Learning a Robust Multiagent Driving Policy for Traffic Congestion Reduction

Yulin Zhang
The University of Texas at Austin
United States
yulin@cs.utexas.edu

William Macke
The University of Texas at Austin
United States
wmacke@cs.utexas.edu

Jiaxun Cui
The University of Texas at Austin
United States
cuijiaxun@utexas.edu

Daniel Urieli
General Motors R&D Labs
Israel
daniel.urieli@gm.com

Peter Stone
The University of Texas at Austin and
Sony AI
United States
pstone@cs.utexas.edu

ABSTRACT

In most modern cities, traffic congestion is one of the most salient societal challenges. Past research has shown that inserting a limited number of autonomous vehicles (AVs) within the traffic flow, with driving policies learned specifically for the purpose of reducing congestion, can significantly improve traffic conditions. However, to date these AV policies have generally been evaluated under the same limited conditions under which they were trained. On the other hand, to be considered for practical deployment, they must be robust to a wide variety of traffic conditions. This paper establishes for the first time that a multiagent driving policy can be trained in such a way that it generalizes to different traffic flows, AV penetration, and road geometries, including on multi-lane roads.

1 INTRODUCTION

According to Texas A&M's 2021 Urban Mobility Report, traffic congestion in 2020 in the U.S. was responsible for excess fuel consumption of about 1.7 billion gallons, an annual delay of 4.3 billion hours, and a total cost of $100B [9]. A common form of traffic congestion on highways is *stop-and-go waves*, which have been shown in field experiments to emerge when vehicle density exceeds a critical value [15]. Past research has shown that in human-driven traffic, a small fraction of automated or autonomous vehicles (AVs) executing a controlled multiagent driving policy can mitigate stop-and-go waves in simulated and real-world scenarios, roughly double the traffic speed, and increase throughput by about 16% [14]. Frequently, the highest-performing policies are those learned by deep reinforcement learning (DRL) algorithms, rather than hand-coded or model-based driving policies.

Any congestion reduction policy executed in the real world will need to perform robustly under a wide variety of traffic conditions such as traffic flow, AV penetration (percentage of AVs in traffic, referred to here as "AVP"), AV placement in traffic, and road geometry. However, existing driving policies have generally been tested in the same conditions they were trained on, and have not been thoroughly tested for robustness to different traffic conditions. Indeed, their performance can degrade considerably when evaluated outside of the training conditions (Figure 1). Therefore, it remains unclear how to create a robust DRL congestion-reduction driving policy that is practical for real-world deployment.

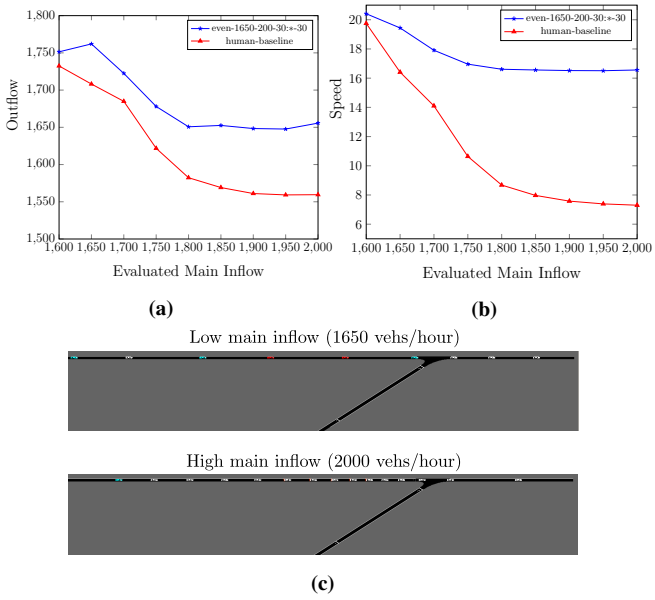

Figure 1: Increasing incoming vehicle flow (the demanded *inflow*) degrades performance of a policy trained with inflow of 1650 veh/hour**, with respect to both throughput (a) and speed (b). A visual representation (c) is given that shows what this decreased efficiency looks like.**

In this paper, we establish for the first time the existence of a robust DRL congestion-reduction driving policy that performs well across a wide variety of traffic flows, AVP, AV placement in traffic, and several road geometries. Moreover, we investigate the question of how to come up with such a policy and what degree of robustness it can achieve. We create a benchmark with a diverse, pre-defined collection of test traffic conditions of real-world interest including the single-lane merge scenario shown in Figure 1c. Such merge scenarios are a common source of stop-and-go waves on highways [10]. While there are different approaches to training robust DRL policies in other domains with different levels of success, our approach is

