# Deep Reinforcement Learning for Autonomous Vehicles Collaboration at Unsignalized Intersections

Jian Zheng, Kun Zhu, Ran Wang
College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China
Emails: {161640226, wangran, zhukun}@nuaa.edu.cn

GLOBECOM 2022 - 2022 IEEE Global Communications Conference | 978-1-6654-3540-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/GLOBECOM48099.2022.10001133

***Abstract*—As conservative intersection management, signalized intersection has a significant bottleneck in improving traffic efficiency when it comes to connected autonomous vehicles (CAVs). In this paper, to make the intersection management more fine-grained, a decentralized conflict-free coordination scheme is tailed for CAVs at intersections without traffic signals. First, the problem of multiple vehicles navigation through an unsignaled intersection is formulated as a Partially Observable Stochastic Game (POSG). Second, we propose a cooperative multi-agent proximal optimization algorithm (CMAPPO) to make driving-decision for each CAV agent and achieve collaboration in a distributed manner. Finally, simulations are carried out on SUMO to evaluate the proposed method. The results show that the CMAPPO has significant effectiveness in solving the multi-vehicle coordination at intersections.**

***Index Terms*—Multi-agent Coordination, Connected and Autonomous Vehicles, Reinforcement Learning, Unsignalized Intersection, Decision-making.**

## I. Introduction

THE connected autonomous vehicles (CAVs) are regarded as the future of the transportation system [1]. CAVs are a new type of intelligent vehicle that integrates the connected perception capabilities of the connected vehicles (CVs) and the self-driving decision-making capabilities of the autonomous vehicles (AVs). Once CAVs become a reality, they will reshape the existing transportation system. Compared with ordinary highways, intersections, as the crux of the road network, are more complicated and challenging for collaboration among vehicles. Traffic signal control is one of the most classic and effective solutions to multiple vehicles cooperation at intersections via the optimization of traffic signal timing [2]–[4]. However, with the soaring traffic flow on the road, the existing deployed traffic lights may not meet the requirements of real-time intersection management, because of the limitation of lane-level intersection regulation.

The emergence of CAVs enables vehicle-road coordination via vehicle-to-everything (V2X) communications, which makes it possible to develop innovative and efficient traffic management solutions in the modern intelligent transportation systems. Such as unsignalized control can realize vehicle-level intersection management through multi-vehicle collaboration, which is more real-time and accurate than signalized control. In research [5], traffic flow is divided into batches and collaboration within one batch is done by centralized optimal control, then the batches are managed via reservation-based approach. Pan et al. adopted the Decentralized Model Predictive Control (DMPC) method through convex modeling method to solve multi-vehicle coordination problem at unsignaled intersections [6]. Although existing works have effectively alleviated traffic congestion in the context of unsignalized control, their predefined rules or models are limited by the traffic uncertainty, which becomes a bottleneck for performance improvement.

Reinforcement learning (RL) can find out optimal policy by interacting with the environment without relying on specific rules or models, which make it competent for the tasks of autonomous driving decision-making at intersections. Kai et al. unified the three-direction navigation at intersection via a multi-task RL algorithm [7]. The end-to-end architecture developed in [8] can obtain the mapping table between local observations containing traffic images and autonomous vehicle control actions. David et al. presented two action space representations in the framework of deep Q-networks (DQN) to learn active perception behaviors to enable safe navigation at occluded intersections [9]. Seong et al. introduced an attention mechanism to enable policy to learn to focus on more important spatial and temporal features in its egocentric observations [10].

In the above studies, these single-agent RL algorithms are used to control a single autonomous vehicle to pass the intersection, they only concentrate on one agent, but the transportation is inherently multi-agent system. This paper focus on the scenario of multiple CAVs collaboration at unsignalized intersections. Then a multi-agent RL method is proposed for the task of multi-vehicle navigation through unsignalized intersections. The main contribution of this paper is, thus, the problem of cooperatively scheduling of CAVs passing an unsignaled intersection safely, efficiently and comfortably is modeled as a partially observable stochastic game (POSG). Then the Cooperative Multi-agent Proximal Policy Optimization algorithm (CMAPPO) is proposed to seek the optimal control policy for multi-vehicle coordination at unsignalized intersections. Finally, the sufficient simulations are conducted, and the results show that our proposed method has significant effectiveness in solving the multiple vehicles coordination at signal-free intersections.

