Supporting Information for

Urban Eco-Network Traffic Control via MARL-Based Signal and Speed Coordination

Lanping Chen [1,*], Fan Yang [2,*], Chenyuan Chen [1], Yue Zhu [1], Ziyuan Xu [1], Ying Xu [2] and Lin Zhu [1],

[1] Department of Automation, Changzhou University, Changzhou 213159, China; qq644050356@gmail.com (C.C.); s23060854047@smail.cczu.edu.cn (Y.Z.); s24060854056@smail.cczu.edu.cn (Z.X.); zhulin200112@163.com (L.Z.)
[2] School of Computer Science and Artificial Intelligence, Changzhou University, Changzhou 213159, China; s23150812051@smail.cczu.edu.cn (F.Y.); s24150812055@smail.cczu.edu.cn (Y.X.)
* Correspondence: chenlp@cczu.edu.cn

3.1. Vehicle speed controller

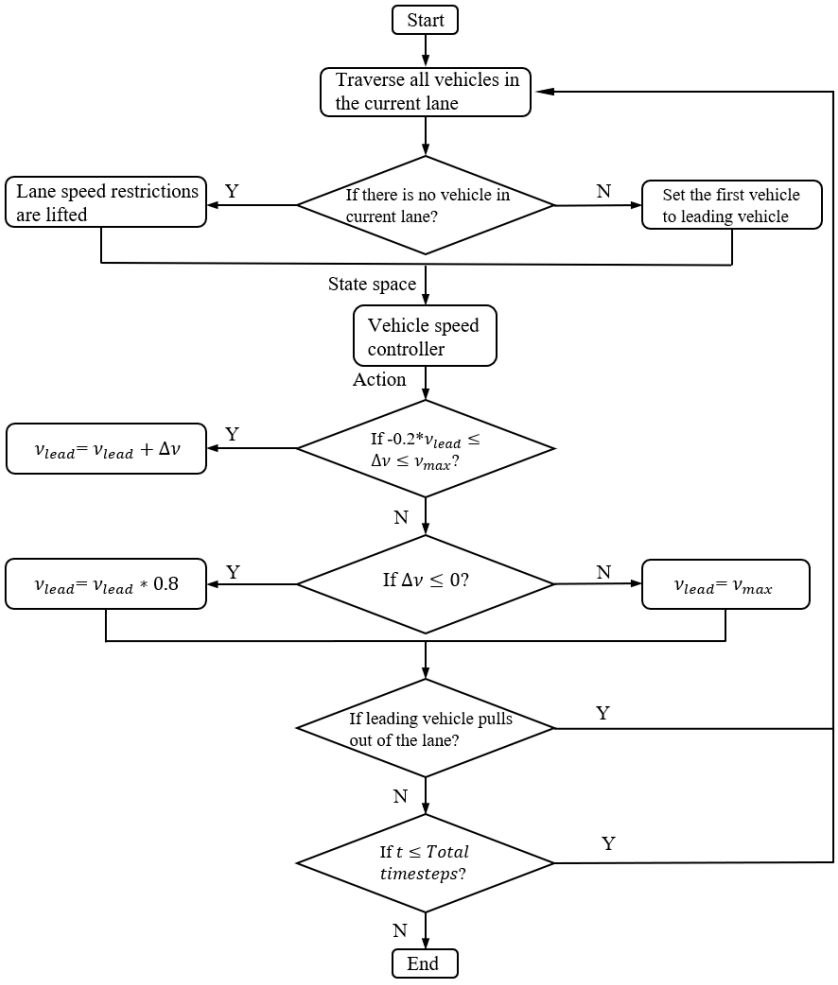


Figure S1. The flowchart of Vehicle Speed Controller System.

4.2. Parameters details of algorithms:

The detailed configurations of SAC and other algorithms are as follows:

SAC: In this experiment, the value network of SAC uses a double Q-network, where each Q-network consists

of three fully connected layers with a dimension of 256. The policy network consists of two fully connected layers with the same structure. SAC adopts a dynamic temperature adjustment method, where the temperature is dynamically adjusted based on the difference between the target entropy and the actual entropy of the current policy. The activation function of SAC is LeakyReLU, and the learning rates for the Actor and Critic are 5e-3 and 5e-2, respectively. The detailed parameters are shown in Table S1.

