# Imagination-Augmented Reinforcement Learning Framework for Variable Speed Limit Control

Duo Li[1,2], *Senior Member*, *IEEE*, Joan Lasenby[1]

***Abstract*** **– Variable Speed Limit (VSL) is a commonly applied active traffic management measure for urban motorways. In recent years, model-based and model-free approaches have been extensively adopted to solve VSL optimization problems. However, the success of model-based VSL relies heavily on the nature of the environmental model adopted (e.g., traffic flow model). Implicit environment models may result in inappropriate control actions. Although model-free approaches are able to directly map raw measurements to control actions without a need for an environment model, they usually require large amounts of training data. In order to address these issues, we propose an Imagination-Augmented Agent (I2A) for VSL control. The I2A consists an imagination path and a model-free path, which work together to generate appropriate control actions. The simulation results show that the proposed I2A agent outperforms other tested Reinforcement Learning (RL) agents in terms of Total Time Spent and bottleneck volume.**

***Index Terms*****—Motorway Control; Variable Speed Limit; Deep Reinforcement Learning**

## I. Introduction

TRAFFIC congestion is considered an urgent and growing challenge today. In Europe, the costs due to traffic congestion are estimated to be approximately 1% of annual Gross Domestic Product (GDP) [1], and road traffic emissions are responsible for 72% of total greenhouse gas emissions from the transport sector [2]. Nevertheless, increasing road capacity through extending existing infrastructure is usually associated with high costs and may induce additional traffic demand.

Nowadays, active traffic management measures have been widely used to improve traffic conditions. This improvement is made through better utilization of existing infrastructure. For motorway traffic management, two of the commonly used measures are ramp metering and Variable Speed Limit (VSL). Ramp metering prevents traffic congestion by regulating the flow of traffic entering motorways [3] [4]. The focus of this study is on VSL which is designed to improve traffic safety and reduce traffic congestion through a better harmonization of traffic flow. VSL was implemented, for the first time, in Germany more than four decades ago. Later, various rule-based control algorithms have been developed. The threshold parameters can be based on traffic volume, speed, occupancy, or a combination of the three [5] [6] [7] [8]. However, designing thresholds on an ad-hoc basis, commonly done in practice, does not fully utilize the potential of VSL. To address this issue, more complex rule-based algorithms have been designed, such as the SPECIALIST algorithm [9] which aims to resolve shock waves and was evaluated in a field test. In addition, a number of optimal-control-based VSL algorithms have been developed. One noteworthy example of these algorithms was proposed by Hegyi et al. [10]. They proposed a Model Predictive Control (MPC) framework to suppress shockwaves at motorway bottlenecks. The control objective was to minimize total travel time of all vehicles in the network. A general second-order traffic flow model METANET [11] was modified to incorporate the influence of VSL into the optimization process. Researchers continued the study of Hegyi et al. [10] by adopting various macroscopic traffic flow models (e.g., Cell Transmission Model [12]) and objective functions (e.g., emissions and crash risk [13] [14] [15] [16]). In recent years, Reinforcement Learning (RL) methods, that are able to rapidly adapt to new circumstances and “achieve goals in a wide range of environments” [17], have drawn increasing attention. Various RL algorithms, such as Q-Learning, Deep Q-Learning and Actor-Critic, have been successfully introduced to motorway traffic control [18] [19] [20] [21] [22] [23] [24].

As introduced above, currently the majority of research on VSL control algorithms concentrate on either optimal control or RL frameworks. Optimal-control-based algorithms require an explicit model of traffic flow dynamics. A traffic flow model has to be carefully calibrated for each road section to ensure optimal control. However, in the domain of RL-based VSL, the majority of literature uses model-free approaches, where raw observations directly map to actions. Model-free RL approaches are promising alternatives for modeling VSL and addressing the problem of requiring explicit traffic flow models. Nevertheless, model-free RL approaches usually require large amounts of training data and the resulting policies do not readily generalize to novel tasks in the same environment, as they lack the behavioral flexibility constitutive of general intelligence [25].

This study presents an Imagination-Augmented Agent (I2A) [25] for VSL control. The I2A consists of an imagination path and a model-free path, which work together to generate control actions. It should be noted that the definition and usage of the term ‘model’ can exhibit variations across different

[1] Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, United Kingdom

[2] Department of Engineering, Nottingham Trent University, Nottingham NG1 4FQ, United Kingdom

Corresponding author: Duo Li (email: dl655@cam.ac.uk)

academic disciplines. In the context of traffic control, a model of the environment generally refers to a mathematical or computational representation that characterizes the dynamics of traffic flow (e.g., METANET and Cell Transmission Model). In model-based reinforcement learning, a model of the environment can take on various forms, such as Markov decision processes (MDPs), Bayesian models, neural network models (in our case), and mathematical models, that represent the possible states, transitions, and rewards of the environment. Henceforth, the phrase 'environment model' will be used to denote its meaning in the context of reinforcement learning.

The main contributions of this research can be summarized as follows

- Conventional Optimal-control-based methods lacks the ability to deal with an imperfect environment model. We provide an end-to-end way to interpret and extract useful information from imperfect environment models, where real-time traffic states are converted into image-like input and processed by the I2A agent to generate appropriate control actions for VSL.
- We improve the performance of model-free VSL controllers by augmenting with imaginations. More specifically, the environment model generates imagined trajectories for all possible actions at each time step. These imagination trajectories are interpreted and then provided as additional context to a model-free agent for control action computation.
- I2A was first proposed in [25] for video game tasks and has since been applied to various fields, such as natural language processing [26] and robotics [27]. Our work is the first to apply I2A in the context of traffic control and introduce such novel and versatile architecture that can work with most existing environment models and RL methods. Our findings demonstrate the potential of I2A in this domain and fuel the development of model-based, model-free and hybrid algorithms for traffic control.

## II. Related Works

A number of RL algorithms have been proposed for solving VSL optimization problems. The most widely studied algorithm is the Q-Learning algorithm [28], which identifies an optimal action selection policy for any given Finite Markov Decision Process (FMDP) based on its Q-table. Various studies demonstrated formulations of VSL control as Q-Learning problems. For example, in [19], actions were described by a set of four speed limits {60, 80, 100, 120 km/h}. The state space was composed of two previous speed limits and real-time speeds at four consecutive sections in the vicinity of the congested area. The reward function was formulated as the minimization of Total Travel Time (TTT) with additional conditions. Q-Learning falters with increasing numbers of states/actions since the likelihood of an agent visiting a particular state and performing a particular action is increasingly small. One solution is to combine Q-Learning with function approximation. Several function approximators [20] [22] [19] were introduced to enhance Q-Learning VSL's capability of dealing with exponential growth of the solution space. An example of the k-Nearest Neighbors Temporal Difference (kNN-TD) algorithm presented in [20] has been successfully applied for Q-function approximation in [21]. Although Eligibility Traces based Reinforcement Learning (ETRL) [24] and Reinforcement-Markov Average Reward Technique (R-MART) [18] based VSL stated their benefits over the Q-Learning VSL, there is a lack of direct comparison between these controllers and the Q-Learning VSL controller.

