robotics

MDPI

Article

Large-Scale Urban Traffic Management Using Zero-Shot Knowledge Transfer in Multi-Agent Reinforcement Learning for Intersection Patterns

Theodore Tranos [1,†], Christos Spatharis [1,2,†], Konstantinos Blekas [1,*,†] and Andreas-Giorgios Stafylopatis [3,†]

[1] Department of Computer Science & Engineering, University of Ioannina, 451 10 Ioannina, Greece; thtranos@cs.uoi.gr (T.T.); cspatharis@iit.demokritos.gr (C.S.)
[2] National Centre for Scientific Research 'Demokritos', 153 41 Agia Paraskevi, Greece
[3] School of Electrical & Computer Engineering, National Technical University of Athens, 157 73 Zografou, Greece; andreas@cs.ntua.gr
* Correspondence: kblekas@cs.uoi.gr
† These authors contributed equally to this work.

Abstract: The automatic control of vehicle traffic in large urban networks constitutes one of the most serious challenges to modern societies, with an impact on improving the quality of human life and saving energy and time. Intersections are a special traffic structure of pivotal importance as they accumulate a large number of vehicles that should be served in an optimal manner. Constructing intelligent models that manage to automatically coordinate and steer vehicles through intersections is a key point in the fragmentation of traffic control, offering active solutions through the flexibility of automatically adapting to a variety of traffic conditions. Responding to this call, this work aims to propose an integrated active solution of automatic traffic management. We introduce a multi-agent reinforcement learning framework that effectively models traffic flow at individual unsignalized intersections. It relies on a compact agent definition, a rich information state space, and a learning process characterized not only by depth and quality, but also by substantial degrees of freedom and variability. The resulting driving profiles are further transferred to larger road networks to integrate their individual elements and compose an effective automatic traffic control platform. Experiments are conducted on simulated road networks of variable complexity, demonstrating the potential of the proposed method.

Keywords: multi-agent reinforcement learning; traffic control; intersection patterns; knowledge transfer

check for updates

Citation: Tranos, T.; Spatharis, C.; Blekas, K.; Stafylopatis, A.-G. Large-Scale Urban Traffic Management Using Zero-Shot Knowledge Transfer in Multi-Agent Reinforcement Learning for Intersection Patterns. *Robotics* **2024**, *13*, 109. https://doi.org/10.3390/robotics13070109

Academic Editor: Sunan Huang

Received: 11 June 2024
Revised: 5 July 2024
Accepted: 15 July 2024
Published: 19 July 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Traffic congestion in urban areas is a major problem that significantly affects the daily rhythm of life within cities as transportation demands escalate. Congestion is mainly observed at intersections where conflicting vehicles with disparate trajectories converge within shared road spaces. Intersections constitute an important traffic component that plays a crucial role in the efficiency, safety, and energy consumption of the entire traffic system. Traffic at intersections poses challenges for autonomous driving vehicles and contains considerable dynamic characteristics and uncertainty. Therefore, traffic management at intersections is directly linked to the smooth operation of traffic within urban road networks.

With the significant advancements in vehicle-to-everything technologies that provide real-time information exchange, a promising prospect emerges: the potential for automated vehicles to seamlessly navigate through intersections cooperatively, without relying on traffic lights, i.e., "unsignalized intersections". However, ensuring driving safety while improving intersection traffic management efficiency remains a challenging issue, due to the existence of multifaceted factors, including real-time processing, system robustness, intricate scheduling, and inherent complexity of the task. These issues demand comprehensive

solutions for effective resolution. Furthermore, autonomous driving within unsignalized intersections generates complex environmental state spaces, introducing profound challenges to the scalability of the scheduling algorithms. This requires the design of scalable learning strategies to guarantee driving safety, while concurrently enhancing traffic efficiency for autonomous vehicles [1,2]. Moreover, an important challenge is to manage cooperative driving to address traffic conflicts [3,4].

Machine learning (ML) and especially reinforcement learning (RL) [5] constitutes a framework with tools as leverage for constructing intelligent decision-making algorithms in an attempt to provide a more efficient, comfortable and accident-free traffic system. Several attempts have been made towards providing the use of RL algorithms for intelligent transportation systems in order to generate optimal driving profiles of automated vehicles with less energy consumption and fewer emissions, including approaches to regulate the traffic around an intersection [6].

Agent-based methodologies encompass two essential objectives: (a) providing robustness against environmental variations, and (b) fostering the ability to transfer the acquired knowledge and skills from a specific environment or scenario to new and diverse ones, without necessitating significant modifications or re-training.

Nowadays, in RL agent research, the concept of knowledge transfer has gained much traction [7]. In direct knowledge transfer, distinct models are trained independently for each source task. Subsequently, their policies are directly applied and evaluated on other similar tasks within the broader system. This approach can be particularly effective when the source tasks share substantial similarities with the target tasks. By leveraging pre-trained models on specific tasks, the general system can quickly benefit from the knowledge gained during the individual training processes.

Motivated by these challenging issues, this study introduces a holistic approach to traffic management in complex road networks that unfolds in two stages. Initially, autonomous driving strategies are constructed for specific intersection patterns based on a novel multi-agent deep reinforcement learning scheme. Then, a direct knowledge transfer framework is established that enforces the existing pre-trained driving policies at the intersections of the studied road network. The proposed methodology promotes robust traffic management in diverse scenarios and network structures, and thus naturally facilitates scalability over large road networks and complex environments with a large degree of variability.

In a recent work of ours [8], we proposed a collaborative multi-agent reinforcement learning approach for addressing the traffic control problem, by treating the potential routes taken by vehicles as the agents. This approach offered the flexibility, from both modeling and learning perspectives, to design an economical solution for multi-agent schemes by alleviating increased complexity and constraints of common practices that treat each vehicle as a distinct agent. Nonetheless, the complexity issue remains a substantial obstacle, particularly when handling large-scale networks, given the exponential expansion of available routes that corresponds to the network's size. In addition, the transferability of the acquired knowledge is limited to the level of the respective reference road network and thus the exploitation of the generalization ability is significantly minimized.

Conversely, in this study, the traffic control problem is framed as a multi-tasking problem. It is deconstructed into smaller traffic control sub-problems centered around intersection patterns. In these patterns, autonomous vehicles are trained to collaborate and navigate, while adhering to acceleration policies. This approach fosters straightforward adaptability and transferability.

The main contributions of this work are summarized as follows:

- We introduce an efficient multi-agent reinforcement learning scheme for modeling the traffic flow management at intersections. We consider the incoming roads to the intersections as agents, effectively mitigating the escalating complexity and constraints associated with conventional methods that focus on modeling agents at the vehicle level. Another novelty is the introduction of a priority feature within the state space

to ensure seamless coordination among agents at intersections by incorporating this centralized vision, we aim to enhance the overall efficiency and safety of traffic flow by offering active solutions that can dynamically respond to changing traffic conditions.
- We aim to provide a compact strategy for urban traffic congestion by dividing the road network into homogeneous structures based on intersection patterns. This offers the advantage of independence from traffic fluctuations and immunity to changes of traffic density. Moreover, partitioning the road network enhances coordination among traffic patterns at a higher level, leading to improved traffic flow and a more consistent and reliable traffic management system.
- We promote an advantageous zero-shot knowledge transfer process that associates intersections in road networks with learned multi-agent policies. Our approach creates adaptable driving profiles for existing intersection patterns with exceptional generalization capabilities, facilitating seamless deployment of a traffic management system to address urban congestion effectively and rapidly.

The structure of this manuscript is as follows. Section 2 summarizes state-of-the-art reinforcement learning and knowledge transfer strategies for autonomous traffic control. Then, Section 3 specifies the problem to be solved and describes analytically all stages of the proposed method. Section 4 presents the simulation cases based on the SUMO environment and illustrates the depicted results for measuring the performance of our method. Finally, Section 5 concludes the paper and discusses future directions.

2. Related Work

Ongoing research in autonomous vehicle control and planning explores the problem from various perspectives, including rule-based, reinforcement learning, and imitation learning schemes. The majority of the studies focus on employing traffic lights as a means to model the agents, with the objective to minimize waiting time [9–14]. Nevertheless, there has been a notable surge in research interest within the field of autonomous vehicle navigation. Optimum control of smooth and safe driving at intersections is an important element of traffic that drastically affects the efficiency, safety, and energy consumption of all traffic [2]. These emerging approaches involve considering the vehicles as the agents within the multi-agent system, shifting the focus away from traffic lights.

The most significant rule-based approach in autonomous vehicle control is the time-to-collision (TTC) method [15]. Other commonly used approaches involve trajectory optimization-based methods [16–19], although hand-crafted heuristics can be extremely difficult and often lead to suboptimal results. Another solution is given by Carlino et al. [20], where the authors propose running auctions at each intersection to decide the sequence in which drivers execute conflicting moves. In Wang et al. [21], there is an intersection access management method that optimizes vehicle crossing sequences using a cooperative intersection control strategy and a high-level hybrid queuing model. Furthermore, Levin and Rey [22] developed a Mixed Integer Linear Programming (MILP) model to determine the optimal coordination strategies for a single intersection. On the other hand, Li et al. [23] frame the problem of coordinating multiple vehicles as a Multi-Agent Path Finding (MAPF) problem. Finally, Lu et al. [24] proposed a fail-safe buffer area for intelligent intersections, enabling the system to handle emergencies effectively and allowing vehicles to brake in time. In contrast to the previous works, our approach deals with large-scale traffic networks and complex scenarios containing multiple intersections and vehicles, while also being less costly to implement.

Studies involving imitation learning, such as behavior cloning (BC) [25], employ supervised learning approaches on expert data. However, they face challenges such as compounding errors and distribution shifts. An online interactive data collection method has been developed in Menda et al. [26], which uses an online demonstrator during training to act as a teacher. Nevertheless, the requirement of an expert adds extra complexity to the training process.

