IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 July 2024, accepted 23 July 2024, date of publication 29 July 2024, date of current version 7 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3434965

RESEARCH ARTICLE

Distributed Highway Control: A Cooperative Reinforcement Learning-Based Approach

BÁLINT KŐVÁRI[1,2], ISTVÁN GELLÉRT KNÁB[3], DOMOKOS ESZTERGÁR-KISS[4], SZILÁRD ARADI[1], (Member, IEEE), AND TAMÁS BÉCSI[1], (Member, IEEE)

[1]Department of Control for Transportation and Vehicle Systems, Faculty of Transportation Engineering and Vehicle Engineering, Budapest University of Technology and Economics, 1111 Budapest, Hungary
[2]Asura Technologies Ltd., 1122 Budapest, Hungary
[3]Systems and Control Laboratory, HUN-REN Institute for Computer Science and Control (SZTAKI), 1111 Budapest, Hungary
[4]Department of Transport Technology and Economics, Budapest University of Technology and Economics, 1111 Budapest, Hungary

Corresponding author: Tamás Bécsi (becsi.tamas@kjk.bme.hu)

This work was supported in part by European Union within the Framework of the National Laboratory for Autonomous Systems under Grant RRF-2.3.1-21-2022-00002; and in part by the Ministry of Innovation and Technology of Hungary from the National Research, Development and Innovation Fund, financed under the TKP2021 Funding Scheme under Project BME-NVA-02. The work of Tamás Bécsi was supported by the János Bolyai Research Scholarship of Hungarian Academy of Sciences under Grant BO/00233/21/6.

ABSTRACT With increasing realised traffic on transport networks, greenhouse gas emissions show a similar trend. Reducing them is a modern aspiration, creating a better place to live and moving towards sustainability. Expanding the infrastructure is often not an appropriate solution, as the system would only be fully utilised at peak times, while at less frequent times it would not even approach capacity and would require huge investment costs. An alternative to further construction work is the implementation of intelligent traffic systems, where smoother flows can achieve higher capacity by reducing the variability in the system. In a motorway environment, a common approach is Variable Speed Limit Control, where the road is divided into zones and individual speed limits are used to increase or decrease the load on the cells. This paper proposes a solution in which individual cells make decisions cooperatively, in contrast to classical state machine-based methods. Thanks to the jointly formulated goal of the agents, a predictive control method is created that leads to a reduction in emissions due to avoided shock waves and reduced waiting times. This paper presents a solution that provides a universal solution across multiple application lengths, illustrating the power of deep learning. https://github.com/istvan-knab/Variable-speed-limit-control.

INDEX TERMS Variable speed limit control, multi-agent reinforcement learning, machine learning, traffic simulation.

I. INTRODUCTION

A key guiding principle in the design of today's cities, including high-speed suburban areas, is to reduce noise and greenhouse gas emissions to create a more liveable environment. According to the report from [1], 23% of greenhouse emissions come from the transport sector. In terms of transportation, the aim is to make the traffic flow as smooth as possible to reduce the time vehicles spend idling. When traffic on the network exceeds the limits of the physical infrastructure, further improvements can be achieved through the use of traffic control solutions in the absence of physical expansion [2]. Static traffic controllers with adjustable values are most commonly used in urban environments [3], such as traffic lights and adaptive systems [4], but the need is similar in highly congested time windows to set up dynamically controllable systems on high-speed motorways [5]. Both Traffic Signal Control [3] and the Variable Speed Limit Control that appears on highways fall under the category of Intelligent Transportation Systems, thereby implementing event-driven control.

On highways, other types of anomalies are responsible for the loss of capacity due to the potential for quicker change from higher speed ranges. This can generate the shock wave effect shown in Figure 1, a moving jam due to the cumulative nature of the reaction times along the

The associate editor coordinating the review of this manuscript and approving it for publication was Qiang Li.

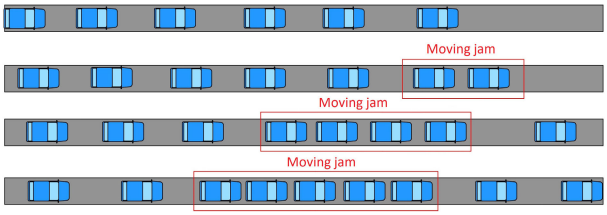

FIGURE 1. Moving jam.

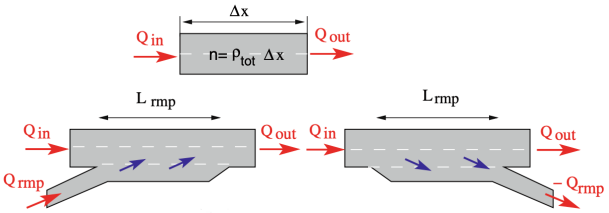

FIGURE 2. Continuity in traffic flow application [10].

vehicle column. To avoid these, it is advisable to minimise sudden changes in traffic flows, thus aiming for a steady and constant flow. This is the purpose of vehicle platoons in microscopic traffic management, focusing on the acceleration and deceleration of vehicles in specific traffic situations [6]. In the macroscopic case, this can be done with Variable Speed Limit Control(VSLC), which provides a number of algorithms to solve the problem [7].

For traffic flows where the objective is not to plan the trajectory of individual vehicles [8], but to achieve flow stability, it is entirely appropriate to use a macroscopic model that takes into account traffic characteristics such as traffic density (ρ), average speeds (v), and flow magnitude [9]. These measures are the same as those used in fluid mechanics to characterise the state of each medium, so the behaviour of the traffic network can be described in physical terms as an analogy. The most important of these laws is the continuity theorem, which defines the flow rate at a given cross-section in terms of velocity and density. In traffic situations, the cross-section is represented by the number of lanes that can be used by vehicles. The aforementioned relationship, which establishes a connection between the various cross-sections and the main flow indicators, can be formulated as follows:

$$\rho_1 v_1 A_1 = \rho_2 v_2 A_2, \tag{1}$$

where v_i denotes the average speed of the vehicles in the considered zone. Additionally, as previously mentioned, the density, denoted as ρ_i, is also an important indicator, since Figure 3 demonstrates that after a certain point, the increase in density negatively impacts the flow rate. As can be seen, the successive cross-sections(A_i) are in a linear relationship with the other two variables, so the constrictions of that value lead to a performance drop, which motorway entrances and urban network connections can cause, as well as lane closures due to construction or accidents. Since the cross-sections only change rapidly in the event of accidents, the formulated control task focuses on influencing the other two variables.