Prior studies have shown that Deep Learning (DL) networks can be useful when there are a large number of state-action pairs in the RL model. The authors in [23] proposed a Deep RL (DRL) model-free framework for VSL under the automated vehicle environment. The proposed DRL VSL can directly change the speeds of automated vehicles within specific traffic lanes. A policy gradient method, Trust Region Policy Optimization (TRPO) [29], was used to optimize the parameters of the neural network. The proposed framework was tested using FLOW [30], which is a library for applying RL to automated vehicles in microscopic traffic simulators. Wu et al. [31] developed a Differential VSL (DVSL) system that can generate different speed limit values for each lane separately. An actor-critic architecture [32] was adopted to train the agent for DVSL, where the actor yields a speed limit, and the critic assesses the executed action of the actor. In their optimization problem, the number of action combinations is too large for Q-Learning or Deep Q-Learning [33]. For example, given a motorway section with five lanes and six speed limit options, the number of combinations can be $6^5$. Thus, they used Deep Deterministic Policy Gradient (DDPG) [34] to optimize parameters in the actor-critic architecture. In addition, efforts have been made to enhance the performance of RL-based VSL control. For example, in [35], the transfer learning algorithm was used to enhance the transferability of RL-based VSL control. Han et al., [36] employed an iterative training framework to alleviate model mismatch through online/offline learning.

A few studies investigated multi-agent RL VSL control algorithms. In [37], a W-Learning VSL was proposed based on the idea that there is no need for any global controller. In the study, two agents were trained using the W-Learning algorithm [38] to jointly control two motorway sections upstream of a congested area. The reward function for each agent was only based on its local performance and there was no additional communication between agents. In [39], a distributed RL approach was proposed to improve motorway mobility and safety under the Vehicle-to-Infrastructure (V2I) environment. The coordinated VSL agents were trained by a Deep Q-Learning framework. As presented in [20], a hierarchical multi-agent RL framework was adopted for coordinated ramp metering and VSL control. In the framework, an agent with a higher hierarchy is the first one to take an action, and then the second agent receives this action and determines its own action based on the received action.

The topic about how to mimic the human imagination and to benefit the industries is open and studied by a variety of researchers. For example, Hafner et al. [40] presented a agent, Dreamer that learns long-horizon behaviors purely by latent imagination to solve visual control tasks. In [41], semantic labels are processed by the imagination model to produce additional data for the enhancement of facial expression recognition. Imagination augmented models are also used for mobile robots [42] and natural language understanding [26]. The I2A architecture adopted in this study have been used in different fields, such as spoken dialogue system [43], video games [44] and robotic applications [27], but have not been introduced to the field of transportation yet.

In this section we have given an overview of RL VSL studies; a more comprehensive survey can be found in [45]. In general, progress has been made in developing effective agents for VSL control using model-free RL frameworks in conjunction with DL neural networks. However, as model-free approaches purely sample from experience, they lack some form of planning and require large amounts of training data in order to reach a level of expected performance.

## III. VSL Control Problem

Fig. 1 demonstrates a typical VSL control example. Note that we use different colors to draw detectors simply to distinguish their locations from one another. The colors do not carry any additional meaning beyond this purpose. To ensure consistency, detectors with the same color indicate the same location in the following figures. In the merging area, the interference between on-ramp and mainline traffic causes speed reductions, leading to the formation of a merging bottleneck. Traffic congestion occurs when the demand volume exceeds the bottleneck capacity. The purpose of applying VSL is to limit the number of upstream vehicles entering the bottleneck, and therefore relieve merging difficulties and keep bottleneck traffic operating near its capacity. As such, traffic congestion can be delayed or even prevented [10].

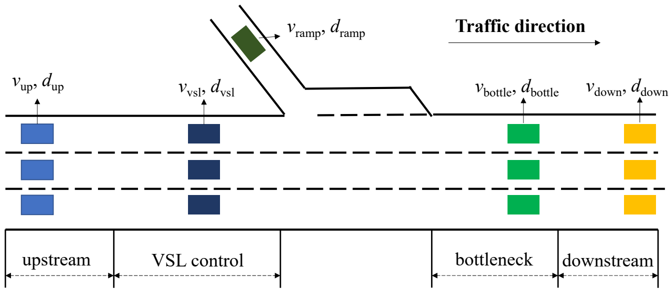

$v_{\text{ramp}}$, $d_{\text{ramp}}$
Traffic direction
$v_{\text{up}}$, $d_{\text{up}}$
$v_{\text{vsl}}$, $d_{\text{vsl}}$
$v_{\text{bottle}}$, $d_{\text{bottle}}$
$v_{\text{down}}$, $d_{\text{down}}$
upstream
VSL control
bottleneck
downstream

Fig. 1: VSL control example (detectors are shown as colored blocks)

To prepare VSL control for an RL setting, a Markov Decision Process (MDP) should be defined. In particular, at each time step $k$, the motorway section is in a certain state $s(k)$, and the agent may choose an action $a(k)$ that is available in the current state. The motorway section responds at the next time step $k+1$ by transitioning into a new state $s(k+1)$ and returning the agent corresponding reward $r(k+1)$. In this study, the state space $S$, action set $A$ and reward function $r$ are defined as follows:

*State Space*

States are used to reflect real-time traffic condition, which can be any available traffic measurement. Considering the complexity of traffic flow dynamics, we build the state space $S$ using the volume $v_{\text{up}}$ and density $d_{\text{up}}$ at the upstream mainline section, volume $v_{\text{vsl}}$ and density $d_{\text{vsl}}$ at the VSL controlled section, volume $v_{\text{bottle}}$ and density $d_{\text{bottle}}$ at the bottleneck section, volume $v_{\text{down}}$ and density $d_{\text{down}}$ at the downstream section, and volume $v_{\text{ramp}}$ and density $d_{\text{ramp}}$ on the on-ramp section. Detector locations are illustrated in Fig. 1.

*Action Set*

VSL control is realized by adjusting the speed limit posted on Variable Message Signs (VMSs). Considering the user acceptance issue, speed limit values should be discrete with a proper increment. In this study, the action set $A$ contains eleven elements ranging from 20 to 70 mph with a 5 mph increment.

