed rewards are more important than immediate rewards, and vice versa.

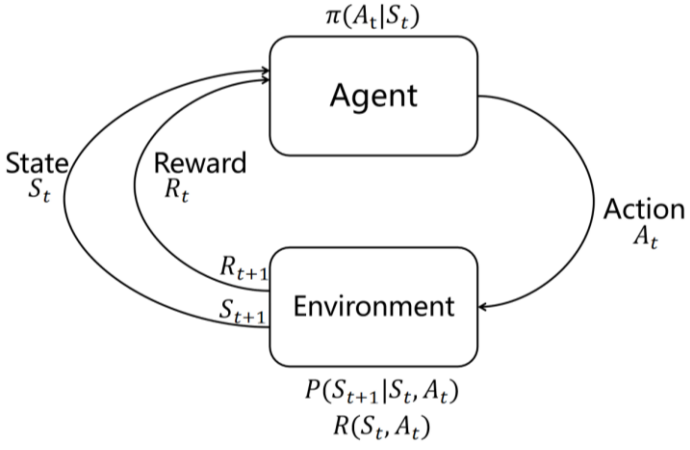



**Figure 7.** MDP process

As shown in Fig.7, when an Agent is performing a task, it first interacts with the environment, generating a new state, while the environment provides a reward. This cycle continues, with the agent and environment continuously interacting to produce more new data.

In the MDP, the connection between actions and states is referred to as a policy, denoted by the symbol Here, the policy $\pi$ refers to the probability distribution of selecting various possible actions under certain conditions, which is the process of choosing and executing actions.

The goal of the MDP is to find the mapping between states and actions that maximizes cumulative rewards. At a given time $t$ , the cumulative reward that the agent can obtain is defined as the state value function expressed by the formula:

$$V^{\pi}(s_t) = r_1 + \gamma\, r_2 + \gamma^2 r_3 + \cdots + \gamma^{n-1} r_n \tag{12}$$

The state-action value function under policy $\pi$ at action $A$ is defined as the expected return when taking action $a$ in state $s$ , and subsequently following policy $\pi$, typically denoted by$Q^{\pi}(s,a)$. Its Bellman equation is:

$$Q^{\pi}(s,a) = r_t + \gamma\big(\, V^{\pi}(s_{t+1})\big) \tag{13}$$

### 3.2. Q-learning algorithm

Q-learning is a typical model-free RL method, with its most notable feature being its independence from a model, relying instead on temporal difference learning.

The core principle of this method is to construct a state-action table (Q-table) that records reward values for different actions under various conditions. Through iterative interaction with the environment, the Q-table is updated based on the current state, the action taken, the environmental feedback, and the subsequent state. The Bellman equation is employed to refine the Q-values, enabling the Q-table to progressively converge toward optimal values. This process allows the agent to make optimal decisions in uncertain environments.

The update method of Q-learning can be expressed by the following formula:

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha[r_t + \gamma\, max\, Q(s_{t+1}, a) - Q(s_t, a_t)] \tag{14}$$

In the formula, $Q(s_t, a_t)$ represents the reward value function obtained by taking the corresponding action $a_t$ in state $s_t$ , $max\ Q(s_{t+1}, a)$represents the best action value function obtained in the next state, and $r_t$ represents the instantaneous reward obtained at the moment.

The algorithmic process is shown in Fig.8:

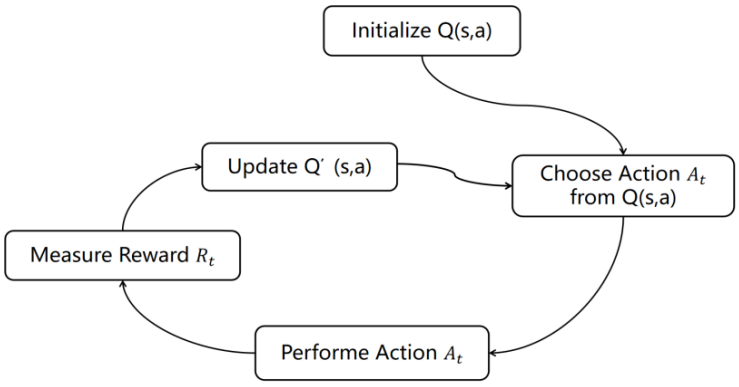



**Figure 8.** Q-learning process

## 4. RL-IAPF Algorithm

To meet the accuracy requirements for obstacle avoidance in UAV path planning tasks, this paper proposes a Q-learning-based UAV path planning algorithm combined with an improved APF method. Based on the MDP model, the algorithm integrates and optimizes the APF method, utilizing known

map obstacles and threat information to plan a global path on the map. The flow of the algorithm is shown in the Fig.9.

