

State Design in Reinforcement Learning-Based Traffic Signal Control Using Similarity Metrics

Chao Wan¹; Weiqiang Wu²; Xiaoning Liu³; and Weibin Zhang^{4*}

¹ School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; E-Mail: wanc@njust.edu.cn

² School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; E-Mail: wuweiqiang55@163.com

³ School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; E-Mail: liuxiaoning@njust.edu.cn

⁴ (Corresponding Author) School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; E-Mail: wbin.zhang@outlook.com

ABSTRACT

In traffic signal control tasks based on reinforcement learning, the design of observation forms presents a challenging issue. Simple observation forms may result in information loss, while complex observation forms can lead to the curse of dimensionality. This paper addresses this challenge by employing bisimulation metrics to calculate the similarity between heterogeneous observations, thereby decomposing the entire state space and achieving a more compact observation form. By leveraging the limited fitting capability of shallow neural networks, we compensate for the discrepancies caused by the observation forms.

Moreover, simulation software acts as a data source for models; however, the bounded resources of these simulations impede the comprehensive exploration of the entire state space. To tackle these challenges, a novel simulation platform named LiikeSim has been developed. It is designed to enhance simulation speeds, enabling more extensive exploration of the state space within the constraints of limited simulation resources.

INTRODUCTION

Traditional traffic signal control algorithms are finding it increasingly challenging to adapt to the growing complexity of traffic environments. Consequently, reinforcement learning (RL) has emerged as a viable approach to address signal control

problems. Figure 1 illustrates the interaction between the agent and the environment, where the environment returns full state information, which is actually only partially observed (Observations) by the agent. When developing an RL-based algorithm, the design of the form of observation, action, and reward must be initiated.

The significance of rewards cannot be overstated; they serve as the optimization objective for the model. Consequently, rewards often lack generalizability across different tasks. Even in optimization problems such as maximizing road network throughput or minimizing travel time, the design of rewards can vary significantly. However, since the design of rewards is typically closely tied to the specific task, it becomes challenging to create a universally applicable reward paradigm.

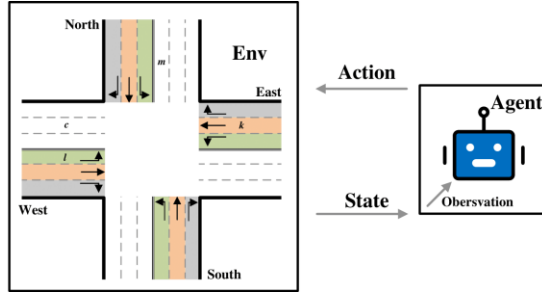


Figure 1. Interaction between the agent and the environment.

For the design of observation, it tends not to be directly associated with the task. Therefore, it can be reused in different tasks. Observation denotes the perception of the environment by the agent. In practice, the agent often does not have access to all the information. So, we need to be careful with the observation design to ensure that the information acquired by the intelligences is effective. The traffic parameters involved in the observations designed in previous work are summarized in Table I. Observations are often designed to involve one or more traffic flow parameters.

Table 1. Traffic Observation Statistics

Observation	Article
Average waiting time	(Oroojlooy et al. 2020)
Running vehicle num	(Oroojlooy et al. 2020) (Zhang et al. 2021)
Total vehicle num	(Wang et al. 2021) (Wang et al. 2019) (Wei et al. 2019) (Liang et al. 2022) (Wu et al. 2021) (Xu et al. 2021) (Wei et al. 2018) (Xiong et al. 2019) (Zang et al. 2020)
Cumulative delay	(Chu et al. 2019)
Average speed	(Mao et al. 2023) (Wang et al. 2019) (Liang et al. 2022)
Queue length	(Oroojlooy et al. 2020) (Wang et al. 2019) (Wei et al. 2018) (Yoon et al. 2021)
Pressure	(Chen et al. 2019)
Density	(Liang et al. 2022)

To design better observations, we need to measure the similarity between different observations. Observations with greater variability combined tend to contain more information. Directly calculating the variability between observations is very complicated. Therefore, we consider starting with the observation of transfer probabilities. The variability between observations is measured indirectly by calculating the variability of the probability distribution. In summary, the main contribution of this paper as follows:

- A framework has been proposed for calculating the distances between heterogeneous observations within the same state space.
- By classifying various observational forms, we have identified more effective forms of observation.

