applied sciences

MDPI

*Article*

# Cooperative Decision-Making for Mixed Traffic at an Unsignalized Intersection Based on Multi-Agent Reinforcement Learning

**Huanbiao Zhuang, Chaofan Lei, Yuanhang Chen and Xiaojun Tan ***

School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen 518107, China
* Correspondence: tanxj@mail.sysu.edu.cn

**Abstract:** Despite rapid advances in vehicle intelligence and connectivity, there is still a significant period in mixed traffic where connected, automated vehicles and human-driven vehicles coexist. The behavioral uncertainty of human-driven vehicles makes decision-making a challenging task in an unsignalized intersection scenario. In this paper, a decentralized multi-agent proximal policy optimization (MAPPO) based on an attention representations algorithm (Attn-MAPPO) was developed to make joint decisions at an intersection to avoid collisions and cross the intersection effectively. To implement this framework, by exploiting the shared information, the system was modeled as a model-free, fully cooperative, multi-agent system. The vehicle employed an attention module to extract the most valuable information from its neighbors. Based on the observation and traffic rules, a joint policy was identified to work more cooperatively based on the trajectory prediction of all the vehicles. To facilitate the collaboration between the vehicles, a weighted reward assignment scheme was proposed to focus more on the vehicles approaching intersections. The results presented the advantages of the Attn-MAPPO framework and validated the effectiveness of the designed reward function. Ultimately, the comparative experiments were conducted to demonstrate that the proposed approach was more adaptive and generalized than the heuristic rule-based model, which revealed its great potential for reinforcement learning in the decision-making of autonomous driving.

**Keywords:** cooperative decision-making; connected and automated vehicles; multi-agent reinforcement learning; unsignalized intersection



**Citation:** Zhuang, H.; Lei, C.; Chen, Y.; Tan, X. Cooperative Decision-Making for Mixed Traffic at an Unsignalized Intersection Based on Multi-Agent Reinforcement Learning. *Appl. Sci.* **2023**, *13*, 5018. https://doi.org/10.3390/app13085018

Academic Editor: Rosario Pecora

Received: 10 March 2023
Revised: 5 April 2023
Accepted: 13 April 2023
Published: 17 April 2023



## 1. Introduction

In recent years, with the development of vehicle-to-everything (V2X) and sensor technologies, connected and automated vehicles (CAVs) received extensive attention regarding their ability to reduce the occurrence of traffic congestion and accidents [1]. Using environmental sensing sensors, such as cameras, lidar, and radar, the CAVs are able to obtain the state of roads and vehicles around them. The continuous development of V2X (vehicle-to-everything) technology has accelerated the development of the automobile intelligent network. V2X, including vehicle-to-vehicle (V2V), vehicle-to-infrastructure (V2I), and vehicle-to-pedestrians (V2P), enables each CAV to share information about the vehicle position, speed, acceleration, orientation, destination, etc. with any traffic participating entity, which enhances the range of perception for these vehicles [2]. Equipped with perception systems and vehicular communications, a CAV acquires a wider range of information, which is beneficial for cooperative driving to evade crashes and achieve better traffic performance [3].

According to the intelligent degree of vehicles, vehicles are divided into six levels, from L0 to L5, by the Society of Automotive Engineers (SAE). By 2030, 82 million L4/L5 intelligent vehicles will be in operation in China, the United States, and Europe [4]. Despite the dramatic advances in autonomous driving, the predictable transition from purely conventional vehicles to a purely intelligent and connected environment will require a

sustained investment in infrastructure and technology development. In the coming years, the transportation environment will exist in a transition stage of mixed transportation where the human-driven vehicles (HDVs) and CAVs coexist.

In general, significant research regarding cooperative decision-making and control is under the assumption that all the vehicles on the road are CAVs [5,6]. Cooperative decision-making and control in mixed traffic is a challenge in the field of intelligent transportation systems (ITSs) [7]. This is the reason why the CAVs need to interact with the HDVs, however the behavior of the HDVs is uncertain. To ensure a safe collaboration in mixed traffic environments, it is more practical to implement strategies for the CAVs that take the driver's behavior into account than expecting drivers to interact with the CAVs with caution. To imitate the human driving behavior, Treiber et al. developed an intelligent driver model (IDM) [8]. As a safe distance model, the IDM is able to describe the behavior of vehicles from free flow to congested flow with fewer parameters and reflects the dynamic changes of the vehicle position and speed in real time. Peng et al. [9] modeled the HDVs using the IDM, which were guided by the CAVs to avoid collision and congestion and achieved a significant traffic efficiency. Li et al. [10] extended the IDM using the Ornstein–Uhlenbeck process to describe the perceptual error dynamically. Wang et al. [11] proposed a model that combined a first principles nominal model with a Gaussian process model to predict human behaviors and then implemented a mixed platoon. However, the HDVs normally follow traffic rules to drive, rather than following the CAVs. Based on this view, the CAVs will anticipate the trajectories of the HDVs and cooperate to upgrade their safety and crossing efficiency in this paper.

An unsignalized intersection scenario does not have traffic lights to govern it. The existence of conflict areas at the lane interchanges and the changes in the vehicle driving behaviors can easily cause disorder, resulting in traffic safety accidents, such as rear-end collisions and traffic jams, which is where a large number of accidents occur [12]. In this paper, we assumed that all the vehicles, including the HDVs, will be given the right of way according to the traffic regulations. Based on this assumption, the CAVs learned to adapt to the HDVs and cooperate with them for safe and efficient crossing. Since multiple CAVs are required for cooperative driving, it is natural to adopt a multi-agent reinforcement learning framework to achieve the cooperative goal. The multi-agent proximal policy optimization (MAPPO) in a cooperative setting [13], which had a surprising effectiveness, was used as the benchmark algorithm in this paper. The original algorithm directly took the information from the agent's neighbors as its own observations. As the number of agents increased, learning became more difficult [14]. Therefore, an attention representation was utilized to select the most relevant information from the vehicle's neighbors. In the context of the cooperative multi-agent setting, the reward assignment problem concentrated on how to assign a global return to each agent that accurately reflected the agent's contribution to the overall behavior. We noted that the distance between a vehicle and the intersection had different effects on the safe operation of the system, and thus proposed a weight reward assignment scheme.

Therefore, to address the above problems in this paper, a decentralized MAPPO based on the attention representations (Attn-MAPPO) is proposed to make joint decisions at the intersection based on the trajectory prediction of all the vehicles. The contributions of this paper are fourfold.

First, decision-making at the intersection where the CAVs and HDVs coexist was formulated as a decentralized MARL problem. Based on the traffic rules and the most valuable neighbor information extracted by an attention module, the Attn-MAPPO algorithm was developed to allow vehicles to cross the intersection safely and effectively.

Secondly, a weighted reward assignment scheme was proposed. According to the position of the vehicles, the contribution of each CAV could be measured to enhance the cooperation between the vehicles.

Thirdly, an effective reward function was designed. Due to the uncertain behavior of the HDVs, all the vehicle's trajectories at every time step were predicted in a forward predictive time horizon, which reflected precisely how well the action was taken.

Fourthly, we conducted the comparative experiment about the traditional heuristic rule-based and our proposed approaches, and the results showed that our proposed approach was more adaptive and generalized in a complex traffic environment.

The remainder of the paper is organized as follows. Section 2 briefly reviews several approaches to settle the decision-making problem at an intersection. Section 3 states some of the problems, including the research scenario, the right of way the rules, and the vehicle intersection model. The problem formulation and the proposed MARL framework are described in Section 4. The experiments, results, and discussions are presented in Section 5. The paper is concluded, and the future works are discussed in Section 6.

## 2. Literature Review

Due to the complex interactions, the decisions made by the CAVs at the unsignalized intersections are a critical issue. Several approaches have been recommended for a settlement, namely the rule-based, optimization-based, and data-driven algorithms.

Earlier research mainly used a rule-based approach, where the main idea was to determine the order of the vehicle passage based on the rules or experience. The most direct approach was to carry out the first-in-first-out rule reservation scheme based on the centralized controller [15]. Another distributed approach based on fuzzy logic controllers was proposed. Milanes et al. [16] used V2V communication to determine the position and speed of the other vehicles in the intersection and then utilized a fuzzy controller to adjust the speed according to the speed of the vehicles with right of way. The rule-based approach was simple to implement, but not optimal.

Optimization-based algorithms take the decision factor as the objective and formulate the optimization under the constraints. Bian et al. [17] presented a distributed optimization to schedule the arriving times for the trajectory planning and achieved a satisfactory cooperative management with a 8.8–18.1% growth in the average passing times. However, the real-time optimization was limited to a high computational load and could not be achieved. In recent work, the combination of the above two approaches realized the approximations that dealt with the optimization via some heuristic rules, leading to a good trade-off between the performance and the computational complexity. For example, Xu et al. [18] solved a nearly globally optimal coordinated decision using the Monte Carlo tree search algorithm based on the feasible passing order that was selected using two heuristic rules. Vaio at el. [19] reformulated the vehicle coordination as the equivalent virtual platoon control problem based on the ascending order of the distance to the intersection. Due to the lack of learning and adaptability, these two methods were intractable to effectively tackle the task without accurate models.

