sustainability

MDPI

Article

Modeling Interactions of Autonomous/Manual Vehicles and Pedestrians with a Multi-Agent Deep Deterministic Policy Gradient

Weichao Hu [1,2], Hongzhang Mu [3,4], Yanyan Chen [1,*], Yixin Liu [5] and Xiaosong Li [2]

[1] School of Metropolitan Transportation, Beijing University of Technology, Beijing 100124, China; hu.weichao@outlook.com
[2] Research Institute for Road Safety of the Ministry of Public Security, Beijing 100062, China; lxs199602@126.com
[3] Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100085, China; muhongzhang@iie.ac.cn
[4] School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100085, China
[5] Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China; guangmingzui2de@126.com
* Correspondence: cdyan@bjut.edu.cn; Tel.: +86-010-6739-1680

Abstract: This article focuses on the development of a stable pedestrian crash avoidance mitigation system for autonomous vehicles (AVs). Previous works have only used simple AV–pedestrian models, which do not reflect the actual interaction and risk status of intelligent intersections with manual vehicles. The paper presents a model that simulates the interaction between automatic driving vehicles and pedestrians on unsignalized crosswalks using the multi-agent deep deterministic policy gradient (MADDPG) algorithm. The MADDPG algorithm optimizes the PCAM strategy through the continuous interaction of multiple independent agents and effectively captures the inherent uncertainty in continuous learning and human behavior. Experimental results show that the MADDPG model can fully mitigate collisions in different scenarios and outperforms the DDPG and DRL algorithms.

Keywords: autonomous–manual vehicle; multi-agent; intersection risk; driving behavior

Citation: Hu, W.; Mu, H.; Chen, Y.; Liu, Y.; Li, X. Modeling Interactions of Autonomous/Manual Vehicles and Pedestrians with a Multi-Agent Deep Deterministic Policy Gradient. *Sustainability* **2023**, *15*, 6156. https://doi.org/10.3390/su15076156

Academic Editors: Quan Yuan, Cong Chen, Weiwei Qi and Tao Wang

Received: 2 March 2023
Revised: 25 March 2023
Accepted: 30 March 2023
Published: 3 April 2023

1. Introduction

According to recent statistics, the worldwide annual average of traffic accident fatalities is approximately 1 million. The majority of these reported accidents involve cars and pedestrians [1]. Meanwhile, pedestrians are vulnerable in traffic, especially when crossing a street where other participants (drivers and cyclists) are moving. Not only are they in danger themselves, but they can also cause accidents amongst other participants as they might dodge or break abruptly. Furthermore, a recent research board on traffic safety reported that the important factors of a dangerous crossing include the distraction for pedestrian [2], the type of street [3], the implementation of the autonomous vehicle [4], etc. Of late, a new element that can make traffic safer is the implementation of the autonomous vehicle (AV). The cars are computer-operated, instead of having a human driver. The computer follows the valid traffic laws, and different sensors can estimate the distance to other vehicles or pedestrians. AVs thus stop as soon as something approaches them that could cause a collision. Since the presence of AVs in our everyday traffic situations can positively affect traffic safety, it is of importance to study this pedestrian–vehicle interaction [4]. To do so, this study extensively tests the effect of AVs on pedestrian crossing behavior. Autonomous vehicles have the potential to reduce traffic congestion and emissions by optimizing routes and reducing idle time. By incentivizing the adoption of autonomous vehicles through policy or financial means, sustainability can be improved. By using intelligent transportation systems that optimize traffic flow, such as signal timing and

route optimization, traffic congestion can be reduced, leading to improved air quality and reduced emissions. Pedestrians are vulnerable road users who need to be protected. By implementing infrastructure changes such as pedestrian crossings, sidewalks, and bike lanes, pedestrian safety can be improved.

In order to reduce the risk of traffic accidents and traffic collisions, AVs are configured to have a PCAM system, which has the ability of path awareness planning [5]. This method can only describe pedestrian movement through dynamic planning, ignoring the interaction between the AV and the pedestrian, so it lacks objectivity in the real world [6]. With the development of neural networks, RL and neural networks complement each other's advantages. Through the interaction of AV action and pedestrian action, there is no need for prior domain knowledge, which effectively improves the generalization performance of AV. This field is gradually becoming a research hotspot [7]. In an intelligent vehicle–pedestrian reaction environment, there are three key agents: AVs, manual vehicles (MVs), and pedestrians. In the MA (multi-agent) scenario, the operation of an agent affects the state of other agents. Recently, relevant scholars proposed an MA-DRL method to achieve multi-agent joint training through the combination of DNN and MA-DRL. The success of MA-DRL depends on the availability of high computing power and the effect of large-scale neural networks, such as DeepMind [8] and OpenAI [9]. However, the previous work only used an easy AV–pedestrian model but did not analyze the risk of pedestrians crossing the road. As a variable that will exist for a long time in the development of intelligent intersections, manual vehicles are often ignored. As a result, the existing algorithm cannot reflect the actual intelligent intersection interaction and risk status.

Multi-agent deep deterministic policy gradient (MADDPG) is a reinforcement learning algorithm that can be utilized to address challenges in multi-agent systems, where several agents interact with both their environment and each other. Proximal constrained actor–critic with multiple critics (PCAM) is an improvement of the actor–critic algorithm that incorporates multiple critics to improve the stability and robustness of the learning process. The application of MADDPG to PCAM without signals refers to using these algorithms to solve multi-agent problems where there is no explicit communication or signaling between the agents. In such cases, agents must learn how to cooperate or compete independently based on information they can observe from the environment.

For instance, imagine a scenario where autonomous vehicles are navigating a busy intersection without traffic lights. Each vehicle has to decide whether to stop, yield to other cars, or proceed through the intersection based solely on the position, speed, and direction of nearby vehicles, without explicit communication between them. To apply MADDPG to PCAM without signals, we would train each agent using the multi-agent version of the deep deterministic policy gradient (DDPG) algorithm, which is another variant of the actor–critic algorithm.

The agents would learn to select actions based on their observations of the environment, as well as those of other agents nearby. The multiple critics in PCAM would then provide feedback to the agents regarding the quality of their actions, facilitating improved learning stability and robustness. Ultimately, the combination of MADDPG and PCAM without signals proves to be a powerful approach for solving multi-agent problems where explicit communication or signaling is not feasible or desirable.

Therefore, the objective of this paper is to better understand pedestrians' crossing behavior by developing an AV PCAM system through the MADDPG algorithm that exploits the continued interaction of three independent agents:, AVs, MVs, and pedestrians. In this approach, the PCAM system is updated, and the pedestrian can cross the street fast and safely. To sum up, our contributions in this paper are as follows:

We studied the interaction between MVs and AVs before entering the intersection, which improved the accuracy of risk prediction of the pedestrian–vehicle interaction.

