Safe and Robust Multi-Agent Reinforcement Learning for Connected Autonomous Vehicles under State Perturbations

Zhili Zhang

Yanchao Sun

Furong Huang

Fei Miao

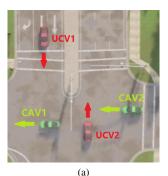
Abstract-Sensing and communication technologies have enhanced learning-based decision making methodologies for multi-agent systems such as connected autonomous vehicles (CAV). However, most existing safe reinforcement learning based methods assume accurate state information. It remains challenging to achieve safety requirement under state uncertainties for CAVs, considering the noisy sensor measurements and the vulnerability of communication channels. In this work, we propose a Robust Multi-Agent Proximal Policy Optimization with robust Safety Shield (SR-MAPPO) for CAVs in various driving scenarios. Both robust MARL algorithm and control barrier function (CBF)-based safety shield are used in our approach to cope with the perturbed or uncertain state inputs. The robust policy is trained with a worst-case Q function regularization module that pursues higher lower-bounded reward in the former, whereas the latter, i.e., the robust CBF safety shield accounts for CAVs' collision-free constraints in complicated driving scenarios with even perturbed vehicle state information. We validate the advantages of SR-MAPPO in robustness and safety and compare it with baselines under different driving and state perturbation scenarios in CARLA simulator. The SR-MAPPO policy is verified to maintain higher safety rates and efficiency (reward) when threatened by both state perturbations and unconnected vehicles' dangerous behaviors.

I. INTRODUCTION

The application of machine learning models assisted by more accurate on-board sensors, such as camera and LiDARs have enabled intelligent driving to a certain degree. Meanwhile, advances in wireless communication technologies also make it possible for information sharing beyond an individual's perception [1], [2]. Through vehicle-to-everything (V2X) communications, it has been shown that such shared information can contribute to connected autonomous vehicles' (CAVs) decision-making [3], [4], [5], and improve the safety and coordination of CAVs [6], [7], [8].

However, noisy sensor measurements or vulnerable communication channel may lead to uncertain or perturbed state inputs to agents' learning-based decision models and should not be ignored [9], [10]. As is shown in Fig. 1b during our testing, the perturbed state of unconnected vehicle (UCV1) in disguise is received by agent 1 and input to its policy. It successfully deceives the non-robust agent and results in an accident. Motivated by such imperfections regarding robustness of the current safe MARL approaches, we aim to develop a safe and robust MARL algorithm for CAV

This work was supported by NSF 1932250 and NSF 2047354 grants. Z. Zhang, and F. Miao are with the Department of Computer Science and Engineering, University of Connecticut, Storrs Mansfield, CT, USA 06268. Y. Sun, and F. Huang are with the Department of Computer Science, University of Maryland, College Park, MD, USA 20742. Email: {zhili.zhang, fei.miao}@uconn.edu., {ycs, furongh}@umd.edu



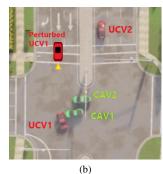


Fig. 1. *Intersection* (a, b): two unconnected vehicles run the red light when two CAVs are passing the box in *Intersection* scenario. Connected autonomous vehicles are in green; unconnected vehicles are in red. (1a): CAVs successfully avoid collision 1a; (1b): CAV2 collides with an UCV in 1b because the perturbed location of UCV1 (with yellow triangle) mislead the CAVs to consider "no UCVs in the intersection" as the current state.

systems that can handle challenging driving scenario while withstanding input state perturbations.

In this work, we design the Safe and Robust MAPPO algorithm with robust Safety Shield (SR-MAPPO) for CAVs' cooperative policy-learning in driving environments with state perturbations. To handle the potential errors and uncertainties within the state inputs, we adopt a worst-case aware Q network [11] as a robustness module for each CAV agent. Specifically, our approach can cope with challenging driving scenarios in intersection and highway with unconnected vehicles (UCVs) posing threats of collision, without having to generate these safety critical situations during training; meanwhile, ego or shared observations as state inputs suffer from bounded error as defined in Sec. III-B. We adopt the MARL framework with each CAV learning its robust policy through PPO and the worst-case Q (Fig. 2). We also equip each agent with the Safety Shield based on robust Control Barrier Functions (CBFs) to enhance safety and robustness against the error in vehicle dynamic state. We further introduce our reward design for MARL, which considers both rewards from achieving its driving goals and the safety rewards as the feedback from the Safety Shield. In summary, the main contributions of this work are:

- We design the Safe and Robust Multi-Agent Proximal Policy Optimization algorithm with robust Safety Shield for CAV system to enhance safety and efficiency of agents under state perturbation.
- We define the robust Control Barrier Functions for CAV system which can withstand bounded state perturbations and is applicable on common driving environments.
- We validate through experiment on CARLA simulator that the proposed Safe-Robust MARL framework

significantly improves the collision-free rate and the CAV agents achieves higher overall returns compared to baselines. Our results show contributions from robust MARL algorithm and robust Safety Shield respectively by ablation study.