On the other hand, deep RL schemes have also gained attention recently in the field of traffic control. In Bouton et al. [27], Monte Carlo was employed to optimize decision-making at unsignalized intersections. Similarly, deep recurrent Q-networks (DRQNs) are used by Tram et al. [28] to train a vehicle to pass through an intersection, by adjusting its acceleration through a negotiation phase with other human-controlled vehicles. In a subsequent work by the same authors [29], a hierarchical framework encompassing RL-based high-level decisions and a lower-level controller was proposed. In Isele et al. [30], the intersection traversal problem was studied using a deep Q-network (DQN) and then was extended to include situations with occluded intersections [31]. Moreover, in Li and Czarnecki [32], a thresholded lexicographic Q-learning variant was designed for collision avoidance in complex scenarios involving multi-lane roads and intersections. A recent work, presented by Shao et al. [33], addresses collaborative decision-making for autonomous vehicles in unsignalized intersections using a MARL algorithm based on PPO, incorporating parameter sharing, local rewards, and an action masking mechanism. While quite similar, our work operates on much larger and more densely trafficked scenarios, and demonstrates effective knowledge reuse capabilities. Finally, in our recent study [8], we explored navigating multiple vehicles in unsignalized traffic networks, treating the possible routes as intelligent agents of the multi-agent system.

A significant drawback of RL methods is their prolonged training periods, as RL agents lack awareness of training environment dynamics. Transfer learning (TL) involves pre-training a network on a different task before fine-tuning it on the target task [7]. By leveraging the skills acquired during pre-training, the network can benefit and potentially overcome the challenges of RL training. Furthermore, TL allows reusing learned agent knowledge on more intricate traffic scenarios than those encountered in initial training. In Akhauri et al. [34], a model trained on simulated data was transferred to train another model using real-world data. Additionally, in Chiba and Sasaoka [35], pre-trained model vehicles are transferred and further trained in modeling autonomous driving. Finally, in Shu et al. [36] a knowledge transfer deep RL method is presented, where a learned single-agent intersection-traversing policy is applied to three similar driving tasks.

In the recent bibliography, there are some other alternative transfer learning schemes. Zero-shot learning refers to the ability of agents to perform tasks they have never encountered before, relying on learned generalizations from related tasks for autonomous vehicles [37,38]. A similar scheme is curriculum learning, which involves training an agent by gradually increasing the complexity of tasks, used to learn driving behavior at 4-way intersections [39] or create an end-to-end driving policy in combination with deep RL [40]. Finally, other schemes from federated RL perspectives have been presented, where multiple distributed agents collaboratively learn and share knowledge while training on local data [41–43].

Lastly, building upon single-agent RL, multi-agent reinforcement learning (MARL) produces agents that not only optimize their individual policies but also collaborate and share knowledge, enabling the emergence of improved solutions [44,45]. However, in comparison with single-agent settings, research on the generalization domain of MARL is either of increased complexity [46] or is focused on transferring policies learned in simulations to real-world scenarios [47]. Specifically, Candela et al. [47] created a simulated multi-agent environment that imitates the dynamics of a real-world robotic track, and successfully transferred the obtained policies of the former to the vehicles of the latter. Finally, in Jang et al. [48], a pair of autonomous vehicles are initially trained using deep RL to guide a fleet onto a roundabout, and then the authors directly transfer the policy from a simulated to a real city environment, encompassing policies with and without injected noise for training.

3. Proposed Method

3.1. Overall Description of the Method

In this study, we address the problem of navigating multiple vehicles in road traffic networks containing unsignalized intersections using multi-agent reinforcement learning

policies and knowledge transfer capabilities. The key idea of our approach lies in perceiving the road network as a dynamic environment characterized by an intricate interlinking of multiple roads through unsignalized intersections. Central to our analysis are intersections, which represent complex and heavily trafficked zones characterized by recurring patterns and phenomena, requiring traffic control strategies and active solutions to ensure the secure and efficient passage of vehicles so as to prevent collisions. In our traffic scenarios, vehicles seamlessly integrate into the traffic network to be served automatically, subject to a stochastic entry rate.

According to our scenario, each vehicle follows a pre-defined route towards its destination (in real-time scenarios, vehicles' paths can be automatically designed through a route-generator mechanism based on the current traffic or environmental conditions).

Thus, every route can be seen as a sequence of intersections that the vehicles visit and pass through. Based on this framework, we can deconstruct the broader task of vehicle movement into distinct movements between intersections, which we define as sub-tasks. This provides us with two notable advantages:

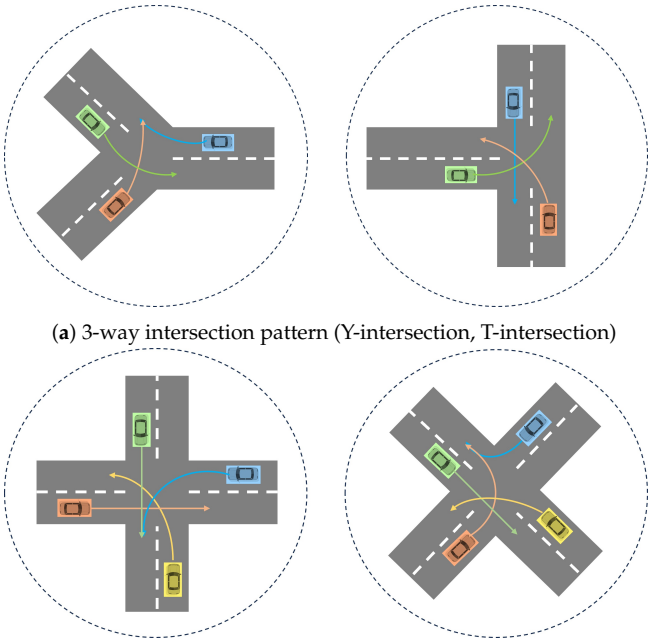
1. Establishment of a well-organized and compact system for managing traffic flow and optimizing transportation networks.
2. Introduction of traffic and congestion patterns for modeling specific types of intersections.

By concentrating on movements between intersections, we can gain valuable insights into how traffic behaves at various types of intersections. This knowledge has the advantage of enhancing our ability to model and simulate large-scale traffic networks more accurately.

This holistic approach encompasses the development of traffic patterns on intersections that cover various configurations, each possessing unique characteristics. These patterns can include, for example, Y-intersections, T-intersections, crossroads, X-intersections, roundabouts, and more, as shown in Figure 1a. More specifically, in this work we considered two general intersection patterns: 3-way and 4-way intersections. In 3-way intersections, three roads intersect at a single point. This includes the "Y-intersection", where two roadways join to form a single road, and the "T-intersection", where a minor roadway meets a major roadway, as shown in Figure 1. Despite their differences in topology and geometry (road lengths), our scheme treats these cases equivalently and manages them similarly though a unified general-purpose 3-way intersection pattern. A 4-way intersection is the most common traffic type that involves the crossing of two roadways. This pattern used in our analysis covers not only cases where the roads appear perpendicular, but also various design configurations where roads can intersect at any angle. Two such examples are shown in Figure 1b.

After introducing intersection patterns, the subsequent stage involves implementing an intelligent decision support system tailored to each intersection's unique characteristics and peculiarities. We propose a multi-agent reinforcement learning (MARL) framework to facilitate smooth and secure vehicle passage through intersections and develop driving profiles. Emphasis is placed on robust training of the generated MARL policies, ensuring adaptability to various topological structures within the same intersection pattern. This enables the policies to effectively accommodate alternative configurations, enhancing their versatility and applicability in real-world scenarios. Moreover, the MARL scheme offers active solutions in traffic management by generating adaptive driving policies that can dynamically respond to changing traffic conditions.

Finally, we propose a zero-shot knowledge transfer process designed to facilitate the reuse of acquired knowledge across multiple intersections within the road network. Learned MARL policies are directly applied to similar intersections, streamlining traffic management and minimizing the need for extensive adaptation. Figure 2 depicts the overall structure of the proposed method for traffic flow management in urban areas with multiple unsignalized intersections.



(**a**) 3-way intersection pattern (Y-intersection, T-intersection)

(**b**) 4-way intersection pattern (crossroad, X-intersection)

Figure 1. Examples of different configurations covered by the 3-way and 4-way intersection patterns.

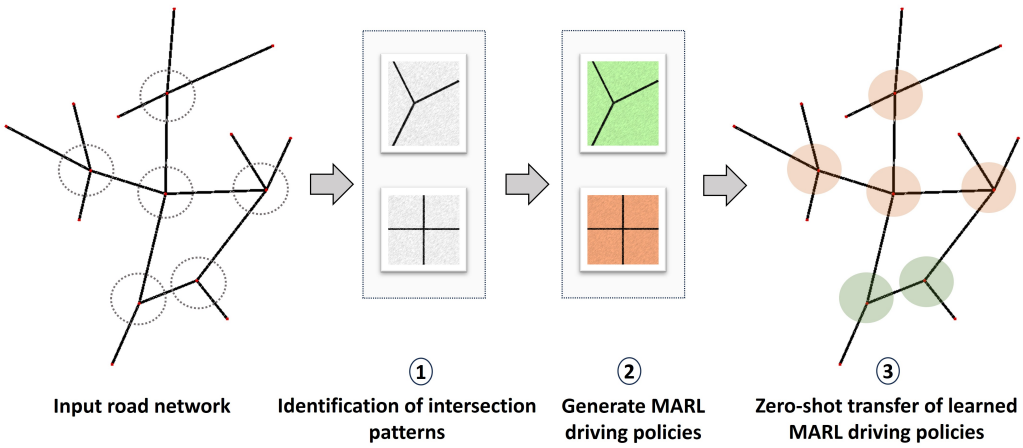


Figure 2. Overview structure of the proposed method for traffic control in road networks consisting of three major modules.

