Mastering Arterial Traffic Signal Control With Multi-Agent Attention-Based Soft Actor-Critic Model

Feng Mao, Zhiheng Li, *Member, IEEE*, Yilun Lin, and Li Li, *Fellow, IEEE*

***Abstract*— Recent studies have made dozens of attempts to apply multi-agent deep reinforcement learning (MARL) for large-scale traffic signal control. However, most related studies have ignored how to master arterial traffic signal control. We cannot easily extract useful information and search solution space because the arterial traffic control problem has large state-action spaces. Here we tackle these issues by proposing a multi-agent attention-base soft actor-critic (MASAC) model to master arterial traffic control. Specifically, we implement the attention mechanism in the actor and critic network to enhance traffic information extraction ability. More importantly, we are the first to apply the soft actor-critic (SAC) algorithm to train the arterial traffic control model to search more solution spaces. Testing results indicate that the MASAC method significantly outperforms existing MARL algorithms and the multiband-based method. These findings can help researchers to design better model structures for other MARL problems.**

***Index Terms*— Arterial traffic signal control, multi-agent reinforcement learning, soft actor-critic, attention.**

I. Introduction

TRAFFIC signal control is critical to building a smart city. With the development of deep reinforcement learning (DRL) techniques, extensive studies [1], [2], [3], [4], [5] utilized the DRL method for isolated traffic signal control. Recently, numerous researchers started to apply DRL approaches for arterial traffic signal control [6] and large-scale traffic signal control [7], [8], [9], [10], [11], [12], [13], [14], [15]. This paper focuses on how to master arterial traffic signal control because coordinating traffic signals in the arterial road is one of the keys to improving traffic efficiency. Moreover, how to design a MARL model for the arterial traffic control remained to be answered when we intend to solve multi-intersection traffic signal control tasks with larger state-action spaces.

Due to the high dimension of state-action spaces [9], [16] of the arterial traffic signal control problem, we need to design a MARL model to extract useful information and seek solution spaces. Generally, such a MARL model is hard to design and is mainly affected by three categories of factors: model settings, model structure design, and DRL training algorithm.

The first problem for MARL model design is to select the appropriate model settings to describe the arterial traffic control problem [6]. At present, most approaches selected similar state (e.g., queue length, the number of approaching vehicles, and the current phase as the state [3], [6], [8], [9], [11], [12], [13], [14]), reward (e.g., queue length [11], [12], [13] or pressure [6], [7]), and action settings (phase shift [2], [4], [17] or phase selection [6], [8], [11], [12], [13], [14]). Therefore, we focus on the latter two categories of factors and investigate what mechanisms can help to boost the performance of MARL models.

The second problem is to design a good model structure to extract useful information. We cannot easily extract useful information because arterial traffic control suffers from local observation and high-dimension of state-action spaces. To solve these problems, there are mainly three categories of techniques to design a MARL model structure: centralized learning [18], [19], communication module [8], [10], [12], [13], [14], and the attention mechanism [18]. However, few studies took a deep insight into what mechanism in the MARL model structure design help to master arterial traffic control.

The third problem is to choose a powerful DRL training algorithm to search solution spaces for the arterial traffic control problem. Most previous studies used simple DRL algorithms to train their MARL models for large-scale traffic signal control, such as deep Q-network (DQN) [12], [13], [20], policy gradient [14], or advantage actor-critic (A2C) [8], [9], [21]. However, no traffic signal control-related studies applied the SAC algorithm [22], which achieves the state-of-the-art (SOTA) performance in dozens of control benchmark tasks and behaves with strong robustness because of its superior solution space search capability [22].

In addition, few studies made a comparison between the MARL model and the SOTA arterial traffic signal optimization method, the multiband [23] and the AM-band method [24].

Manuscript received 18 January 2022; revised 3 June 2022 and 5 October 2022; accepted 28 November 2022. This work was supported in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2020B0909050003 and in part by the Shenzhen Science and Technology Innovation Committee under Grant JSGG20211029100204006. The Associate Editor for this article was S. C. Wong. *(Corresponding author: Li Li.)*

Feng Mao is with the Department of Automation, Tsinghua University, Beijing 100084, China, and also with the Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China (e-mail: mf19@mails.tsingua.ed.cn).

Zhiheng Li is with the Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China, and also with the Research Institute of Tsinghua, Pearl River Delta, Guangzhou 510530, China (e-mail: zhhli@tsinghua.edu.cn).

Yilun Lin is with the Shanghai AI Laboratory, Shanghai 200232, China (e-mail: linyilun@pjlab.org.cn).

Li Li is with the Department of Automation, BNRist, Tsinghua University, Beijing 100084, China (e-mail: li-li@tsinghua.edu.cn).

Digital Object Identifier 10.1109/TITS.2022.3229477

We are curious about that can the MARL model outperformed the multiband-based method.

To conclude, we need to thoroughly answered the following questions for the arterial traffic control problem:

(1) Which mechanism in model structure contributes most to master arterial traffic signal control, centralized training technique, communication module, or attention mechanism?
(2) Do existing MARL models have better performances by applying the SAC algorithm to train models?
(3) Can the MARL algorithms outperform the multiband-based method in the arterial traffic signal control problem?

To answer these questions, we propose a **M**ulti-agent **A**ttention-based **S**oft **A**ctor-**c**ritic (MASAC) model for arterial traffic signal control. Specifically, we apply the centralized training technique suggested by many previous MARL algorithms [18], [19] to mitigate the local observation issue and extract useful information. We also utilize the attention mechanism in the actor and critic networks to extract useful information from large state-action spaces. More importantly, we select the SAC algorithm [22], [25] to train the MARL model to seek more solution spaces.

To give answers to the first two questions, we add different combination modules of centralized training technique, communication module, attention mechanism, and SAC algorithm to various MARL models, and conduct extensive ablation experiments to investigate the effect of each module. To give answers to the third question, we compare the MASAC model with the AM-band method and several SOTA MARL models.

The remainder of this paper is organized as follows. *Section II* reviews recent related studies for MARL algorithms. *Section III* describes the model settings for the arterial traffic control problem. We introduce detailed implementations of our MASAC method to master arterial traffic control in *Section IV*. In *section V*, we compare the experimental results for various DRL algorithms and present the ablation results to illustrate the effect of each module in MARL algorithms. Finally, we conclude this paper in *Section VI*.

II. Related Works

In this section, we will introduce isolated traffic signal control first, then present the related studies on multi-intersections traffic signal control based on MARL models.

In recent years, dozens of studies [2], [3], [4] have attempted to apply DRL algorithms for isolated traffic signal control because DRL can adaptively optimize policy through learning experience samples generated by interaction with the traffic environment. The experimental results showed that the DRL algorithm outperformed the conventional traffic signal control method. At present, isolated traffic signal control has been done well. However, when extending isolated traffic signal control to multi-intersections traffic control, the multi-intersections traffic control suffers from the high dimension of state-action space [9], [16].

To mitigate the curse of dimension, researchers have proposed the MARL algorithms by distributing the global controller to each isolated agent. However, it is hard to extract useful information in many MARL algorithms due to local observations of each agent. Therefore, it is critical to design an elaborate MARL model to extract useful information and search solution space from large state-action spaces for the multi-intersections traffic signal control problem.

There are mainly three categories of techniques to design a MARL model structure to help extract useful information: centralized learning, communication module, and the attention mechanism.

Centralized learning is an effective method to mitigate local observation by utilizing global observations for the critic network and using the local observations for the actor network during training. In this way, the centralized critic can help the actor calculate a more accurate gradient [18]. Lowe et al. [18] made one of the earliest attempts to apply centralized learning to mitigate the non-stationary issue of MARL. Lately, numerous research [19], [21] applied the centralized learning mechanism in various MARL problems and achieved good performance. Centralized learning can not apply to large-scale MARL problems and may fall into a local optimal solution because centralized learning needs to use global observations to train the critic network.

The communication module [20], [26] is a promising way to address the problem of local observation by training a differentiable communication network to model the information interaction among agents. Extensive recent MARL models designed for multi-intersections traffic signal control [8], [10], [12], [13], [14] implemented communication modules. It should be noted that the MARL models may not be robust by applying the communication module because agent communication still can not remedy the local observation.

The attention mechanism [27] is a practical approach to extracting useful information by learning the importance weight of features. Oroojlooy et al. [3] proposed a universal attention-based DRL model (AttendLight) for isolated traffic signal control. The attention mechanism has been proved to be the key to achieving the universal capability of AttendLight. Besides, recent studies [12], [13] also utilized the attention mechanism in their proposed MARL algorithms. Both studies demonstrated that the attention mechanism can help to capture the importance of adjacent traffic lights to the target lights, as well as capture the key observation information over time.