The remainder of this paper is structured as follows. Section

This work is supported by National Natural Science Foundation of China (62071230)

978-1-6654-3540-6/22/$31.00 ©2022 IEEE

II illustrates the problem statement and POSG formulation of multi-vehicle coordination at signal-free intersections. Section III proposes CMAPPO, which is an improved multi-agent RL algorithm based on proximal policy optimization (PPO). Section IV presents the experimental setting and analyzes the results and section V summarizes the work of this paper.

## II. Problem Statement And Formulation

### A. Problem Statement

In this paper, we focus on a signal-free and four-way intersection shown in Fig. 1. There are four entrances and four exits, all with only one lane. Each direction is respectively marked as right, left, up and down. The area covered by the intersection is specific. The deployed coordinator is not involved in decision-making, just serves as the transportation system database storing the intersection geometric and kinetic information of all CAVs. The intersection covers the control zone where the coordinator can communicate with CAVs and all CAVs can communicate with other. We call the convergence area of the roads in four directions of the intersection as the merge zone in which lateral collisions between vehicles may occur. The distance from the entry of the control zone to the entry of the merging zone is $L \in \mathbb{R}^+$, and it is the identical for all four entrances. The length of the merge zone is denoted as $D \in \mathbb{R}^+$. As depicted in Fig. 1, the CAVs enter the intersection area in different lanes and will cross the intersection according predetermined paths, including three types of going straight, turning left and turning right. Their trajectories may converge at conflicting points in the merging zone. In our problem of interest, the CAVs that are likely to collide with each other need to find appropriate gaps according to their time to arrive at the conflict points, making cooperative driving decisions to pass the intersection.

Several assumptions are set in our problem as follows. Firstly, the vehicle-to-vehicle and vehicle-to-infrastructure communication can be done without errors or delays, and the kinetic information of CAVs can be shared in real time. Then, all CAVs are homogeneous and their dynamics information such as location and velocity can be obtained to support driving decision-making to pass the intersection without collision.

### B. Problem Formulation

The problem of multi-vehicle collaboratively passing the intersection can be formulated as a POSG which extends stochastic game to problem with partially observability, partly on account of inadequate perception and communication capabilities, the physical embodiment of the agent [11]. Similar to Markov Games, POSG is re-defined as Partially Observable Markov Games(POMG) in the context of RL, as a tuple $\langle \mathcal{N}, \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{R} \rangle$ for $N$ agents. Unsignalized intersection coordination policy learning can be model as a POSG, in which each CAV acts as an agent with self-learning ability to improve its behavior by interacting with other CAVs. The detailed definition of state and observation space, action space and reward function as follows.

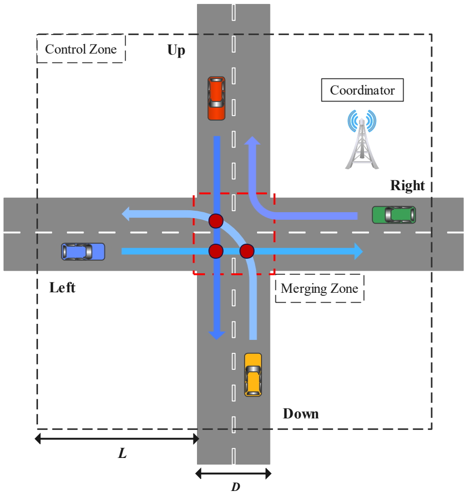


Fig. 1. An intersection with a coordinator communicating with CAVs inside the control zone.

*1) State and Observation Space:* According to the aforementioned assumptions, each CAV is able to obtain kinetic information of other CAVs at the intersection. In order realize the coordination among CAVs, the observation space of each CAV should include the dynamics of the series of vehicles most likely to collide with it. The vehicles introduced into the observation space are called the reference vehicles. The impact between vehicles can be measured by the difference in the distance from the vehicle to the center of merging zone. We choose $m$ vehicles with the greatest impact on CAV $i$ as its reference vehicles so as to make the dimension of observation space fixed. The observation space of CAV $i$ is defined as

$$\mathcal{O}_i = \{s^i_{\text{own}}, s^i_{\text{other},1}, ..., s^i_{\text{other},m}\}, \tag{1}$$

$$s^i_{\text{own/other},*} = \{v, x, y, \sin\theta, \cos\theta\}. \tag{2}$$

Each state is a tuple: $\langle v, x, y, \sin\theta, \cos\theta \rangle$ shown in (2), where $v, (x, y), \theta$ are, respectively, the speed, the Cartesian coordinate and the heading angle of the current vehicle.