As the field of artificial intelligence evolved, data-driven approaches have been a research hotspot for model-free problems due to their unparalleled data processing and generalization ability. Game theory [20] and deep reinforcement learning (DRL) [21] play a significant role in the decision-making at intersections. DRL addresses the cooperative decision-making tasks depending on the impressive learning capabilities based on the continuous interaction. Isele et al. [22] used DRL to identify the strategy that outperforms the common approach based on the heuristic rules. Lin et al. [23] discovered that when the scenario was modeled inaccurately, the policy trained by DRL performed better than the optimization approach of the interior point optimization (IPO). Shi et al. [24] proposed a coordinated control method with a proximal policy optimization (PPO) to make the CAV adapt to the HDVs. Liu et al. [25] employed a DRL to guide the expected speed and converge to the planned decision.

However, these studies involved only a single vehicle, disregarding the fact that concurrent interaction cooperative decision-making is a typical multi-agent system (MAS). Hence, it was logical to extend to a multi-agent reinforcement learning (MARL) framework. The decision at the intersection can be articulated as a fully cooperative MARL problem. MARL has already been used for research in the field of intelligent transportation, such as traffic

signal control [26] and highway decision-making [27]. Decision-making at the intersection can be articulated as a fully cooperative MARL problem, for which very few works exist, even though it is still a new area of research [21]. Based on the model accelerated proximal policy optimization (PPO), Guan et al. [28] proposed a centralized coordination method to globally coordinate the CAVs approaching the intersection by considering their states altogether, which achieved an increased efficiency. Zhou et al. [29] applied a centralized control for all CAVs that shared the same learned controller and then enabled the CAVs to form an appropriate behavior using the deep deterministic policy gradient (DDPG) algorithm. Antonio et al. [30] used MARL to identify complex real-life traffic scenarios and collaboratively regulate the CAVs at the intersection, which reduced 59% of the travel time and 95% of the congestion time compared to the traffic light control method. However, some issues in the MARL settings were not considered in the aforementioned works. The first issue was how to learn the most relevant information from other agents in a partially observable environment. The second was the reward assignment problem. In this study, we attempt to address the abovementioned issues by adopting a multi-agent reinforcement learning-based approach.

## 3. Problem Statement

### *3.1. Scenario Description*

The management of an unsignalized intersection is a challenging problem due to the multiple vehicles with potential conflicts and variable driving behavior. We focused on a traffic scenario at an intersection where the CAVs and HDVs coexisted and interacted with the surrounding vehicles to cross safely and efficiently.

As shown in Figure 1, a typical single-lane four-way unsignalized intersection consisting of four entrances and four exits was introduced. The areas upstream of the entrances and downstream of the exits were combined with a straight lane of the distance $L_s$, where only a rear-end collision was possible since overtaking was not feasible. When the vehicle reached the end of the entrance lane, it reached the stop line, indicating that the vehicle was about to enter the intersection. Inside the intersection, there could be a head-on collision, a rear-end collision, or other hidden dangers. Suppose that a vehicle pre-plans a path $p$ based on its initial lane and target lane before entering the intersection and follows this path to cross the intersection. Accordingly, there are 12 paths in total at the intersection, excluding the U-turn. There are a total of 20 conflicting points on these paths, including 16 crossing points and four converging points. The possibility of overtaking does not exist at the intersection, hence the diverging points are not the critical conflicting points. Assume that the preset path can be tracked perfectly by the vehicle, and thus only the longitudinal velocity along the path needs to be adapted. Our objective was to select the actions about the acceleration at every time step for the CAVs to cross the intersection safely and efficiently.

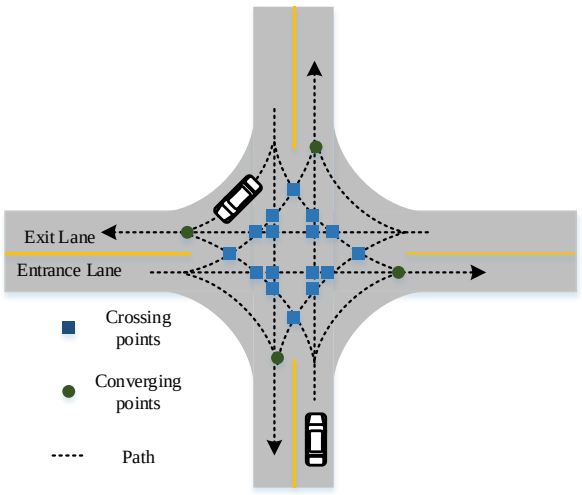


**Figure 1.** Illustration of the considered single-lane four-way unsignalized intersection scenario.

### 3.2. Right of Way Rules

Human drivers mainly avoid crash conflicts by following the traffic rules at unsignalized intersections, which helps drivers determine how they should proceed to pass through successively, i.e., the right of way assignment. In this paper, the traffic rules in China were used to regulate the vehicles at the intersections. More specifically, each vehicle $i$ that is about to enter the intersection should yield to the vehicle $j$ that conforms to any of the following conditions (in order of priority): (a) vehicle $j$ is already engaged in the intersection or significantly closer to the intersection; (b) vehicle $j$ is on the right side of vehicle $i$; (c) vehicle $j$ is proceeding straight and encounters the turning vehicle $i$; (d) vehicle $j$ is about to turn and interacts with the vehicle $i$ to turn right, when in the opposite direction.

According to the aforementioned rules, if vehicle $i$ goes before vehicle $j$, it means $i \prec j$. $ps_{i,j}$ is used to represent the passage order. Thus, a priority state between these two vehicles can be defined as follows.

$$ps_{i,j} = \begin{cases} 1, i \succ j, \\ -1, i \prec j, \\ 0, \text{otherwise}, \end{cases} \tag{1}$$

If there is no potential collision risk between vehicles $i$ and $j$, the two vehicles have the same priority, that is, the priority state is 0.

### 3.3. Vehicle Interaction Model

In the considered scenario, the modeling of the CAV-to-CAV and the CAV-to-HDV interactions were developed in this paper. The CAVs equipped with V2X communication devices could exchange information, including the vehicle position, speed, acceleration, orientation, destination with the vehicles within the communication range, and road infrastructure, i.e., vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication. Based on the road infrastructure, the HDVs' state information could be observed and transferred to the CAVs [31]. Unlike the CAVs, the HDVs have the liberty to maneuver themselves under the premise of obeying the traffic regulations. However, they cannot send any acknowledgment of their interactions. In this paper, the HDVs were modeled as cautious drivers and they slowed in response to the dangerous movements of the vehicle ahead. The intelligent driver model (IDM), which imitates human driving behavior, was used to control the HDVs. Consequently, the acceleration of an HDV could be anticipated based on its current speed and headway distance.

## 4. Cooperative Decision-Making in a Multi-Vehicle Cooperative Task

### 4.1. Problem Formulation

In this paper, the unsignalized intersection environment, where the CAVs and HDVs coexist, was modeled as a model-free multi-agent system to solve the cooperative decision-making and control problems. The system was described as $G = (V, E)$, where $V = \{v_1, v_2, \ldots, v_N\}$ represents a non-empty finite with $N$ CAVs and the edge set $E \subseteq V \times V$ represents the connections among the CAVs. The $\text{CAV}_i$ makes decisions based on its observations of the local sensors, such as cameras and lidars, and its communication with its neighbors, denoted as $N_i \doteq \{j | \varepsilon_{i,j} \in E, i \neq j\}$, to enhance the range of perception. Considering that the decisions of $N$ CAVs are interactive, the problem was considered a fully cooperative multi-agent task, which was modeled as a partially observed Markov decision process (POMDP). For the POMDP, each agent received a partial observation $o_i \in O_i$ from the global state $s \in S$. Based on the observation, the agents took joint action, $a_1, a_2, \cdots, a_N \in A$ from the action set $A \doteq \times_{i \in V} A_i$ to interact with the environment and receive a reward from all the agents $R = \{r_1, r_2, \ldots, r_N\} : S \times A_1 \times \ldots A_N \to \mathbb{R}^N$. The goal was that the agents would attempt to learn an optimal joint policy $\pi_i : O_i \to A_i$ to maximize the expected return $G = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r_t^i\right]$ in the interaction with the environment, where $\gamma \in (0, 1]$ is the accumulated discount factor to quantify the importance of the

future reward and T denotes the total steps of an episode. It can be defined by a tuple $(S, A, P, R, O, N, \gamma)$, where $P$ denotes a state transition function.

Action: In this paper, the control system of the vehicle was split into two levels, the high-level decision-making and the low-level controller, as depicted in Figure 2. According to the local observation, the high-level decision-making selected an action in the action space $A_i$ defined as $A_i$ = {*hard acceleration, acceleration, idle, deceleration and hard deceleration*}. Then, based on the decision taken, the lower level controller, i.e., a PID controller, generated the corresponding throttle signals to maneuver the CAV.