METHODOLOGY

In this section, we present our method for calculating the distance between various observations. The compositions of various forms of observations are thoroughly described. Additionally, the methodology for calculating distances is delineated in this section.

Observations, Action and Reward Design

Observations. As various forms of observation have selected, we list them in the following. Unless specified, the statistics are for incoming approach.

Traffic movement is defined as the traffic traveling across an intersection from one incoming approach to an outgoing approach, denote a traffic movement from lane l to lane m as (l, m) . The whole state space is formed as below:

- Vehicles num. The count of vehicles excluding those that are stopped. Denoted as $x_r(l)$.
- Waiting vehicles num. The number of stopped vehicles on the road, denoted as $x_s(l)$.
- Total vehicles num. Number of all vehicles travelling or stopping on the incoming approach, equal to $x_r(l) + x_s(l)$.
- Signal phase. Current signal phase p , the next signal phase in the fixed phase sequence is not considered here as an observation.
- Average speed. Average speed of vehicles in the incoming approach $s(l) = \sum_{i \in [x_r(l)]} v(i) / x_r(l)$. $v(i)$ denote the speed of vehicle i .
- Density. Ratio of number of vehicles to lane length, calculated as $x(l) / x_{max}(l)$, where $x_{max}(l)$ is the maximum permissible vehicle number on l .
- Pressure. Phase-based features P_i , denoted as:

$$P_i = \left| \sum_{(l,m) \in i} (w(l,m)) \right|, w(l,m) = \frac{x(l)}{x_{max}(l)} - \frac{x(m)}{x_{max}(m)} \quad (1)$$

- Cumulative delay. Sum of the incoming delay of each vehicle, i.e. $\left(1 - \frac{v_{\{t_d\}}}{v_{\{max\}}}\right)$ the moment of phase selection t_d .
- Average waiting time. Average waiting time of queued vehicles, calculate as $w(l) = \sum_{i \in [x_s(l)]} \frac{w(i)}{x_r(l)}$. $w(i)$ denote the waiting time of vehicle i .

Action. The action of each agent at each time step is to choose one of the phases of the traffic signal. Actions are chosen every twenty seconds; 15 seconds are allocated to meet pedestrian crossing requirements, followed by a 5-second yellow light phase to facilitate vehicle clearance.

Reward. While the design of rewards is significant, it does not constitute the primary focus of this study. We introduce various reward designs, and select one for implementation as the reward system in this research.

- Delay. Calculate as $\left(1 - v_{t_d}/v_{max}\right)$ at the moment of phase selection t_d .
- Queue length. The total queue length of all incoming lanes is considered when selecting actions.
- Pressure. Calculate the same as observation pressure.
- Queue length difference. The change in vehicle queue lengths before and after executing an action.
- Waiting time difference. The change in waiting time before and after executing an action.

Here, the performance of various rewards at a single intersection is presented in terms of travel time and throughput.

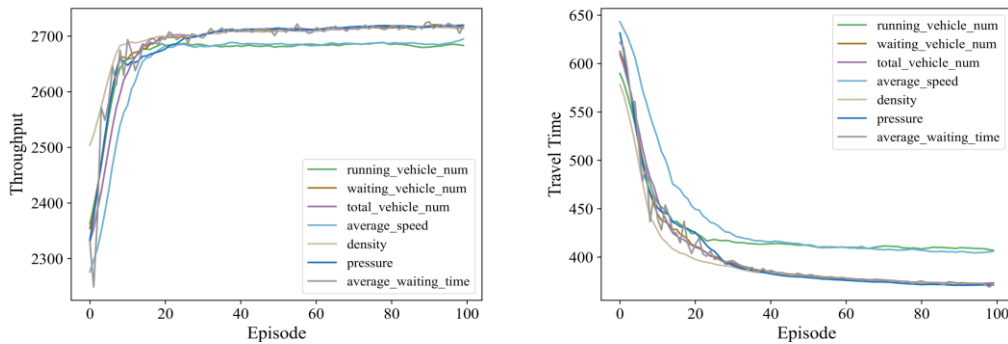


Figure 2. Performance of rewards in terms of throughput and travel time.