*Reward Function*

Total Time Spent (TTS) is commonly used to reflect mobility performance of a network, which can be expressed as

$$\mathrm{TTS}(k) = T\sum_{k=1}^{K} N(k) \tag{1}$$

where, $T$ is the time interval, $K$ is the total number of time steps, $N(k)$ is the total number of vehicles in the network (both the mainline and on-ramp) at time $k$. This study also takes into account vehicles waiting to enter the network when calculating TTS. In this study, the reward function is formulated as:

$$r(k) = \begin{cases} [-\mathrm{TTS}(k) - p]/\mathrm{TTS}_{\mathrm{c}}, & \text{if } a(k) - a(k-1) \geq 10 \text{ mph} \\ -\mathrm{TTS}(k)/\mathrm{TTS}_{\mathrm{c}}, & \text{otherwise} \end{cases} \tag{2}$$

with

$$\mathrm{TTS}(k) = T\sum_{k=1}^{K}\sum_{m=1}^{M} d_m(k) L_m N_{\mathrm{ln},m} \tag{3}$$

$$\mathrm{TTS}_{\mathrm{c}} = T\sum_{k=1}^{K}\sum_{m=1}^{M} d_{\mathrm{c}} L_m N_{\mathrm{ln},m} \tag{4}$$

$$p = \lambda_p \mathrm{TTS}_{\mathrm{c}} \tag{5}$$

where, $d_m(k)$ is the density (veh/km/lane) of the $m^{th}$ section at time $k$; $L_m$ is the length of the $m^{th}$ section; $N_{\mathrm{ln},m}$ is the number of lanes of the $m^{th}$ section ; $\mathrm{TTS}_{\mathrm{c}}$ is the TTS under congested traffic condition; $d_{\mathrm{c}}$ is the critical density;

$p$ is the penalty for sudden changes in speed limits; and $\lambda_p$ is the scaling parameter used to adjust the magnitude of the penalty. Note that sudden changes in speed limits that could be potentially dangerous, which may result in rear-end collisions. More details about agent settings can be found in Appendix D.

## IV. Imagination-Augmented Agent (I2A)

I2A [25] is a novel RL architecture making use of an approximate environment model to embed imagined trajectories in the policy learning of a model-free agent. In this section, we will illustrate the I2A framework in a top-down manner. We will begin by presenting its high-level architecture, and then we will explain its two paths, which can also be viewed as two distinct blocks or parts that process the input individually. Finally, we will delve into the details of each path. Fig. 2a illustrates the high-level architecture of I2A which consists of a model-free path, an imagination path and a policy module. The model-free and imagination paths are used to process the input state. Then, the policy module receives the information from both paths and produces the policy $\pi$ and estimated value $V$.

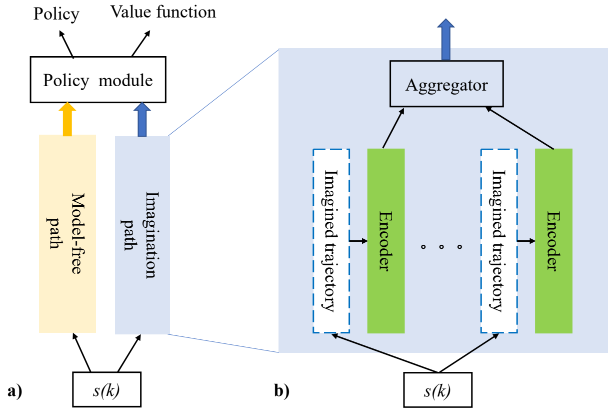


Fig. 2: I2A architecture: a) high-level architecture and b) imagination path

As depicted in Fig. 2b, the imagination path consists of a set of rollout encoders and an aggregator that converts rollout embeddings into a single imagination code. Each encoder is responsible for one imagined trajectory $\tilde{\tau}$. At each time step, the imagination path generates 11 imagined trajectories, each corresponding to one of the 11 VSL actions. Each trajectory includes predicted actions, states, and rewards for multiple future time steps. The process of generating imagined trajectories is described in detail in the following paragraph. With the help of imagined trajectories, the agent can know how the existing policy will effect future performance and learn to interpret information from these imaginations when selecting actions. Although these trajectories may contain information beyond reward sequence or even yield unexpected results, they are still informative. As they can cover various situations that the agent will confront in the future, which increases the chance of getting higher reward.

Fig. 3 demonstrates how the imagined trajectories are generated. This process is on the basis of an environment model that predicts future states, and a rollout policy $\tilde{\pi}$ that generates corresponding actions. Given a state-action pair, the environment model can predict the next state and reward. This process rolls out over multiple time steps into the future and produces a set of trajectories by initializing the trajectory with the current observation. Note that the rollout policy $\tilde{\pi}$ here is different from the policy $\pi$ mentioned earlier. More detailed explanations about imagined trajectories can be found in Appendix C. In the following subsections, the detailed descriptions of I2A components are given.

### A. Environment Model

An environment model can be any recurrent architecture. In this study, we build a Convolutional Neural Network (CNN) [46] based environment model, as shown in Fig. 4. To fit the input format of CNN layers, the traffic state is represented as an image-like shape with two channels (density and volume), and the action is one-hot encoded and broadcasted into a corresponding image. The state and action images are concatenated and fed into a CNN block containing two CNN layers: $conv\_1$ and $conv\_2$. Then, a CNN layer $conv\_out$ is added to output a prediction of the next state $s(k+1)$. The corresponding reward $r(k+1)$ can be computed via Eq.2. The CNN network is optimized based on Mean Squared Error (MSE) between model predictions and real observations. It should be noted that I2A allows pitfalls within the environment model and learns to extract useful knowledge gathered from imperfect predictions.

### B. Standard Model-Free Agent

Any model-free RL agent can be augmented with imaginations. In this study, we use an Advantage Actor Critic (A2C) [47] architecture that is a synchronous, deterministic variant of Asynchronous Actor-Critic Agents (A3C) [47]. In this subsection, a brief description of standard A2C is given. The next subsection explains how to augment the standard A2C with imaginations. Note that the standard A2C presented here is also used as a baseline algorithm to verify the effectiveness of the proposed I2A in the case study section.

In RL, two typical categories of methods are 1) value-based methods (e.g., Deep Q-learning) which maps each state-action pair to a value by learning a value function, and 2) policy-based methods (e.g., policy gradient [48]) that optimizes the policy without using a value function. As illustrated in Fig. 5a, A2C combines the value-based and policy-based methods through a "critic" network and an "actor" network. The actor (policy-based) network generates a control action, and the critic (value-based) network evaluates the selected action. In this study, a CNN block with the same structure to the one in the environment model takes current state $s(k)$ as input, followed by a Fully Connected (FC) layer $fc\_1$. This FC layer feeds into two heads: into a FC layer $fc\_2$ computes the value function $V(s(k);\theta_V)$, and into another FC layer $fc\_2$ that generates the policy logits $\log\pi(a(k)|s(k),\theta)$. The