Due to continuity, the total demand can be given as the sum of the traffic flows interpreted on the lanes, both at entry and exit, as shown in the Figure 2 following the equation:

$$\sum_{i=1}^{n} Q_{in_i} = \sum_{j=1}^{m} Q_{out_j}, \tag{2}$$

indexed by the current incoming lanes with i and the outflow of the previous sector j.

As stated above, the cross-section is a static indicator of the infrastructure, so the manipulation of the velocity and density values can be used to determine the control as specified in Equation 1, where, due to the macroscopic approach, velocities are not understood as individual interventions, but as the definition of speed limits.

As will be shown in the next section, using the fundamental diagram (Figure 3), by looking at the relationship between flow and density on the two axes, the aim is to target the top of the increasing branch with density, where the flow rate is still influenced in a positive direction for small fluctuations by the increase in this indicator. Still keeping in mind the constraint of satisfying the Equation 1, the objective function is the ideal choice of density interventions through speed limits as a driver-interpretable action.

A. RELATED WORK

Given the static nature of the cross-sections and the inability of the infrastructure to expand spatially, an event-driven traffic control solution must be implemented to ensure traffic stability. As discussed in the previous section, assuming that density is not meaningful for drivers in a traffic situation due to human sensory limitations, the remaining variable in the Equation 1 is the average speed, which can be approximated by introducing ideal constraints in allowed speed, divided by sector. The speed limits introduced in this way can be important for both safety and flow aspects within a network [3].

VSLC [11], [12] is a widely used macroscopic traffic management approach for which several algorithms have been implemented so far. This method is based on the spatial distribution of the flow, which is achieved by dividing the road network into segments and setting individual speed limits for each segment, thus varying the temporal flow of traffic.

The advantage of this system lies in its dynamic nature, which intervenes on an event-driven basis by monitoring traffic in specific congestion situations since the dynamics of a column of vehicles are different during peak traffic than during a period when it is operating at a fraction of its capacity [13]. Among the conventional approaches, the SPECIALIST [14], [15] and Motorway Control System (MCS) algorithms play an important role, the latter already in use on Swedish and Dutch motorways [7], so this will be taken as a basis for evaluating the effectiveness of the algorithm.

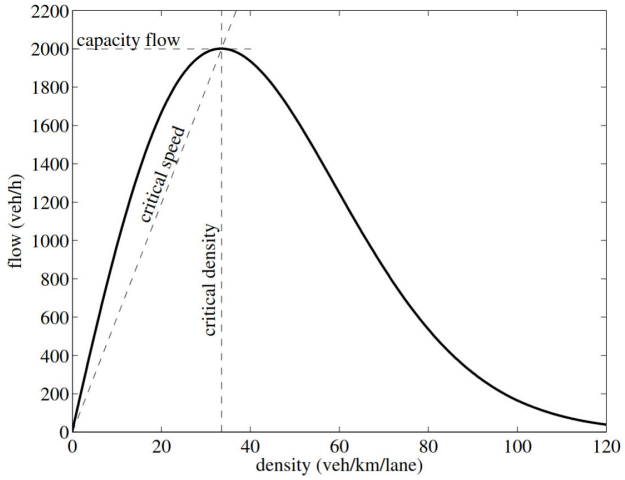

FIGURE 3. Fundamental diagram to represent the relationship between density and flow magnitude. [2].

MCS operates by measuring the speed of a given cross-section and adjusting the speed limit to a slightly higher value and also adjusting two previous limits to a ramping speed value to achieve smoother deceleration and resolve established schockwaves.

SPECIALIST operates via shock wave identification by coupling time-location and flow-density diagrams. It describes a larger set of current situations due to its four phases and six states, thus providing a more effective control due to better segmentation of states. In both cases, the algorithms help to resolve emerging congestion, but the reduction of existing congestion is still to be achieved, and although they increase traffic efficiency, they are not an optimal solution.

In order to fully assess these shock waves, some look-forward response is needed. In the case of pre-estimated interventions, model-based solutions such as Model Predictive Control (MPC) can also be applied by defining physical equations and state-space-based control, but today's research directions are increasingly dominated by data-based solutions using Machine Learning (ML). Among such solutions, the one of most dynamically developing are the algorithms based on Neural Networks (NN) [16]. Their applicability shows great diversity, ranging from machine vision to various sequential decision-making processes. For this problem, Reinforcement Learning (RL) algorithms are mostly given in the context of freeway onramps [17], [18], [19], [20], but in an earlier phase of research, prior to the implementation discussed in this paper, an emission reduction model was developed to demonstrate that a significant improvement over uncontrolled systems can be achieved using a Deep Learning-based(DL) approach.

Building on this research, the need arose to test suitability in environments with arbitrary sensor density. Roadside units require significant investments [5], so their applicability to already existing diverse equipment represents a major advancement. Wider applicability implies that a trained model can be applied to multiple lengths, so achieving controllability of motorway sections with the same character could lead to further improvements. If a model could do this across multiple environments, the offline training process could be significantly shortened, making the solution more financially successful. This would also demonstrate that it is not just an overfitted model, but a more generalised and workable solution.

