Courteous Behavior of Automated Vehicles at Unsignalized Intersections via Reinforcement Learning

Shengchao Yan, Tim Welschehold, Daniel Büscher, Wolfram Burgard

***Abstract*—The transition from today's mostly human-driven traffic to a purely automated one will be a gradual evolution, with the effect that we will likely experience mixed traffic in the near future. Connected and automated vehicles can benefit human-driven ones and the whole traffic system in different ways, for example by improving collision avoidance and reducing traffic waves. Many studies have been carried out to improve intersection management, a significant bottleneck in traffic, with intelligent traffic signals or exclusively automated vehicles. However, the problem of how to improve mixed traffic at unsignalized intersections has received less attention. In this paper, we propose a novel approach to optimizing traffic flow at intersections in mixed traffic situations using deep reinforcement learning. Our reinforcement learning agent learns a policy for a centralized controller to let connected autonomous vehicles at unsignalized intersections give up their right of way and yield to other vehicles to optimize traffic flow. We implemented our approach and tested it in the traffic simulator SUMO based on simulated and real traffic data. The experimental evaluation demonstrates that our method significantly improves traffic flow through unsignalized intersections in mixed traffic settings and also provides better performance in a wide range of traffic situations compared to the state of the art.**

***Index Terms*—Intelligent transportation systems, reinforcement learning, deep learning methods.**

I. Introduction

OVER the past decades we observed a strong increase in the mobility of the population around the world. While, in general, this can be regarded as an indication of an improved quality of life, it does come with a strong increase in overall and individual traffic, creating a variety of problems including increased travel duration, high energy consumption, and increased pollution. A promising and practical solution to this problem is to increase the efficiency of the traffic. As intersections represent one of the major bottlenecks of traffic flow [1], optimizing the management of intersections is a highly important task. In the past, intersection management relied on traffic polices, semaphores, traffic lights, traffic signs and sets of rules. Furthermore, drivers also use turn signals, brake lights and even hand signals to communicate and cooperate with other traffic participants. Traffic control signals are not panacea for intersection problems [2]. For example, they may reduce traffic efficiency for low or unbalanced traffic demand. Although recent works [3], [4] developed more intelligent adaptive traffic signal control methods, for the majority of all intersections, which often have only one lane per road and mostly small traffic volume [5], the use of static road signs assigning priority has proven to be more efficient [2].

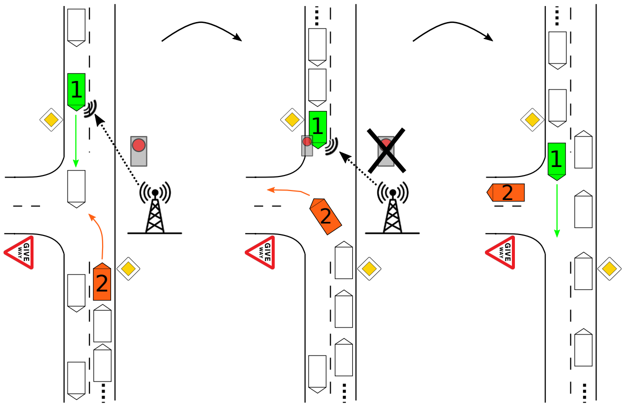


Fig. 1: Our intersection management agent optimizes traffic flow by assigning virtual red traffic lights to connected autonomous vehicles (vehicle number 1). Once vehicle 2 is released, the vehicles following it can also proceed through the intersection.

Nowadays, the first automated vehicles are mingling with the traffic and it is to be expected that their share will steadily increase in the future. Besides overcoming human limitations in driving and reducing accidents, these automated vehicles will supposedly be interconnected and thus offer new, more efficient ways of communication and traffic management. Based on the expectation that future traffic will consist of connected autonomous vehicles (CAVs), a large majority of current research excludes human-driven vehicles (HVs) in their development of traffic management approaches. However, it might take decades for the technology, the infrastructure and the users to be ready for traffic with only connected autonomous vehicles [6]. We therefore believe that, for the near future, applicable traffic management solutions must *i)* consider various degrees of mixed traffic, *ii)* pose no complications or major adjustment requests for human-driven vehicles, and *iii)* not present a traffic disturbance or danger when the communication between the connected autonomous vehicles

Manuscript received: June, 10, 2021; Revised September, 1, 2021; Accepted October, 4, 2021. This letter was recommended for publication by Editor Youngjin Choi upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by the Priority Programme "Cooperative Interacting Automobiles" of the German Science Foundation DFG.

The authors are with the Department of Computer Science, University of Freiburg, Germany. {yan, twelsche, buescher, burgard} @cs.uni-freiburg.de

Digital Object Identifier (DOI): see top of this page.

fails.

One might argue that HVs lack means of efficient communication and coordination with other road users so that unsignalized intersections with mixed traffic cannot benefit from the introduction of CAVs [7]. However, Ulbrich *et al.* [8] showed that humans cooperate with other traffic participants to improve the whole traffic utility. Consider, as an example, the situation shown in Fig. 1. Even though vehicle 1 has higher priority and can proceed through the intersection before vehicle 2, its driver might prefer to yield to vehicle 2 so that the traffic behind vehicle 2 can be released sooner.

In this paper, we propose a novel centralized method to improve intersection management in mixed traffic. Our approach learns a policy for CAVs that maximizes the overall utility while at the same time showing courteous behavior [9]. We make the following contributions:

- We present a centralized intersection management method based on deep reinforcement learning that improves traffic performance at unsignalized intersections in mixed traffic scenarios.
- We introduce *return scaling* for training in environments with a large imbalance of cumulative rewards at different states. In our case, this helps to balance policy updating of states with different traffic densities, in particular to counteract the large cumulative reward collected in heavy traffic, which would otherwise dominate the stochastic gradient descent process and make the policy unstable for states in sparse traffic.
- We present a comprehensive performance comparison for various traffic densities and changing rates of CAVs to demonstrate the potential of our approach.

