sensors

MDPI

Article

Reinforcement Learning-Based Approach for Minimizing Energy Loss of Driving Platoon Decisions †

Zhiru Gu [1], Zhongwei Liu [1], Qi Wang [2], Qiyun Mao [1], Zhikang Shuai [2] and Ziji Ma [2,*]

[1] College of Railway Transportation, Hunan University of Technology, Zhuzhou 412007, China; guzhiru@hut.edu.cn (Z.G.); 16401700232@stu.hut.edu.cn (Z.L.); 17419002071@stu.hut.edu.cn (Q.M.)
[2] College of Electrical and Information Engineering, Hunan University, Changsha 410082, China; wangqii@hnu.edu.cn (Q.W.); szk@hnu.edu.cn (Z.S.)
* Correspondence: zijima@hnu.edu.cn
† This paper is an extension version of the conference paper: Zhiru Gu, Zhongwei Liu, Ziji Ma, Feilong Wang, Xiaogang Zhang; Minimizing Energy Loss Decisions for Green Driving Platoon; 2022 32nd International Telecommunication Networks and Applications Conference (ITNAC); Wellington, New Zealand; 30 November–2 December 2022.

Abstract: Reinforcement learning (RL) methods for energy saving and greening have recently appeared in the field of autonomous driving. In inter-vehicle communication (IVC), a feasible and increasingly popular research direction of RL is to obtain the optimal action decision of agents in a special environment. This paper presents the application of reinforcement learning in the vehicle communication simulation framework (Veins). In this research, we explore the application of reinforcement learning algorithms in a green cooperative adaptive cruise control (CACC) platoon. Our aim is to train member vehicles to react appropriately in the event of a severe collision involving the leading vehicle. We seek to reduce collision damage and optimize energy consumption by encouraging behavior that conforms to the platoon's environmentally friendly aim. Our study provides insight into the potential benefits of using reinforcement learning algorithms to improve the safety and efficiency of CACC platoons while promoting sustainable transportation. The policy gradient algorithm used in this paper has good convergence in the calculation of the minimum energy consumption problem and the optimal solution of vehicle behavior. In terms of energy consumption metrics, the policy gradient algorithm is used first in the IVC field for training the proposed platoon problem. It is a feasible training decision-planning algorithm for solving the minimization of energy consumption caused by decision making in platoon avoidance behavior.

Keywords: carbon emissions; green driving; green eco; reinforcement learning; policy gradient; platoon

Citation: Gu, Z.; Liu, Z.; Wang, Q.; Mao, Q.; Shuai, Z.; Ma, Z. Reinforcement Learning-Based Approach for Minimizing Energy Loss of Driving Platoon Decisions. *Sensors* **2023**, *23*, 4176. https://doi.org/10.3390/s23084176

Academic Editors: Mustafa Ilhan Akbas and Jun Chen

Received: 9 March 2023
Revised: 12 April 2023
Accepted: 18 April 2023
Published: 21 April 2023

1. Introduction

The aim of green autonomous driving is to enable a vehicle to navigate and make decisions based on its surroundings without human intervention while adhering to environmental protection standards. A critical component of achieving this goal is developing driving strategies that can automatically output control signals such as steering, throttle, and brake in response to the observed environment.

The behavioral decision-making methods of automatic driving are divided into traditional methods and reinforcement learning methods. Traditionally, the decision control software system of an automatic driving system includes environmental prediction, behavior decision, action planning, path planning, and other functional modules. Traditional methods are often rule-based state control algorithms, including fuzzy logic [1], PID Bayesian control [2], and so on. Although these estimation algorithms are very accurate, such as Kalman filter [3], Kalman estimating IMU [4,5], Kalman estimating GNSS [5,6], and YOLO for RGB image detection [7], this paper focuses more on behavioral decision algorithms. For traditional behavioral decision making, behavioral decision algorithms

have the advantages of easy construction and adjustment, good real-time performance, simple application, etc. However, because it is difficult to adapt to all situations, they need to make targeted adjustments and their behavioral rule base can easily overlap and fail, and it is difficult for a finite state machine to cover all the conditions that the vehicle may encounter, resulting in decision making errors. For behavioral decision making based on reinforcement learning, the influence of environmental uncertainties can be reduced by simulating and learning various unexpected situations due to the strong computing power. The traditional approach is to utilize multiple sensors, such as cameras [8], radar [9,10], and lidar [11], to map visual inputs directly to action outputs [12,13]. However, traditional methods for developing driving strategies, such as utilizing multiple sensors, can generate excessive heat and consume a significant amount of energy, which is not conducive to the green and low-carbon aims of autonomous driving. Furthermore, these methods can be cumbersome and require significant resources to develop and implement [14]. In contrast, reinforcement learning offers a promising alternative to traditional methods for developing driving strategies. Unlike these traditional methods, which rely on human supervision and can be cumbersome and resource-intensive, reinforcement learning is achieved through an iterative trial-and-error approach that does not require explicit human supervision. This technique is well suited to action planning and has shown promise in developing effective driving strategies [15]. On the other hand, little research has been conducted on whether autonomous driving is more energy efficient than manual driving, particularly with regard to obstacle avoidance strategies [16].

Veins [17] is an open-source framework that enables the simulation of wireless communication in mobile in-vehicle environments. It is a useful tool for studying topics such as autonomous vehicle driving, formation driving, path planning, and coordination in signalized areas within connected vehicle environments. The framework's underlying structure can be used directly, allowing researchers to avoid wasting time and effort on non-research elements that can still significantly impact simulation results [18]. On the other hand, by exporting Veins simulations as OpenAI Gyms, Veins-Gym enables the use of reinforcement learning algorithms to address problems in the domain of vehicular ad hoc networks (VANETs) [19].

The connection established by Max Schettler between Veins tools and reinforcement learning by coupling Veins with OpenAI Gym provides a bridge for researchers who possess expertise in either field to leverage their knowledge in the other. This interface enables VANET researchers to access a compatible platform based on the generic RL framework [20]. Researchers can concentrate on studying algorithms and communication within the framework without the need to develop complicated interfaces for both Veins and reinforcement learning from scratch [21]. For platoon control, vehicle speed and acceleration are important state inputs, but these states cannot be obtained directly. Through sensor data such as GNSS, IMU, and camera, many scholars design robust estimation methods to obtain the states indirectly. However, in the Veins-Gym platform of this study, researchers do not need to pay much attention to environment building and state acquisition beyond reinforcement learning algorithms.

The field of reinforcement learning for energy conservation, green economy, and reducing CO_2 emissions is extensive and encompasses a wide range of studies [22]. For example, in the optimal regulation of microgrids, carbon emissions allowances take into account volume prediction [23], electrical energy consumption management of household appliances [24], hybrid clouds harnessing renewable energy, task scheduling [25], and so on, which all concern green and low-carbon environmental protection. However, despite the significant attention given to reinforcement learning applications for energy conservation and carbon emission reduction, limited studies have addressed the issues related to autonomous driving, connected vehicle communication, and energy consumption. These issues include carbon emissions from the decision-making behavior of autonomous vehicles, energy consumption during connected vehicle communication, etc., and how to reduce greenhouse gas emissions from vehicles. Max Schettler's development of Veins-Gym has

made it easy to perform data statistics and analysis, offering an excellent simulation and development tool for investigating fleet behavior decision making [26,27], particularly the obstacle avoidance problem on which this paper concentrates [28,29].