II. RELATED WORK

a) Planning and Decision-Making of Autonomous Vehicles: A branch of robust control methods for autonomous vehicles includes rule-based methods. Upgrading from CBF-based approaches [12], measurement uncertainty was also considered in CBF formulation solution in [13] for a single-agent driving problem, while this approach assumes accurate model of the system and also require extra effort to adapt to new environment compared to learning-based solutions. End-to-end autonomous driving using deep learning [14] also considers robustness that targets noisy input from vision-based sensors for one single agent. Yet end-to-end approaches without explicit safety requirements has inherently questionable capability of safety maintenance for multi-agent systems (CAV) under complicated interactions.

b) Safe RL: Different approaches in Safe RL have been proposed to guarantee or improve safety of the system, such as defining a safety module assisting RL or MARL algorithm in either training or execution stage [15], learning a separate 'risk' network aiming to shape and learn a safe policy with constrained policy optimization [16], and decentralized policy optimization with safety constraints [17], etc. For multi-agent CAV system, [8] adopts an individual safety-checking module evaluating behavior safety; [18] uses Signal Temporal Logic to generate rewards and guide policy learning for safety requirements' satisfaction. However, the above works assume accurate state inputs to RL or MARL algorithm from the driving environment and cannot tolerate noisy or inaccurate state input. In this work, we consider safe and robust decision-making for multi-agent CAV system in complicated driving environments; the robust MARL and robust Safety Shield work collaboratively to enhance agents' safety and robustness to input state perturbations.

c) Robust RL: Robust RL and robust MARL have made progress in theories and algorithm design recently. In [11], the authors propose a general framework for robust RL, using a worst-case-aware action-value function to evaluate the lower-bound of return with perturbed state so as to raise robustness to perturbation. Theoretical analysis of robust MARL solutions under state perturbations have been analyzed in [19], [20]. Despite the current effort mentioned above on exploring possibilities of robust RL, the algorithms are mostly applied or tested on simple RL environments without hard safety requirements or complex interactions among multi-agents. Considering both safety and robustness requirements with complicated dynamics or maneuvers in single or multi-agent autonomous driving environment still remain challeging. We design robust MARL algorithm that does not require adversaries during training; meanwhile the robust Safety Shield checks safety requirements and returns feedback to MARL with safety reward.

III. PROBLEM FORMULATION

A. Problem Description

We consider the robust cooperative policy-learning problem under uncertain state inputs for CAVs in mixed traffic environments including unconnected vehicles that do not communicate or coordinate with CAVs, and various driving scenarios such as multi-lane intersection and highway (as shown in Fig.1&4). We assume that each CAV can get shared information from V2V and V2I communications. We consider that a CAV agent i has accurate self-observation of its driving state but potentially perturbed observations of the other vehicles; and the two parts collectively constitute its state s_i in reinforcement learning, explained in Sec. III-B.

B. State-perturbed MARL Problem Formulation for CAV

A State-perturbed Multi-Agent Reinforcement Learning for CAV is defined as a tuple $G = (S, A, P, \{r_i\}, e_l, e_v, \mathcal{G}, \gamma)$ where $\mathcal{G} := (\mathcal{N}, \mathcal{E})$ is the communication network of all CAV agents. S is the joint state space of all agents: $\mathcal{S} := \mathcal{S}_1 \times \cdots \times \mathcal{S}_n$. The state space of agent $i: \mathcal{S}_i = \mathcal{S}_i$ $\{o_i, o_{\mathcal{N}_i}, o_{\mathcal{N}^{UV}}\}$ contains self-observation o_i , observations $o_{\mathcal{N}_i} = \{o_j | j \in \mathcal{N}_i\}$ shared by neighbor connected agents \mathcal{N}_i and lastly observations $o_{\mathcal{N}^{UV}}$ of unconnected vehicles \mathcal{N}_{i}^{UV} made by either agent i itself or shared by other agents or infrastructures. Observations in $\{o_i\} \bigcup o_{\mathcal{N}_i}$ each contains location, velocity, acceleration and Lane-detection (l, v, α, LD) , a CAV's measurement; while observations of unconnected vehicles \mathcal{N}_i^{UV} contains only location and velocity $(\boldsymbol{l}, \boldsymbol{v})$ of the vehicle. Here $\boldsymbol{l} = (l_x, l_y)$ and $\boldsymbol{v} = (v_x, v_y)$ is defined in 2D space, x is vehicle's traveling direction and yis perpendicular to x. In this work, we consider perturbed observation $\tilde{o} = (\tilde{l}_x, l_y, \tilde{v}_x, v_y), \tilde{l}_x = l_x + e_l(l_x), \tilde{v}_x =$ $v_x + e_v(v_x)$ on location l_x and velocity v_x along x-direction.The perturbation $e := (e_l, e_v)$ defined by bounded errors e_l, e_v is further explained in Sec. V-A. In concise, agent i's state s_i contains accurate o_i but potentially perturbed $o_{\mathcal{N}_i}$ and $o_{\mathcal{N}^{UV}}$, whose perturbations are only on vehicle's traveling direction.

The joint action set is $A := A_1 \times \cdots \times A_n$ where $A_i = \{a_{i,1}, a_{i,2}, \cdots, a_{i,4+k}\}$ is the discrete finite action space for agent i. $a_{i,1}$: KEEP-LANE-SPEED - the CAV i maintains current speed in the current lane. $a_{i,2}$: CHANGE-LANE-LEFT - the CAV i changes to its left lane. In experiment, by taking $a_{i,2}$ we set a target waypoint trajectory onto its left neighboring lane [21]. $a_{i,3}$: CHANGE-LANE-RIGHT - the CAV i changes to its right lane. In experiment, we set a target waypoint trajectory onto its right neighboring lane. $a_{i,4}$: BRAKE. In the experiment, the CAV i's actuator will compute a brake value within range $brake_{i}^{t} \in [0, 0.5]$ at time $t. a_{i,5}, a_{i,6}, \dots, a_{i,4+k}$ are k discretized throttle intervals. Given the available throttle value set in the simulator as [0,1], we set $a_{i,4+j} = \left[\frac{j-1}{k}, \frac{j}{k}\right]$. By choosing the action $a_{i,5}$, for example, the actuator of the vehicle i will maintain in current lane and compute a throttle value $throttle_i \in [\frac{j-1}{k}, \frac{j}{k}]$ according to the controller design.