Our method is extended when the initial TTC (time to collision) value, street width, and pedestrian walking speed are different.

Our MADDPG algorithm is compared with the intelligent traffic control system [10] and AV system [11]. The experimental results show that our model is ahead of the current algorithms and uses the existing methods with different evaluation indicators.

2. The Literature

2.1. Interactions of Autonomous Vehicles and Pedestrians

An innovation that has been studied a lot but is still not part of our public road network is the autonomous vehicle (AV). These cars do not have a human who drives them or controls the gas and brake and thus looks out for other people on the road; instead, a computer system must estimate the situation and adapt their speed to their surroundings. Most research on pedestrian and AV interaction has been to improve the computer of AVs. The more various studies you do on human behavior, the more elaborated your computer will be for different scenarios. The social side of AVs and pedestrians has been modeled by Gupta et al. based on three challenges regarding self-driving vehicles [12]. The first challenge is intent perception, in which pedestrians intend to try to cross the road and during this operation engage with car drivers [12]. Sobrinho et al. [13] evaluated public streets with vehicles, pedestrians, lights, and noise. Through multivariate analysis, the author established that using smart phones while walking may pose risks to pedestrians, so it is necessary to perceive pedestrian intentions. The second one is about social behavior in traffic zones, where specific social rules are shared. The agreement in negotiation is the last challenge, which is an important aspect of social behavior in traffic. It is about reaching an agreement with other road users to get the right of way. By modeling this interaction, an improved algorithm can be applied to the computer that drives the vehicle. Andreai et al. [14] proposed a method to simulate the interaction between people and vehicles to make decisions through the traffic volume, service level, driver's compliance with the right of way of pedestrians on the zebra crossing, age-driven pedestrian crossing behavior risk, and decision-making. Florentine et al. [15] proposed a method to convey perceptual information by equipping autonomous vehicle with light-emitting diode (LED) lights, build trust and participation with pedestrians through LEDs, and improve the passing of autonomous vehicles through a street to avoid collisions. The pedestrian–AV interaction has been simplified by Millard-Ball, in his game theory [16]. The game of cross-walk chicken as the author calls it, is about two drivers moving toward each other at full-speed and neither of them wants to yield or get in a crash. This would mean there is no solution because one of those scenarios is going to happen anyway. This has been applied to the interaction between a pedestrian, who wants to cross the street in the same walking flow as the rest of their walk, and the driver, who does not want to yield or cut their speed. When applying this to AVs, the situation changes; they are programmed to avoid collisions. A pedestrian thus does not have to stop; they will not get hit anyway [16]. To specifically study the effect of AVs on pedestrians' wait time, this vehicle type was included in our experiment.

2.2. Interactions of Autonomous/Manual Vehicles and Pedestrians

In previous studies, researchers used several traffic safety analysis techniques [17]: simultaneous equation model, negative binomial model, random-effects ordered logic model, ordered probability model, random-effects negative binomial model, and Bayesian hierarchical binomial logic model. Recent methods for analyzing the severity of accident injuries [18] include the generalized ordered method, zero expansion model, fractional segmentation method, copula method, and panel hybrid method. Logistic regression [19] can measure correlation and control the effect of mixed variables and has been widely used in previous pedestrian crash analysis studies to determine the correlation between injury severity and contributing factors, to identify risk factors related to fatal crashes. Camara et al. [20] put forward a game theory method in which vehicles and pedestrians pass the crosswalk to study the joint behavior of pedestrians and drivers from the perspective of safety, complete the decision through the expectation maximization algorithm, and reduce the conflict risk of the crosswalk. Wu et al. [21] proposed trajectory planning and CAV

control methods to achieve effective vehicle passing. However, the researchers measured the correlation between vehicles and pedestrians through conventional methods, such as the respective characteristics of vehicles and pedestrians, analyzed the collision risk by controlling hybrid variables, and did not consider the automatic method of mining the correlation of hidden factors.

2.3. The Deep Reinforcement Learning Method

Several deep reinforcement learning methods are used in modeling the interaction of vehicles and pedestrians based on conflict and cooperation. Chae et al. [22] first proposed a DRL-based PCAM system, which effectively avoids collisions, but the AV agent only contains simple tasks, ignoring complex actions in real scenes. Inspired by Chae et al. [22], Papini et al. [23] proposed a DRL system based on which pedestrians can safely cross the road by learning the AV's speed, but there will still be some collisions. In [24], a grid-based state representation model is proposed. The state representation supports multiple AV and pedestrian operations. Although the trained model is evaluated in CARLA, the model can control multiple agents; however, the impact of uncertainties such as pedestrian distribution and behavior in the scene has not been reflected, so the model cannot run in the real world. In order to solve the challenges in our work, we fully studied the impact of uncertainty factors on multi-agent agents. Deshpande et al. [25] proposed that AVs realize the decision of crossing the road with pedestrians through navigation, rather than interacting with pedestrians through multiple objectives and multiple factors. A general summary of DRL can be discovered in [26]. During such DRL, the DDPG is wildly used in multi-agent modeling. Vasquez et al. [27] proposed a deep reinforcement learning–based multi-objective autonomous braking system that is a continuous action space that seeks to maximize pedestrian safety and perception. Then, the MADDPG algorithm is applicable to the deep reinforcement learning algorithm as single-agent DDPG is extended to a multi-agent system. There are methods to apply the single-agent algorithm to the multi-agent field, but the strategy of the agent only depends on its own state, which will cause the instability of the state transition probability of the agent, and there are some defects such as large variance of the value function. Wu et al. proposed to coordinate each agent based on the MADDPG network in the traffic light control scenario in the vehicle network to alleviate the problem of poor learning performance caused by an unstable external environment [28]. However, these methods consider an all-autonomous, cooperative agent environment without considering the PCAM decision policy.

To fill the gap above, in this paper, we develop a new-strategy PCAM algorithm to simulate the interaction between driverless vehicles and pedestrians on unsignalized crosswalks, which features a priority-based safety supervisor, parameter sharing, and local reward shaping. Performance comparisons between the proposed algorithm and the above benchmarks are presented in Sections 3 and 4.

3. Methods

The PCAM system is established in a simulated cockpit of an AV and an MV facing a single pedestrian at a no-signal crossroads. It is notable that a large number of accidents are reported at no-signal crosswalks [29].

In this paper, other road users are neglected, and there is no priority given to the pedestrian. The agents are described as follows:

AV: According to a system developed by SAE International, an AV is usually divided into six levels, Level 0 represents no automation, Level 1 represents shared control support, Level 2 represents partial automation, Level 3 represents conditional automation, Level 4 represents high automation, Level 5 represents full automation. In this paper, the AV is fully autonomous in its driving capabilities (Level 5). In this paper, the AV's behavior follows the rule of the AV model, which decided based on several variables, such as following car, TTC, road environment, etc. (for more details, see Section 3.3).