Proc. of the Adaptive and Learning Agents Workshop (ALA 2022), Cruz, Hayes, Silva, Santos (eds.), May 9-10, 2022, Online, https://ala2022.github.io/. 2022.

to systematically search for a robust policy by varying the training conditions, evaluating the learned policy on our proposed test set in a single-lane merge scenario, and selecting the highest performing one. The highest performing policy outperforms the human-only baseline with as few as 1 % AVs across different traffic conditions in the single-lane merge scenario. We further investigate the policy's generalization to more complex roads it has not seen during training, specifically with two merging ramps at a variety of distances, or on a double-lane main road, with cars able to change lanes. Notwithstanding negative prior results showing that a policy developed in a single-lane ring road fails to mitigate the congestion on a double-lane ring road [3], the learned policy outperforms human-only traffic and effectively mitigates congestion in all of these scenarios defined by our benchmark. Taken together, this paper's contributions and insights take us a step closer towards making the exciting concept of traffic congestion reduction through AV control a practical reality.

2 RELATED WORK

Traffic optimization has long been a challenging research area with direct real-world impact [4]. An important research question is how to mitigate highway *stop-and-go waves*, which have been demonstrated to emerge when vehicle density exceeds a critical value, and to result in reduced throughput and increased driving time [15]. In small-scale field experiments, vehicles controlled autonomously by hand-designed driving policies successfully dissipated stop-and-go waves, thus reducing congestion [14]. The industry-wide development of autonomous vehicles (AVs) has inspired researchers to tackle this problem at a larger scale.

Recent progress in Reinforcement Learning (RL) [16] has made it possible to learn congestion reduction AV driving policies that perform well in simulation. Using state-of-the-art algorithms, significant congestion reduction was achieved both in circular roads with a fixed set of vehicles (referred to as *closed* road networks), and acyclic roads with vehicles entering and leaving the system (referred to as *open* road networks) [8, 19, 22], as compared with simulated human-driven traffic implemented with accepted human driving models [18]. Most of these past successful driving policies controlled AVs in a *centralized* manner, where a single controller simultaneously processes all available sensing information and sends driving commands to the AVs. More recent efforts focused on developing *decentralized* driving policies which might be harder to learn, but are considered a more realistic option for real-world deployment, as they mostly rely on local sensing and actuation capabilities [2, 19]. This paper continues the line of research on decentralized policies but aims to develop one that is robust to real-world traffic conditions of practical interest.

Recent RL techniques for developing robust policies include adversarial training [12] and domain randomization [17]. Existing research uses these ideas to build congestion reduction policies that are robust to some particular traffic conditions. Wu et al. present policies that can generalize on a closed ring road to traffic densities higher and lower than the ones they were trained on, by randomizing densities during training [23]. Parvate et al. evaluate the robustness of a *hand-coded* controller over different AV penetration and driving aggressiveness [11]. This paper focuses on learning a driving policy that is robust to different traffic flows, AV penetrations, AV placement within traffic, and road geometries.

In a parallel unpublished work [20], Vinitsky et al. studied a similar setup. In particular, similarly to our work, they developed a robust, decentralized policy that is shared among all AVs for an open road network scenario. On the other hand, our work differs from theirs in several ways. First we focus on merge scenarios, while they focus on bottleneck scenarios. Second, they developed a robust policy by randomizing the training conditions, while we did a systematic sweep of the training conditions to understand how each training condition contributes to the performance of the trained policy. Third, we further examined the robustness of the policy trained from a merge scenario on a more complex road with multiple merging ramps and multiple lanes.

3 BACKGROUND AND SETUP

We start by introducing the background and the problem of learning a robust traffic congestion reduction policy.

3.1 Road-merge congestion reduction

Consider a network with a main highway and a merging road, as shown in Figure 1c. There are vehicles joining and leaving the network, and the traffic consists of both human-driven and autonomous vehicles. The human drivers are assumed to be self-interested and optimize their own travel time, while autonomous vehicles (AVs) are assumed to be altruistic and have a common goal of reducing traffic congestion. Our goal is to come up with a driving policy that controls each AV such that traffic performance is improved.

We measure the performance of policies in terms of both *outflow* and *average speed*. Outflow is the number of vehicles per hour exiting the simulation, representing system-level throughput. The average speed represents the time delay it takes an average driver to drive the simulated road. We note that it is important to report both metrics, since scenarios with low and high average speeds could have the same system throughput, such that one is considered congested while the other is not.