This paper focuses on examining how reinforcement learning algorithms can be utilized to train the behavior of member vehicles in a CACC platoon [30,31] consisting of vehicles with varying parameters in the event of a serious collision involving the leading vehicle [32]. Additionally, the paper aims to determine the most energy-efficient and eco-friendly solution that consumes the least amount of energy while also fulfilling the requirements of avoiding collisions or minimizing collision damage. The conventional approach relies on using sensors that operate independently, without any communication, to assess the surroundings and determine the appropriate course of action.

The main contributions of this paper are as follows:

1. A hypothetical situation is created to depict a scenario where a line of vehicles is present on a two-lane highway. When there are no other vehicles present, the front vehicle of the platoon suffers a serious traffic accident.
2. To prevent further damage, reinforcement learning (policy gradient algorithm) is used to obtain the most efficient strategy to be adopted by the member vehicles to reduce the impact of the collision on the team.
3. While solving the collision avoidance problem, the reinforcement learning algorithm also examines the damage caused by the vehicle behavior and computes the strategy that minimizes the damage.
4. In order to break the limitations of traditional algorithms, reinforcement learning algorithms (policy gradients) are applied to the behavioral decision of the fleet, which is a leap forward and a hot spot for future research.

2. Platoon Algorithms with RL

This section focuses on several aspects, including inferring the formula of the policy gradient (PG) algorithm for reinforcement learning, modeling the Veins simulation model, and assuming a universal and typical simulation scenario [33].

2.1. The CACC Car-Following Model

x_i is the displacement of the following vehicle, v_i is the speed of the following vehicle, e is the error between the actual distance and the desired distance, T is the minimum safe headway time distance, $i-1$ is the front car, and v_{kprev} is the speed of the vehicle in the previous moment.

$$\begin{aligned} v_i &= v_{kprev} + k_{pe} + k_d\dot{e} \\ e &= x_{i-1} - x_i - Tv_i \end{aligned} \tag{1}$$

2.2. Proposed Car Dynamics Cost Model

This paper assumes that the nature of the collision between vehicles is inelastic:

$$\begin{cases} m_2v_0 = m_1v_1 + m_2v_2 \\ \frac{1}{2}m_2v_0^2 = \frac{1}{2}m_1v_1^2 + \frac{1}{2}m_2v_2^2 + E \end{cases} \tag{2}$$

For individual vehicle loss functions within a fleet:

$$J_i = E_i = \frac{1}{2}m_2\left(v_0^2 - v_2^2\right) - \frac{1}{2}m_1v_1^2 \tag{3}$$

In contrast to the actual scenario, Wentao Chen posits that the nature of vehicle collisions is completely inelastic. However, the analysis and reconstruction of heavy goods vehicle traffic accidents categorizes the nature of vehicle collisions based on three distinct properties associated with three different vehicle speeds. This approach aligns more

accurately with real-world situations. In this paper, the relationship between mass and collision loss is proposed as:

$$\frac{v_2}{v_1} = \frac{1}{2}\frac{1}{1+e^{\ln\frac{m_1}{m_2}}} = \frac{1}{2}\frac{m_2}{m_1+m_2} \tag{4}$$

To facilitate the calculation, we specify:

- When $m_1 < m_2$, it is elastic collision; $v_2/v_1 = 1/2$.
- When $m_1 = m_2$, it is inelastic collision; $v_2/v_1 = 1/4$.
- When $m_1 > m_2$, it is completely inelastic collision; $v_2/v_1 = 0$.

The platoon cost function is defined as:

$$J = \sum_{i=1}^{N} J_i \tag{5}$$

2.3. Markov Decision Process (MDP)

For a complete sequence of state behavior trajectories $\tau = (s_0, a_0, \ldots, s_{T-1}, a_{T-1}, s_T)$, i.e., the process of obtaining the next state s_{i+1} after obtaining behavior a_i at state s_i, there are single-step reward functions $R(s_t, a_t)$ and total reward functions $R(\tau) = \sum_{t=0}^{T-1} \gamma^i R(s_t, a_t)$ for obtaining the maximum expected reward, where π_θ is the parameterization strategy of the neural network composition. Under the strategy π_θ, the expected value τ is used for the trajectory, so the problem can be transformed into finding the optimal parameter θ.

The gradient of the desired reward can be obtained from the gradient descent $\nabla_\theta \mathbb{E}_{\pi\theta} R(\tau)$, and the parameters are updated by the hyperparametric learning rate α.

$$\theta \leftarrow \theta + \alpha \nabla_\theta \mathbb{E}_{\pi\theta} R(\tau) \tag{6}$$

Let $P(\tau|\theta)$ be the probability of the trajectory τ under the strategy π_θ. Then, the gradient can be calculated:

$$\begin{aligned} \nabla_\theta \mathbb{E}_{\pi\theta} R(\tau) &= \nabla_\theta \sum_{\tau} P(\tau|\theta) R(\tau) \\ &= \sum_{\tau} \nabla_\theta P(\tau|\theta) R(\tau) \\ &= \sum_{\tau} \tfrac{P(\tau|\theta)}{P(\tau|\theta)} \nabla_\theta P(\tau|\theta) R(\tau) \\ &= \sum_{\tau} P(\tau|\theta) \nabla_\theta \log P(\tau|\theta) R(\tau) \\ &= \mathbb{E}_{\pi\theta(\nabla_\theta \log P(\tau|\theta) R(\tau))} \end{aligned} \tag{7}$$

Therefore, the probability of the trajectory τ can be calculated:

$$P(\tau|\theta) = p(s_0) \prod_{t=0}^{T-1} p(s_{t+1}|s_t, a_t) \pi_{\theta(a_t|s_t)} \tag{8}$$

where $p(s_{t+1}|s_t, a_t)$ is the probability of transitioning to state s_{t+1} after taking behavior a_t at the moment of state s_t.

2.4. Policy Gradient Algorithm

Consider the optimization model from the gradient of the objective function:

$$\nabla J_\theta(\theta) = \int \nabla_{\pi\theta}(\tau) r(\tau) \mathrm{d}r = \mathbb{E}_{\tau\sim\pi\theta(\tau)} \nabla_\theta \log \pi_\theta(\tau) r(\tau))$$

to obtain the gradient:

$$\mathbb{E}_{\tau\sim\pi\theta(\tau)} \left(\sum_t \nabla_\theta \log \pi_\theta(a_t|s_t) \right) \left(\sum_t r(s_t, a_t) \right) \tag{9}$$

Strategy function:

When designing policy functions, it is important to address the discrete state space and the continuous state space separately due to their substantial differences in the quantity and definition of states.