Algorithm 1: SR-MAPPO

```
1 Initialize policy, PPO critic and Worst-case Q
        networks oldsymbol{	heta}_i^0, oldsymbol{\phi}_i^0, oldsymbol{\omega}_i^0; worst-Q weight oldsymbol{\lambda}_i^0=0 ;
 2 for each episode \epsilon do
              Initialize s = \prod_i s_i \in \mathcal{S};
 3
              Initialize safe action set A^{\text{safe}} = \prod_i A_i^{\text{safe}} = A;
 4
             egin{aligned} & \textit{Rollout}(s, \mathcal{A}^{	ext{safe}}) \colon \textbf{for } \textit{each } \textit{step, } \textit{agent } i \ \textbf{do} \\ & \mid \text{ Choose } a_i \in \mathcal{A}_i^{	ext{safe}} \text{ based on } \epsilon\text{-greedy,} \end{aligned}
 5
 6
                        a = \prod a_i;
                     Execute action a, observe rewards r = \{r_i\},\
 7
                        and the new state s' = \prod_i s_i';
                      Update A^{\text{safe}}, r_i^s = \mathbf{Safety\_Shield}(s');
 8
                     Store (s_i, a_i, r_i, c_i, s'_i), \forall i \text{ in } M_i;
 9
                     s \leftarrow s';
10
             end
11
             Training: for each agent i do
12
                     Compute PPO critic loss \mathcal{L}_i^V and worst-Q
13
                       critic loss \mathcal{L}_{i}^{Q}, update \phi_{i}^{\epsilon}, \omega_{i}^{\epsilon} to \phi_{i}^{\epsilon+1}, \omega_{i}^{\epsilon+1} [22], [11];
                    Compute objective \mathcal{L}_{i}(\theta_{i}) with (1);
Update \theta_{i}^{\epsilon} to \theta_{i}^{\epsilon+1} and \lambda_{i}^{\epsilon} to \lambda_{i}^{\epsilon+1};
14
15
             end
16
17 end
```

The state transition function is $P: \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0,1]$. The reward functions $\{r_i \coloneqq \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}\}$ are defined as $r_i(s,a) = \sum_j \mu^v_{i,j} \|v_j\|_2 + \sum_j \mu^l_{i,j} \|l_j - d_j\|_2 + \sum_j \mu^s_{i,j} r^s_j(s,a)$, in which v_j is vehicle j's velocity; d_j is j's default destination and r^s_j is the safety reward concerning vehicle's collision and action safety, defined in Sec. IV-C. μ^v, μ^l, μ^s are non-negative weights. In experiment, we set $\mu^v = \mu^l = \mu^s = \frac{1}{|\mathcal{N}|}$, making every agent have an identical reward function and collaborate. Agent i's policy with parameter θ_i is defined as: $\pi^{\theta_i}(a_i|\tilde{s_i})$. As shown in Fig. 2, selected actions are examined by the robust $\mathit{Safety Shield}$ discussed in IV-B in order to guarantee the satisfaction of safety constraints and only safe actions will be implemented by lower level controllers.

IV. METHODOLOGY

In this section, we introduce our major contribution, Safe-Robust Multi-Agent Proximal Policy Optimization (SR-MAPPO), a robust MARL algorithm with *Safety Shield*. We equip each PPO agent with a worst-case Q network [11] to raise the robustness of trained policy, and we also design a robust *Safety Shield* for CAVs based on robust Control Barrier Function [13] implemented by quadratic programming. The *Safety Shield* also provides a feedback term to the reward function IV-C. The proposed algorithm improves the safety and efficiency of CAVs under state uncertainties. We will introduce the **SR-MAPPO** Algorithm 1 in section IV-A, followed by detail of the robust *Safety Shield* and reward design in IV-B and IV-C.

A. Safe-Robust MAPPO

The proposed robust MARL algorithm (Alg 1; Fig. 2) uses centralized-training decentralized-execution design, similar

to the MAPPO [23] algorithm structure. Each robust PPO agent maintains a policy network π^{θ_i} ("actor") with parameter θ_i , a value network ("critic") V(s) with parameter ϕ_i approximating state value function, and the second critic $Q^{\omega_i}(s,a_i)$ network with parameter ω_i approximating the worst-case action values with parameter ω_i . We are inspired by the Worst-case-aware Robust RL framework [11] that was originally designed for single-agent RL like PPO or DQN. During training, both critics account for the policy update so that the trained policy can balance the goals between maximum reward and higher robustness to state perturbations. The robustness is realized through value-based state regularization technique [24], [23] that can distinguish and highlight "crucial state" during training, making them less "vulnerable" to perturbation and achieve relatively higher worst-case value.

The Safety Shield (Alg. 2; Sec. IV-B) interacts with the MARL during Rollout process. As the algorithm starts, the CAV agent i interacts with other vehicles in the environment and observes s_i as the state input. The agent's actor applying stochastic policy outputs the probability distribution $\mathcal{P}(\mathcal{A}_i)$ over action set; its safety shield computes the safe action set using $\mathcal{A}_i^{\text{safe}} \subset \mathcal{A}_i$ with robust CBF-QP [13]. Having all agents selected and actuated their actions, participants in environment step to next time-frame by observing the new state $\{s_i'\}$ and the rewards $\{r_i\}$ which is associated with agents' velocity, accomplishment of preset destination and the safety score (Sec. IV-C).