A policy can be hand-programmed or learned. Reinforcement learning (RL) has been shown to produce superior policies [8, 19, 22] and is therefore our method of choice. Congestion reduction driving policies can either be *centralized*, controlling all vehicles simultaneously based on global system information, or *decentralized*, controlling each vehicle independently based on its local observations. Decentralized policies with no vehicle-to-vehicle communication are most realistic, since they mostly rely on local sensing and actuation capabilities [2, 20], and are therefore the focus of this paper.

This multiagent traffic congestion reduction problem can be modelled as a discrete-time, finite-horizon decentralized partially observable Markov decision process (Dec-POMDP) [1], denoted as a tuple $(\mathcal{S}, \{\mathcal{A}_i\}, P, R, \{\Omega_i\}, \mathcal{O}, T, \gamma)$ where,

- $\mathcal{S}$ is a state space representing the location and speed of every vehicle in the network,
- $\{\mathcal{A}_i\}$ is a joint action space for all agents, where $\mathcal{A}_i$ specifies an acceleration action for agent i,
- $P : \mathcal{S} \times \{\mathcal{A}_i\} \times \mathcal{S} \to [0, 1]$ is a state transition probability distribution, which is realized via a traffic simulator,
- $R : \mathcal{S} \times \{\mathcal{A}_i\} \to \mathbb{R}$ is a global reward function,

- $\{\Omega_i\}$ is a collection of local observations for each agent (see Section 3.2),
- $O : \mathcal{S} \times \{\mathcal{A}_i\} \times \{\Omega_i\} \rightarrow [0, 1]$ outputs the probability that each agent receives a specific observation given the next state and the joint action just taken,
- T is the episode length,
- $\gamma \in [0, 1]$ is the discount factor of reward.

A decentralized, shared *driving policy* is a probability density function over the action space $\pi_\theta : \Omega \times \mathcal{A} \rightarrow [0, 1]$ parameterized by θ that stochastically maps each agent's local observations to its driving actions.

Throughout this paper we use the SUMO traffic simulator [6] as the state transition function. SUMO is a micro simulator that includes accepted human driving models [7, 18], configurable traffic networks and flows, and mechanisms for enforcing traffic rules, safety rules, and basic physical constraints. To learn AV driving policies, we use the RLlib library [5]. We interface with SUMO and RLlib using UC Berkeley's Flow software [21].

3.2 RL-based decentralized driving policy

To learn a decentralized driving policy we use the Proximal Policy Optimization (PPO) algorithm [13]. To facilitate data and computational efficiency and reduce the risk of overfitting, all AVs learn and execute a single, shared driving policy. The observation space and reward design used in this paper are modeled after those used by Cui et al. [2], which were shown to be effective for decentralized policies. The observation for each AV includes

- the speed and distance of the closest vehicles in front of and behind it,
- the AV's speed,
- the AV's distance to the next merging point,
- the speed of the next merging vehicle and its distance to the merge junction (assumed to be obtained by the vehicle's cameras/radars, or be computed by some global infrastructure and then shared with all the vehicles).

The reward of the ith AV at time step t is defined as:

$$r_{i,t} = (1 - \mathbb{I}\{done\})\left(-\eta + (1-\eta) \times \frac{\sum_{j=1}^{n_t} v_j}{n_t V_{max}} \right) + \mathbb{I}\{done\} \cdot Bonus$$

where $\mathbb{I}\{done\}$ is an indicator function of whether an AV is leaving the network; *Bonus* is a constant reward for an AV if it exits the network; the term $\frac{\sum_{j=1}^{n_t} v_j}{n_t V_{max}}$ represents the normalized average speed, where v_j is the speed of vehicle j, n_t is the total number of vehicles in the network at time t, $V_{\max}$ is the max possible speed, and η is a constant that weights the individual and the global reward.

3.3 Robustness evaluation conditions and metrics

Similarly to past work, our baseline setup consists of simulated human-driven vehicles only. In contrast to past work, which typically showed improvement over this baseline in a *single* combination of traffic conditions, our goal is to develop a robust AV driving policy that improves over this baseline across a *range* of realistic traffic conditions, characterized by:

- *Main Inflow Rate*: the amount of incoming traffic on the main artery (veh/hour),
- *Merge Inflow Rate*: the amount of incoming traffic on the merge road (veh/hour),
- *AV Placement*: the place where the AVs appear in the traffic flow; the AVs can either be distributed evenly or randomly among the simulated human-driven vehicles.
- *AV Penetration*: the percentage of vehicles that are controlled autonomously,
- *Merge road geometry:* the distance between two merge junctions (in relevant scenarios), and the number of lanes.