Although the above MARL models have achieved success for large-scale traffic signal control, most studies did not deeply analyze what mechanisms help to boost the MARL algorithm performance. More importantly, most previous MARL models neglected to select powerful DRL training algorithms to search solution space.

The DRL training algorithm also plays an important role in boosting the performance of MARL models by seeking better solution space. Most of the existing studies utilized simple DRL algorithms to train their MARL model. Numerous studies [12], [13], [20] used the DQN algorithm [28], [29] to train their model. In this way, these studies cannot implement the framework of centralized training with decentralized execution, because the DQN algorithm only has the critic network to update policy. Besides, some studies [14], [26] chose the policy gradient algorithm [30] as the training method, which

led to sampling inefficiency and cannot use the framework of centralized training with decentralized execution. Recent studies [8], [9], [21] utilized the advantage actor-critic (A2C) algorithm [31] to train their proposed model, thus they could apply the framework of centralized training with decentralized execution. However, existing MARL models had low sample efficiency and unstable training performance. In this sense, we want to investigate can the powerful SAC algorithm help boost the performance of MARL models.

III. Model Settings for Arterial Traffic Control

A. The General MDP Description for MARL Problem

In this paper, the road network is represented by a directed graph $G = (\mathcal{V}, \mathcal{E})$, where $v_i \in \mathcal{V}$ denotes the control traffic light and $\mathcal{E} = \left\{e_{i,j} = \langle v_i, v_j \rangle \mid v_i, v_j \in \mathcal{V}\right\}$ denotes the directed link set. We model the arterial traffic signal control problem as a Decentralized Partially Observable Markov Decision Process (Dec-POMDP) [32], which can be defined as a tuple $\langle N, \mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$, and each element is described as follows:

(1) The number of agents N: $N = \|\mathcal{V}\|$ denotes the number of control traffic lights.
(2) State Space $\mathcal{S}$: $S_t \in \mathcal{S}_t$ denotes the road network state at time t, which comprises all the information in the road network at time t.
(3) Observation Space $\mathcal{O}$: We view the arterial traffic signal control as a Dec-POMDP problem and assume the agent can only observe the traffic information on the entrance lanes of its intersection, $o_{i,t} \in \mathcal{O}_t$ denotes the partially observable information for the traffic light i at time t.
(4) Action space $\mathcal{A}$: $a_{i,t} \in \mathcal{A}_i$ denotes the action for the traffic light i at time t, the action is calculated by a policy network π_{θ_i}, where θ_i are the parameters. $\mathcal{A} = \{a_1, \cdots, a_N\}$, $\pi = \{\pi_1, \cdots, \pi_N\}$ are the joint action and the joint policy of all traffic lights in the road network, respectively.
(5) State Transition Probability $\mathcal{P}$: In the POMDP system, $\mathcal{P}$ is controlled by the current state S_t and the joint action A_t, that is $S_t \times A_t \to S_{t+1}$.
(6) Reward $\mathcal{R}$: $r_{i,t} \in \mathcal{R}_t$ is the reward for the traffic light i at time t, and $r_{i,t}$ is controlled by the current state S_t and the joint action A_t, that is $S_t \times A_t \to r_{i,t}$.
(7) Discount Factor $\gamma \in [0, 1]$ indicates the impact of the current action selection on the future, the closer it is to 1, the greater the impact of the current action selection on the future.

In the arterial traffic signal control problem, we aim to maximize the discounted cumulative reward of the whole road network as follows:

$$R^{\pi}(S_0) = E\left(\sum_{t=0}^{T}\sum_{i=1}^{N}\gamma^t r_{i,t} \mid S_0, \pi\right) \tag{1}$$

where T denotes the time horizon of an episode, S_0 is the initial state of the traffic light i.

B. The Special MDP Settings for Arteral Traffic Control

For the traffic signal control problem, we have conducted comprehensive experiments to determine the following choices of appropriate state, action, and reward setting. We find that most studies [3], [6], [8], [9], [11], [12], [13]) have similar model settings through years of exploration.

1) Observation Setting: Observation $o_{i,t}$ describes the local information of the traffic light i. We combine ideas in recent studies [6], [17], and select the number of queued vehicles on each segment of every incoming lane $l(k)$ $(k = 1, \cdots, K)$, mean speed in each lane $v_{i,t}^l$, and the current phase $p_{i,t}$ as the local observation. In this paper, each lane is divided into 2 segments evenly $(K = 2)$, thus the observation space is defined as:

$$o_{i,t} = \left\{\left\{q_{i,t}^{l(1)}, q_{i,t}^{l(2)}\right\}, \left\{v_{i,t}^{l}\right\}, p_{i,t}\right\}, \quad l \in L_i \tag{2}$$

where L_i denoted the lane set of the intersection i.

Note that we use one hot form to represent the current phase p_t and describe the observation by lane characteristics. In concretely, we utilize a $(\|L_i\|, 4)$ matrix to represent $o_{i,t}$, where $\|L_i\|$ is the number of lanes for the intersection i. Fig. 1 illustrates how to describe the local observation of a intersection.

2) State Setting: The state $S_t = \left\{o_{1,t}, o_{2,t}, \cdots, o_{N,t}\right\}$ represents the global observation of the road network. S_t is the superposition of the observation of all agents. We extend the local observation of a single intersection to global observation situations. Fig. 1 shows the diagram of the state setting.

3) Action Setting: The most frequently used two action settings for traffic signal control are phase shift and phase selection. The former setting is to decide whether to keep or change the current phase, which is represented as $a_{i,t} \in \{0, 1\}$. The agent learns to adjust the green time of each phase and switch the phase in a fixed order. For the latter action setting the agent learns to select phase from the predefined phase set, which is denoted as $a_{i,t} \in \mathcal{A}_i$, where $\mathcal{A}_i$ is the predefined phase set for the intersection i. In this paper, we adopted the phase selection setting suggested by most studies [8], [9], [12], [13] owing to its flexibility.

4) Reward Setting: The most frequently used two reward choices for traffic light control are queue length and pressure as follows, respectively.

$$r_{i,t} = -\frac{1}{\|L_i\|}\sum_{l \in L_i} q_{i,t+1}^{l} \tag{3}$$

$$r_{i,t} = -\sum_{(l,m) \in M_i}\left|n_{i,t+1}^{l} - n_{i,t+1}^{m}\right|, \quad l \in L_i,\ m \in L_{i,\text{ out}} \tag{4}$$

where $n_{i,t+1}^{l}$ is the number of vehicles on lane l, $L_{i,out}$ is the out lane set for the intersection i, (l, m) is the available movement for the intersection i, M_i is the available movement set.

In this paper, we follow the reward setting in recent studies [6], [8], [11], [12], [13], [14] and apply the queue length denoted by Eq.(1) as the reward.