*2) Action Space:* Considering that CAVs may trun right or left at the intersection, the driving decision should involve the acceleration and steering. Therefore, A hierarchical controller is applied to regulate the behaviors of CAVs when driving at intersection area. The high-level is based on the proposed RL-enabled driving decision-making policy to adjust the longitudinal acceleration, and the low-level is the tracking control model for lateral control. The lateral controller is used for two aspects of control, the position control and heading control. The expression of position control is represented as follows:

$$v_l^{ex} = -K_p d_l, \tag{3}$$

$$\Delta\theta = \arcsin(\frac{v_l^{ex}}{v}), \tag{4}$$

where $v_l^{ex}$ is the expected speed, $K_p$ is the position control gain and $d_l$ is the distance from the vehicle to the center line of the target lane. $\theta$, $v$ are the current heading angle and speed respectively. The heading control is calculated by the proportional-derivative controller as:

$$\theta = \theta_L + \Delta\theta, \tag{5}$$

$$\dot{\theta} = K_h(\theta^{ex} - \theta), \tag{6}$$

$$\delta = \arcsin(\frac{1}{2}\frac{l_r}{v}\dot{\theta}), \tag{7}$$

where $\theta_L$ is the lane heading, $\theta_{ex}$ is the target heading angle, and $K_h$ is the heading control gain. The steering angel $\delta$ is computed by (7), in which $l_r$ is the distance between the center of gravity of the front wheels.

The Actions determined by RL method are only used to control the longitudinal accelerations of CAVs. Hence, the continuous action space for longitude acceleration $a_i \in \mathcal{A}_i$:

$$a_i \in [-a_{max}, a_{max}]m/s^2, \tag{8}$$

where $a_{max}$ is the maximum acceleration. If $a_i$ is greater than zero, it means CAV $i$ accelerates, less than zero, it decelerates, and equal to zero it maintains the current speed.

*3) Centralized Reward Shaping:* In our hypothetical single-lane intersection scenario, we can make all vehicles at the intersection share rewards, as the centralized reward. According to the driving goal of passing the intersection efficiently, safely and comfortably. The proposed centralized reward is the sum of three parts: efficiency-related, comfortability-related and terminal-related ones:

(i) Efficiency-related reward: For the definition of intersection delay, we refer to the seminal work in [12]. There are $N$ vehicles crossing through the intersection within a period of time. Ideally, if there are no other vehicles, and the vehicle $i$ would pass the intersection at the maximum speed allowed within the expected time ($ET_i$). However, due to the adjustment of speed caused by emergence of other vehicles, the actual travel time is $AT_i$. Thus, the intersection travel delay is computed by:

$$T_{delay} = \sum_{CAV i \in N} (AT_i - ET_i). \tag{9}$$

In order to improve the traffic efficiency of the intersection, consistent with minimizing the intersection travel delay, so we set the negative value of (9) as a reward. Assuming that the distance traveled by CAV $i$ with the speed of $v_i$ in step $k$ is $v_i\Delta t$, then the expected travel time with the maximum speed $V_m$ is $v_i\Delta t/V_m$, so the efficiency-related reward can be calcalated as:

$$R_v = \sum_{CAV i \in N} -(\Delta t - \frac{v_k^i \Delta t}{V_m}). \tag{10}$$

(ii) Comfortability-related reward: To improve travel comfortability, we use the $L^2$-norm of acceleration actuated by the vehicles as a penalty to reduce the degree of speed change and improve travel comfortability.

$$R_c = \sum_{CAV i \in N} -\frac{\|a_k^i\|^2}{\|a_{max}\|}. \tag{11}$$

(iii) Terminal-related reward: Termination conditions include CAVs passing through the intersection successfully, collision and simulation ending:

$$R_t = \begin{cases} +2, & \text{if all CAVs success;} \\ -5, & \text{if collision occurs;} \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

The final Centralized reward is the weighted sum of the above rewards shown as:

$$r_k = w_1 R_v + w_2 R_c + w_3 R_t, \tag{13}$$

where $w_1, w_2, w_3 \in \mathbb{R}^+$ are the weight coefficients of the corresponding components and need to be fine-tuned according to research purpose and testing needs. Therefore, several CAVs in the intersection area share the same reward function according to (13) with the aim of cooperatively passing through the intersection.