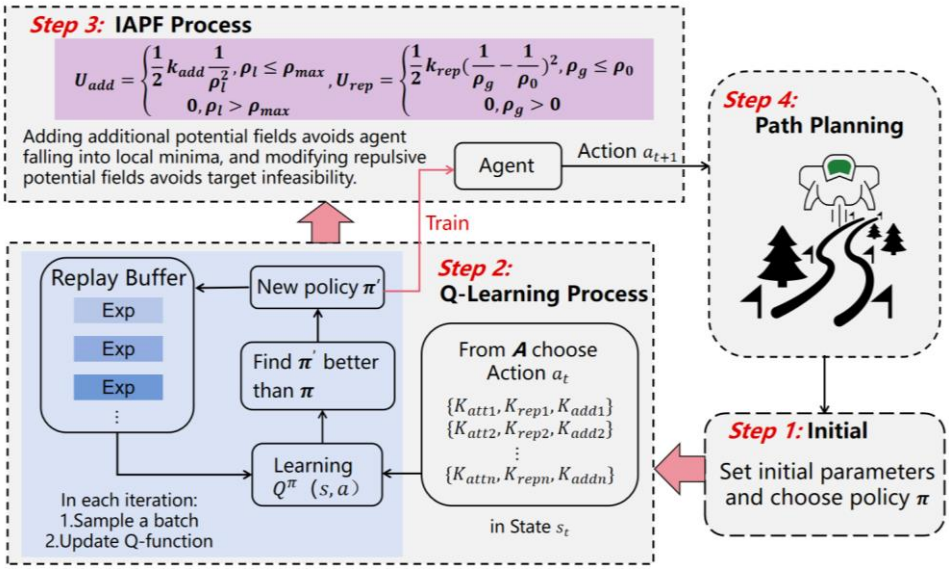



**Figure 9.** RL-IAPF flow

### 4.1. Definition of RL Elements

From Section 3. 1, it is known that a complete MDP can be described by a quintuple $(S, A, P, R, \gamma)$.

In this paper, the state space $S = \{s_1, s_2, \cdots, s_i\}$ is mainly composed of the relative position information between the UAV and the target point, as well as the relative position information of obstacles. It is defined as follows:

$$S = [s_1^T, s_2^T, s_3^T]^T$$

represents the current position of the UAV in the coordinate system; $s_2$ represents the relative position between the UAV and obstacles or threat zones; $s_3$ represents the relative position between the UAV and the target point.

Based on the kinematic model of the UAV and the APF method, the action space is defined as follows:

$$A = [{k_{att}}^T, {k_{rep}}^T, {k_{add}}^T]^T$$

$k_{att}$ is the attraction gain coefficient, $k_{rep}$ is the repulsion gain coefficient, and $k_{add}$ is the additional gain coefficient, which are the attraction, repulsion, and additional gain coefficients optimized in the APF method in this paper.

The reward function is the feedback signal received by the agent in the environment when executing actions in RL. This value quantifies the immediate return of the agent taking a specific action in a specific state, directly affecting the learning process of the agent. Daniel Dewey advocates the "reward engineering principle" [11]: as AI systems based on RL become more common and autonomous, designing reward mechanisms that elicit desired behaviors becomes more important and challenging.

Before designing the reward mechanism, the main event should be clarified. For example, the main event in this paper is: in the navigation task, the UAV reaches the target point in the safest and shortest distance.

The reward for this goal is the mainline reward, i. e., the reward function in path planning, which is usually expressed as follows [12]:

$$\mathrm{R} = \begin{cases} 1, \text{reach target} \\ -1, \text{reach obstacle} \end{cases}$$

However, using only the mainline reward often leads to the sparse reward problem. To address the sparse reward problem, some dense reward mechanisms are added on the basis of the mainline reward

to optimize the reward function, so that the UAV can continuously receive feedback throughout the path planning process, thereby accelerating the learning speed.

Therefore, the reward function set in this paper mainly includes the following rewards:

Mainline reward: Distance Reward, Obstacle Avoidance Reward [10]

Sparse reward: Direction Reward [13], Escaping Reward.