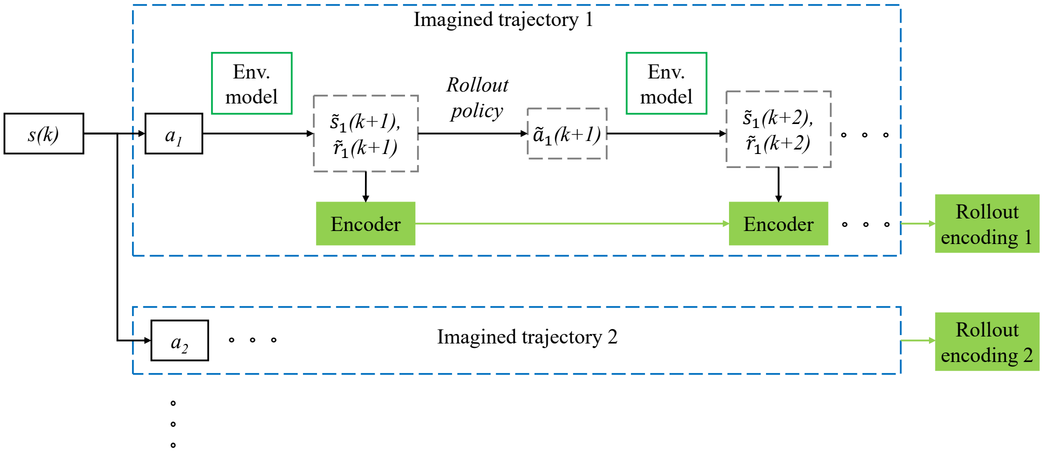


Fig. 3: Imagined trajectory generation and encoding

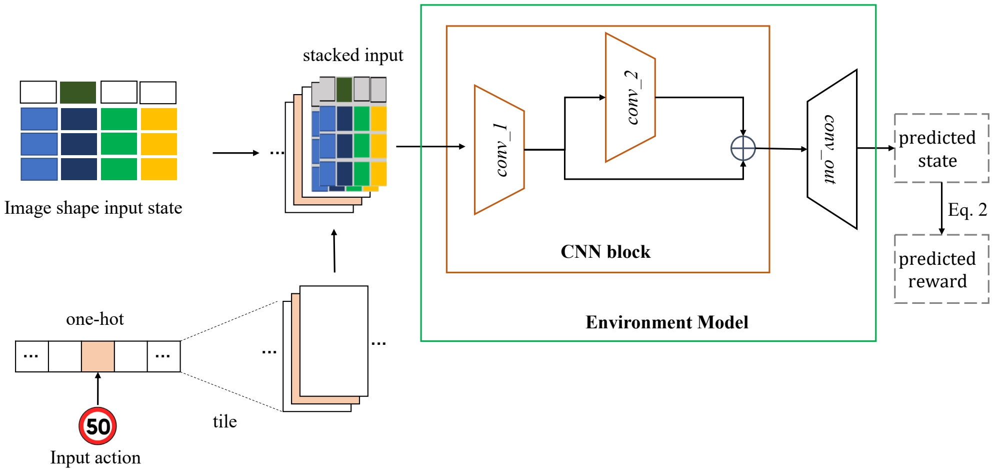


Fig. 4: Environment model

$\theta$ and $\theta_V$ are the parameters of the actor and critic networks, respectively.

### C. Combining Model-Free and Imagination Paths

Fig. 5b shows how to embed imaginations in the policy learning of the standard A2C agent. On the imagination path, an encoder concatenates predicted states and rewards from the environment model, and encodes them via a Long short-term memory (LSTM) [49] network (see Fig. 3). This process is repeated for all eleven rollouts (one per action). The last output of the LSTM for all rollouts are concatenated into a single vector $c_{\text{im}}$. Then, this feature vector is concatenated with the output $c_{\text{free}}$ of the model-free path and is passed into the FC layers to calculate value function $V(s(k);\theta_V)$ and policy logits $\log \pi(a(k)|s(k),\theta)$.

### D. Agent Training and Rollout Policy Distillation

For the standard A2C and the proposed I2A described above, the policy logits $\log \pi(a(k)|s(k),\theta)$ are generated by the policy network with parameters $\theta$. During training, we update the parameters $\theta$ using policy gradient $g(\theta)$:

$$g(\theta) = \nabla_{\theta \log \pi(a(k)|s(k),\theta)} \mathrm{Adv}(s(k),a(k)) \tag{6}$$

where, $\mathrm{Adv}(s(k),a(k))$ is an advantage function. $\mathrm{Adv}(s(k),a(k))$ is computed as the difference between the return $\mathrm{RT}(k)$ received by the agent and and the estimated value $V(s(k);\theta_V)$:

$$\mathrm{Adv}(s(k),a(k)) = \mathrm{RT}(k) - V(s(k);\theta_V) \tag{7}$$

Here, $V(s(k);\theta_V)$ is estimated by the value network with parameters $\theta_V$. The update equitation of the value network is given by:

$$g(\theta_V) = -\mathrm{Adv}(s(k),a(k)\partial_{\theta_V} V(s(k);\theta_V)) \tag{8}$$

As mentioned at the beginning of this section, a shared rollout policy $\tilde{\pi}$ is needed for action generation on the imagination path. After testing different types of rollout policies (e.g., random, pretrained), the authors in [25] suggested a distillation strategy. The policy distillation is realized by adding a cross entropy auxiliary loss between the imagination-augmented policy $\pi$ and the rollout policy $\tilde{\pi}$:

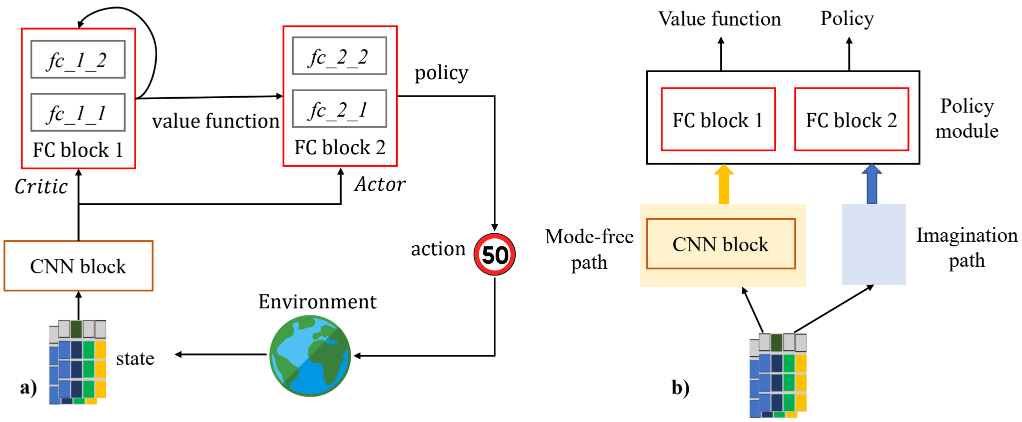


Fig. 5: a) Standard A2C, and b) A2C augmented with imaginations

$$l_{\text{dist}} = \lambda_{\text{dist}} \sum_{a} \pi(a|s(k)) \log \tilde{\pi}(a|s(k)) \qquad (9)$$

with scaling parameter $\lambda_{\text{dist}}$. In addition, an entropy regularizer is included to encourage exploration:

$$l_{\text{env}} = \lambda_{\text{env}} \sum_{a(k)} \pi(a(k)|s(k);\theta) \log \pi(a(k)|s(k);\theta) \qquad (10)$$

with the parameter $\lambda_{\text{env}}$ thoughout all experiments.