## III. Cooperative Multi-agent Algorithm

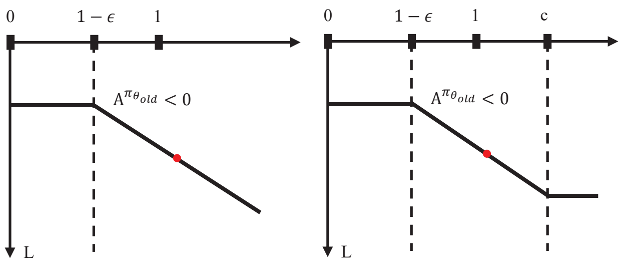
In the above POSG, it can be found that it is difficult for CAVs to obtain complete model of the state transition of surrounding environment. Therefore, we need a model-free algorithm that does not require the transition functions for all states to solve the resulting problem. We solve the formulated POSG by resorting to the dual-clip PPO algorithm and adopt the paradigm of centralized training and decentralized learning to stabilize the learning process.

### *A. Dual-clipped PPO*

The standard PPO algorithm in [13] adopts the objective function involving an importance sampling ratio clip is shown as,

$$\begin{aligned} \mathcal{L}^{\text{ppo}}(\theta) = \mathbb{E}_{s,a}[\min(&r_t(\theta)A^{\pi_{\theta_{\text{old}}}}, \\ &\text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)A^{\pi_{\theta_{\text{old}}}})], \end{aligned} \tag{14}$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ is the probability ratio between new and old policies, The motivation of the objective has been revealed, that is, to keep $r_t(\theta)$ within a small interval around 1, accurately $[1-\epsilon, 1+\epsilon]$, and as a result, the incentive of the large policy update is removed. However, in off-policy training environments such as autonomous driving decision-making scenario, the sampled trajectories come from various policies, which may have a certain deviation from the current policy $\pi_\theta$.



(a)

(b)

Fig. 2. (a) Standard PPO (clip with $\epsilon$); (b) Dual-clip PPO (clip with $\epsilon$ and $c$ when $A^{\pi_{\theta_{\text{old}}}} < 0$)

**Algorithm 1** CMAPPO Based Multi-vehicle Driving Cooperation at Unsignalized Intersections.

```
Input:
Intersection Environment $\mathbf{E}$;
Observation Space $\mathcal{O}$;
Action Space $\mathcal{A}$;
Process:
Initialize the actor network parameters $\theta$, critic network parameters $\omega$ and memory bufer $\mathcal{B}$ of each CAV agent;
for each iteration do
    while CAVs are within the control zone do
        for $t = 0$ to $T$ do
            for each CAV agent $i$ do
                Observe $o_i[t]$ and choose action $a_i[t]$ based on policy $\pi_i$
            end for
            All CAV agents take joint action $\mathcal{A}_t$ to interact with the environment and receive a centralized reward $r_t$
            for each CAV agent $i$ do
                Share the centralized reward $r_t$ and store tuples $\tau = (o_i[t], a_i[t], r_t, o_i[t+1])$ to the buffer $\mathcal{B}_i$
            end for
        end for
    end while
    for each CAV agent $i$ do
        Calculate cumulative discounted reward $\hat{R}_i[t]$ in buffer $\mathcal{B}_i$
        Compute advantages $\hat{A}_i[t]$ using GAE
        Update the policy by Adam:
            $\theta_i[t+1] = \arg\max_{\theta} \frac{1}{|\mathcal{B}_i|T} \sum_{\tau \in \mathcal{B}_i} \sum_{t=0}^{T} L(\theta_i)$
        Update the value function by Adam:
            $\omega_i[t+1] = \arg\min_{\omega} \frac{1}{|\mathcal{B}_i|T} \sum_{\tau \in \mathcal{B}_i} \sum_{t=0}^{T} L(\omega_i)$
    end for
end for
Output: Policies for all CAV agents
```

