sustainability

MDPI

Article

Risk-Aware Travel Path Planning Algorithm Based on Reinforcement Learning during COVID-19

Zhijian Wang [1,*], Jianpeng Yang [1], Qiang Zhang [2] and Li Wang [2]

[1] School of Electrical and Control Engineering, North China University of Technology, Beijing 100144, China
[2] Beijing Aerospace Measurement & Control Technology Co. Ltd., Beijing 100041, China
* Correspondence: wzjian0722@163.com

Abstract: The outbreak of COVID-19 brought great inconvenience to people's daily travel. In order to provide people with a path planning scheme that takes into account both safety and travel distance, a risk aversion path planning model in urban traffic scenarios was established through reinforcement learning. We have designed a state and action space of agents in a "point-to-point" way. Moreover, we have extracted the road network model and impedance matrix through SUMO simulation, and have designed a Restricted Reinforcement Learning-Artificial Potential Field (RRL-APF) algorithm, which can optimize the Q-table initialization operation before the agent learning and the action selection strategy during learning. The greedy coefficient is dynamically adjusted through the improved greedy strategy. Finally, according to different scenarios, our algorithm is verified by the road network model and epidemic historical data in the surrounding areas of Xinfadi, Beijing, China, and comparisons are made with common Q-Learning, the Sarsa algorithm and the artificial potential field-based reinforcement learning (RLAFP) algorithm. The results indicate that our algorithm improves convergence speed by 35% on average and the travel distance is reduced by 4.3% on average, while avoiding risk areas, compared with the other three algorithms.

Keywords: urban traffic; path generation; reinforcement learning; resident travel; COVID-19; restricted search

check for updates

Citation: Wang, Z.; Yang, J.; Zhang, Q.; Wang, L. Risk-Aware Travel Path Planning Algorithm Based on Reinforcement Learning during COVID-19. *Sustainability* **2022**, *14*, 13364. https://doi.org/10.3390/su142013364

Academic Editors: Ciyun Lin, Bowen Gong and Dayong (Jason) Wu

Received: 25 July 2022
Accepted: 29 September 2022
Published: 17 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

At the beginning of 2020, a pneumonia epidemic caused by a new type of coronavirus broke out globally. Due to the latent and asymptomatic nature of the virus, the epidemic has spread widely around the world. The outbreak of COVID-19 has brought great inconvenience to people's daily travel. Before traveling, people must consider the epidemic risk and travel as healthily and safely as possible. During COVID-19, the Chinese government resolutely chose travel control measures, such as traffic bans and travel restrictions, to stop the spread of the new crown pneumonia epidemic. However, it is important to note that necessary travel plays an irreplaceable role in maintaining the normal life of residents. Taking Beijing, China, as an example (Figure 1), during the outbreak of the new crown epidemic, the travel intensity of urban residents dropped sharply. With the implementation of epidemic prevention and control in China, and hospital treatment, the epidemic was effectively controlled, and the travel intensity of residents in the later stage of the epidemic also increased. However, for residents' travel, the implementation of undifferentiated restrictions and prohibitions will not only cause huge losses to residents' daily life, but it will also be detrimental to the commuting needs of front-line staff in the fight against the epidemic. Therefore, the study of the risk-averse travel path planning algorithm during COVID-19 is of great significance to people's travel path planning.

The spread of the new crown epidemic has brought a great impact to people's travel. Before travel, people must consider travel risks and travel as healthily and safely as possible. Necessary travel plays an irreplaceable role in maintaining the normal life of residents.

Therefore, the study of risk-averse travel path planning algorithms during the epidemic is of great significance for residents' travel path planning. In this regard, some domestic scholars have proposed a safe path planning method based on the epidemic situation and related indicators [1,2]; Ma Changxi [3] constructed an emergency customized bus route optimization model under the epidemic, and designed a genetic algorithm to solve it; Tu Qiang et al. [4] proposed a label modification algorithm, which can calculate the path selection results in the face of different risk attitudes; Jia Fuqiang [5] et al. described the risk-averse attitude of travelers based on the cumulative prospect theory and established a multi-objective decision-making model considering risk-averse travel paths; incorporating decision theory into path planning, Subramani et al. [6] propose a step-by-step process for calculating the optimal path for risk; in order to solve the dynamic random shortest path problem, LiPing Fu [7] proposes a heuristic algorithm based on K-shortest path algorithm; A. Khani [8] proposed an iterative labeling algorithm to solve the problem of non-additive reliable paths from a starting point to all ending points; and combining decision theory with basic stochastic-time optimal path planning.

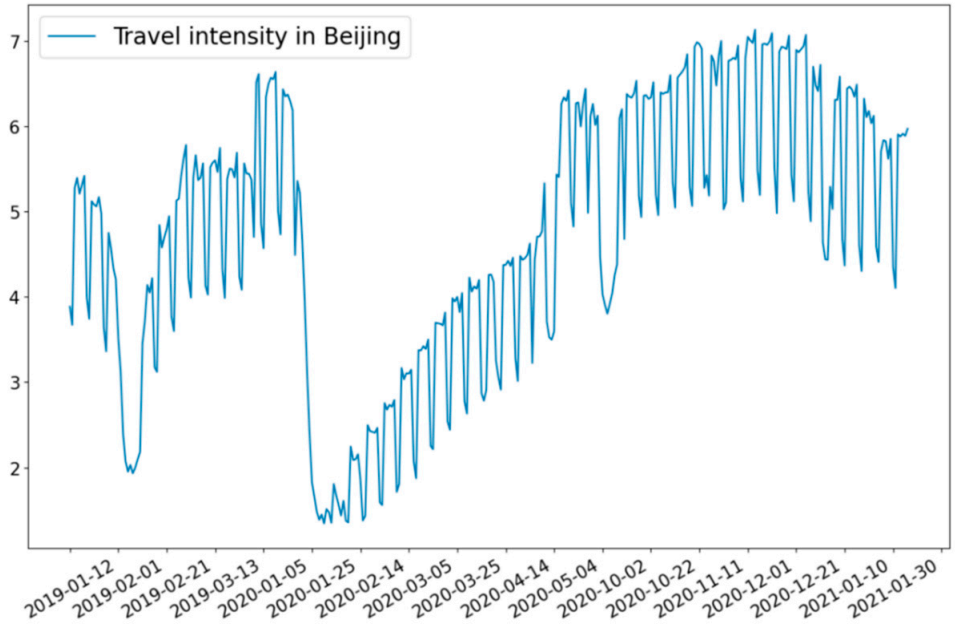


Figure 1. Travel intensity of residents in Beijing, China.

Reinforcement learning is one of the paradigms and methodologies of machine learning, mainly used to describe and solve problems in which agents learn strategies to maximize rewards, or achieve specific goals in the process of interacting with the environment. In recent years, scholars at home and abroad have conducted research on the use of reinforcement learning to solve problems related to path planning. Some Chinese scholars [9–12] have optimized and improved guiding agents to choose actions and reward function settings. Although the results are better than traditional methods, they still need some improvement in the work. A. Wang et al. [13] proposed a cross-regional customized bus route planning algorithm based on improved Q-learning during the epidemic. Sharon Levy [14] of the United States introduced a new path generation method through deep reinforcement learning, which was able to successfully optimize the multi-criteria path selection problem. Cian Ryan et al. [15] proposed a proactive approach to assess autonomous driving risk based on driving behavior and safety-critical events, and process behavior data through telematics for autonomous driving risk-aware path planning. Last year, Bdeir A. [16] proposed an autoregressive strategy of sequentially inserting nodes to solve the problem of constructing a solution to the capacity constrained vehicle routing problem (CVRP). Xu [17] built a MRAM model with multiple relationships based on the AM model [18]

to better obtain the dynamic characteristics of vehicle routing problems. Zhang Ke [19] proposes an encoding decoding framework with attention layer to iteratively generate multiple vehicle tours.

At present, the classification of epidemic risk levels in China is mainly determined by information released by the National Health Commission. Relevant scholars at home and abroad only consider the current epidemic indicators when planning travel routes during COVID-19. Relevant studies [20] have shown that in confined spaces, the new coronavirus can indeed be transmitted through aerosols, and the transmission distance can reach 4.5 m. Since the new coronavirus can be transmitted in the air, we must consider the risks of epidemic-related areas. When using reinforcement learning to solve the path planning problem, some scholars optimize action selection. Although the convergence speed is improved, due to the limitation of the reward function, the desired path will still pass through the risk-related area. Reinforcement learning is a method in which an agent continuously interacts with the environment to adapt to the environment. At present, the application of path planning has shortcomings, such as slow convergence speed and unstable calculation path results. Table 1 compares the limitations of related work.

Table 1. Overview table.

Author	Limitations
Zhou Jibiao, Zhang Haisu et al. [1,2]	Risks in the surrounding areas of COVID-19 were not considered
Ma Chang-xi et al. [3]	Other modes of transportation are not considered
Tu Qiang et al. [4]	Only the car network is considered
Jia Fuqiang et al. [5]	The objective function is relatively simple
Subramani et al. [6]	The risk avoidance path has a high probability of error
Liping Fu et al. [7]	Models and algorithms to be expanded
A. Khani et al. [8]	Risk factors in travel are not considered
Xu Ke, Liu Sijia, Luo Fei et al. [9–11]	The algorithm and model need to be further improved
Wang Keyin et al. [12]	The model is not suitable for urban traffic path planning
Wang A. et al. [13]	Different preferences and requirements of passengers are not considered
Levy S. et al. [14]	The setting of reward function needs further improvement

Therefore, we use the SUMO simulator to build the actual road network model and design a method to extract the road network impedance matrix, which greatly improves the efficiency and accuracy of road network modeling. We have established an enhanced learning path planning model in the context of urban traffic, and designed a search mechanism to avoid risk related areas of the new epidemic. We designed a restrictive search machine initialization Q table to improve the exploration efficiency in the early stage of agent learning. We also introduced the artificial constant function to guide the direction of the agent during learning with different probabilities, and explored the influence of the probability on the convergence result. We used the improved reinforcement learning algorithm to solve the residents' travel path planning problem during the new epidemic situation, and made comparisons with other algorithms. The results show that the algorithm can take both the rapidity and security of travel path into consideration, and it is better in term of convergence speed and stability of convergence.

2. Problem Description and Modeling

2.1. Criteria for Risk Classification of COVID-19 in China

At present, the Chinese Health Commission's classification criteria for low, medium, and high risk areas of the epidemic are as follows [21].

Low risk: no confirmed cases or no new confirmed cases for 14 consecutive days;

Medium risk: There are new confirmed cases within 14 days, the cumulative confirmed cases do not exceed 50, or the cumulative confirmed cases exceed 50, and no cluster epidemic occurs within 14 days;

High risk: The cumulative number of cases exceeds 50, and there is a clustered outbreak within 14 days.

The academician of the Chinese Academy of Engineering [22] believes that the new coronavirus can survive on the surface of objects in a dry environment for up to 48 h, and can survive in the air for about 2 h. It can be seen that the areas near medium and high risk areas also have a certain risk of infection, so people should try to avoid areas near medium and high risks when traveling.

Taking Beijing as an example, "Beijing Health Kit" is a small program that is convenient for individuals to check their own health status related to epidemic prevention [23]. Its query results can be used as a reference for people's health status related to epidemic prevention in scenarios such as the resumption of work and production, and access to public places, etc. [24]. If the health kit turns into a red state of "concentrated observation", it means that people belong to the confirmed, suspected, close, or asymptomatic infected persons, that the Beijing health and health department has grasped since the outbreak of the epidemic. It will require you to perform the duties of centralized isolation and observation. If the health kit turns to yellow "home observation" status, it indicates that people have been checked by the community and have visited medium, high risk areas, and other related places. The community will require people to be isolated at home. However, due to mobile phone signal deviation and other reasons, there is a great possibility of pop-up windows in the areas near the high-risk areas of COVID-19 (hereinafter referred to as relevant areas). Therefore, whether it is from the perspective of personal safety or from the perspective of travel convenience, we should try to avoid related areas as much as possible to ensure travel safety and travel efficiency.