3.2. Multi-Agent Reinforcement Learning for Traffic Management in Intersection Patterns

In our study, each road coming into the intersection pattern is treated as an individual agent, called a road-agent. The goal of the road-agent is to temporarily guide a vehicle along its designated road segment until it reaches the next intersection. Once a vehicle has traversed the intersection and moved onto the outgoing road, it is no longer within the scope of this agent and, by extension, the MARL system of the pattern to which that intersection belongs. This is due to the assumption that the vehicle has successfully reached its local destination. Subsequently, control of the vehicle is handed over to the multi-agent road-agent of the next intersection (and the matching intersection pattern that may be similar to the previous one), who drives it to the next intersection. This process repeats until the vehicle reaches its final destination. Notably, when the vehicle traverses the last intersection of its route, a deterministic rule is applied: maximum acceleration is used to expedite the vehicle's exit from the road network via the final road segment.

3.2.1. MDP Formulation

Reinforcement learning has demonstrated significant potential in addressing challenges related to decision making and control. A fundamental issue in reinforcement learning is the concept of the Markov Decision Process (MDP) [5,49], which defines a mathematical framework that represents decision-making scenarios involving intelligent agents. As the complexity of decision-making scenarios increases, containing multiple interacting entities, the multi-agent Markov Decision Process (MA-MDP) emerges as a pivotal framework. Specifically, it extends the MDP to accommodate scenarios where multiple intelligent agents co-act within a shared environment, each possessing distinct objectives and actions.

In our study, we formulate the task of intersection traffic management as an MA-MDP that can be represented as a tuple $< Ag, S, A, T, R, \gamma >$, where:

- Ag is a society of agents $i = 1 \ldots N$;
- S denotes the state space of the environment. Each agent, i, observes its local state, s_i;
- A denotes the set of possible actions of any agent. Each agent, i, takes a local action, a_i;
- $P(s_i'|s_i, a_i) : S \times A \times S \rightarrow [0, 1]$ specifies the probability of the agent, i, to transition to a new state, s_i';
- $R(s_i, a_i)$ is the reward function for a state action pair $R : S \times A \rightarrow \mathbb{R}$;
- $\gamma \in (0, 1)$ is the discount factor across time for future rewards.

Under the MA-MDP formulation, the agent adheres to a policy within a given environment. A policy in RL is a strategy that determines the actions of an agent based on its current state. Let us consider a vehicle following its corresponding road-agent, $i \in Ag$. At each time step, this vehicle is found in a specific state, denoted as $s_i \in S$, and it chooses an action, $a_i \in A$, according to the policy, π_i, implemented by its corresponding road-agent. Subsequently, the vehicle transitions to a new state, s_i', and yields a reward, $R(s_i, a_i) \in \mathbb{R}$. Figure 3 illustrates an example of how the proposed MARL scheme is applied to a 4-way intersection pattern (crossroad) for modeling traffic control.

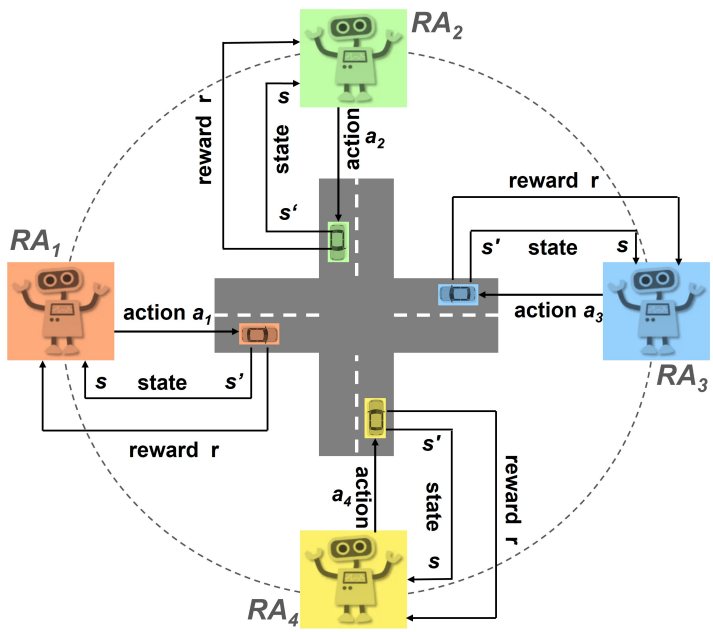


Figure 3. Traffic control in a 4-way intersection pattern using the proposed MARL scheme. Every road-agent (RA_i) is responsible for safely guiding vehicles through its designated road segment, while cooperatively coordinating with the other agents (in our case three).

Value functions are essential components in RL, providing estimates of the expected cumulative rewards associated with different states and actions. Specifically, the state-value function is defined as the expected cumulative discounted future reward at state s_i:

$$V^{\pi_i}(s_i) = \mathbb{E}_{\pi_i}\left[R(s_i, a_i) + \gamma V^{\pi_i}(s_i')\right] . \tag{1}$$

Moreover, the action-value function, or Q-function, is the expected cumulative reward starting from state s_i, executing action a_i and following policy π_i thereafter:

$$Q^{\pi_i}(s_i, a_i) = \mathbb{E}_{\pi_i}\left[R(s_i, a_i) + \gamma Q^{\pi_i}(s_i', , a_i')\right] \tag{2}$$

In actor-critic methods [5], the agent is composed of the actor and the critic components. The actor is responsible for learning the policy, while the critic approximates the value function. A fundamental element is the advantage function, which measures the advantage of taking a particular action in a specific state compared with the average action value in that state according to the policy. Advantage is defined as the difference between the *action-value* and the *state-value* function under policy π_i:

$$A^{\pi_i}(s_i, a_i) = Q^{\pi_i}(s_i, a_i) - V^{\pi_i}(s_i) \tag{3}$$

In our study, we employed the proximal policy optimization (PPO) [50] method to learn the policy and the value function simultaneously. PPO is a state-of-the-art on-policy actor-critic algorithm, which operates by alternating between data sampling through interactions with the environment and optimizing a clipped surrogate objective function, using stochastic gradient descent. The clipped surrogate objective function enhances training stability by constraining the magnitude of policy changes at each step. In PPO, the clipped surrogate objective function is formulated as follows:

$$L^{CLIP}(\theta_i) = \mathbb{E}\left[\min\left(\rho(\theta_i)A^{\pi_{\theta_i}}(s_i, a_i), \text{clip}(\rho(\theta_i), 1-\epsilon, 1+\epsilon)A^{\pi_{\theta_i}}(s_i, a_i)\right)\right] \tag{4}$$

$$\rho(\theta_i) = \frac{\pi_{\theta_i}(a_i|s_i)}{\pi_{\theta_i}(old)(a_i|s_i)}, \tag{5}$$

where ${\pi_{\theta_i}}^{(old)}$ represents the policy before the update, $\pi_{\theta_i}(a_i|s_i)$ is the probability of taking action a_i in state s_i under current policy π_{θ_i}, and θ_i are the weights of the policy's neural network. The advantage function is calculated using the generalized advantage estimator (GAE).

It is worth noting that simultaneously managing multiple instances of the same road-agent (vehicles on the same road) provides additional flexibility to the learning process. This has the advantage of enhancing its generalization capability, without compromising the decision-making effectiveness of each individual agent. In what follows, we outline all the essential elements that constitute the definition of the MDP specific to our problem.

3.2.2. State and Action Representations

Following our proposed profiling scheme, each vehicle undergoes a process in which a particular road-agent is designated to observe its driving behavior and guide it through the intersection. This approach ensures that each vehicle is appropriately managed and supervised by the relevant road-agent, facilitating smooth and efficient traffic flow within the intersection.

The proposed state space for road-agents in an intersection pattern is designed to be both cost-efficient and easily manageable. This design facilitates efficient knowledge transfer across diverse intersections (instances) of the same pattern. More specifically, the state is described by the 5-dimensional vector:

$$s = [v,\ pos,\ fveh_v,\ fveh_dist,\ pindex] , \tag{6}$$

comprising four continuous values capturing the velocity and position of the examined vehicle (called *ego* vehicle) and its front vehicle, and a priority term that takes discrete values. Specifically, it consists of the following quantities:

- v: normalized velocity (in m/s) of the ego vehicle, where we consider a maximum allowed velocity, v_{max} (in our simulation studies, we set $v_{max} = 20\ m/s$);
- pos: normalized relative position (in m) of the ego vehicle on the road it traverses with respect to the road's length;
- $fveh_v$: normalized velocity of the *front* vehicle. We consider the *front* vehicle to be the closest vehicle in front of the ego vehicle, within a visible range (in our experiments, we consider a visible range of 100 m). If there is no vehicle within this range, this variable is set to -1;
- $fveh_dist$: normalized distance from the *front* vehicle. It is calculated as the real distance between two vehicles divided by the visible range. Similarly, in the absence of a *leading* vehicle, this variable is set to -1;
- $pindex$: priority term that takes three possible values (-1, 0, and 1). This is a centralized feature that acts as a factor of coordination and cooperation among road-agents, and is calculated based on the distances of incoming vehicles to the intersection's center. To underscore the significance and meaningful contribution of this variable, we extended the designated area surrounding the intersection so as to guarantee the utmost precision and reliability in identifying the exact position of each vehicle. This enhances the agent's ability to implement proactive measures and prevent potential collisions. Vehicles that are currently inside the intersection area will have their $pindex$ values set to -1. All other vehicles will have their $pindex$ values set to 0. The vehicle closest to the intersection's center will be assigned a value of $pindex = -1$, indicating it has the highest priority to cross.

As already mentioned, the road-agents are responsible for controlling the acceleration of their supported vehicles to ensure safe traversing of the intersection. Consequently, the vehicles' acceleration (continuous value) is used as the agent's action, and the task of autonomous driving involves applying a sequence of acceleration values to cross through the intersection in an optimal manner. (Acceleration is bounded from a negative value (break pedal) to a positive value (gas pedal).)

3.2.3. Reward Function

The reward function plays a crucial role in shaping the behavior of vehicles as they pass through the intersection. It is designed and finely tuned to promote correct and safe actions, while providing an evaluation of their effectiveness based on the current vehicle's state.