In this paper, we fix the merge inflow rate to be 200 veh/hour (small enough to cause traffic congestion on the main road) and set the range of the main inflow to be [1600, 2000] veh/hour (resulting in minimal to maximal congestion in our simulations), AV penetration (AVP) to be within [0, 40] percent (for a realistic amount of controllable AVs in the coming years). The placement of the AVs can either be random or even. For *even placement*, AV are placed every N human-driven vehicles in a lane. For *random placement*, AVs are placed randomly among simulated human-driven vehicles. Merge road geometries include one or two merges at distances that vary between [200, 800] meters, and the main road can have one or two lanes.

4 LEARNING A ROBUST POLICY IN THE SINGLE-LANE MERGE SCENARIO

While real-world congestion-reducing driving policies need to operate effectively in a wide variety of traffic conditions, most past research has tested learned policies under the same conditions on which they were trained. Since in the real world it is impractical to deploy a separate policy for each combination of conditions, our primary goal is to understand whether it is feasible to learn a *single* driving policy that is robust to real-world variations in traffic conditions.

The performance of an RL-based driving policy depends on the traffic conditions under which it is trained. We hypothesize that the policy trained under high inflow, medium AV penetration, and random vehicle placement is robust in a range of traffic conditions defined in Section 3.3 for a single-lane merge scenario. We test this hypothesis by comparing 30 policies, each of which is trained under a combination of traffic conditions specified below in Section 4.1. The training of each policy takes about 7 hours on a 3.7 GHz Intel 12 Core i7 processor. Each policy, including human-only baseline, is evaluated 100 times using the same 100 random seeds, and each evaluation takes about 1 hour. After identifying a policy that generalizes well across training conditions, we then evaluate it on road geometries different from its training scenario.

4.1 Discretization of traffic conditions for training

Since there is an innumerable set of possible traffic conditions, for the purpose of training we discretize traffic conditions along their defining dimensions to a total of 30 representative combinations of conditions, as follows. We consider main inflows of 1650, 1850, and 2000 veh/hour which result in low, medium, and high congestion. We discretize AV placement in traffic to be random or even-spaced. Finally, we discretize the training AV penetration into 5 levels: 10 %, 30 %, 50 %, 80 %, 100 %. Based on this $3 \times 2 \times 5$ discretization, we train 30 policies, one for each combination.

Each trained policy is then evaluated across the range of traffic conditions described in Section 3.3, leading to two performance values (outflow and average speed) on each testing condition for each policy. We plot these results using the following convention. The label of a data point consists of two parts: (i) the training conditions of the policy to be evaluated, and (ii) the policy's evaluation conditions. The policy's training conditions indicate the vehicle placement, main inflow, merge inflow, and AVP, separated by "-". For example, "random-2000-200-30" denotes the policy trained under random vehicle placement with main inflow 2000 veh/hour, merging inflow 200 veh/hour, and 30 % AVP. The evaluation conditions also consist of vehicle placement, main inflow, merging inflow, and AVP. In this paper, the merging inflow is always fixed to be 200 veh/hour and the vehicle placement is specified separately from the graph label. Therefore we only specify the evaluation-time main inflow and AVP to indicate the evaluation condition for each data point. Hence, each evaluation result is labeled as a 6-tuple, where the first four elements describe the training conditions and the remaining two describe the evaluation conditions. For example, "random-2000-200-30:1800-10" labels the result of policy "random-2000-200-30" evaluated under main inflow 1800 veh/hour and AVP 10 %. We further use "$*$" in the evaluation condition to denote which evaluation condition varies in a plot. For example, "random-2000-200-30:1800-$*$" indicates that the policy "random-2000-200-30" was evaluated under main inflow of 1800 and varying AVPs; "random-2000-200-30:$*$-10" indicates that policy "random-2000-200-30" was evaluated under AVP 10 % and varying main inflows.

4.2 Robustness to vehicle placement, AV penetration and inflow

In this section, we will test our hypothesis that training with high inflow, medium AV penetration, and random vehicle placement yields a robust policy, by showing representative slices of the evaluation results.