2.2. Objective Function Establishment

This paper studies the problem of travel risk avoidance path planning during the epidemic. It is necessary to consider the length and safety of the path at the same time, that is, in addition to finding a shorter path, it is necessary to avoid risks and related areas as much as possible. Therefore, we set a safe distance for agents to learn, as follows:

$$L_{safe} = \min(\sqrt{R_{ix} - S_{jx}} + \sqrt{R_{iy} - S_{jy}}) \quad (1)$$

In the formula, L_{safe} is the safe distance for path planning, R_{ix} and S_{jx} are the abscissas of risk area R_i and node S_j, R_{iy} and S_{jy} are the ordinates of risk area R_i and node S_j. The agent initializes the reward function through a safe distance, thereby avoiding the risk area. During the calculation, the total length of the path from the specified start point to the end point is defined as follows:

$$L = \|S_{start} - S_1\|_2 + \sum_{k=1}^{n-1} \|S_k - S_{k+1}\|_2 + \left\|S_n - S_{target}\right\|_2 \quad (2)$$

In the formula, S_{start} and S_{target} are the two-dimensional coordinates of the start and end points. S_k is the two-dimensional coordinate of the kth intersection in the path, S_{k+1} is the coordinate of the next intersection of the second intersection k in the path. That is, the total length L of the path is the sum of the lengths of each edge that the agent passes through. The purpose of this paper is to make the required L the shortest under the premise of ensuring safety.

2.3. Markov Decision Process Construction

The Markov Decision Process [25] (MDP) is a mathematical model of sequential decision, which is used to simulate the achievable random policies and rewards of an agent in an environment where the system state has Markov properties. In this paper, the entire reinforcement learning process of the agent can be represented as a Markov decision process consisting of four-tuples $< S, A, P, R >$. The four parts represent the state, action, state transition probability, and reward function, in the urban traffic environment, which will be further explained below.

(1) Status

In the urban traffic environment, this paper assumes that the vehicle cannot turn around or stop on a road, so the state of the vehicle is the set of all nodes in the road network, as shown below:

$$S = \{J_0, J_1, J_2, \ldots, J_n\} \tag{3}$$

In the formula, J_i is the number of intersections in the road network.

(2) Action

In the urban traffic path planning based on reinforcement learning, the action of the agent is the movement process of the vehicle from node i to node j in the road network. Different from the directional fixed-length motion of the robot, the optional actions of the object in the road network are closely related to its state, that is, the direction and step size of the optional action set of the vehicle in each state are different. When a vehicle is at a certain intersection, it can only choose the road connected to it as the action at the next moment, so this paper constructs the selection action set of the vehicle as follows:

$$A = \{J_i : [i \to i_1, i \to i_2, \ldots, i \to i_k]\} \tag{4}$$

In the formula, J_i is the state of the agent, i_k is the adjacent node of state J_i. The specific route of path planning in single reinforcement learning is spliced by each selected action.

(3) State transition probability

Markov's state transition probability $P(s_{t+1}|s_t, A_t = a)$ is action-related, and the agent adopts a certain strategy to select actions in each state. For example, the above formula represents the probability that the agent selects action a and transitions to the next state in the case of states. We assume that when the agent selects an action, it will certainly execute the action and update the status. The agent must move to the next node of the selected action, and it will not happen that the two adjacent moments are in the same state.

(4) Reward function

In reinforcement learning, reward refers to the reward that the environment feeds back to the agent after the agent performs an action. In the scenario of this paper, after the vehicle selects and executes different actions in different states, the rewards of the environment feedback are not the same. In order to keep the agent away from the risk area, the reward function is set as follows:

$$R = \begin{cases} r_1, S_{next} = S_{target} \\ R(r), S_{next} \neq S_{target} \end{cases} \tag{5}$$

$$R(r) = \begin{cases} 1/(\sum_{i=1}^{k}\sum_{j=1}^{n} L_{ij}), d_{next} < d_{now} \\ r_2, d_{next} \geq d_{now} \text{ and } R_{s_next} \leq R_{s_now} \\ \xi r_2, d_{next} \geq d_{now} \text{ and } R_{s_next} > R_{s_now} \end{cases} \tag{6}$$

In the formula, S_{next} is the next state to which the agent moves after selecting an action in the current state and executing it. S_{target} is the end of the trip set by the user. d_{now} and d_{next} are the distance between the current state and the next state of the agent and the target position. L_{ij} is the distance from epidemic risk area i to road j. R_{s_now} and R_{s_next} are the sum of the distance between the current state and the next state and the risk area. r_1 is a positive constant, and r_2 is a negative constant, representing reward and punishment, respectively, ξ is reward transfer coefficient.

3. Algorithm Design

3.1. Impedance Matrix Generation Method

Since we combine the model-free reinforcement learning algorithm and the road network model to improve the algorithm, the road network needs to be physically modeled.

Since the manual modeling method is too cumbersome, this paper proposes a road network impedance matrix based on SUMO simulation. SUMO is a reality-based open-source software for micro-traffic simulators with strong portability and secondary development capabilities. The specific method of generating the road network impedance matrix using SUMO simulation is as follows:

Step 1: Use the SUMO simulator to build a road network consisting of nodes and edges;

Step 2: Extract the number and coordinates of nodes in the road network;

Step 3: Parse the road network configuration file and extract the list combination of [edge, start point, end point] of the road network;

Step 4: Calculate the length of each side according to the coordinates and assign the matrix impedance;

Step 5: Set the non-existing road impedance to infinity.

The generative matrix style is shown in Figure 2.

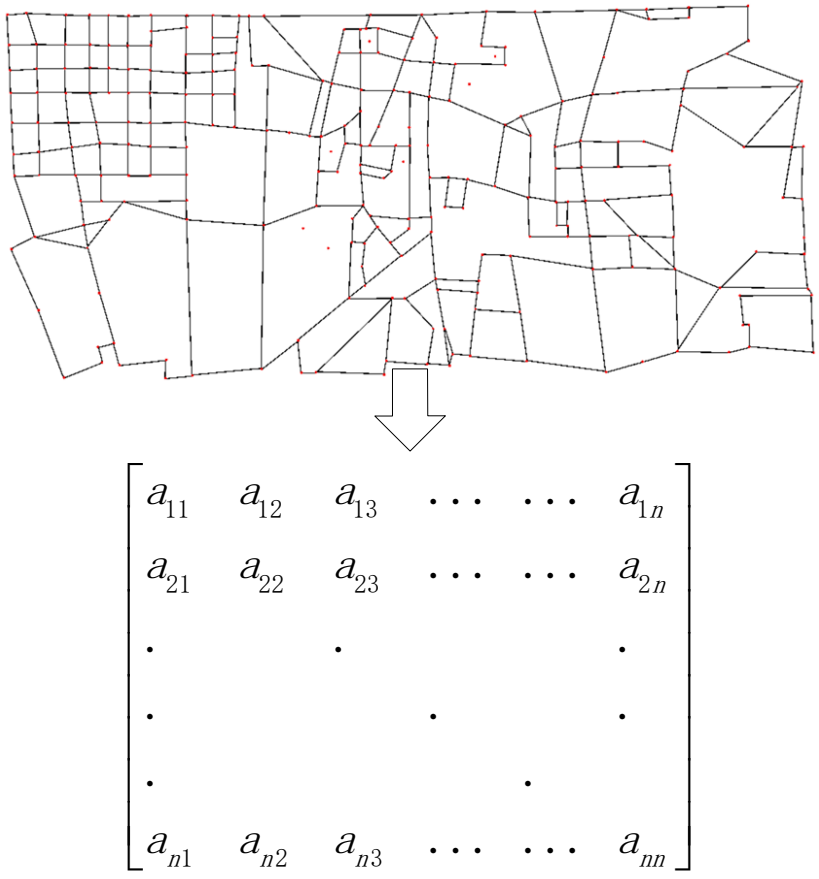


Figure 2. Generate matrix style. (The "n" in the figure represents the number of intersections in the road network).

When we use the SUMO simulator to build the road network model, the OSM open-source map is usually used to define the location of each intersection in the road network and the connection mode between intersections. The location and coordinates of the risk area are recorded in the real epidemic data. We mapped the location of the actual risk area to the SUMO simulator at the same scale as the map, thus defining the risk area.

3.2. Restrictive Search Mechanism for Initializing Q Table

In order to improve the exploration efficiency of the agent in the early stage of reinforcement learning and accelerate the convergence speed of the algorithm, we design a restricted search mechanism to initialize the Q-table operation.

Before calculating the path, it is necessary to set the start and end points. We regard the starting point and the end point as the focus of the ellipse, and the long axis of a certain length forms an ellipse. As shown in Figure 3, with point A as the starting point and point B as the end point, the reinforcement learning path planning is calculated. Then, calculate the ellipse with A and B as the focus according to the focal length, and find all the vectors in the ellipse. Instead of the traditional 0 setting, we increase the q value of the vector corresponding to the action of the vector $\overrightarrow{AB}$ angle less than or equal to 90 degrees to a normal number instead of the traditional setting of 0.

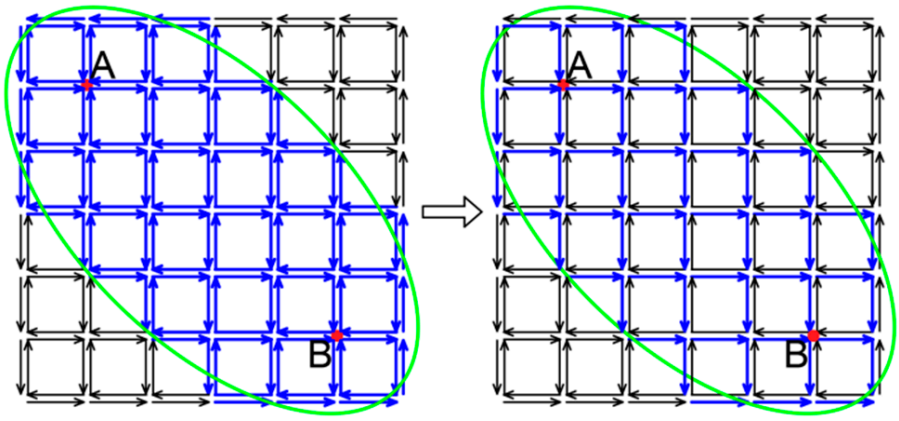


Figure 3. Restricted search method.

3.3. Improved Artificial Potential Field Method to Optimize Action Selection Strategy

In order to speed up the exploration efficiency in the agent learning, we optimize the agent's action selection strategy by improving the Artificial Potential Field (APF) method. In the early stage, the artificial potential field method was proposed by foreign scholar Khatib [26]. Its basic idea is that the target point has a "gravitational force" on the agent, and obstacles, roads or environmental boundaries act as opposite sex to generate a "repulsive force" on the agent. Under the joint action of attraction and repulsion, the agent is guided to move towards the direction of the resultant force. Since the risk factor has been embedded in the reward to the agent when setting the reward function in this paper, only the improved gravitational field function is embedded in the reinforcement learning algorithm. The commonly used gravitational field functions are as follows:

$$U_{att}(q) = \frac{1}{2}\phi\rho^2(q, q_{goal}) \tag{7}$$

In the formula, ϕ is scale factor, $\rho(q, q_{goal})$ represents the distance between the current state of the object and the target. Referring to its definition, the gravitational field function in this scenario is as follows:

$$G(s) = \frac{1}{2}\cdot\vec{V}\cdot(S_{now} - S_{target})^2{}_{(p)} \tag{8}$$

In the formula, $G(s)$ represents the gravitational potential field of the agent under the state S, S_{now} and S_{target} represent the coordinates of the current position of the agent and the coordinates of the target position. $\vec{V}$ is gravitational field unit vector, its direction is that the current state of the agent points to the state of the target node. p is the probability of using the artificial potential field function to guide the decision, and the strength of the gravitational potential field is represented by the second norm of the gravitational potential field.

Figure 4 is a schematic diagram of an agent's decision guidance using the improved artificial potential field method. In the figure, {$o\rightarrow$a, $o\rightarrow$b, $o\rightarrow$c, $o\rightarrow$d} are all the action sets

that the agent can choose in the current state. The direction of the gravitational field is from the current position o to the target position t, and the size is $\|G(s_o)\|_2$. We take the direction of the gravitational field as the baseline and make 45° and 90° rays on both sides, respectively, namely the vectors $\vec{oA}, \vec{oB}, \vec{oC}, \vec{oD}$. We use the improved APF method to guide the agent to choose an action policy is as follows:

$$a = \begin{cases} \{o \to a, o \to b\}, \|G(s_o)\|_2 \geq 1/2\|G(s_{start})\|_2 \\ \{o \to b\}, \|G(s_o)\|_2 < 1/2\|G(s_{start})\|_2 \end{cases} \quad (9)$$

In the formula, a is the action selected by the agent in the current state, $\to$ represents the movement from one point to another. We use the 2-norm of $\|G(s_o)\|_2$ to express the value of the gravitational potential field of the agent at point o. We take half of $\|G(s_{start})\|_2$ as the boundary for action screening, this is to allow the agent to have more choices in the early stage of the calculation path. It can let the agent choose better actions in the later stage, so as to improve the exploration efficiency when approaching the target.