We conduct experimental studies in the traffic simulation environment SUMO [10] and show that our method outperforms the state-of-the-art intersection management method on a wide range of traffic densities with varying traffic distributions on the incoming lanes.

II. Related Work

Among the first ones to propose an intelligent intersection management system were Dresner and Stone whose reservation-based approach [11], [12] divides the junction with intersecting trajectories into a grid of tiles. Their autonomous intersection management approach, realized as a centralized controller, applies a first-come-first-served (FCFS) strategy to deal with the requests by CAVs for time slots of the tiles along their trajectories. To accommodate HVs they employ traffic lights and the so-called FCFS-light policy [13], [14]. Later, this framework was extended to allow for the centralized intersection management to set the speed profiles of vehicles with cruise control [15]. To improve the performance of FCFS-light, Sharon and Stone introduced hybrid autonomous intersection management [16]. With this extension, requests of CAVs can be approved regardless of the traffic lights if there are no HVs in the intersecting routes.

In general, the methods based on autonomous intersection management [11] provide a relative advantage to CAVs over HVs, which, in our opinion, should be avoided as it might cause the public to repel automated vehicles. Furthermore, human drivers will be more sensitive to stopping and waiting than the passengers in CAVs. We therefore suggest that the benefit brought by intersection management and CAVs in general should be evenly shared with human drivers.

Lin *et al.* developed a method similar to the FCFS-light policy [17]. It reserves conflicting sections among different routes instead of the grid of tiles. Another first-come-first-served reservation based method has been proposed by Bento *et al.* [18]. They suggest to control both CAVs and HVs via speed profiles sent by the intersection management unit. This places an undesirable burden on human drivers to follow a given speed profile and additionally even requires all HVs to be connected.

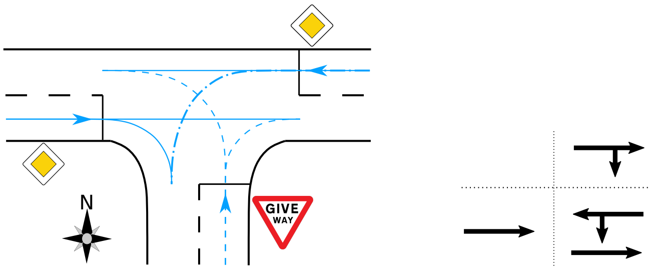
The described approaches make the vehicles roughly follow a first-come-first-serve strategy to traverse intersections. However, as shown by Meng *et al.* [19], the performance of an intersection management strategy mainly depends on the passing order of vehicles and not so much on the individual trajectory planning algorithms. As the computation time grows exponentially with the number of considered vehicles [19], often simplifying assumptions are made including linear constraints, no overtaking, no lane changing, constant speed, and constant traffic input. The coordination of the passing order can mitigate control uncertainties, which makes it more suitable for mixed traffic. Based on this idea, our work is aimed at finding better passing orders, while having vehicles drive based on their own trajectory planning model.

Qian *et al.* [20] assign priorities representing the passing order to vehicles. While CAVs receive the priority from a central control unit and plan trajectories accordingly, the passing order of HVs is regulated by traffic lights. With high rates of HVs, this potentially results in an inefficient, mostly first-come-first-served control. Fayazi *et al.* [21] propose to formulate the intersection management problem as a mixed-integer linear program. Their controller assigns times of arrivals to a virtual access area around the junction to CAVs, while HVs are regulated by traffic lights.

The approaches of these related works are already outperformed by Webster's method or fixed-time traffic signal controllers when over 10% to 20% of the vehicles are driven by humans [11], [13], [17], [21]. The exception is our previous state-of-the-art learning-based adaptive traffic signal controller, which further outperforms these two controllers in any traffic flow range and reduces the average travel time by up to 30% to 60% in the experiment with real-world traffic input [4]. Therefore, we evaluate our proposed method mainly against [4] in a wide range of dynamic traffic demands and show that the performance gain is available even with a small portion of CAVs in the traffic system.

III. Methods

Deep reinforcement learning has shown great potential for solving complex decision making and controlling problems [22], [23]. Accordingly, we model the intersection management task at unsignalized intersections as a Markov Decision Process, where the agent follows a policy $\pi(a \mid s)$



(a) Three-way intersection with six routes. (b) Four actions.

Fig. 2: Common regulation of a right-hand traffic three-way intersection (a). The high-priority-routes are W-E, W-S and E-W. The low-priority-routes are S-W and S-E. Route E-S has intersecting routes with higher and lower priority. The proposed set of actions (b) stops CAVs on routes along the indicated directions.

in a specific environment. Based on its state s_t the agent selects an action $a_t \in \mathcal{A}$ according to the policy, transits to a successor state s_{t+1} and receives a reward $r_t \in \mathbb{R}$. The agent is aimed at maximizing the expectation of the return (discounted cumulative reward) $G(s_t) = \sum_{i \geq t} \gamma^{i-t} r_i$, where $\gamma \in [0,1]$ is the discount factor.