Assume a set of trajectories $\mathcal{D}=\{\tau_k\}$ were sampled from agent i's memory \mathcal{M}_i after $\mathit{Rollout}$, and the trajectory $\tau_k \coloneqq \{(s^t, a_i^t, r_i^t, s^{t+1}) | t \in [\tau_k]\}$, then agent i's expected rewardsto-go at time t_0 in τ_k is defined as $\hat{R}_i^{t_0} \coloneqq \sum_{t=t_0}^{|\tau_k|} \gamma^{(t-t_0)} r_i^t + \gamma^{(|\tau_k|-t_0)} V^{\phi_i}(s^{|\tau_k|})$ and the instant advantage is defined as $\hat{A}_i^t \coloneqq \hat{R}_i^t - V^{\phi_i}(s^t)$. The robust advantage is defined as $\hat{A}_i^t = \hat{A}_i^t + \kappa_{wst} \underline{Q}^{\omega_i}(s^t, a_i^t)$, in which we append the worst-case action value $\underline{Q}^{\omega_i}(s^t, a_i^t)$ for robustness to state perturbations. We also inherit the "ratio" $\rho_{\theta_i}(a_i^t|s_i^t) \coloneqq \frac{\pi^{\theta_i(a_i^t|s_i^t)}}{\pi^{\theta_i^{old}}(a_i^t|s_i^t)}$ and "clipping" $c(\rho) = clip(\rho, 1-\epsilon, 1+\epsilon)$ from PPO [22]. Abbreviating the average-over- \mathcal{D} operator $\frac{1}{|\mathcal{D}_i||\tau_k|}\sum_{\tau_k\in\mathcal{D}_i}\sum_{t=0}^{|\tau_k|}$ as $\frac{1}{N}\sum_t^N$, the new $\mathit{robust-clipped-surrogate}$ objective \mathcal{L}_i^{RCS} of agent i can be defined as:

$$\mathcal{L}_{i}^{RCS}(\boldsymbol{\theta}_{i}) = \frac{1}{N} \sum_{t}^{N} \min(\rho_{\boldsymbol{\theta}_{i}}(a_{i}^{t}|s_{i}^{t}) \underline{\hat{A}}_{i}^{t}, c(\rho_{\boldsymbol{\theta}_{i}}(a_{i}^{t}|s_{i}^{t})) \underline{\hat{A}}_{i}^{t})$$

Meanwhile, we also apply the state regularization technique [11] to pay extra attention to "vulnerable" states s_i . Considering the value-based state-regularization term

$$\mathcal{L}_{i}^{reg}(\boldsymbol{\theta}_{i}) = \frac{1}{N} \sum_{t}^{N} w(s^{t}) \max_{\tilde{s}_{i}^{t}} KL(\boldsymbol{\pi}^{\boldsymbol{\theta}_{i}}(s_{i}^{t}), \boldsymbol{\pi}^{\boldsymbol{\theta}_{i}}(\tilde{s}_{i}^{t}))$$

in which $w(s^t) = (V^{\phi_i}(s^t) - \min_{a_i} \underline{Q}^{\omega_i}(s^t, a_i))$ is the "importance" of s_i^t . Higher importance indicates a more significant decision at s_i^t for the agent, while a higher " $\max KL$ " manifests the maximum divergence of policy incurred by perturbed state \tilde{s}_i^t . The objective function to

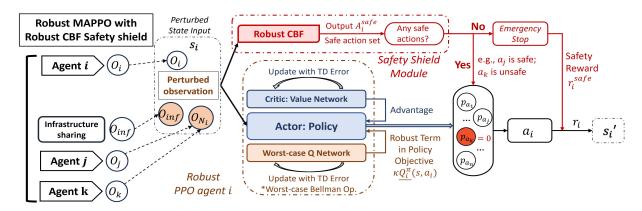


Fig. 2. Robust MAPPO algorithm with Robust Safety Shield. The figure demonstrates an agent's decision pipeline while other agents share the same procedure. Agent *i* takes potentially perturbed state, and inputs it to the actor and safety shield separately. The two modules collectively decide the final executed action. During training, both Value network and worst-Q network join the update of actor's policy.

optimize for policy π^{θ_i} , $i \in \mathcal{N}$ is defined as:

$$\mathcal{L}_i(\boldsymbol{\theta}_i) = \mathcal{L}_i^{RCS}(\boldsymbol{\theta}_i) + \kappa_{reg} \mathcal{L}_i^{reg}(\boldsymbol{\theta}_i)$$
 (1)

The PPO critic V^{ϕ_i} and the worst-case critic \underline{Q}^{ω_i} are updated according to [22] and [11], respectively. We remark that both V^{ϕ_i} and \underline{Q}^{ω_i} are centralized critics and take collective state of all agents as input.

B. Robust CBF-based Safety Shield

CBF-based safety shield has been integrated with MARL algorithm to prevent collisions and violation of safety requirements in CAV systems [8], [25]. However, two major challenges remain unsolved to further improve the safety rate of all agents in reality. First, the safety shield requires accurate states of the agents and lack of robustness to error of state information. Another limitation is that it CBF mainly defines constraints related moving objects on the same road, it is not clear how to define a barrier function for traffic scenarios with a crossing road (intersections) or a merging curb. Therefore, a robust Safety Shield that tolerates state perturbation and handles more common driving scenarios is highly anticipated. Motivated by these challenges, we design a Safety Shield for the proposed SR-MAPPO based on robust CBF [13] for CAV systems, considering more general driving scenarios. The overall design logic is shown in Alg.2. Given the state s_i , Safety Shield loops through all candidate actions $a_{i,\kappa} \in \mathcal{A}_i$ and decides which satisfies safety requirements based on CBFs that implemented by quadratic programming (CBF-QP (4)). The safety evaluation results of actions combined with the eventually executed action at each time step will be a feedback to the MARL algorithm as a "safety score" in return (Sec. IV-C).