## V. Evaluation

In this study, a 2.8-km stretch around the Junction 12 of the motorway M25 (also known as the London Orbital Motorway) in the UK was selected as the test bed (see Fig. 6). The traffic data was obtained from Highways England[1], which provides traffic volume and average speed within the 15-min time slice. Preliminary analysis showed that the highest traffic demands were in June. Therefore, 6:30-9:00 AM on June 10, 2019, a typical weekday peak period, was chosen as the simulation period.

We simulated the selected motorway stretch using the SUMO micro-simulator [50]. The simulation model was calibrated against the data collected on June 10, 2019 and then validated against the data on June 17, 2019 using the GEH (Geoffrey E. Havers) index [51]. As depicted in Fig. 6, input to the I2A agent was converted into image shape where on-ramp measurements were averaged. Model parameters are shown in Table 1. The parameters were modified based on the original I2A model in [25]. In the original model, the input of size 15x19x3 was processed by CNNs with a 3x3 kernel, while CNNs with a 2x2 kernel were used to process the input of size 4x4x2 in this study. In addition, the FC layer $fc\_3$ with 11 units were used to handle 11 VSL actions in our case instead of the FC layer with 5 units in the original model. Each motorway section is 500 meters, which was chosen based on the distance traveled by vehicles at 70 mph in one detection interval of 15 seconds. Note that the control interval is 60 seconds, which takes means of four 15-second measurements as input.

TABLE I: Parameters of different layers in I2A

| Layer | | Configuration |
|---|---|---|
| CNN block | *conv_1* | 2 × 2 kernels, 32 output channels |
| | *conv_2* | 2 × 2 kernels, 32 output channels |
| *conv_out* | | 2 × 2 kernels, 2 output channels |
| *fc_1* | | 256 units |
| *fc_2* | | 1 unit |
| *fc_3* | | 11 units |
| LSTM | | 256 units |

The RMSprop optimizer [52] with learning rate=0.001 was used for agent training. The I2A agent was trained on 300 episodes of simulation, each lasting 2.5 hours from 6:30am to 9:00am. During each simulation, the demand values were randomly sampled between 85% and 115% of a demand profile randomly selected from 10 weekdays in June 2019. Our CNN-based environment model was jointly trained with the I2A agent by adding a negative loglikelihood loss to the total loss as an auxiliary loss. The simulation results presented in the following paragraphs are the means of 10 simulation runs.

In order to verify the effectiveness of the I2A agent, we compared the proposed agent against two extensively used model-free RL agents, namely

- **Deep Q-Learning (DQL)** that has the same state space and action set to the I2A agent; the DQL agent consists of a CNN block and two FC layers that are the same as the CNN block, $fc\_1$ and $fc\_3$ descibed in Table 1;
- **A2C** that has the same structure to the one described in the section 4.2.

Fig. 7 shows the TTS under different control scenarios. Here, the no-control scenario serves as the baseline without any VSL control. Fig. 8 displays the traffic volumes at the motorway bottleneck under different control scenarios. Combining the information from both figures, a typical morning peak period was observed in the no-control scenario: the traffic volume at the bottleneck surged after 7:00 and peaked around 15 minutes later. The capacity drop occurred shortly after the

[1] http://tris.highwaysengland.co.uk/detail/monthlysummarydata

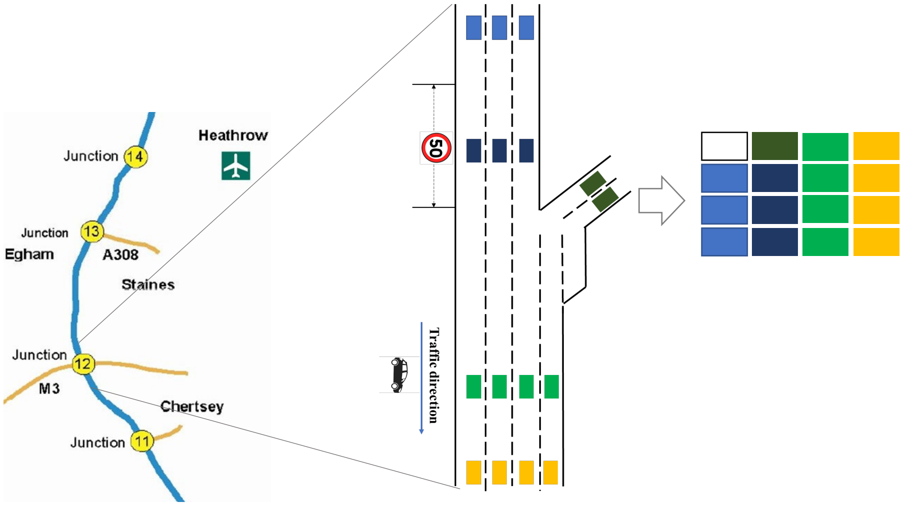


Fig. 6: Study area: Junction 12 of M25

formation of congestion, resulting in increased TTS. This severe deterioration in TTS lasted over an hour. As seen in Fig 8, with the RL-based VSL control, the capacity drop was prevented and the bottleneck volume was increased, leading to reduced congestion duration and TTS displayed in Fig 7. More specifically, the DQL and A2C VSL agents recorded 9.7% and 8.9% improvements in TTS respectively, compared with the no-control scenario. When the A2C agent was augmented with imaginations, a remarkable improvement in TTS was witnessed, which was 11.9% compared with the TTS (538 veh · h) in the no-control scenario. These TTS reductions could be mainly attributed to the increased traffic volumes at the bottleneck section. The DQL, A2C and I2A increased the average volume at the bottleneck by 7.9% 7.5% and 9.0%, respectively, compared against the average volume (6930 veh/h) in the no-control scenario. These improvements proved the I2A-based VSL's capacity of ameliorating traffic condition and increasing motorway productivity.