We start by showing that the policies trained under random vehicle placement outperform others that are trained under even vehicle placement. The performance of a subset of these policies is depicted in Figure 2a and 2b. The red curves represent the evaluation results for the policies trained under random vehicle placement, and the blue curves represent the results for the policies trained under even vehicle placement. These policies are evaluated using the outflow and average speed metrics under both random vehicle placement (Figure 2a) and even vehicle placement (Figure 2b). When evaluating on either random placement or even placement, the policies trained with random placement outperform the human baseline as well as their counterparts trained with even placement. Specifically, the results in Figure 2a confirm the intuition that when evaluated with random vehicle placement, the policies trained under random vehicle placement should have better performance than their counterparts trained with even vehicle placement. However, counter-intuitively, random placement at training time also results in more robust policies when testing under *even* placement. We hypothesize that this performance increase is due to the more diverse data collected when RL vehicles are randomly placed.

Next, we confirm the intuition that the polices trained under medium AV penetration are better than others. Figure 2c show when fixing the main inflow, the policies trained under AVP 30 % (red curve) are competitive in both their outflow and average speed when evaluated under varying AVPs. They have the best performance across a large range of the evaluation AVPs. The same conclusion also holds if we fix the AVP but instead vary the main inflow during evaluation. We hypothesize that these mid-range AVP values during training perform best since (i) the policies are well-trained with sufficient AVs collecting training data; (ii) there are a certain amount of human-driven vehicles and the learned policies are useful to reduce traffic congestion created by these human-driven vehicles.

Finally, we test the hypothesis that the policies trained under high inflow are robust. When fixing the AVP and varying main inflow during evaluation, Figure 2d shows that our proposed policy trained under main inflow 2000 veh/hour (green curve) has better performance in both outflow and average speed than other policies trained with different main inflows. The same conclusion also holds if we fix the main inflow but vary the AVPs during evaluation. We hypothesize that the policies trained under the highest inflow outperform others because a higher main inflow yields more diverse vehicle densities at training time. Specifically, the simulation dynamics can lead high inflow to include both dense vehicle placements and sparse vehicle placements, while a lower main inflow tends to mostly result in sparse vehicle distribution.

Verifying our hypothesis, we find that the policy "random-2000-200-30", which is trained under random vehicle placement, main inflow 2000 veh/hour, merge inflow 200 veh/hour, and AVP 30 %, outperforms the alternatives. In the single-lane merge scenario, this policy achieves significant improvement over the human-only baseline when the AVP is greater than or equal to 1 % during deployment (with p-value 0.05 as the cutoff for significance).

5 DEPLOYING THE LEARNED POLICY TO MORE COMPLEX ROADS

We learned a robust policy in a single-lane merge scenario. To push this policy one step further toward a real-world deployment, we test this policy's robustness to more complex road structures: roads with two merging roads, and roads with two lanes.

5.1 Deployed to roads with two merging ramps

We first deploy the selected policy on more complex road structures, which have two merging roads at varying distances, and evaluate the performance of the learned policy with respect to the distance between these two ramps. An example road with two merging on-ramps is shown in Figure 3.

Consider the merge scenario with two merging ramps: the first merging ramp is located 500 meters from the simulated main road's start, the second merging ramp is located 200, 400, 600, or 800 meters after the first, the total length of the main road is 1500 meters, and the total length of the merging roads is 250 meters. We tested the random-2000-200-30 policy with random AV placement, main inflow of 1800 veh/hour, merge inflow 200 veh/hour, across a range of AV penetrations and the above gaps between the two merging roads.

The results are shown in Figure 4, where the blue curves show the performance of the policy to be tested, and the red curve shows the human baseline's performance. The random-2000-200-30 policy