To find the optimal policy, we use proximal policy optimization [23] due to its stability, good performance and ease of implementation. For a policy π_θ parameterized by θ, the algorithm maximizes the following objective:

$$\mathcal{J}_\theta = \mathbb{E}_t\Big[\min\Big(\rho_t(\theta)A_t, \text{clip}\left(\rho_t(\theta), 1-\epsilon, 1+\epsilon\right) A_t\Big) + \beta_{\text{entropy}} \cdot H\Big(\pi_\theta(s_t)\Big)\Big], \tag{1}$$

where the expectation is taken over samples collected by following $\pi_{\theta_{\text{old}}}$, and $\rho_t(\theta) = {}^{\pi_\theta(a_t|s_t)}/_{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ is the importance sampling ratio. The function H represents the entropy of the current policy and β_{entropy} adjusts the strength of entropy regularization. The term A_t is a truncated version (on trajectory segments of length up to K) of the generalized advantage estimator [24], which is an exponentially-weighted average (controlled by λ):

$$A_t = \delta_t + (\gamma\lambda)\delta_{t+1} + \cdots + (\gamma\lambda)^{K-1-t}\delta_{K-1}, \tag{2}$$

where $\delta_t = r_t + \gamma V_{\phi_{\text{old}}}(s_{t+1}) - V_{\phi_{\text{old}}}(s_t)$. The value function V_ϕ, parameterized by ϕ, is learned by minimizing the following loss (with coefficient β_{value}):

$$\mathcal{L}_\phi = \beta_{\text{value}} \cdot \mathbb{E}_t\left[\|V_\phi(s_t) - G(s_t)\|_2^2\right]. \tag{3}$$

Our work is aimed at training a centralized agent for an intersection that timely stops the CAVs on the routes with higher priority to let the vehicles on conflicting routes with lower priority pass, so that the performance of the whole system is optimized. Since this is similar to red traffic lights for CAVs on the routes with higher priority, we denote our method as *Courteous Virtual Traffic Signal Control* (CVTSC). We evaluate our proposed approach on the most common type of three-way intersections as illustrated in Fig. 2. By adjusting the state and action representations, our approach can easily be generalized to other intersection layouts, as we show for the real-world intersection in Sec. IV-E.

A. Background

As we focus on an isolated intersection, we assume that the vehicles can drive freely after they passed the junction and entered the outgoing lanes. Thus the vehicles on the outgoing lanes do not influence the intersection management. However, unlike in our previous work [4], in which we only considered vehicles in front of the stop lines, we here also take the vehicles into account, which already passed the stop line but not yet entered the outgoing lanes. This is necessary as at unsignalized intersections vehicles very often choose to wait after stop lines and coordination may happen there inside the junction.

In the following we give some definitions of quantities relevant to our approach:

- Throughput (N^{TP}): The number of vehicles that enter outgoing lanes during step t is denoted N_t^{TP}.
- Travel time (T_{travel}): For each vehicle passing a junction, its travel time is measured as the time period starting from its scheduled spawning time in the simulator (accounting for potential delays caused by traffic jams at the intersection) and ending when it enters an outgoing lane. For vehicles not released at the end of an episode, the travel time is counted until the episode ends.
- Traffic flow rate (F) and Saturation flow rate (F_s): F represents the number of vehicles (in vehicles per hour ${}^{\text{v}}/_{\text{h}}$) that pass through a point, e.g., an intersection or one lane, in unit time. The term F_s is a constant representing the theoretical upper limit for the traffic flow rate.

B. Action Space

For the intersection in Fig. 2a we assume that vehicles drive according to the priorities predefined by the road signs, where the diamond indicates priority roads and the triangle indicates yield. Vehicles on the routes with lower priority have to wait until there is enough gap on the conflicting routes with higher priority before passing the junction. Note that in Fig. 2a the route E-S has intersecting routes with higher and lower priority.

To obtain courteous behavior for CAVs on routes with higher priority, without loss of generality, we define a discrete set of four actions {(), (W-E), (W-E, W-S), (W-E, E-W, E-S)} as the action space $\mathcal{A}$ in relation to Fig. 2b. The indicated directions show the corresponding routes on which the intersection management unit commands CAVs to halt before the respective stop lines to give priority to vehicles waiting on intersecting routes with lower priority. The action restricting no routes uses the default priorities to manage the intersection. We set the duration of each action to 1 second. A categorical policy is learned: during training the actions are sampled according to the output distribution, while during testing the action with the highest probability is always chosen. When a new action a_t is chosen, CAVs on the routes indicated in a_t will receive stopping commands, while the instruction for the routes restricted by a_{t-1} is canceled, if they are not regulated by a_t. If a CAV receives a stopping command while being too close to the stop line, it will continue through the intersection thus ignoring the command. Acceleration, collision avoidance and safe distance are managed by the low-level controllers of the individual vehicles (both CAVs and HVs).

C. State Space

Due to the restriction of sensors and wireless communication, we assume that the intersection management unit can collect information of vehicles that are within a distance of $150\mathrm{m}$ along the road measured from the center of the intersection. We assume that every vehicle's state, composed of continuous values (its position along the road, velocity and time since entering intersection) and discrete values (a binary value for CAV or HV and optionally a route index indicating the driving direction if the lane contains more than one route), is available to the control unit. Similar to our previous work [4], the state s_t of the intersection at time t is given by a vector that contains the structured information of vehicles in it. The intersection is divided into several lane segments. The capacity of each segment is the maximum amount of vehicles in it during a traffic jam. The states of all vehicles in one segment ordered by their distances to the stop line constitute a part of s_t with a fixed length. Default values are given when fewer vehicles are present than the capacity. The states of all lane segments are concatenated into s_t in a fixed order.

As described in Sec. III-B, only CAVs are controlled by the agent. Every 1 second a new action should be chosen according to the new state. However, at certain points in time there are no CAVs in the intersection and including these states in training hinders the learning process. We therefore remove states without CAVs from the training data. As a result, the influence of actions is not limited to a fixed interval and the duration of one step in the learning process can be any positive integer in seconds. To deal with this variable step length, we employ the method of *adaptive discounting* as proposed by Yan *et al.* [4].