1) Robust CBF Formulation: CBF primarily supports normal driving behaviors [21] for CAV agents in both highway and intersection scenarios, and robust CBF under model uncertainties for dynamic systems has been analyzed [13]. We consider a nonlinear system for the continuous dynamics of each CAV: $\dot{x} = f(x) + g(x)u$ with state $x \in \mathbb{R}^n$, and assume each agent has accurate observation of self-state x but imperfect observation of target vehicle's state: $\tilde{x}_T = x_T + e(x_T)$, in which $e(x_T)$ is the bounded perturbation

Algorithm 2: Safety_Shield

```
1 Input: s = \prod s_i; initialize \mathcal{A}^{\text{safe}} = \emptyset;

2 for each agent i do

3 | for each action a_{i,\kappa} \in \mathcal{A}_i do

4 | if a_{i,\kappa} is safe, i.e. CBF-QP has a feasible solution then append a_{i,\kappa} to \mathcal{A}_i^{\text{safe}};

5 | end

6 | if \mathcal{A}_i^{\text{safe}} = \emptyset then \mathcal{A}_i^{\text{safe}} = [Emergency\_stop];

7 end

8 Return: \mathcal{A}_i^{\text{safe}}, r_i^s (Safe-action set, safety reward)
```

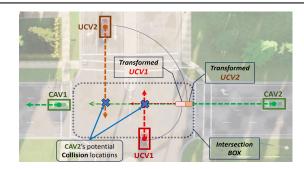


Fig. 3. CBF in crossing scenario. Two UCVs are transformed to CAV2's path as *pseudo* cars. CAV1 is not affected by either UCV.

defined in section III-A,with the 2-norm upperbound $\epsilon = \max_e ||e||_2$. The input is $u \in \mathcal{U} \subset \mathbb{R}^m$, where \mathcal{U} is the admissible input set of the system, f and g are locally Lipschitz. Define a superlevel set $\mathcal{C} \subset \mathbb{R}^n$ of a differentiable function $h: \mathcal{C} = \{x \in \mathbb{R}^n : h(x, \tilde{x}_T, t) \geq 0\}$. A set $\mathcal{C} \subset \mathbb{R}^n$ is forward invariant if for every $x_0 \in \mathcal{C}$, the solution x(t) to the system satisfies $x(t) \in \mathcal{C}$ for all $t \geq 0$. The system is safe with respect to the set \mathcal{C} if the set \mathcal{C} is forward invariant [12]. The function h is a control barrier function (CBF) for the system on \mathcal{C} if there exists $\gamma \in \mathcal{K}_{\infty}$ [26]:

$$\sup_{\boldsymbol{u}\in\mathcal{U}}\left[\frac{\partial h}{\partial t} + L_f h + L_g h \cdot \boldsymbol{u}\right] \ge -\gamma h \tag{2}$$

We adopt the kinematic bicycle model for vehicles [27] and state of the system $x = [x, y, v, \psi]^T$ are the coordinates, velocity, orientation of the vehicle's center of gravity (c.g.). The inputs $u = [\alpha, \varphi]^T$ to the system are

acceleration at the vehicle's c.g. and the steering angle. The adopted functions of safety-following and safety-leading distances are $\mathcal{D}_{SF}(v,v_f) \coloneqq c_1 v + c_2 (\frac{v^2}{2 \mid \max(\alpha) \mid} - \frac{v_f^2}{2 \mid \max(\alpha_f \mid)}) + c_3$ and $\mathcal{D}_{SL}(v,v_r) \coloneqq c_1 v_r + c_2 (\frac{v_r^2}{2 \mid \max(\alpha_f \mid)} - \frac{v_f^2}{2 \mid \max(\alpha_f \mid)}) + c_3$ [8]. Both functions combine the reaction delay $c_1 v$ and the hardbraking distance $c_2 (\frac{v^2}{2 \mid \max(\alpha) \mid} - \frac{v_f^2}{2 \mid \max(\alpha_f \mid)})$ when considering safety requirements for following and leading scenarios. Barrier functions for both scenarios h_f, h_l can thus be given as $h_f(\boldsymbol{x}, \tilde{\boldsymbol{x}}_{\mathcal{T}}, t) \coloneqq (x_f - x) - \mathcal{D}_f(v, v_f)$ and $h_l(\boldsymbol{x}, \tilde{\boldsymbol{x}}_{\mathcal{T}}, t) \coloneqq (x - x_r) - \mathcal{D}_l(v, v_r)$.

Suffering from state perturbations e(x), the inequality 2 cannot guarantee satisfaction. Hence we consider the approach in [13] for Measurement-Robust CBF and append the term $\mathcal{A}(h,\epsilon)$, being the product between the sum of $\frac{\partial h}{\partial t}$ and $(\mu \circ h)$'s Lipschitz and the error bound ϵ .

$$\mathcal{A}(h,\epsilon) = (\mathfrak{L}_{L_t h} + \mathfrak{L}_{L_{u \circ h}})\epsilon \tag{3}$$

It serves as a conservative "buffer" related to perturbation bound ϵ . Let $\mathcal H$ be the set of barrier functions as constraints. The CBF-QP and its surrogate constraints are given as (4). $\mathcal H$ is the set of all barrier functions as constraints and h_j being h_f and h_l for a following and leading vehicle, respectively. (4) was proven to infer (2) if the input u is accurate.