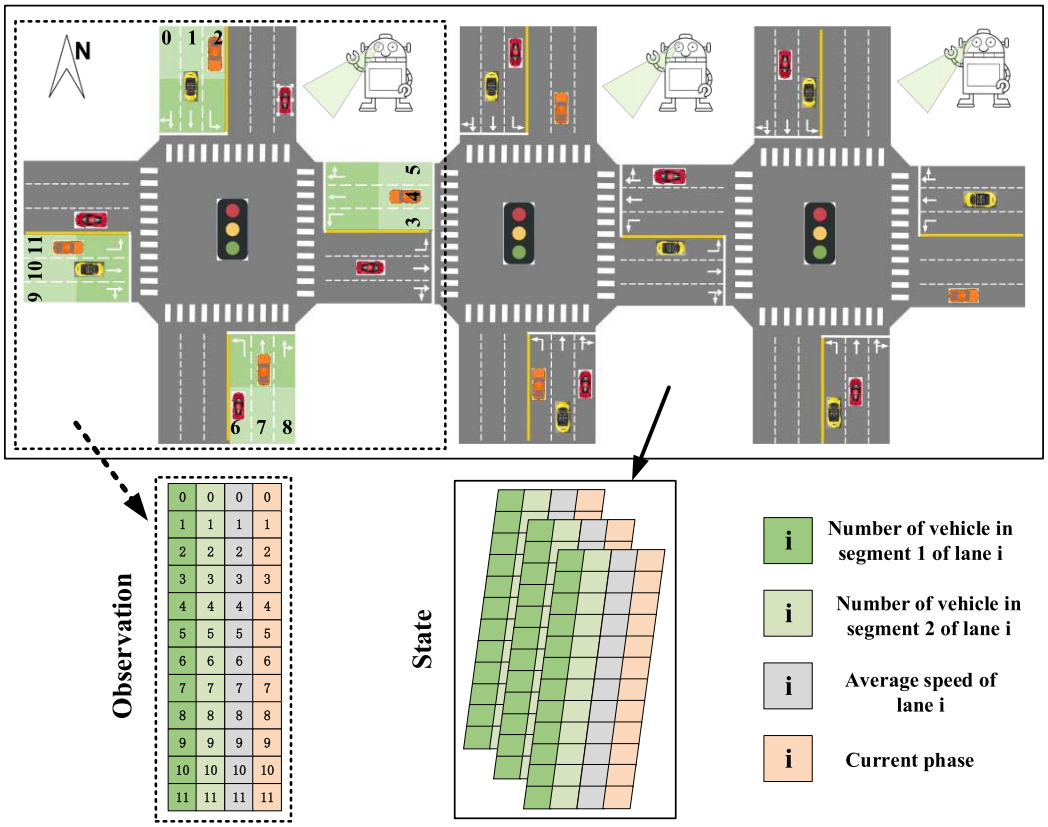


Fig. 1. The diagram of observation and state settings. For the observation setting, the first two vectors represent the number of approaching vehicles in segment1 and segment2. The third vector denotes the mean speed in each lance. The fourth vector is the current phase.

C. Our SAC-Based DRL Setting

We consider the arterial traffic control problem as a Dec-POMDP process, each intersection is modeled as an agent with the actor-critic framework to apply the centralized learning and decentralized execution method [18] to mitigate local observation. Specifically, the actor network can only obtain the local observation $o_{i,t}$, while the critic model can use the global observation S_t to update the network

We need to consider three aspects to design a MARL model for arterial traffic signal control:

(1) How to design the actor network to extract useful information to select the phase?
(2) How to design the critic network to help the actor network to reduce the bias of gradient estimation?
(3) How to choose the DRL training algorithm to search solution space from large state-action spaces?

To answer these questions, we propose our MASAC model as shown in Fig. 2. We use the SAC algorithm to train the model because SAC has a strong solution space search capability [22] and achieves the SOTA performance in dozens of control benchmark tasks. Each agent has an ALight network and two ACritic networks. The ALight network is an actor network with an attention mechanism and use to calculate the probability of selecting each phase. The ACritic network is an integrated critic network with attention mechanisms and use to calculate the Q value of each phase. It should be noted that we have compared the individual network and the weight-sharing network settings, the former outperformed the weight-sharing network in most experiments. The reason may be that the weight-sharing network need to use the ID of intersection as input, the adjacent intersections have strong correlations in arterial traffic control problem, so the weight-sharing network setting may not fit.

IV. Model Details and Training Algorithms

A. The Actor Network

First, we design an attention-based actor network called ALight to better extract useful information from the perspective of the phase feature. A recent study [3] showed that the DRL model had a better performance by extracting features based on the phase. The ALight network uses the local observation $o_{i,t}$ as input and outputs the probability of choosing each phase (a matric $(1, \|p_i\|)$) as follows:

$$\pi_{\theta_i}\left(o_{i,t}\right) = \text{softmax}\left(\text{ALight}\left(o_{i,t}\right)\right) \tag{5}$$

where θ_i denotes the parameters of the ALight network for the intersection i.

To allow more capacity in extracting features for each phase, we use the embedding1 layer [3] to embed into a higher dimension space. Moreover, we use the attention layer [27] to capture the importance of each embedded phase feature and outputs the weighted sum of embedding features as $h_{i,t}$, then the $h_{i,t}$ will be fed into a fully connected layer and a Softmax function, and outputs the probability of choosing each

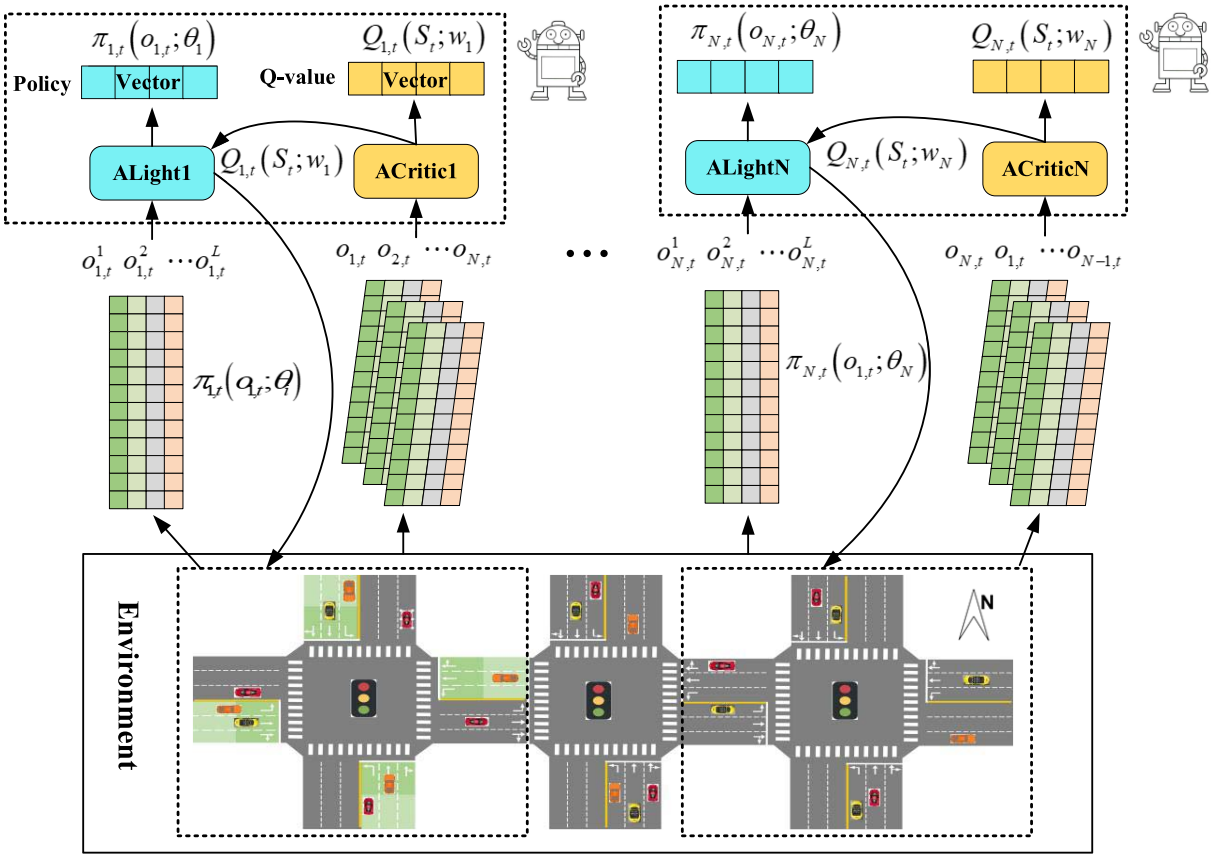


Fig. 2. The structure diagram of the MASAC model. Each agent has an ALight network and an ACritic network.

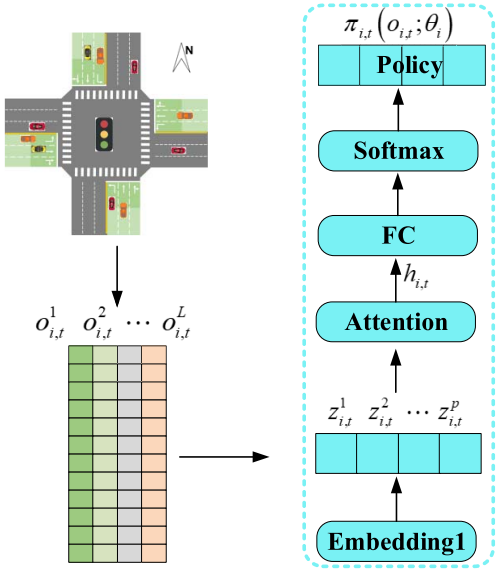


Fig. 3. The structure diagram of the ALight network.

phase $\pi_{\theta_i}\left(o_{i,t}\right)$. The detail of the ALight network is presented in Fig. 3.

Specifically, the embedding feature of the phase p can be calculated as:

$$z_{i,t}^{p}=\sum_{l\in L_{p}^{p}}\text{relu}\left(fc\left(o_{i,t}^{l}\right)\right) \tag{6}$$

where L_i^p is the lane set of the phase p for the intersection i, $f_c(.)$ denotes the fully connected function.