D. Reward Function

The common objective of intersection management methods is to improve the efficiency while keeping a certain level of fairness for all vehicles. In this work, we extend the idea of a reward function with equity factor [4]. Instead of using ${T_{travel}}^{\eta}$, we propose to use $\eta_\mathrm{a} \cdot T_{travel} + \eta_\mathrm{b}$ as the reward for each released vehicle, where η, η_a and η_b are equity factors. Due to the flexible step lengths discussed above, the reward of each step r_t is calculated by accumulating discounted rewards generated during step t which might contain up to k environment steps (each one second). I.e., we accumulate the contribution of N_t^{TP} released vehicles by

$$r_t = \sum_{i=0}^{k-1} \gamma^i \sum_{j=1}^{N_{t\_i}^{\mathrm{TP}}} (\eta_\mathrm{a} \cdot \tau_j + \eta_\mathrm{b}), \tag{4}$$

where $N_{t\_i}^{\mathrm{TP}}$ is the throughput of the ith second in step t and τ_j is the travel time of the jth released vehicle in the ith second.

The values of η_a and η_b are selected as by Yan *et al.* [4] based on two heuristics. First, we favor releasing each vehicle as soon as possible for the purpose of efficiency. The second heuristic aims at equity by considering a traffic situation, where one vehicle waits for saturated traffic flow on an intersecting route. Since efficient traffic flow on the high priority route should not be achieved on the expense of accumulating too large waiting time on the single vehicle, we increase the reward contributed by each released vehicle according to its travel time. This linear relation between reward and travel time is more intuitive than the previous exponential formulation. Moreover, the additional free variable in this formulation can be used to scale the rewards of single released vehicles to keep them around unity, which is beneficial for hyperparameter tuning in common deep reinforcement learning setups.

E. Return Scaling

According to the reward definition, the return $G(s_t)$ is mainly influenced by the throughput and the travel time of released vehicles. Since both of them increase with the traffic input, the scale of $G(s_t)$ could vary from less than 5 to over 100 if the state of the intersection changes from nearly empty s_low to saturated s_high. Consequently, s_high would have a much larger impact on π_θ and V_ϕ during the update phase, making the learning process of a policy for light traffic less stable.

We introduce *return scaling* to resolve the issues caused by imbalanced return of states, which has shown to be critical for convergence with low traffic volumes in our experiments. In order to reduce the difference between $G(s_\text{low})$ and $G(s_\text{high})$, we scale the cumulative rewards before the update phase with

$$G(s_t) = \rho(s_t) \cdot \sum_{i \geq t} (\gamma^{\sum_{j=t}^{i-1} k_j}) r_i, \tag{5}$$

where k is the number of environment steps (each one second) in one step of learning process. The scaling factor ρ is defined as

$$\rho(s_t) = ({N_\mathrm{c}^\mathrm{V}}/{n^\mathrm{V}})^{\xi}, \tag{6}$$

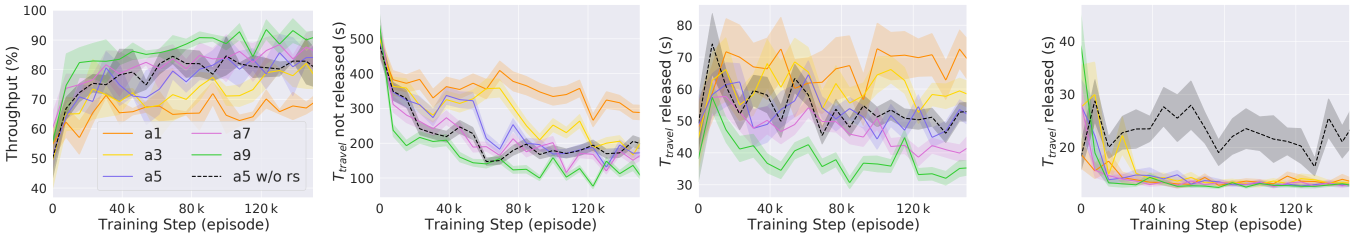
where n^V and N_c^V are the current number of vehicles in the intersection and its capacity, respectively, and ξ is a hyperparameter.

IV. Experiments

A. Experimental Setup

We use the open-source traffic simulator SUMO [10] to train and evaluate various intersection management agents. Besides simulated traffic episodes we also evaluate our approach on real-world rush hour traffic demand. For all roads we set a speed limit of $50\,{}^{\mathrm{km}}/_{\mathrm{h}}$. We compare our approach CVTSC to baselines managing the intersection with *road signs (RS)* defining static priorities for routes and with *traffic lights (TL)* controlled by a deep reinforcement learning agent according to our previous work [4]. A possible set of green phases for the three-way intersection is shown in Fig. 4.

Two fully connected networks θ and ϕ are used as the policy and value function estimators. They both have an input layer of size 343 and two hidden layers of size 2,048 (ReLU) and 1,024 (ReLU). The output layer is of size 4 for θ and 1 for ϕ. A grid search was used to select the hyperparameters.



(a) $2\,000 \sim 3\,000$ v/h

(b) $0 \sim 1\,000$ v/h

Fig. 3: Results obtained in evaluation during training for all agents with varying CAV rates in traffic (solid lines) and an ablation study for the usage of the return scaling (dashed black line). The plots show the mean with standard deviation, where the latter is scaled by $\pm 1/10$ for the travel times (for clearer visualization), over three non-tuned random seeds. By the end of each episode there are still some vehicles, which have not passed the junction. The travel time for such a vehicle is calculated with $T_{\text{episode}} - T_{\text{spawn}}$, where T_{episode} is the episode duration and T_{spawn} is its scheduled spawning time in the simulator.

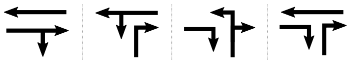


Fig. 4: Traffic light green phases for the intersection in Fig. 2a.