The standard PPO generally works, but it ignores the case of $A^{\pi_{\theta_{\text{old}}}} < 0$. For example, when $\pi_{\theta_{\text{old}}}(a_t|s_t) \ll \pi_\theta(a_t|s_t)$, then the ratio $r_t(\theta)$ is very large. At this time, such a huge ratio $r_t(\theta)$ will bring a huge and unbounded variance because $r_t(\theta)A^{\pi_{\theta_{\text{old}}}} \ll 0$. Therefore, using the objective of PPO cause the new policy to be significantly different from the old policy, which will make it difficult to guarantee the policy convergence. We thus adopt a dual-clip PPO algorithm proposed by Ye et al., which is used in the complexe action control of Multi-player Online Battle Arena(MOBA) [14]. It clips the ratio $r_t(\theta)$ by setting a hyperparameter as the lower bound of $r_t(\theta)A^{\pi_{\theta_{\text{old}}}}$, as shown in (14). When $A^{\pi_{\theta_{\text{old}}}} < 0$, the objective function of dual-clipped PPO is:

$$\mathbb{E}_{s,a}\left[\max(\min(r_t(\theta)A^{\pi_{\theta_{\text{old}}}}, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)A^{\pi_{\theta_{\text{old}}}}), \right. \\ \left. cA^{\pi_{\theta_{\text{old}}}}\right],$$

in which $c > 1$ and is a constant indicating the lower bound.

### B. Generalized Advantage Estimation

Using the advantage function is an extremely important strategy for DRL, especially for the PPO policy gradient and computed by

$$\hat{A}(s_t, a_t) = \hat{Q}(s_t, a_t) - V(s_t),$$

where $\hat{Q}(s,a)$ is the state value function estimated by sampling, $V(s)$ is the expression of the state value function, and the difference between the two is used to characterize the advantage function. Several policy-based methods approximate $\hat{Q}(s,a)$ using Monte Carlo methods, i.e.,

$$\hat{Q}(s_t, a_t) = \sum_{k=0}^{\infty} \gamma^k r_{t+k},$$

which has no bias but suffers from excessive variance. TD$(0)$ methods construct $\hat{Q}^\pi(s,a)$ through one-step bootstrapping, i.e.,

$$\hat{Q}(s_t, a_t) = r_t + \gamma V(s_{t+1}),$$

which is biased but with low variance.

Generalized advantage estimation (GAE) [15] borrows the idea from TD$(\lambda)$ and trades off bias and variance through n-step bootstrapping, embodying a compromise from one-step bootstrapping to Monte Carlo methods, which is as follows.

$$\hat{A}_t^{\text{GAE}(\gamma,\lambda)} = \sum_{k=0}^{\infty} (\gamma\lambda)^k \delta_{t+k},$$

where $\delta_{t+k}$ denotes the TD error,

$$\delta_{t+k} = r_{t+k} + \gamma V(s_{t+k+1}) - V(s_{t+k}).$$

### C. Cooperative Multi-agent PPO

The PPO is mainly applied to solve the single-agent RL problem, in order to make it suitable for the multi-vehicle collaboration problem at signal-free intersection, this paper makes some adaptations of PPO. Based on dual-clip PPO and adopting a decentralized structure, our proposed cooperative multi-agent PPO (CMAPPO) trains a pair of critic network and actor network for each CAV agent, allowing all CAV agents to share a centralized reward. To achieve cooperation between vehicles and overcome non-stationarity in multi-agent environments, CMAPPO adopts the paradigm of centralized training and distributed execution (CTDE). Similar to MADDPG [16] and COMA [17], each agent learns a centralized value function. During training, critic can obtain the observations of the reference vehicles and guide the actor network to update. At execution time the agent acts only based on its own observations.