In this paper, we embed local observation $o_{i,t}$ into the 128 dimensions to capture useful information. We conduct dozens of experiments and found that the 128 dimensions embedded space is enough, you can also use higher dimension embedded space but it requires more time to train the embedding layer.

By implementing the attention layer, the output $h_{i,t}$ is calculated as a weighted sum of the embedding features based on the importance weight:

$$h_{i,t}=\text{Attention}\left(fc_{\theta_Q}\left(z_{i,t}\right),fc_{\theta_K}\left(z_{i,t}\right),fc_{\theta_V}\left(z_{i,t}\right)\right) \tag{7}$$

where $z_{i,t}=\left\{z_{i,t}^{1},z_{i,t}^{2},\cdots,z_{i,t}^{p}\right\}$ is the embedding features of all phases. $fc_{\theta_Q}(.)$, $fc_{\theta_K}(.)$, $fc_{\theta_V}(.)$ are the embedding layer of queries, keys, and values, respectively, and θ_Q, θ_K, θ_V are parameters of the corresponding network layer.

Then $h_{i,t}$ can be calculated as:

$$h_{i,t}=\text{softmax}\left(\frac{fc_{\theta_Q}\left(z_{i,t}\right)\left(fc_{\theta_K}\left(z_{i,t}\right)\right)^{T}}{\sqrt{d_{k_i}}}\right)fc_{\theta_V}\left(z_{i,t}\right) \tag{8}$$

where d_{k_i} is the scaling factor for the agent i.

Through the previous modules, $h_{i,t}$ has extracted features from the perspective of phase traffic information. Finally, we can calculate the probability that traffic light i selects each phase as follows:

$$\pi_{\theta_i}\left(o_{i,t}\right)=\text{softmax}\left(fc\left(h_{i,t}\right)\right) \tag{9}$$

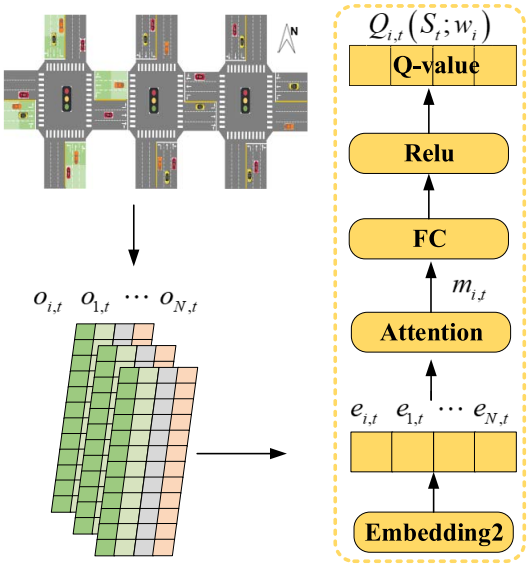


Fig. 4. The structure diagram of the ACritic network.

B. The Critic Network

Second, we design an attention-based critic network called ACritic to estimate a more accurate Q value $Q_w(S_t)$ (a $(1, \|p_i\|)$ matric denotes the accumulative reward of choosing each phase) for the agent itself. In this way, the critic network can help the actor reduce the bias of gradient estimation $\nabla_\theta \log \pi_\theta(a_t \mid o_t) Q_w(S_t, a_t)$. The ACritic network uses the global observation S_t and outputs the Q value as follows:

$$Q_{w_i}(S_t) = \text{relu}(\text{ACritic}(S_t)) \tag{10}$$

where w_i denotes the parameters of the ACritic network for the intersection i.

Similar to the network design, to allow more capacity in extracting features for each intersection, we use the embedding2 layer to embed S_t into a higher dimension space. To better capture the contribution of each intersection to a certain intersection, we apply the attention layer and outputs the integrated intersection features $m_{i,t}$, then the $m_{i,t}$ will be fed into a fully connected layer and a Relu function, and outputs the Q value of choosing each phase $Q_w(S_t)$. The detail of the ACritic network is presented in Fig. 4.

Specifically, the embedded feature of each intersection can be calculated as:

$$e_{i,t} = \text{relu}\left(fc\left(o_{i,t}\right)\right) \tag{11}$$

The attention module aims to integrate the features of all traffic lights. The output of the attention module is represented as:

$$m_{i,t} = \text{softmax}\left(\frac{fc_{\theta_Q}\left(e_{i,t}\right)\left(fc_{\theta_K}\left(e_t\right)\right)^T}{\sqrt{d_{k_i}}}\right) fc_{\theta_V}\left(e_t\right) \tag{12}$$

where $e_t = \{e_{1,t}, e_{2,t}, \cdots, e_{N,t}\}$ is the embedding features of all traffic lights.

Finally, the soft Q-value is calculated as:

$$Q_{w_i}(S_t) = \text{relu}\left(f_c\left(m_{i,t}\right)\right) \tag{13}$$

Note that the ACritic is used to estimate the separate reward of the intersection itself, other than the global reward of all intersections. We have conducted dozens of experiments to compare the performance between the separate reward and global reward settings, and the separate reward setting achieves a better performance in the arterial traffic control problem. The reason may be that the arterial traffic control problem is a competitive-cooperative scenario, the model can learn to reach the desired Nash Equilibrium point in most cases under the individual reward setting. Therefore, our MASAC does not include an explicit communication mechanism. Of course, our MASAC model may achieve better performance by designing an elaborate communication mechanism, we will investigate it in the near future.

C. The Training Algorithm

Third, we apply the SAC algorithm [25] to train the MARL model to search for better solution space. The original actor-critic algorithm cannot keep a proper balance between solution space exploration and optimal solution exploitation and may fall into a bad local optimal solution because of inappropriate gradient updates. As an alternative, the SAC algorithm attempts to maximize the cumulative while also maximizing entropy to search for more solution space.

Specifically, we apply the discrete SAC algorithm [25], [33] to train our MASAC model for arterial traffic control. The whole model includes one ALight network, two ACritic networks, and two target ACritic networks. Note that the double Q network is used to address the overestimation issue and the target network is used to make the model train more stable [25].

The objective of SAC is to maximize the expected sum of rewards augmented with the expected entropy of the policy:

$$J\left(\pi_{\theta_i}\right) = \sum_{t=0}^{T} \pi_{\theta_i}\left(o_{i,t}\right)^T \left[r_{i,t} - \alpha \log\left(\pi_{\theta_i}\left(o_{i,t}\right)\right)\right] \tag{14}$$

where T is the time horizon, and α is the temperature parameter controls the importance of the entropy term against the reward.

We can optimize the ACritic network by minimizing the loss function as follow:

$$J\left(w_{i,j}\right) = E_{(S_t,a_t)\sim D} \left[\frac{1}{2}\left(\pi_{\theta_i}\left(o_{i,t}\right)^T Q_{w_{i,j}}(S_t) - y\left(r_{i,t}, S_{t+1}\right)\right)^2\right], j = 1, 2 \tag{15}$$

where D is the replay buffer, $Q_{w_{i,j}}(.)$ is the Q value calculated by the ACritic network j of the agent i, and $y(.)$ is the target value calculated by the target ACritic networks.

The target value is given by:

$$\begin{aligned} y\left(r_{i,t}, S_{t+1}\right) &= r_{i,t} + \pi_{\theta_i}\left(o_{i,t+1}\right)^T \\ &\quad \times \left(\min_{j=1,2} Q_{w_{tar,i,j}}(S_{t+1}) - \alpha \log \pi_{\theta_i}\left(o_{i,t+1}\right)\right) \end{aligned} \tag{16}$$

where $\pi_{\theta_i}\left(o_{i,t+1}\right)$ is the probability of choosing each action at the next state $o_{i,t+1}$, $w_{\text{tar},i,j}$ is the parameters for the target ACritic network j of the agent i.