B. CONTRIBUTION

As presented in the literature review, RL has already been applied to the VSLC problem. However, the proposed RL framework has several key differences and novelties compared to the existing solutions. These innovations are the following:

- The paper proposes a novel state representation of the VSLC problem that enables scalability in terms of the size of the controlled highway section. In the literature, the papers use two main approaches; the first is describing the state of the currently controlled highway zone only [20], [21], [22], or showing the state of all the controlled zones at once [17]. Compared to that, we propose a sliding windows-based approach where an agent state representation is composed of the agent's highway zone and its neighboring zones; this concept makes the agent invariant to the control highway length and supports better performance since an agent can see beyond its highway zones.
- The proposed action space is also novel compared to the literature since all the papers use fixed discrete speed limits [17], [20], [21], [23] or continuous speed limits [22]. The proposed action space represents the change amount to the currently used speed limit. Thanks to that, the agent can not change the speed limit drastically in one step, while it can completely cover any speed limit interval without re-training the model. Furthermore, the amount of change can be altered without re-training since it can be redefined in the configuration, making the proposed concept more flexible than existing solutions.
- The reward functions in the literature use mostly TTS (Total Time Spent) [24], mean speed [23], traffic density distribution [19], and gas emission [25]. Compared to that, the proposed reward function uses a novel concept that gives maximum reward to the agent when the waiting time is minimal and the average speed is maximal in the entire network

With these three key considerations, the paper formulates a new RL abstraction for the VSLC problem. The potential of the proposed method is demonstrated by comparing the performance to algorithms that are actually deployed on highways in Europe. The results show that the presented approach outperforms the baseline algorithms in every emission metric and also in classic metrics such as waiting time and queue length.

The paper is organized as follows: First, the environment is detailed with its essential components to define the utilized

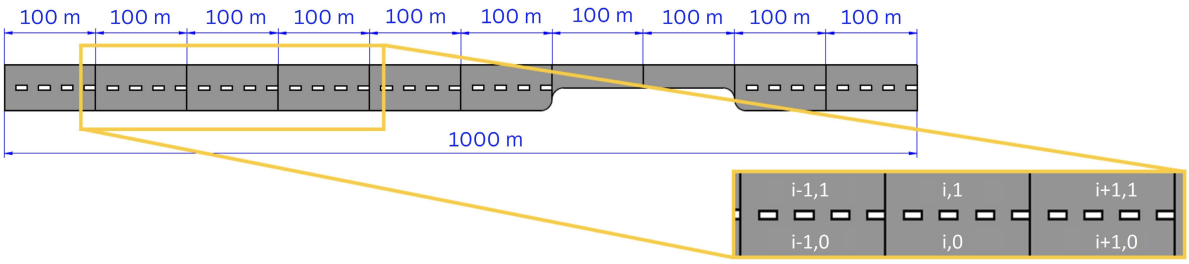

FIGURE 4. The highway environment and agent separation used in the simulation.

state representation, action space, and reward function, and also the SUMO simulator is introduced. Secondly, the utilized algorithms are presented to detail the core concept of RL and MARL. After the methodology, the results are presented to provide a comprehensive evaluation of the performance of the agent compared to the baseline methods. Finally, the possible future developments and directions are discussed in the conclusion.

II. ENVIRONMENT

By analyzing Figure 3, it can be seen that the permeability of a highway cross-section decreases as the density increases. This phenomenon can be caused by several factors, such as cross-section shrinkage from road accidents, lane disappearing, or by road construction. In such scenarios, shock wave formation is inevitable if the environment is not controlled.

For simulation purposes, a highway section is built in the SUMO simulator [26], which is shown in Figure 4. In Figure 4 a zone is the $100m$ long lane pair, which is controlled individually. A cell is a lane of a zone, and the three consecutive zones represent the state representation information for the agent that will control the speed limit in the middle zone. In the modeled highway part, one of the zones' lanes is closed to simulate the cross-section shrinkage that can trigger shock waves.

The highlighted zones in Figure 4 are the ones that are encoded in the state representation and provide the only information to an agent for decision-making, while the chosen action is only applied to the middle highway zone. The neighboring zones in the highway section provide a horizon for the agent to understand the scenario around its controlled zone. Thanks to this concept, the state representation will operate as a sliding window over the entire controlled highway section.

As mentioned above, the SUMO simulator is used to create the highway part mainly because it can measure all the necessary traffic metrics that can characterize the quality of the control methods, such as emissions and waiting time. SUMO also has a great Python interface that allows any kind of method to control the built traffic network. In the created traffic network, the sections' speed limit can be set individually between the $130km/h$ and $10km/h$ interval. The agents are allowed to change the speed limit for every $10s$, and the change can only be $10km/h$ at once. It is important to mention that the generated traffic load always saturates the network but still allows all vehicles to enter. This traffic flow generation scheme is essential to make the evaluations reliable because, with that, all control methods have to push the same amount of vehicles through the network. This is also the terminal criteria of each training episode.

A. STATE REPRESENTATION

The goal of the agent is to produce an appropriate decision based on the sensor data that enables the maximization of the cumulated reward during the episode. Typically, the task of state representation involves organizing relevant sensor data, thereby providing input for decision-making at a given step. The proposed state representation consists of macroscopic traffic indicators, such as density and the average vehicle speed in a given section. The reason for this can be seen in Equation 1; these metrics can characterize the magnitude of the traffic flow. For clarification, the state representation of an agent that controls a highway zone consists of the cells of it is own zone and the cells of the neighboring zones as depicted in Figure 4.

A primary consideration in this formulation is to ensure generalizability. The length of the controlled highway section can be increased, and the agents' performance will last since an agent is provided with information from the neighboring zones along with the controlled one, thanks to the state representation. The implementation is based on an analogy of a sliding window, wherein, at each time step, all agents within the VSLC zone are afforded the opportunity to make decisions. Although the agents do not know about the decisions of other agents. Nevertheless, they can indirectly understand the decisions of others through the sliding window-based state representation, thanks to the horizon.

From a differentiation standpoint, it is advantageous to incorporate into the state representation not only the attributes of the focal segment but also those of its immediate surroundings. This enables the algorithm to discern differences between the active cell and its neighboring cells, thereby facilitating decision-making regarding the load of the respective cell. Taking into account the neighboring cells to ensure continuous monitoring of those in front, behind, and

beside, the state vector for zone i is presented as follows:

$$state_i = \{v_{x,y}, \rho_{x,y} \mid x \in (i-1, i, i+1), y \in (0, 1)\} \quad (3)$$

where $v_{x,y}$ and $\rho_{x,y}$ are the velocity and density values of the cells for lane y of zone x, as indexed in Figure 4, and $i \in (1..n)$, where n is the number of controlled zones. The reason for the state vector consisting of 5 cells is that the neighboring cell alternates between being in front and behind at different time steps.