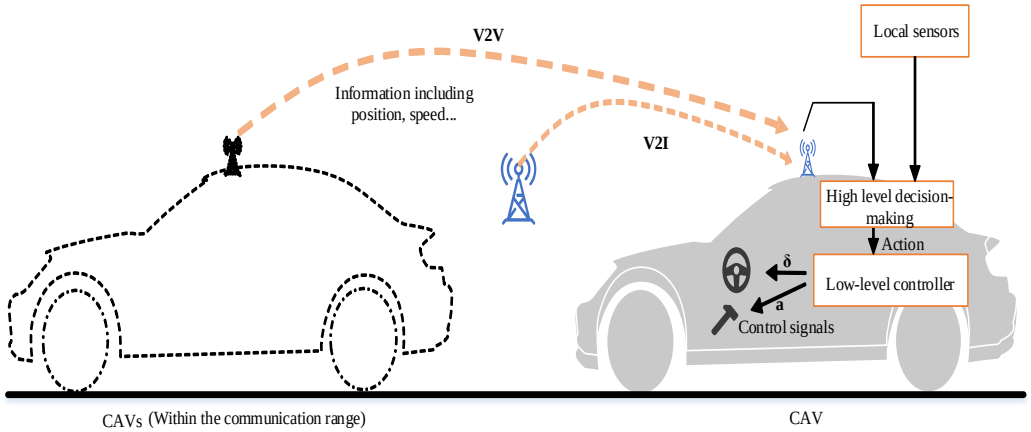


**Figure 2.** Schematic diagram of the decision control for multiple intelligent vehicles.

Observation: The observation $o_i$ of the vehicle $v_i$ included the information required to make decisions effectively. It was assumed that each vehicle could sense its state and exchange information with its neighbors, which is denoted as $N_i$. The neighbors of a vehicle were defined as the vehicles that were within $L_c$ meters and with the potential of a collision, i.e., a different priority on their respective routes. The observation $o_i$ was defined as a matrix of $b \times W$, where $b$ denotes the upper bound of the number of its neighbors and $W$ is the number of features $c_i$ representing the state of a vehicle. Specifically, the observation feature is defined as $c_i = [ispresent, x, y, v_x, v_y, h, ps]^T$, where *ispresent* is a binary flag denoting whether a vehicle can be observable. $x, y, v_x, v_y$ represent the absolute longitudinal position, the lateral position, the longitudinal speed, and the lateral speed for the ego vehicle, while the relative to the ego vehicle for observed vehicles; $h$ denotes the heading; and $ps$ represents the priority of crossing the intersection. The entire state of the system is the Cartesian product of the individual observation, that is, $S = O_1 \times O_2 \times \cdots \times O_N$..

Reward: The basic goals of a vehicle at the intersection include driving safely, crossing the intersection, and reaching its target lane under the right of way rules and in a timely manner. Several rewards are specified as follows.

(a) The occurrence of a collision is detected by the body circle model. It reduces the vehicle to a circle with the center of the vehicle as the center and the diagonal of the vehicle as the diameter. When two vehicles are tangent to each other, a collision is considered to have occurred. The collision reward $r_c$ is defined to penalize the occurrence of a collision and reward the successful pass, expressed as follows.

$$r_c = \begin{cases} -1, \text{if a collision happens,} \\ 1, \text{if all CAVs pass successful,} \\ 0, \text{otherwise} \end{cases} \quad (2)$$

(b) For safety, the vehicle is penalized when the minimum time headway $t_h$ with other vehicles is less than desired time headway $t_d$. In this paper, $t_d$ was set to 2 s. Our study scenario was a mixed traffic scenario where the HDVs and CAVs coexisted. The uncertain behavior of the HDVs prevented us from directly calculating the time to

collision (TTC). Therefore, within a prediction horizon $T_h$, all the vehicles' trajectories are predicted. The trajectories of the CAVs are predicted via the execution of the current action, while the HDVs are estimated from the IDM model. At each step of the prediction, the collision of vehicle $i$ is detected according to the body circle model. If a collision occurs at step $t$, the time headway $t_h$ is determined according to $t_h = t/f$, where $f$ represents the control frequency of the vehicle. When no collisions are detected, $r_h$ is set to 1. Thus, the headway reward $r_h$ is defined as follows.

$$r_h = \begin{cases} \ln \frac{t_h}{t_d}, \text{if collisions are detected,} \\ 1, \text{ otherwise} \end{cases}. \tag{3}$$

(c) To pass though the intersection safely and effectively, the speed should be considered as appropriate and is punished too low. The speed reward is defined as follows.

$$r_s = \min\left(\frac{v_t - v_{min}}{v_{max} - v_{min}}, 1\right), \tag{4}$$

where $v_t, v_{\min}, v_{\max}$ are the current speed and the minimum and maximum speeds to be rewarded.

(d) For the vehicle to make decisions under the rules demonstrated in Section 3.2 the rule reward $r_r$ is set to 1 if the rule is obeyed, otherwise $r_r = -1$.

Based on these definitions above, the total reward for the $v_i$ at time step $t$ is defined as follows.

$$r_t^i = \omega_c r_c + \omega_h r_h + \omega_s r_s + \omega_r r_r, \tag{5}$$

where $\omega_c, \omega_s, \omega_h$, and $\omega_r$ are all weighting factors that account for each part of the reward.

### *4.2. Wighted Reward Assignment*

In the fully cooperative MARL setting, each agent was provided with the same goal and assigned the same reward after executing the action at every step. The same reward can be represented by the average global reward as $r_t = \frac{1}{N}\sum_{i=1}^{N} r_{i,t}$. However, the shared rewards scheme is intricate in order to infer each vehicle's contributions to the system cooperation. Further, instead of embracing a global reward, a local reward assignment strategy could alleviate the issue of the credit assignment problem. Specifically, each host vehicle only focuses on its surrounding vehicles, which considerably impacts the smooth interaction between the vehicles. Thus, the reward for the ego vehicle $v_i$ at the step $t$ is defined as follows.

$$r_{i,t} = \frac{1}{|T_i|}\sum\nolimits_{j \in T_i} r_{j,t}, \tag{6}$$

where $T_i = i \cup N_i$ is a set whose elements include the ego vehicle $v_i$ and its neighbor vehicles and $|\cdot|$ represents the cardinality operator of the set. The local reward assignment has two advantages. First, the communication burden can be reduced by focusing only on the nearby vehicles, establishing a more real-time system. Second, the contribution of the vehicles to their cooperation can be indicated more accurately.

Nevertheless, the regional reward assignment strategy still doesn't accurately differentiate between the contributions of each ego vehicle at different positions on the road. Collisions are more likely to occur inside the intersection than outside it, meaning that a vehicle closer to the intersection should be assigned more rewards, i.e., a weighted reward assignment according to the vehicle's position. A weight factor $\eta$ was defined to measure the contribution for crossing the intersection safely and efficiently, as shown in Equation (7).

$$\eta_i = \frac{L_s - d_i}{\sum_{j \in T_i}(L_s - d_j)}, \tag{7}$$

where $d_i$ denotes the distance from the vehicle $v_i$ not entering the intersection to the entrance or from $v_i$ not exiting the intersection to the exit, and $L_s$ is the length of the entrance or exit

straight lane. When $v_i$ enters the intersection, $d_i = 0$. Thus, the reward for the ego vehicle $v_i$ at the step $t$ is as follows.

$$r_{i,t} = \sum_{j \in T_i} \eta_i r_{j,t}. \tag{8}$$

When a vehicle and its neighbors are in the intersection, Equation (8) will degenerate to the local reward assignment strategy, i.e., Equation (6).

### *4.3. Cooperative Learning Algorithm for Multi-Agent Task*

The MARL method was leveraged to deal with the unsignalized intersection management due to the fact that it could produce an optimal policy through a continuous interaction with the traffic environment. A cooperative PPO-based decision-making method (Attn-MAPPO) is proposed in this section, as shown in Figure 3. The algorithm was based on a centralized training decentralized execution (CTDE) framework to reduce the environmental instability; that is, all the information of the agents was utilized during training and the agents only made decisions according to their own local observation after the training. Specifically, after choosing the actions from the policy to interact with the environment, all the vehicles executed them and reached a new state. Based on the new observation, an attention module was used to aggregate the information from the neighbors of a vehicle. Then, the output of the module was used to update the critic network. The critic network evaluated the taken actions and the agent network was updated.

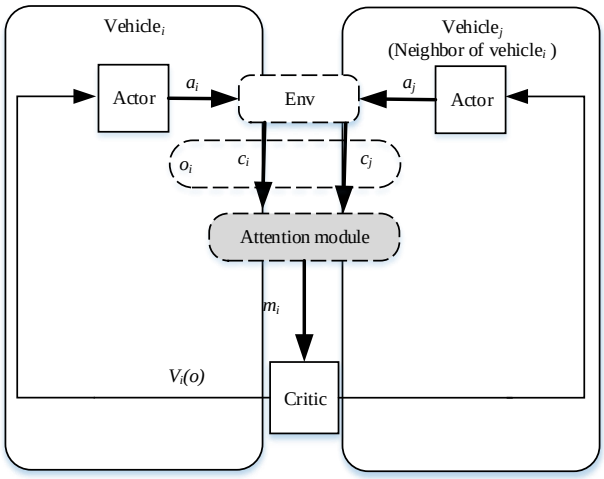


**Figure 3.** The framework of the proposed Attn-MAPPO algorithm.

In this paper, each vehicle could observe its neighbors and collect their state information to make decisions. Although the relevant vehicle was selected as the focus, the aggregated information was not the most valuable information. For example, the state information of the neighbors which were more likely to collide required more attention. The attention mechanism was introduced in this paper to selectively pay attention to the neighbors' observations.