MV: The car is controlled by the driver, it can also be presented as Level 0—no automation. In this paper, the MV's behavior is described as natural driving behavior, which can be influenced by several variables, such as driver characteristics, road environment, surrounding cars, etc. (for more details, see Section 3.2).

Pedestrian: The pedestrian's action is mainly to cross the road from the left or right sidewalks. Due to the change of human consciousness, the conditions for setting the pedestrian state are limited.

No-signal crosswalk scenario: In this study, the 200 m before entering the unsignalized intersection is selected as the simulation scene. The road is three-lane, and the vehicle speed limit is 50 km/h. The MV and AV are randomly generated and interact before entering the intersection. The reward of the scenario is calculated based on the interaction between vehicle and pedestrian as shown in Figure 1. After $T \in N$ time steps, one vehicle–pedestrian interaction episode ends. In this scenario, the vehicle position is x_v, and x_{pe} is the pedestrian's position. When an epoch starts, the AV and MV are generated randomly at the start of road, each vehicle is facing the crosswalk in front. The vehicles' first position is randomly chosen as one of three lanes. The width of the three-lane street is w_{stteet}. The pedestrians cross the crosswalk at a certain speed v_{walk}^{pe}; the speed of the AV is v_{walk}^{av}; and the speed of the MV is v_{walk}^{mv}. The initial distance of pedestrian from the curb is ζ^{pe}. When the vehicle has passed the crosswalk by a distance of ζ^{av}, the vehicle's goal position x^{pe} is reached. A collision is defined as the collision between vehicles and pedestrians in a certain area, which is usually the vehicle boundary and additional safety margin η. The inequality is described as follows:

$$\eta > x^{av} - x^{pe} \,\&\, \eta > x^{mv} - x^{pe} \tag{1}$$

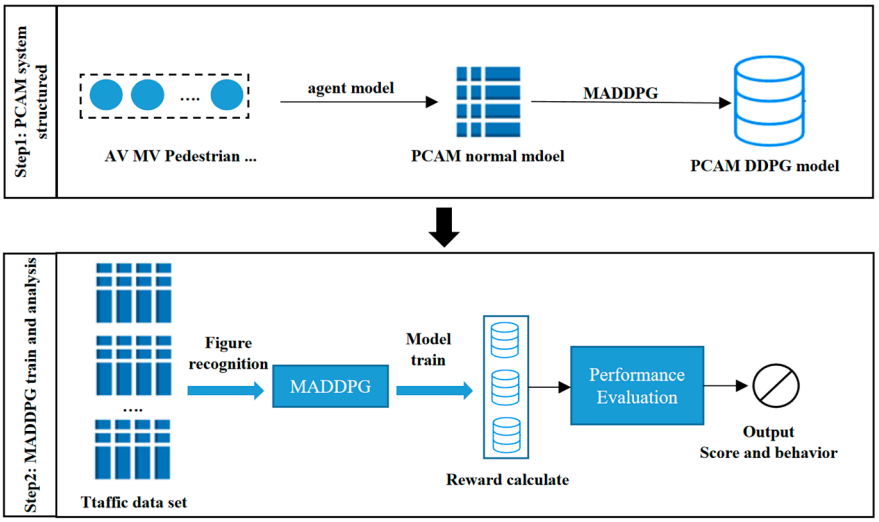

Figure 1. The flow chart.

For simplicity, TTC value is calculated from the vehicle's center point to the pedestrian's position, i.e., the main performance index is the collision rate, because in reality, collision between vehicles and pedestrians is not allowed. At this time, the TTC value can also be greater than 0. The goals of each agent are the same. They are all to reach the target location as soon as possible. The second is the traffic flow performance index, which represents the average time of the vehicle–pedestrian interaction. This utility function is described as follows:

$$F_w = -T_{end}^{t} \tag{2}$$

Each agent $W = \{w^{av}, w^{mv}, w^{pe}\}$. t_{end}^{av}, t_{end}^{mv}, and t_{end}^{pe} represent the time required for each agent to reach its target location. The final objective function is obtaining large traffic flow benefits and the shortest combined travel times without collision.

The details of the model of each agent are as follows.

3.1. Pedestrian Models

To simulate basic but rational human road crossing behavior, we defined a pedestrian strategy based on the *TTC* estimation at each moment, according to

$$Agent_{t}^{pe} = \begin{cases} walk, \text{ if } TTC > 3 \text{ s} \\ wait, \text{ if } TTC \leqq 3 \text{ s} \end{cases} \tag{3}$$

When pedestrians decide to walk, they will maintain the average speed to reach the target position of pedestrians. In this paper, the pedestrian's velocity is reflected by selecting walk consistently from $v_{walk}^{pe} \in \{1.15, 1.39, 1.48, 1.54, 1.56\}$ m/s, it represents typical pedestrian walking speeds. Note that the agent will also start walking when the vehicle has passed the crossing by 4 m [6]. The speed of the pedestrian crossing the road at time t, with a velocity v_{t+1}^{pe} at the next moment is v_{walk}^{pe}.

3.2. MV Models

As a part of the multi-agent model, the MV model can realize interactive training of other agents through the MADDPG algorithm. The MV's state at time t is presented as s_{t}^{mv}. The s_{t}^{mv} changes according to the vehicle's position (leading or following). The models are shown as follows:

$$s_{t}^{mv} = \begin{cases} \begin{bmatrix} TTC_t, |v_t^{lv}|, lv_{type}, |v_t^{mv}|, |a_t^{mv}|, R_{time}, RE, \\ HF, \Delta x_t^{rel}, L_{num}, F_t \end{bmatrix} & if\ is\ following \\ \begin{bmatrix} TTC_t, |v_t^{pe}|, |v_{walk}^{pe}|, |v_t^{mv}|, |a_t^{mv}|, R_{time}, RE, HF, \\ \Delta x_t^{rel}, L_{num}, PDTC_t, b_{str}, b_{s-str} \end{bmatrix} & if\ is\ not\ following \end{cases} \tag{4}$$

TTC_t is the current *TTC* value.

$|v_t^{lv}|$ is the current speed of the leading vehicle-

lv_{type} is the type of leading vehicle: AV or MV.

$|v_t^{mv}|$ is the current speed of the target manual vehicle.

$|a_t^{mv}|$ is the current acceleration of the target manual vehicle. The acceleration parameter list a_t^{mv} is granted by the action behavior $U_{vehicle}$ as $\{-9.8, -5.8, -3.8, 0, 1, 3\}$ m/s^2.