In order to make the artificial potential field method more effective and without loss of generality, we design a scheme to use the artificial potential field to guide the strategy according to the probability. In the case where the agent randomly chooses an action, the artificial potential field method is used with a bootstrap probability p. We set the bootstrap probability $p \in [0.5, 1)$. The larger the p, the higher the probability of using the APF method, and vice versa.

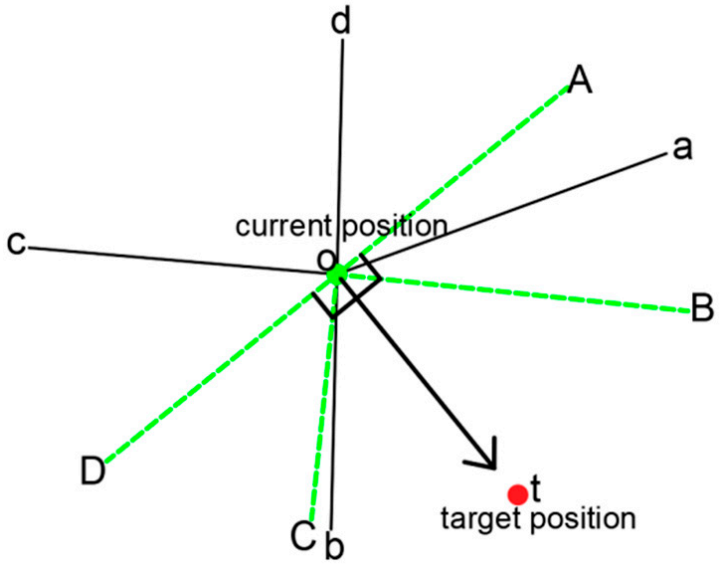


Figure 4. Improved APF method to guide agents to choose actions.

3.4. Dynamic Greedy Strategy

In the whole process of reinforcement learning, how to balance the exploration and the utilization of the exploration efficiency of the agent, and the convergence speed of the whole reinforcement learning, is very important. Traditional methods usually adopt a stage-growing $\varepsilon - greedy$ greedy strategy to balance exploration and utilization, but the stability of convergence is not very good. In order to improve the smoothness and stability of convergence, we design a dynamic $\varepsilon - greedy$ greedy strategy, and defines the greedy rate as follows:

$$\varepsilon = 1 - b^{-x} (b > 1) \quad (10)$$

In the formula, x is the number of times the agent learns, b is coefficient of variation. ε is the probability that an agent selects an action according to the maximum action value. Because it needs to comply with the basic law of agents exploring first, and then use, it is proved by experiments that the value of B should be greater than 1.

When other parameters are the same, the convergence diagram of reward value before and after adjustment is shown in Figure 5. It can be seen that the dynamic greedy strategy converges more stably, than the traditional greedy strategy with staged growth.

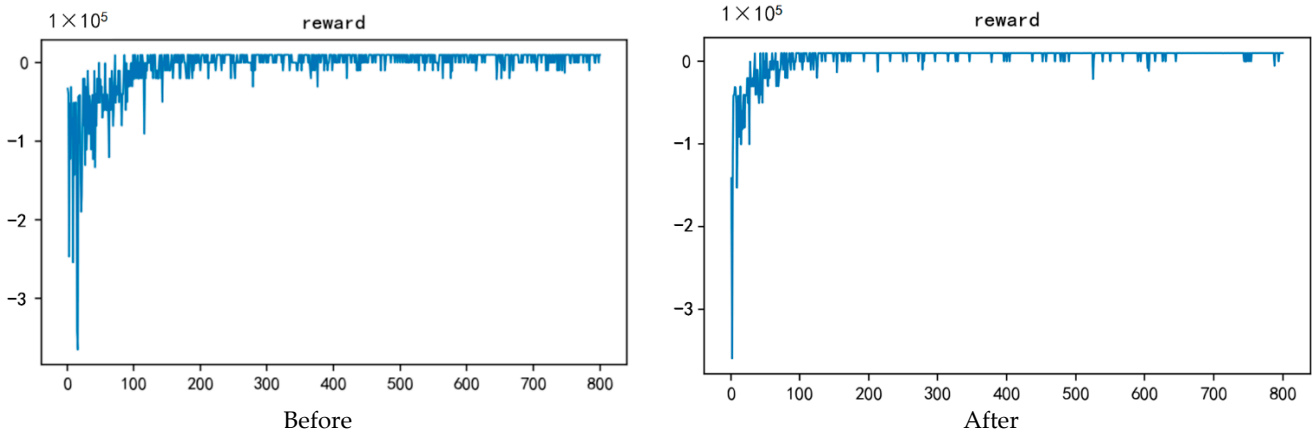


Figure 5. Before and after adjustment.

3.5. Algorithm Process

Based on the above method, we propose a Restricted Reinforcement Learning-Artificial Potential Field (RRL-APF), and its process is shown in Figure 6. Before the agent performs reinforcement learning, it needs to extract relevant information about the road network and epidemic risk areas. The initial reward matrix was established according to the road network model, and sets the reward value corresponding to the nonexistent road as a negative constant with a large absolute value. According to the set safety distance, the reward corresponding to the road in the epidemic risk related area is set to a negative constant with a small absolute value. At the same time, the optional action set of the agent in each node state is constructed. Then, use the restrictive search mechanism to initialize the q-table, and set the Q-values corresponding to all directional edges within the ellipse that are less than 90 degrees from the starting point, to the end point as normal numbers. After setting the start and end point information, the reward value of the last step before establishing the matrix distance from the end point can be enlarged, to improve the convergence speed. Then, reinforcement learning starts, initializing the current state of the agent as the starting point, and filtering the optional action set according to the current state. Due to the high exploration rate and low utilization rate of the agent in the early stage, when the agent randomly selects an action, it uses the RRL-APF method to guide the selection action with P probability, and then immediately executes the action and updates the status, calculates the reward value is calculated and the Q table is updated.

Through experimental comparison, we use the Sarsa algorithm similar to the Q-Learning algorithm [27] to update the Q table. The update mechanism is as follows:

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha[r + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)] \tag{11}$$

In the formula, $Q(s_t, a_t)$ is the current state and the Q value of the corresponding action, $Q(s_{t+1}, a_{t+1})$ is the Q value corresponding to the next state and the actual action taken, α is learning efficiency, γ is Attenuation factor.

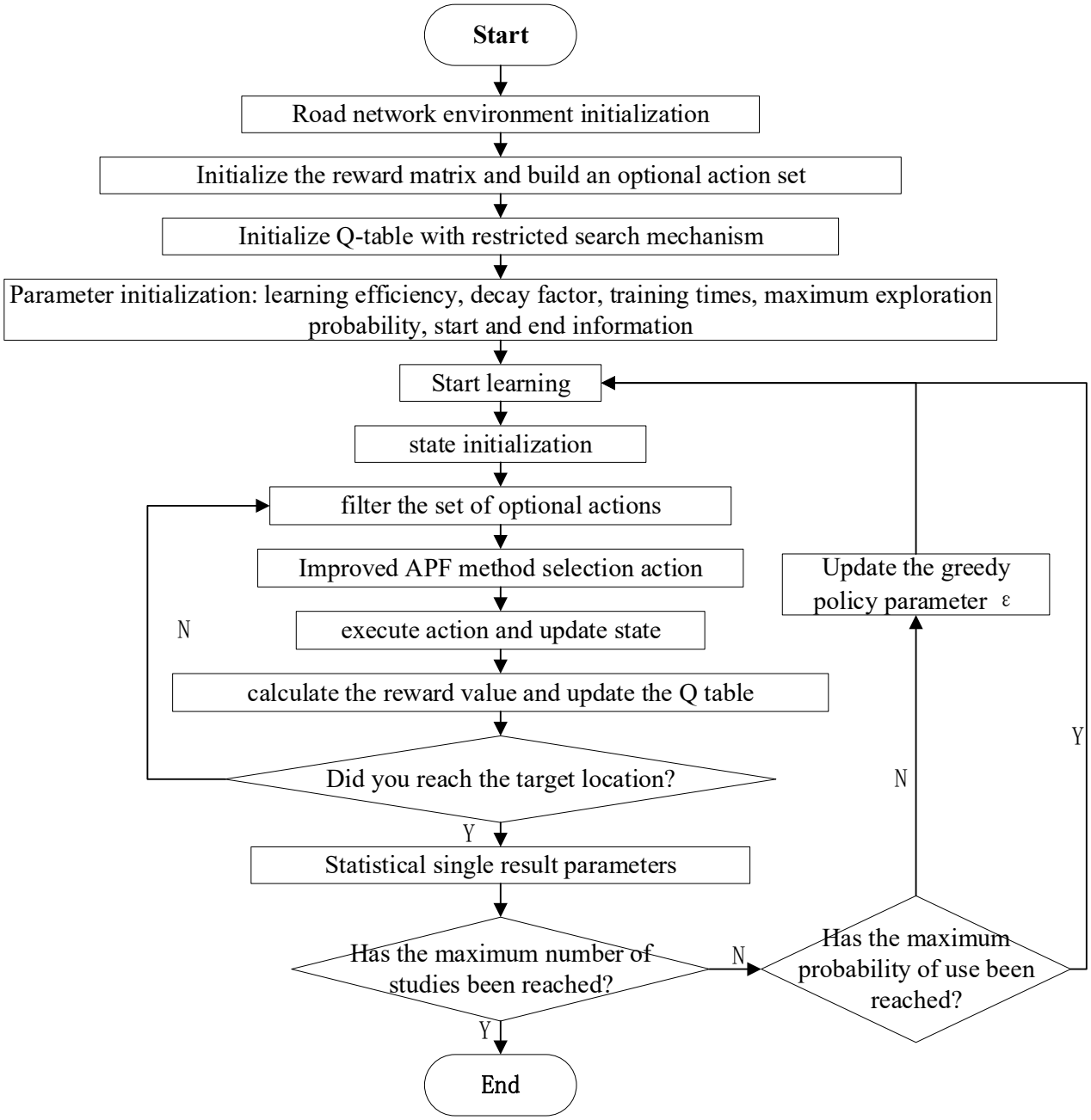


Figure 6. RRL-APF Algorithm flow.

When the agent learns, if the current state is the target position state, the current round of training ends, otherwise, it continues to select actions according to the current state in this cycle. After each training, that is, after the agent reaches the target point from the starting point, it records the relevant results of this training, updates the parameters of the greedy strategy according to the conditions, and then carries out the next training. The detailed flow of the algorithm is shown in the Algorithm 1.

Algorithm 1. Pseudo code of RRL-APF algorithm.

RRL-APF Algorithm

According to the road network model extracted by SUMO, the epidemic risk location information is determined and the safety distance is set.
Initialize the reward matrix, build an optional action set, and initialize the Q table by using the restrictive search mechanism.
Parameter initialization: learning efficiency, attenuation factor, training times, maximum exploration probability.
Initialization starting point: S_{start}, S_{target}.
For episode = 1 to N do:
    Update parameter ε in greedy policy.
    Build the initial state according to s_{start}: $S_{now} = S_{start}$.
    While True:
        Filter the action set as (A1, A2, A3, ...) according to the status.
        If random number < ε:
            Select the action a_i with the largest Q value in a_s.
        Else:
            If random number < p:
                Select action a_i according to the improved APF method.
            Else:
                Random selection action a_i.
        Execute the action and get the next state S_{next}.
        Calculate the distance between S_{now} and S_{next} from the target, observe the location of epidemic risk areas in the environment, and calculate the reward r of the action according to the reward function.
        If $S_{next} = S_{target}$:
            $Q(s_t, a_t) = Q(s_t, a_t) + \alpha[r - Q(s_t, a_t)]$
        Else:
            $Q(s_t, a_t) = Q(s_t, a_t) + \alpha[r + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$
        Update status: $S_{now} = S_{next}$.
        If $S_{now} = S_{target}$:
            Break.
    Count the cumulative rewards, path length, distance from the risk area and learning times of this training
end.
Output the optimal path scheme and algorithm convergence diagram in the whole training process.