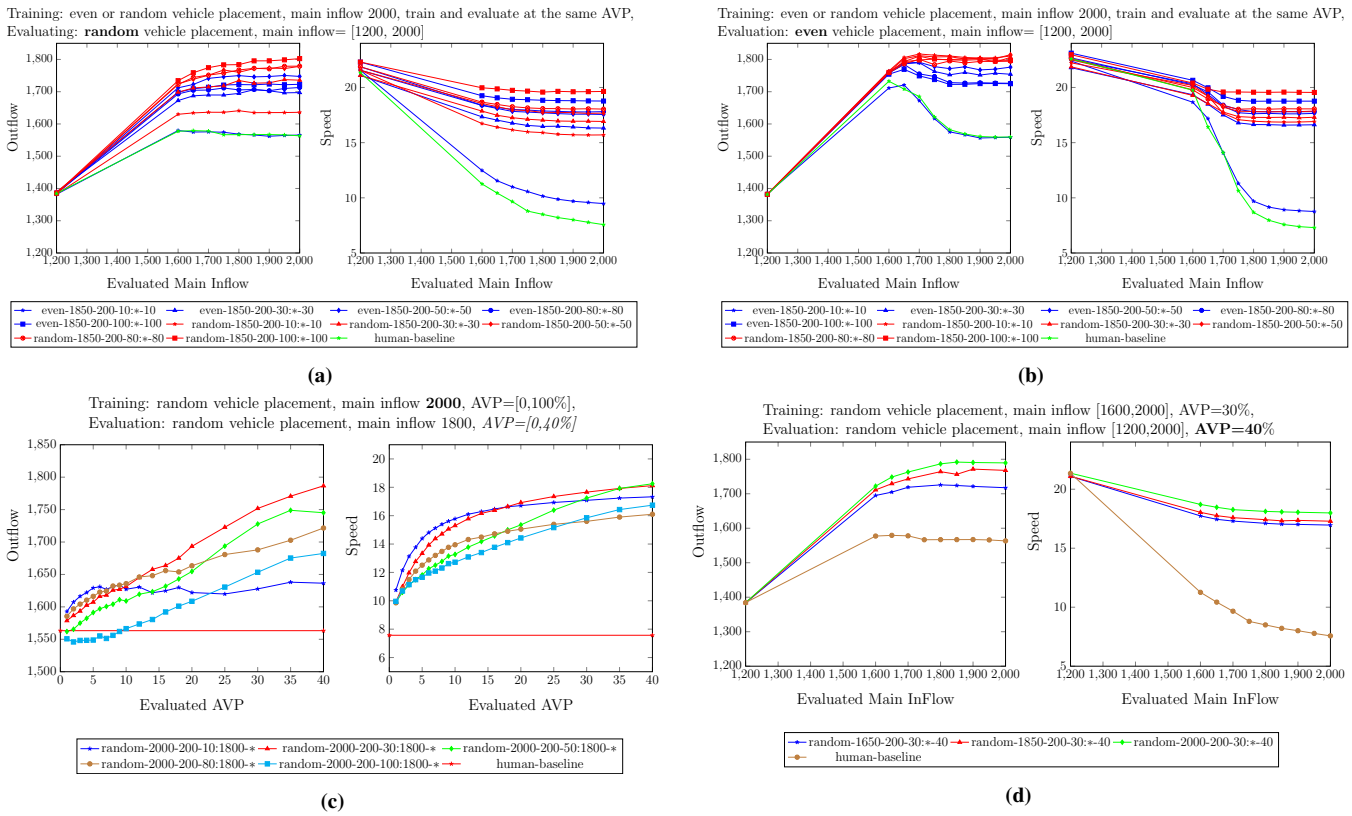

Figure 2: Results of policies trained under different AV placements, AV penetrations, main inflows. Figure (a)–(b): we show that the the policies under random vehicle placement is robust when evaluated under random or even vehicle placement. Figure (c): we fix the evaluation inflow, and find an AVP 30 % **that is robust when varying evaluation AVPs; Figure (d): we fix the evaluation AVP, and verify that main inflow** 2000 veh/hour **is also robust when varying evaluation inflows.**

Figure 3: A more complex road with two merging on-ramps.

is better than the human baseline even when the merging ramps are just 200 meters away. As we increase the distance between these two on-ramps, the performance increases. When this distance is small, the traffic congestion at the second merging ramp interferes with the traffic flow at the first merging ramp, but is not observable to the RL vehicles approaching the first ramp. As we increase the distance between these two merging ramps, such interference decreases and the traffic flow approaching these two merging ramps can be treated by the AVs increasingly independently. As a consequence, when these two merging ramps become further away from each other, the decision making processes for the AVs are similar to those on the single-lane merge roads — they only need to consider the traffic flow at the next incoming junction. Accordingly, the selected policy effectively reduces traffic congestion in the two-ramp scenario; and its performance improves as the distance between these two ramps increases.

5.2 Deployed to double-lane merge roads

Urban highway often consists of multiple lanes. Thus past research suggesting that AVs might *increase* traffic congestion on multi-lane roads [3] has (rightfully) raised concerns about the practical deployability of systems like the one considered in this paper. Contrary to those results, we find that AVs can reduce congestion even in multi-lane scenarios. Specifically, we consider a double-lane merge road as shown in Figure 5, by adding a second lane in the main road. Similar to that of the single-lane merge scenario, the vehicles in the right lane must yield to the vehicles from the merging lane and may cause potential congestion in the right lane. But the vehicles in the left lane have the right of way when passing the junction. As a consequence, the vehicles in the left lane tend to move at a faster speed, and there will be more vehicles changing from right to left for speed gain than the number of vehicles changing from left to