To calculate the reward function, after executing the agent's action, we initially examine whether or not the vehicle reached a terminal state, i.e., either successfully left the intersection or collided with other vehicles. In this case, the agent receives a positive reward relative to the current vehicle's velocity (to promote higher values), or a constant negative reward, respectively. This is shown in the following equation:

$$R(s,a) = \begin{cases} v, & \text{if successfully leaves intersection} \\ -L, & \text{if collides} \\ r(s,a), & \text{otherwise} \end{cases}, \tag{7}$$

where v_{max} indicates the maximum allowed velocity, and the threshold value, L, is set to $L = 5$ (experiments did not show strong sensitivity to its choice).

In case of no terminal states, the reward function, $r(s,a)$, is shaped accordingly, mainly to the vehicle's priority indicator ($pindex$) value. There are three possible cases:

- The vehicle is currently found inside the intersection sector ($pindex = -1$). Then,

- if its previous state was of similar ($pindex^{(old)} = -1$) or highest priority ($pindex^{(old)} = 1$), then a positive reward would have been received to encourage the maintenance of its priority status:

$$r(s,a) = 1 + v \tag{8}$$

- if it forcefully finds itself inside the intersection ($pindex^{(old)} = 0$), a penalty will be given as a deterrent against "stealing" priority:

$$r(s,a) = -(1 + v) \tag{9}$$

- The vehicle currently takes maximum priority ($pindex = 1$) and prepares to cross the intersection. Then, it is positively rewarded in a manner proportional to its normalized velocity:

$$r(s,a) = v \tag{10}$$

- The vehicle currently does not have any priority ($pindex = 0$). The reward depends on the distance to the vehicle in front ($fveh_dist$), and whether it exceeds a safety threshold value ($safe_dist$):

$$r(s,a) = \begin{cases} v, & \text{if } fveh_dist > safe_dist \\ -\frac{safe_dist}{min(fveh_dist, safe_dist)} * v, & \text{otherwise} \end{cases} \tag{11}$$

Overall, the proposed reward function facilitates the development of efficient and safe policies for the agents. However, its ultimate purpose is to promote priority for the closest vehicles by automatically determining the appropriate priority rules for the vehicles to cross the intersection.

3.2.4. Training

The proposed multi-agent methodology establishes an efficient collaborative framework that focuses on two key features:

- The road-agents that enable the simultaneous control of multiple vehicles in the same road in order to construct optimal intersection traversing driving policies.
- The priority term that promotes coordination among vehicles from different road-agents, aiming to prevent collisions inside the intersection.

The overall training process is conducted independently for each road-agent within the dedicated multi-agent system of the intersection patterns and is based on the PPO algorithm. At the beginning, the road-agents initialize their policy and value networks, and their memory buffers to store the generated vehicles' trajectories. According to the traffic scenario, the vehicles sequentially enter into the road network (intersection pattern) from the "entrance lanes", where they are placed at a random position in order to safely traverse the intersection. We initially assign to them a random velocity, and the corresponding road-agent controls their acceleration until they cross the intersection. At each time step, the road-agent takes an action (acceleration) based on its current policy, and then a low-level controller is activated that handles the transition from the current to the desired velocity. At the end of this procedure, the neural networks' parameters are updated using the clipped objective function of PPO.

A noteworthy aspect to highlight is that the training process employs a dynamic approach, wherein the traffic scenario is modified at each episode of the training cycle. This dynamic scenario adaptation plays a pivotal role in enhancing the system's adaptability and generalization capabilities. Indeed, by subjecting the multi-agent system to a diverse range of scenarios during each training episode, we expose it to a rich variety of traffic situations and challenges. This exposure allows the system to learn robust and flexible driving strategies, ensuring that it can effectively handle a wide array of unknown scenarios

that may vary in complexity, uncertainty, and novelty. Algorithm 1 summarizes the overall training scheme.

Algorithm 1 Multi-agent PPO for traffic control in an intersection pattern.

Initialize policy $\pi_{\theta_i}(a|s)$ and value $V_{\phi_i}(s)$ networks with parameters θ_i and ϕ_i for each road-agent
Initialize empty replay buffer D_i for each road-agent
for $k = 1 \ldots$ *max episodes* **do**
  Create random traffic scenario with variable vehicles' arrival rate
  Match incoming vehicles to road-agents i
  Generate trajectories using road-agents' policies $\pi_{\theta_i}(a_i|s_i)$ and store them to the corresponding buffers D_i
  Update θ_i and ϕ_i using mini batches of trajectories sampled from buffers D_i using PPO
end for
Store the learned agents' policies to a driving profile for this intersection pattern

It must be noted that, after experimentation, we chose PPO for the control of multiple vehicles at intersections over other actor-critic algorithms like Deep Deterministic Policy Gradient (DDPG), Lillicrap et al. [51]; Soft Actor-Critic (SAC), Haarnoja et al. [52,53]; and Twin Delayed Deep Deterministic Policy Gradient (TD3), Fujimoto et al. [54]. We opted for that algorithm due to its performance and stability in complex and dynamic environments, where the state and action spaces are continuous. PPO prevents the policy from making large updates, which can lead to unstable driving behaviors, and furthermore it is easier to tune compared with the others, which are sensitive to hyperparameters and require careful balancing. Finally, PPO's on-policy nature ensures that the policy is consistently updated in a direction that improves performance according to the most recent data, which is crucial when expert data are not involved in training.

3.3. Zero-Shot Transfer of Learned MARL Policies for Automated Traffic Control

After training the MARL intersection patterns and constructing their driving policies, we can then transfer them to road networks in order to achieve their traffic control management. Our primary objective is to directly apply the acquired knowledge via policies without the need for any further fine-tuning or adjustments, specific to the configurations of the network's intersections. This procedure not only saves time and computational resources, but also ensures a smooth integration of the trained road-agents into the existing road network.

A significant step in the knowledge transfer process is addressing the challenge of aligning arbitrary intersections in the road network with the default intersection patterns used during the learning phase. This alignment is crucial to ensure seamless integration and compatibility between existing intersections and the learned patterns. In our case, the matching process is considerably simplified and focuses on the correct assignment of roads. Each intersection then fully adopts the learned MARL policy of the corresponding intersection pattern without any further adaptation, ensuring smooth integration of the learned patterns into the existing traffic network. It is important to note that other factors, such as the road length, geometry, and the traffic volume or flow, are not considered, as they do not influence the learned policy.

Figure 4 illustrates an example of how a road network, consisting of four "4-way intersections" and two "3-way intersections", inherits the driving policies of two default intersection patterns. The procedure followed by each vehicle to complete a route (e.g., light-colored thin strip in the road network of Figure 4) within the road network is as follows. At the beginning of its navigation, the MARL policy associated with the first visited intersection (a 4-way pattern) is activated to guide the vehicle through smoothly and securely. Upon exiting the intersection, the vehicle transitions to the outgoing lanes and continues its journey towards its destination, passing through a sequence of five inter-

sections. As the vehicle enters the field of view of another intersection, it comes under the control of a new multi-agent system, which governs its path through this intersection. The new policy may resemble those applied in previous interactions, with similar characteristics that the vehicle has encountered (e.g., the first two intersections of the route in Figure 4).

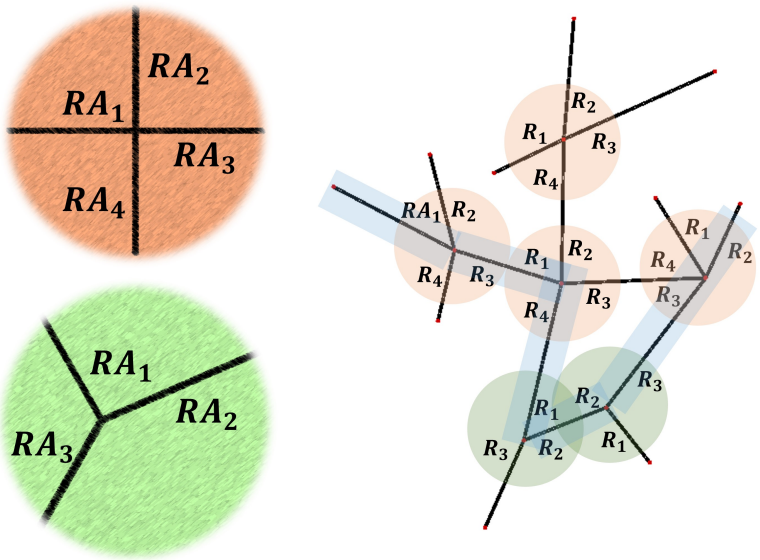


Figure 4. Matching the network's intersections with two default intersection patterns. The road network contains four and two copies of the default "4-way" and "3-way" intersection patterns, respectively. The light-colored thin strip corresponds to a route that a vehicle may follow within this network.

This handover process, from one multi-agent system to another, ensures a smooth and continuous traffic flow, allowing vehicles to smoothly navigate through various intersections until they reach their final destinations. Also, it facilitates scalability, since the learned policies can be efficiently transferred to any other traffic network, fostering in such a way their consistency across multiple occasions. Thus, the proposed transfer process optimizes traffic management while accommodating the dynamic nature of vehicle movement within a network of interconnected unsignalized intersections. By allowing intersections to employ policies that align with their specific characteristics and traffic demands, we can enhance the efficiency and safety of autonomous driving systems in urban environments.

4. Experimental Results

Our experimental study is guided by three primary objectives:

- Efficiently solving individual intersection patterns: the proposed multi-agent reinforcement learning platform aims to demonstrate its capability in effectively resolving congested situations in different intersection patterns.
- Optimizing the learning strategy for intersection patterns' traffic management: we seek to develop an efficient learning strategy that enables the developed multi-agent systems to acquire a diverse set of generalized capabilities. These capabilities should allow the use of their learned driving policies directly and effectively across various scenarios and traffic network configurations.
- Ensuring efficient knowledge transfer: one of our key goals is to establish a robust and scalable knowledge transfer process that enables its application to complex road network structures and traffic scenarios satisfactorily.

By addressing these questions, we aim to establish the effectiveness and practicality of our proposed platform in solving intersection-related challenges across various traffic network scenarios.