More specifically, the structure of the attention representation is defined as Figure 4. The relevant message from other vehicles, $m_t^i$, is summed by assigning the attention weights to the embedding of each agent, which is mathematically expressed as the following.

$$m_t^i = \sum_{j \neq i} \alpha_t^{i,j} W_v e_t^j, \tag{9}$$

where $e_t^j = f_j(o_t^i)$ is the embedding where a multi-layer perceptron (MLP) $f_j(\cdot)$ is used as the embedding function and $W_v$ is a matrix to linearly transform $e_t^j$ into a "value". The

attention weight $\alpha_t^{i,j}$ between the vehicles $i$ and $j$ was calculated using a *softmax* function, as shown in the following.

$$\alpha_t^{i,j} = \frac{\exp(\beta_t^{i,j})}{\sum_{j \neq i} \exp(\beta_t^{i,j})}, \tag{10}$$

$$\beta_t^{i,j} = W_q e_t^i \cdot (W_k e_t^j)^T, \tag{11}$$

where $\beta_t^{i,j}$ computes the correlation between the vehicles $i$ and $j$, the matrix $W_q$ linearly transforms $e_t^i$ into a "query", and $W_k$ transforms $e_t^j$ into a "key".

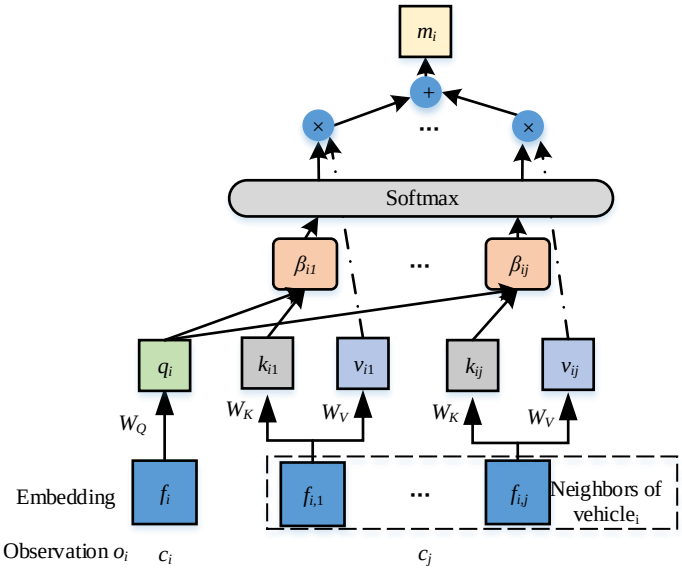


**Figure 4.** Flowcharts of the attention representation.

In the cooperative multi-agent setting, each agent shared an actor network (i.e., policy) and a critic network (i.e., policy evaluation), which were trained using the individual trajectory $(o_t^i, a_t^i, r_t^i)$. The actor network was trained to maximize the expected return. Each agent generated its action $a_t^i$ from the policy $\pi_\theta$ based on its local observations $o_t^i$ (i.e., $\pi(a_i^t | o_t^i; \theta)$ at the time slot $t$. The most valuable information which an attention representation aggregated from the neighbors of an agent was access to the input of the critic network. An episodic setting was considered with each vehicle until a crash occurred or until it proceeded $T$ seconds. Our objective was to obtain the optimal joint policy that maximized the accumulated reward.

The PPO algorithm based on the actor–critic framework was an improvement of the policy gradient (PG) algorithm. The objective function of the PG algorithm is expressed as the following.

$$\nabla J(\theta) = \mathbb{E}_{(s_t, o_t) \sim \pi_\theta}[\nabla_\theta \log \pi_\theta(a_t | s_t) A_{\pi_\theta}(s_t, a_t)], \tag{12}$$

where

$$A_{\pi_\theta}(s_t, a_t) = Q_{\pi_\theta}(s_t, a_t) - V_\phi(s_t) \tag{13}$$

is an estimation of the advantage function at the step $t$ and $\theta, \phi$ are the parameters for identifying the policy and state value functions, respectively.

By modifying the objective, the PPO constructs a clipped surrogate objective function to confine the policy updates to a small scope of around 1. The objective and update functions are illustrated as the following.

$$J^{PPO}(\theta) = \mathbb{E}_{(o_t^i, a_t^i) \sim \pi_{\theta_{old}}}[\psi(\rho_\theta^i, A_t^i)], \tag{14}$$

$$\theta_{t+1} = \theta_t + l\nabla_\theta J^{PPO}(\theta), \tag{15}$$

where $l$ is the learning rate and $\rho_\theta^i = \frac{\pi_\theta(a_t^i|o_t^i)}{\pi_{\theta_{old}}(a_t^i|o_t^i)}$ specifies the probability ratio of the old and new policies to select the action and

$$\psi(\rho_\theta^i, A_t^i) = \min(\rho_\theta^i A_t^i, \text{clip}(\rho_\theta^i, 1-\varepsilon, 1+\varepsilon)A_t^i), \tag{16}$$

where $\varepsilon$ is the penalty factor to prevent the policy from changing extremely. When $A \geq 0$, the return for taking the action $a_t^i$ is greater than the expected observation $o_t^i$. Therefore, the updated policy should increase the probability of the action, but the increase amplitude should be restrained to $(1+\varepsilon)\pi_{\theta_{old}}(a_t^i|o_t^i)$. The opposite is true when $A < 0$.

To decrease the variance, the advantage function $A_{\pi_\theta}$ in Equation (13) is replaced by the generalized advantage estimation (GAE), which is expressed as the following.

$$\hat{A}_t^i = \delta_t^i + \sum_{l=t+1}^{T} (\gamma\lambda)^{l-t}\delta_l^i, \tag{17}$$

where $\delta_t^i(\gamma) = r_t^i + \gamma V_\phi(o_{t+1}^i) - V_\phi(o_t^i)$ is the TD error and $\lambda$ is a factor that balances the variance and bias of the estimation. When $\lambda = 0$, $\hat{A}_t^i = r_t^i + \gamma V_\phi(o_{t+1}^i) - V_\phi(o_t^i)$, which is unbiased with a high variance. The GAE achieves both a low bias and a low variance by linearly integrating the $n$-step bootstrapping. The critic network $V_\phi(\cdot)$ is updated using the loss function as follows.

$$J(\phi) = \min \mathbb{E}_{(o_t^i, a_t^i)\sim\pi_{\theta_{old}}}[r_t^i + \gamma V_\phi(o_{t+1}^i) - V_\phi(o_t^i)]^2, \tag{18}$$

$$\phi_{t+1} = \phi_t + l\nabla_\phi J(\phi). \tag{19}$$

The whole pseudo-code for the proposed method is presented in Algorithm 1.

**Algorithm 1**: Attn-MAPPO

```
Initialize the actor network and target the actor network using the parameters θ and θ⁻;
Initialize the critic network and target the critic network using the parameters φ and φ⁻;
Initialize the memory buffer D_i and hyper-parameters lr, τ, ϵ.
for episode = 1, ... , M do
    for t ≤ T and not terminal do
        for i ∈ V do
            Observe o_i and select an action a_i ∼ π_θ⁻ using the ϵ-greedy strategy.
            All agents execute the actions and receive their own reward r_i.
            Store trajectories (o_i, a_i, r_i) in D_i.
        end for
        for i ∈ V do
            Obtain the attention representation of each agent using Equation (9).
            Update the critic network θ and the actor network θ using a randomly sampled
        mini batch from D_i in Equations (15) and (19), respectively.
            Update the target networks:φ⁻ = τφ + (1 − τ)φ⁻, θ⁻ = τθ + (1 − τ)θ⁻.
        end for
    end for
    Initialize D_i ← ∅, and reset the environment.
end for
```

## 5. Results and Discussion

In this section, the proposed Attn-MAPPO algorithm was evaluated at the unsignalized intersection on an open source platform, called highway-env [32]. A total of 30 episodes of the evaluation for each contrast were executed using 30 various seeds. Two metrics were used for the performance of the algorithm, namely the collision rate and the average

speed. The collision rate was defined as the ratio of the number of episodes where the collision occurred to the total number of the test episodes. The average speed was defined as the average speed of all the vehicles in all the episodes. In the scenario, a straight lane had a distance of $L_s$ = 200 m, the right turn radius was 9 m, and the left turn radius was 13 m. The vehicle's policy frequency was set to 5 Hz, i.e., the CAVs took an action every 0.2 s. The speed range of receiving a reward was [8,10] m/s. For the actions, the normal acceleration and deceleration were to add or subtract 1.5 m/s from the current speed, as the desired speed and the hard were 3 m/s. The PID algorithm was used as the low-level controller to change the current speed to the expected speed. To avoid the excessive speed, the maximum speed of CAVs was set to 10 m/s. The HDVs' desired speed was set to 10 m/s. The communication range $L_c$ was set to 120 m. The setting of the training process was presented as follows. The model was evaluated every 20 episodes during the training. ADAM was used as the optimizer and the learning rate was set to $8 \times 10^{-5}$. The soft update weighting factor $\tau$ was set to 0.001. The trade-off coefficient of the GAE $\lambda$ was set to 0.95 and the discounting factor was set to 0.99.