The pseudocode of CMAPPO is shown in Algorithm 1. At first, we construct the corresponding critic network and actor network and initialize network parameters for each CAV agent, and set up replay buffers to store experience. After starting training, each CAV agent feeds its own observation into the actor network and chooses action at each time step. Then all agents take a joint action to interact with the environment and get a immediate centralized reward $r_t$ according to (13), and the environment moves to the next state. Each CAV agent exploits its policy to adjust the longitudinal acceleration. The critic network approximating the state value function $V_\omega(s_i^t)$ has the same structure as that of the actor network estimating the policy function $\pi_\theta$. The output of the critic network is a component of the actor network loss function, which is used to approximate the advantage $\hat{A}_i[t]$ via GAE. The actor network generates the policy, and the critic network evaluates the current policy through $\hat{A}_i[t]$.

When the network parameters need to be updated, the sampled experiences in the buffer are used to construct the loss function which consists of two parts, i.e, $L(\theta_i)$ and $L(\omega_i)$. The loss function of the critic network $L(\omega_i)$ is represented as

$$L(\omega_i) = (V_{\omega_i[t]}(o_i[t], o_1[t], ..., o_m[t]) - R_i[t])^2, \qquad (15)$$

which is in the form of mean squared error and optimized by Adam optimizer. In our multi-agent off-line training setting, we find that actions with negative advantage can negatively affect the policy. In particular, when the policy ratio $r_t(\theta)$ is a large value, such a huge ratio will lead to unbounded variance due to $r_t(\theta)\hat{A}_{\theta[t]} \ll 0$. Obviously, old and new policies will diverge, making it challenging to ensure policy convergence. Therefore, we adopt the dual-clip PPO based algorithm, and the loss function of the actor network is shown as

$$\begin{aligned} L(\theta_i) = \mathbb{E}_{s,a}[\max(\min(r_t(\theta)\hat{A}_{\theta[t]}, \text{clip}\,(r_t(\theta), \\ 1-\epsilon, 1+\epsilon)\,\hat{A}_{\theta[t]}), c\hat{A}_{\theta[t]})], \end{aligned} \qquad (16)$$

when $\hat{A}_{\theta[t]} < 0$. Otherwise, calculate $L(\theta_i)$ according to (14).

## IV. Simulation Results and Analysis

In this section, the experiments that we conduct are all evaluated in SUMO [18], where the intersection without traffic signal scenario is shown in Fig. 1. The intersection contains four directions and allows CAVs to go straight, turn right, and turn left. In addition, all vehicles are spawned at a random distance $d^i \sim \mathcal{N}(\mu = 70m, \sigma = 3m)$ from the intersection center with an initial velocity. When they reach the central area of the intersection, they choose one of the directions of going straight, turning right and turning left according to a uniform distribution. Considering a small intersection in our experiments, all vehicles are in low-speed mode with specific speed constraints. Table I shows the parameter settings in detail. With time step of $0.2s$ in simulation, four CAVs appear in the lanes of different directions. The actor network outputs two values to parameterize the normal distribution of the acceleration action for each CAV, while the critic network outputs only one value to approximate to the value of current state. And we use the Adam optimizer with learning rate of $LR = 0.0001$. In addition, according to multiple trials, the policies of multi-agent system have stabilized after training 800 iterations.

In this experiment, we train multi-agent policies with CMAPPO and IPPO in the same intersection environment and compare the two in terms of convergence and traffic performance. In IPPO algorithm, all agents use PPO to train their own policies in a decentralized manner. Fig. 3 illustrates the results of replay stable policy after 500 training iterations showing success multiple vehicles collaboration. In this episode, CAV1, CAV2 and CAV3 reach the intersection center almost simultaneously at timestep 24, only CAV4 slows down until timestep 8. From Fig. 3(c) we can see that CAV4 keeps accelerating to the maximum speed after timestep 8, and reaches the center of the intersection at timestep 33, which shows that CAV4 has learned the policy of decelerating in advance to avoid other vehicles. The acceleration action curves shown in 3(d) is no longer random, where three vehicles take seemingly synchronized acceleration actions, and CAV4 only maintains accelerating after it completes its avoidance. In other words, the velocity profiles of CAV1, CAV2, and CAV3 illustrate that they have a tendency to maintain a high velocity so as to cross the intersection quickly, and when there is no risk of collision with other vehicles at intersection center, CAV4 then accelerates to the maximum speed and pass through the intersection as well.