Their definitions are shown in the following Table. 1:

**Table 1.** Setting of Reward Functions

| Type | Formula |
|---|---|
| Distance Reward | $r_{dis} = -k_{att} \times (\frac{d_{u,g}}{d_{s,g}}) + r_c$ |
| Obstacle Avoidance Reward | $r_{avo} = \begin{cases} k_{rep} \times (\frac{d_{u,o} - R_{obs}}{\rho_o}), d_{u,o} > R_{obs} \\ k_{rep} \times (\frac{d_{u,o} - R_{obs}}{R_{obs}}), d_{u,o} \le R_{obs} \end{cases}$ $(\rho_o > R_{obs})$ |
| Direction Reward | $r_{dir} = -\frac{\lvert\alpha_t - \alpha_{t+1}\rvert}{\pi}$ |
| Escaping Reward | $r_{esc} = k_{add} \times (\frac{d_{u,o} - \rho_{amax}}{\rho_{amax}}), d_{u,o} > \rho_{amax}$ |

In Table. 1,$k_{att}$is the attraction gain coefficient, $k_{rep}$is the repulsion gain coefficient, $k_{add}$is the additional gain coefficient, $d_{u,g}$ is the distance between the UAV and the target point, $d_{u,o}$, $d_{s,g}$is the distance between the starting point and the target point, $R_{obs}$ is the radius of the obstacle and threat zone,$\alpha_t$ and $\alpha_{t+1}$represent the UAV heading angles at time *t* and *t+1*, respectively. And $\rho_{amax}$ is the radius of the additional potential field.

The reward obtained by the UAV from the environment is expressed as follows:

$$R = r_{dis} + r_{avo} + r_{dir} + r_{esc} \tag{15}$$

### 4.2. Simulation results and analysis

The proposed algorithm is simulated and analyzed. And algorithm parameters shown in Table.2. The working area of UAV is limited to a two-dimensional space of 100(m)*100(m), and the starting point is [0 0], the target position is [100 100], and step size is 0. 05m.

**Table 2.** RL-IAPF parameters

| Parameters | Value |
|---|---|
| Learning rate | 0. 1 |
| Discount Factor $\gamma$ | 0. 9 |
| ε-greedy probability $\varepsilon$ | 0. 9 |
| Action space $A$ | $\begin{cases} a_1 = \{k_{att} = 0.5, k_{rep} = 0.5, k_{add} = 5\} \\ a_2 = \{k_{att} = 0.7, k_{rep} = 0.5, k_{add} = 5\} \\ a_3 = \{k_{att} = 0.5, k_{rep} = 0.7, k_{add} = 5\} \\ a_4 = \{k_{att} = 0.5, k_{rep} = 0.5, k_{add} = 7\} \\ a_5 = \{k_{att} = 0.5, k_{rep} = 0.5, k_{add} = 3\} \\ a_6 = \{k_{att} = 0.6, k_{rep} = 0.6, k_{add} = 6\} \end{cases}$ |

#### 4.2.1. Target unreachability case

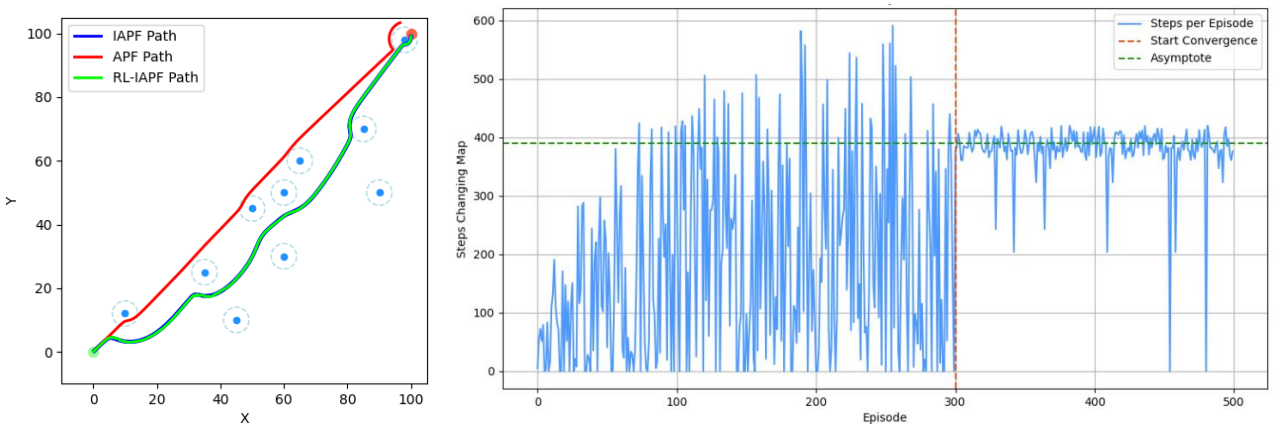
Experiment on the problem of target unreachability.