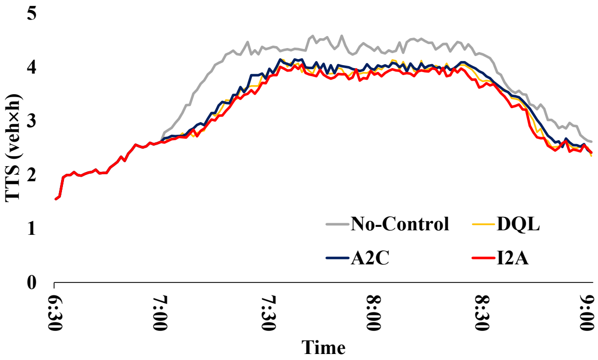


Fig. 7: TTS for different control scenarios

Fig. 9 depicts speed limits generated by different RL algorithms. The speed limit profile generated by the I2A was more stable than those generated by the A2C and DQL agents, which might help to reduce over- or under-control cases, where the system is either over-regulated or under-regulated, leading to undesirable outcomes. As shown in the reward function, we included a penalty to prevent sudden changes in speed limits. In this study, we used a penalty scaling parameter $\lambda_p$=5. We tested the effect of this penalty and the simulation results showed that the I2A without the penalty produced sudden speed limit changes, for example, the speed limit dropped from 60 mph to 40 mph around 7:05 am in response to congestion. However, using I2A with the penalty ensured that the difference in speed limits between two consecutive time steps never exceeded 10 mph.

Fig. 10 depicts density contour plots for the tested RL agents. It is observed that traffic congestion mainly occurred on the merging area and propagated to the upstream sections. The red and yellow spots (representing high traffic density) reduced significantly for VSL scenarios when compared with the no-control case, representing significant improvement in traffic condition near the merging area. This improvement might be attributed to reduced inflow to the merging area: slightly higher density on the VSL controlled area was witnessed in VSL cases, indicating more evenly distributed traffic flow.

The driver's compliance with speed limits plays a vital role in the success of VSL control. Therefore, we tested the performance of the proposed I2A VSL at different compliance levels. The simulation results presented above were generated using 100% driver compliance rate. The performance of the I2A decreased as the compliance rate to the posted speed limits decreased. Specifically, the TTS computed using the I2A VSL increased from 473 veh · h (100% compliance) to 482 veh · h at 80% compliance rate and 506 veh · h at 60% compliance rate.

## VI. Conclusions

Most of the existing studies adopt model-based or model-free approaches to solve the VSL optimization problem. However, the success of model-based VSL highly relies on the employed traffic flow model. Implicit traffic flow models may result in inappropriate control actions (see Appendix

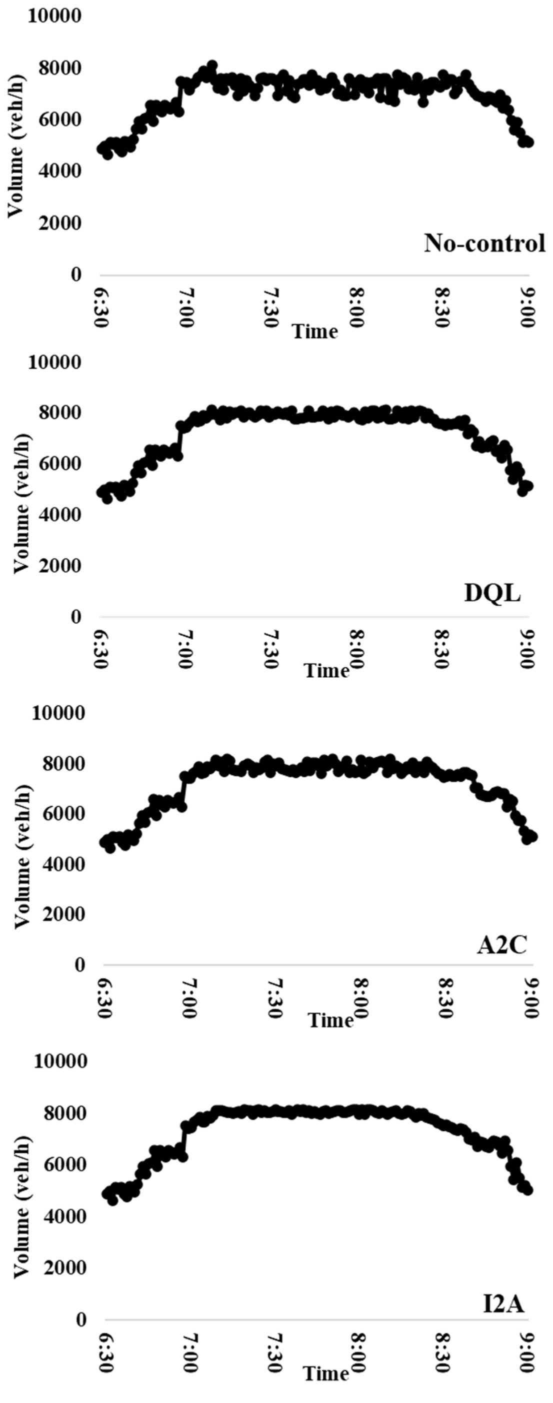


Fig. 8: Bottleneck traffic volumes under different control scenarios

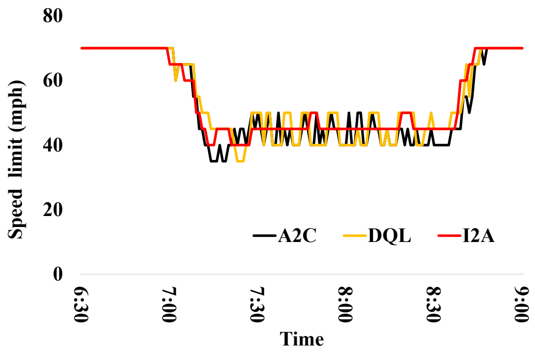


Fig. 9: Speed limits generated by different algorithms

A). Although model-free RL approaches are able to directly map raw measurements to control actions without a need for traffic flow model, they usually require large amounts of training data. In order to address these issues, we introduced an Imagination-Augmented Agent (I2A) for VSL control. The I2A consists an imagination path and a model-free path. Imagination trajectories produced on the imagination path and output of the model-free path are combined to generate control actions. The proposed I2A-based VSL was assessed for a critical bottleneck of the motorway M25 in the UK using the SUMO micro-simulator and verified against two well-known RL algorithms. The amount of training data required by the proposed agent is at the same level of most optimal-control-based methods (see Appendix B). The following conclusions can be drawn from the simulation results:

- The I2A-based VSL shows the capacity of delaying and relieving traffic congestion, ameliorating traffic conditions and increasing motorway productivity.
- The I2A-based VSL outperforms two well-known model-free RL approaches, namely, Deep Q-Learning and Advantage Actor Critic in terms of TTS and bottleneck volume.
- The I2A architecture offers the promising potential to improve the performance of a model-free agent by augmenting it with imaginations.

In this study, only the combination of CNN layers, FC layers, LSTM layers, and A2C approach was tested for the I2A-based VSL. One advantage of I2A architecture is its flexibility. In future research, we will explore a variety of combinations to further improve the performance of I2A agent. For example, conventional macroscopic traffic flow models, such as the METANET model and the cell transmission model, can be used as environment models; transformer models [53] can be introduced to encode imagination trajectories; and other model-free RL approaches, such as Proximal Policy Optimization (PPO) [54] and Soft Actor-Critic (SAC) [55], can be augmented with imaginations. It is worthy noting that the RL algorithm provides a solution that is optimized based on the available information, but it may not necessarily represent the absolute optimal solution in all circumstances.