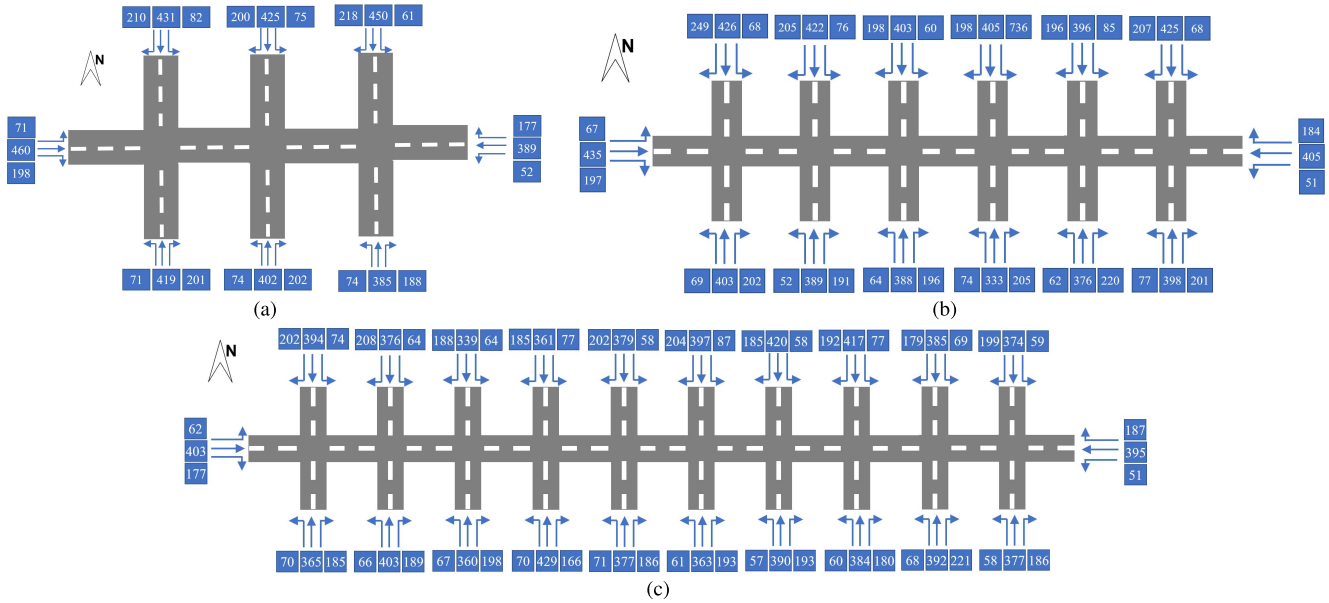


Fig. 5. The diagram of traffic flow: (a) Env-1 × 3grid, (b) Env-1 × 6 grid, (c) Env-1 × 10 grid.

We can update the ACritic by one step of gradient using:

$$\nabla_{w_{i,j}} J\left(w_{i,j}\right) = E_{(S_t,a_t)\sim D}\left[\nabla_{w_{i,j}} \frac{1}{2}\left(\pi_{\theta_i}\left(o_{i,t}\right)^T Q_{w_{i,j}}\left(S_t\right) - y\left(r_{i,t}, S_{t+1}\right)\right)^2\right], \quad j = 1, 2 \tag{17}$$

In the meantime, the Q value of ACtiric networks can help the ALight network to calculate the gradient as follows:

$$\nabla_{\theta_1} J\left(\theta_i\right) = E_{(\theta_2,a_j)\sim D_i}\left[\nabla_{\theta_i}\pi_{\theta_i}\left(o_{i,t}\right)^T\left(-\min_{j=1,2} Q_{w_{i,j}}\left(S_t\right) + \alpha\nabla_{\theta_i}\log\left(\pi_{\theta_i}\left(o_{i,t}\right)\right)\right)\right] \tag{18}$$

The gradient of the temperature parameter can be written as:

$$\nabla_{\alpha_i} J\left(\alpha_i\right) = \pi_{\theta_1}\left(o_{i,t}\right)^T\left[-\nabla_{\alpha_i}\alpha_i\left(\log\left(\pi_{\theta_i}\left(o_{i,t}\right)\right) + \bar{H}_i\right)\right] \tag{19}$$

where $\bar{H}_i$ is a constant vector that represents the target entropy for the agent i.

The target ACritic network is updated as:

$$w_{tar,i,j} = \tau w_{i,j} + (1-\tau) w_{tar,i,j}, \quad j = 1, 2 \tag{20}$$

where τ is the parameter of the soft update method.

To better describe how to apply the SAC algorithm to train the MASAC model, We present the pseudo code of the training algorithm in ***Appendix A-A***.

V. Experimental Results and Comparison

We conduct our experiments in a traffic micro-simulator, simulation of urban mobility (SUMO) [34]. All of the MARL algorithms in this paper are deployed on a machine with Intel Core i9-9920X CPU, 64GB RAM, and 2 NVIDIA GeForce RTX 2080 Ti.

A. Experimental Settings

The experimental settings mainly comprise two parts: the road network setting and traffic flow setting.

For the experimental road network, each intersection is connected with four three-lane roads and the length of each road is 500m. The speed limit on all road lanes is set to be 15m/s. Such a setting is typical and appears in many related studies. It should be pointed out that we conducted dozens of experiments and found that our model and algorithm also work for other arterials. Constrained by the length limit, only the testing results for the above arterial settings are reported.

In the experiments, we utilize three public arterial road network traffic flow datasets[1] to test the DRL algorithms: Env-1 × 3 grid, Env-1 × 6 grid, and Env-1 × 10 grid. The total traffic flows for each dataset are 5514veh/h, 9399veh/h, and 14131 veh/h, respectively. Fig. 5 illustrates the traffic demand of each dataset in detail.

B. Baselines and Performance Indices

1) Baselines: In this paper, to prove the effectiveness of the proposed MASAC method, we choose the multiband [23] and the recent AM-band method [24], which are the SOTA arterial road traffic signal optimization methods in transportation approaches. Besides, to give answers to the last two questions raised in Section I, we select several SOTA multi-agent MARL algorithms as baselines. The implementations of these algorithms are presented in ***Appendix A-B***.

a) MIAC: An individual A2C approach [31] that does not consider the information of adjacent intersections. Each agent updates network parameters with its local observations independently.

b) MASAC: Our proposed DRL method is based on the paradigm of "centralized learning with decentralized execution" in the MDDPG model [18]. Different from the MDDPG

[1] https://github.com/wingsweihua/presslight/tree/master/data/template_lsr

TABLE I
SUMMARY OF THE COMPARISON METHODS

Methods	Centralized learning	Communication module	ALight+ACritic	SAC
MIAC	×	×	×	×
CommLight	√	√	×	×
NeurLight	√	√	×	×
MALight	√	√	×	×
STLight	×	√	×	×
MASAC	√	×	√	√

model, we apply the attention mechanism to extract traffic information and use the SAC training algorithm to stable the method.

c) CommLight: A DRL method based on CommNet model [26]. Compared with the MASAC method, the agent can share local observation with its adjacent intersections. The CommLight method utilize the A2C training algorithm to realize centralized learning.

d) NeurLight: To the best of our knowledge, NeurComm [8] is the SOTA DRL algorithm for large-scale traffic signal control so far. Compared with the CommLight model, NeuroLight learns a more flexibly and accurately communication module through an elaborate neural network.

e) MALight: A DRL method with the attention mechanism to build the communication module. Compared with the CommLight model, MALight learns a communication module with adjacent intersections by an attention mechanism.

f) STLight: A DRL method based on the SOTA DQN-based method STMARL [12], which uses the attention mechanism to build the communication module for multi-intersection traffic signal control.

Note that we consider the attention mechanism in MALight as a part of the communication module, and all DRL algorithms apply the individual network with the individual reward settings. Table I summarizes the characteristic of the comparison DRL algorithms.

2) Performance Indices: We will evaluate the DRL algorithm from three perspectives: training speed, algorithm execution performance, and algorithm transferring ability.

Training speed is mainly used to evaluate the performance of MARL algorithms in the training process. We adopt the following criteria to reflect the training speed:

a) Best convergence episode (BCE): The BCE [8], [11] indicates the sampling efficiency of a DRL algorithm.

b) Best convergence time (BCT): The BCT [13] implies the time complexity of the method.

c) Best convergence point (BCP): The BCP [8], [11] denotes the best value of the reward when the algorithm converges.

To evaluate the execution performance and the transferring ability of the MARL algorithm, we utilize the following criteria:

d) Average queue length (AQL) [9], [11], [12], [13]: The average queue length of all vehicles traveling in the road network.