We use $5\text{e}{-6}$ as the learning rate for the Adam optimizer and 0.001 as the coefficient for weight decay. For proximal policy optimization algorithm, we use 32 actors, the clipping threshold of $\epsilon = 0.001$ and the discount factor of $\gamma = 0.98$. For the return scaling factor, we use $\xi = 0.2$, which is found to be the optimal value in the range of $(0, 1]$. In each learning step mini-batches of size 100 are used to update the agents in 8 epochs. The number of mini-batches in each learning step is, however, variable due to the varying step lengths. The equity factors η_{a} and η_{b} for reward calculation are set to 0.0027 and 0.946. The training process of $150\,\text{k}$ steps takes about 40 to $60\,\text{h}$ (depending on the corresponding CAV rate) running on four NVIDIA TITAN X GPUs, while CPU computation is not a limiting factor.

B. Training Setup

Most current related work has been developed and tested with simplified traffic demand, such as constant traffic input to the intersection. We challenge our approach and train it with more dynamic traffic input ranges to cover as many real traffic scenarios as possible. For the three-way junction in Fig. 2a the saturation flow rate F_s of each incoming lane is $1\,670$ v/h and as it is very rare that two non-conflicting routes are simultaneously saturated, we set the traffic demand range to $[F_{\min}, F_{\max}] = [0, 3\,000]$ v/h.

We train our agents online on simulated traffic episodes with a duration of $1\,200$s. First, the total traffic input F_{begin} is randomly sampled in $[F_{\min}, F_{\max}]$. Then F_{end} is sampled uniformly within $[\max(F_{\min}, F_{\text{begin}} - 1\,500), \min(F_{\max}, F_{\text{begin}} + 1\,500)]$. After that the beginning and ending traffic flow for each route is randomly sampled from an uniform distribution, such that they sum up to F_{begin} and F_{end}, respectively. Finally, the traffic flow during the episode is generated by linear interpolation between these two values for each route. We train five agents (*a1, a3, a5, a7, a9*), each corresponding to a fixed CAV rate of $[10, 30, 50, 70, 90]\,\%$, corresponding to the expected increasing CAV rates in the future traffic.

C. Evaluation during Training

To monitor the learning process the performance is evaluated for traffic input of different ranges: $[0, 1\,000]$, $[500, 1\,500]$, $[1\,000, 2\,000]$, $[1\,500, 2\,500]$, $[2\,000, 3\,000]$. The generation of traffic demand is analogous to that of training episodes except that the total traffic inputs at the beginning F_{begin} and end F_{end} are sampled independently in the five given ranges.

The plots in Fig. 3 show the performance of agents trained with different CAV rates and present an ablation study for the usage of the *return scaling*. The agent *a5 w/o rs* is trained with a CAV rate of 50% without using return scaling. We analyze the throughput in percentage of released vehicles among all spawned vehicles, the travel time of released and not released vehicles at the highest traffic density level and the travel time of released vehicles at the lowest level. The calculated travel time is the mean among all released or not released vehicles during three evaluation episodes. We analyze the throughput and travel times instead of the accumulated reward as they give us a better estimate of the overall performance. The variance of the travel times is of particular interest as it is a good indicator for the equity. Large variances correspond to some vehicles with long waiting times at the intersection.

As illustrated in Fig. 3a, CVTSC with higher CAV rate leads to more throughput, more efficient clearance (lower average T_{travel}) of the intersection and more fairness (shown by lower standard deviation of T_{travel}) to all the vehicles. As expected, from Fig. 3b and the travel time plots of Fig. 3a, we observe that the agent without return scaling fails to learn an efficient policy for light traffic, although its performance is similar to that of *a5* in heavy traffic. We plan to conduct further investigation on return scaling, in particular whether it is applicable to a broader class of problems or can be replaced with other methods like γ-tuning.

D. Evaluation on Simulated Traffic Demand

We first test our agents with simulated traffic episodes, each with a duration of one hour. For each of the five traffic demand levels described above, we first create 50 traffic episodes with spawning time of each vehicle following the procedure to that for evaluation during training. Then we generate five sets of mixed traffic episodes with different CAV rates by randomly setting each vehicle as CAV or HV according to the

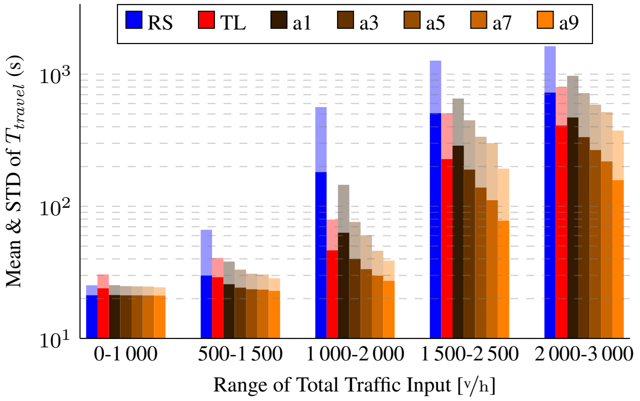


Fig. 5: Performance comparison of our CVTSC with baselines RS and TL in traffics with different CAV rates. For each controller with each traffic density, the mean (opaque bars) and positive standard deviation (translucent bars) of T_{travel} are calculated over all vehicles (including released and not released) of 50 simulated traffic episodes. Each CVTSC agent is trained and evaluated in traffics with its corresponding CAV rate.

Traffic Input (v/h)	Throughput (%)						
	RS	a1	a3	a5	a7	a9	TL
0 ∼ 1 000	99.4	99.4	99.4	99.4	99.4	99.4	99.4
500 ∼ 1 500	99.2	99.3	99.3	**99.4**	99.3	**99.4**	99.2
1 000 ∼ 2 000	91.1	97.7	98.6	99.0	99.1	**99.2**	98.5
1 500 ∼ 2 500	72.2	85.3	90.6	93.5	94.7	**96.8**	88.5
2 000 ∼ 3 000	59.8	74.6	82.1	85.8	88.5	**91.9**	77.9

TABLE I: Throughput (%) of considered methods in Fig. 5.

penetration rate. Note that the baseline methods *road sign (RS)* and *traffic light (TL)* do not distinguish between CAV and HV. Following this setup, we test both baselines and our trained agents with identical number of vehicles and same spawning times. In the following, the five agents are first tested with their corresponding CAV rates to evaluate their performance against the baseline methods. Then we cross-evaluate them on settings corresponding to different CAV penetration rates.