R_{time} is the drivers' reaction time; in this paper, the reaction time is set to 3.5 s [30].

RE is the road environment's influence; in this paper, the *RE* is set to 1, which means the road environment has no influence on the driver.

HF is the human factor influence such as gender and age. In this paper, the *HF* is also set to 1, which means the driver has good situational awareness when driving the vehicle.

Δx_t^{rel} is the relative distance between the vehicle and the other target (leading vehicle or pedestrian) at the current time.

L_{num} is the current number of lanes that the vehicle is in.

F_t is the following time of vehicle.

$|v_t^{pe}|$ is the current speed of the pedestrian.

$|v_{walk}^{pe}|$ ensures that when the pedestrian intelligent agent wants to walk, the pedestrian speed remains unchanged.

$PDTC_t$ is the remaining distance for pedestrians to reach the target position.

b_{str} is the width of the street.

b_{s-str} is to guide the pedestrian on which side to start walking. It is randomly chosen from {left, right}.

3.3. AV Models

The MV's state at time t is presented as s_t^{av}. The s_t^{av} has different models that change according to the vehicle's position. The models are shown as follows:

$$s_t^{av} = \begin{cases} \begin{bmatrix} TTC_t, |v_t^{lv}|, lv_{type}, |v_t^{av}|, |a_t^{av}|, \\ RE, \Delta x_t^{rel}, L_{num}, F_t \end{bmatrix} \; if\ is\ following \\ \begin{bmatrix} TTC_t, |v_t^{pe}|, |v_{walk}^{pe}|, |v_t^{av}|, |a_t^{av}|, RE, \Delta x_t^{rel}, \\ L_{num}, PDTC_t, b_{str}, b_{s-str} \end{bmatrix} \; if\ is\ not\ following \end{cases} \tag{5}$$

Most components have the same description as mentioned above. The different variables are as follows:

$|v_t^{av}|$ represents the recent speed of the AV.

$|a_t^{av}|$ represents the recent acceleration of the AV, which is also given by the action space $U_{vehicle}$ with $\{-9.8, -5.8, -3.8, 0, 1, 3\}$ m/s^2.

3.4. DDPG and MADDPG

Deep reinforcement learning (DRL) is a subfield of machine learning that combines reinforcement learning with deep neural networks to enable agents to learn from high-dimensional and complex input spaces. The basic elements of DRL include the following:

Agent: The agent is the learner that interacts with the environment and takes actions based on its policy. The policy is a function that maps observations to actions.

Environment: The environment is the external system that the agent interacts with and from which it receives observations and rewards. The environment is typically modeled as a Markov decision process (MDP), where the current state depends only on the previous state and the current action.

Reward: The reward is a scalar value that the agent receives from the environment after taking an action. The reward indicates how good or bad the action was in terms of achieving the agent's goals.

State: The state is a representation of the environment that the agent observes. The state can be a high-dimensional sensory input, such as an image or a sound, or it can be a lower-dimensional representation of the environment.

Action: The action is the output of the agent's policy that it takes in response to the observed state. The action can be discrete, such as moving left or right, or continuous, such as the speed and direction of movement.

Policy: The policy is the function that maps observations to actions. The policy can be represented by a neural network that takes the state as input and outputs the action.

Value function: The value function is an estimate of the expected future reward that the agent can obtain from a given state or state–action pair. The value function is used to guide the agent's learning by estimating the quality of its actions.

In DRL, the agent learns to optimize its policy by interacting with the environment, receiving rewards, and updating its neural network parameters based on the observed state and the estimated value function. The goal is to learn a policy that maximizes the expected cumulative reward over time.

To optimize the multi-agent policy, we used MADDPG to model the agent formula and simulate the intersection scenario. The MADDPG is an evolution of the depth deterministic strategy gradient (DDPG) algorithm, mainly including the actor and critic control method. This method is composed of two sub-modules. One is to predict the action to be taken at the next moment according to the state of the previous moment, and the other is to calculate the expected return based on the prediction results.

The actor network and the critic network of the DDPG is composed of the online network and the target network. The original actor network is the state as input, and the output is the action executed at the current time; the input of the target actor network is the state at the next time, and the output is the action to be executed at the next time. The input of the original evaluation network is the state and action at the current time, the output is

the desired reward (state action value function) obtained by executing the current action at the current time; the input of the target critic network is the changed state and action at the next time, and the output is the state action value function at the next time. Then, the critic network uses the TD difference between the current Q value of the agent and the next Q value to update. The original actor network uses the learned Q value to update, while the parameters of the target actor network are obtained by copying the updated original actor network parameters after a certain time step. DDPG is an off-policy algorithm, which samples from the experience buffer pool. The experience buffer pool stores the historical game track of each step, and the strategy track is a quad (s_t, a_t, r_t, s_{t+1}), that is, the state, action, reward, and $t+1$ state of the agent at time t.

The MADDPG adopts a framework of centralized training and decentralized execution. The observation information of all agents in this framework trains one or more centralized critic networks, and each agent has its own actor network. The input of each actor network is the observation value of each agent, and the output is the action performed by each agent. In the test phase, the strategies executed by a single agent no longer depend on the observations of other agents. This framework avoids the challenges of non-Markov and non-stationary environments in the learning process and reduces the variance of the value function. For each agent, the algorithm essentially still uses the DDPG algorithm.

For the actor network, the inputs of the DDPG algorithm and the MADDPG algorithm are the state of the current agent, and the action performed by the current agent is the output result. For critic network, the DDPG algorithm takes the state and action of the current agent as input, while the input of MADDPG algorithm is the state and action of all agents. In the MADDPG algorithm, each agent also uses the car's track s_{it} sampled in the experience buffer pool to train the actor network and critic network, where i, t represent the state of agent i at time t, a_{it} represents that agent i adopts strategy $\pi_i(a_{it}|s_{it})$ at time t, r_{it} represents the rewards obtained when agent i takes action at time t, s_{it+1} represents the state of agent i at time $t+1$. Actor network is θ parameterized as $\theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$,

In the multi-agent system, for the cooperative environment, the rewards obtained by the team are easy to get. Therefore, compared with the MADDPG algorithm, each agent has its own critic Network, and the evaluation network is centralized because all agents share the same critic network. It adopts a framework of centralized training and implementation. Each agent in the system has its own actor network, but all agents have a common critic network, as shown in Figure 2. The actor network generates corresponding actions according to the current state of agent i, and the critic network evaluates the expected benefits obtained by all agents executing the current joint actions. The gradient update formula for the actor network is as follows:

$$\nabla_{\theta_i} J(\theta_i) = E_{s_{i,t}\sim p, a_{i,t}\sim\pi_i}\left[\nabla_{\theta_i} \log \pi_i(a_{i,t} \mid s_{i,t}) Q_i^{\pi}(x, a_{1,t}, \cdots, a_{n,t})\right] \tag{6}$$