### *5.1. Performance Comparison between the Proposed Attn-MAPPO Algorithm and the Benchmark*

In this subsection, we compared the proposed Attn-MAPPO approach with the MAPPO benchmarks [13]. The architecture was the same for both, but the Attn-MAPPO used an attention mechanism to extract the most valuable information as the input for the network. To validate the performance of the proposed Attn-MAPPO algorithm, two traffic scenarios were set up with (1) two CAVs and three HDVs, and (2) four CAVs and five HDVs. The first setting was simple due to the sparse traffic density, while the second was complex. The comparison results of the episode reward in the training is illustrated in Figure 5. The results showed that in the simple setting, namely with the two CAVs, three HDVs, both the algorithms performed comparably but moderately better than the benchmark. Yet, as the number of vehicles increased, the benchmark performance declined significantly and the Attn-MAPPO algorithm showed a stronger performance. In the 30 tests, one collision occurred using the proposed method and four collisions occurred using the benchmark in the complex setting. Table 1 shows that for the proposed algorithm, the collision rate could be considerably diminished in the complex traffic setting. This implies that the added attention module was able to extract the critical information for decision-making, whereas the benchmark did not have this capability.

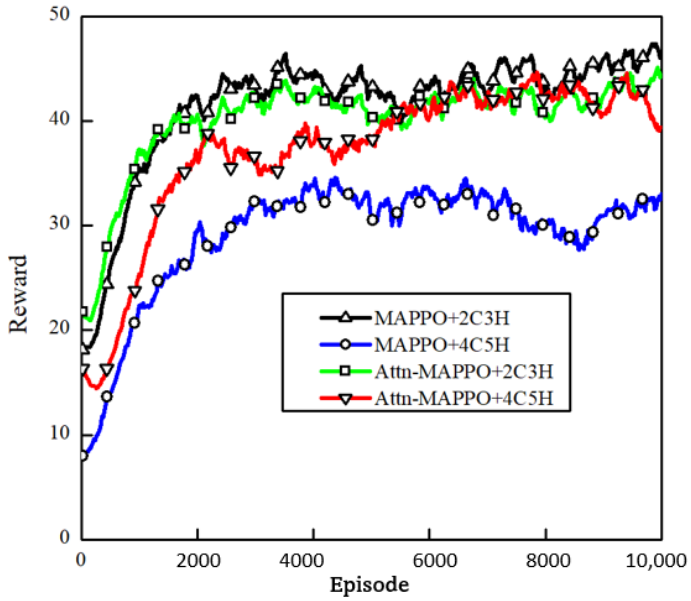


**Figure 5.** Comparison of the episode reward in the training between the benchmark and the proposed Attn-MAPPO (proposed) algorithm.

**Table 1.** Performance comparison between the benchmark and the proposed Attn-MAPPO (proposed) algorithm.

| Metrics | Algorithm and Settings | | | |
|---|---|---|---|---|
| | **Benchmark +2C3H** | **Benchmark +4C5H** | **Attn-MAPPO +2C3H** | **Attn-MAPPO +4C5H** |
| Collision rate | 0 | 0.13 | 0 | 0.03 |
| Average speed (m/s) | 10.08 | 8.77 | 9.91 | 8.95 |

### *5.2. Performance of the Proposed Reward Scheme Designs*

In this paper, the setting of the reward mechanism had two main aspects. The first was the design of the reward function and the second was the design of the reward assignment scheme.

There were four components in the reward function, including the collision evaluation, headway evaluation, speed evaluation, and rule evaluation. In order to validate the effectiveness of the reward function for decision-making, we set up an intersection crossing scenario. There were three vehicles in this scenario, where one HDV shown in blue was going straight and two CAVs shown in green were turning left, denoted as $CAV_1$ and $CAV_2$, respectively. In order to present the spatial relationship of the vehicles more intuitively, the positions of the vehicles were represented by the distance from the stop line along their path, rather than their world coordinates. The positions were negative when the vehicles crossed the stop line. The position and speed curves of all the vehicles during the whole process are shown in Figure 6. Figure 7 shows a sequence of the time slices for demonstrating the inter-vehicle interaction. Apparently, all the vehicles crossed through the conflict zone safely. As shown in Figure 6a, when there was no interaction between vehicles, they could accelerate to the maximum speed of 10m/s to obtain a higher reward. In accordance with the traffic rules introduced in Section 3.2, the vehicle going straight goes first, and the one on the right has the right of way. Thereby, the passage sequence of the three vehicles was HDV, $CAV_1$, and $CAV_2$, which was exactly what is shown in Figure 7. There were two interactions during the crossing. $CAV_1$ and the HDV first interacted in the same lane, where $CAV_1$ slowed down and pulled away from the HDV ahead of it. The second interaction took place between the two CAVs. Since $CAV_2$ should have crossed later than $CAV_1$, $CAV_2$ decelerated to make way at 12.4 s. All the above analyses demonstrated that the designed reward function was qualified for the effective decision-making for the CAVs.

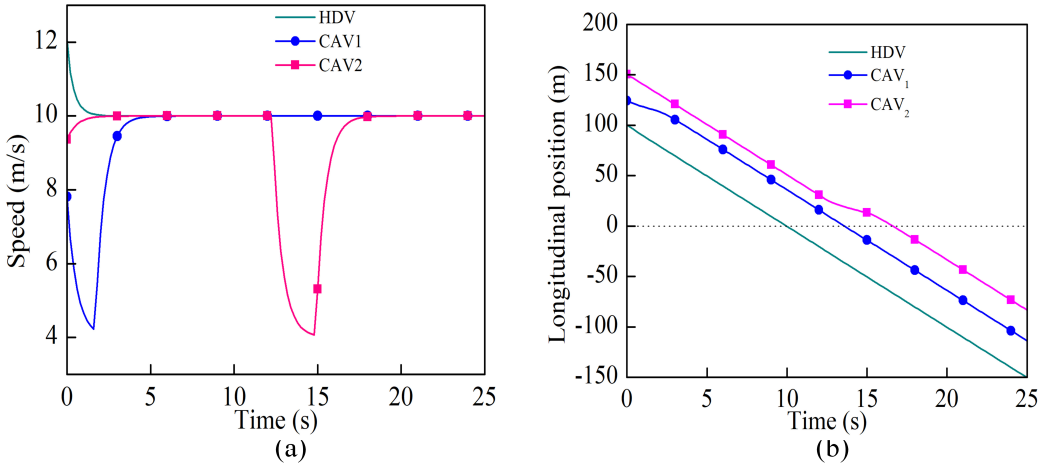


**Figure 6.** The speed and position curves of all the vehicles during the whole crossing; (**a**) the speed, (**b**) the position.