4. Algorithm Verification

We take the surrounding area of Xinfadi, Fengtai District, Beijing, as the actual road network, and take the residential areas or places where COVID-19 cases have been active in 2020 as the actual data to verify the RRL-APF algorithm. Then, in order to verify the superiority of this algorithm, we compare the proposed algorithm with the traditional Q-learning algorithm, the Sarsa algorithm, and the reinforcement learning algorithm, based on artificial potential field. We make a comparative analysis with other algorithms in terms of path length and travel risk. The experimental environment is the Windows 10 operating system, AMD Ryzen 7-4800H-2.90GHz processor, 16GB running memory, the programming language is Python, version 3.8.2, program compiler is PyCharm, version 2021_X64.

The data is provided by the Beijing Municipal Health Commission. The actual road network in the study area is shown in Figure 7. The road network has a total of 249 intersections, 770 directional roads, and seven epidemic risk areas and activity venues. The location of road network model and risk area constructed through SUMO simulation is shown in the Figure 8.

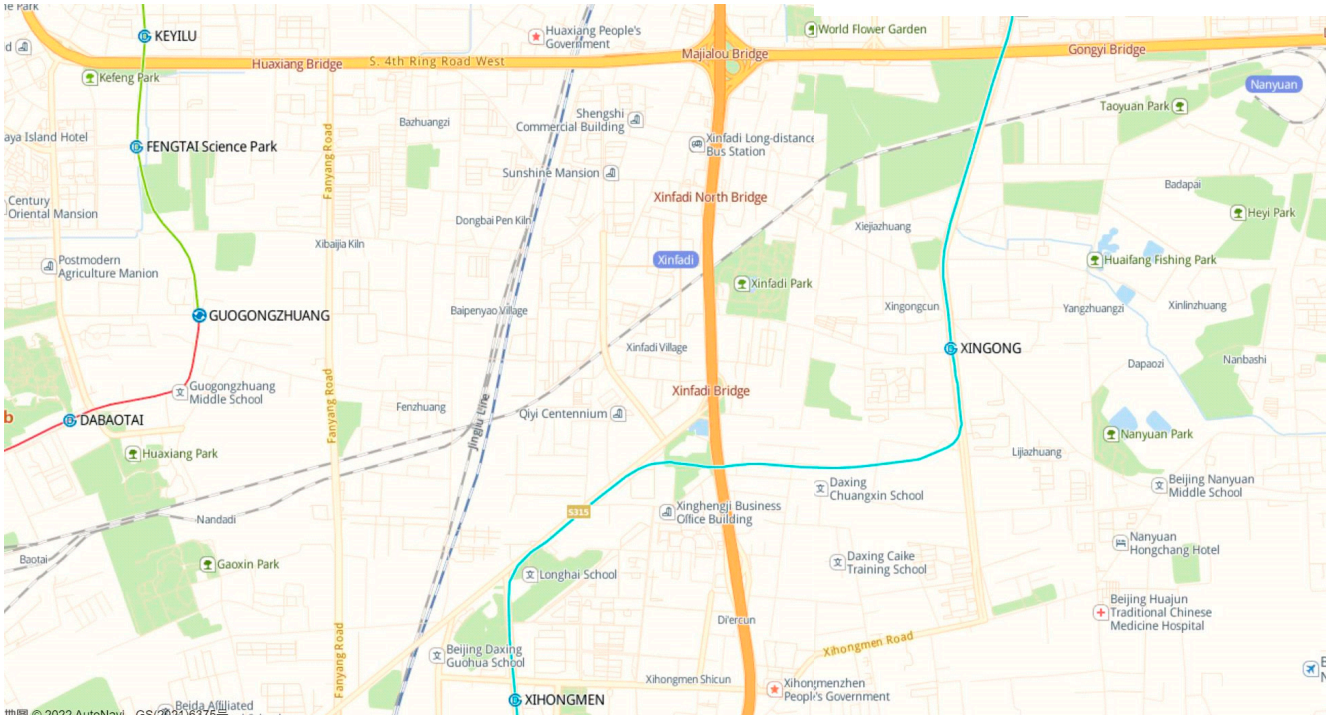


Figure 7. Actual road network scene.

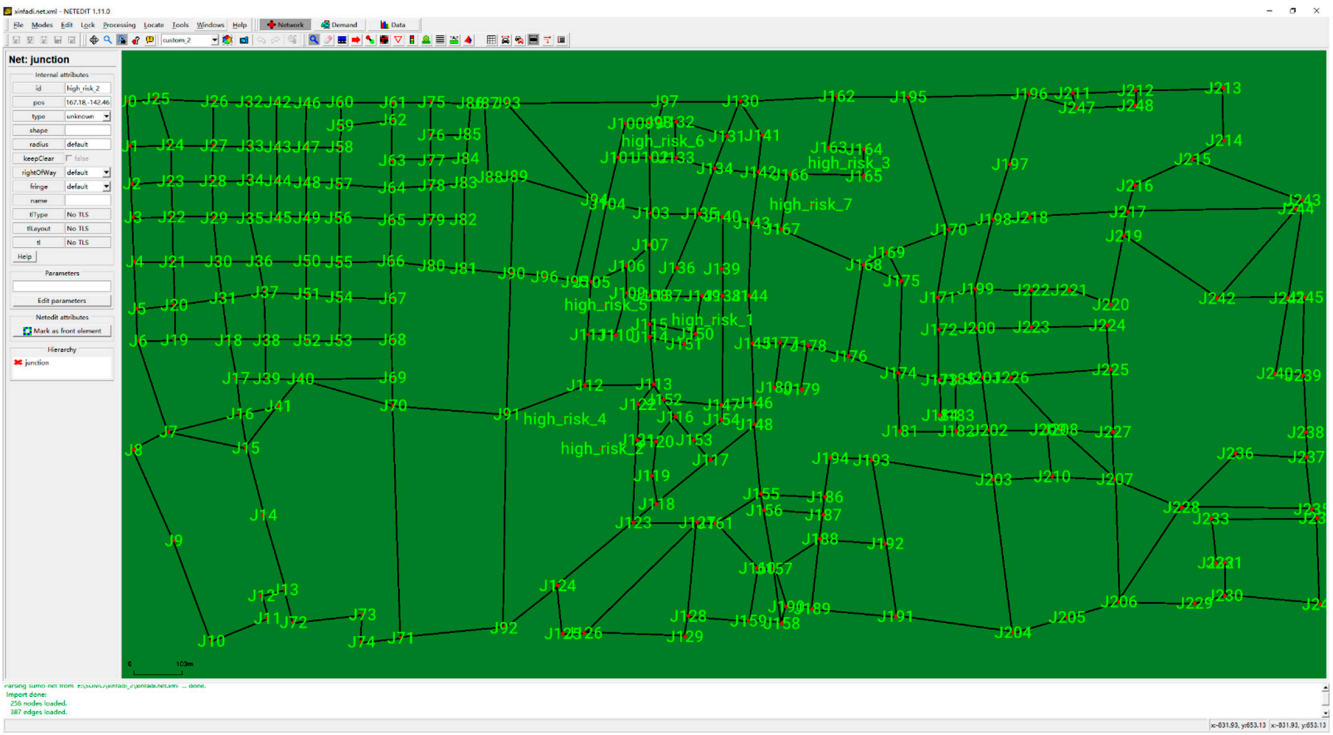


Figure 8. Construction of road network model by SUMO simulator.

The impedance matrix generation algorithm is used for calculation. The results are shown in Table 2:

Table 2. Road network impedance matrix.

	0	1	2	3	4	5	6	7	8	9	10	11	12	…	247	248
0	0	92.8	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	…	inf	inf
1	92.8	0	77.8	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	…	inf	inf
2	inf	77.8	0	69.5	inf	inf	inf	inf	inf	inf	inf	inf	inf	…	inf	inf
3	inf	inf	69.5	0	91.6	inf	inf	inf	inf	inf	inf	inf	inf	…	inf	inf
4	inf	inf	inf	91.6	0	97.4	inf	inf	inf	inf	inf	inf	inf	…	inf	inf
5	inf	inf	inf	inf	97.4	0	66.3	inf	inf	inf	inf	inf	inf	…	inf	inf
6	inf	inf	inf	inf	inf	66.3	0	197.8	inf	inf	inf	152.7	163.8	…	inf	inf
7	inf	inf	inf	inf	inf	inf	197.8	0	81.3	inf	inf	inf	inf	…	inf	inf
8	inf	inf	inf	inf	inf	inf	inf	81.3	0	204.9	inf	inf	inf	…	inf	inf
9	inf	inf	inf	inf	inf	inf	inf	inf	204.9	0	221.2	inf	inf	…	inf	inf
10	inf	inf	inf	inf	inf	inf	inf	inf	inf	221.2	0	48.2	inf	…	inf	inf
11	inf	inf	inf	inf	inf	inf	152.7	inf	inf	inf	48.2	0	50.1	…	inf	inf
12	inf	inf	inf	inf	inf	inf	163.8	inf	inf	inf	inf	50.1	0	…	inf	inf
…	…	…	…	…	…	…	…	…	…	…	…	…	…	0	…	…
247	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	…	0	121.1
248	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	…	121.1	0

In the white area of Table 2, each number represents the distance between two nodes. The diagonal is 0 because there is no distance between the same node. There are a lot of "inf" in the table. Its numerical meaning represents infinity, which means that the corresponding road does not exist. For example, the distance between node 0 and node 2 is inf, which means that node 0 and node 2 are not directly connected. That is to say, the agent cannot choose the action from 0 to 2, because the path from 0 to 2 does not exist. At the same time, under the premise of ensuring accuracy, in order to simplify the calculation, this paper assumes that each road is two-way and the road length in both directions is the same. For example, the distances from 0 to 1 and from 1 to 0 are the same.

The extraction of epidemic risk areas and the activity places of epidemic personnel are shown in Table 3. The coordinates of each risk area are the result of the mapping of the centroid of the COVID-19 risk area on the SUMO simulator.

Table 3. The coordinates of the epidemic site.

Id	Coordinate
risk1	(61.99, 123.22)
risk2	(−167.18, −142.46)
risk3	(343.29, 447.28)
risk4	(−244.26, −81.63)
risk5	(−159.75, 154.4)
risk6	(−41.88, 492.77)
risk7	(264.45, 361.53)

Since there are 249 nodes in the road network, the dimension of the reward matrix is defined as a two-dimensional matrix of 249 × 249. When initializing the reward matrix, we set the reward value corresponding to the nonexistent road and repeating node to a large negative number, and build an optional action set based on it, as shown in Table 4. By assuming three different travel scenarios, the coordinate information of different starting and end points in the SUMO simulator is extracted, as shown in Table 5.

Table 4. Optional action set.

State	Optional Action Set
0	0→1, 0→25
1	1→0, 1→2, 1→24
2	2→1, 2→3, 2→23
3	3→2, 3→24, 3→22
...	...
246	246→230, 246→234
247	247→196, 247→211, 247→248
248	248→212, 248→247

Table 5. Different scene information.

Scenes	Starting Point Coordinates	End Point Coordinates
1	(−1148.9, 574.7)	(1317.7, −464.2)
2	(−600.0, 572.3)	(682.4, −522.3)
3	(−975.7, −540.0)	(1116.6, 601.8)

After the starting and end point information is determined, a Q table is established and initialized according to the restricted search mechanism. The Q-table has the same dimensions as the reward matrix. We take the starting point and the end point as the focal point, and use the focal length of the square root of 2 as the long axis to establish an ellipse, and filter out the vectors in the ellipse. In each scene, we find the vector whose included angle is less than 90 degrees with the vector from the start point to the end point, and set the corresponding q value to 100.

We used the RRL-APF algorithm to conduct simulation experiments for up to 300 times of learning, and the process of an agent from the starting point to the end point through exploration is defined as a complete learning. In order to verify the superiority of the RRL-APF algorithm in terms of convergence speed and other aspects, the RRL-APF algorithm was compared with the Q-Learning [27] algorithm, the Sarsa [28] algorithm and the RLAPF [9,12] algorithm under the same starting and end point, and epidemic risk location information.