2.5. Gauss Policy Function

The Gauss strategy function for the continuous behavior space is generated from a Gaussian distribution with a fractional function:

$$\nabla_\theta \log_{\pi\theta}(s, a) = \frac{\left(a - \phi(s)^{T_\theta}\right)\phi(s)}{\sigma^2} \tag{10}$$

2.6. Softmax Policy Function

Softmax for discrete spaces:

$$\pi_\theta(s, a) = \frac{e^{\phi(s,a)^{T_\theta}}}{\sum_b e^{\phi(s,a)^{T_\theta}}} \tag{11}$$

The odds of a behavior occurring are weighed using a linear combination of the features $\phi(s, a)$ describing the state and the behavior with the parameter θ. The corresponding score function is its derivative:

$$\nabla_\theta \log_{\pi\theta}(s, a) = \phi(s, a) - \mathbb{E}_{\pi\theta}[\phi(s, \cdot)] \tag{12}$$

3. Proposed Model

3.1. Vehicle Dynamics and Network Parameters

Consider a smooth and empty two-lane, one-way straight highway with a row of N convoys with the following parameters.

The fleet uses the CACC follow-the-leader model such that the fleet member i $\in$ N can maintain a relative speed of 0.

Let the lead vehicle in the convoy with marker i $= 0$ at the moment t $= 0$ have a larger vehicle collision, so that:

$$v_0 = \begin{cases} 120, & t < 0 \\ 110 - 8t, & t \geq 0 \\ 0, & t \geq 13.75 \end{cases} \tag{13}$$

Vehicle 1, when vehicle 0 experiences the accident, brakes in response when the relative distance is reduced to 6.8 m; when the vehicle speed is reduced to 0, the relative speed is reduced to 57.16 m, which is greater than the head time distance of 1.5 s, that is, 50 m, resulting in vehicle 1 causing a rear-end accident and a chain of further rear-end accidents from the cars behind it.

Consider the workshop communication using protocol IEEE802.11p for short-distance communication. To verify and confirm the network adjacency, hello packets are sent

periodically by the vehicle. To ensure the timeliness of the communication, the transmission protocol uses the faster UDP and the packet interval is set to 0.2 s. Learning rate α and exploration rate ε are the reinforcement learning training parameters. The detailed simulation parameters are shown in Table 1.

Table 1. Assumed speed and routing protocol.

Parameter	Value
Car speed	120 Km/h
Headway	1.5 s
Response time	1 s
Deceleration	8 m/s^2
Routing packet size	512 Bytes
Simulation distance	1000 m
Maximum rate	10 MB/s
Number of nodes	4
Communication distance	250 m
Hello packet interval	Ls
Transfer Protocol	UDP
Packet interval	0.1 s
MAC layer protocol	802.11
Channel transmission rate	3 Mbps
Learning rate α	0.001, 0.01, 0.05, 0.1
Exploration rate ε	0.01, 0.05, 0.1, 0.2

Additionally, considering the scenario, the mass of the motor vehicle does not satisfy $m_i = m_j$ and the deceleration does not satisfy $a_i = a_j$; therefore, in scenarios where traffic accidents are caused by the leading vehicle, braking may not be the optimal approach. The growing volume of road traffic necessitates further reduction in driving distances, which in turn calls for increased attention to the safety of autonomous vehicles. It is essential to focus on developing more intelligent approaches for path planning and safe decision making. Policy gradient algorithm for platoon is as follows (Algorithm 1).

Algorithm 1 Policy gradient for platoon

Input: a differentiable policy parameterization $\pi(a|s,\theta)$
Algorithm parameter: step size $a > 0$
Initialize policy parameter $\theta \in \mathbb{R}^{d'}$, environment and state S_0

Loop for each episode:
Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot|\cdot,0)$
Loop for each step of the episode $t = 0, \ldots, T-1$:
$S \leftarrow$ return from step t (S_t)
$R \leftarrow R + \gamma^i R(s_t, a_t)$
$\theta \leftarrow \theta + \alpha\gamma^t \nabla \bar{R}_\theta$
If episode is complete:
Break
Train and learn for agent:
Return θ

3.2. Strategic Gradient Decision-Making Behavior

3.2.1. State Space and Action Space

In certain public platforms, such as Gym, the state space for most domains is readily accessible, allowing scholars to compare the performance and convergence speed of various algorithms. However, in real-world projects, state space design work must be carried out independently. Based on the author's personal experience, adding new state information can significantly improve performance, more so than other aspects (such as tuning), which

is very cost effective, so the optimization of the state space is almost always carried out in the project.

3.2.2. Mission Analysis

The following situation was established: the foremost vehicle of a convoy on a highway becomes engaged in a significant collision, resulting in a severe reduction in speed. At this point, the convoy members are compelled to make crucial choices to evade the collision or decrease the harm caused by an unavoidable collision.

3.2.3. Observation Spatial Information Filtering

The information in the environment is passed to the intelligence (agent) for generating the reward function (reward). In this paper, we consider that for the ith member car, the key point is the relative position, the relative velocity of the previous car i-1 and i-2. Moreover, if the car decides to change lanes, the environmental feedback data of the next state, which include the relative position and velocity with the preceding car after changing lanes, are analyzed. The state space is defined as [x1, x2, x3, v1, v2, v3], where x and v denote the distance and speed, respectively, and the labels 1, 2, and 3 denote the physical quantities with the car in front, the car behind, and the car in front of the next lane, respectively. The behavior space comprises [lane change, deceleration], and their elements are Boolean values.

3.2.4. Reward Function Settings

The fleet collision loss function is utilized to consider the environment space key points. The reward in the next moment within the current lane is dependent on the state space, where any behavior a_i must result in a new change in the environment s_i. To ensure the safety of all vehicles and that each vehicle remains within the safety threshold, the reward function must be designed accordingly. Additionally, the reward function takes into account the fuel consumption that results from the vehicle's braking and lane-changing behaviors.

4. Analysis of Simulation Results

The convergent process of each loss function value in the whole training process is compared under different learning and exploration rates. The energy loss function and the collision loss function are slightly different. The two are not strongly correlated. When the energy loss is low, the collision loss may be larger; sometimes, the opposite result is produced. The selection of super parameters is extremely important. The training system is chaotic, and any small change in the parameter value leads to great changes in the convergence characteristics, even in the non-convergence situation. The value of the learning rate parameter used for the constraint training convergence curve is set to [0.001, 0.1]. The exploration rate parameters used to explore more decisions have values of [0.001, 0.2].

Considering the complexity of the scenario, it is reasonable to assume that there is no need to dynamically adjust the learning rate. Figures 1–4 show the loss calculated by the strategy gradient algorithm according to the training time under the condition of different learning rates and exploration rates in the target scenario. In order to obtain faster convergence time and reduce training time in the whole training process, the value of the learning rate should be larger. A low learning rate leads to a difficult training process and long convergence time. Too high a learning rate causes a serious loss value jump, makes it difficult to obtain the overall downward trend, and even causes the problem of non-convergence. In order to obtain more strategy choices during the initial training, the value of the exploration rate needs to be set high. Too low an exploration rate makes it difficult to jump out of the current decision, that is, the loss gradient drops to a minimum point that is not relatively good and the existing decision remains unchanged. Too large an exploration rate causes the agent to hesitate, try repeatedly among many possible choices,

and fail to make the optimal decision. Based on the training results, the optimal learning rate, $\alpha = 0.05$, and the optimal exploration rate, $\varepsilon = 0.1$, are obtained.