The gradient update formula for the critic network is as follows:

$$\begin{aligned} L(\theta_i) &= E_{x,a,r,x}\left[\left(Q_i^{\pi}(x, a_{1,t}, \cdots, a_{n,t}) - y\right)^2\right] \\ y &= r_{i,t} + \gamma Q_i^{\pi'}(x', a_{1,t+1}, \cdots, a_{n,t+1}) \end{aligned} \tag{7}$$

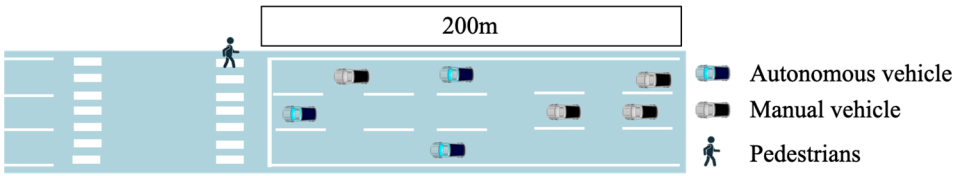

Figure 2. The no-signal crosswalk scenario.

The specific update process is shown in the Figure 3.

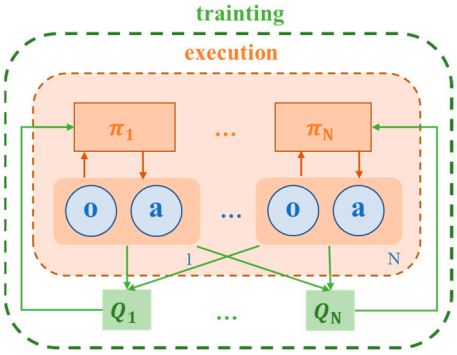

Figure 3. The update process of MADDPG.

The MADDPG has a training target based on the reward function R^t with the reward r_t^{ve-pe} is based on

$$R^t = r_t^{pe} + r_t^{ve} \tag{8}$$

where r_t^{pe} is the reward of the pedestrian, which can be calculated based on the formula

$$r_t^{pe} = -a - \begin{cases} \delta \text{ if is collision} \\ 0 \text{ otherwise} \end{cases} \tag{9}$$

where a represents punishment each time a step is taken, and usually we set $a = 0.01$. If a collision occurs, a penalty of $\delta = 10$ is added. The pedestrian agent's strategy needs to keep the two penalty conditions relatively balanced, but the absolute value is set according to the real scene. In any case, the goal of the pedestrian is to cross the road as soon as possible without collision.

Furthermore, r_t^{ve} is the reward of the vehicle, which can be calculated based on the formula

$$r_t^{ve} = -b - \begin{cases} \partial \text{ if is collision} \\ 0 \text{ otherwise} \end{cases} - \begin{cases} \varnothing \text{ if is speeding} \\ 0 \text{ otherwise} \end{cases} \tag{10}$$

For each epoch, the constant penalty is defined as $b = 0.01$; $\partial = 10$ is the collision penalty. The speed penalty $\varnothing = 0.05$ means that when the AV's speed exceeds the speed limit sign, the vehicle should learn to obey the traffic rules, but the speed limit can also be lifted in case of emergency.

3.5. Simulation Setup

The road length was set to 200 m, all of the vehicles' size was set as 2 m. The initial velocity of the vehicle (MV and AV) was sampled from a uniform distribution ∼{18, 50} km/s, representing the standard speed of urban roads in China. The initial TTC value was randomly sampled from ∼(1.0, 5.0) s. The goal state of the vehicle goal was reached when the vehicle passed the crosswalk by $\zeta^{av} = 10$ m. The variability of the pedestrians' walking speed was 0.5 m/s, and a value of 0.5 m was used for the safety margin ζ^{ped}. Generally, the street width is selected as $\{6.0, 7.5\}$ m, and other environmental uncertainties are also introduced. A collision will occur when the inequality (1) is satisfied, the collision margin was defined as: $\eta = 0.5$ m.

When it began to train, the super parameter setting were as follows: The discount factor was set to 0.8; the minibatch size was 128; the update factor was set to 0.9; the epoch was 300; the learning rate was 0.0001; the learning rate attenuation was set to 0.001; the hidden layer node was 64; and the regularization mode was chosen as Adam.

4. Results and Discussion

In order to prove our MADDPG algorithm, all multi-agents went through more than 8000 episodes, of which 800 episodes were used for exploration, and the first 250 episodes were randomly selected. This paper used the vehicle detection dataset as recommended to

build a traffic scenario, both experiments in this paper did not have real images. The detection dataset is a dataset of 20,000 images of vehicles in different environments, captured from ground-based platforms; these are available for download at http://www.gti.ssr.upm.es/data/Vehicle_database.html (accessed on 18 December 2022). Reward is an important evaluation index for reinforcement learning. It can not only continuously optimize the strategy but also reflect the degree to which the agent completes the goal. Therefore, to verify the accuracy of our agent policy model, we chose reward as the evaluation index and chose three environments to conduct comparative experiments on the model. For this paper, we selected DRL, DDPG, and MADDPG as the baseline model and recorded the reward during the model training phase. The results in Figure 4 show that under the same goal, the reward score of the agent MADDPG model was superior to that of DDPG and DRL. Moreover, we calculated the fluctuation of the reward. The fluctuation of the reward score of the DRL model was higher than that of DDPG and MADDPG. This means that MADDPG had the best performance among the models.

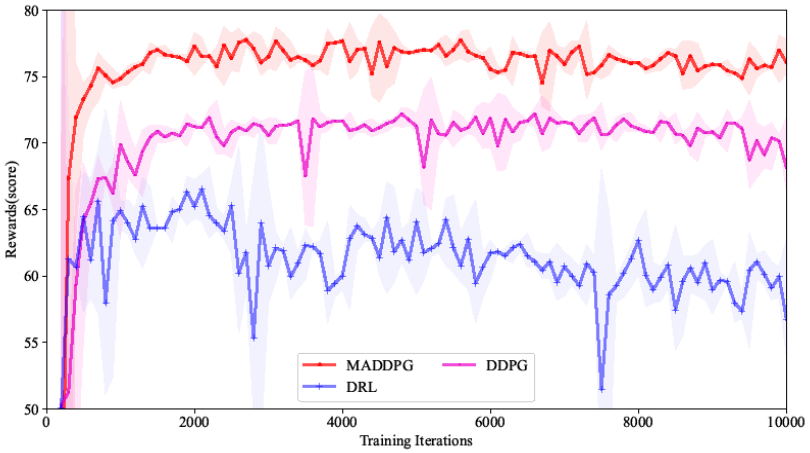

Figure 4. Training total reward score.