We summarize the specific performance values of the rewards in Figures 2 as Table 2.

Table 2. Reward Performance

Reward Form	Travel Time	Throughput
Delay	440.31	2651
Pressure	386.17	2659
Queue Length	397.21	2710
Queue Length Difference	370.72	2721
Waiting Time Difference	362.37	2754

For travel time, the smaller the better, and for total throughput, the bigger the better. During testing, the most effective reward was determined to be the waiting time difference; consequently, this metric has been selected as the reward for our study. It can be calculated as $r(s_t, a_t) = w_t - w_{t+T}$. w_t represents the accumulate waiting time of vehicles on L_i at time t .

Distance Calculation

Previous research (Ferns et al. 2006; Castro 2019) has subdivided the state space within a single observation environment for clearer analysis. In this section, we introduce a method for measuring similarity within an isolated observation space. This similarity measurement will be expanded to heterogeneous observation spaces.

In the TSC problem, the set of observable state information is assumed to be S . In general, the similarity between observations does not account for differences in observation forms. Hence, the observation space is continued to be represented as s . Various forms of states contribute to the observation s of the environment. It is evident that s is a non-empty subset of S . Assume that an optimal observation representation, denoted by s_i , exists, the following conditions hold:

$$\forall s_j \in S \setminus s_i, \forall s'_j \in S \setminus s_j, \exists s'_i \in s_i: \pi(s'_i) = \pi(s'_j) \quad (2)$$

This formula can be decomposed into a combined representation of the reward function and the state transition function:

$$\begin{aligned} R(s'_i, a) &= R(s'_j, a) & \forall a \in A \\ P(G \mid s'_i, a) &= P(G \mid s'_j, a) & \forall a \in A, \forall G \in S \end{aligned} \quad (3)$$

And $P(G \mid s, a) = \sum_{s' \in G} P(s' \mid s, a)$. However, strictly adhering to an optimal representation is impractical due to its high sensitivity to any changes in the

environment or the model. For this reason, modified this expression to be less rigid by defined a metric d . With $c \in [0,1]$, mathematical expression for d as:

$$d(s_i', s_j') = \max_{a \in A} (1 - c) \cdot |R(s_i', a) - R(s_j', a)| + c \cdot W_1(P(G | s_i', a), P(G | s_j', a); d)$$

W_1 is the 1th Wasserstein metric as:

$$W_1(P_i, P_j; d) = \left(\inf_{\gamma' \in \Gamma(P_i, P_j)} \int_{S \times S} d(s_i, s_j) \gamma'(s_i, s_j) \right) \quad (4)$$

Heterogeneous Observation Space Similarity Calculation

Take two distinct types of observations as examples, each associated with its own observation space as o_i and o_j . If a state transition process occurs, the following Markov process can be derived:

$$\begin{aligned} &\{o_i', a, o_i'', r\} \\ &\{o_j', a, o_j'', r\} \end{aligned} \quad (5)$$

Rewrite the above equation as:

$$\{(o_i', o_j') \times a \rightarrow (o_i'', o_j'')\} \quad (6)$$

If the two observation spaces are the same, then there exists a function f , for all Markov process, it has the following properties:

$$f(o_j') = o_i' + \varepsilon' \quad (7)$$

while $\varepsilon = 0$, then the two observation forms are completely equivalent. However, this formula will ignore transfer probabilities within the observation space. Considering the Observation transfer probability, the loss function for computing the similarity of heterogeneous observations can be defined as:

$$d(C[f(o_i'), f(o_i'')], C[f(o_j'), (o_j'')]) \quad (8)$$

The model for calculating the Wasserstein metric is as follows:

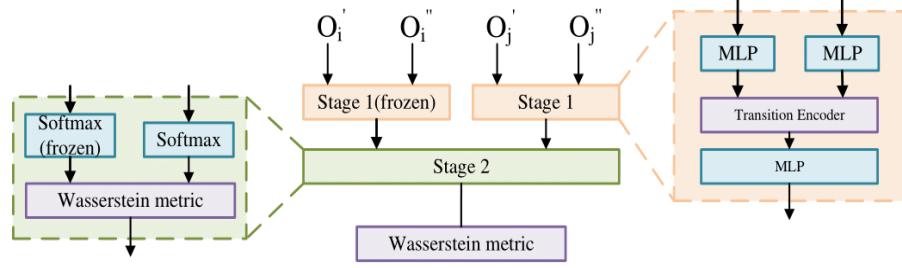


Figure 3. Wasserstein Metric Calculation.

To measure multiple observation spaces, one observation space (stopped vehicle num) is assumed as an anchor. Measurement of the distance from other observation spaces. The parameters of the anchor are not involved in the update.

RESULT

Experimental data

We tested the proposed structure on both standard and non-standard intersections. The intersection model was built using the in-house road simulation software LiikeSim. To verify the robustness of the proposed module, it has been tested on both standard and non-standard single intersections. The standard intersection has homogeneous approaches, all of which allow left turns, right turns, and straight travel.

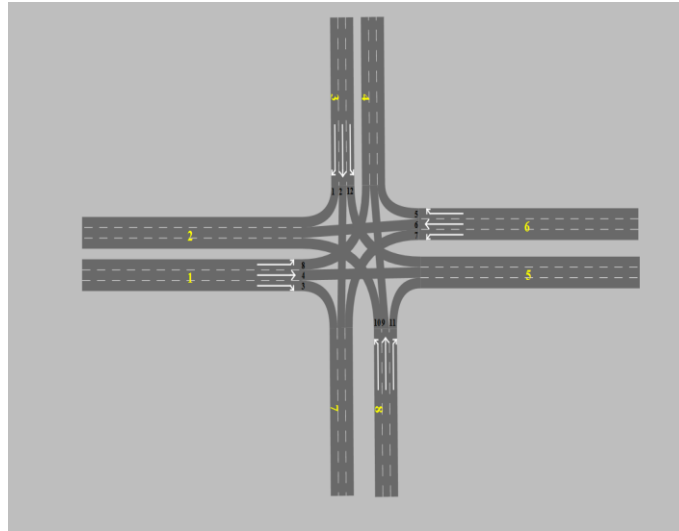


Figure 4. Standard single-intersection structure.

The non-standard single intersection was taken from an intersection in Hangzhou. The structures of intersections in urban roads vary, and not all approaches at this intersection allow left and right turns as well as straight travel. This is a common scenario in urban traffic.

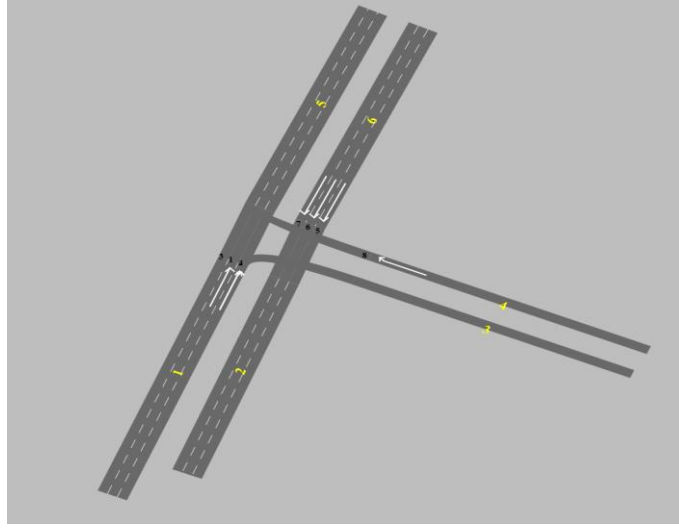


Figure 5. Non-standard single-intersection structure.