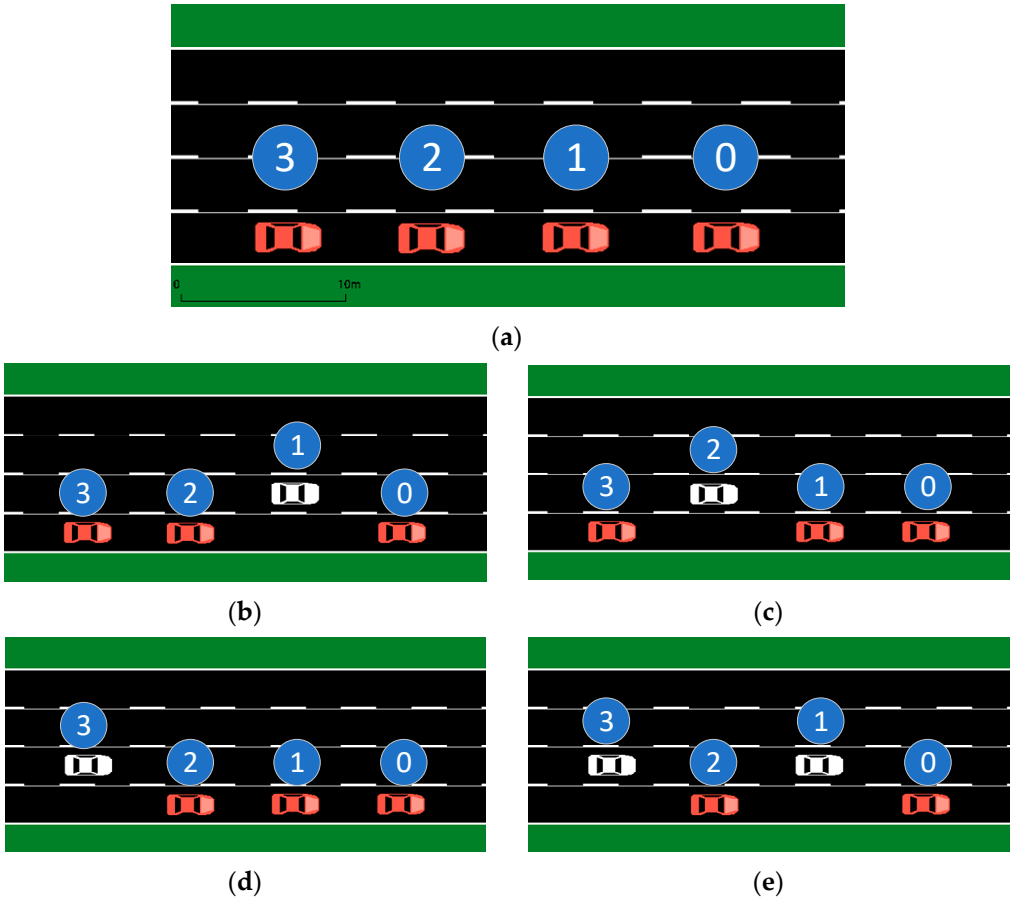


Figure 1. Decision making: slowing down and changing lanes. (**a**) No action; (**b**) the 1st member car changing lanes; (**c**) the 2nd member car changing lanes; (**d**) the 3rd member car changing lanes; (**e**) the 1st and 3rd member cars changing lanes.

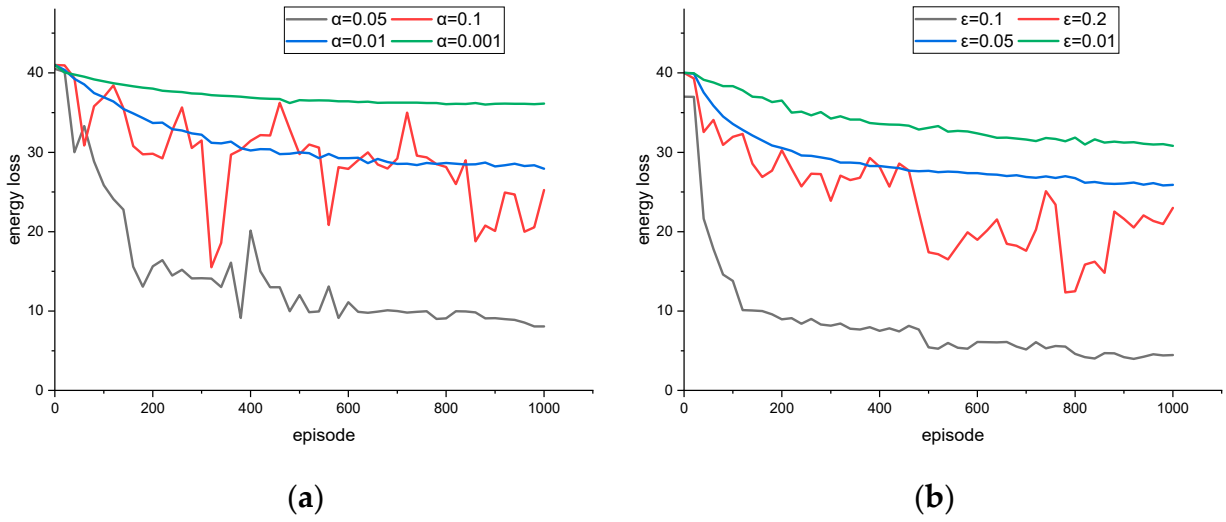
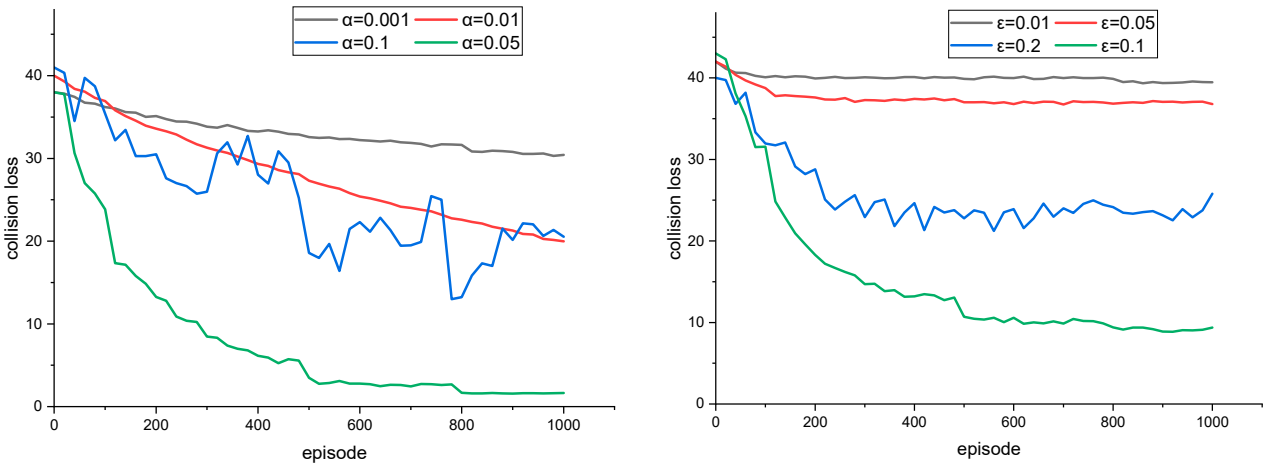


Figure 2. In order to obtain the optimal energy loss strategy, different hyperparameters were used for training energy loss with (**a**) different α; (**b**) different ε.