Through experimental comparison, we set the learning rate α = 0.01, attenuation factor γ = 0.9, coefficient of variation b = 1.02, the maximum greedy coefficient is 0.9. We use the RRL-APF method to set the probability of selecting an action with p = 0.5 and a safe distance of 200m, the reward transfer coefficient ξ = 0.1.

The specific parameter settings of each algorithm are shown in the Table 6. Among them, b is the coefficient of variation in the dynamic greedy strategy proposed in this paper. The vector filter range is the angle range of the "good" vectors filtered out when initializing the Q table. P is the probability that agents use the improved artificial potential field method to select actions when they randomly select actions.

Table 6. Setting of experimental parameters.

Algorithm	α	γ	Maximum Greedy Coefficient	Safe Distance	ξ	b (Coefficient of Variation)	Vector Filter Range	P
RRL-APF	0.01	0.9	0.9	200	0.1	1.02	90°	0.5
RLAPF	0.01	0.9	0.9	200	0.1	null	null	0.5
Q-Learning	0.01	0.9	0.9	200	0.1	null	null	null
Sarsa	0.01	0.9	0.9	200	0.1	null	null	null

Due to the reinforcement learning algorithm randomly selecting actions under certain conditions, there is a certain randomness in each learning process, and a single execution cannot explain the pros and cons of the algorithm. Therefore, when the parameters such as efficiency and attenuation factor are the same, we run the four algorithms for 10 times, respectively. We obtain the average value of reward value, total length of path and training time in each learning process, as shown in Figures 9–11.

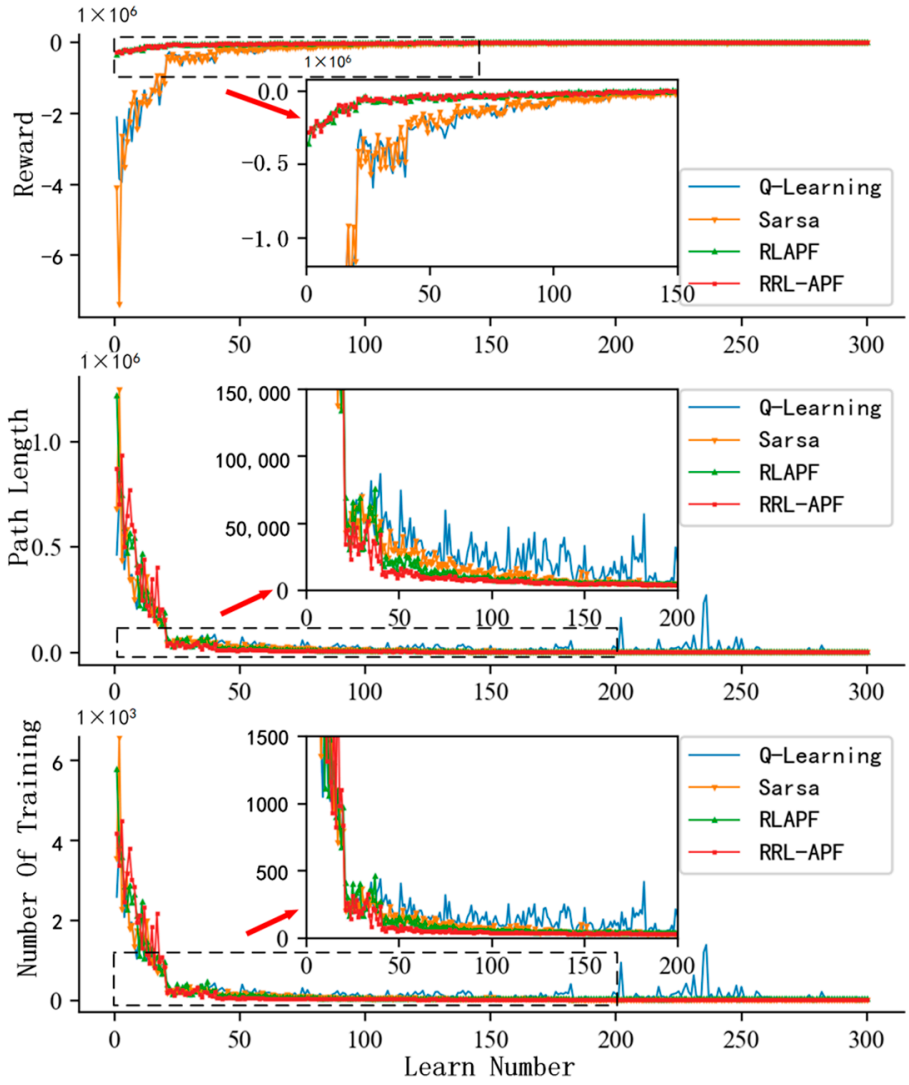


Figure 9. Results of Scenario 1.

Taking Scenario 1 as an example, it can be seen from Figure 9 that the RRL-APF algorithm proposed in this paper is superior to the other three algorithms, in terms of the convergence speed of the three parameters.

(1) In the reward convergence graph, the RRL-APF algorithm and the RLAPF algorithm have the same convergence trend, and are significantly higher than the Q-Learning algorithm, and the Sarsa algorithm in the initial stage. It indicates that compared with the traditional algorithm, the artificial potential field method will be more obvious after adding the method. By reducing the early trial and error behavior, the reward value has increased by approximately 1167%.

(2) In terms of path length, compared with the other three algorithms, the path length of the Q-Learning algorithm increases significantly when the number of learning is between 80 and 170, and the convergence process fluctuates greatly in the later stage of training. This shows that the Q-Learning algorithm is more volatile than the other

three algorithms, so for the sake of simplicity, a comparative analysis of the other three algorithms is carried out.

We take the total path length of 10,596 m as the benchmark. In order to achieve the goal, Sarsa needs to be trained 100 times, and RLAPF needs to be trained 82 times, while the algorithm in this paper is trained only 45 times, and the training times are reduced by 55% and 45%, respectively. This is because the RRL-APF algorithm uses the restricted search method to initialize the Q table before the agent starts training. During the training process, the improved artificial potential field method guides the agent to choose more favorable actions. This can effectively shorten the convergence time of the whole training process.

(3) According to the results of a single training time, it can be seen that the algorithm in this paper has an average reduction of 39% and 26% in the number of training times, compared with the Sarsa algorithm and the RLAPF algorithm. It indicates that the algorithm proposed in this paper has higher computational efficiency and faster convergence speed.

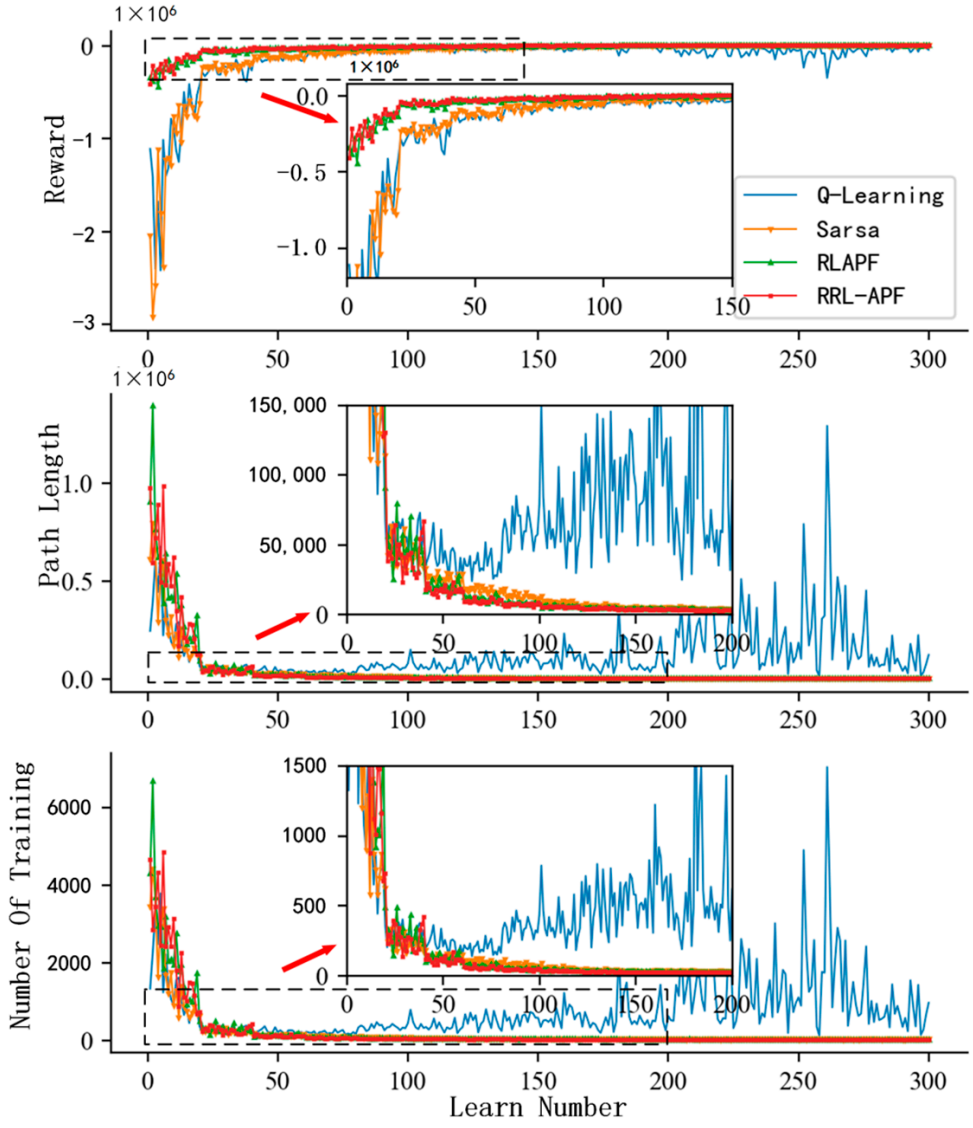


Figure 10. Results of Scenario 2.

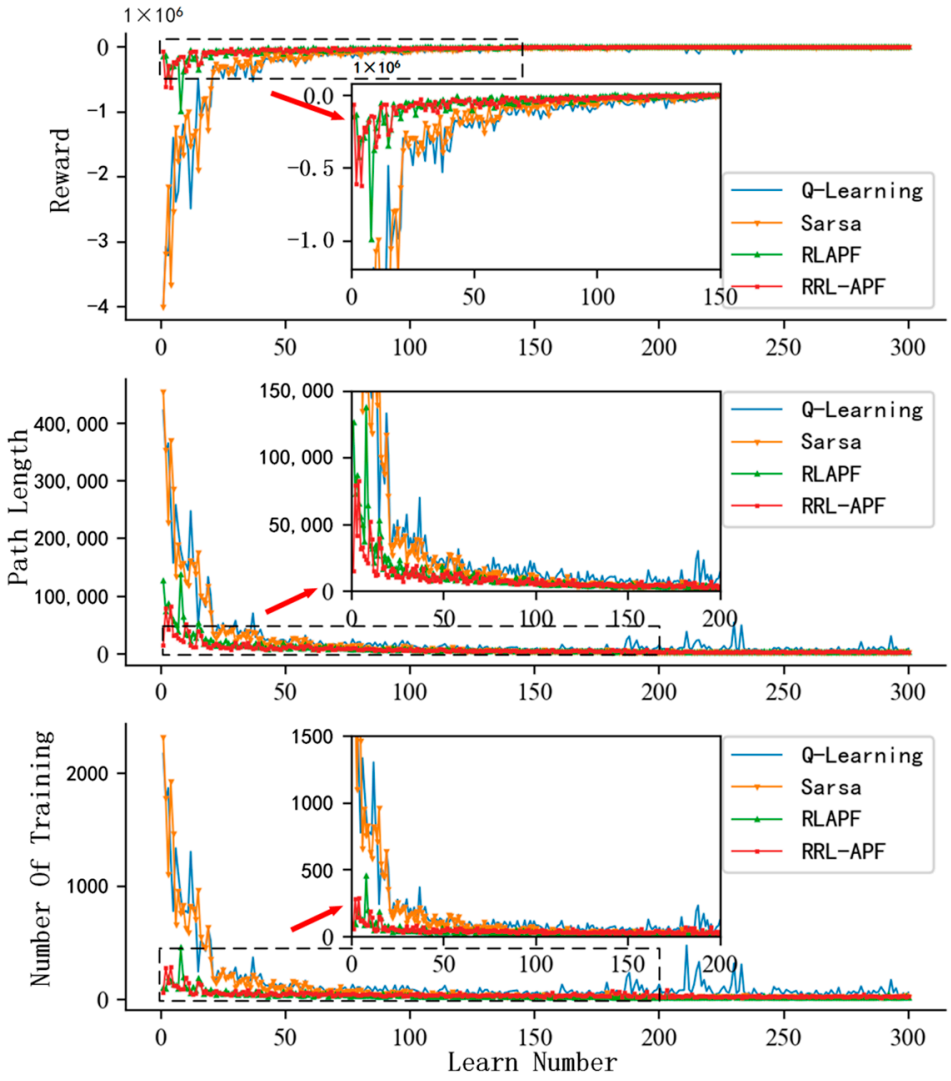


Figure 11. Results of Scenario 3.