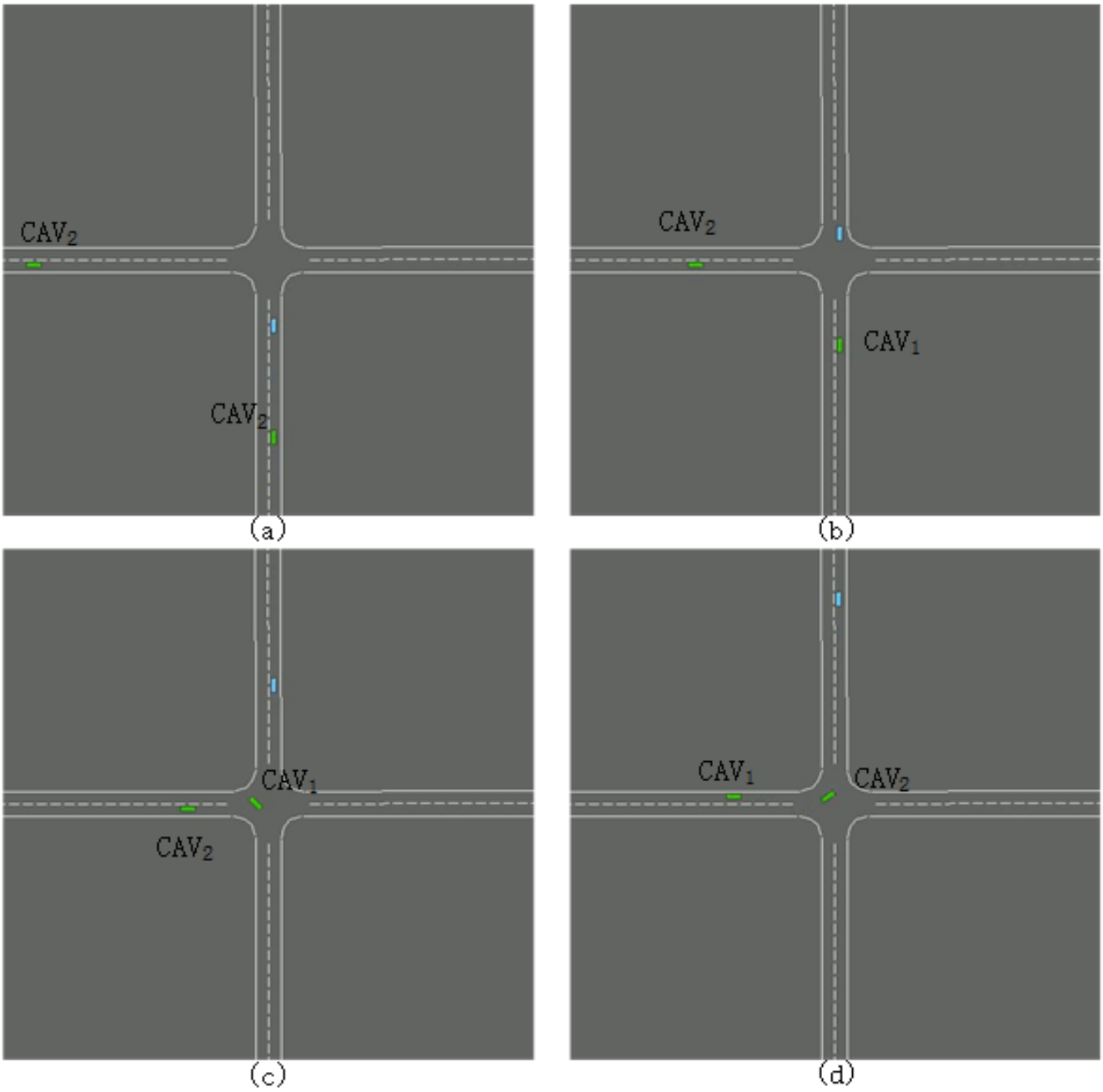


**Figure 7.** The frames of the time slices showing the interactions performed by the simulation platform (blue: HDV; green: CAVs); (**a**) 9 s, (**b**) 12 s, (**c**) 15 s, and (**d**) 18 s.

To validate the performance of the proposed reward assignment scheme, the global reward, local reward, and the proposed weighted reward scheme were used for the training in the simple setting, i.e., the two CAVs and three HDVs. As many people are aware, collisions are more likely to occur in the areas within the intersections than in the straight lanes. When the system was penalized, the vehicle nearest the intersection should have been assigned more penalties to help it execute the correct decision. The same went for receiving a reward. The results are shown in Figure 8. As excepted, the results confirmed that the proposed weighted assignment outperformed the other two assignment schemes. Table 2 shows a clear improvement in the collision rates and the average speed metrics. This was because the global reward was not an accurate representation of the contribution of a CAV. Although the local reward improved on this, the reward could not be assigned based

on an agent's location. The weight reward assignment scheme assigned more rewards to the vehicles closer to the intersection, which facilitated the cooperation between the agents.

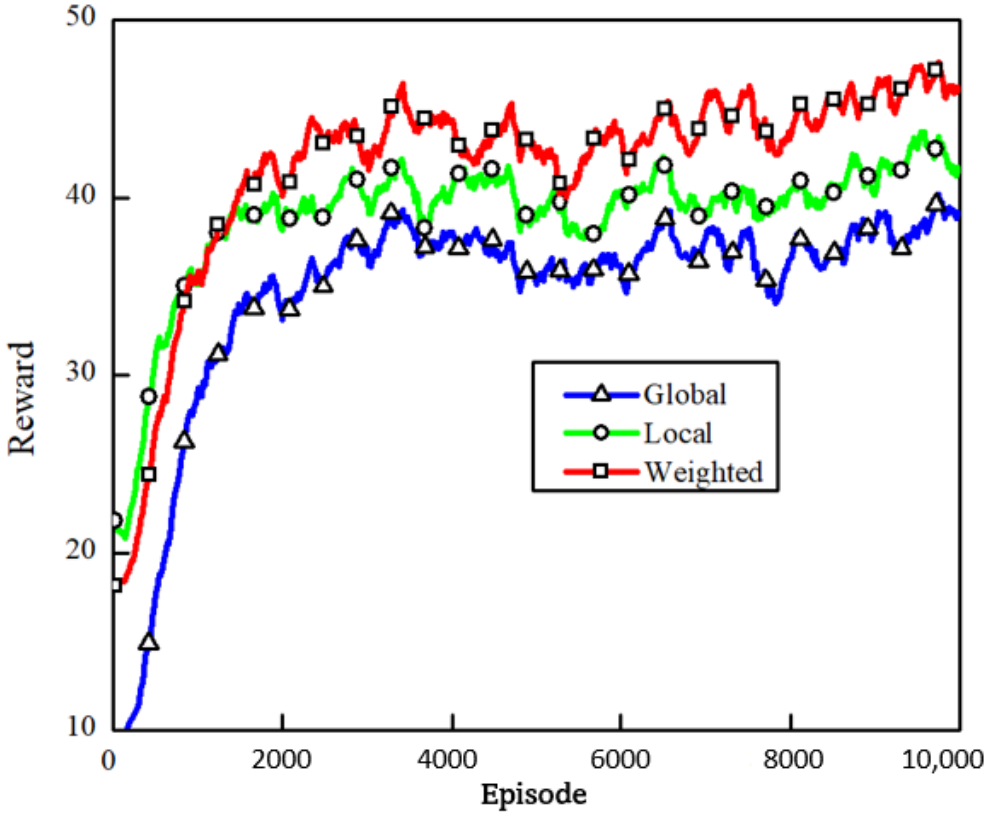


**Figure 8.** Comparison of the episode reward in the training using the different reward assignment schemes.

**Table 2.** Performance comparison using the different reward assignment schemes.

| Metrics | Reward Assignment Scheme | | |
|---|---|---|---|
| | **Global** | **Local** | **Proposed** |
| Collision rate | 0. 07 | 0.03 | 0 |
| Average Speed (m/s) | 9.06 | 9.04 | 9.91 |

### *5.3. Performance Comparison of the Different Forward Predictive Horizons*

Our research scenario was a mixed traffic scenario where the HDVs and CAVs coexisted. In this paper, the CAVs made decisions based on the predicted trajectories of all the vehicles over time. To validate the impact of the different time horizon for predicting the trajectories of all the vehicles, three settings were used for the training. The results are shown in Figure 9 and the performance metrics are presented in Table 3. The results indicated that a larger prediction time horizon did not imply a better performance. Although a larger predictive time horizon could reduce the collision rate, it would significantly reduce the metrics of the average speed. The performance depended on the design of the headway evaluation in the reward function and the desired headway time, as introduced in Section 4.1. The desired headway time $t_d$ was set to 2 s during the training. When $T_h = 1$ s, even though the current headway time was less than $t_d$, the vehicle agent still received a positive reward because no collision was detected within the predictive horizon. This could easily result in the vehicle anticipating a collision when the brakes would not be able to prevent a collision. Compared to $T_h = 3$ s, the prediction horizon of $T_h = 5$ s was conservative for the decision-making of the vehicle. As long as the collision was not detected within 3 s, the vehicle received a reward of one with a 3 s prediction horizon. In

the case of $T_h = 5$ s, the vehicle would be rewarded with the value of $\ln(5/2) < 1$, even though the collision was predicted at 5 s. This indicated that a suitable prediction time horizon could achieve a better performance. $T_h = 3$ s achieved a good trade-off between the collision rate and the prediction efficiency.

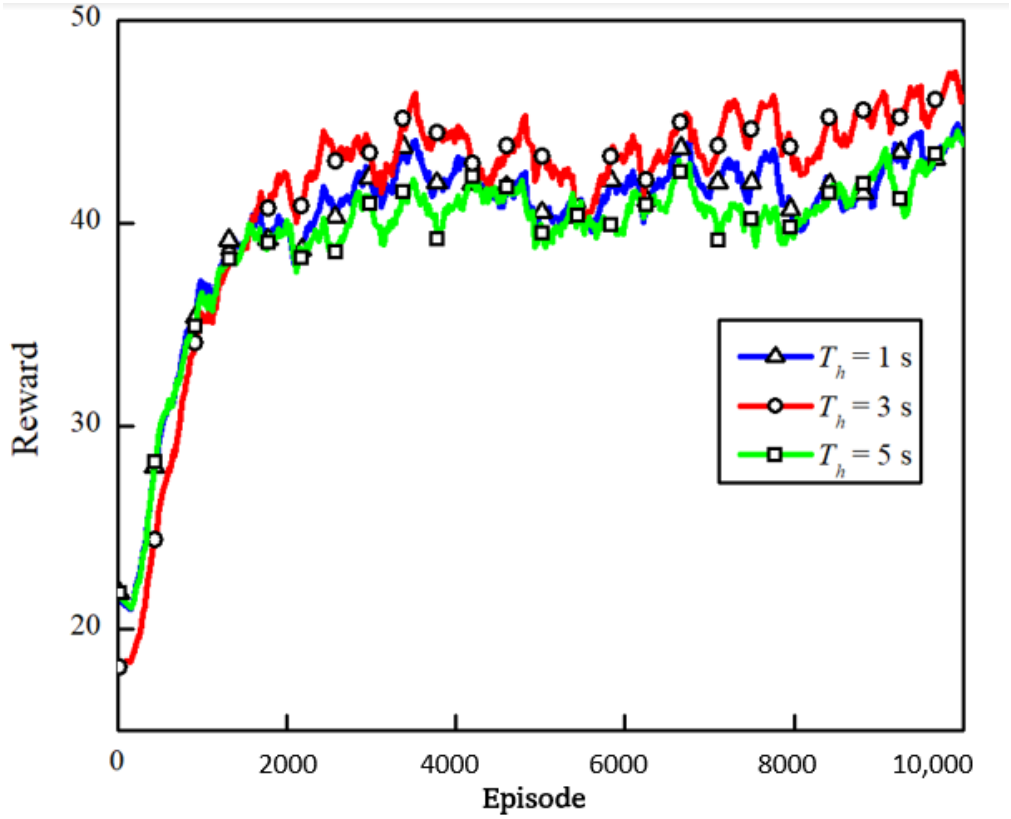


**Figure 9.** Comparison of the episode reward in the training for the different forward predictive horizons.

**Table 3.** Performance comparison for the different forward predictive horizons.

| Metrics | Horizon Time | | |
|---|---|---|---|
| | $T_h$ = 1 s | $T_h$ = 3 s | $T_h$ = 5 s |
| Collision rate | 0.03 | 0 | 0 |
| Average Speed (m/s) | 9.10 | 9.91 | 8.67 |