For the current experiment, the positions of 10 obstacles are shown in Table 3**.**

**Table 3.** Positions of 10 Obstacles

| Index | Position | Index | Position |
|---|---|---|---|
| Obs 1 | [10 12] | Obs 2 | [98 98] |
| Obs 3 | [35 25] | Obs 4 | [50 45] |
| Obs 5 | [60 50] | Obs 6 | [85 70] |
| Obs 7 | [60 30] | Obs 8 | [90 50] |
| Obs 8 | [65 60] | Obs 10 | [45 10] |

With traditional APF algorithm, the UAV starts to move with the attractive force, and the repulsive force will be acted on the UAV when UAV moves into the influence range of obstacle. The target unreachable phenomenon occurs as what mentioned above, and Obs2 caused this trouble.




(a) Path in Target unreachability case (b) Steps changing map

**Figure 10.** Results in Target unreachability case

In order to solve the problem, the IAPF algorithm and RL-IAPF algorithm is applied in the same environment, the result is shown as shown in Fig.10, the UAV can reach the destination. The steps of IAPF are 432 whilst the steps of RL-IAPF are 390.

The training process begins at episode 1 with gradually increasing oscillation amplitudes, indicating the agent's initial learning phase. Around the 200th episode, the oscillation reaches its peak magnitude while maintaining a continuous oscillatory pattern. Between the 200th and 300th episode, the oscillation amplitude begins to decrease, showing a trend toward convergence at step=390, albeit with persistent oscillations. After the 300th episode, the oscillation amplitude significantly reduces, demonstrating clear convergence around step=390. This behavior indicates that the agent has successfully identified a near-optimal path suitable for the given environment.

### 4.2.2. Local minima phenomenon

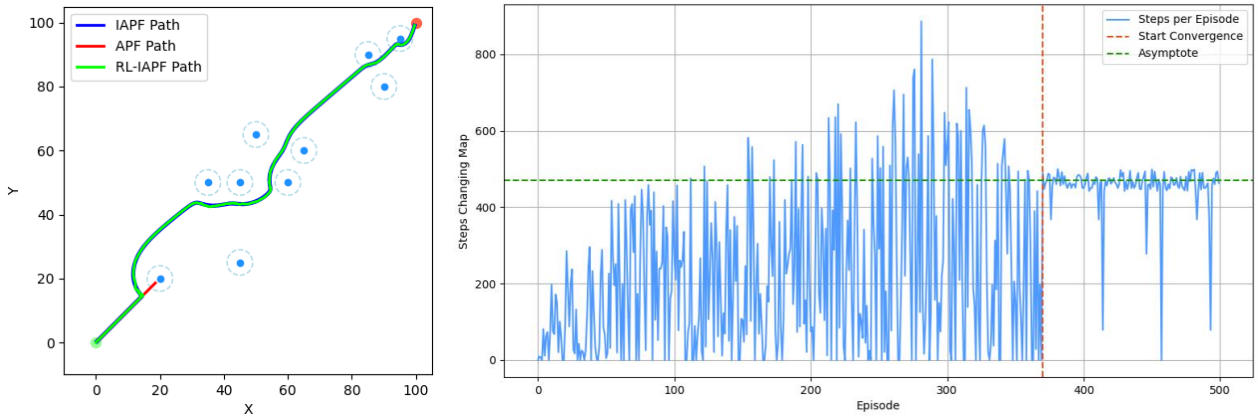
Experiment on the problem of problem of local minima.

The positions of 10 obstacles in current environment are shown in Table. 4. And the parameters are $k_{add}$= 0. 5,$\rho_{amax}$=25, local minima phenomenon occurs when attractive force equals to the repulsive force, and the UAV will be stuck in the local minima position.

**Table 4.** Positions of 10 Obstacles

| Index | Position | Index | Position |
|---|---|---|---|
| Obs 1 | [20 20] | Obs 2 | [45 25] |
| Obs 3 | [35 50] | Obs 4 | [45 50] |
| Obs 5 | [60 50] | Obs 6 | [50 65] |
| Obs 7 | [65 60] | Obs 8 | [90 80] |
| Obs 8 | [85 90] | Obs 10 | [95 95] |

Apply the IAPF algorithm and RL-IAPF algorithm to plan the path, the UAV arrives at our destination as what is shown in Fig.11. The steps of IAPF are 514 whilst the steps of RL-IAPF are 474.