With this description, the cells before the bottleneck are provided with intervention opportunities, as there is no need for further speed reduction after the decrease in cross-section in terms of continuity. Moreover, the very first cells do not receive intervention opportunities since their state representation cannot be generated. However, in real-world scenarios, this does not pose an issue; it merely signifies the simulated environment's boundary at the first cell.

In summary, the main advantage of this custom-designed state representation is its independence from the position of the segment, describing the relevant environment, and its scalability.

B. ACTION SPACE

Due to the nature of the problem, discrete actions are implemented, as only certain values can be taken by the individual speed-limiting signs. The aspirations are similar to those in creating the state representation, seeking a universal solution that covers as much ground as possible to provide the agent with opportunities to help maintain traffic continuity. In addition to the discrete action space, another factor is the density of interventions, as discussed during the examination of environmental constraints. For this, a 10-second value is specified, both during the testing of the proposed algorithm and as a reference for the MCS serving as a comparative baseline. The action space contains three discrete choices: the first is increasing the speed limit in the given highway zone by $10km/h$, the second is not changing the speed limit in the given zone, and the last one is decreasing the speed limit by $10km/h$. For clarity, the actions can be seen in Equation 4.

$$action = \begin{bmatrix} +10 & km/h \\ 0 & \\ -10 & km/h \end{bmatrix} \quad (4)$$

The main advantage of this concept is that any real speed limit can be created by the agent over time, and it enables faster convergence because the output dimensionality of the neural network is small. The incremental implementation has an additional benefit. Notably, by only allowing slow changes in the system, the likelihood of developing moving traffic waves is further reduced, as this speed adjustment better accommodates drivers' capabilities.

C. REWARD DESIGN

The main concept utilized in subsequent methodologies is to evaluate the quality of decisions made by an agent through feedback in the form of a scalar value provided by the environment. Indeed, the reward signal serves this purpose perfectly. It is derived from the state of the traffic, generating a value that indicates how good a decision is made by the agent.

Several approaches have been considered in the development of this. Initially, simply using the average speed of the network is proposed. However, a problem arose as the network could achieve higher values by sacrificing certain road segments. To address this issue, waiting time is incorporated into the system, with the aim of decreasing it to align with the desired behavior expected from the agent. Since its increase is particularly detrimental to the environment, it is represented not as a linear but as a quadratic term in the reward function, thereby penalizing its growth more heavily. The final reward function for the training process looks as follows:

$$R = \frac{v_{avg}}{(1+w)^2}, \quad (5)$$

- v_{avg} represents the average speed across the entire network.
- w denotes the sum of waiting times across the entire network.

Consequently, the formulated objective function serves the advancement of two tasks. On one hand, increasing the average speed enhances the throughput capacity. On the other hand, reducing waiting times aims to ensure that greenhouse gases are emitted into the air only when actual progress is made.

III. METHODOLOGY

A. REINFORCEMENT LEARNING

In certain scenarios, conventional algorithms and physical models may fail to accurately approximate reality to the desired degree. Furthermore, often the required computational capacity completely precludes real-time usage. The models generated by learning algorithms, however, often provide solutions to such problems. Due to the aforementioned characteristics, Reinforcement Learning [27] is a particularly favored research area for sequential decision-making tasks.

The basic idea behind RL is the communication between an agent and its environment (Figure 5), where the agent takes action at each time step and, in response, receives its next state and a reward that quantifies the quality of the decision by a scalar. In an MDP (S, A, R, S', T) [28], there is a tight connection between the initial states, actions taken, resulting states, and the obtained rewards. From these, so-called state transitions can be formed, which unambiguously define the quality of a decision in a given situation. The agent's objective is to acquire a behavioral strategy that maximizes the cumulative reward attained across successive episodes. In order to make good decisions not only for a given situation but also in the long term, one must take into account the potential rewards available during the subsequent steps. The calculation for this is done as

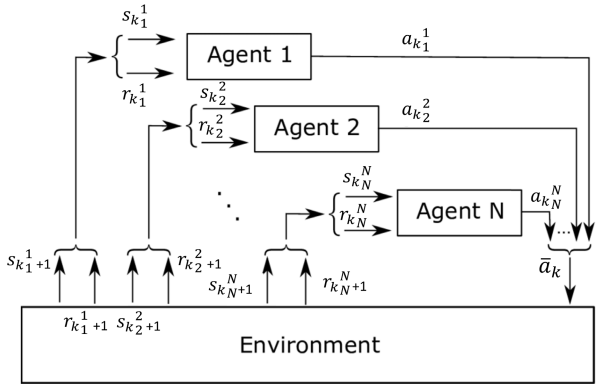

FIGURE 5. The interactions between agents and the environment in case of MARL.

follows:

$$G_t = \sum_{t=0}^{T} \gamma^t \cdot r_t, \tag{6}$$

where the γ discount factor is responsible for reducing the influence of states further in the future on the agent's current decision.

In the initial phase, the agent is not aware of the rewards associated with each state transition; thus, it needs to gather experiences to enable later conscious decision-making. Over time, based on acquired experiences, it can distinguish between successful and less successful decisions by assessing the resulting rewards. In order to explore the environment while striving to learn, two types of decisions can be made: exploratory ones and decisions based on current knowledge, aiming for optimal actions. Due to the initial lack of knowledge, the maximum number of exploratory decisions gradually transforms into fully consciously made decisions over time.

B. DEEP Q NETWORK

The fundamental model of value-based algorithms is provided by Deep Q-Network (DQN) [29], which evaluates state transitions not only based on rewards but also on the predicted Q-values associated with state-action pairs. Learning here is accomplished with the help of the Bellman equation, which looks as follows:

$$Q(s,a;\theta) = Q(s,a;\theta) + \alpha(r + \gamma max_{a'} Q(s',a';\theta^-) - Q(s,a;\theta)). \tag{7}$$

Here, it can be observed that the reward value directly influences the resultant value, while γ modulates the future Q-values attainable. Additionally, there's another parameter observed alongside, known as α, which influences the learning rate, determining the extent of learning.