4.1. Simulation Environment

We chose SUMO (Simulation of Urban MObility) [55] as the simulation package to generate a range of artificial scenarios and carry out our experimental analysis. SUMO is an open-source package designed for microscopic urban traffic simulation, with the ability to manage extensive road networks featuring multiple intersections and vehicles. Traffic scenarios in SUMO can be configured by defining the road network, along with several parameters that govern traffic conditions, such as the number of intersections, roads, road length, maximum velocity, arrival rates, and traffic density. In our experiments, we employed vehicles with common characteristics regarding the maximum velocity and the vehicle length. However, we introduced variability in the arrival rate to simulate real traffic conditions with different level of stochasticity.

The Krauss model [56] serves as the default car-following model within SUMO and was adopted as a baseline in our comparative study. This model introduces a microscopic traffic flow framework that depicts how vehicles engage with one another while navigating through a traffic flow. It operates on the principle that drivers adjust their velocity and maintain a specific distance from the vehicle ahead, taking into account the disparities between their desired velocity and the velocity of the leading vehicle.

4.2. Experimental Design and Implementation Issues

The proposed approach was applied on two intersection patterns representing the fundamental structure of 4-way and 3-way intersections (see left side of Figure 4). All roads connected to these intersections were of the same length (200 m) and comprised a single lane in each direction. For each intersection pattern, we trained five distinct MARL policies through independent runs of the proposed multi-agent PPO algorithmic scheme (Algorithm 1), maintaining the same structure of the intersections across all runs.

The generation of traffic scenarios followed the next setting. All vehicles enter sequentially onto one of the "entrance lanes", assuming a random arrival rate spanning from 1 to 6 s. Initially they are located at a random position within a range of $[0, 20]$ m from the start of the lane. We set the maximum allowed velocity as $v_{max} = 20$ m/s, while all engaged vehicles started navigating with a random initial velocity of $[10, 20]$ m/s. The continuous range of action values (acceleration or deceleration) was $[-5, 3]$ m/s^2. Lastly, a time step of 1 s was applied in all cases.

For the learning process of the multi-agent RL system, we followed the next configuration. The actor and critic networks of each road-agent were both deep neural networks consisting of two (2) hidden layers of size 256×128 with ReLU neurons. We used the Adam optimizer for the adjustment of their weights (parameters). In each simulated run, training consisted of a cycle of 2500 epochs. The duration of each (randomly generated) scenario within every epoch was approximately 900 time steps (15 min), where the total number of served vehicles approached 280. Furthermore, the learning rate was set to 10^{-5}, and a weight decay regularizer was applied with a coefficient of 10^{-8}. Additionally, the discount factor, γ, was set to 0.99, and the clip factor of PPO was set to 0.1. Finally, a mini batch size of 250 was utilized, and a standard deviation of 0.3, which was gradually reduced, was incorporated to control the degree of output variability in the policy's continuous actions.

The effectiveness of the proposed method was evaluated based on its capacity to efficiently manage traffic within the intersection, ensuring the safe and optimal navigation of all incoming vehicles within the scenario. Three evaluation metrics were considered (we decided not using the traffic throughput as an evaluation criterion since it can be easily derived by the number of collisions):

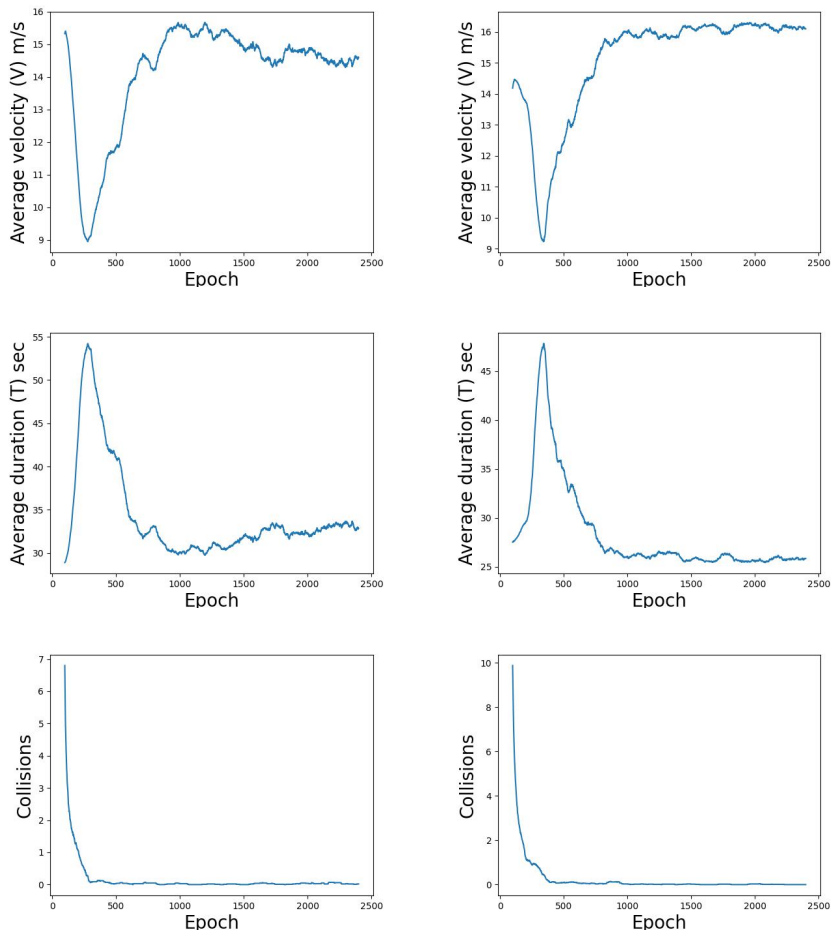
- Average velocity (V) (in m/s) of all vehicles across their routes;
- Average duration (T) (in s) of completion of the vehicle routes;
- Number of collisions during the execution of the scenario.

Finally, all reported results are the mean values of the above quantities.

4.3. Results

4.3.1. Performance of Multi-Agent RL for Traffic Control on Intersection Patterns

At first, we examined the performance of the proposed multi-agent RL framework for the traffic control management of a single intersection. Figure 5 illustrates the learning curves received during the training process for two intersection patterns in terms of three (3) evaluation criteria. As it can be observed, initially the average velocity constantly decreases, reaching a lower value in the range of $[9, 10]$ m/s. This corresponds to an almost safe policy (approximately zero collisions). However, as the training progresses, the method stabilizes to higher values in the range of $[14, 15]$ m/s for the 3-way intersection pattern and $[16, 17]$ m/s for the 4-way intersection pattern. This discovered policy covers both safeness and efficiency in traffic management. Similar observations concern the metric of average duration (see Figure 5).



(**a**) 3-way intersection pattern (**b**) 4-way intersection pattern

Figure 5. Learning curves of the 3-way and 4-way intersection patterns in terms of the average velocity, duration, and collisions per epoch, created by using a rolling window of fifty (50) episodes.

The reason why we do not observe solid convergence in the learning curves is found in the learning process. Specifically, as it was mentioned previously, in each training epoch the model is fed with a new randomly generated scenario of the vehicle traffic flow, through a stochastic generator. As it was expected, this strategy slows down the learning process and convergence, but on the other hand it significantly expands the space of situations and enriches the model with great generalization capabilities.

Additional experiments were conducted by evaluating the performance and the scalability of the proposed method to larger- scale scenarios. In particular, we created ten (10) independent traffic scenarios for both intersection patterns of 30 min (1800 s) du-

ration that consisted of approximately 500 vehicles. We used the same arrival rate of the trainable scenarios. In Table 1, we present the evaluation results (average velocity and duration) of all learned driving policies in these scenarios (total of $5 \times 10 = 50$ independent simulation runs for every one of two available intersection patterns). It is interesting to observe that the performance of the learned policies in these evaluation cases did not significantly decline from their corresponding performance during training, despite the variation in the duration of the scenarios that caused more intense congestion conditions.

Table 1. Mean performance of the five learned 3-way and 4-way intersection patterns' driving profiles to 10 test simulated half-hour scenarios. The best performance found is also shown.

Environment	*Average Velocity (m/s)*	*Average Duration (s)*
3-way intersection (best)	13.0 ± 0.5 (13.6)	33.1 ± 1.6 (32.0)
4-way intersection (best)	15.4 ± 0.3 (15.9)	27.1 ± 0.8 (26.3)

Furthermore, in Figure 6, we depict the unfolding stages of a (representative) execution sequence within a test scenario. Specifically, we present the dynamic progression of both the average velocity within the subset of vehicles currently under the multi-agent's control, and the real-time count of vehicles effectively served, each measured per second. As expected, the velocity of the vehicles decreases as the scenario unfolds, and multiple vehicles gather around the intersection to traverse it. Inevitably, the number of served vehicles is increasing. Comparing the 3-way to the 4-way intersection (with identical arrival rates and road length), the former exhibits a higher susceptibility to congestion. This is normal due to the intricate nature of the 3-way setup, which tends to amplify congestion-related challenges compared with the relatively more balanced traffic distribution in a 4-way configuration.