(**a**) (**b**)

Figure 3. In order to obtain the optimal collision loss strategy that is different from energy loss, different hyperparameters were used for training collision loss with (**a**) different α; (**b**) different ε.

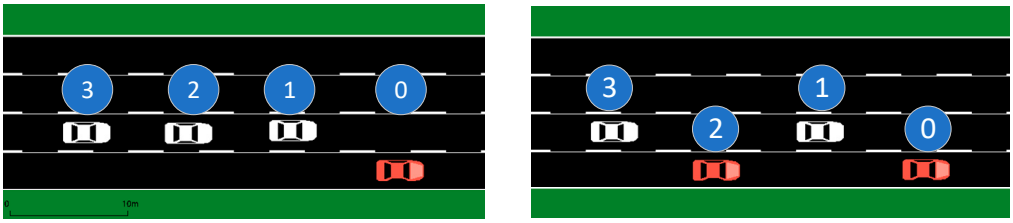


Figure 4. Result of traditional and proposed method.

In the simulation scenarios and algorithm models of Veins-Gym, the vehicle under test continuously tests and trains the deep neural network. Sensor-based vehicle distance decision algorithms need no training. The decision of CACC is determined by the headway, min-Gap (the distance between the front bumper and the rear bumper of the vehicle in the queue), and the reaction time τ. The optimal headway = 1.5 s, min-Gap = 2.5 m, and τ = 1 s are adopted. Comparing the traditional algorithm with the proposed algorithm, the following is obtained.

In one thousand training sessions, a target vehicle exhibits a pattern of both overall convergence and local oscillation in the collision loss generated by the fleet and the reward value it receives. The proposed method differs from the traditional approach in that lane changes occur between vehicles instead of all vehicles changing lanes. The loss convergence process of the agent, which is highly correlated with the reward function, can be evaluated using multiple metrics. When evaluating a single intelligent entity, the most crucial factor to consider is the trend of collision loss and reward value changes.

There is no strong correlation between energy loss and collision loss, which is particularly obvious in the traditional algorithm; that is, all CACCs of the team members may adopt the same lane change strategy at the same time, which makes the decision without safety significance and increases energy consumption. According to the Figure 5, the policy gradient algorithm-based reinforcement learning method outperforms conventional methods in anticipating autonomous driving situations, devising optimal routes, and issuing timely warnings. This superiority is evident in both subjective impressions and objective assessments.

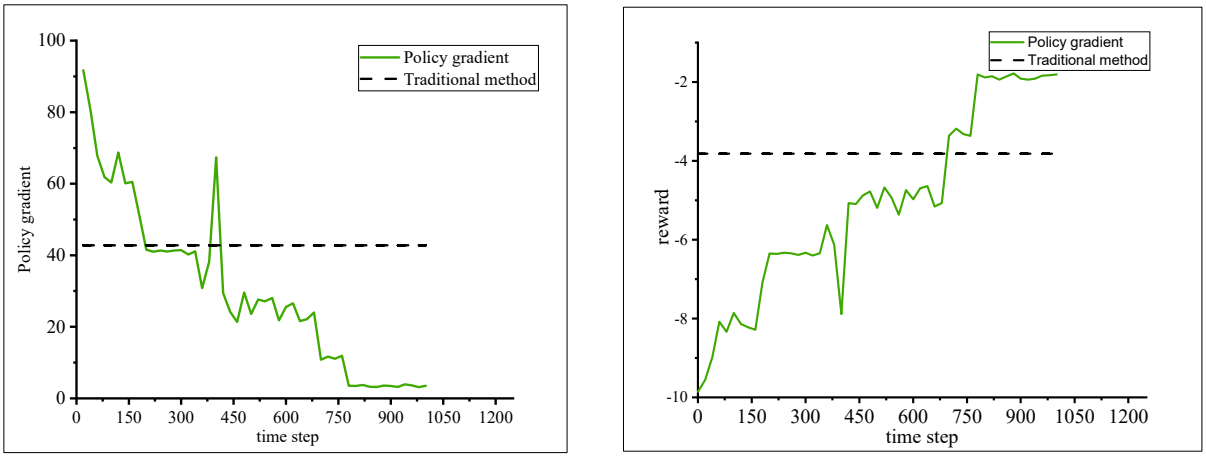


Figure 5. PG algorithm and traditional algorithm cost function graph. (**a**) Collision loss convergence; (**b**) reward convergence.

Analysis of Three-Lane Scenario Simulation Results

In the three-lane experiment, the parameters and the design of the model were different compared to those for the two-lane experiment. According to Figure 6,the distance and steering model was introduced here, which is more in line with the actual situation. The expansion region is defined as follows: the centroid of the top view of the car is the center of the expansion rectangle; the side length is the rectangular edge, which is determined by the car edge and the speed of the two vehicles. The minimum distance between the rectangles corresponds to the safety car distance of the two vehicles. When the following vehicle's speed is greater than that of the leading vehicle and the lane changes, the velocity vector at the minimum distance point is used to predict the collision and decide the change in the reward function. When the speed of the following vehicle is close to or less than the speed of the leading vehicle, the velocity vector has little significance. It is used to share the speed of the forward direction and increase the distance between the two vehicles.

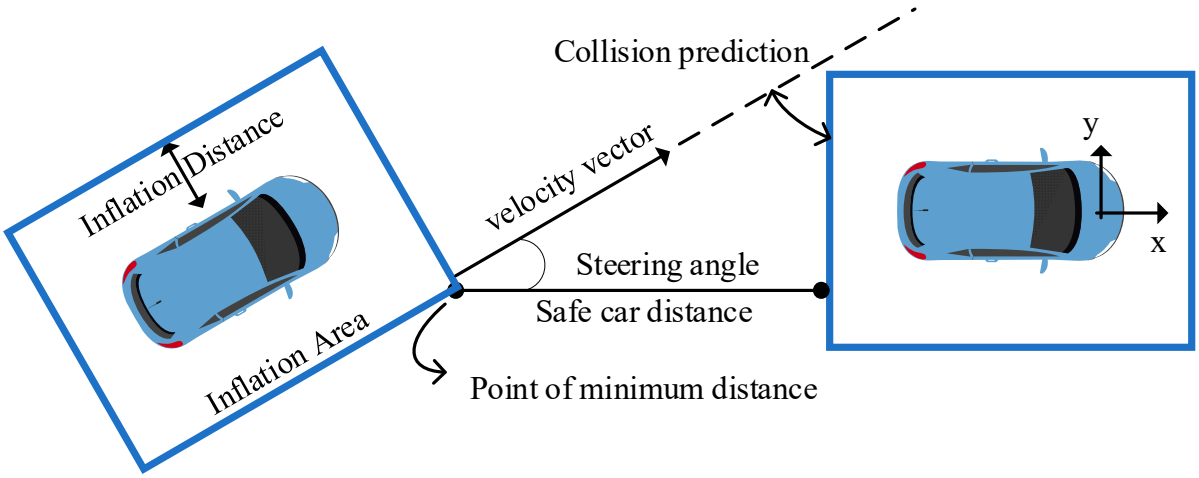


Figure 6. Distance and steering model.