Table S1. Hyperparameters used for SAC with Beta.

hyperparameters	value	hyperparameters	value	hyperparameters	value
actor lr	5e-3	activation	LeakyReLU	memory size	1e6
critic lr	5e-2	entropy coeff	auto	batch size	4000
gamma	0.99	entropy lr	1e-3	action space	box

PPO: Proximal Policy Optimization (PPO) is one of the most advanced MARL algorithms currently applied in this field. One of the major advantages of PPO is its ability to handle both discrete and continuous action spaces, making it widely applicable. PPO limits the update magnitude using a clipped probability ratio (Clip), ensuring stability. Additionally, PPO uses Generalized Advantage Estimation (GAE) to compute the advantage estimate function based on a global value network. To increase the exploration of the policy, PPO also introduces policy entropy. The PPO network model in this experiment consists of four hidden layers, with each hidden layer containing 256 neurons. The detailed parameters are shown in Table S2.

Table S2. Hyperparameters used for PPO.

hyperparameters	value	hyperparameters	value	hyperparameters	value
lr	1e-5	activation	LeakyReLU	batch size	4000
gae lambda	0.99	entropy coeff	1e-2	action space	box
gamma	0.99	grad clip	10	clip param	0.3

A3C: Asynchronous Advantage Actor-Critic (A3C) combines Actor-Critic with an asynchronous update mechanism, demonstrating excellent performance in addressing the stability and convergence speed issues of traditional RL methods. A3C accelerates training by using multiple parallel worker threads. Each thread interacts independently with the environment and updates its local policy and value functions by receiving reward signals. Each thread maintains a local copy of the network, which is updated asynchronously with the global network. A3C also uses Gradient Clipping to prevent gradient explosion and restricts updates to the global network. The detailed parameters are shown in Table S3.

Table S3. Hyperparameters used for A3C.

hyperparameters	value	hyperparameters	value	hyperparameters	value
lr	1e-4	optimizer	Adam	batch size	4000
gae lambda	0.99	entropy coeff	1e-2	action space	discrete
gamma	0.99	grad clip	40		

DQN: Deep Q-Network (DQN) is a value-based reinforcement learning algorithm that integrates deep learning techniques. It improves algorithm performance by introducing the separation of the value function and the advantage function. This method decomposes the Q-value into two parts: the state value function and the action

advantage function. In this experiment, the DQN consists of two Q-networks, each with two fully connected layers of size 256. DQN uses Huber loss to stabilize the prediction of the value function, ensuring both the convergence speed of the model and the robustness of the training process. The detailed parameters are shown in Table S4.

Table S4. Hyperparameters used for DQN.

hyperparameters	**value**	**hyperparameters**	**value**	**hyperparameters**	**value**
lr	1e-6	optimizer	Adam	memory size	5e5
v min	-10	v max	10	batch size	4000
gamma	0.99	grad clip	40	action space	discrete

Table S5 shows the application scenarios of the above algorithms. Since the action space of the traffic signal control system is discrete and the action space of the CAV speed control system is continuous, only SAC and PPO, which can handle mixed action spaces, are suitable for the proposed CTS. The other algorithms are only suitable for TSC.

Table S5. The usage scenarios of each algorithm.

	SAC	PPO	A3C	DQN
Traffic signal controller	√	√	√	√
Cooperative traffic controller	√	√	×	×