The proposed agent was only trained and evaluated based on a merge bottleneck. However, motorway congestion may result from various factors, such as accidents, adverse weather conditions, work zones, slow-moving/breakdown vehicles, and bottlenecks due to merging/diverging traffic, lane drops, and grade changes. In future research, the effectiveness of the proposed I2A agent will be verified against a range of different locations and conditions. In Fig. 10, the reduced density with VSL control suggests a more evenly distributed traffic flow and less congestion, potentially leading to a safer driving experience by allowing for smoother driving and less braking. Our current study focuses on enhancing the mobility performance of VSL control. However, we recognize the significance of

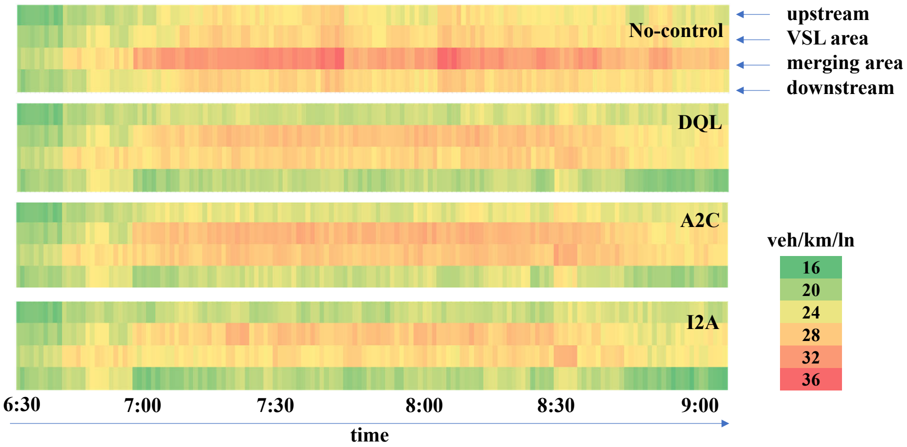


Fig. 10: Density contour plots for different control scenarios

safety considerations and, therefore, plan to design and test the proposed RL agent for safety improvement in our future work. Moreover, the I2A VSL controller proposed in this study is only implemented on a local level. Attempts will be made to develop an multi-agent I2A framework for coordinated VSL control.

## Acknowledgements

This study was sponsored by the Engineering and Physical Sciences Research Council (EPSRC) (Project No.EP/R035199/1).

## References

[1] P. Christidis, J. N. I. Rivas *et al.*, "Measuring road congestion," *Institute for Prospective and Technological Studies, Joint Research Centre, Brussels*, 2012.

[2] E. E. Agency, "Quality of transport: Eb82.2 directorate-general for communication," Copenhagen, Denmark, Tech. Rep., 2019.

[3] M. Papageorgiou and A. Kotsialos, "Freeway ramp metering: An overview," *IEEE transactions on intelligent transportation systems*, vol. 3, no. 4, pp. 271–281, 2002.

[4] D. Li, P. Ranjitkar, and Y. Zhao, "Efficiency and equity performance of a coordinated ramp metering algorithm," *Promet-Traffic&Transportation*, vol. 28, no. 5, pp. 507–515, 2016.

[5] P. Allaby, B. Hellinga, and M. Bullock, "Variable speed limits: Safety and operational impacts of a candidate control strategy for freeway applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 4, pp. 671–680, 2007.

[6] L. Elefteriadou, S. S. Washburn, Y. Yin, V. Modi, C. Letter *et al.*, "Variable speed limit (vsl)-best management practice," University of Florida. Transportation Research Center, Tech. Rep., 2012.

[7] D. Li, P. Ranjitkar, and A. Ceder, "Integrated approach combining ramp metering and variable speed limits to improve motorway performance," *Transportation Research Record*, vol. 2470, no. 1, pp. 86–94, 2014.

[8] D. Li and P. Ranjitkar, "A fuzzy logic-based variable speed limit controller," *Journal of Advanced Transportation*, vol. 49, no. 8, pp. 913–927, 2015.

[9] A. Hegyi and S. P. Hoogendoorn, "Dynamic speed limit control to resolve shock waves on freeways-field test results of the specialist algorithm," in *13th International IEEE Conference on Intelligent Transportation Systems*. IEEE, 2010, pp. 519–524.

[10] A. Hegyi, B. De Schutter, and H. Hellendoorn, "Model predictive control for optimal coordination of ramp metering and variable speed limits," *Transportation Research Part C: Emerging Technologies*, vol. 13, no. 3, pp. 185–209, 2005.

[11] A. Messmer and M. Papageorgiou, "Metanet: A macroscopic simulation program for motorway networks," *Traffic Engineering Control*, vol. 31, pp. 466–470, 1990.

[12] P. Mao, X. Ji, X. Qu, L. Li, and B. Ran, "A variable speed limit control based on variable cell transmission model in the connecting traffic environment," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 17 632–17 643, 2022.

[13] Y. Han, A. Hegyi, Y. Yuan, S. Hoogendoorn, M. Papageorgiou, and C. Roncoli, "Resolving freeway jam waves by discrete first-order model-based predictive control of variable speed limits," *Transportation Research Part C: Emerging Technologies*, vol. 77, pp. 405–420, 2017.

[14] R. Yu and M. Abdel-Aty, "An optimal variable speed limits system to ameliorate traffic safety risk," *Transportation research part C: emerging technologies*, vol. 46, pp. 235–246, 2014.

[15] M. Hadiuzzaman and T. Z. Qiu, "Cell transmission model based variable speed limit control for freeways," *Canadian Journal of Civil Engineering*, vol. 40, no. 1, pp. 46–56, 2013.

[16] D. Li, X. Zhao, and P. Cao, "An enhanced motorway control system for mixed manual/automated traffic flow," *IEEE Systems Journal*, 2020.

[17] S. Legg and M. Hutter, "Universal intelligence: A definition of machine intelligence," *Minds and machines*, vol. 17, no. 4, pp. 391–444, 2007.

[18] F. Zhu and S. V. Ukkusuri, "Accounting for dynamic speed limit control in a stochastic traffic environment: A reinforcement learning approach," *Transportation research part C: emerging technologies*, vol. 41, pp. 30–47, 2014.

[19] E. Walraven, M. T. Spaan, and B. Bakker, "Traffic flow optimization: A reinforcement learning approach," *Engineering Applications of Artificial Intelligence*, vol. 52, pp. 203–212, 2016.