Part of the Q table after reinforcement learning under Scenario 1 is shown in Table 7. Each value in the table represents the value of every action of the agent after reaching the maximum number of learning times.

Figures 12–14 shows the calculation results of the path length and the average distance from the risk area of the four algorithms in each scenario. Taking Scenario 1 as an example, the corresponding travel route map is shown in Figure 15. Among them, the "risk area" is announced by the China Health and Health Commission. China implemented closed control measures for high-risk areas in the epidemic, and specifically adopts the measures of "can't get in, can't get out" or "can only go in, can't go out". At the same time, taking the residents of Beijing, China, as an example, if people did not enter but passed there or passed the nearby area, their "Beijing Health Kit" may appear abnormal. During the epidemic, citizens in Beijing need to show the "green code" (a description of personal information under normal conditions) when entering all public places (including shopping malls, supermarkets, etc.). According to the results, it can be seen that the results of the algorithm without considering the epidemic risk will still pass through the epidemic-related area. This may bring the risk of virus infection and abnormal personal health information to people traveling. At the same time, compared with the other three algorithms, the average total path length of the RRL-APF algorithm under the same starting and end point conditions decreased by 3.58%, 1.82%, and 7.51%, respectively. And the average distance

from the risk area increased by 7.77%, 4.33%, and 10.32%. It can be seen that the RRL-APF algorithm proposed in this paper can take both the travel distance and travel risk into account at the same time, and ensure that the travel distance is reduced to a lower level on the basis of sufficient safety.

Table 7. Final action value of Scenario 1.

	0	1	2	3	4	5	6	7	8	9	10	11	12	…	247	248
0	0	81.5	0	0	0	0	0	0	0	0	0	0	0	…	0	0
1	−0.6	0	86.5	0	0	0	0	0	0	0	0	0	0	…	0	0
2	0	−0.5	0	90.7	0	0	0	0	0	0	0	0	0	…	0	0
3	0	0	−0.1	0	94.0	0	0	0	0	0	0	0	0	…	0	0
4	0	0	0	−3	0	96.4	0	0	0	0	0	0	0	…	0	0
5	0	0	0	0	0	0	97.9	0	0	0	0	0	0	…	0	0
6	0	0	0	0	0	0	0	98.7	0	0	0	0	0	…	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	…	0	0
8	0	0	0	0	0	0	0	100	0	100	0	0	0	…	0	0
9	0	0	0	0	0	0	0	0	0	0	100	0	0	…	0	0
10	0	0	0	0	0	0	0	0	0	−0.1	0	100	0	…	0	0
11	0	0	0	0	0	0	0	0	0	−0.1	−0.3	0	0	…	0	0
12	0	0	0	0	0	0	0	0	0	−0.2	0	100	99.9	…	0	0
…	…	…	…	…	…	…	…	…	…	…	…	…	…	0	…	…
247	0	0	0	0	0	0	0	0	−0.3	0	0	0	0	…	0	97.1
248	0	0	0	0	0	0	0	−0.2	0	0	0	0	0	…	−197	0

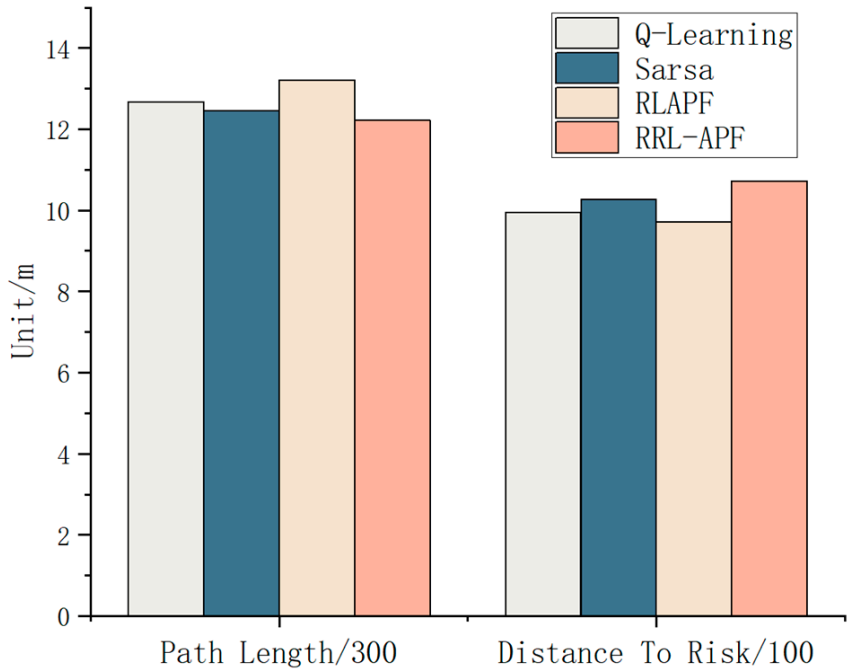


Figure 12. Scenario 1.

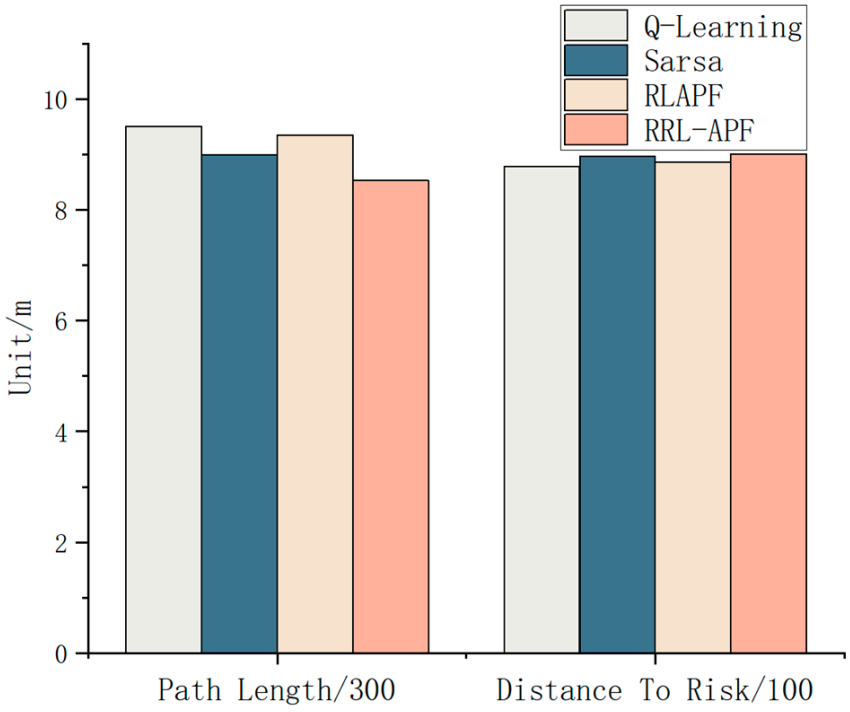


Figure 13. Scenario 2.

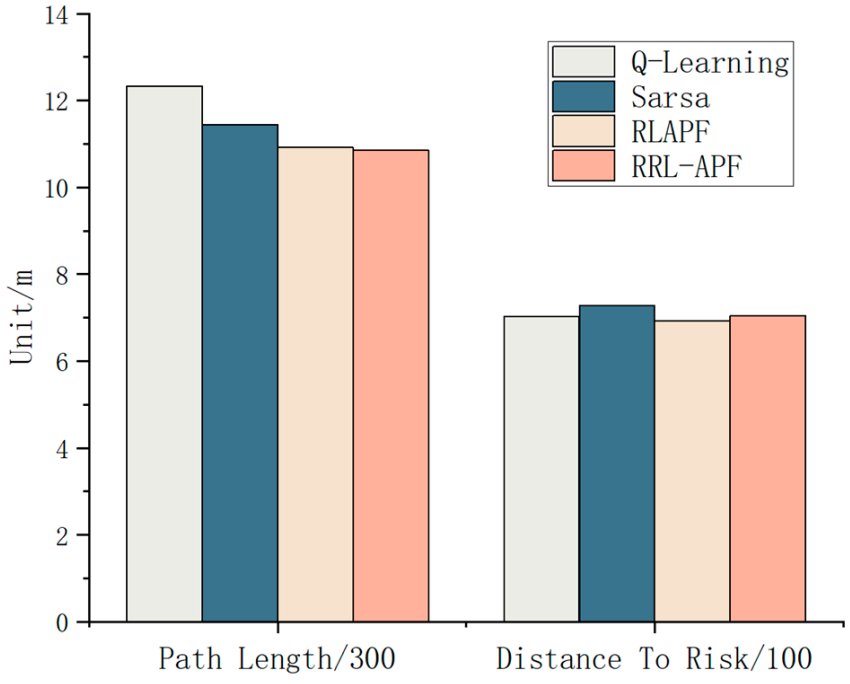


Figure 14. Scenario 3.

Figures 16 and 17 show the path calculation results in Scenario 2 and Scenario 3. It can be seen from the figures that when the distance between the start and end points is short or the available paths are few, our algorithm and the other three algorithms may produce the same path results. For example, in Scenario 3, although the optimal paths obtained by the four algorithms are the same, according to the results in Table 8, the average path length and the average distance from the risk area are still different. Compared with the other three algorithms, the average total path length of the RRL-APF algorithm under the same

starting and ending point conditions decreased by 7.93%, 6.52%, and 7.33%, respectively, and the average distance from the risk area increased by 19.9%, 11.8%, and 9.62%. However, compared with the algorithm that does not consider the epidemic risk and the algorithm that avoids the epidemic risk area, the results of the path calculation are still very different. It shows that our algorithm can ensure that travelers avoid epidemic risks as much as possible during travel, and maximize their own safety.

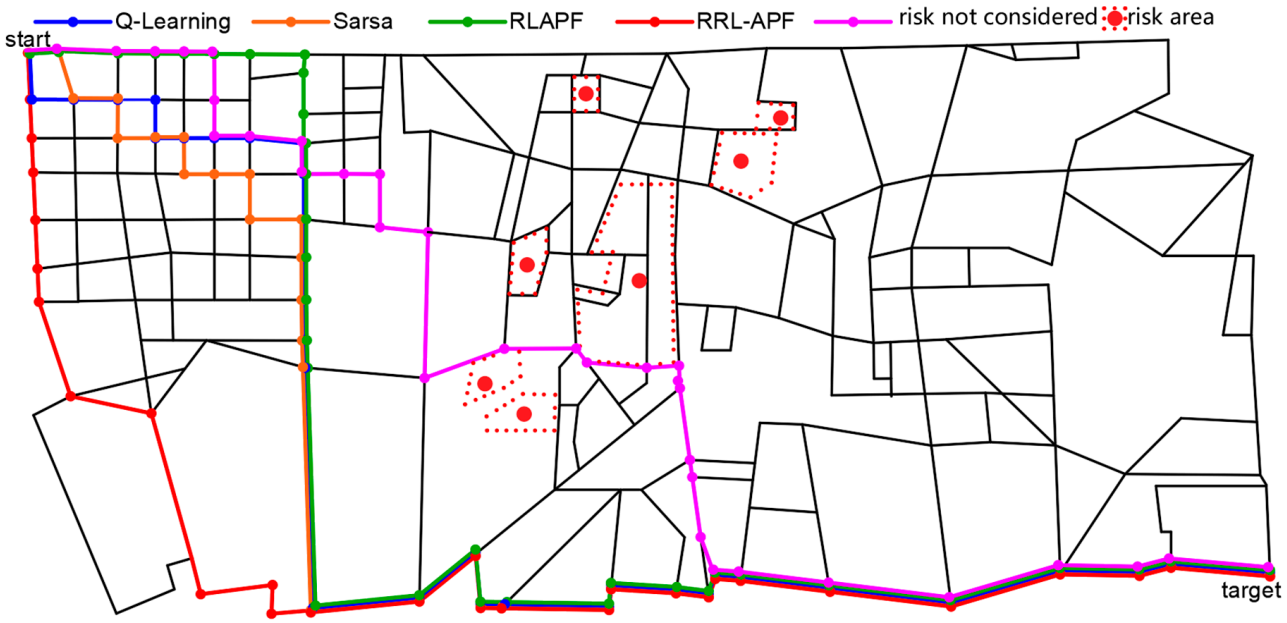


Figure 15. Comparison of travel routes of different algorithms for Scenario 1.