1) Performance of Intersection: The performance is shown in Fig. 5 and Table I. For all the tested traffic density levels, our CVTSC agents can improve the performance of the unsignalized intersection. Not only more vehicles are released during the same period, but also the mean and standard deviation of their travel times are reduced. The higher the CAV rate is, the better our approach performs. The performance gain of CVTSC on the lowest traffic density is not obvious, because nearly no vehicles have to stop at the junction. When there is little traffic, employing *TL* can cause unnecessary stopping due to the transition phase (amber or red lights). In heavier traffic over 1 500v/h *TL* outperforms *a1* by a little margin. However, it is outperformed by CVTSC when 30% or more vehicles are CAVs.

2) Performance of Vehicle Groups: In contrast to the relative advantage of CAVs over HVs suggested by the methods based on autonomous intersection management [13], our CVTSC tends to share the performance gain evenly between the two types of vehicles. Fig. 6 shows how CVTSC can increase the intersection management performance while keeping the balance between different vehicle categories. Since the actions are executed only for CAVs on the main road, we divide vehicles on the main road into *Main road CAV* and *Main road HV* and assign all vehicles on the side road to a third group *Side road all*. As illustrated, the performance gain against *RS* is mainly caused by the improvement of the traffic on the side road. With only 10% CAVs the throughput of the side road traffic is increased from 74.3% to 95.6% and the median travel time is decreased by 61%. As a necessary side effect, the courteous behavior adds about 13 s to the median travel time of CAVs on the main road and slows down some HVs following the CAVs consequently. However, the median travel time of Main road HV and the throughput of both vehicle groups on the main road are nearly not influenced. With growing rate of CAVs in traffic, the performance of the traffic on the side road continues to be improved while the initial disadvantage for the main road is compensated.

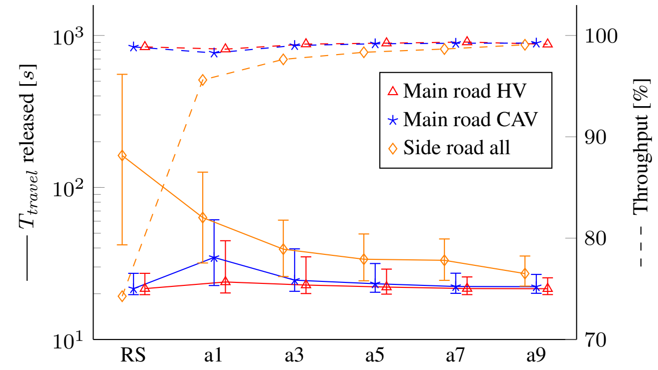


Fig. 6: Performance comparison of different vehicle groups at traffic demand 1 000 ∼ 2 000 v/h. The plotted travel times show the median, lower quartile and higher quartile over all released vehicles among all evaluated episodes. The plotted throughput is the percentage of released vehicles among all spawned vehicles throughout all episodes.

3) Comparison of Agents: To cross-evaluate their performance on other traffic settings than their natives, we further test each agent (a1 to a9) on the five different CAV rates on 50 simulated episodes on each of the five traffic densities. Since CVTSC brings nearly no measurable difference for the lowest traffic density, only the results for the other four traffic densities are listed in Table II.

We observe that all trained CVTSC agents outperform *RS* in any mixed traffic setting. Furthermore, two significant patterns can be observed in the results. First, for each CAV rate the agents trained with similar rate values are among the best, as expected. Second, as the CAV rate increases the performance of all agents is continuously improved. Interestingly, *a5*, the one trained with CAV rate of 50%, outperforms or performs equally well as *a7* and *a9* even in settings where CAVs are the majority. We suppose this is because *a5* during training is exposed to more diverse traffic situations, especially ones with fewer CAVs in the intersection. As shown in Fig. 5 and Fig. 6, the margin of the performance gain decreases with increased CAV rate. Even though *a7* and *a9* can handle highly automated traffic better than *a5*, the performance gain is so small that it can not compensate the performance loss when occasionally more HVs drive in the intersection.