TABLE I
Hyperparameters of the experiments.

| Parameter | Value |
|---|---|
| *Simulator* | |
| Lane length | $100m$ |
| Vehicle length | $5m$ |
| Initial position | $\mathcal{N}(\mu = 70, \sigma = 3)m$ |
| Initial velocity | $10m/s$ |
| Velocity constrain | [$5m/s$,$15m/s$] |
| Acceleration | [-$3m/s^2$,$3m/s^2$] |
| *CMAPPO & IPPO* | |
| Discount factor $\gamma$ | 0.99 |
| GAE lambda $\lambda$ | 0.95 |
| Clip range $\epsilon$ | 0.2 |
| Total iterations | 800 |
| Time horizon per iteration | 6000 |
| Seed number | 112 |
| Batch size $B$ | 1024 |
| Minibatch size $MB$ | 32 |
| Epoch $U$ | 10 |
| SGD iterations | 10 |
| Learning rate $LR$ | 0.0001 |
| Hidden layers | $32 \times 32 \times 32$ |
| Optimizer | Adam |

Fig. 4(a) illustrates the average episode reward of CMAPPO and IPPO over training process. We can clearly find that both CMAPPO and IPPO obtain the highest cumulative rewards near 2, which stands for that all CAVs pass the intersection safely. Compared with IPPO, CMAPPO converges slightly faster and is more stable. The success rate and collision rate curves of vehicles passing through intersections shown in Fig. 4(c) and Fig. 4(d) respectively, tend to be consistent with the corresponding average reward curves. The average delay curve in Fig. 4(b) illustrates the difference between the time all vehicles actually spend and the time they expect to pass through the intersection. Compared with IPPO, CMAPPO always has lower travel delay during the training process and tends to stabilize faster, which shows that CMAPPO is more effective in improving traffic efficiency.

## V. Conclusion

In this paper, we present a decentralized conflict-free collaboration scheme of multiple CAVs at signal-free intersections

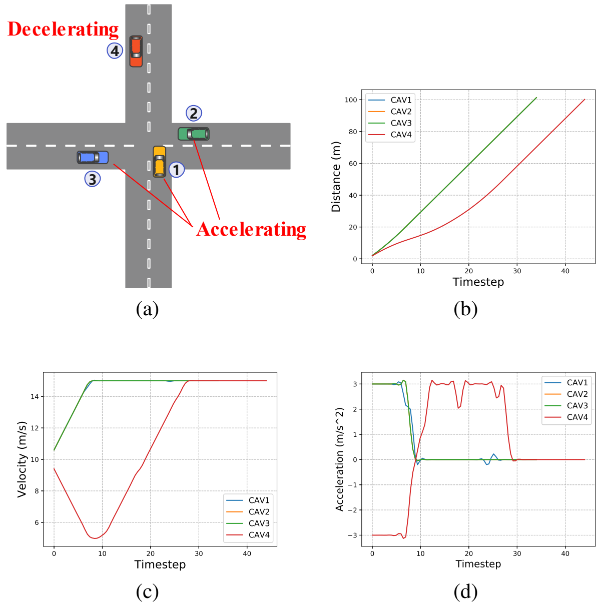


Fig. 3. Results from the replay of the policy after 500 training iterations, including the distance traveled, the velocity and acceleration of profiles each CAV during one episode. (a) An episode in which all CAVs pass successfully. (b) Distance (c) Velocity. (d) Acceleration.

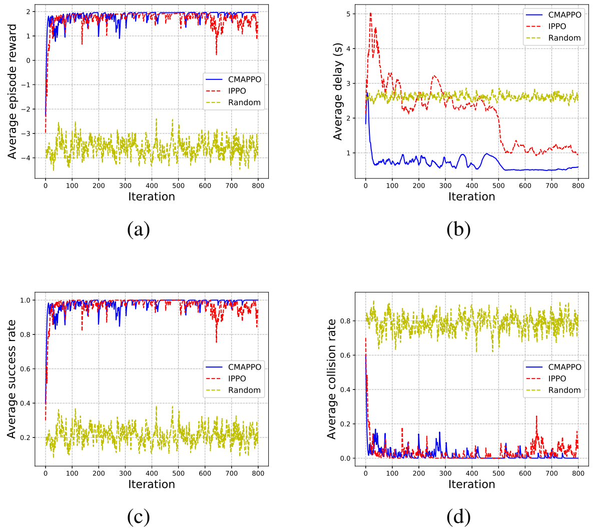


Fig. 4. Comparison of CMAPPO, IPPO and Random during total 800 iterations training process. (a) Average episode reward. (b) Average delay. (c) Average success rate. (d) Average collision rate.