CBF-QP:
$$\min_{\boldsymbol{u} \in \mathbb{R}^m} \frac{1}{2} \parallel \boldsymbol{u} - \boldsymbol{u}_i \parallel^2 \quad \text{s.t. } \forall h_j \in \mathcal{H}$$
$$\frac{\partial h_j}{\partial t} + L_f h_j + L_g h_j \cdot \boldsymbol{u} - \mathcal{A}(h_j, \epsilon) \ge -\gamma h_j \qquad (4)$$

2) General Driving Scenario: We further adapt the CBF to general driving scenarios such as crossing an intersection or merging from curb, by transforming a target vehicle to a *pseudo* following or leading vehicle. For example, in Fig. 3, two unconnected vehicles (UCVs) are driving across the intersection from the perspective of CAV2. Assume the two CAVs' driving direction has green light, the agent CAV2 targets two UCVs as potential threat of collision (e.g., violating red light). The CBF of CAV2 will first transform both UCVs locations and velocities onto CAV2's path and generate two pseudo cars in front of CAV2. The crossing scenario is then reduced to one direction road scenario, and the transformed locations and velocities of the pseudo cars are considered as constraints for the leading vehicles of CAV2 in the CBF (4). The CBF should be capable of preventing collisions within the "intersection box" as long as the target crossing vehicle drives on the road section that intersects with the CAV's path.

C. Reward Design and Safety Feedback to R-MARL

As mentioned in III-B, reward functions $\{r_i\}$ are defined as $r_i(s,a) = \sum_j \mu_{i,j}^v \|v_j\|_2 + \sum_j \mu_{i,j}^l \|l_j - d_j\|_2 + \sum_j \mu_{i,j}^s r_j^s(s,a)$. μ^v, μ^l, μ^s are coefficients balancing self-return and group-return. In experiment, we take the same values for $\mu_{i,j}$ for all agents, so that every agent coordinates by having identical reward. The safety reward $r_j^s(s,a) = \sum_{i,t} P^{Col}(i,t) + P^{SAS}(\mathcal{A}, \mathcal{A}^{\text{safe}})$ consists of the accumulated collision penalty $\sum_{i,t} P^{Col}(i,t)$ and the safe-action score $P^{SAS}(\mathcal{A}, \mathcal{A}^{\text{safe}})$, which takes discontinuous manner

and penalizes the agent when no action satisfies *CBF-QP* and *Emergency_stop* has to be executed.

V. EXPERIMENTS AND EVALUATIONS

We deploy our experiment in the CARLA Simulator environment [28], where each vehicle is configured with inborn GPS and IMU sensors and a collision sensor that detects the collision with other objects. We set the communication range of all CAVs in simulation as 200m. The k-discretized throttle ranges in action space A_i is set as k=3 and the discount factor as $\gamma = 0.99$. Every model was trained 200 episodes and tested 50 episodes in each scenarios, while each episode lasts 200 steps. Agents are trained in a relatively 'safe' setup without state perturbation, while in testing we include more challenging edge cases for UCVs behaviors and agents also suffer from perturbed state inputs. Scenario-related and perturbation details are introduced in Sec. V-A. The training and testing of our algorithm and baselines took place in a server configured with AMD Ryzen 3970X processor and NVIDIA Quadro RTX 6000 GPU. The experiments are performed with CARLA 0.9.14, Python 3.8, PyTorch 1.10.

A. Simulation with Perturbations and Challenging Scenarios

We consider two challenging scenarios in daily driving, respectively at *Intersection* (Fig. 4a-4c) and on *Highway* (Fig. 4d), where 3 CAVs and some UCVs are spawned. We adopt three types of perturbations exclusively for testing. One is the random error $e^{\text{rand}} \sim U(-2,2)$ (U: uniform distribution); and two kinds of perturbation strategies, $perturb_over_time$: PTB^T and $perturb_target_vehicles$: PTB^V are given as:

$$\begin{split} & \text{PTB}^T = \{(e^t, \frac{e^t}{2}) | e^t \sim \textit{U}(e^0 - \frac{1}{2}, e^0 + \frac{1}{2}), \pm e^0 \sim \textit{U}(-9,11), t \in T\} \\ & \text{PTB}^{\mathcal{V}} = \{(e^{\nu}, \frac{e^{\nu}}{2}) | e^{\nu} \sim \textit{U}(e^0 - \frac{1}{2}, e^0 + \frac{1}{2}), \pm e^0 \sim \textit{U}(-9,11), \nu \in \mathcal{V}\} \end{split}$$

The two strategies aim to exert consistent perturbation values to CAVs' states, so as to affect their behavior patterns. The former PTB^T targets all cars in time duration T while the latter $PTB^{\mathcal{V}}$ have the target subset of vehicles \mathcal{V} wrongly observed by others throughout simulation.

- 1) Highway: Figures. 4d illustrates the scenario, where three CAVs (green) are spawned behind three UCVs (red) on a multi-lane highway. During training, UCVs keep in their lanes at random speed from [8-10] m/s; in testing, the middle UCV rehearses a real-life broken vehicle by suddenly braking to a random speed in [3-4] m/s. CAVs aim to avoid any collision while reaching the preset destination at further.
- 2) Intersection: The Fig. 1 and Fig. 4 presents intersection scenarios with three CAVs (green) passing the intersection and two UCVs (red) from both sides crossing the box recklessly in the meantime. The velocities of UCVs passing the intersection are random-uniformly sampled for each episode from [9,11] m/s in training and from [7.5,12.5] m/s in testing. CAVs aim to avoid collision and reach the preset destination after intersection. Intersection is a much more challenging scenario for collision avoidance as the two crossing UCVs synchronously pose collision threats apart from intra-CAV safety requirements.

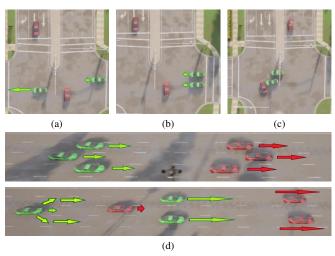


Fig. 4. Intersection (a,b,c) and Highway (d) scenarios for testing. Different collision-free cases in (4a) and (4b) and a collision in (4c). (4d): Highway scenario initialization and in progress when one UCV suddenly brakes.