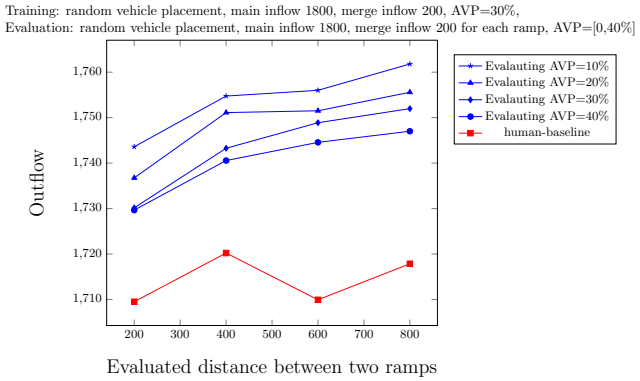

Figure 4: Results of deploying the selected training policy on roads with two on-ramps.

right. Those lane-changing vehicles cause additional stop-and-go waves in the left lane. To test the robustness of the selected policy in this new road structure, we deploy the learned policy to control the AVs on the right lane. During evaluation, there are only human-driven vehicles in the left lane with inflow 1600 veh/hour, and 10 % of the vehicles in the right lane are AVs, each of which is controlled by our learned policy. Figure 6 shows that the performance of the deployed policy is always significantly better than that of the human-only traffic, regardless of the right main inflow. We find that the learned policy, mitigating the congestion in the right lane, also reduces the amount of lane-changing vehicles since the right lane is less congested. Hence, the policy trained on the single-lane merge road generalizes well in the double-lane merge scenario.

Figure 5: A double-lane merge road.

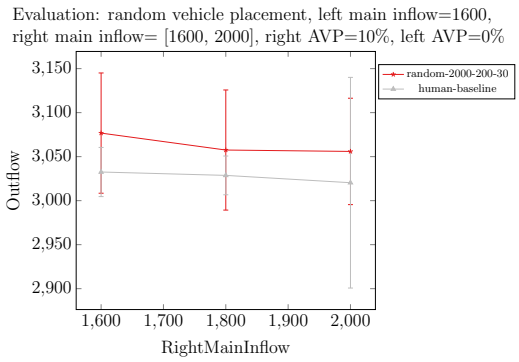

Figure 6: Results of deploying the selected training policy on the double-lane merge roads.

6 CONCLUSION AND FUTURE WORK

We presented an approach for learning a congestion reduction driving policy that performs robustly in road merge scenarios over a variety of traffic conditions of practical interest. Specifically, the resulting policy reduces congestion in AV penetrations of 1 %–40 %, traffic inflows ranging from no congestion to heavy congestion, random AV placement in traffic, single-lane single-merge road, single-lane road with two merges at varying distances, and double-lane single-merge road with lane changes. The process of finding this policy involved identifying a single combination of training conditions that yields a robust policy across different evaluating conditions in a single-lane merge scenario. We find, for the first time, that the resulting policy generalizes beyond the training conditions and road geometry it was trained on.

Recently there has been an increasing interest in developing RL training methods that result in robust policies. In our domain we find that randomizing AV placement and searching for an effective training setup over the space of traffic conditions achieve robustness effectively. The straightforward nature of our method and its limited set of assumptions and tuning parameters make it a potential candidate for real-world deployments. Given that RL algorithms have been shown to be brittle in many domains, finding an RL-based policy that performs robustly across a wide variety of traffic conditions in the challenging domain of multiagent congestion reduction is both encouraging and somewhat surprising.

Nonetheless, our work has a few limitations that could serve as important directions for future research. First, the question of whether there exists a driving policy that reduces congestion when deployed on the left lane of multilane scenarios still open. Second, our tests used the same aggressiveness level for all simulated human-driven vehicles. Testing with a variety of human behaviors would further increase the simulation results' applicability. Third, there is room to investigate a wider variety of road geometries beyond the ones we investigated. Finally, even after investigating these extensions, there will likely be a sim2real gap to close, due to noisy/limited sensing and actuation delay. These limitations notwithstanding, this paper's contributions and insights advance our ongoing effort to reduce traffic congestion via AV control in the real world.

ACKNOWLEDGMENTS

This work has taken place in the Learning Agents Research Group (LARG) at the Artificial Intelligence Laboratory, The University of Texas at Austin. LARG research is supported in part by the National Science Foundation (CPS-1739964, IIS-1724157, FAIN-2019844), the Office of Naval Research (N00014-18-2243), Army Research Office (W911NF-19-2-0333), DARPA, Lockheed Martin, General Motors, Bosch, and Good Systems, a research grand challenge at the University of Texas at Austin. The views and conclusions contained in this document are those of the authors alone. Peter Stone serves as the Executive Director of Sony AI America and receives financial compensation for this work. The terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its policy on objectivity in research.