C. MULTI AGENT REINFORCEMENT LEARNING

Single-Agent Reinforcement Learning can provide solutions only to a fraction of the control problems since, in real life, multiple entities can interact through cooperation [30] or competition [30] to achieve some goal [31]. The interaction means that every agent can change the environment's state, influencing individual agents' state representations. In such cases, Multi-Agent Reinforcement Learning is used to tackle the control problem. MARL can be distinguished from RL in many ways, such as the mathematical framework, which is Markov Games [30] instead of a simple MDP. The utilized MARL approach in this paper is the independent learner concept [32]. This concept means that a single neural network is used during training, and this model is trained to understand every agent's situation and make decisions from their perspective. The approach has several benefits:

- The first one is that employing one neural network and training from the data of every agent perspective can gather more diverse training data that extensively supports the success of the training.
- The second one is about agent communication. Thanks to the novel problem formulation, the agents that have to share information have a shared part in their state representation, and all of them have a shared reward as well. This attribute allows them to understand each other's decisions without dedicated communication. The avoidance of dedicated communication significantly mitigates the complexity of the training, which has a positive effect on convergence and performance.

D. INTERPRETATION OF MARL TO VSLC

MARL is utilized in the formalized VSLC problem since every highway zone requires individual simultaneous decision-making regarding the applied speed limits. The time steps in the environment are discretized, and in every time step, all agents decide the speed limit of their highway zone. Consequently, every controlled highway zone has an agent that can control the speed limit of the zone. Thanks to the novel state representation, a zone agent can understand the highway's local traffic load; hence, direct communication between agents can be avoided, which makes the training process more stable and faster. Even though an individual agent controls every highway zone, one neural network is trained, which is the independent learner MARL concept itself. The state representation of every agent is created at the end of every time step, and then a single neural network predicts the created states, which means that it decides the speed limit in every zone as if it controls the zone. Then, the new speed limits are applied to the controlled highway section for a fixed amount of time, and the process repeats itself as displayed in Figure 6.

E. TRAINING PROCESS

The conditions of the training are crucial from the aspect of the final performance. The agents are trained on $1km$ long highway part, where each controlled zone is $100m$ long. The traffic load is randomly generated for every episode, and the traffic flow always saturates the network, but every vehicle can enter the network; hence, there isn't a jam at the beginning of the highway part. An episode terminates

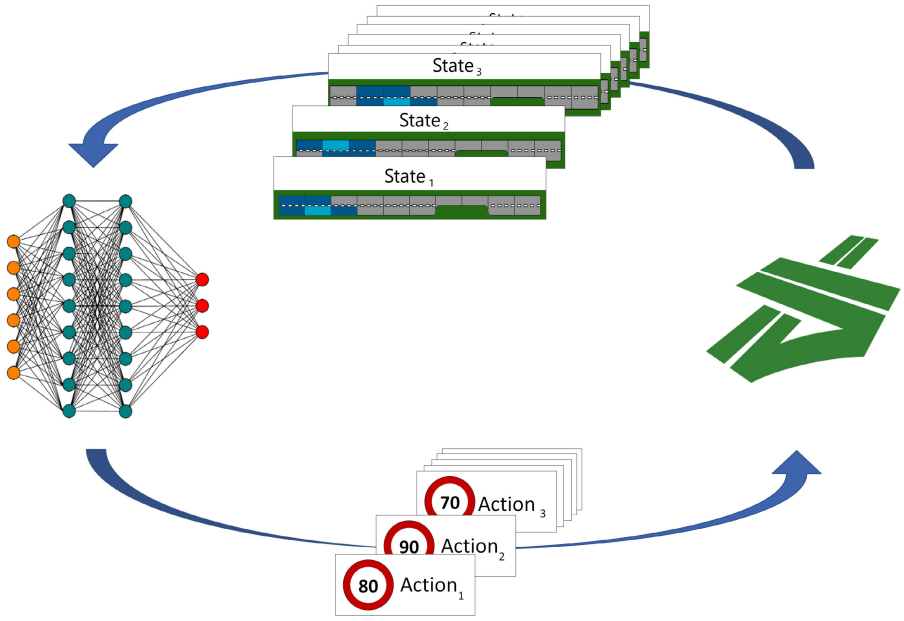

FIGURE 6. Multi-agent decision making.

if the last vehicle leaves the network. During training, the immediate reward concept is utilized. Consequently, after each step, the agents get a reward from the environment to characterize the consequences of their actions. It is also important to mention that the episodes are discretized in time, and the agents can change another speed limit for every $10s$.

IV. RESULTS

A. DATA FOR EVALUATION

The evaluation of competing methods is crucial to make the process reliable. This attribute is ensured through seeding the test simulations, where every competing method has to solve the same scenario. Furthermore, an additional aspect to the evaluation is added that can strongly demonstrate the scalability of the proposed method. Notably, the agents trained on a $1km$ long highway part are evaluated on a $3km$ and a $10km$ long highway part to articulate that the novel problem formulation makes the agents invariant to the length of the controlled highway section. It is also important to mention that the same criteria are applied to the test simulations regarding the training phase in terms of time discretization, terminal events, etc.

As discussed in the introduction, a crucial aspect of the sustainable development of modern cities and their environments is the implementation of intelligent traffic management systems. The main goal of ITS is to make transportation more efficient, and this endeavor can be characterized by sustainability metrics and classic metrics such as waiting time. The computation of these metrics is done by SUMO. The carbon emissions are not realized through specific sensors measuring individual vehicles but rather through various approximate models available within the discussed simulator. Among these, the default HBEFA v2.1 [33] model is utilized in SUMO, which is calculated as follows:

$$E = c_0 + c_1 va + c_2 va^2 + c_3 v + c_4 v^2 + c_5 v^3, \qquad (8)$$

where the c_i constants take different values for certain emissions. These values can be obtained on a per-lane basis, akin to waiting times, and can be aggregated across all sectors to accumulate the total waiting times.

The reduction of these indicators unequivocally defines the energy efficiency of the control method. As a result, these are compared across scenarios, including uncontrolled environments, motorway networks managed by MCS, and scenarios representing the research topic, namely MARL-based solutions deployment.