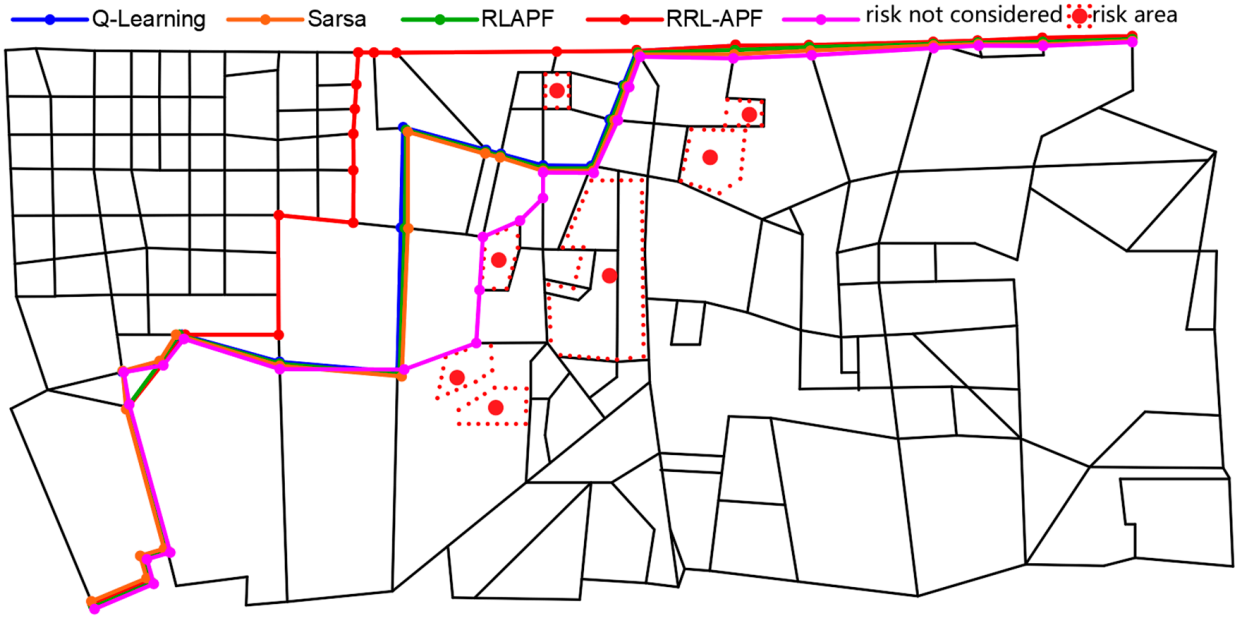


Figure 16. Comparison of travel routes of different algorithms for Scenario 2.

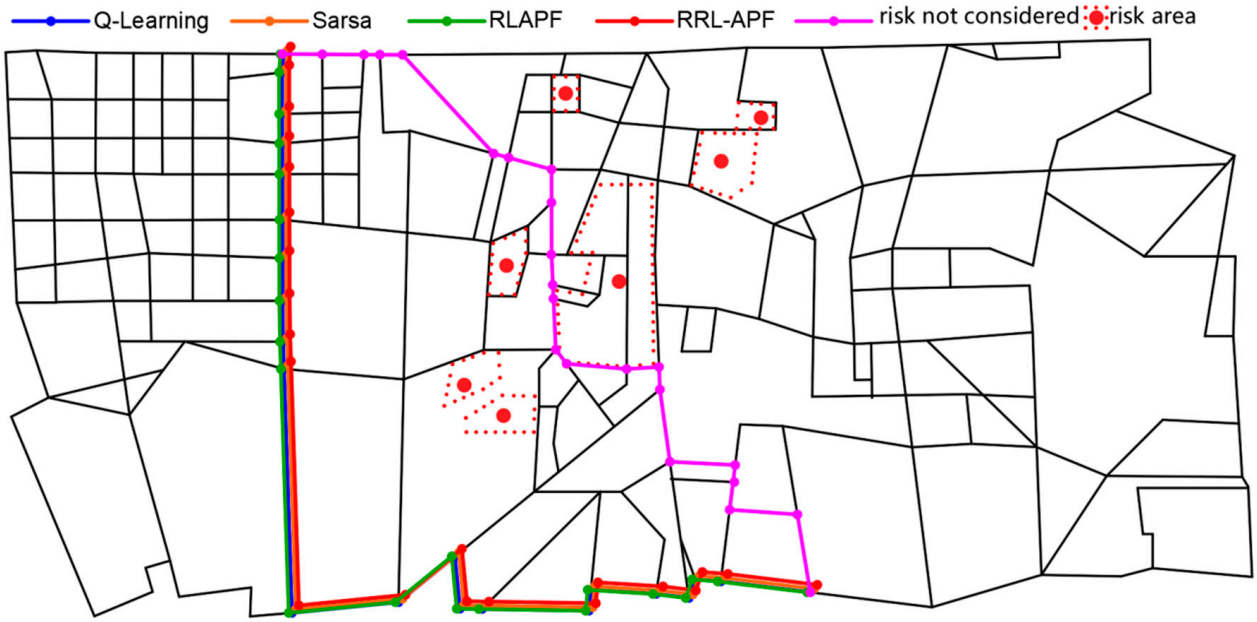


Figure 17. Comparison of travel routes of different algorithms for Scenario 3.

Table 8. Average calculation results in different scenarios.

Scenario	Result	Q-Learning	Sarsa	RLAPF	RRL-APF
1	PL	2851.82	2696.27	2804.31	2559.19
	D2R	877.48	896.35	885.28	900.89
2	PL	3698.67	3433.46	3275.21	3255.83
	D2R	702.59	727.77	692.94	703.79
3	PL	2398.72	2362.45	2383.2	2208.53
	D2R	473.1	521.16	523.83	591.08

The RRL-APF method designed in this paper mainly includes the restricted search mechanism and the improved artificial potential field method. In order to verify the effectiveness of the two methods, we set up ablation experiments for experimental verification. Figure 18 shows the ablation experiments results in Scenario 1. The RRL-APF algorithm in the figure is designed in this paper. RR represents the experimental results after removing the artificial potential field method, and APF represents the results after removing the Q table initialized by the restrictive search mechanism. From the convergence diagram of the three results in the figure, it can be seen that the convergence speed of the artificial potential field method is significantly faster than that of the method without artificial potential field. In terms of reward value, path length and single training times, the convergence speed is reduced by 71%, 52%, and 60%, after using the artificial potential field method. This proves the effectiveness of the artificial potential field method. In terms of reward value and training times, the convergence trends of APF and RRL-APF in the figure are roughly the same. However, in terms of path length, it can be seen from the figure that the green lines are almost entirely below the orange lines. After calculation, the convergence speed of using the restrictive search mechanism is approximately 38% lower than that of not using it, which proves the effectiveness of the restrictive search mechanism.

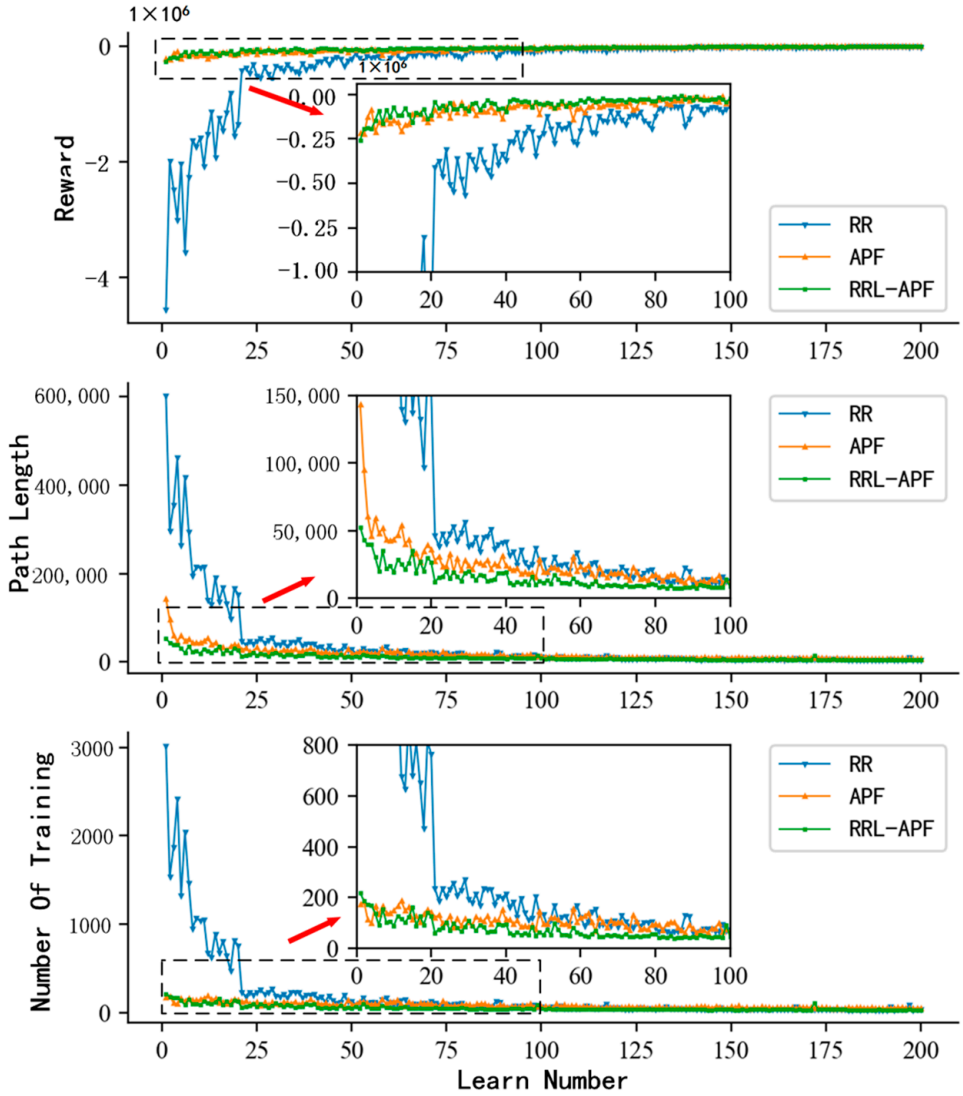


Figure 18. Ablation experiments result.

Under the premise of randomly selecting actions, in order to explore the influence of the probability of the agent using the artificial potential field method to select actions on the convergence results, we added a group of parameter control experiments under different probabilities. To make the experimental results clearer, and intuitive, we fit the convergence curve. The fitting curve of the experimental results is shown in Figure 19. From the convergence graph of reward value, it can be seen that the reward value at the early stage with a probability of 0.9 is significantly lower than that with the probability of 0.5 and 0.7. However, from the middle to late stage, especially after the agent has learned 25 times, the reward value is almost ($p = 0.9$) > ($p = 0.7$) > ($p = 0.5$). This shows that in the middle and late stages of agent learning, the reward value increases with the probability of using APF. According to the path length and the number of single trainings, the result of $p = 0.9$ in the early stage of agent learning is larger than that in the other two cases, but the convergence trend of the three in the middle and late stages of agent learning is roughly the same.