Traffic Input		Average T_{travel} [s]					Throughput [%]				
CAV Rate	Input Flow [$^{\mathrm{v}}/_{\mathrm{h}}$]	a1	a3	a5	a7	a9	a1	a3	a5	a7	a9
10%	$500 \sim 1\,500$	25.8	**25.7**	25.8	26.3	27.5	99.3	99.3	99.3	99.3	99.3
	$1\,000 \sim 2\,000$	**63.2**	69.9	80.8	102.1	131.4	**97.7**	97.4	96.9	95.8	94.2
	$1\,500 \sim 2\,500$	**287.8**	299.3	339.0	364.2	432.3	**85.3**	84.7	82.1	80.6	77.1
	$2\,000 \sim 3\,000$	**471.0**	482.5	517.8	554.4	610.1	**74.6**	73.9	72.0	69.3	65.9
30%	$500 \sim 1\,500$	24.7	**24.3**	24.4	24.9	24.8	99.3	99.3	99.3	99.3	99.3
	$1\,000 \sim 2\,000$	42.1	**40.0**	43.4	49.4	58.7	98.5	**98.6**	**98.6**	98.2	98.0
	$1\,500 \sim 2\,500$	213.4	**190.0**	204.2	237.9	274.1	89.5	**90.6**	90.0	88.1	85.9
	$2\,000 \sim 3\,000$	367.3	**334.2**	347.0	411.7	430.6	80.3	**82.1**	81.4	77.5	76.6
50%	$500 \sim 1\,500$	24.1	**23.6**	**23.6**	23.9	23.9	99.3	99.3	**99.4**	99.3	99.3
	$1\,000 \sim 2\,000$	36.0	33.7	**33.5**	35.6	38.9	98.9	**99.0**	**99.0**	98.9	98.7
	$1\,500 \sim 2\,500$	191.0	145.5	**138.8**	159.7	174.7	90.6	93.2	**93.5**	92.3	91.8
	$2\,000 \sim 3\,000$	346.9	269.0	**267.0**	308.6	313.4	81.4	**85.9**	85.8	83.5	83.3
70%	$500 \sim 1\,500$	23.6	23.2	**23.1**	23.4	23.3	**99.4**	**99.4**	**99.4**	99.3	99.3
	$1\,000 \sim 2\,000$	32.6	29.8	**28.6**	29.9	30.0	99.0	**99.1**	**99.1**	**99.1**	**99.1**
	$1\,500 \sim 2\,500$	176.3	120.7	**101.1**	111.2	112.1	91.2	94.4	**95.6**	94.7	95.0
	$2\,000 \sim 3\,000$	323.2	234.2	**203.5**	219.0	217.0	82.6	87.3	**89.3**	88.5	88.5
90%	$500 \sim 1\,500$	23.1	22.8	**22.6**	23.1	22.9	**99.4**	**99.4**	**99.4**	99.3	**99.4**
	$1\,000 \sim 2\,000$	30.3	27.9	**26.7**	27.4	27.3	99.0	**99.2**	**99.2**	**99.2**	**99.2**
	$1\,500 \sim 2\,500$	164.8	105.5	77.0	**76.5**	77.9	91.8	95.2	**96.8**	96.7	**96.8**
	$2\,000 \sim 3\,000$	311.5	192.0	161.6	**154.5**	157.8	83.2	90.0	91.9	**92.2**	91.9

TABLE II: Performance comparison of different agents with different traffic input settings. For each agent with each traffic setting, the average T_{travel} is calculated over all vehicles (including released and not released) of 50 simulated traffic episodes.

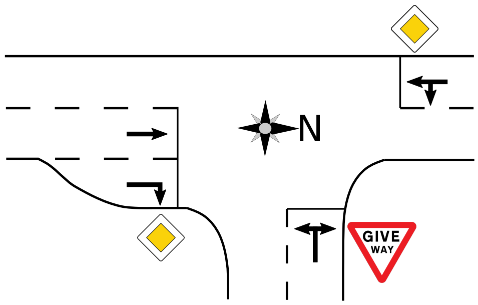


Fig. 7: Intersection of Tullastrasse and Hans-Bunte-Strasse in Freiburg, Germany.

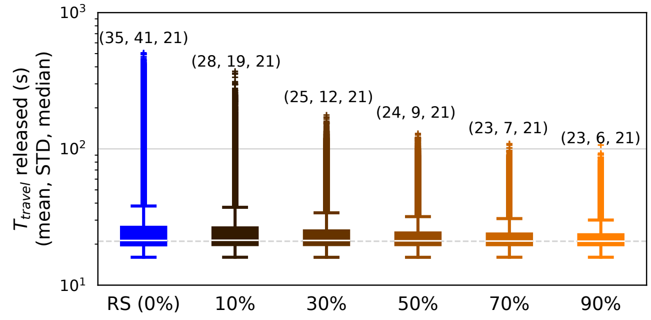


Fig. 8: Box plot of travel times with different CAV rates over all released vehicles in the simulation based on the real-world intersection of Fig. 7. The whiskers extend $1.5 \cdot$ IQR (interquartile range) from the upper and lower quartiles.

E. Evaluation on Real-world Traffic Demand

To further evaluate CVTSC in more realistic traffic situations, we conduct additional tests with real-world traffic demand recorded at an intersection in Freiburg, Germany, which is sketched in Fig. 7. Unlike the intersection in Fig. 2a, one part of the main road (Tullastrasse) forks before the stop line. After adjusting the state representation and the intersection structure in the simulator we trained two new agents *a3* and *a5* and employ them in the test. The traffic demand, listed in Table III, was manually recorded on October 19, 2017 by the traffic department of Freiburg. The total traffic input was about $1\,000 \sim 1\,500\,{}^{\mathrm{v}}/_{\mathrm{h}}$ with roughly 20% on the side road.

Fig. 8 shows box plots of the travel times of released vehicles controlled by *RS* and CVTSC agents in traffic scenarios with different CAV rates. The agent *a3* is employed for 10% and 30% automated traffic, while *a5* is employed for the other three. In all scenarios over 99.7% of all vehicles traverse the intersection. Our method continuously improves the traffic flow with increasing rate of CAVs in traffic. We notice that the median of travel times in all scenarios stay similar, which means the performance gain comes mainly from the vehicles with long travel times on the side road. CVTSC agents manage to release them faster without delaying the traffic on the main road.

V. Conclusion

In this paper we present a novel approach to managing mixed traffic at unsignalized intersections using deep reinforcement learning. Our proposed method CVTSC creates courteous behavior for automated vehicles in order to optimize the overall traffic flow at intersections. Furthermore, we introduce return scaling to counteract the imbalance of cumulative rewards at different states and to stabilize training. We validate the effectiveness of CVTSC using simulated and real-world traffic data and show that CVTSC improves the traffic performance continuously with increasing percentage of automated vehicles. With more than 10% of automated vehicles it also outperforms the state-of-the-art adaptive traffic signal controller. Besides the performance gain, our method does not require a change of the current driving habits of humans. Moreover it is fault-tolerant, since the method is an add-on to the existing traffic rules and thus the intersection will

Direction	Traffic Input (Number of Vehicles every $15\,\mathrm{min}$)															
	7:15	7:30	7:45	8:00	8:15	8:30	8:45	9:00	16:15	16:30	16:45	17:00	17:15	17:30	17:45	18:00
N-S	55	63	101	80	98	85	60	111	102	104	79	97	148	122	104	67
N-E	44	29	38	44	32	44	28	31	32	44	26	28	32	37	38	19
S-N	71	76	96	111	78	86	80	65	105	88	119	116	112	86	100	108
S-E	35	41	32	53	68	42	52	43	29	32	29	36	33	30	29	27
E-N	11	26	29	20	40	29	20	22	58	48	56	35	55	50	47	35
E-S	16	25	51	26	31	21	32	22	53	32	43	23	32	19	31	25

TABLE III: Traffic in rush hours on the morning and afternoon of October 19, 2017 at the intersection of Fig. 7.

still be fully functional even if the intersection management unit fails. Last but not least, our method can be easily adopted to different intersections.