The traffic flow data corresponding to the standard intersection was quantitatively generated directly using the API provided by LiikeSim. The traffic flow data corresponding to the non-standard intersection was collected by video capture devices and then converted.

Wasserstein distance metric for heterogeneous and homogeneous single intersections

Figure 6 shows the Wasserstein distance metric between various observation forms and the anchor (number of waiting vehicles) at the standard intersection. The current signal phase observation form is not included in the calculation, as this parameter is often provided by the agent and is typically represented by an integer constant, making it meaningless to calculate the distance metric. It is evident from the figure that there is a significant distance between the observation forms of average waiting time and number of waiting vehicles. The distances between the other types are roughly similar. Therefore, it can be concluded that combining the observation forms of average waiting time and number of waiting vehicles can yield greater information content. That is, combining $w(l)$ and $x_w(l)$.

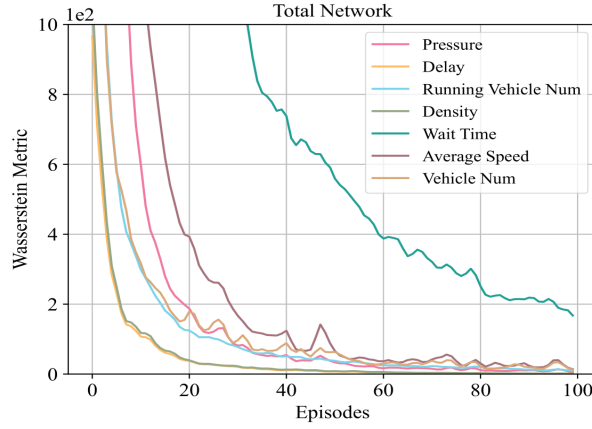


Figure 6. Wasserstein distance metric at a standard intersection.

Specifically, the distance metric between various observation forms at a single intersection may lack robustness. Therefore, we supplemented the distance metric results for heterogeneous intersections as shown in Figure 7.

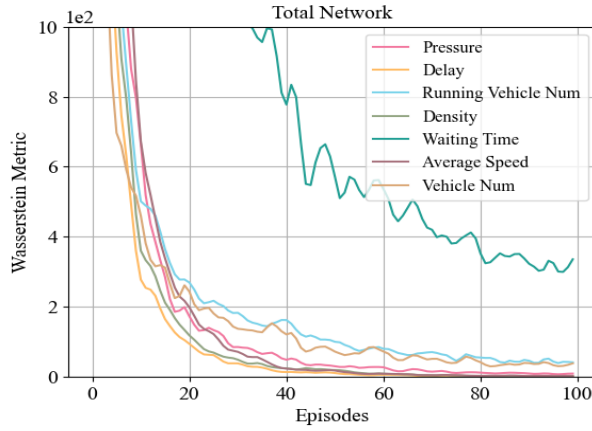


Figure 7. Wasserstein distance metric at a non-standard intersection.

Wasserstein distance metric between various observation forms at the network level

In urban traffic, intersections are closely interconnected with each other, so this distance metric is also expected to be robust at the network level. Therefore, in this section, the distance metric module is also tested at the network level.