5.3. Performance analysis of CTS under intersections

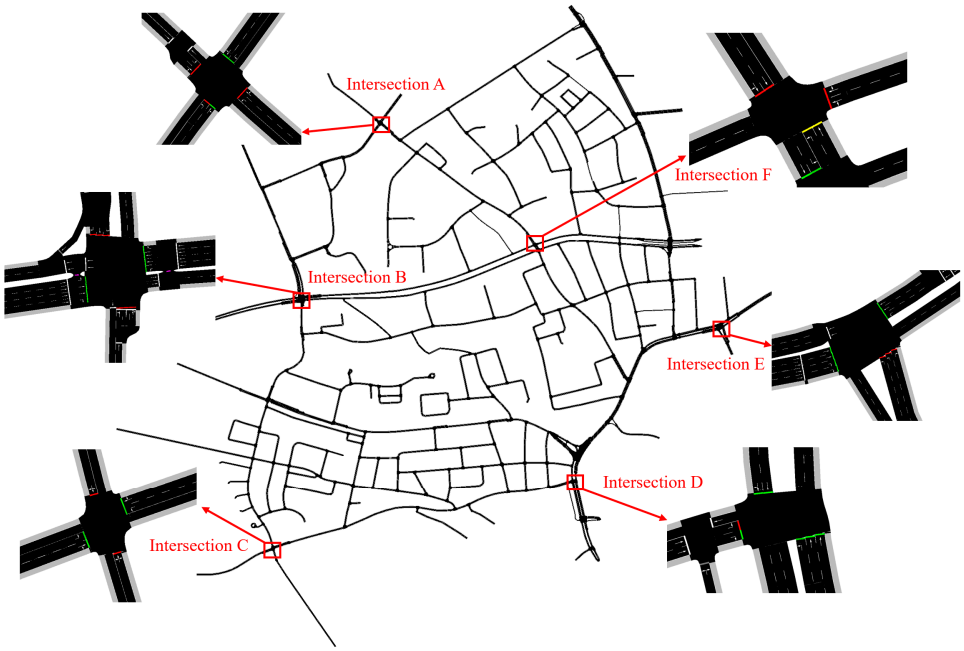
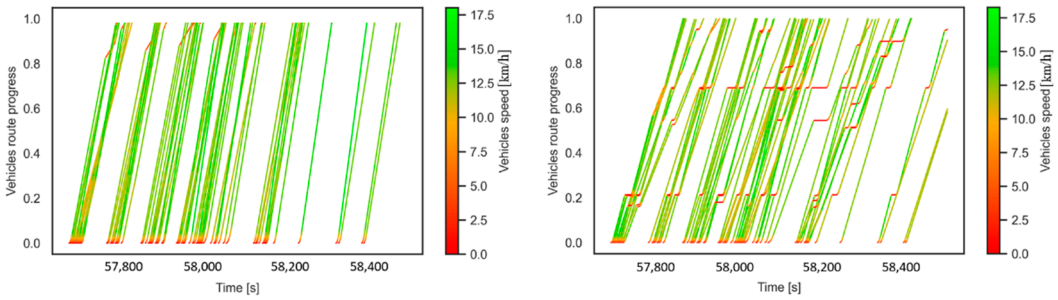
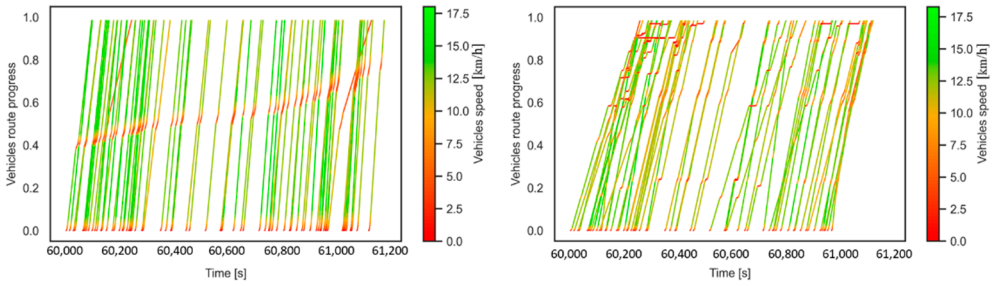


Figure S2. Schematic diagram of the six major intersections from Ingolstadt.



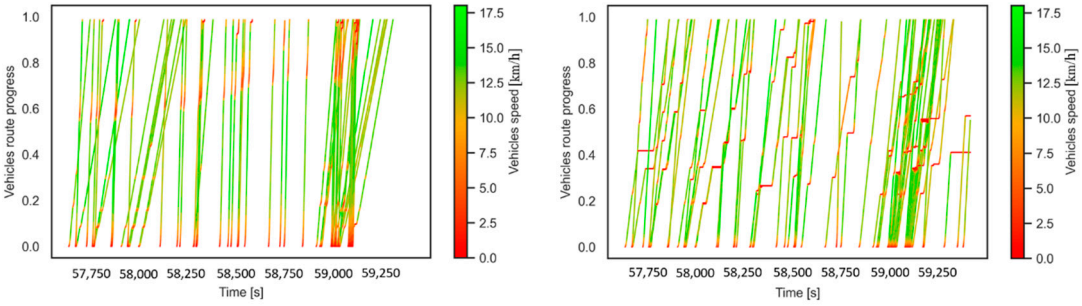
(a) SAC with Beta (b) PPO

Figure S3. Diagram of the change in vehicle speed when passing through the Intersection B.



(a) SAC with Beta (b) PPO

Figure S4. Diagram of the change in vehicle speed when passing through the Intersection C.



(a) SAC with Beta (b) SAC with Gaussian

Figure S5. Diagram of the change in vehicle speed when passing through the Intersection E.