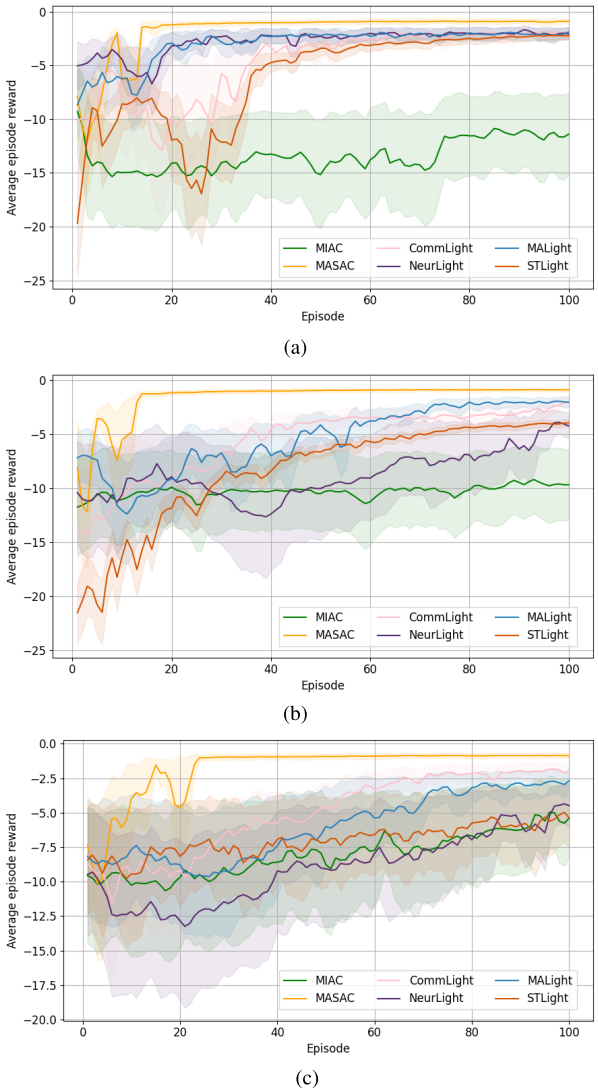


Fig. 6. Training performance comparison of various DRL algorithms on three datasets: (a) Env-1 × 3 grid, (b) Env-1 × 6 grid, (c) Env-1 × 10 grid.

e) Average travel time (ATT) [3], [4], [6], [13]: The average travel time of all vehicles from entering the road network to leaving the road network.

f) Average waiting time (AWT) [35]: the average waiting time of all vehicles in the road network, where the waiting is defined as the vehicle speed $v < 0.1$ m/s.

g) Average speed (AS) [35]: The average speed of all vehicles in the road network.

h) Average number of stops (ANS): The average number of stops of all vehicles in the road network. One of the important purposes of arterial traffic signal control is to reduce the ANS in the road network. We define a stop if the vehicle speed decelerates to less than 0.1m/s.

C. Analysis of the Experimental Results

1) Training Performance Comparison: We summarize the training results of various MARL algorithms over three datasets in Table II, the best values are in bold. For better illustration, we visualize the training curve results in Fig. 6.

TABLE II

TRAINING PERFORMANCE RESULTS OF DIFFERENT MARL ALGORITHMS: ENV-1 × 3 GRID(TOP), ENV-1 × 6 GRID(MIDDLE), ENV-1 × 10 GRID(BOTTOM). BEST VALUES ARE IN BOLD WITH '*'

	MIAC	MASAC	CommLight	NeurLight	MALight	STLight
BCE [epi]	75	**16***	42	26	25	27
BCT [s]	4.86	**0.72**	1.98	0.87	0.87	0.88
BCP	-11.76	**-0.87**	-3.07	-3.56	-3.60	-3.67
BCE [epi]	100	**17***	55	90	79	81
BCT [s]	9.15	**1.61***	4.76	8.06	5.95	7.87
BCP	-9.73	**-0.83***	-3.43	-5.79	-1.91	-2.89
BCE [epi]	100	**24***	68	85	74	84
BCT [s]	13.01	**2.71***	7.80	12.42	8.91	14.87
BCP	-5.27	**-0.81***	-2.70	-5.17	-3.00	-5.23

TABLE III

EXECUTION PERFORMANCE RESULTS OF DIFFERENT MARL ALGORITHMS: ENV-1×3 GRID(TOP), ENV-1×6 GRID(MIDDLE), ENV-1×10 GRID(BOTTOM). BEST VALUES ARE IN BOLD WITH '*'. THE NUMBER IN PARENTHESIS REPRESENTS THE IMPROVEMENT OVER THE AM-BAND METHOD

	Multiband	AM-band	MIAC	MASAC	CommLight	NeurLight	MALight	STLight
AQL [veh]	1.43	1.39	10.80	**0.87*(39%)**	2.15	1.97	1.96	2.19
ATT [s]	112.77	111.73	182.75	**102.07*(9%)**	133.59	128.43	127.75	139.07
AWT [s/veh]	29.58	28.72	106.38	**21.23*(28%)**	49.05	43.81	44.10	53.08
AS [m/s]	10.97	10.21	6.04	**12.08*(10%)**	9.27	9.64	9.68	8.87
ANS	0.98	0.96	1.06	**0.78*(20%)**	1.06	1.05	0.99	1.08
AQL [veh]	1.24	1.18	9.48	**0.84*(32%)**	2.38	2.10	2.09	2.47
ATT [s]	119.42	118.15	210.83	**110.82*(7%)**	147.12	154.00	143.76	159.25
AWT [s/veh]	30.42	30.01	116.39	**23.15*(24%)**	55.00	79.33	52.49	86.41
AS [m/s]	11.16	10.69	6.02	**12.05*(8%)**	9.03	7.57	9.26	7.13
ANS	1.03	1.01	1.74	**0.86*(17%)**	1.25	1.53	1.16	1.67
AQL [veh]	1.33	1.27	5.90	**0.82*(38%)**	1.96	1.95	2.09	2.27
ATT [s]	131.77	130.82	194.52	**117.30*(10%)**	155.00	152.69	167.84	175.82
AWT [s/veh]	36.73	35.11	101.63	**25.89*(30%)**	58.76	56.62	67.62	75.79
AS [m/s]	10.55	10.02	6.64	**11.80*(12%)**	8.91	8.94	7.69	7.08
ANS	1.23	1.09	1.98	**0.95*(23%)**	1.34	1.33	1.74	1.86

The MASAC method achieves the best training performance on BCE, BCT, and BCP metrics in three traffic scenarios.

Besides, the MASAC algorithm possesses a relatively-fixed training episode on three datasets, while the CommLight, NeurLight, MALight, and STLight require more training episodes to converge in complex scenarios. This result indicates that the SAC training algorithm is an effective way to improve the training speed. Note that the MIAC algorithm can not converge in three datasets. The reason may be that each agent only updates its network parameters independently, and did not consider coordination with other intersections.

We present the training speed comparison results in Fig. 7. It is obvious that the MASAC method can converge with the lowest training episodes and the least training time. In addition, the MASAC method shows a greater superiority of training speed in more complex traffic scenarios.

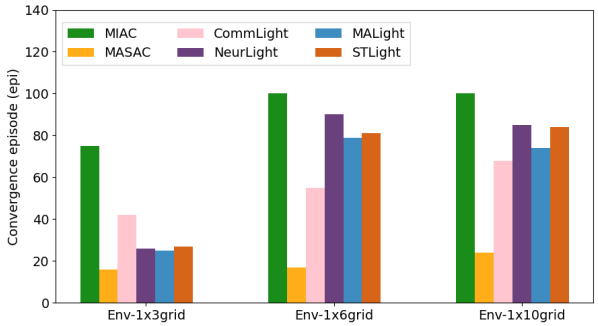
2) Execution Performance Comparison: Table III presents the execution performance of the mutiband-based method and various DRL algorithms in three traffic scenarios. The MASAC method notably outperforms other MARL methods in a big margin. Besides, the MASAC method outperforms the multiband method by 36%, 8%, 27%, 10%, and 20% improvement on AQL, ATT, AWT, and ANS metrics. Besides, the A2C-based DRL algorithms outperform the STLight model in most cases, the reason may be that A2C-based DRL algorithms have stronger solution space ability.

It is interesting to find that the multiband and AM-band method outperform other MARL algorithms other than the MASAC method. The possible reason is that the CommLight, NeurLight, MALight, and STLight method utilize the A2C or DQN to train the algorithm, the algorithm fall into a local optimal solution. We can see from Fig. 6(b) and Fig. 6(c) that CommLight, NeurLight, and the MALight method do not converge to the optimal solution after training 100 episodes.

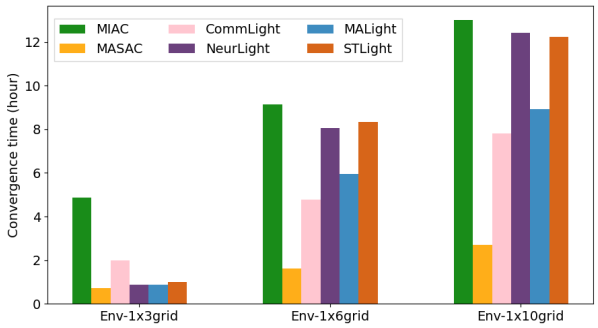
D. Ablation Experiments

In this section, we discuss the effect of centralized learning technique, communication module, attention mechanism, and SAC algorithm contribute to arterial traffic control in detail. For better display, we only present the AQL (veh/lane) index of various MARL algorithms.