The operation loss parameter of the reward function is set to a larger value, that is, the reward value is more affected by the operation loss. The vehicle trajectories over a period of time are shown in the Figures 7 and 8 below. The agent should pay as little attention as possible to the influence of the distance change outside the expansion interval on the security. Agents are trained several times, and the three agents with the maximum

reward value are taken as the training results. There are some commonalities among these three results. The first and second cars are always in different lanes, while the third car has no obvious correlation with the position of the car in front. This is because the third car is furthest away from the accident vehicle and therefore has more maneuvering time and distance.

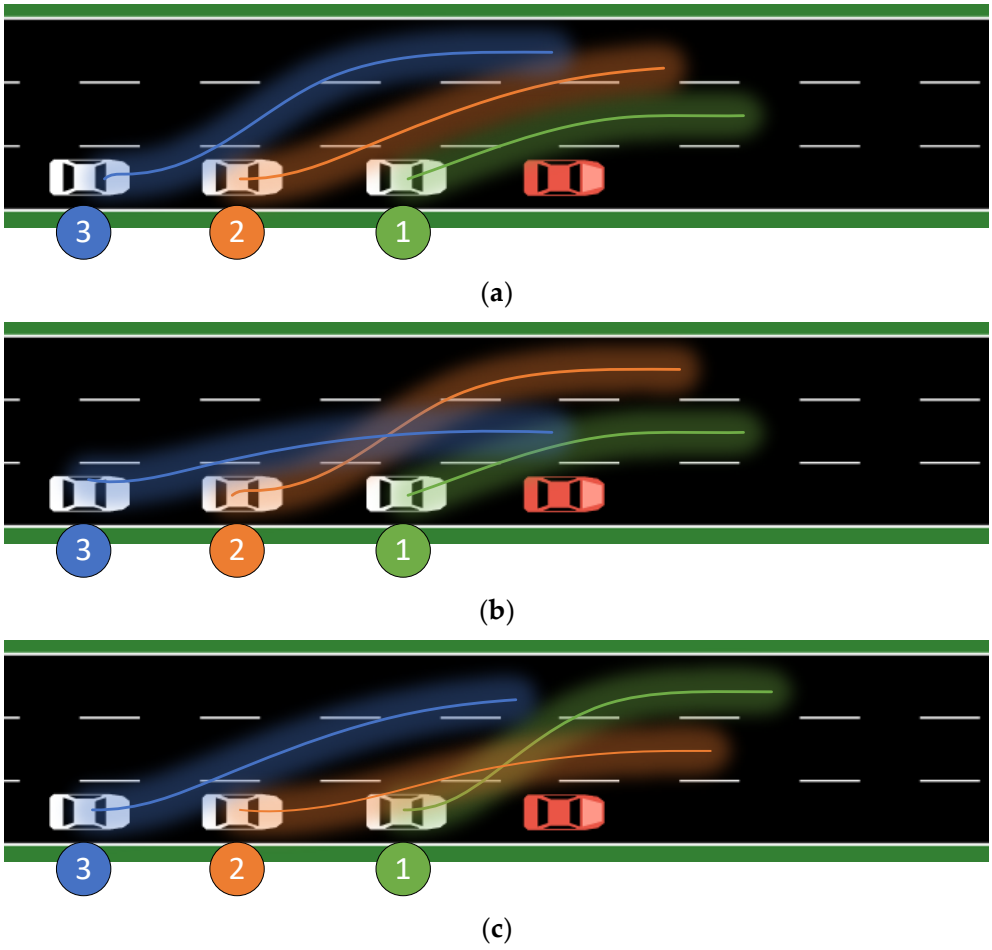


Figure 7. Results of agent training with minimum operational loss priority. The trajectories of (**a**) car 1 changing to lane 2; car 2, 3 to 3; (**b**) car 1, 3 changing to lane 2; car 2 to 3; (**c**) car 1 changing to lane 2; car 2, 3 to 3.

Examining the relationship between the sequence of fleet members and important physical quantities over time is of great interest. Even though vehicles make many steering movements during high-speed driving, the results of experiments conducted in a reinforcement learning agent simulation environment are worth investigating. Even if the directional change resulting from this movement is very small, typically less than 10 degrees, it can have catastrophic consequences. For instance, in rainy or snowy weather, the friction between the tire and the ground in the vertical direction to the direction of travel can cause rolling friction which transforms into sliding friction. Once this occurs, the vehicle's state becomes uncontrollable, leading to serious traffic accidents. In this three-lane experiment, the behavior of each member is intrinsically linked to the entire system. Due to communication delay and uncontrollable hardware control random delay, the behavior of each vehicle has a reasonable lag and delay in response to the previous vehicle's actions.