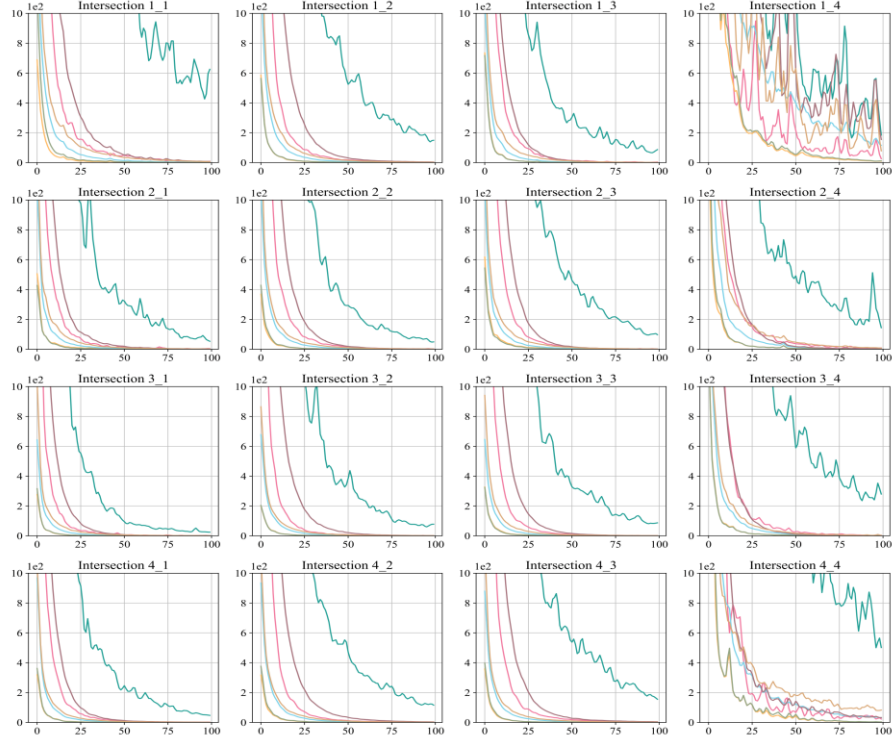


Figure 8. Wasserstein distance metric for various observation forms at the network level containing the standard intersection.

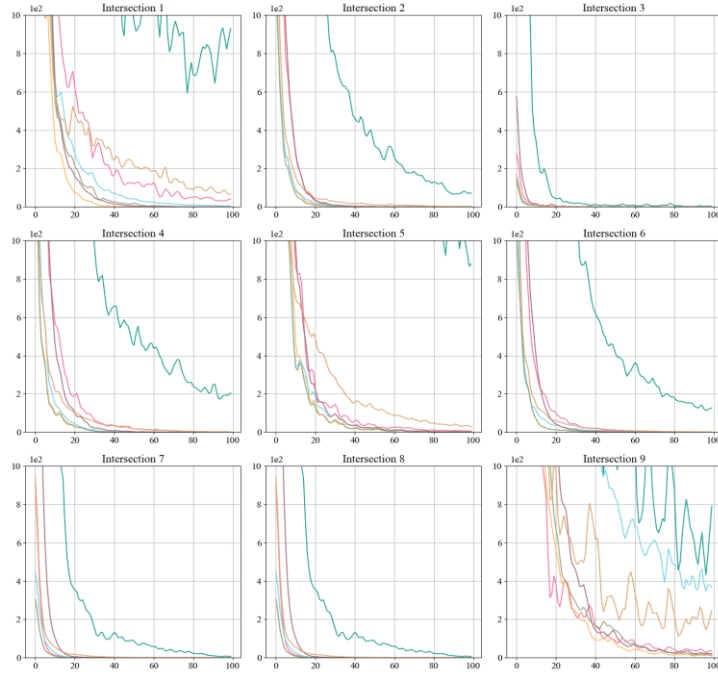


Figure 9. Wasserstein distance metric for various observation forms at the network level containing the non-standard intersection.

Evidently, there is still a significant distance between the observation forms of average waiting time and number of waiting vehicles. The experimental results demonstrate the robustness of the distance metric module proposed in this chapter. This means that in both single-intersection and network-level signal control problems, using the combination of average waiting time and number of waiting vehicles can encompass most of the information. In previous works, few articles have used this combination.

CONCLUSION

In this paper, we leveraged the limited fitting capability of shallow neural networks to reduce the representation error between observations with different representations. While reducing the size of the state space that needs to be explored, we ensured maximum information content. We extended the bisimulation metric method, which was previously only applicable to homogeneous observations, to heterogeneous observation forms, making it possible to calculate the distribution distance between heterogeneous observations. Additionally, our results on single intersections validated the effectiveness of the distance metric module and obtained the optimal combination of observation forms as the state. Furthermore, we also validated the robustness of the proposed module at the network level. Even at the network level, the results obtained were still stable.