1) Centralized Learning: Extensive research [18, 21] has proved centralized learning is a useful tool to improve the DRL algorithms. Therefore, we only conduct simple experiments to analyze the effect of the centralized learning



(a)



(b)

Fig. 7. Training speed comparison of various DRL algorithms: (a) convergence episode, (b) convergence time.

TABLE IV

EFFECT OF CENTRALIZED LEARNING ON EXECUTION RESULTS

Methods	Env-1x3grid	Env-1x6grid	Env-1x10grid
MIAC	10.80	9.48	5.90
MIAC+SAC+ALC	1.21	1.25	2.96
MASAC	**0.87***	**0.84***	**0.82***

Note: ALC: ALight + ACritic.

technique. We compare our proposed MASAC method with the MIAC+SAC+ALC model, the difference between the two models is that MASAC implements the centralized learning technique. The experimental results in Table IV reflect that centralized learning is a promising approach to improve traffic efficiency.

2) Communication Module: To explore the effect of the communication module, we add the SAC and ALight+ACritic module in the CommLight, NeurLight, and MALight algorithms. Table V compares the execution performance of these four models, it is interesting to observe that the communication module does not improve the performance of DRL algorithms. This finding is inconsistent with the conclusion in a recent study [8]. The reason may be that all MARL algorithms in the study [8] do not use the centralized learning technique. By applying the communication module in the arterial traffic control problem, most intersections only added the local observation of two adjacent intersections, the centralized learning module is sufficient to achieve this effect.

3) Attention Mechanism: To analyze the effect of the attention mechanism used in the ALight and ACritic modules,

TABLE V

EXECUTION PERFORMANCE COMPARION OF VARIOUS COMMUNICATION MODULES

Methods	Env-1x3grid	Env-1x6grid	Env-1x10grid
AM-band	1.39	1.18	1.27
MASAC	**0.87***	**0.84***	**0.82***
CommLight+SAC+ALC	0.88	0.86	0.84
NeurLight+SAC+ALC	0.88	0.87	0.86
MALight+SAC+ALC	0.94	0.86	0.85

Note: ALC: ALight + ACritic.

TABLE VI

EXECUTION PERFORMANCE COMPARION OF VARIOUS DRL ALGORITHMS WITH DIFFERENT MODULE COMBINATION

Methods	Env-1x3grid	Env-1x6grid	Env-1x10grid
AM-band	1.39	1.18	1.27
MASAC-SAC-ALC	1.97	2.05	1.97
MASAC-SAC	0.93	1.15	1.03
MASAC-ALC	0.94	0.88	0.86
MASAC	**0.87***	**0.84***	**0.82***
CommLight	2.15	2.38	1.96
CommLight+ALC	0.97	1.17	1.01
CommLight+SAC	0.93	0.86	0.86
CommLight+SAC+ALC	0.88	0.86	0.84
NeurLight	1.97	2.10	1.95
NeurLight+ALC	1.08	0.89	1.09
NeurLight+SAC	0.92	0.87	0.90
NeurLight+SAC+ALC	0.88	0.87	0.86
MALight	1.96	2.09	2.09
MALight+ALC	1.17	1.02	1.15
MALight+SAC	0.93	0.91	0.87
MALight+SAC+ALC	0.94	0.86	0.85

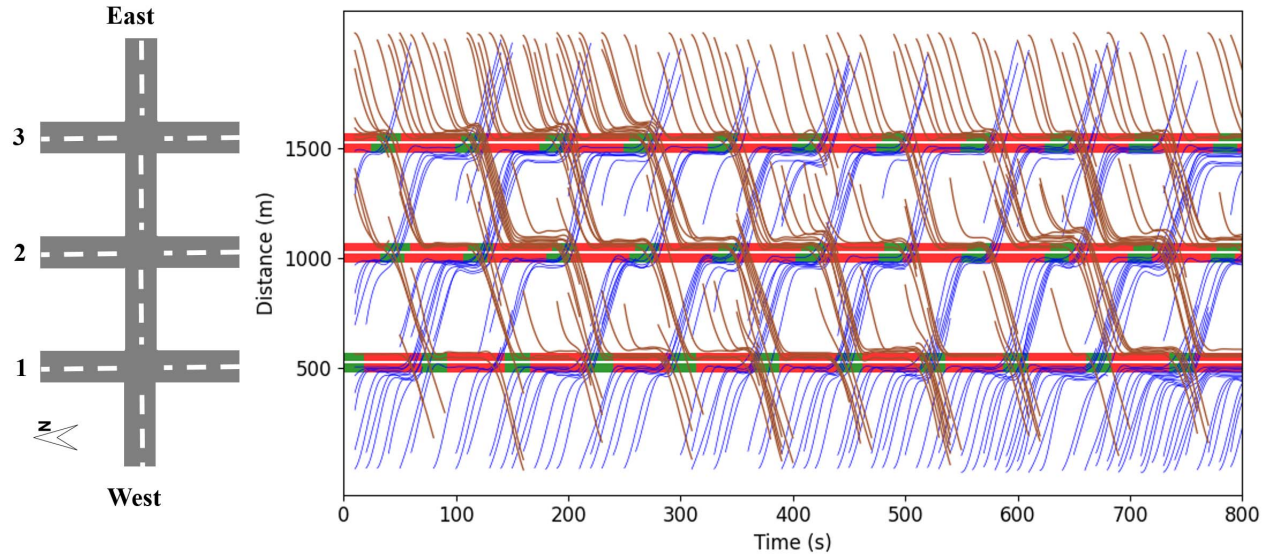
Note: ALC: ALight + ACritic.

we add the ALight+ACritic module for the CommLight, NeurLight, and MALight models. Table VI indicates that the attention mechanism can significantly improve the execution performance of various DRL algorithms. Besides, all DRL algorithms outperform the AM-band method by applying the ALight and ACritic modules.

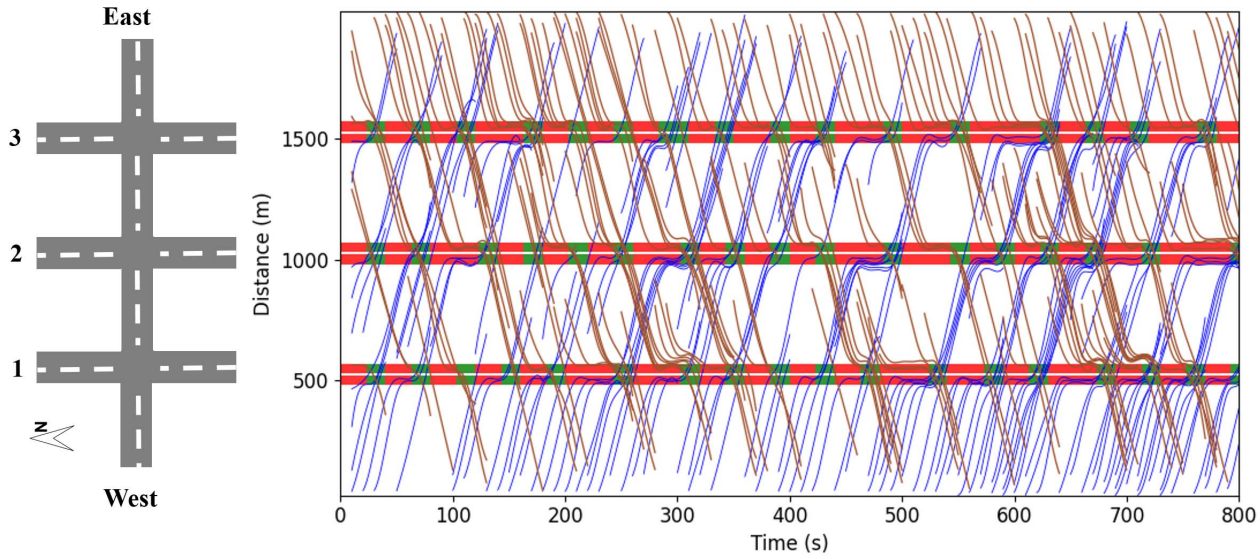
4) SAC Algorithm: By adding the SAC algorithm into the base model of the CommLight, NeurLight, and MALight methods, we can analyze the effect of the SAC algorithm. The experimental results in Table VI reveal that the SAC algorithm can reduce the AQL index in a big margin, all MARL algorithms yield better execution performances than the AM-band method by utilizing the SAC training algorithm. The SAC algorithm plays a greater role in improving the control performance of arterial traffic control than the attention module. Besides, by combining the ALight+ACritic module and the SAC algorithm, most multi-agent DRL algorithms can achieve better execution performances.