(a) Path in Local minima case (b) Steps changing map

**Figure 11.** Results in Local minima case

The training process initiates at episode 1, with oscillation amplitudes gradually increasing, reflecting the agent's ongoing learning. By approximately the 300th episode, the oscillation reaches its maximum peak while continuing to oscillate. From the 300th to the 370th episode, the oscillation amplitude begins to diminish, showing a tendency to converge toward step=470, though oscillations persist. After the 370th episode, the oscillation amplitude becomes notably smaller, with clear convergence observed around step=470. This indicates that the agent has effectively identified a path well-suited to the environment.

## 5. Conclusion

This study proposes a hybrid path planning algorithm for quadrotor UAVs by integrating Q-learning with an IAPF method, aiming to address the challenges of local optima and target unreachability in dynamic environments. The research begins with an analysis of the limitations of traditional APF methods, such as target unreachability and local minima, and introduces an enhanced APF approach that incorporates relative distance and auxiliary potential fields to mitigate these issues. The integration of MDP theory with the IAPF method allows for the generation of global waypoints, ensuring safer and more efficient navigation. Simulation results demonstrate that the proposed RL-IAPF algorithm significantly improves trajectory accuracy and adaptability, particularly in scenarios involving target unreachability and local minima. The study's findings highlight the potential of combining reinforcement learning with traditional path planning techniques to enhance UAV navigation in complex environments. However, the research is limited to two-dimensional simulations with a limited number of maps, which may not fully capture all real-world scenarios. Additionally, the generated paths exhibit some oscillations, indicating room for further refinement. Future work should focus on smoothing the planned trajectories and extending the algorithm to three-dimensional environments to enhance its practical applicability in real-world UAV operations.

## References

[1] L. C. Shen (2013). Theories and methods of Autonomous Cooperative Control for Multiple UAVs, 1st ed; National Defence Industry Press, Beijing, China. 2003; Volume 1, pp. 1 - 25.

[2] Gong Youmin. Research on trajectory tracking and autonomous landing control of quadrotor UAV [D]. Harbin Institute of Technology, 2017.

[3] Wang Shuwei, Li Jia, Feng Jian, et al. Path optimization of quadrotor aircraft based on BOA-BP neural network [J]. Modern Defense Technology, 1 - 7 [2024 - 07 - 31].

[4] Lu Yanjun, Li Yueru. Trajectory planning of quadrotor aircraft based on improved APF method [J]. Fire and Command Control, 2018, 43 (11): 119 - 122.

[5] Si Bingshan, Dong Zhiming, Sun Maofan. Research on path optimization algorithm of unmanned vehicle group based on RL [J]. computer simulation, 2024, 41 (02): 455 - 461.

[6] Kan Huang, Xin Changfan, Tan Zheqing, et al. Research on collision avoidance path planning of UAV Based on MDP [J]. Computer Measurement and Control, 2024, 32 (06): 292 - 298. DOI:10. 16526/j. cnki. 11-4762/tp. 2024. 06. 042.

[7] Guo Jing, Li Xiang, Xian Yong. Adaptive APF path planning method based on RL [J]. Journal of Ordnance and Equipment Engineering, 1 - 9 [2024 – 07 - 31].

[8] YAO Q, ZHENG Z, QI L, et al. Path planning method with improved APF—a RL perspective [J]. IEEE access, 2020 (8): 135513 - 135523.

[9] O. Khatib, "Real-time obstacle avoidance for manipulators and mobile robots," The international journal of robotics research, vol. 5, no. 1, pp. 90 - 98, 1986.

[10] Chen Kangxiong, Liu Lei. Path planning algorithm for UAV Based on disturbed fluid and td3 [J]. Electro Optic and Control, 2024, 31 (01): 57 - 62.

[11] Dewey, Dan. “RL and the Reward Engineering Principle.” AAAI Spring Symposia (2014).

[12] XIN J, ZHAO H, LIU D, et al. Application of deep RL in mobile rot path planning [C] Chinese Automation Congress (CAC) Jinan, IEEE,2017: 7112 - 7116.

[13] YAO J, LI X, ZHANG Y, et al. Path planning of unmanned helicopter in complex environment based on heuristic deep Q-network [J]. International Journal of Aerospace Engineering, 2022, 2022: 1360956.