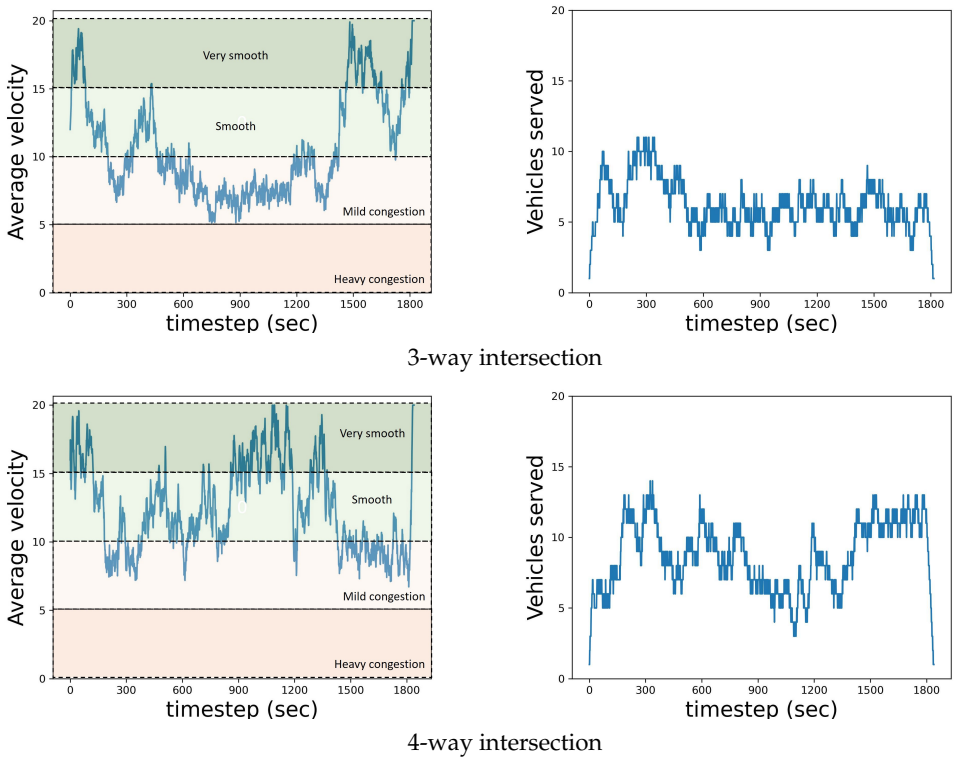


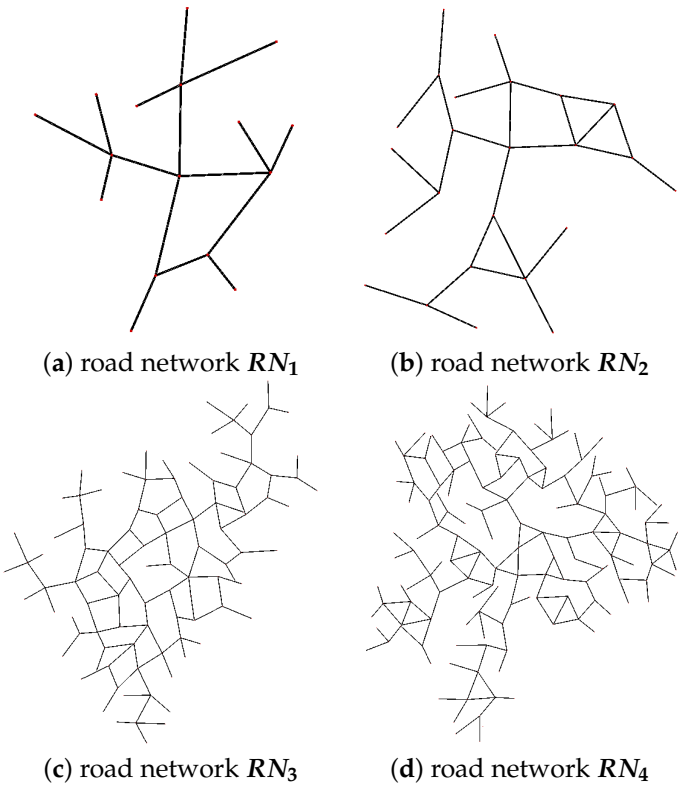
Figure 6. Dynamic evolution of both the average velocity and the frequency of vehicles being served per second, obtained from the implementation of learned multi-agent policies on intersection patterns within a designated test scenario. Traffic state colored zones are also shown.

The same observations can be obtained by adopting the speed performance index (SPI) used to evaluate urban road traffic conditions [57]. This measure quantifies the congestion level (ranging from 0 to 100) and can be defined by the ratio between vehicles' velocity (v) and the maximum permissible velocity (v_{max}), $SPI = (\frac{v}{v_{max}}) \times 100$. According to this index, the traffic state level can be classified into four zones with three threshold values (25, 50, and 75): heavy ($[0, 25]$), mild ($[25, 50]$), smooth ($[50, 75]$), and very smooth ($[75, 100]$) congestion levels. Since in our study we have considered $v_{max} = 20$ m/s, then the congestion zones are divided into value intervals of length 5. The depicted velocity diagrams in Figure 6 show this division into (colored) zones in both intersections. In summary, the congestion level never reached a heavy degree throughout the execution of the scenarios. The most common traffic state in the 4-way intersection was the smooth, while halfway through the scenario of the 3-way intersection the status was found to be mild congestion, due to its structure and its inability to offer a balanced distribution of traffic.

4.3.2. Performance of Zero-Shot Transfer Knowledge to Large Road Networks

Additional experiments were performed in order to measure the proposed knowledge transfer process and study its capability to successfully reuse the learned policies of intersection patterns in large-scale (unknown) traffic networks. For this reason, we constructed multiple simulated environments where we considered only two constraints: (a) all intersections are either 3-way or 4-way, and (b) all roads have variable length falling within a specified range, $[200, 400]$ m. No other structural constraints were assumed during the generation process.

Figure 7 illustrates four such artificial road networks that were created, while in Table 2 we present some of their particular characteristics, such as the number of roads and the number of 3-way and 4-way intersections. It is important to emphasize that the learned driving policies of both intersection patterns were directly applied to all available intersections of any road network.



(**a**) road network $\boldsymbol{RN_1}$ (**b**) road network $\boldsymbol{RN_2}$

(**c**) road network $\boldsymbol{RN_3}$ (**d**) road network $\boldsymbol{RN_4}$

Figure 7. Four artificial road networks of increasing complexity that were generated for evaluating the knowledge transfer process. Every intersection is a noisy copy of either the default 3-way or 4-way intersection pattern.

Table 2. Specifications of the four artificial road networks used in our experimental study.

Road Network	*# Roads*	*# Intersections*	
		3-Way	*4-Way*
RN_1	32	2	4
RN_2	54	9	4
RN_3	230	35	21
RN_4	334	43	35

Subsequently, we created several noisy traffic scenarios on these networks, each lasting approximately two (2) hours (7200 timesteps). These scenarios exhibited varying levels of randomness in the arrival rate of vehicles. While the uniform rate of $[1, 6]$ s remained consistent with the scenarios used during training, we additionally considered three other arrival rates with lower upper limits (5, 4, and 3), introducing in such a way further diversity to the continuous rate of vehicles' arrival and thus scenarios of increased complexity. Overall we considered the following scenario cases (of 2 h traffic control):

- SC_6: uniform arrival rate on interval $[1, 6]$—total of around 2100 vehicles (about 18 vehicles per minute);
- SC_5: uniform arrival rate on interval $[1, 5]$—total of around 2400 vehicles (about 20 vehicles per minute);
- SC_4: uniform arrival rate on interval $[1, 4]$—total of around 2900 vehicles (about 24 vehicles per minute);
- SC_3: uniform arrival rate on interval $[1, 3]$—total of around 3600 vehicles (about 30 vehicles per minute).

This experimental setup offers a dual advantage. Firstly, it enables us to measure the effectiveness of the learned policies in efficiently managing larger networks and handling massive scenarios with multiple vehicles, in contrast to the trained ones. Secondly, it allows us to assess their ability to generalize satisfactory and remain robust in complex environments, which may involve unseen, or partially seen, cases.

For the evaluation study, we relied on the following procedure. For each one of the four (4) types of scenario (SC_k, $k = \{6, 5, 4, 3\}$), we generated ten (10) different traffic scenarios on each road network. We used five (5) different pairs of the trained 3-way and 4-way intersection policies randomly selected from those generated during template training, called P_1, P_2, P_3, P_4, and P_5, correspondingly (each one of the 5 + 5 = 10 available policies was selected only once).

In addition, we examined the performance of an hlensemble MARL policy, denoted as $\overline{\boldsymbol{P}}$, which was derived from the joint application of the five (5) available policies. Specifically, the ensemble strategy was located on the action selection mechanism, where the continuous-valued action of the ensemble multi-agent was calculated as the average of all five (5) available trained policies for each intersection pattern. Thus, we performed 240 ($6 \times 10 \times 4$) simulation runs on every road network, for the six (6) policy combinations, the ten (10) noisy scenarios, and the four (4) scenario cases. Finally, a comparison was also conducted using the Krauss default model of SUMO.

Tables 3–6 show the obtained comparative results in terms of two (2) evaluation criteria: mean velocity, $\boldsymbol{V}$ (m/s), and mean duration, $\boldsymbol{T}$ (s), of all vehicles in the four (4) road networks, RN_1–RN_4, respectively. It is important to note that, in all cases, the proposed method manages to successfully resolve all tasks without any collisions. Again, it is evident that this performance aligns with the results obtained during the training of MARL policies, even in the cases SC_4 and SC_3, which represent scenarios with heavy traffic conditions.

Table 3. Comparative results of the five independent learned policies, the average policy, and the Krauss model in the case of traffic network RN_1.

Policy	**Scenario SC_6**		**Scenario SC_5**		**Scenario SC_4**		**Scenario SC_3**	
	V (m/s)	T (s)	V (m/s)	T (s)	V (m/s)	T (s)	V (m/s)	T (s)
P_1	17.0	49.1	16.5	50.7	15.4	54.5	11.9	70.4
P_2	18.0	46.1	17.5	47.6	16.5	51.0	12.6	67.1
P_3	17.4	47.9	16.9	49.5	15.8	53.3	12.0	70.2
P_4	17.9	46.5	17.4	47.8	16.4	51.2	12.5	67.9
P_5	17.9	46.5	17.4	47.9	16.4	51.2	12.5	67.7
$\overline{\boldsymbol{P}}$	18.0	46.3	17.5	47.7	16.4	51.2	12.5	68.1
Krauss	13.8	58.7	13.7	59.1	13.6	59.8	13.4	60.4
(collisions)	(17)		(29)		(34)		(51)	

Table 4. Comparative results of the five independent learned policies, the average policy, and the Krauss model in the case of traffic network RN_2.