TABLE I
TRAINING RESULTS IN TWO SCENARIO

Scenario	Baselines		Ours
	MAPPO-SS ¹	MAPPO-rSS ²	SR-MAPPO ³
Intersection	79.5%; -356.2	79.5%; -104.2	82.5%; -3.5
Highway	93%; -370.0	91%; -477.7	96.5% ; -450.0

¹MAPPO-SS: MAPPO with (non-robust) Safety Shield; ²MAPPO-rSS: MAPPO with robust Safety Shield; ³SR-MAPPO: Robust MAPPO (worst-case aware) with robust Safety Shield

Each entry above is (collision-free rate; mean episode return).

TABLE II
TESTING RESULTS IN SIX SCENARIO-PERTURBATIONS.

Testing Scenario	Baselines		Ours
	MAPPO-SS	MAPPO-rSS	SR-MAPPO
Intersection-e ^{rand}	88% ; -358.5	68%; -162.2	84%; -232.8
$\textit{Intersection-} \mathtt{PTB}^T$	68%; -430.4	64%; -275.1	86%; -273.9
$\textit{Intersection-PTB}^{\mathcal{V}}$	64%; -389.9	74%; -166.5	88%; -98 . 5
Highway-e ^{rand}	100%; -380.6	96%; -445.6	100%; -369.5
$\mathit{Highway} ext{-}\mathtt{PTB}^T$	88%; -402.7	100%; -378.5	100%; -346.7
$\mathit{Highway} ext{-PTB}^\mathcal{V}$	72%; -465.0	100%; -380.4	100%; -356.2

Each entry above is (collision-free rate; mean episode return). Our method **outperforms** baselines significantly in both scenarios and 5/6 cases, while maintaining remarkable collision-free rate.

B. Experiment Results

We trained our model ('SR-MAPPO' in the table I, II), a baseline using MAPPO with robust $Safety\ Shield$ ('MAPPO-rSS' in tables) and another baseline 'MAPPO-SS' using MAPPO with non-robust $Safety\ Shield$ respectively on Intersection and Highway scenarios. Training and testing experiment results are presented in table I and II. We **highlight** our method's leading performances. In all testing with perturbations except $Intersection\text{-}e^{\text{rand}}$, our method achieves the highest collision-free rates and mean episode returns, which is defined as the mean of agents' sums over stepwise rewards: $\frac{1}{m}\sum_{\epsilon=1}^{m} mean_i(\sum_t r_i^t).$ Examples of episode return values from our model in $Intersection\text{-PTB}^{\mathcal{V}}$ are given in Fig. 5.

1) Robustness from Worst-case Q in MARL: In almost all testing scenarios with state perturbation, our approach outperforms baselines by its highest overall returns and

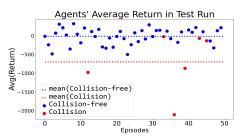


Fig. 5. Scatter-plot of returns in testing SR-MAPPO on *Intersect*-PTB $^{\mathcal{V}}$ collision-free rate, and presents remarkable robustness to random error or perturbation strategies PTB T and PTB $^{\mathcal{V}}$. Compared with the baseline 'MAPPO-rSS' regarding the safety performances in *Intersection*, we see that robust MARL algorithm seems to have good "chemistry" with robust *Safety Shield* such that the safety rates with three tested perturbations did not drop as significant as 'MAPPO-rSS' does, but even increased a bit by sacrificing a little driving efficiency. 'MAPPO-rSS' on the contrary becomes relatively more 'aggressive' for returns in testing compared with training, yet its collision-free rates decreased by approximately 10% and the agents suffer from more safety threats.

2) Benefits of Robust Safety Shield: The robust Safety Shield together with robust MARL shows significant robustness to perturbations in comparison with its nonrobust counterpart. In *Intersection*, this combination achieves overall higher collision-free rates during testings. The only exception is MAPPO-SS with random noise, yet the baseline sacrifices greatly with its apparently lower returns for being conservative. In Highway's testings, we see our method and the MAPPO-rSS achieve overwhelmingly better safety compared to MAPPO with non-robust Safety Shield. As a remark, we also found by ablation study that our method seems to be relatively more robust to the strategies PTB^T and $PTB^{\mathcal{V}}$. As is mentioned, MAPPO with non-robust Safety Shield is trained to be conservative in Intersection. However, such conservativeness fails with PTB^T and PTB^V, which generates stronger, more targeting and consistent perturbations compared to random noises, and deceptive enough to fail the non-robust Safety Shield.

VI. CONCLUSION

In this work, we study the safe and robust decision-making problem for connected autonomous vehicles in common driving scenarios under state perturbations. We propose the Safe-Robust MAPPO algorithm and the robust Safety Shield for CAVs based on information-sharing and coordination. The robust MAPPO algorithm with worst-case consideration is verified to significantly raise agents' performance under perturbed state input. The Safety Shield with robust Control Barrier Functions tolerating bounded perturbations also provides safety rewards as feedback to MARL and help shape the policy. In experiments, we verify the effectiveness of the Safety Shield and the advantage of robust MARL algorithm in challenging testing scenarios. Future work could extend to consider safe-robust MARL frameworks tackling model uncertainties, perturbed actions and dynamic model uncertainties underlying the vehicles.