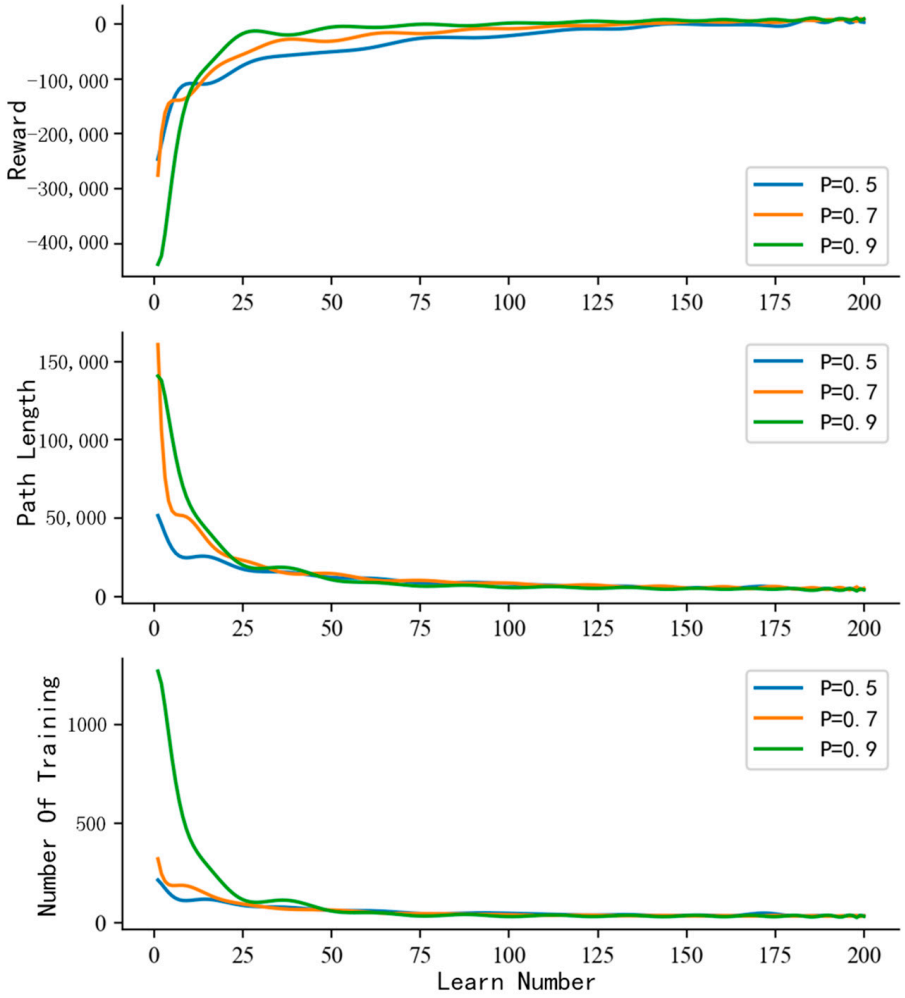


Figure 19. Convergence curve fitting results under different selection probabilities.

5. Conclusions

The main conclusions of this paper are as follows:

(1) Based on the reinforcement learning model, we established a path planning model to avoid epidemic risk in urban traffic scenarios, and designed a constrained search reinforcement learning algorithm based on improved artificial potential field. In specific scenarios, the convergence speed of our algorithm can be improved by 20–55%. At the same time, it can be seen that the intervention of artificial potential field method can greatly reduce the trial-and-error steps of the traditional reinforcement learning algorithm in the initial stage. The method of initializing the Q-table by using ellipses for restrictive search can also enable the agent to have a general understanding of the environment before starting to learn.

(2) Taking the surrounding area of Xinfadi, Fengtai District, Beijing, China, as an example, we verified the algorithm with historical epidemic data. Compared with the existing algorithm research, the RRL-APF algorithm can provide a more secure short-distance path scheme. The total path length in the scene is reduced by 4.3% on average, and the distance from the risk area is increased by 7.5% on average.

(3) The model and algorithm we built can be used for residents' safe travel path planning considering epidemic and other risks. In the context of intelligent transportation system, our proposed method can provide effective path planning suggestions for advanced transportation information service system. More importantly, it can provide a theoretical basis for travel management after major public health events and sudden disasters in the future.

(4) In real life, people may travel through multiple modes of transportation. Therefore, in the next stage, it is necessary to consider combining multimodal transport and dynamic road network information. This may involve the connection between different modes of transportation. We need to further study the coordination between public transport, such as buses and subways, with fixed time and on-line car hailing and taxis with variable time. At the same time, we should also consider the traffic capacity of the road network, the traffic efficiency of road intersections, and the travel habits and preferences of residents. In the near future, developing a travel route planning application with intuitive user interface is the research content of the next stage. In addition, it is also very important to identify the tracks of people infected with COVID-19. Some algorithms and network models (such as Balance Learning, PFGFE-Net, G-CNN) proposed by some scholars [29–33] for ship detection have also inspired us greatly. In the future, we can use relevant target detection methods to identify and track the movements of COVID-19 infected people, determine the location of COVID-19 risk related areas in the shortest possible time, and minimize the risk of COVID-19 transmission.

Author Contributions: Data curation, Q.Z. and L.W.; investigation, Q.Z. and L.W.; methodology, Z.W. and J.Y.; project administration, L.W.; resources, Z.W.; software, J.Y.; validation, J.Y.; visualization, J.Y.; writing—original draft, J.Y.; writing—review and editing, Z.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: https://data.beijing.gov.cn/yqdt/index.htm (accessed on 10 December 2021).

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

References

1. Zhou, J.; Ma, C.; Dong, S.; Zhang, M. Unconventional anti-epidemic strategies of urban public transportation under the new crown pneumonia epidemic: Taking Ningbo as an example. *J. China Highw.* **2020**, *33*, 1–10.
2. Zhang, H.; Qi, C.; Zhang, S.; Xu, F.; Wang, L.; Ji, C.; Yu, J.; Deng, K.; Li, X.; Li, T.; et al. A Walking Path Planning Method and System Based on Community Epidemic Situation. CN202010315880.7. 21 April 2020. Available online: http://so.ip.top/patentDetail?q=61445b51bfc1a10bc0a6968d (accessed on 10 December 2021).
3. Ma, C.-X.; Wang, C.; Hao, W.; Liu, J.; Zhang, Z.-L. Emergency customized bus route optimization under public health emergencies. *Chin. J. Transp. Eng.* **2020**, *20*, 89–99.
4. Tu, Q.; Cheng, L.; Lin, F.; Sun, C. Optimal path search considering traveler's risk attitude. *J. Jilin Univ.* **2019**, *49*, 720–726.
5. Jia, F.; Yang, X.; Dai, C.; Ma, C. Analysis of Travel Route Selection Behavior Considering Risk Avoidance. *J. Shenzhen Univ. (Sci. Technol. Ed.)* **2022**, *39*, 111–236.
6. Subramani, D.N.; Lermusiaux, P.F. Risk-optimal path planning in stochastic dynamic environments. *Comput. Methods Appl. Mech. Eng.* **2019**, *353*, 391–415. [CrossRef]
7. Fu, L.; Rilett, L.R. Expected shortest paths in dynamic and stochastic traffic networks. *Transp. Res. Part B Methodol.* **1998**, *32*, 499–516. [CrossRef]
8. Khani, A.; Boyles, S.D. An exact algorithm for the mean-standard deviation shortest path problem. *Transp. Res. Part B: Methodol.* **2015**, *81*, 252–266. [CrossRef]
9. Xu, K. Research and Implementation of Intelligent Travel Route Planning Algorithm Based on Reinforcement Learning. Master's Thesis, Xidian University, Xi'an, China, 2020.
10. Liu, S.; Tong, X.R. Urban Traffic Path Planning Based on Reinforcement Learning. *Comput. Appl.* **2021**, *41*, 185–190.
11. Luo, F.; Bai, M. Decision Optimization of Traffic Scenario Problems Based on Reinforcement Learning. *Comput. Appl.* **2022**, *42*, 2361.
12. Wang, k.; Shi, Z.; Yang, Z.; Yang, Y.; Wang, S. Application of improved reinforcement learning algorithm to mobile robot path planning. *Comput. Eng. Appl.* **2021**, *57*, 270–274.
13. Wang, A.; Guan, H.; Wang, P.; Peng, L.; Xue, Y. Cross-regional customized bus route planning considering staggered commuting during the COVID-19. *IEEE Access* **2021**, *9*, 20208–20222. [CrossRef]
14. Levy, S.; Xiong, W.; Belding, E.; Wang, W. SafeRoute: Learning to Navigate Streets Safely in an Urban Environment. *arXiv* **2018**, arXiv:1811.01147. [CrossRef]
15. Ryan, C.; Murphy, F.; Mullins, M. Spatial risk modelling of behavioural hotspots: Risk-aware path planning for autonomous vehicles. *Transp. Res. Part A* **2020**, *134*, 152–163. [CrossRef]

16. Bdeir, A.; Boeder, S.; Dernedde, T.; Tkachuk, K.; Falkner, J.K.; Schmidt-Thieme, L. RP-DQN: An application of Q-Learning to Vehicle Routing Problems. In *KI 2021: Advances in Artificial Intelligenc, Proceedings of the 44th German Conference on AI, Virtual Event, 27 September–1 October 2021*; Springer: Cham, Switzerland, 2021.
17. Xu, Y.; Fang, M.; Chen, L.; Xu, G.; Du, Y.; Zhang, C. Reinforcement Learning With Multiple Relational Attention for Solving Vehicle Routing Problems. *IEEE Trans. Cybern.* **2022**, *52*, 11107–11120. [CrossRef]
18. Kool, W.; van Hoof, H.; Welling, M. Attention, learn to solve routing problems! In Proceedings of the 2019 International Conference on Learning, New Orleans, LA, USA, 6–9 May 2019.
19. Zhang, K.; He, F.; Zhang, Z.; Lin, X.; Li, M. Multi-vehicle routing problems with soft time windows: A multi-agent reinforcement learning approach. *Transp. Res. Part C* **2020**, *121*, 102861. [CrossRef]
20. Available online: https://baijiahao.baidu.com/s?id=1660610997152195541&wfr=spider&for=pc (accessed on 12 May 2021).
21. Available online: https://www.kepuchina.cn/article/articleinfo?business_type=100&classify=2&ar_id=88958 (accessed on 12 May 2021).
22. Available online: https://baijiahao.baidu.com/s?id=1657640223323611905&wfr=spider&for=pc (accessed on 12 May 2021).
23. Available online: https://baike.baidu.com/item/%E5%8C%97%E4%BA%AC%E5%81%A5%E5%BA%B7%E5%AE%9D/24487219?fr=aladdin#reference-[1]-24927247-wrap (accessed on 10 December 2021).
24. Available online: https://baike.baidu.com/reference/24487219/9760zcg80zh_eCwyDXeO_T2OoL4dx3oX5NWrkWnQxtB0Z7-Pc_7MCIDmwRLziCfWsa-oPgyjChk2gsNbOiy2M9stUWl_BE0xFGn6TA (accessed on 10 December 2021).
25. Sutton, R.; Barto, A. *Reinforcement Learning:An Introduction*; MIT Press: Cambridge, MA, USA, 1998.
26. Khatib, O. Real-time obstacle avoidance for manipulators and mobile robots. *IEEE Int. Conf. Robot. Autom.* **1985**, *2*, 500–505.
27. Watkins, C.J.C.H.; Dayan, P. Q-learning. *Mach. Learn.* **1992**, *8*, 279–292. [CrossRef]
28. Sutton, R.S. Generalization in reinforcement learning: Successful examples using sparse coarse coding. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 1996; pp. 1038–1044.
29. Zhang, T.; Zhang, X.; Shi, J.; Wei, S. Depthwise Separable Convolution Neural Network for High-Speed SAR Ship Detection. *Remote Sens.* **2019**, *11*, 2483. [CrossRef]
30. Zhang, T.; Zhang, X.; Liu, C.; Shi, J.; Wei, S.; Ahmad, I.; Zhan, X.; Zhou, Y.; Pan, D.; Li, J.; et al. Balance learning for ship detection from synthetic aperture radar remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *182*, 190–207. [CrossRef]
31. Zhang, T.; Zhang, X. A Polarization Fusion Network with Geometric Feature Embedding for SAR Ship Classification. *Pattern Recognit.* **2022**, *123*, 108365. [CrossRef]
32. Zhang, T.; Zhang, X. High-Speed Ship Detection in SAR Images Based on a Grid Convolutional Neural Network. *Remote Sens.* **2019**, *11*, 1206. [CrossRef]
33. Zhang, T.; Zhang, X.; Zhu, P.; Tang, X.; Li, C.; Jiao, L.; Zhou, H. Semantic Attention and Scale Complementary Network for Instance Segmentation in Remote Sensing Images. *IEEE Trans. Cybern.* **2021**, *52*, 10999–11013. [CrossRef] [PubMed]