Policy	**Scenario SC_6**		**Scenario SC_5**		**Scenario SC_4**		**Scenario SC_3**	
	V (m/s)	T (s)	V (m/s)	T (s)	V (m/s)	T (s)	V (m/s)	T (s)
P_1	17.8	78.2	17.6	79.3	17.1	81.4	15.5	88.9
P_2	18.5	75.1	18.3	76.2	17.8	78.1	16.5	84.9
P_3	18.1	77.1	17.8	78.3	17.0	80.9	15.7	88.1
P_4	18.2	76.6	17.9	77.7	17.5	79.6	16.3	86.1
P_5	18.1	77.0	17.8	78.1	17.4	79.9	16.2	86.4
$\overline{\boldsymbol{P}}$	18.6	74.9	18.3	75.9	17.9	77.8	16.5	84.7
Krauss	14.3	96.2	14.2	96.7	14.1	97.1	14.0	98.0
(collisions)	(17)		(17)		(28)		(40)	

In all cases, we took the best results using the ensemble policy, $\overline{\boldsymbol{P}}$, showing a remarkable stability and tolerance in environments with variable noise levels. It is interesting to note that there were cases where its performance was slightly improved in comparison with the optimal combined policy (P_i) found, revealing its inherent qualities and its capacity to provide traffic solutions that are remarkably robust and adaptable.

Table 5. Comparative results of the five independent learned policies, the average policy, and the Krauss model in the case of traffic network RN_3.

Policy	**Scenario SC_6**		**Scenario SC_5**		**Scenario SC_4**		**Scenario SC_3**	
	V (m/s)	T (s)	V (m/s)	T (s)	V (m/s)	T (s)	V (m/s)	T (s)
P_1	18.5	169.3	18.4	170.0	18.3	171.4	18.1	172.9
P_2	19.4	161.3	19.3	161.9	19.2	163.1	19.1	164.3
P_3	18.4	170.0	18.3	170.7	18.1	172.0	17.9	174.0
P_4	18.9	165.2	18.9	165.8	18.8	167.0	18.6	168.2
P_5	18.8	166.7	18.7	167.2	18.6	168.5	18.5	169.6
$\overline{\boldsymbol{P}}$	19.5	160.7	19.4	161.3	19.3	162.5	19.2	163.7
Krauss	14.6	213.8	14.6	214.0	14.5	215.5	14.4	215.5
(collisions)	(9)		(11)		(14)		(20)	

Table 6. Comparative results of the five independent learned policies, the average policy, and the Krauss model in the case of traffic network RN_4.

Policy	Scenario SC_6		Scenario SC_5		Scenario SC_4		Scenario SC_3	
	V (m/s)	T (s)	V (m/s)	T (s)	V (m/s)	T (s)	V (m/s)	T (s)
P_1	18.5	183.5	18.4	183.6	18.3	184.9	18.2	185.4
P_2	19.4	174.6	19.4	174.5	19.3	175.7	19.2	176.0
P_3	18.7	180.9	18.7	181.0	18.6	182.4	18.4	183.1
P_4	19.0	178.5	18.9	178.3	18.9	179.5	18.8	179.8
P_5	18.8	180.1	18.8	179.9	18.7	181.2	18.6	181.5
$\overline{P}$	19.5	174.0	19.4	173.9	19.3	175.1	19.2	175.4
Krauss	14.6	231.2	14.6	231.0	14.5	231.9	14.5	231.3
(collisions)	(11)		(10)		(17)		(23)	

Based on the illustrated driving profiles in the learning diagrams, it is apparent that the average duration for traversing individual intersections, whether they are 3-way or 4-way, stands at roughly 16 s. Upon implementation within the experimental road networks and integration with their corresponding intersections, the observed outcomes, shown in Tables 3–6, reveal not only a high degree of resemblance to their learned behavior, but also an ability to extrapolate and accommodate diverse traffic scenarios and instances of congestion. It can be easily observed, for example, that the average duration for passing through any intersection in the network RN_1 (Table 3) ranges from around 12 (case SC_6) to 20 s (case SC_3), depending on the level of the vehicles' arrival rate. This proves that the multi-agent policies are designed in such a way that they exhibit both resilience and adaptability, thereby guaranteeing an optimal level of control on traffic management across a spectrum of scenarios and varying intersection layouts within road networks. The proposed knowledge transfer process enhances the scalability and the consistency of these policies.

On the other hand, the Krauss method seemed to reveal weak performance in all cases, since it consistently struggled to navigate scenarios without encountering collisions. In addition, its policy resulted in inducing a traffic state of mild congestion in all road networks and scenarios that persisted for a notably extended duration, showing its limited applicability in achieving smooth traffic flow across diverse scenarios.

The simulation results for several test scenarios can be found in the Supplementary Materials with download links.

5. Conclusions and Future Work

In this study, we proposed an active method for traffic management in large-scale urban networks with unsignalized intersections. Introducing the notion of intersection patterns and road-agents, we established efficient multi-agent reinforcement learning schemes for traffic control at intersections, and learning strategies that construct economical, robust, and scalable modeling structures. The generated multi-agent policies are then reused in large traffic networks, which are compositions of intersection patterns, thereby offering compact and secure traffic regulation solutions. Several experiments were conducted with increased levels of stochasticity that enhanced the robustness of the traffic profiles and measured the performance of our zero-shot knowledge transfer approach on unknown larger networks and traffic scenarios.

It is essential to acknowledge the limitations inherent in our study. While our experiments were limited to single-lane road networks, the implications of our findings can encompass multi-lane road setups. Additionally, although our primary emphasis focused on specific interconnecting intersection patterns (3-way and 4-way junctions), it is worth noting the potential for broader applicability to encompass a diverse range of network

configurations, spanning from roundabouts to staggered intersections and other designs. Moreover, investigating techniques for transferring the learned policies across different urban environments or cities and under different traffic conditions constitutes an interesting issue for research study.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/robotics13070109/s1.

Author Contributions: Conceptualization, K.B. and A.-G.S.; methodology, all; software, T.T. and C.S.; validation, T.T. and C.S.; formal analysis, all; investigation, all; resources, all; data curation, T.T. and C.S.; writing—original draft preparation, all; writing—review and editing, all; visualization, T.T. and C.S.; supervision, K.B. and A.-G.S. All authors have read and agreed to the published version of the manuscript

Funding: This research received no external funding.

Data Availability Statement: Python source code and SUMO's simulated road networks are available under request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Qian, B.; Zhou, H.; Lyu, F.; Li, J.; Ma, T.; Hou, F. Toward Collision-Free and Efficient Coordination for Automated Vehicles at Unsignalized Intersection. *IEEE Internet Things J.* **2019**, *6*, 10408–10420. [CrossRef]
2. Lianzhen, W.; Zirui, L.; Jianwei, G.; Cheng, G.; Jiachen, L. Autonomous Driving Strategies at Intersections: Scenarios, State-of-the-Art, and Future Outlooks. In Proceedings of the International Intelligent Transportation Systems Conference, Indianapolis, IN, USA, 19–22 September 2021; pp. 44–51.
3. Dresner, K.; Stone, P. A multiagent approach to autonomous intersection management. *J. Artif. Intell. Res.* **2008**, *31*, 591–656. [CrossRef]
4. Lee, J.; Park, B. Development and evaluation of a cooperative vehicle intersection control algorithm under the connected vehicles environment. *IEEE Intell. Transp. Syst. Mag.* **2012**, *13*, 81–90. [CrossRef]
5. Sutton, R.; Barto, A. *Reinforcement Learning: An Introduction*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2018 .
6. Haydari, A.; Yılmaz, Y. Deep Reinforcement Learning for Intelligent Transportation Systems: A Survey. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 11–32. [CrossRef]
7. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A Comprehensive Survey on Transfer Learning. *Proc. IEEE* **2020**, *109*, 43–76. [CrossRef]
8. Spatharis, C.; Blekas, K. Multiagent reinforcement learning for autonomous driving in traffic zones with unsignalized intersections. *J. Intell. Transp. Syst.* **2022**, *28*, 103–119. [CrossRef]
9. Camponogara, E.; Kraus, W. Distributed Learning Agents in Urban Traffic Control. In Proceedings of the Portuguese Conference on Artificial Intelligence, Beja, Portugal, 4–7 December 2003.
10. Salkham, A.; Cunningham, R.; Garg, A.; Cahill, V. A collaborative reinforcement learning approach to urban traffic control optimization. In Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology, Sydney, Australia, 9–12 December 2008; pp. 560–566.
11. Arel, I.; Liu, C.; Urbanik, T.; Kohls, A. Reinforcement learning based multi-agent system for network traffic signal control. *Intell. Transp. Syst.* **2010**, *4*, 128–135. [CrossRef]
12. Chen, C.; Wei, H.; Xu, N.; Zheng, G.; Yang, M.; Xiong, Y.; Xu, K.; Li, Z. Toward A Thousand Lights: Decentralized Deep Reinforcement Learning for Large-Scale Traffic Signal Control. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 3414–3421. [CrossRef]
13. Rasheed, F.; Yau, K.L.A.; Noor, R.; Wu, C.; Low, Y.C. Deep Reinforcement Learning for Traffic Signal Control: A Review. *IEEE Access* **2020**, *8*, 208016–208044. [CrossRef]
14. Bálint, K.; Tamás, T.; Tamás, B. Deep Reinforcement Learning based approach for Traffic Signal Control. *Transp. Res. Procedia* **2022**, *62*, 278–285. [CrossRef]
15. Lee, D. A Theory of Visual Control of Braking Based on Information about Time-to-Collision. *Perception* **1976**, *5*, 437–459. [CrossRef] [PubMed]
16. Zohdy, I.; Rakha, H. Optimizing driverless vehicles at intersections. In Proceedings of the 19th ITS World Congress, Vienna, Austria, 22–26 October 2012.
17. Ji, J.; Khajepour, A.; Melek, W.W.; Huang, Y. Path Planning and Tracking for Vehicle Collision Avoidance Based on Model Predictive Control with Multiconstraints. *IEEE Trans. Veh. Technol.* **2017**, *66*, 952–964. [CrossRef]
18. Rodrigues de Campos, G.; Falcone, P.; Hult, R.; Wymeersch, H.; Sjöberg, J. Traffic coordination at road intersections: Autonomous decision-making algorithms using model-based heuristics. *IEEE Intell. Transp. Syst. Mag.* **2017**, *9*, 8–21. [CrossRef]
19. Pan, Y.; Lin, Q.; Shah, H.; Dolan, J. Safe Planning for Self-Driving Via Adaptive Constrained ILQR. In Proceedings of the International Conference on Intelligent Robots and Systems, Las Vegas, NV, USA, 24 October 2020–24 January 2021; pp. 2377–2383.