using RL to improve traffic efficiency and comfort. We model the problem of CAVs cooperatively passing a unsignalized intersection as a POSG, where each CAV acts as an agent and each acceleration decision corresponds to the action taken by the CAV. Furthermore, we propose a multi-vehicle collaboration algorithm based on our proposed CMAPPO to seek the optimal policy for each CAV. Finally, we train the co-operative multi-agent policy in a typical four-direction single-lane intersection environment using the SUMO simulator. The simulation results show that the trained policy can ensure that all CAVs pass through the intersection safely, efficiently and comfortably.

## References

[1] D. J. Fagnant and K. Kockelman, "Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations," *Transportation Research Part A: Policy and Practice*, vol. 77, pp. 167–181, 2015.

[2] H. Wang, M. Zhu, W. Hong, C. Wang, G. Tao, and Y. Wang, "Optimizing signal timing control for large urban traffic networks using an adaptive linear quadratic regulator control strategy," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 1, pp. 333–343, 2022.

[3] G. Long, A. Wang, and T. Jiang, "Traffic signal self-organizing control with road capacity constraints," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–10, 2022.

[4] C. Jiang, X. Hu, and W. Chen, "An urban traffic signal control system based on traffic flow prediction," in *2021 13th International Conference on Advanced Computational Intelligence (ICACI)*, 2021, pp. 259–265.

[5] B. Li, Y. Zhang, T. Acarman, Y. Ouyang, C. Yaman, and Y. Wang, "Lane-free autonomous intersection management: A batch-processing framework integrating reservation-based and planning-based methods," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 7915–7921.

[6] X. Pan, B. Chen, L. Dai, S. Timotheou, and S. A. Evangelou, "Decentralized model predictive control for automated and connected electric vehicles at signal-free intersections," in *2021 60th IEEE Conference on Decision and Control (CDC)*, 2021, pp. 2659–2664.

[7] S. Kai, B. Wang, D. Chen, J. Hao, H. Zhang, and W. Liu, "A multi-task reinforcement learning approach for navigating unsignalized intersections," in *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2020, pp. 1583–1588.

[8] G. Li, S. Li, S. Li, Y. Qin, D. Cao, X. Qu, and B. Cheng, "Deep reinforcement learning enabled decision-making for autonomous driving at intersections," *Automotive Innovation*, vol. 3, no. 4, pp. 374–385, 2020.

[9] D. Isele, R. Rahimi, A. Cosgun, K. Subramanian, and K. Fujimura, "Navigating occluded intersections with autonomous vehicles using deep reinforcement learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 2034–2039.

[10] H. Seong, C. Jung, S. Lee, and D. H. Shim, "Learning to drive at unsignalized intersections using attention-based deep reinforcement learning," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, 2021, pp. 559–566.

[11] P. Palanisamy, "Multi-agent connected autonomous driving using deep reinforcement learning," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–7.

[12] K. Dresner and P. Stone, "Multiagent traffic management: A reservation-based intersection control mechanism," in *Autonomous Agents and Multiagent Systems, International Joint Conference on*, vol. 3. IEEE Computer Society, 2004, pp. 530–537.

[13] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017.

[14] D. Ye, Z. Liu, M. Sun, B. Shi, P. Zhao, H. Wu, H. Yu, S. Yang, X. Wu, Q. Guo *et al.*, "Mastering complex control in moba games with deep reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 6672–6679.

[15] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," *arXiv preprint arXiv:1506.02438*, 2015.

[16] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *NIPS*, 2017.

[17] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[18] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wießner, "Microscopic traffic simulation using sumo," in *The 21st IEEE International Conference on Intelligent Transportation Systems*. IEEE, 2018. [Online]. Available: https://elib.dlr.de/124092/