4.1. Performance Evaluation

For the evaluation process, we selected eight different test results, and their deviation was expressed in the form of 80% confidence interval from 10% quantile to 90% quantile. The results of the MADDPG algorithm are shown in Figure 5. When the pedestrian risk behavior is determined, the MV has the ability not to collide with pedestrians. The collision rate was 0.0% when α_{pe} = 0.0, 0.1, and 0.2. When the risk is more uncertain, the probability of collision is greater. Notably, at α_{pe} = 0.5, the MV could still avoid collision when the behavioral risk was extremely uncertain. When noise increased, pedestrians could also pass the road smoothly, thus reducing waiting time. The duration increased from 4.46 s to 5.51 s, an increase of about 23%.

As shown in Figure 6, our MADDPG algorithm has the ability to make vehicles and pedestrians quickly cross the road. It is obvious that, when the agent strategy was in a lower risk scenario, namely α_{pe}= 0.0 and 0.1, AV vehicles completely avoided collision. At α_{pe} = 0.3, the highest median collision rate was obtained, and the collision rate was 0.123%. When α_{pe} = 0.3, the collision rate was the highest, 0.200%, when the α_{pe} exceeded 0.3, although there was high uncertainty in the scene, our agent strategy made correct response to avoid collision. From the perspective of AV, this agent strategy can adapt to complex scenes rather than simply determine the scene.

To compare the proposed method with other papers [31,32], we use the hidden Markov model, which guarantees the accuracy of long-term prediction. At the same time, we used a human behavior sequence model, which can ensure the consistent segmentation of continuous points in behavioral actions. Our model can ensure that vehicles can pass the intersection in the shortest time, that pedestrians can also safely pass the intersection, and that better performance and accuracy are provided.

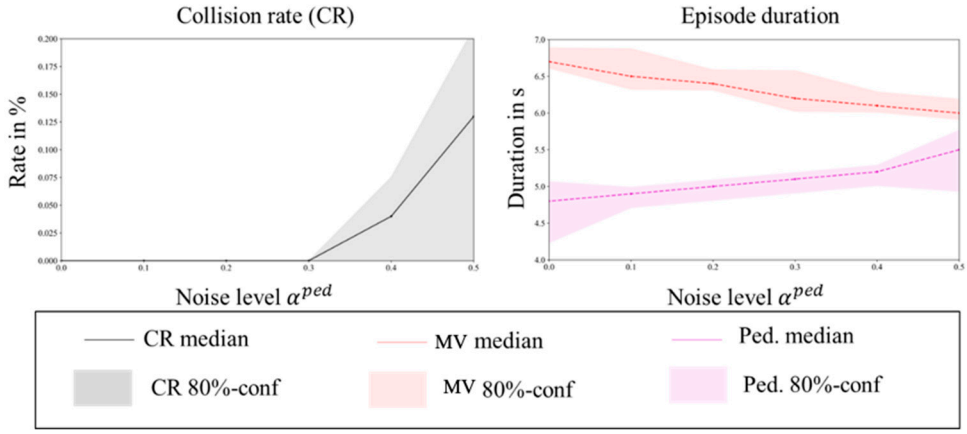

Figure 5. Training total reward score of MV model.

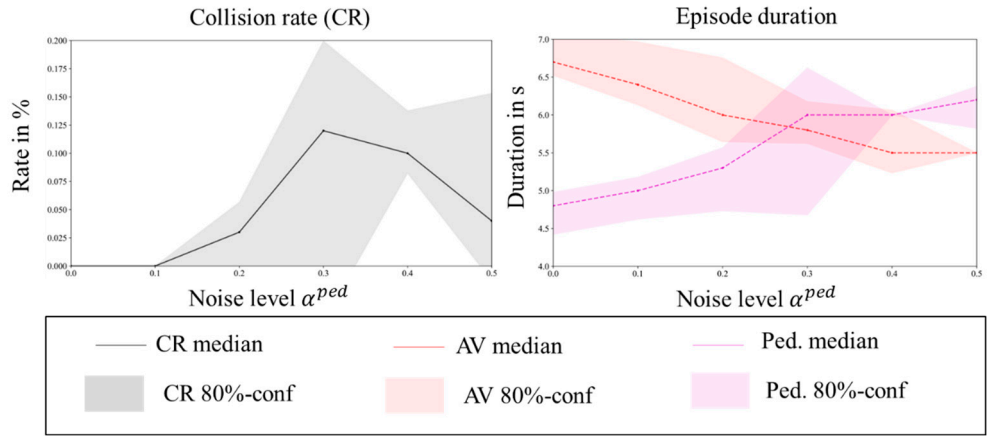

Figure 6. Training total reward score based on AV model.

4.2. Behavior Analysis

In this subsection, we attempt to interpret the learned AV behaviors. For example, after m times of simulation test, we counted the vehicle AV1 and AV2 speed changes. Figure 7 shows the snapshots of distances as well as the speeds of agents 1–2. It can be observed that, at a distance of 50 m, vehicle1 and vehicle2 start to slow down; vehicle 2 slowed to its lowest speed at a distance of 87 m, and vehicle 1 slowed to its lowest speed at a distance of 126 m. Then vehicle1 and vehicle2 accelerated slowly. Finally, vehicle1 and vehicle2 decelerated again 10 m before the crosswalk.

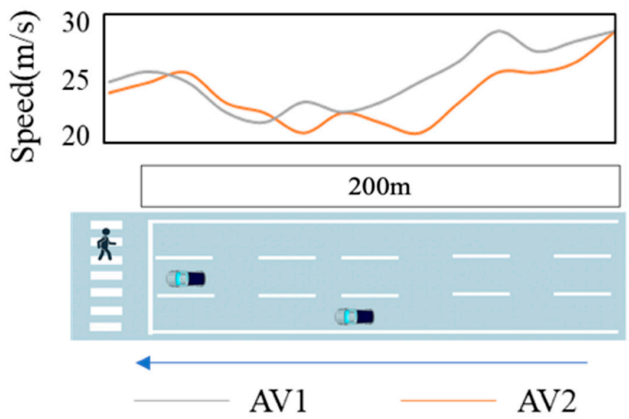

Figure 7. AV1 and AV2 speed change during AV–pedestrian interaction.

After training, the autopilot model learns a strategy that effectively reduces event duration while maintaining speed limits and minimizing collisions. We carried out m simulations for our model. During the simulation of the vehicle–pedestrian interaction, as shown in Figure 8, since the scene initialization TTC value was 4.5 s, the learning AV model accelerated slightly at the beginning of the event. When TTC < 3 s, the autonomous vehicle started to reduce its speed constantly, and when the pedestrian was facing the front, the vehicle almost stopped.