### *5.4. Performance Comparison of the Proposed and Heuristic Rule-Based Algorithm*

In this subsection, we verified the adaptability of our proposed algorithm by comparing it with the heuristic rule-based algorithm using the same scenario in Section 5.2. The traditional ruled-based or model-based decision control approaches were capable of generating stable decision results, accurate control curves, and were easy to implement. For example, the vehicles coordination of negotiating the access in an intersection was reformulated as a virtual platoon control problem in [9]. According to the passage sequence, the HDV acted as the leader and the two CAVs kept the desired following distance. Figure 10 shows the results using a virtual platoon. The HDV traveled at the desired speed of 10 m/s. $CAV_1$ adjusted its speed to keep the desired distance from the HDV and $CAV_2$ followed $CAV_1$.

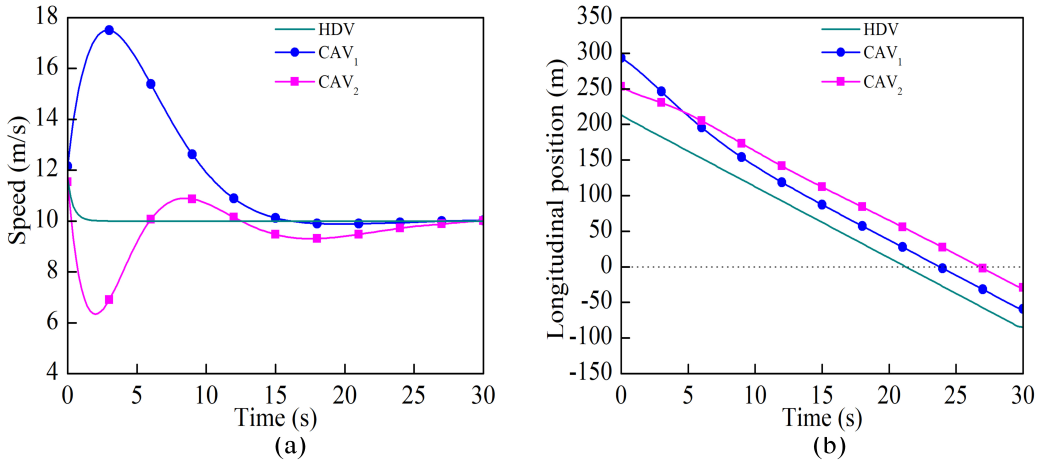


**Figure 10.** The speed and position curves of all the vehicles in the ideal traffic using the heuristic rule-based algorithm; (**a**) the speed, (**b**) the position.

Although the vehicles could cooperatively and efficiently cross the intersections under the ideal traffic conditions using the heuristic rule-based model, some of the idealized condition assumptions were difficult to satisfy in many cases. Generally speaking, the traffic scenarios involving the HDVs were more complex and uncertain for the uncontrolled human driving. For example, some drivers were conservative and traveled at a low speed for fear of getting scratched in intersections. The speed and position curves of all the vehicles when the HDV traveled at a low speed using a heuristic rule-based algorithm are shown as Figure 11. In order to form a virtual platoon, the speed of both CAVs converged to that of the leading HDV. Practically, the $CAV_2$ was not in the same lane as the HDV, so it did not need to keep up with the speed of the HDV, which could greatly enhance the road efficiency. The results demonstrated that the heuristic rule-based approaches, while easy to implement, were not adaptive to some special cases.

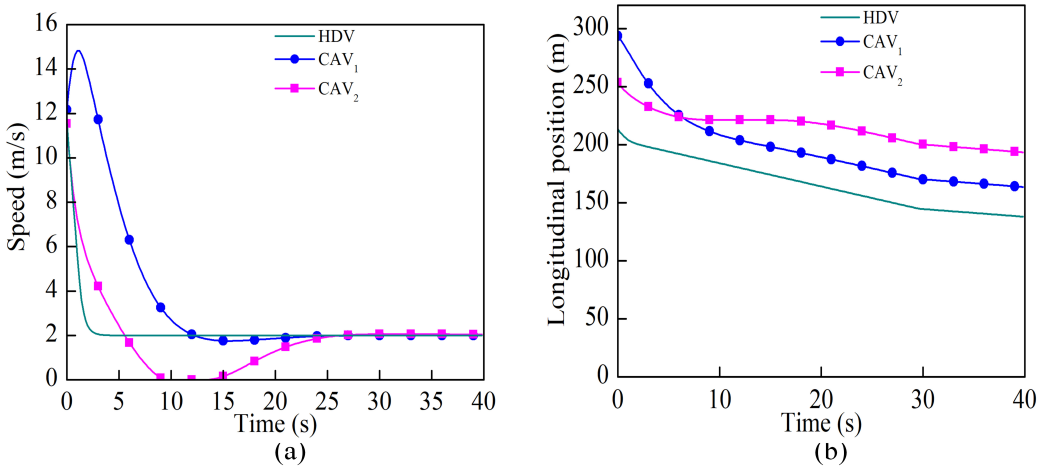


**Figure 11.** The speed and position curves of all the vehicles when the HDV traveled at a low speed using the heuristic rule-based algorithm; (**a**) the speed, (**b**) the position.

A learning-based approach enabled more efficient decision-making. The results are presented in Figure 12 using the proposed algorithm when the HDV traveled at a low speed. Based on the trajectory prediction of the other vehicles, $CAV_2$ predicted that the slow speed of the HDV would not cause a collision. $CAV_2$ made the decision to breaks traffic rules and pass first, since the reward for following the traffic rules was smaller than the penalty for crossing the intersection at a low speed. Thus, if no collision was predicted, the HDV only affected the decision of $CAV_1$ behind it in the same lane but not $CAV_2$'s

in other lane. As shown in Figure 12b, after the HDV crossed the stop line, that is, when its position was less than 0, the CAV began to accelerate. This accounted for the fact that the proposed approach could deal with some special cases, which were more adaptive and generalized.

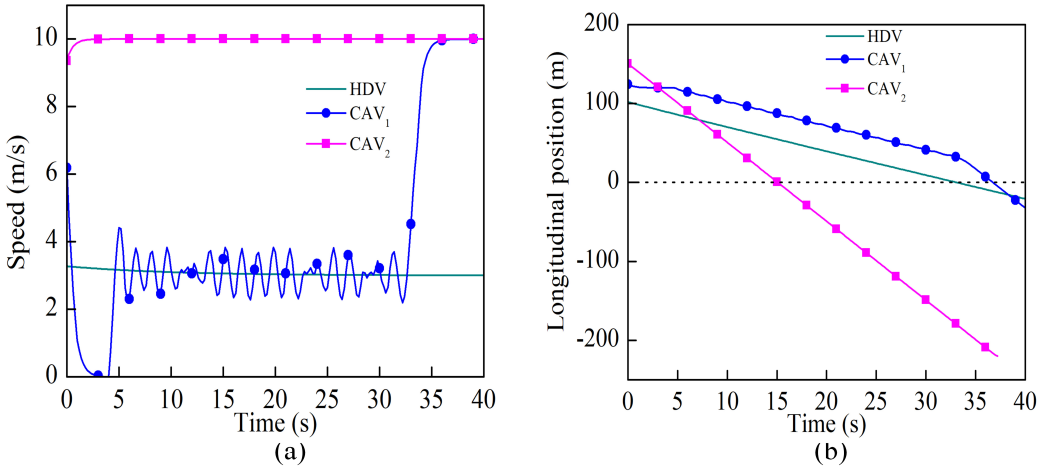


**Figure 12.** The speed and position curves of all the vehicles when the HDV traveled at a low speed using the proposed algorithm; (**a**) the speed, (**b**) the position.

## 6. Conclusions

In this paper, we modeled the decision-making at the intersection of the CAVs and HDVs in coexistence as a model-free and fully cooperative multi-agent system. Based on this, we presented the design of the observation, action, and reward functions to formulate the cooperative decision-making as a MARL problem. Then, a decentralized MAPPO based on the attention representations algorithm (Attn-MAPPO) was developed to make joint decisions to avoid collisions and cross the intersection effectively. Finally, the policy was trained and evaluated via an open source simulation platform. The results showed the advantages of our proposed algorithm and the designed reward scheme. In addition, the comparison of the results using the three prediction time horizons also suggested that a suitable horizon could achieve a better performance. We compared the performance of the Attn-MAPPO to an equivalent virtual platoon control, a heuristic rule-based method, which indicated that the proposed approach could deal with some special cases more adaptively and generically.