B. MODEL EVALUATION

During the evaluation, every sustainability metric cannot be examined, yet it is also unnecessary since in emission models only the constants vary thus, there is no magnitude difference in their trends. The same model is used for evaluating CO, CO_2, NO_x, PM_x and HC as well [33]. Among these, CO_2 emissions, as a major contributor to greenhouse gas effect, are one of the significant emissions under scrutiny, along with NO_x. Additionally, illustrating the minimization of idle time spent in shock waves, both the number of vehicles forced to stop and the resulting accumulated waiting times are part of the comparison among the evaluations.

C. EVALUATION BASED ON SUSTAINABILITY METRICS

As a baseline, the MCS algorithm and the control-free scenario are used. As Tables 1-3 demonstrates, the proposed MARL-based solution significantly outperforms both

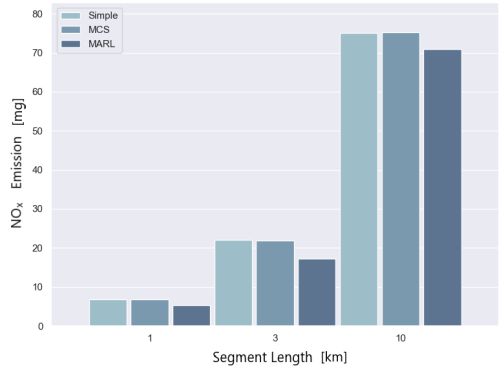

FIGURE 7. NO_X emission in different segment lengths.

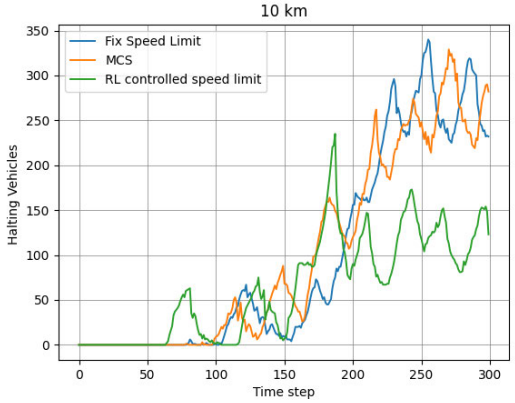

FIGURE 8. Number of halting vehicles.

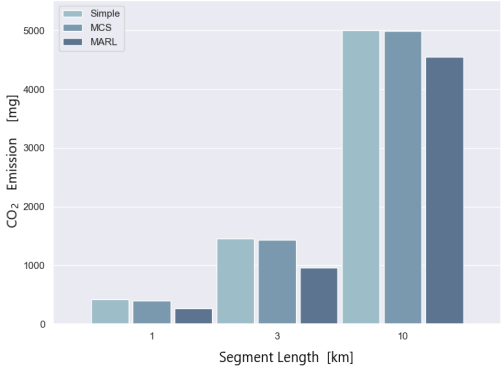

FIGURE 9. CO_2 emission in different segment lengths.

baseline methods in terms of all measured sustainability metrics, which means CO_x and NO_x. The tendencies between the different highway lengths in the sustainability metrics are shown in Figures 7 and 9. The tendencies suggest that there is a great necessity for methods that can adapt to different traffic volumes since the MCS algorithm only slightly defeats the control-free scenario. At the same time, the MARL-based approach holds its 7% reduction in the sustainability measures.

D. EVALUATION BASED ON CLASSIC METRICS

In this evaluation, the same baselines are used as in the sustainability metrics, and the accumulated waiting time

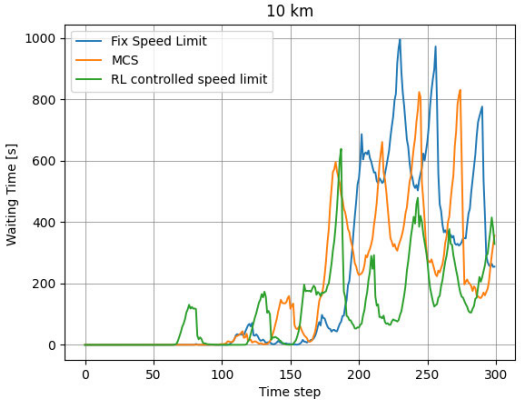

FIGURE 10. Waiting time on the whole system.

TABLE 1. Results compared in a 1 km long highway section.

1 km	Simple	MCS	MARL
CO_2 Emission [mg]	420.38321	402.28047	260.66972
NO_x Emission [mg]	6.87358	6.77200	5.30981
Waiting Time	5.42857	5.09791	0.25446

TABLE 2. Results compared in a 3 km long highway section.

3 km	Simple	MCS	MARL
CO_2 Emission [mg]	1458.40422	1431.63130	955.66162
NO_x Emission [mg]	21.99125	21.8748	17.25388
Waiting Time	24.54404	22.85565	0.007142

TABLE 3. Results compared in a 10 km long highway section.

10 km	Simple	MCS	MARL
CO_2 Emission [mg]	5003.25073	4987.89459	4556.32901
NO_x Emission [mg]	75.02422	75.24406	70.85701
Waiting Time	205.43928	173.36738	109.90261

is measured during the simulations for all three highway lengths. The results are shown in Tables1-3. The results suggest that the proposed MARL-based solution outperformed the control-free scenario and the MCS algorithm. These results demonstrate that trained agents can jointly decrease greenhouse gas emissions and the accumulated waiting time in the network, which means that with the liberty of changing the speed limits in the individual highway zones, the harmful effect of shock waves can be mitigated.

Another evaluation is conducted, which focuses on the characteristics of a single test simulation. The results can be seen in 8 and 10. These Figures articulate that the MARL agent effectively mitigates the effect of the shock waves. The number of halting vehicles can be decreased along with the cumulated waiting time in the traffic network, which significantly benefits road users and the environment.

Based on the results, it can be seen that the MARL-based solution reaches superior performance, compared to already deployed solutions both in terms of greenhouse gas emmisions and classic metrics such as waiting time and halting vehicles.

V. CONCLUSION

Recently, deep learning solutions in ITS have become attractive since they can provide real-time solutions to complex control problems with excellent performance. The disadvantages of black-box-based operations can be neglected since some cases are not safety-critical.