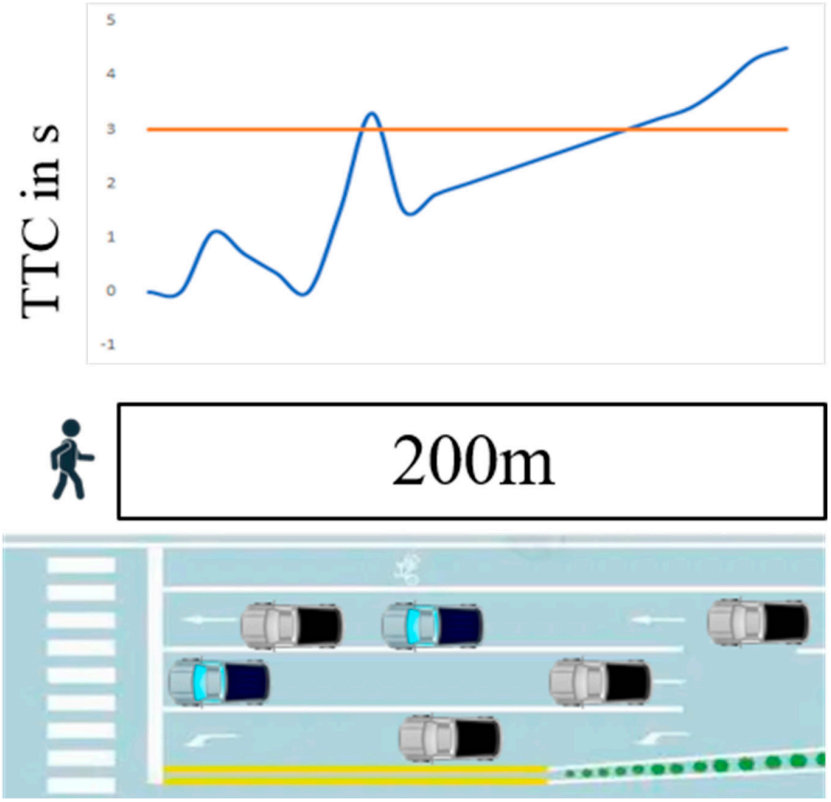

Figure 8. The corresponding TTC during AV–pedestrian interaction.

5. Conclusions

This paper proposes a new AV PCAM system, through the solution of MADDPG based on DMARL, which utilizes the continuous interaction of three independent agent policy that observes the global critic to guide actor training. In this paper, we studied the interaction between MVs and AVs before entering the intersection, which improved the accuracy of risk prediction for the pedestrian–vehicle interaction. Then, the proposed method was generalized to different complex scenes when there were large differences in initial TTC values and the basic information of vehicles and pedestrians. The results show that the MADDPG-based method achieved robust performance in the autonomous/manual vehicles and pedestrian scenario, although there were many uncertainties in the scenario. Moreover, under the same goal, the performance of the agent MADDPG algorithm was better than that of DDPG and DRL.

The Innovative contributions of the paper are as follows:

Proposing a novel approach that models the interactions between autonomous/manual vehicles and pedestrians using a multi-agent deep deterministic policy gradient (MADDPG) algorithm.

Introducing a new state representation that includes the relative position, velocity, and acceleration of all agents, as well as the distance to the nearest intersection.

Developing a three-agent policy that accounts for the interactions between autonomous vehicles, manual vehicles, and pedestrians and enables them to cooperate in navigating through complex urban scenarios.

Conducting comprehensive experiments that demonstrate the effectiveness of the proposed approach in various scenarios, such as unsignalized intersections, signalized intersections, and pedestrian crossings.

Overall, the paper contributes to the field of autonomous vehicles and pedestrian safety by proposing a new approach that models the interactions between different agents in a more realistic and effective way.

In future work, we will analyze the behavior of pedestrians and vehicles in the simulated environment and the real scene to find similarities between the two scenes; our preliminary analysis shows that they have similar characteristics. At the same time, in order to further improve the performance of the PCAM system, we will use more complex reinforcement learning strategies and expand the scene to more complex road scenes. It may also be challenging to enhance distributed DMARL training strategies; then, we will also study the rationality of the overall simulation model, which is composed of pedestrians, AVs, and MVs, and compare it with real, complex environments to verify the effectiveness of the system.

Author Contributions: Conceptualization, W.H.; data curation, Y.L.; formal analysis, H.M.; investigation, H.M.; methodology, W.H.; software, W.H.; supervision, Y.C.; validation, H.M.; visualization, Y.L. and X.L.; writing—review and editing, Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China, grant number 2020YFB1600304.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Krul, I.; Nijman, S. Voetgangers op de SEH-afdeling Oorzaken en Risicogroepen. 2018. Available online: https://www.veiligheid.nl/sites/default/files/2022-04/Voetgangers%202018%20%281%29.pdf (accessed on 18 December 2022).
2. SWOV. Factsheet Voetgangers. SWOV. 2020. Available online: https://www.swov.nl/feiten-cijfers/factsheet/voetgangers (accessed on 18 December 2022).
3. Brosseau, M.; Zangenehpour, S.; Saunier, N.; Miranda-Moreno, L. The impact of waiting time and other factors on dangerous pedestrian crossings and violations at signalized intersections: A case study in Montreal. *Transp. Res. Part F Traffic Psychol. Behaviour.* **2013**, *21*, 159–172. [CrossRef]
4. Kalatian, A.; Farooq, B. Deepwait: Pedestrian wait time estimation in mixed traffic conditions using deep survival analysis. In Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference (ITSC), Auckland, New Zealand, 27–30 October 2019; pp. 2034–2039.
5. Schratter, M.; Hartmann, M.; Watzenig, D. Pedestrian collision avoidance system for autonomous vehicles. *SAE Int. J. Connect. Autom. Veh.* **2019**, *2*, 279–293. [CrossRef]
6. Trumpp, R.; Harald, B.; David, S. Modeling interactions of autonomous vehicles and pedestrians with deep multi-agent reinforcement learning for collision avoidance. In Proceedings of the 2022 IEEE Intelligent Vehicles Symposium (IV), Aachen, Germany, 4–9 June 2022.
7. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [CrossRef] [PubMed]
8. Vinyals, O.; Babuschkin, I.; Czarnecki, W.M.; Mathieu, M.; Dudzik, A.; Chung, J.; Choi, D.H.; Powell, R.; Ewalds, T.; Georgiev, P.; et al. Grandmaster level in starcraftII usingmulti-agentrein-forcement learning. *Nature* **2019**, *575*, 350–354. [CrossRef] [PubMed]
9. Bowen, B.; Ingmar, K.; Todor, M.; Yi, W.; Glenn, P.; Bob, M.; Igor, M. Emergent tool use from multi-agent autocurricula. *arXiv* **2019**, arXiv:1909.07528.
10. Guillen-Perez, A.; Cano, M. Intelligent IoT systems for traffic management: A practical application. *IET Intell. Transp. Syst.* **2021**, *15*, 273–285. [CrossRef]