REFERENCES

[1] Daniel S Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The complexity of decentralized control of markov decision processes. *Mathematics of Operations Research*, 27(4):819–840, 2002.

[2] Jiaxun Cui, William Macke, Harel Yedidsion, Aastha Goyal, Daniel Urieli, and Peter Stone. Scalable multiagent driving policies for reducing traffic congestion. *arXiv preprint arXiv:2103.00058*, 2021.

[3] Liam Cummins, Yuchao Sun, and Mark Reynolds. Simulating the effectiveness of wave dissipation by followerstopper autonomous vehicles. *Transportation Research Part C: Emerging Technologies*, 123:102954, 2021.

[4] Anthony Downs. *Stuck in traffic: Coping with peak-hour traffic congestion*. Brookings Institution Press, 2000.

[5] Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning*, pages 1329–1338, 2016.

[6] Daniel Krajzewicz, Jakob Erdmann, Michael Behrisch, and Laura Bieker. Recent development and applications of sumo-simulation of urban mobility. *International Journal on Advances in Systems and Measurements*, 5(3&4), 2012.

[7] Stefan Krauß. Microscopic modeling of traffic flow: Investigation of collision free vehicle dynamics. 1998.

[8] Abdul Rahman Kreidieh, Cathy Wu, and Alexandre M Bayen. Dissipating stop-and-go waves in closed and open networks via deep reinforcement learning. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 1475–1480, 2018.

[9] Tim Lomax, David Schrank, and Bill Eisele. 2021 urban mobility report. https://mobility.tamu.edu/umr/. Accessed: 2021-10-07.

[10] Namiko Mitarai and Hiizu Nakanishi. Convective instability and structure formation in traffic flow. *Journal of the Physical Society of Japan*, 69(11):3752–3761, 2000.

[11] Kanaad Parvate. On training robust policies for flow smoothing. (UCB/EECS-2020-197), 2020.

[12] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 2817–2826, 2017.

[13] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.

[14] Raphael E Stern, Shumo Cui, Maria Laura Delle Monache, Rahul Bhadani, Matt Bunting, Miles Churchill, Nathaniel Hamilton, Hannah Pohlmann, Fangyu Wu, Benedetto Piccoli, et al. Dissipation of stop-and-go waves via control of autonomous vehicles: Field experiments. *Transportation Research Part C: Emerging Technologies*, 89:205–221, 2018.

[15] Yuki Sugiyama, Minoru Fukui, Macoto Kikuchi, Katsuya Hasebe, Akihiro Nakayama, Katsuhiro Nishinari, Shin ichi Tadaki, and Satoshi Yukawa. Traffic jams without bottlenecks—experimental evidence for the physical mechanism of the formation of a jam. *New Journal of Physics*, 10(3):033001, 2008.

[16] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[17] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 23–30. IEEE, 2017.

[18] Martin Treiber and Arne Kesting. The intelligent driver model with stochasticity-new insights into traffic flow oscillations. *Transportation Research Procedia*, 23:174–187, 2017.

[19] Eugene Vinitsky, Kanaad Parvate, Aboudy Kreidieh, Cathy Wu, and Alexandre Bayen. Lagrangian control through deep-rl: Applications to bottleneck decongestion. In *21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 759–765, 2018.

[20] Eugene Vinitsky, Nathan Lichtle, Kanaad Parvate, and Alexandre Bayen. Optimizing mixed autonomy traffic flow with decentralized autonomous vehicles and multi-agent rl. *arXiv preprint arXiv:2011.00120*, 2020.

[21] Cathy Wu, Aboudy Kreidieh, Kanaad Parvate, Eugene Vinitsky, and Alexandre M Bayen. Flow: Architecture and benchmarking for reinforcement learning in traffic control. *arXiv preprint arXiv:1710.05465*, page 10, 2017.

[22] Cathy Wu, Aboudy Kreidieh, Eugene Vinitsky, and Alexandre M Bayen. Emergent behaviors in mixed-autonomy traffic. In *Conference on Robot Learning*, pages 398–407, 2017.

[23] Cathy Wu, Abdul Rahman Kreidieh, Kanaad Parvate, Eugene Vinitsky, and Alexandre M. Bayen. Flow: A modular learning framework for mixed autonomy traffic. *IEEE Transactions on Robotics*, pages 1–17, 2021.