REFERENCES

- [1] D. Martín-Sacristán, S. Roger, D. Garcia-Roger, J. F. Monserrat, P. Spapis, C. Zhou, and A. Kaloxylos, "Low-latency infrastructurebased cellular v2v communications for multi-operator environments with regional split," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 2, pp. 1052–1067, 2020.
- [2] H. Mun, M. Seo, and D. H. Lee, "Secure privacy-preserving v2v communication in 5g-v2x supporting network slicing," *IEEE Trans. Intell. Transp. Syst.*, 2021.
- [3] N. Buckman, A. Pierson, S. Karaman, and D. Rus, "Generating visibility-aware trajectories for cooperative and proactive motion planning," in *ICRA*. IEEE, 2020, pp. 3220–3226.
- [4] A. Miller and K. Rim, "Cooperative perception and localization for cooperative driving," in *ICRA* 2020. IEEE, 2020, pp. 1256–1262.
- [5] S. Han, H. Wang, S. Su, Y. Shi, and F. Miao, "Stable and efficient shapley value-based reward reallocation for multi-agent reinforcement learning of autonomous vehicles," in 2022 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2022, pp. 8765– 8771.
- [6] J. Rios-Torres and A. A. Malikopoulos, "A survey on the coordination of connected and automated vehicles at intersections and merging at highway on-ramps," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 5, pp. 1066–1077, May 2017.
- [7] J. Lee and B. Park, "Development and evaluation of a cooperative vehicle intersection control algorithm under the connected vehicles environment," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 1, pp. 81–90, March 2012.
- [8] Z. Zhang, S. Han, J. Wang, and F. Miao, "Spatial-temporal-aware safe multi-agent reinforcement learning of connected autonomous vehicles in challenging scenarios," in 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023, pp. 5574–5580.
- [9] Y. Zhu, C. Miao, F. Hajiaghajani, M. Huai, L. Su, and C. Qiao, "Adversarial attacks against lidar semantic segmentation in autonomous driving," in *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, 2021, pp. 329–342.
- [10] Y. Zhu, C. Miao, T. Zheng, F. Hajiaghajani, L. Su, and C. Qiao, "Can we use arbitrary objects to attack lidar perception in autonomous driving?" in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 1945–1960.
- [11] Y. Liang, Y. Sun, R. Zheng, and F. Huang, "Efficient adversarial training without attacking: Worst-case-aware robust reinforcement learning," Advances in Neural Information Processing Systems, vol. 35, pp. 22547–22561, 2022.
- [12] A. D. Ames, X. Xu, J. W. Grizzle, and P. Tabuada, "Control barrier function based quadratic programs for safety critical systems," *IEEE Transactions on Automatic Control*, vol. 62, no. 8, pp. 3861–3876, 2017.
- [13] R. K. Cosner, A. W. Singletary, A. J. Taylor, T. G. Molnar, K. L. Bouman, and A. D. Ames, "Measurement-robust control barrier functions: Certainty in safety with uncertainty in state," in 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2021, pp. 6286–6291.
- [14] A. Amini, I. Gilitschenski, J. Phillips, J. Moseyko, R. Banerjee, S. Karaman, and D. Rus, "Learning robust control policies for end-toend autonomous driving from data-driven simulation," *IEEE Robotics* and Automation Letters, vol. 5, no. 2, pp. 1143–1150, 2020.
- [15] L. Brunke, M. Greeff, A. W. Hall, Z. Yuan, S. Zhou, J. Panerati, and A. P. Schoellig, "Safe learning in robotics: From learning-based control to safe reinforcement learning," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, pp. 411–444, 2022.
- [16] L. Wen, J. Duan, S. E. Li, S. Xu, and H. Peng, "Safe reinforcement learning for autonomous vehicles through parallel constrained policy optimization," in 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2020, pp. 1–7.
- [17] S. Lu, K. Zhang, T. Chen, T. Başar, and L. Horesh, "Decentralized policy gradient descent ascent for safe multi-agent reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, 2021, pp. 8767–8775.
- [18] J. Wang, S. Yang, Z. An, S. Han, Z. Zhang, R. Mangharam, M. Ma, and F. Miao, "Multi-agent reinforcement learning guided by signal temporal logic specifications," arXiv preprint arXiv:2306.06808, 2023.
- [19] S. Han, S. Su, S. He, S. Han, H. Yang, and F. Miao, "What is the solution for state adversarial multi-agent reinforcement learning?" arXiv preprint arXiv:2212.02705, 2022.

- [20] S. He, S. Han, S. Su, S. Han, S. Zou, and F. Miao, "Robust multiagent reinforcement learning with state uncertainty," *Transactions on Machine Learning Research*, 2023.
- [21] S. He, J. Zeng, B. Zhang, and K. Sreenath, "Rule-based safety-critical control design using control barrier functions with application to autonomous lane change," in 2021 American Control Conference (ACC). IEEE, 2021, pp. 178–185.
- [22] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv* preprint *arXiv*:1707.06347, 2017.
- [23] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of ppo in cooperative multi-agent games," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24611–24624, 2022.
- [24] H. Zhang, H. Chen, C. Xiao, B. Li, M. Liu, D. Boning, and C.-J. Hsieh, "Robust deep reinforcement learning against adversarial perturbations on state observations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21024–21037, 2020.
- [25] R. Cheng, G. Orosz, R. M. Murray, and J. W. Burdick, "End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 3387–3395.
- [26] G. Wu and K. Sreenath, "Safety-critical control of a 3d quadrotor with range-limited sensing," in *Dynamic Systems and Control Conference*, vol. 50695. American Society of Mechanical Engineers, 2016, p. V001T05A006.
- [27] J. Kong, M. Pfeiffer, G. Schildbach, and F. Borrelli, "Autonomous driving using model predictive control and a kinematic bicycle vehicle model," in *Intelligent Vehicles Symposium*, Seoul, Korea, 2015.
- [28] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, 2017, pp. 1–16.