However, there were still some unsolved problems in this paper. The driving behavior of the HDVs was modeled using the IDM model. In practice, more accurate models may be needed. Moreover, the scalability of the algorithm needs to be improved, i.e., increasing the number of CAVs. At the same time, in this paper, only the full cooperation between the vehicles was considered. In fact, there was also a competition between the vehicles at the intersection. In future works, we will pay more attention to these issues and continue to expand from this study.

**Author Contributions:** Conceptualization, H.Z.; methodology, H.Z.; software, H.Z.; validation, C.L.; formal analysis, C.L.; investigation, Y.C.; resources, X.T.; data curation, H.Z.; writing—original draft preparation, H.Z.; writing—review and editing, C.L. and Y.C.; visualization, C.L.; supervision, X.T.; project administration, X.T.; funding acquisition, X.T. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Key-Area Research and Development Program of Guangdong Province under Grant 2020B0909030005 and 2020B090921003.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data sharing not applicable. No new data were created or analyzed in this study. Data sharing is not applicable to this article.

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

## References


1. Eskandarian, A.; Wu, C.; Sun, C. Research Advances and Challenges of Autonomous and Connected Ground Vehicles. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 683–711. [CrossRef]
2. Ghorai, E.A.P.; Kim, Y.K.; Mehr, G. State Estimation and Motion Prediction of Vehicles and Vulnerable Road Users for Cooperative Autonomous Driving: A Survey. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 16983–17002. [CrossRef]
3. Loke, S.W. Cooperative Automated Vehicles: A Review of Opportunities and Challenges in Socially Intelligent Vehicles Beyond Networking. *IEEE Trans. Intell. Veh.* **2019**, *4*, 509–518. [CrossRef]
4. Chen, L.; Li, Y.; Huang, C.; Li, B.; Xing, Y.; Tian, D.; Li, L.; Hu, Z.; Na, X.; Li, Z.; et al. Milestones in Autonomous Driving and Intelligent Vehicles: Survey of Surveys. *IEEE Trans. Intell. Veh.* **2023**, *8*, 1046–1056. [CrossRef]
5. Wang, H.; Meng, Q.; Chen, S.; Zhang, X. Competitive and cooperative behaviour analysis of connected and autonomous vehicles across unsignalised intersections: A game-theoretic approach. *Transp. Res. Part B Methodol.* **2021**, *149*, 322–346. [CrossRef]
6. Xue, Y.; Zhang, X.; Cui, Z.; Yu, B.; Gao, K. A platoon-based cooperative optimal control for connected autonomous vehicles at highway on-ramps under heavy traffic. *Transp. Res. Part C Emerg. Technol.* **2023**, *150*, 104083. [CrossRef]
7. Aoki, S.; Lin, C.W.; Rajkumar, R. Human-Robot Cooperation for Autonomous Vehicles and Human Drivers: Challenges and Solutions. *IEEE Commun. Mag.* **2021**, *59*, 35–41. [CrossRef]
8. Treiber, M.; Hennecke, A.; Helbing, D. Congested traffic states in empirical observations and microscopic simulations. *Phys. Rev. E* **2000**, *62 Pt A*, 1805–1824. [CrossRef]
9. Peng, B.; Keskin, M.F.; Kulcsár, B.; Wymeersch, H. Connected autonomous vehicles for improving mixed traffic efficiency in unsignalized intersections with deep reinforcement learning. *Commun. Transp. Res.* **2021**, *1*, 100017. [CrossRef]
10. Li, C.; Hu, Z.; Lu, Z.; Wen, X. Cooperative Intersection with Misperception in Partially Connected and Automated Traffic. *Sensors* **2021**, *21*, 5003. [CrossRef] [PubMed]
11. Jie, W.; Zhihao, J.; Vardhan, P.Y. Improving Safety in Mixed Traffic: A Learning-based Model Predictive Control for Autonomous and Human-Driven Vehicle Platooning. *arXiv* **2023**, arXiv:2211.04665.
12. Chen, L.; Englund, C. Cooperative Intersection Management: A Survey. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 570–586. [CrossRef]
13. Yu, C.; Velu, A.; Vinitsky, E.; Wang, Y.; Bayen, A.; Wu, Y. The surprising effectiveness of ppo in cooperative, multi-agent games. *arXiv* **2021**, arXiv:2103.01955.
14. Long, Q.; Zhou, Z.; Gupta, A.; Fang, F.; Wu, Y.; Wang, X. Evolutionary Population Curriculum for Scaling Multi-Agent Reinforcement Learning. *arXiv* **2020**, arXiv:2003.10423.
15. Dresner, K.; Stone, P. A multiagent approach to autonomous intersection management. *J. Artif. Intell. Res.* **2008**, *31*, 591–656. [CrossRef]
16. Milanes, V.; Perez, J.; Onieva, E.; Gonzalez, C. Controller for Urban Intersections Based on Wireless Communications and Fuzzy Logic. *IEEE Trans. Intell. Transp. Syst.* **2010**, *11*, 243–248. [CrossRef]
17. Bian, Y.; Li, S.E.; Ren, W.; Wang, J.; Li, K.; Liu, H.X. Cooperation of Multiple Connected Vehicles at Unsignalized Intersections: Distributed Observation, Optimization, and Control. *IEEE Trans. Ind. Electron.* **2020**, *67*, 10744–10754. [CrossRef]
18. Xu, H.; Zhang, Y.; Li, L.; Li, W. Cooperative Driving at Unsignalized Intersections Using Tree Search. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 4563–4571. [CrossRef]
19. Vaio, M.D.; Falcone, P.; Hult, R.; Petrillo, A.; Salvi, A.; Santini, S. Design and Experimental Validation of a Distributed Interaction Protocol for Connected Autonomous Vehicles at a Road Intersection. *IEEE Trans. Veh. Technol.* **2019**, *68*, 9451–9465. [CrossRef]
20. Nan, J.; Deng, W.; Zheng, B. Intention Prediction and Mixed Strategy Nash Equilibrium-Based Decision-Making Framework for Autonomous Driving in Uncontrolled Intersection. *IEEE Trans. Veh. Technol.* **2022**, *71*, 10316–10326. [CrossRef]
21. Aradi, S. Survey of deep reinforcement learning for motion planning of autonomous vehicles. *IEEE Trans. Intell. Transp. Syst.* **2020**, *23*, 740–759. [CrossRef]
22. Isele, D.; Rahimi, R.; Cosgun, A.; Subramanian, K.; Fujimura, K. Navigating Occluded Intersections with Autonomous Vehicles Using Deep Reinforcement Learning. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia, 21–25 May 2018; pp. 2034–2039.
23. Lin, Y.; McPhee, J.; Azad, N.L. Comparison of Deep Reinforcement Learning and Model Predictive Control for Adaptive Cruise Control. *IEEE Trans. Intell. Veh.* **2021**, *6*, 221–231. [CrossRef]
24. Shi, Y.; Liu, Y.; Qi, Y.; Han, Q. A Control Method with Reinforcement Learning for Urban Un-Signalized Intersection in Hybrid Traffic Environment. *Sensors* **2022**, *22*, 779. [CrossRef]
25. Liu, Y.; Liu, G.; Wu, Y.; He, W.; Zhang, Y.; Chen, Z. Reinforcement-Learning-Based Decision and Control for Autonomous Vehicle at Two-Way Single-Lane Unsignalized Intersection. *Electronics* **2022**, *11*, 1203. [CrossRef]
26. Mao, F.; Li, Z.; Lin, Y.; Li, L. Mastering Arterial Traffic Signal Control with Multi-Agent Attention-Based Soft Actor-Critic Model. *IEEE Trans. Intell. Transp. Syst.* **2022**, *24*, 3129–3144. [CrossRef]

27. Chen, D.; Li, Z.; Wang, Y.; Jiang, L.; Wang, Y. Deep multi-agent reinforcement learning for highway on-ramp merging in mixed traffic. *arXiv* **2021**, arXiv:2105.05701.
28. Guan, Y.; Ren, Y.; Li, S.E.; Sun, Q.; Luo, L.; Li, K. Centralized Cooperation for Connected and Automated Vehicles at Intersections by Proximal Policy Optimization. *IEEE Trans. Veh. Technol.* **2020**, *69*, 12597–12608. [CrossRef]
29. Zhou, M.; Yu, Y.; Qu, X. Development of an efficient driving strategy for connected and automated vehicles at signalized intersections: A reinforcement learning approach. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 433–443. [CrossRef]
30. Antonio, G.P.; Maria-Dolores, C. Multi-Agent Deep Reinforcement Learning to Manage Connected Autonomous Vehicles at Tomorrows Intersections. *IEEE Trans. Veh. Technol.* **2022**, *71*, 7033–7043. [CrossRef]
31. Duan, X.; Jiang, H.; Tian, D.; Zou, T.; Zhou, J.; Cao, Y. V2I based environment perception for autonomous vehicles at intersections. *China Commun.* **2021**, *18*, 1–12. [CrossRef]
32. Leurent, E. An Environment for Autonomous Driving Decision-Making. GitHub Repository. 2018. Available online: https://github.com/eleurent/highway-env (accessed on 31 December 2018).