20. Carlino, D.; Boyles, S.D.; Stone, P. Auction-based autonomous intersection management. In Proceedings of the 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013), The Hague, The Netherlands, 6–9 October 2013; pp. 529–534.
21. Wang, S.; Mahlberg, A.; Levin, M.W. Optimal Control of Automated Vehicles for Autonomous Intersection Management with Design Specifications. *Transp. Res. Rec.* **2022**, *2677*, 1643–1658. [CrossRef]
22. Levin, M.W.; Rey, D. Conflict-point formulation of intersection control for autonomous vehicles. *Transp. Res. Part C Emerg. Technol.* **2017**, *85*, 528–547. [CrossRef]
23. Li, J.; Hoang, T.A.; Lin, E.; Vu, H.L.; Koenig, S. Intersection Coordination with Priority-Based Search for Autonomous Vehicles. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 11578–11585. [CrossRef]
24. Lu, G.; Shen, Z.; Liu, X.; Nie, Y.M.; Xiong, Z. Are autonomous vehicles better off without signals at intersections? A comparative computational study. *Transp. Res. Part B Methodol.* **2022**, *155*, 26–46. [CrossRef]
25. Codevilla, F.; Miiller, M.; López, A.; Koltun, V.; Dosovitskiy, A. End-to-End Driving Via Conditional Imitation Learning. In Proceedings of the International Conference on Robotics and Automation, Brisbane, QLD, Australia, 21–25 May 2018; IEEE Press: New York, NY, USA, 2018; pp. 1–9.
26. Menda, K.; Driggs-Campbell, K.; Kochenderfer, M. EnsembleDAgger: A Bayesian Approach to Safe Imitation Learning. In Proceedings of the International Conference on Intelligent Robots and Systems, Macau, China, 3–8 November 2019; pp. 5041–5048.
27. Bouton, M.; Cosgun, A.; Kochenderfer, M. Belief State Planning for Autonomously Navigating Urban Intersections. In Proceedings of the IEEE Intelligent Vehicles Symposium, Los Angeles, CA, USA, 11–14 June 2017; pp. 825–830.
28. Tram, T.; Jansson, A.; Grönberg, R.; Ali, M.; Sjöberg, J. Learning negotiating behavior between cars in intersections using deep q-learning. In Proceedings of the International Conference on Intelligent Transportation Systems, Maui, HI, USA, 4–7 November 2018; pp. 3169–3174.
29. Tram, T.; Batkovic, I.; Ali, M.; Sjöberg, J. Learning When to Drive in Intersections by Combining Reinforcement Learning and Model Predictive Control. In Proceedings of the Intelligent Transportation Systems Conference, Auckland, New Zealand, 27–30 October 2019; pp. 3263–3268.
30. Isele, D.; Cosgun, A.; Fujimura, K. Analyzing Knowledge Transfer in Deep Q-Networks for Autonomously Handling Multiple Intersections. *arXiv* **2017**, arXiv:1705.01197.
31. Isele, D.; Rahimi, R.; Cosgun, A.; Subramanian, K.; Fujimura, K. Navigating Occluded Intersections with Autonomous Vehicles Using Deep Reinforcement Learning. In Proceedings of the International Conference on Robotics and Automation, Brisbane, Australia, 21–25 May 2018; pp. 2034–2039.
32. Li, C.; Czarnecki, K. Urban Driving with Multi-Objective Deep Reinforcement Learning. In Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems, Montreal, QC, Canada, 13–17 May 2019; pp. 359–367.
33. Shao, C.; Cheng, F.; Xiao, J.; Zhang, K. Vehicular intelligent collaborative intersection driving decision algorithm in Internet of Vehicles. *Future Gener. Comput. Syst.* **2023**, *145*, 384–395. [CrossRef]
34. Akhauri, S.; Zheng, L.; Lin, M. Enhanced Transfer Learning for Autonomous Driving with Systematic Accident Simulation. In Proceedings of the International Conference on Intelligent Robots and Systems, Las Vegas, NV, USA, 24 October 2020–24 January 2021; pp. 5986–5993.
35. Chiba, S.; Sasaoka, H. Basic Study for Transfer Learning for Autonomous Driving in Car Race of Model Car. In Proceedings of the International Conference on Business and Industrial Research, Bangkok, Thailand, 20–21 May 2021; pp. 138–141.
36. Shu, H.; Liu, T.; Mu, X.; Cao, D. Driving Tasks Transfer Using Deep Reinforcement Learning for Decision-Making of Autonomous Vehicles in Unsignalized Intersection. *IEEE Trans. Veh. Technol.* **2022**, *71*, 41–52. [CrossRef]
37. Xu, Z.; Tang, C.; Tomizuka, M. Zero-shot Deep Reinforcement Learning Driving Policy Transfer for Autonomous Vehicles based on Robust Control. In Proceedings of the Proceeding of the 21st International Conference on Intelligent Transportation Systems, Maui, HI, USA, 4–7 November 2018; pp. 2865–2871.
38. Kirk, R.; Zhang, A.; Grefenstette, E.; Rocktaschel, T. A Survey of Zero-shot Generalisation in Deep Reinforcement Learning Systems. *J. Artif. Intell. Res.* **2023**, *76*, 201–264. [CrossRef]
39. Qiao, Z.; Muelling, K.; Dolan, J.; Palanisamy, P.; Mudalige, P. Automatically Generated Curriculum based Reinforcement Learning for Autonomous Vehicles in Urban Environment. In Proceedings of the IEEE Intelligent Vehicles Symposium, Changshu, China, 26–30 June 2018; pp. 1233–1238.
40. Anzalone, L.; Barra, S.; Nappi, M. Reinforced Curriculum Learning for Autonomous Driving in Carla. In Proceedings of the International Conference on Image Processing, Anchorage, AK, USA, 19–22 September 2021; pp. 3318–3322.
41. Jin, H.; Peng, Y.; Yang, W.; Wang, S.; Zhang, Z. Federated Reinforcement Learning with Environment Heterogeneity. In Proceedings of the 25th International Conference on Artificial Intelligence and Statistics, Virtual, 28–30 March 2022; pp. 18–37.
42. Fan, F.X.; Ma, Y.; Dai, Z.; Tan, C.; Low, B.K.H. FedHQL: Federated Heterogeneous Q-Learning. In Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, London, UK, 29 May–2 June 2023; pp. 2810–2812.
43. Liang, X.; Liu, Y.; Chen, T.; Liu, M.; Yang, Q. Federated Transfer Reinforcement Learning for Autonomous Driving. In *Federated and Transfer Learning*; Springer: Cham, Switzerland, 2023; pp. 357–371.
44. Da Silva, F.L.; Taylor, M.; Reali Costa, A.H. Autonomously Reusing Knowledge in Multiagent Reinforcement Learning. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 5487–5493.
45. Da Silva, F.L.; Reali Costa, A.H. A Survey on Transfer Learning for Multiagent Reinforcement Learning Systems. *J. Artif. Intell. Res.* **2019**, *64*, 645–703. [CrossRef]

46. Zhou, Z.; Liu, G.; Tang, Y. Multi-Agent Reinforcement Learning: Methods, Applications, Visionary Prospects, and Challenges. *arXiv* **2023**, arXiv:2305.10091.
47. Candela, E.; Parada, L.; Marques, L.; Georgescu, T.; Demiris, Y.; Angeloudis, P. Transferring Multi-Agent Reinforcement Learning Policies for Autonomous Driving using Sim-to-Real. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Kyoto, Japan, 23–27 October 2022; pp. 8814–8820.
48. Jang, K.; Vinitsky, E.; Chalaki, B.; Remer, B.; Beaver, L.; Andreas, M.; Bayen, A. Simulation to scaled city: Zero-shot policy transfer for traffic control via autonomous vehicles. In Proceedings of the ACM/IEEE International Conference on Cyber-Physical Systems, Montreal, QC, Canada, 16–18 April 2019; pp. 291–300.
49. Kochenderfer, M. *Decision Making under Uncertainty: Theory and Application*; MIT Press: Cambridge, MA, USA, 2015.
50. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal Policy Optimization Algorithms. *arXiv* **2017**, arXiv:1707.06347.
51. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep reinforcement learning. In Proceedings of the 4th International Conference on Learning Representations, San Juan, Puerto Rico, 2–4 May 2016.
52. Haarnoja, T.; Zhou, A.; Abbeel, P.; Levine, S. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 1856–1865.
53. Haarnoja, T.; Zhou, A.; Hartikainen, K.; Tucker, G.; Ha, S.; Tan, J.; Kumar, V.; Zhu, H.; Gupta, A.; Abbeel, P.; et al. Soft Actor-Critic Algorithms and Applications. *arXiv* **2018**.
54. Fujimoto, S.; van Hoof, H.; Meger, D. Addressing Function Approximation Error in Actor-Critic Methods. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; Dy, J.G., Krause, A., Eds.; Volume 80, pp. 1582–1591.
55. Lopez, P.A.; Behrisch, M.; Bieker-Walz, L.; Erdmann, J.; Flötteröd, Y.P.; Hilbrich, R.; Lücken, L.; Rummel, J.; Wagner, P.; Wießner, E. Microscopic Traffic Simulation using SUMO. In Proceedings of the 21st IEEE International Conference on Intelligent Transportation Systems, Maui, HI, USA, 4–7 November 2018.
56. Krauss, S.; Wagner, P.; Gawron, C. Metastable states in a microscopic model of traffic flow. *Phys. Rev. E* **1997**, *55*, 5597–5602. [CrossRef]
57. Afrin, T.; Yodo, N. A Survey of Road Traffic Congestion Measures towards a Sustainable and Resilient Transportation System. *Sustainability* **2020**, *12*, 4660. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.