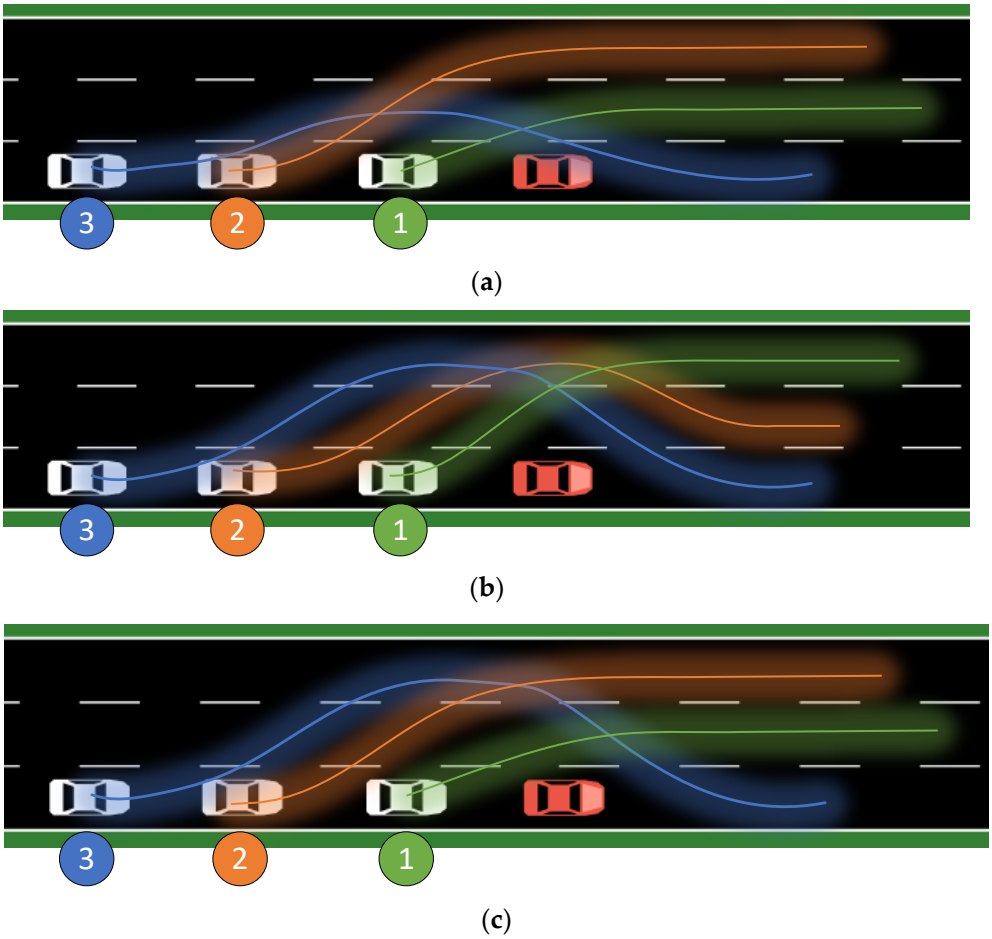


Figure 8. Results of agent training with large distance loss priority. Number of lane changes for vehicle 1, 2, 3: (**a**) 1, 2, 2; (**b**) 2, 3, 4; (**c**) 1, 2, 4.

The function that describes the change in the minimum distance between a vehicle and its leading vehicle as a function of time is always monotonically decreasing, but for vehicles with different indices, the curves exhibit reasonable differences. In general, vehicles with higher indices exhibit greater negative acceleration, i.e., they tend to maintain a greater distance from their leading vehicle. Acceleration, as the second derivative of distance, must have a sufficiently large integral to increase the distance between vehicles; therefore, the acceleration of rear vehicles must be larger and initiated earlier to obtain a larger reward for the agent. To increase the distance between vehicles under loose conditions, vehicles can gain greater distance by making free turns and agents will do whatever it takes, including changing lanes continuously, to achieve this goal. When the vehicle that continuously changes lanes reaches the front of the road and no longer detects a leading vehicle, it is not motivated to change lanes or decelerate. Although the vehicle in this situation should accelerate and overtake surrounding vehicles to leave the accident area as quickly as possible, the simulation environment does not have parameters or behaviors that are less related to obstacle avoidance, minimal energy loss, and minimal safety cost. Therefore, the acceleration of the vehicle is always non-positive, sometimes negative, and sometimes zero.

5. Conclusions

In this paper, we propose a policy gradient algorithm for computing the minimum energy consumption for autonomous driving decisions in the context of a green economy. The approach utilizes the Veins-Gym reinforcement learning and the Telematics training platform. Two experiments were designed to train the proposed algorithm for obstacle

avoidance strategies in two-lane and three-lane highway situations when the team's lead vehicle is involved in a major traffic accident. By training the optimal decision, the vehicles perform better in terms of obstacle avoidance and energy loss performance and achieve minimized cost loss. Due to different parameter settings, the vehicles' obstacle avoidance strategy can be switched between conservative and aggressive operations. In the future deployment of a large number of self-driving vehicles, reinforcement learning has great promise for some decisions, such as environmental prediction and behavioral decisions.

Future research should address the following:

1. This study used short-range communication networks only for small-scale vehicle communication, and when more vehicles and more network types (such as base stations) are added, how reinforcement learning can cope with them should be specifically analyzed.
2. When the road environment is more complex, reinforcement learning intelligences should be multi-agents and consideration should be given to whether to use distributed or centralized agents.
3. The crash environment in this study is only one of many small probability situations; a more generalized decision algorithm should be sought to minimize loss.
4. There is endogeneity in the road passage, traffic calming methods, and crash behavior decisions, and their correlation should be studied.

Author Contributions: Supervision, Q.M., Z.S. and Q.W.; writing—review, Z.M. and Z.G.; conceptualization, writing—original draft, data curation, software, Z.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Nature Science Foundation of China under Grant 61971182; in part by the Hunan Provincial Natural Science Foundation (No.2022JJ5005), Key Projects of Hunan Provincial Education Department (No. 19A139), the Natural Science Youth Foundation of Hunan Province under Grant 2020JJ5144; in part by the Natural Science Youth Foundation of Liaoning Province under Grant 2020-KF-21-03; in part by the Scientific Research Project of Hunan Education Department under Grant 19C0563; and in part by Nature Science Foundation of Hunan Province under Grant 2021JJ30145.

Institutional Review Board Statement: No humans or animals were involved in this study.

Informed Consent Statement: No humans or animals were involved in this study.

Data Availability Statement: Not applicable.

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

References

1. Xia, X.; Meng, Z.; Han, X.; Li, H.; Tsukiji, T.; Xu, R.; Zhang, Z.; Ma, J. Automated Driving Systems Data Acquisition and Processing Platform | DeepAI. *arXiv preprint* **2022**, arXiv:2211.13425. Available online: https://deepai.org/publication/automated-driving-systems-data-acquisition-and-processing-platform (accessed on 9 April 2023).
2. Khosravi, M.; Behrunani, V.N.; Myszkorowski, P.; Smith, R.S.; Rupenyan, A.; Lygeros, J. Performance-Driven Cascade Controller Tuning with Bayesian Optimization. *IEEE Trans. Ind. Electron.* **2022**, *69*, 1032–1042. [CrossRef]
3. Xia, X.; Hashemi, E.; Xiong, L.; Khajepour, A. Autonomous Vehicle Kinematics and Dynamics Synthesis for Sideslip Angle Estimation Based on Consensus Kalman Filter. *IEEE Trans. Control. Syst. Technol.* **2023**, *31*, 179–192. [CrossRef]
4. Xia, X.; Xiong, L.; Huang, Y.; Lu, Y.; Gao, L.; Xu, N.; Yu, Z. Estimation on IMU yaw misalignment by fusing information of automotive onboard sensors. *Mech. Syst. Signal Process.* **2022**, *162*, 107993. [CrossRef]
5. Improved Vehicle Localization Using On-Board Sensors and Vehicle Lateral Velocity | IEEE Journals & Magazine | IEEE Xplore. Available online: https://ieeexplore.ieee.org/document/9707770 (accessed on 9 April 2023).
6. Liu, W.; Xia, X.; Xiong, L.; Lu, Y.; Gao, L.; Yu, Z. Automated Vehicle Sideslip Angle Estimation Considering Signal Measurement Characteristic. *IEEE Sens. J.* **2021**, *21*, 21675–21687. [CrossRef]
7. Liu, W.; Quijano, K.; Crawford, M.M. YOLOv5-Tassel: Detecting Tassels in RGB UAV Imagery with Improved YOLOv5 Based on Transfer Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2022**, *15*, 8085–8094. [CrossRef]
8. Kiran, B.R.; Sobh, I.; Talpaert, V.; Mannion, P.; Al Sallab, A.A.; Yogamani, S.; Perez, P. Deep Reinforcement Learning for Autonomous Driving: A Survey. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 4909–4926. [CrossRef]