11. Qian, X.; Altché, F.; Grégoire, J.; Fortelle, A. Autonomous intersection management systems: Criteria, implementation and evaluation. *IET Intell. Transp. Syst.* **2017**, *11*, 182–189. [CrossRef]
12. Gupta, S.; Vasardani, M.; Winter, S. Negotiation Between Vehicles and Pedestrians for the Right of Way at Intersections. *IEEE Trans. Intell. Transp. Syst.* **2018**, *20*, 888–899. [CrossRef]
13. Sobrinho-Junior, S.A.; de Almeida, A.C.N.; Ceabras, A.A.P.; da Silva Carvalho, C.L.; Lino, T.B.; Christofoletti, G. Risks of accidents caused by the use of smartphone by pedestrians are task-and environment-dependent. *Int. J. Environ. Res. Public Health* **2022**, *19*, 10320. [CrossRef] [PubMed]
14. Gorrini, A.; Crociani, L.; Vizzari, G.; Bandini, S. Observation results on pedestrian-vehicle interactions at non-signalized intersections towards simulation. *Transp. Res. Part F Traffic Psychol. Behav.* **2018**, *59*, 269–285. [CrossRef]
15. Florentine, E.; Ang, M.A.; Pendleton, S.D.; Andersen, H.; Ang, M.H., Jr. Pedestrian notification methods in autonomous vehicles for multi-class mobility-on-demand service. In Proceedings of the Fourth International Conference on Human Agent Interaction, Gothenberg, Sweden, 4–7 December 2016; pp. 387–392.
16. Millard-Ball, A. Pedestrians, Autonomous Vehicles, and Cities. *J. Plan. Educ. Res.* **2018**, *38*, 6–12. [CrossRef]
17. Mahmud, S.M.S.; Ferreira, L.; Hoque, M.S.; Tavassoli, A. Micro-simulation modelling for traffic safety: A review and potential application to heterogeneous traffic environment. *IATSS Res.* **2019**, *43*, 27–36. [CrossRef]
18. AlMamlook, R.E.; Kwayu, K.M.; Alkasisbeh, M.R.; Frefer, A.A. Comparison of machine learning algorithms for predicting traffic accident severity. In Proceedings of the 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), Amman, Jordan, 9–11 April 2019; pp. 272–276.
19. Zhao, J.; Malenje, J.O.; Tang, Y.; Han, Y. Gap acceptance probability model for pedestrians at unsignalized mid-block crosswalks based on logistic regression. *Accid. Anal. Prev.* **2019**, *129*, 76–83. [CrossRef] [PubMed]
20. Camara, F.; Romano, R.; Markkula, G.; Madigan, R.; Merat, N.; Fox, C. Empirical game theory of pedestrian interaction for autonomous vehicles. In Proceedings of the Measuring Behavior 2018, Manchester, UK, 5–8 June 2018; pp. 238–244.
21. Wu, J.; Qu, X. Intersection control with connected and automated vehicles: A review. *J. Intell. Connect. Veh.* **2022**, *5*, 260–269. [CrossRef]
22. Chae, H.; Kang, C.M.; Kim, B.; Kim, J.; Chung, C.C.; Choi, J.W. Autonomous braking system via deep reinforcement learning. In Proceedings of the IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), Yokohama, Japan, 16–19 October 2017.
23. Papini, G.P.R.; Plebe, A.; Da Lio, M.; Dona, R. A Reinforcement Learning Approach for Enacting Cautious Behaviours in Autonomous Driving System: Safe Speed Choice in the Interaction with Distracted Pedestrians. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 8805–8822. [CrossRef]
24. Deshpande, N.; Vaufreydaz, D.; Spalanzani, A. Behavioral decision-making for urban autonomous driving in the presence of pedestrians using deep recurrent Q-network. In Proceedings of the 16th International Conference on Control, Automation, Robotics and Vision (ICARCV), Shenzhen, China, 13–15 December 2020; pp. 428–433.
25. Deshpande, N.; Vaufreydaz, D.; Spalanzani, A. Navigation in urban environments amongst pedestrians using multi-objective deep reinforcement learning. In Proceedings of the 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), Indianapolis, IN, USA, 19–22 September 2021; pp. 923–928.
26. 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.* **2021**, *23*, 4909–4926. [CrossRef]
27. Vasquez, R.; Farooq, B. Multi-objective autonomous braking system using naturalistic dataset. In Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference (ITSC), Auckland, New Zealand, 27–30 October 2019; pp. 4348–4353.
28. Wu, T.; Zhou, P.; Liu, K.; Yuan, Y.; Wang, X.; Huang, H.; Wu, D.O. Multi-agent deep reinforcement learning for urban traffic light control in vehicular networks. *IEEE Trans. Veh. Technol.* **2020**, *69*, 8243–8256. [CrossRef]
29. Zegeer, C.; Stewart, J.R.; Huang, H.H.; Lagerwey, P.A.; Feaganes, J.; Campbell, B.J. *Safety Effects of Marked versus Unmarked Crosswalks at Uncontrolled Locations: Final Report and Recommended Guidelines*; U.S. Department of Transportation Federal Highway Administration: Washington, DC, USA, 2005.
30. Willis, A.; Gjersoe, N.; Havard, C.; Kerridge, J.; Kukla, R. Human Movement Behaviour in Urban Spaces: Implications for the Design and Modelling of Effective Pedestrian Environments. *Environ. Plan. B Plan. Des.* **2004**, *31*, 805–828. [CrossRef]
31. Gao, H.; Qin, Y.; Hu, C.; Liu, Y.; Li, K. An Interacting Multiple Model for Trajectory Prediction of Intelligent Vehicles in Typical Road Traffic Scenario. *IEEE Trans. Neural Networks Learn. Syst.* **2021**, 1–12. [CrossRef] [PubMed]
32. Gao, H.; Lv, C.; Zhang, T.; Zhao, H.; Jiang, L.; Zhou, J.; Liu, Y.; Huang, Y.; Han, C. A Structure Constraint Matrix Factorization Framework for Human Behavior Segmentation. *IEEE Trans. Cybern.* **2021**, *52*, 12978–12988. [CrossRef] [PubMed]