REFERENCES

- Castro, P.S. (2019). Scalable methods for computing state similarity in deterministic Markov Decision Processes. *ArXiv, abs/1911.09291*.
- Chu, T., Wang, J., Codecà, L., & Li, Z. (2019). Multi-Agent Deep Reinforcement Learning for Large-Scale Traffic Signal Control. *IEEE Transactions on Intelligent Transportation Systems*, 21, 1086-1095.
- Chen, C., Wei, H., Xu, N., Zheng, G., Yang, M., Xiong, Y., Xu, K., & Zhenhui (2020). Toward A Thousand Lights: Decentralized Deep Reinforcement Learning for Large-Scale Traffic Signal Control. *AAAI Conference on Artificial Intelligence*.
- Ferns, N., Castro, P.S., Precup, D., & Panangaden, P. (2006). Methods for Computing State Similarity in Markov Decision Processes. *ArXiv, abs/1206.6836*.
- Jiang, Q., Li, J., Sun, W., & Zheng, B. (2021). Dynamic Lane Traffic Signal Control with Group Attention and Multi-Timescale Reinforcement Learning. *International Joint Conference on Artificial Intelligence*.
- Liang, E., Su, Z.C., Fang, C., & Zhong, R. (2022). OAM: An Option-Action Reinforcement Learning Framework for Universal Multi-Intersection Control. *AAAI Conference on Artificial Intelligence*.

- Mao, F., Li, Z., Lin, Y., & Li, L. (2023). Mastering Arterial Traffic Signal Control With Multi-Agent Attention-Based Soft Actor-Critic Model. *IEEE Transactions on Intelligent Transportation Systems*, 24, 3129-3144.
- Oroojlooy, A., Nazari, M., Hajinezhad, D., & Silva, J. (2020). AttendLight: Universal Attention-Based Reinforcement Learning Model for Traffic Signal Control. *Neural Information Processing Systems*, 33, 4079-4090.
- Wei, H., Zheng, G., Yao, H., & Li, Z.J. (2018). IntelliLight: A Reinforcement Learning Approach for Intelligent Traffic Light Control. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.
- Wang, Y., Xu, T., Niu, X., Tan, C., Chen, E., & Xiong, H. (2019). STMARL: A Spatio-Temporal Multi-Agent Reinforcement Learning Approach for Cooperative Traffic Light Control. *IEEE Transactions on Mobile Computing*, 21, 2228-2242.
- Wei, H., Xu, N., Zhang, H., Zheng, G., Zang, X., Chen, C., Zhang, W., Zhu, Y., Xu, K., & Li, Z.J. (2019). CoLight: Learning Network-level Cooperation for Traffic Signal Control. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*.
- Wu, L., Wang, M., Wu, D., & Wu, J. (2021). DynSTGAT: Dynamic Spatial-Temporal Graph Attention Network for Traffic Signal Control. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*.
- Wang, M., Wu, L., Li, J., & He, L. (2021). Traffic Signal Control with Reinforcement Learning Based on Region-Aware Cooperative Strategy. *IEEE Transactions on Intelligent Transportation Systems*, 23, 6774-6785.
- Xiong, Y., Zheng, G., Xu, K., & Li, Z.J. (2019). Learning Traffic Signal Control from Demonstrations. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*.
- Xu, B., Wang, Y., Wang, Z., Jia, H., & Lu, Z. (2021). Hierarchically and Cooperatively Learning Traffic Signal Control. *AAAI Conference on Artificial Intelligence*.
- Yoon, J., Ahn, K., Park, J., & Yeo, H. (2021). Transferable traffic signal control: Reinforcement learning with graph centric state representation. *Transportation Research Part C-emerging Technologies*, 130, 103321.
- Zang, X., Yao, H., Zheng, G., Xu, N., Xu, K., & Zhenhui (2020). MetaLight: Value-Based Meta-Reinforcement Learning for Traffic Signal Control. *AAAI Conference on Artificial Intelligence*.
- Zhang, L., Wu, Q., Shen, J., Lu, L., Du, B., & Wu, J. (2021). Expression might be enough: representing pressure and demand for reinforcement learning based traffic signal control. *International Conference on Machine Learning*.