In future work, we plan to investigate how much the uncertainty of the input states can decrease the performance and how to mitigate this influence. Furthermore, we plan to develop state encoders for variable numbers of vehicles in single lanes, which can be directly used for a new intersection to accelerate the training process, instead of training the whole policy network from scratch. Finally, an exciting area may be extending the centralized controller to a decentralized environment, where no infrastructure for perception and decision making is available and the automated vehicles have to communicate and decide whether to yield according to their incomplete local information.

References

[1] L. Wu, Y. Ci, J. Chu, and H. Zhang, "The influence of intersections on fuel consumption in urban arterial road traffic: A single vehicle test in harbin, china," *PloS one*, vol. 10, no. 9, p. e0137477, 2015.

[2] U.S. Federal Highway Administration, "Manual on Uniform Traffic Control Devices," https://mutcd.fhwa.dot.gov/pdfs/2009r1r2/pdf_index.htm, 2009, [Online; accessed 06-Mar-2021].

[3] P. Varaiya, "The max-pressure controller for arbitrary networks of signalized intersections," in *Advances in Dynamic Network Modeling in Complex Transportation Systems*. Springer, 2013, pp. 27–66.

[4] S. Yan, J. Zhang, D. Büscher, and W. Burgard, "Efficiency and equity are both essential: A generalized traffic signal controller with deep reinforcement learning," in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 5526–5533. [Online]. Available: http://ais.informatik.uni-freiburg.de/publications/papers/yan20iros.pdf

[5] M. Ferreira, R. Fernandes, H. Conceição, W. Viriyasitavat, and O. K. Tonguz, "Self-organized traffic control," in *Proc. of the Seventh ACM International Workshop on VehiculAr InterNETworking*, ser. VANET '10, New York, NY, USA, 2010, p. 85–90.

[6] T. Litman, "Autonomous vehicle implementation predictions: Implications for transport planning," https://www.vtpi.org/avip.pdf, 2021, [Online; accessed 06-Mar-2021].

[7] E. Namazi, J. Li, and C. Lu, "Intelligent intersection management systems considering autonomous vehicles: A systematic literature review," *IEEE Access*, vol. 7, pp. 91 946–91 965, 2019.

[8] S. Ulbrich, S. Grossjohann, C. Appelt, K. Homeier, J. Rieken, and M. Maurer, "Structuring cooperative behavior planning implementations for automated driving," in *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. IEEE, 2015, pp. 2159–2165.

[9] C. Menéndez-Romero, M. Sezer, F. Winkler, C. Dornhege, and W. Burgard, "Courtesy behavior for highly automated vehicles on highway interchanges," in *IEEE Intelligent Vehicles Symposium (IV)*, 2018, pp. 943–948. [Online]. Available: http://ais.informatik.uni-freiburg.de/publications/papers/menendez18iv.pdf

[10] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wiessner, "Microscopic traffic simulation using sumo," in *Proc. of the IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 2575–2582.

[11] 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.

[12] ——, "Multiagent traffic management: An improved intersection control mechanism," in *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, 2005, pp. 471–477.

[13] ——, "Sharing the road: Autonomous vehicles meet human drivers," in *The 20th International Joint Conference on Artificial Intelligence*, January 2007, pp. 1263–68.

[14] ——, "A multiagent approach to autonomous intersection management," *Journal of Artificial Intelligence Research*, vol. 31, pp. 591–656, March 2008.

[15] T.-C. Au, S. Zhang, and P. Stone, "Autonomous intersection management for semi-autonomous vehicles," *Handbook of transportation*, pp. 88–104, 2015.

[16] G. Sharon and P. Stone, "A protocol for mixed autonomous and human-operated vehicles at intersections," in *International Conference on Autonomous Agents and Multiagent Systems*. Springer, 2017, pp. 151–167.

[17] P. Lin, J. Liu, P. J. Jin, and B. Ran, "Autonomous vehicle-intersection coordination method in a connected vehicle environment," *IEEE Intelligent Transportation Systems Magazine*, vol. 9, no. 4, pp. 37–47, 2017.

[18] L. C. Bento, R. Parafita, S. Santos, and U. Nunes, "Intelligent traffic management at intersections: Legacy mode for vehicles not equipped with v2v and v2i communications," in *Proc. of the IEEE International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2013, pp. 726–731.

[19] Y. Meng, L. Li, F.-Y. Wang, K. Li, and Z. Li, "Analysis of cooperative driving strategies for nonsignalized intersections," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 4, pp. 2900–2911, 2017.

[20] X. Qian, J. Gregoire, F. Moutarde, and A. De La Fortelle, "Priority-based coordination of autonomous and legacy vehicles at intersection," in *Proc. of the IEEE International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2014, pp. 1166–1171.

[21] S. A. Fayazi and A. Vahidi, "Mixed-integer linear programming for optimal scheduling of autonomous vehicle intersection crossing," *IEEE Transactions on Intelligent Vehicles*, vol. 3, no. 3, pp. 287–299, 2018.

[22] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[23] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[24] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," in *Proc. of the International Conference on Learning Representations (ICLR)*, 2016.