This paper presented a novel approach for the VSLC problem through MARL, which can significantly outperform solutions already deployed on European highways. A thorough evaluation demonstrated that the proposed method:

- Invariant to the length of the controlled highway section; hence, it is enough to train the agents only once.
- Utilizes a control strategy that mitigates greenhouse gas emissions compared to baseline algorithms.
- Utilizes a control strategy that decreases waiting time and the number of halting vehicles that tremendously impact driver experience and the efficiency of the transportation network.

In our future endeavors, some relevant scenarios must be included in the environment to make the simulation as close to the real world as possible. These modifications would be the following:

- Creating highway section that have more or varying lane counts.
- Creating a highway section that has a ramping lane.
- Creating highway section, where the cross-section shrinkage position is randomly chosen from episode to episode.
- Creating highway sections that are imported from the OpenStreet map to show that real environments can also be controlled.

With these modifications, a real-world environment can be built to evaluate the proposed method's potential fully. It is also worth mentioning that in such a real-world comparison, the portfolio of baseline algorithms should also be made more prosperous, which means using different classic control theory-based methods to see how MARL can outperform these solutions.

REFERENCES

[1] P. IEA. (2020). *IEA (2020), the Role of CCUS in Low-Carbon Power Systems*, IEA, Paris, France. [Online]. Available: https://www.iea.org/reports/the-role-of-ccus-in-low-carbon-power-systems

[2] A. Hegyi, *Model Predictive Control for Integrating Traffic Control Measures*. Delft, The Netherlands: Netherlands TRAIL Research School, 2004.

[3] S. Bouktif, A. Cheniki, A. Ouni, and H. El-Sayed, "Deep reinforcement learning for traffic signal control with consistent state and reward design approach," *Knowl.-Based Syst.*, vol. 267, May 2023, Art. no. 110440.

[4] M. Kolat, B. Kóvári, T. Bécsi, and S. Aradi, "Multi-agent reinforcement learning for traffic signal control: A cooperative approach," *Sustainability*, vol. 15, no. 4, p. 3479, Feb. 2023.

[5] T. Degrande, F. Vannieuwenborg, S. Verbrugge, and D. Colle, "Deployment of cooperative intelligent transport system infrastructure along highways: A bottom-up societal benefit analysis for flanders," *Transp. Policy*, vol. 134, pp. 94–105, Apr. 2023.

[6] D. Pi, P. Xue, W. Wang, B. Xie, H. Wang, X. Wang, and G. Yin, "Automotive platoon energy-saving: A review," *Renew. Sustain. Energy Rev.*, vol. 179, Jun. 2023, Art. no. 113268.

[7] E. F. Grumert, A. Tapani, and X. Ma, "Characteristics of variable speed limit systems," *Eur. Transp. Res. Rev.*, vol. 10, no. 2, pp. 1–12, Jun. 2018.

[8] X. Fang and T. Tettamanti, "Change in microscopic traffic simulation practice with respect to the emerging automated driving technology," *Periodica Polytechnica Civil Eng.*, vol. 66, no. 1, pp. 86–95, 2022.

[9] H.-T. Fritzsche and D.-b. Ag, "A model for traffic simulation," *Traffic Engineering+ Control*, vol. 35, no. 5, pp. 21–317, 1994.

[10] M. Treiber and A. Kesting, "Traffic flow dynamics," in *Traffic Flow Dynamics: Data, Models and Simulation*. Berlin, Germany: Springer, 2013, pp. 983–1000.

[11] B. Khondaker and L. Kattan, "Variable speed limit: An overview," *Transp. Lett.*, vol. 7, no. 5, pp. 264–278, Oct. 2015.

[12] M. Hadiuzzaman, T. Z. Qiu, and X.-Y. Lu, "Variable speed limit control design for relieving congestion caused by active bottlenecks," *J. Transp. Eng.*, vol. 139, no. 4, pp. 358–370, Apr. 2013.

[13] T. Tettamanti and I. Varga, "Distributed traffic control system based on model predictive control," *Periodica Polytechnica Civil Eng.*, vol. 54, no. 1, p. 3, 2010.

[14] A. Hegyi, S. P. Hoogendoorn, M. Schreuder, H. Stoelhorst, and F. Viti, "SPECIALIST: A dynamic speed limit control algorithm based on shock wave theory," in *Proc. 11th Int. IEEE Conf. Intell. Transp. Syst.*, Oct. 2008, pp. 827–832.

[15] A. Hegyi and S. P. Hoogendoorn, "Dynamic speed limit control to resolve shock waves on freeways–field test results of the SPECIALIST algorithm," in *Proc. 13th Int. IEEE Conf. Intell. Transp. Syst.*, Sep. 2010, pp. 519–524.

[16] S. M. Abdullah, M. Periyasamy, N. A. Kamaludeen, S. K. Towfek, R. Marappan, S. K. Raju, A. H. Alharbi, and D. S. Khafaga, "Optimizing traffic flow in smart cities: Soft GRU-based recurrent neural networks for enhanced congestion prediction using deep learning," *Sustainability*, vol. 15, no. 7, p. 5949, Mar. 2023.

[17] K. Kušic, I. Dusparic, M. Guériau, M. Greguric, and E. Ivanjko, "Extended variable speed limit control using multi-agent reinforcement learning," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2020, pp. 1–8.

[18] K. Kušic, E. Ivanjko, F. Vrbanic, M. Greguric, and I. Dusparic, "Dynamic variable speed limit zones allocation using distributed multi-agent reinforcement learning," in *Proc. IEEE Int. Intell. Transp. Syst. Conf. (ITSC)*, Sep. 2021, pp. 3238–3245.

[19] Z. Li, P. Liu, C. Xu, H. Duan, and W. Wang, "Reinforcement learning-based variable speed limit control strategy to reduce traffic congestion at freeway recurrent bottlenecks," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 11, pp. 3204–3217, Nov. 2017.

[20] X. Fang, T. Péter, and T. Tettamanti, "Variable speed limit control for the motorway–urban merging bottlenecks using multi-agent reinforcement learning," *Sustainability*, vol. 15, no. 14, p. 11464, Jul. 2023.