[20] T. Schmidt-Dumont and J. van Vuuren, "A case for the adoption of decentralised reinforcement learning for the control of traffic flow on south african highways," *Journal of the South African Institution of Civil Engineering*, vol. 61, no. 3, pp. 7–19, 2019.

[21] 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 transactions on intelligent transportation systems*, vol. 18, no. 11, pp. 3204–3217, 2017.

[22] K. Kušić, E. Ivanjko, and M. Gregurić, "A comparison of different state representations for reinforcement learning based variable speed limit control," in *2018 26th Mediterranean Conference on Control and Automation (MED)*. IEEE, 2018, pp. 1–6.

[23] E. Vinitsky, K. Parvate, A. Kreidieh, C. Wu, and A. Bayen, "Lagrangian control through deep-rl: Applications to bottleneck decongestion," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 759–765.

[24] S. S. T. Aval, N. S. Ghandeshtani, P. Akbari, N. Eghbal, and A. Noori, "An eligibility traces based cooperative and integrated control strategy for traffic flow control in freeways," in *2019 9th International Conference on Computer and Knowledge Engineering (ICCKE)*. IEEE, 2019, pp. 40–45.

[25] S. Racanière, T. Weber, D. Reichert, L. Buesing, A. Guez, D. J. Rezende, A. P. Badia, O. Vinyals, N. Heess, Y. Li *et al.*, "Imagination-augmented agents for deep reinforcement learning," in *Advances in neural information processing systems*, 2017, pp. 5690–5701.

[26] Y. Lu, W. Zhu, X. E. Wang, M. Eckstein, and W. Y. Wang, "Imagination-augmented natural language understanding," *arXiv preprint arXiv:2204.08535*, 2022.

[27] M. Thabet, *Imagination-Augmented Deep Reinforcement Learning for Robotic Applications*. The University of Manchester (United Kingdom), 2022.

[28] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[29] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International conference on machine learning*, 2015, pp. 1889–1897.

[30] C. Wu, A. Kreidieh, K. Parvate, E. Vinitsky, and A. M. Bayen, "Flow: Architecture and benchmarking for reinforcement learning in traffic control," *arXiv preprint arXiv:1710.05465*, p. 10, 2017.

[31] Y. Wu, H. Tan, L. Qin, and B. Ran, "Differential variable speed limits control for freeway recurrent bottlenecks via deep actor-critic algorithm," *Transportation research part C: emerging technologies*, vol. 117, p. 102649, 2020.

[32] A. G. Barto, R. S. Sutton, and C. W. Anderson, "Neuronlike adaptive elements that can solve difficult learning control problems," *IEEE transactions on systems, man, and cybernetics*, no. 5, pp. 834–846, 1983.

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

[34] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[35] Z. Ke, Z. Li, Z. Cao, and P. Liu, "Enhancing transferability of deep reinforcement learning-based variable speed limit control using transfer learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4684–4695, 2020.

[36] Y. Han, A. Hegyi, L. Zhang, Z. He, E. Chung, and P. Liu, "A new reinforcement learning-based variable speed limit control approach to improve traffic efficiency against freeway jam waves," *Transportation research part C: emerging technologies*, vol. 144, p. 103900, 2022.

[37] K. Kušić, I. Dusparic, M. Guériau, M. Gregurić, and E. Ivanjko, "Extended variable speed limit control using multi-agent reinforcement learning," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2020, pp. 1–8.

[38] M. Humphrys, "Action selection methods using reinforcement learning," *From Animals to Animats*, vol. 4, pp. 135–144, 1996.

[39] 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. 41 947–41 957, 2019.

[40] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," *arXiv preprint arXiv:1912.01603*, 2019.

[41] N. Churamani and H. Gunes, "Clifer: Continual learning with imagination for facial expression recognition," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 322–328.

[42] Z. Shen, L. Kästner, and J. Lambrecht, "Spatial imagination with semantic cognition for mobile robots," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 2174–2180.

[43] Y.-C. Wu, B.-H. Tseng, and M. Gasic, "Actor-double-critic: Incorporating model-based critic for task-oriented dialogue systems," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 854–863.

[44] C.-V. Pal and F. Leon, "A modified i2a agent for learning in a stochastic environment," in *Computational Collective Intelligence: 12th International Conference, ICCCI 2020, Da Nang, Vietnam, November 30–December 3, 2020, Proceedings 12*. Springer, 2020, pp. 388–399.

[45] K. Kušić, E. Ivanjko, M. Gregurić, and M. Miletić, "An overview of reinforcement learning methods for variable speed limit control," *Applied Sciences*, vol. 10, no. 14, p. 4917, 2020.

[46] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," *arXiv preprint arXiv:1404.2188*, 2014.

[47] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *International conference on machine learning*, 2016, pp. 1928–1937.

[48] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Advances in neural information processing systems*, 2000, pp. 1057–1063.

[49] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," 1999.

[50] D. Krajzewicz, J. Erdmann, M. Behrisch, and L. Bieker, "Recent development and applications of sumo-simulation of urban mobility," *International journal on advances in systems and measurements*, vol. 5, no. 3&4, 2012.

[51] R. Dowling, A. Skabardonis, J. Halkias, G. McHale, and G. Zammit, "Guidelines for calibration of microsimulation models: framework and applications," *Transportation Research Record*, vol. 1876, no. 1, pp. 1–9, 2004.

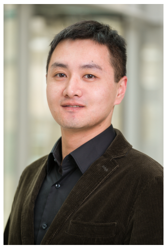
[52] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.

[53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

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

[55] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel *et al.*, "Soft actor-critic algorithms and applications," *arXiv preprint arXiv:1812.05905*, 2018.

[56] M. Gregurić, K. Kušić, and E. Ivanjko, "Impact of deep reinforcement learning on variable speed limit strategies in connected vehicles environments," *Engineering Applications of Artificial Intelligence*, vol. 112, p. 104850, 2022.



**Duo Li** (Senior Member, IEEE) received his PhD, M.S. and B.E. Degrees in Civil Engineering (Transportation) from the University of Auckland in 2015, the University of Queensland in 2011, and the Huazhong University of Science and Technology in 2010, respectively. Since 2022, he has been a Senior Lecturer with the Department of Engineering, Nottingham Trent University. Before this, he had several academic and research positions inculding Research Associate at the University of Cambridge, Humboldt Research Fellow at the German Aerospace Center (DLR), and academic positions at the Chang'an University. His Research interests include Intelligent Transport System (ITS) optimization, microscopic and macroscopic transport modelling and simulation, and data-driven traffic analytics and forecasting.

**Joan Lasenby** received her BA/MMath in Mathematics from the University of Cambridge in 1982, and a PhD in Radio Astronomy from the University of Cambridge in 1986. She is currently Professor of Image and Signal Analysis in the Signal Processing and Communications Group, and Deputy Head of Department (Graduate Studies) in the Department of Engineering, University of Cambridge, and Fellow and Director of Studies at Trinity College Cambridge.