9. Poongodi, M.; Bourouis, S.; Ahmed, A.N.; Vijayaragavan, M.; Venkatesan KG, S.; Alhakami, W.; Hamdi, M. A Novel Secured Multi-Access Edge Computing based VANET with Neuro fuzzy systems based Blockchain Framework-ScienceDirect. *Comput. Commun.* **2022**, *192*, 48–56. Available online: https://www.sciencedirect.com/science/article/abs/pii/S0140366422001669 (accessed on 9 April 2023).
10. Gao, W.; Jiang, Z.-P.; Lewis, F.L.; Wang, Y. Cooperative optimal output regulation of multi-agent systems using adaptive dynamic programming. In Proceedings of the 2017 American Control Conference (ACC), Seattle, WA, USA, 24–26 May 2017; pp. 2674–2679.
11. Park, H.; Lim, Y. Deep Reinforcement Learning Based Resource Allocation with Radio Remote Head Grouping and Vehicle Clustering in 5G Vehicular Networks. *Electronics* **2021**, *10*, 3015. [CrossRef]
12. Reinforcement Learning Based Power Control for VANET Broadcast against Jamming | IEEE Conference Publication | IEEE Xplore. Available online: https://ieeexplore.ieee.org/document/8647273 (accessed on 9 April 2023).
13. Lansky, J.; Rahmani, A.M.; Hosseinzadeh, M. Reinforcement Learning-Based Routing Protocols in Vehicular Ad Hoc Networks for Intelligent Transport System (ITS): A Survey. *Mathematics* **2022**, *10*, 4673. [CrossRef]
14. Wang, J.; Zhu, K.; Hossain, E. Green Internet of Vehicles (IoV) in the 6G Era: Toward Sustainable Vehicular Communications and Networking. *arXiv* **2021**. [CrossRef]
15. Peng, H.; Shen, X.S. Deep Reinforcement Learning Based Resource Management for Multi-Access Edge Computing in Vehicular Networks. *IEEE Trans. Netw. Sci. Eng.* **2020**, *7*, 2416–2428. [CrossRef]
16. Yu, K.; Lin, L.; Alazab, M.; Tan, L.; Gu, B. Deep Learning-Based Traffic Safety Solution for a Mixture of Autonomous and Manual Vehicles in a 5G-Enabled Intelligent Transportation System. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 4337–4347. [CrossRef]
17. Noori, H. Realistic urban traffic simulation as vehicular Ad-hoc network (VANET) via Veins framework. In Proceedings of the 2012 12th Conference of Open Innovations Association (FRUCT), Oulu, Finland, 5–9 November 2012; pp. 1–7.
18. Saravanan, M.; Ganeshkumar, P. Routing using reinforcement learning in vehicular ad hoc networks. *Comput. Intell.* **2020**, *36*, 682–697. [CrossRef]
19. Schettler, M.; Buse, D.S.; Zubow, A.; Dressler, F. How to Train your ITS? Integrating Machine Learning with Vehicular Network Simulation. In Proceedings of the 2020 IEEE Vehicular Networking Conference (VNC), New York, NY, USA, 16–18 December 2020; pp. 1–4. [CrossRef]
20. Zhang, C.; Du, H. DMORA: Decentralized Multi-SP Online Resource Allocation Scheme for Mobile Edge Computing. *IEEE Trans. Cloud Comput.* **2022**, *10*, 2497–2507. [CrossRef]
21. Li, S.; Wu, Y.; Cui, X.; Dong, H.; Fang, F.; Russell, S. Robust Multi-Agent Reinforcement Learning via Minimax Deep Determin-istic Policy Gradient. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33. [CrossRef]
22. Naderializadeh, N.; Hashemi, M. Energy-Aware Multi-Server Mobile Edge Computing: A Deep Reinforcement Learning Approach. In Proceedings of the 2019 53rd Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 3–6 November 2019; pp. 383–387. [CrossRef]
23. Wong, F. Carbon emissions allowances trade amount dynamic prediction based on machine learning. In Proceedings of the 2022 International Conference on Machine Learning and Knowledge Engineering (MLKE), Guilin, China, 25–27 February 2022; pp. 115–120. [CrossRef]
24. Zhang, H.; Wu, D.; Boulet, B. A Review of Recent Advances on Reinforcement Learning for Smart Home Energy Management. In Proceedings of the 2020 IEEE Electric Power and Energy Conference (EPEC), Edmonton, AB, Canada, 9–10 November 2020; pp. 1–6. [CrossRef]
25. Yang, Y.; Shen, H. Deep Reinforcement Learning Enhanced Greedy Algorithm for Online Scheduling of Batched Tasks in Cloud in Cloud HPC Systems. *IEEE Trans. Parallel Distrib. Syst.* **2022**, *33*, 3003–3014. [CrossRef]
26. Ban, Y.; Xie, L.; Xu, Z.; Zhang, X.; Guo, Z.; Hu, Y. An optimal spatial-temporal smoothness approach for tile-based 360-degree video streaming. In Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, USA, 10–13 December 2017; pp. 1–4. [CrossRef]
27. Reinforcement Learning Based Rate Adaptation for 360-Degree Video Streaming | IEEE Journals & Magazine | IEEE Xplore. Available online: https://ieeexplore.ieee.org/document/9226435 (accessed on 9 April 2023).
28. Subramanyam, S.; Viola, I.; Jansen, J.; Alexiou, E.; Hanjalic, A.; Cesar, P. Subjective QoE Evaluation of User-Centered Adaptive Streaming of Dynamic Point Clouds. In Proceedings of the 2022 14th International Conference on Quality of Multimedia Experience (QoMEX), Lippstadt, Germany, 5–7 September 2022; pp. 1–6. [CrossRef]
29. Yazid, Y.; Ez-Zazi, I.; Guerrero-González, A.; El Oualkadi, A.; Arioua, M. UAV-Enabled Mobile Edge-Computing for IoT Based on AI: A Comprehensive Review. *Drones* **2021**, *5*, 148. [CrossRef]
30. Al-Turki, M.; Ratrout, N.T.; Rahman, S.M.; Reza, I. Impacts of Autonomous Vehicles on Traffic Flow Characteristics under Mixed Traffic Environment: Future Perspectives. *Sustainability* **2021**, *13*, 11052. [CrossRef]
31. Yao, L.; Zhao, H.; Tang, J.; Liu, S.; Gaudiot, J.-L. Streaming Data Priority Scheduling Framework for Autonomous Driving by Edge. In Proceedings of the 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC), Madrid, Spain, 12–16 July 2021; pp. 37–42. [CrossRef]

32. A Survey of Multi-Access Edge Computing and Vehicular Networking | IEEE Journals & Magazine | IEEE Xplore. Available online: https://ieeexplore.ieee.org/document/9956993 (accessed on 9 April 2023).
33. Lu, S.; Zhang, K.; Chen, T.; Başar, T.; Horesh, L. Decentralized Policy Gradient Descent Ascent for Safe Multi-Agent Reinforcement Learning. *Proc. Conf. AAAI Artif. Intell.* **2021**, *35*, 8767–8775. [CrossRef]