[21] S. Zheng, M. Li, Z. Ke, and Z. Li, "Coordinated variable speed limit control for consecutive bottlenecks on freeways using multiagent reinforcement learning," *J. Adv. Transp.*, vol. 2023, pp. 1–19, Jun. 2023.

[22] C. Wang, J. Zhang, L. Xu, L. Li, and B. Ran, "A new solution for freeway congestion: Cooperative speed limit control using distributed reinforcement learning," *IEEE Access*, vol. 7, pp. 41947–41957, 2019.

[23] Y. Zhang, M. Quinones-Grueiro, W. Barbour, Z. Zhang, J. Scherer, G. Biswas, and D. Work, "Cooperative multi-agent reinforcement learning for large scale variable speed limit control," in *Proc. IEEE Int. Conf. Smart Comput. (SMARTCOMP)*, Jun. 2023, pp. 149–156.

[24] K. Kušic, E. Ivanjko, and M. Greguric, "A comparison of different state representations for reinforcement learning based variable speed limit control," in *Proc. 26th Medit. Conf. Control Autom. (MED)*, Jun. 2018, pp. 1–6.

[25] Y. Wu, H. Tan, L. Qin, and B. Ran, "Differential variable speed limits control for freeway recurrent bottlenecks via deep actor-critic algorithm," *Transp. Res. C, Emerg. Technol.*, vol. 117, Aug. 2020, Art. no. 102649.

[26] M. Behrisch, L. Bieker, J. Erdmann, and D. Krajzewicz, "SUMO–simulation of urban mobility: An overview," in *Proc. SIMUL 3rd Int. Conf. Adv. Syst. Simulation*, 2011, pp. 1–6.

[27] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[28] B. Kővári, L. Szőke, T. Bécsi, S. Aradi, and P. Gáspár, "Traffic signal control via reinforcement learning for reducing global vehicle emission," *Sustainability*, vol. 13, no. 20, p. 11254, Oct. 2021. [Online]. Available: https://www.mdpi.com/2071-1050/13/20/11254

[29] V. Mnih, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, Feb. 2015.

[30] A. Nowé, P. Vrancx, and Y.-M. De Hauwere, "Game theory and multi-agent reinforcement learning," *Reinforcement Learning: State-of-the-Art*, pp. 441–470, 2012.
[31] L. Buşoniu, R. Babuška, and B. De Schutter, "Multi-agent reinforcement learning: An overview," in *Innovations in Multi-Agent Systems and Applications-1*. Berlin, Germany: Springer, 2010, pp. 183–221. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-14435-6_7#citeas
[32] B. H. K. Abed-Alguni, "Cooperative reinforcement learning for independent learners," Doctoral thesis, Dept. Comput. Sci., Univ. Newcastle, Callaghan, NSW, Australia, Oct. 2014.
[33] D. Krajzewicz, S. Hausberger, P. Wagner, M. Behrisch, and M. Krumnow, "Second generation of pollutant emission models for SUMO," in *Proc. SUMO 2nd SUMO User Conf.*, May 2014, pp. 203–221. [Online]. Available: https://elib.dlr.de/89398/

BÁLINT KŐVÁRI received the B.Sc. degree in vehicle engineering and the M.Sc. degree in autonomous vehicle control engineering from the Faculty of Transportation Engineering and Vehicle Engineering, Budapest University of Technology and Economics, Budapest, Hungary, in 2018 and 2020, respectively. He is currently pursuing the Ph.D. degree with the Budapest University of Technology and Economics. He is also the Machine Learning Team Leader of Asura Technologies Ltd. His research interests include artificial intelligence, reinforcement learning, computer vision, representation learning, and vehicle dynamics.

ISTVÁN GELLÉRT KNÁB received the B.Sc. degree in vehicle engineering from the Faculty of Transportation Engineering and Vehicle Engineering, Budapest University of Technology and Economics, where he is currently pursuing the M.Sc. degree in autonomous vehicle control engineering. He is also a Developer with the Systems and Control Laboratory, HUN-REN Institute for Computer Science and Control (SZTAKI). His research interests include machine learning, reinforcement learning, and intelligent transportation systems.

DOMOKOS ESZTERGÁR-KISS was a Fulbright Scholar with the University of California, Davis. He has been the International Project Coordinator of the Faculty of Transportation Engineering and Vehicle Engineering, since 2014. He is currently a Senior Lecturer with Budapest University of Technology and Economics (BME). He has published more than 50 articles in leading journals with Impact Factor. His research interests include the optimization of multimodal travel chains for passengers, the development of mobility as a service related solutions, and the establishment of workplace mobility plans for promoting sustainable commuting. He is a Council Member of AET. He is the main organizer of several international conferences (e.g., MTITS, 2015; EWGT, 2017; hEART, 2019; and TRA, 2026) and is involved in several Horizon 2020 projects, Interreg projects, and COST Actions (e.g., MoveCit, LinkingDanube, MaaS4EU, Electric traveling, BE OPEN, RegiaMobil, OJP4Danube, and metaCCAZE). He was the Chair of IEEE HS YP and is the Vice-President of ECTRI.

SZILÁRD ARADI (Member, IEEE) received the M.Sc. and Ph.D. degrees from Budapest University of Technology and Economics, Budapest, Hungary, in 2005 and 2015, respectively. He is currently with the Department of Control for Transportation and Vehicle Systems, Budapest University of Technology and Economics, where he has been a Senior Lecturer, since 2016. His research interests include embedded systems, communication networks, vehicle mechatronics, and reinforcement learning. His research and industrial works have involved railway information systems, vehicle on-board networks, and vehicle control.

TAMÁS BÉCSI (Member, IEEE) received the M.Sc. and Ph.D. degrees from Budapest University of Technology and Economics, Budapest, Hungary, in 2002 and 2008, respectively.

He has been an Assistant Lecturer and an Associate Professor with the Department of Control for Transportation and Vehicle Systems, Budapest University of Technology and Economics, since 2005 and since 2014, respectively. His research interests include machine learning, embedded systems, traffic modeling, and vehicle control.