E. Arterial Coordination Control Analysis

To better illustrate the arterial coordination control performance of the algorithm, we select the AM-band method and our proposed MASAC method and plot the spatiotemporal trajectories of vehicles for both the outbound and inbound directions. Fig. 8 and Fig. 9 present the spatiotemporal



(a)



(b)

Fig. 8. Spatiotemporal trajectories for vehicles on Env-1 × 3 grid dataset: (a) AM-band method, (b) MASAC method.

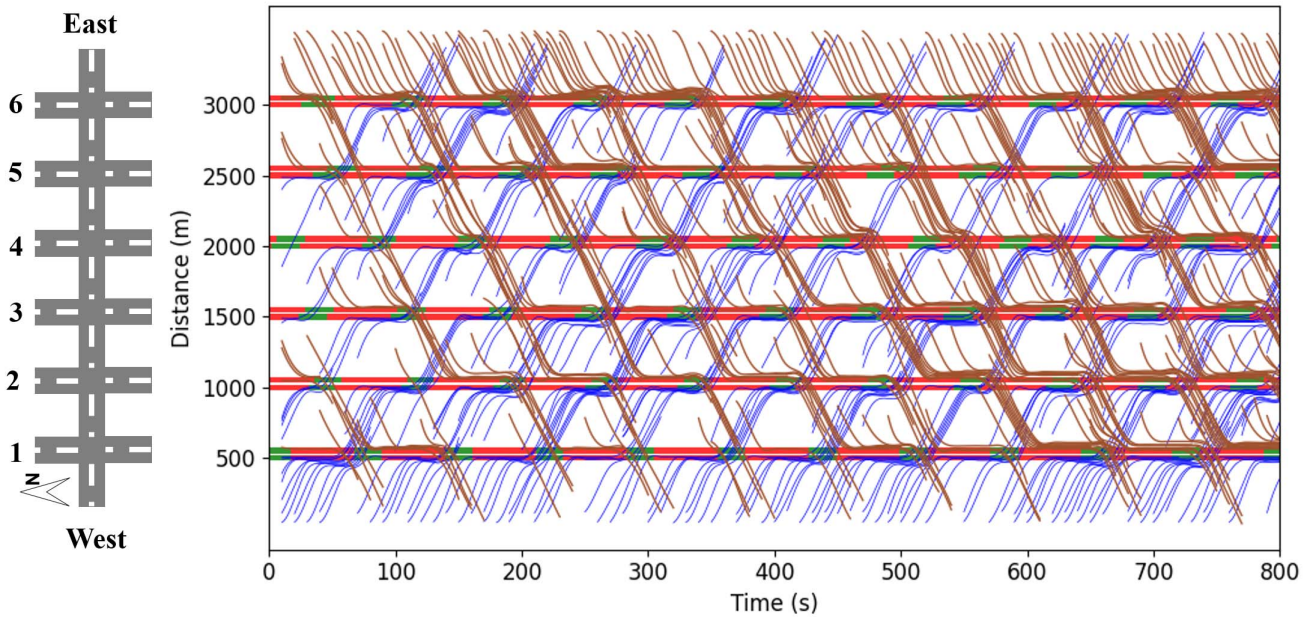
trajectories of the selected two algorithms on Env-1 × 3 grid, Env-1 × 6 grid datasets. In the spatiotemporal trajectories diagram, the larger the slope of the line, the faster the driving speed of the vehicle. It is obvious that the MASAC can learn the phase lag trick to appropriately coordinate arterial traffic signals.

A zoom-in study indicates that our MASAC model does not prefer large yet fixed bandwidths that had been widely adopted by MAXBAND, MULTIBAND, and AM-band models. Here, the portion of the cycle time that allows a platoon of vehicles to pass the arterial without stopping is called the band for the studied direction. The width of the band is called bandwidth.

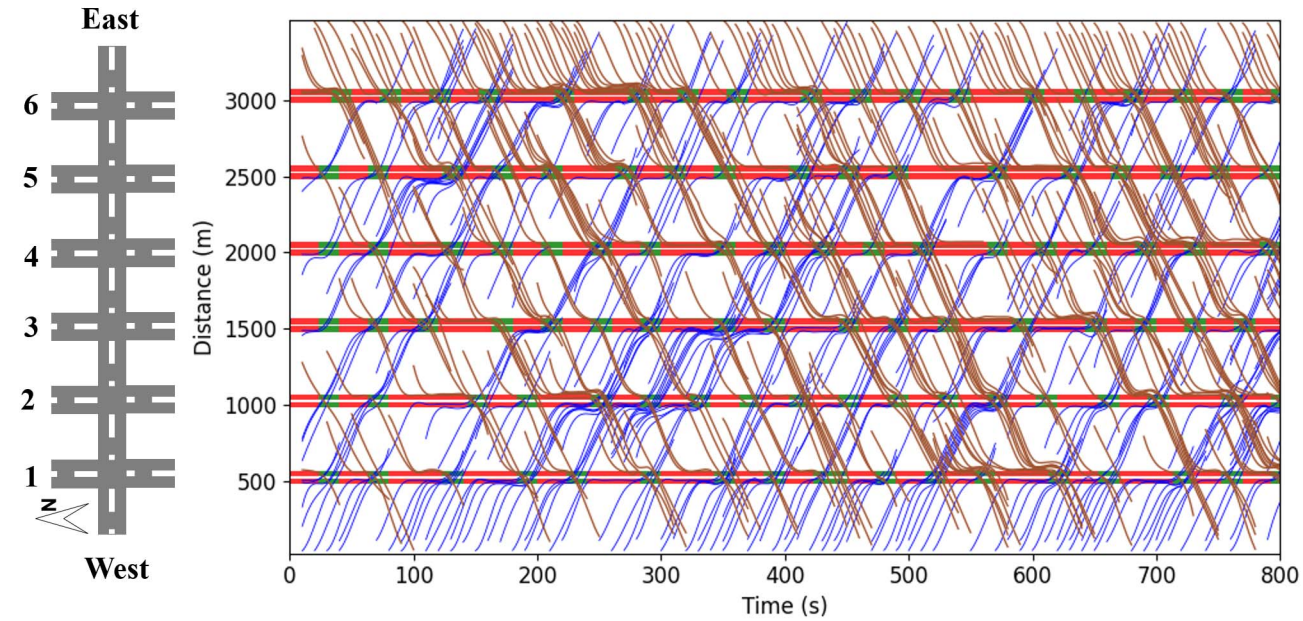
Instead, our MASAC model prefers significantly smaller yet variable bandwidths that allow smaller platoons of vehicles to transverse the arterial. The phase lags between the neighboring intersections are not fixed either. MASAC model will adaptively adjust the bandwidth and phase lags to reduce the delay and stops of vehicles.

We further present the distribution of vehicle travel time and the number of stops in Fig. 10, and Fig. 11, respectively. We observe that the MASAC method can notably reduce the travel time and the queue count of vehicles.The ATT is 111.73 s, 118.15 s, and 130 s on three datasets by applying the AM-band, while the MASAC can reduce the ATT of three datasets to 102.07 s, 110.82 s, and 117.30 s, respectively. The average number of stops (ANS) are 0.96, 1.01, and 1.09 for the AM-band method on three datasets, the MASAC outperforms the AM-band method by 20%, 17%, and 23% on the ANS.

When the inflow rates of the arterial keep constant, the AM model may reach an optimal solution. However, the empirical



(a)



(b)

Fig. 9. Spatiotemporal trajectories for vehicles on the Env-1 × 6 grid dataset: (a) AM-band method, (b) MASAC method.

inflow rate may fluctuate noticeably: see the simulation cases used in this paper and also in many other papers for example. Our MASAC model can adaptively adjust green time length to better handle such flow rate fluctuations and thus achieve better traffic control performance.

TABLE VII
EXECUTION RESULTS OF BEST-TRAINED DRL MODELS IN NEW DATASETS

Methods	Env-1x6grid non-peak hour	Env-1x10grid peak hour
AM-band	1.14	1.38
MASAC	0.87	0.91

F. Search Solution Space Comparison and Algorithm Transferability

To better illustrate the ablation experiments' findings, we visualize the number of optimal solution spaces searched by various DRL algorithms. The state setting in this paper is defined as queue length + average speed + current phase, we consider a state belongs to solution space if the average speed of each lane is over 10m/s.

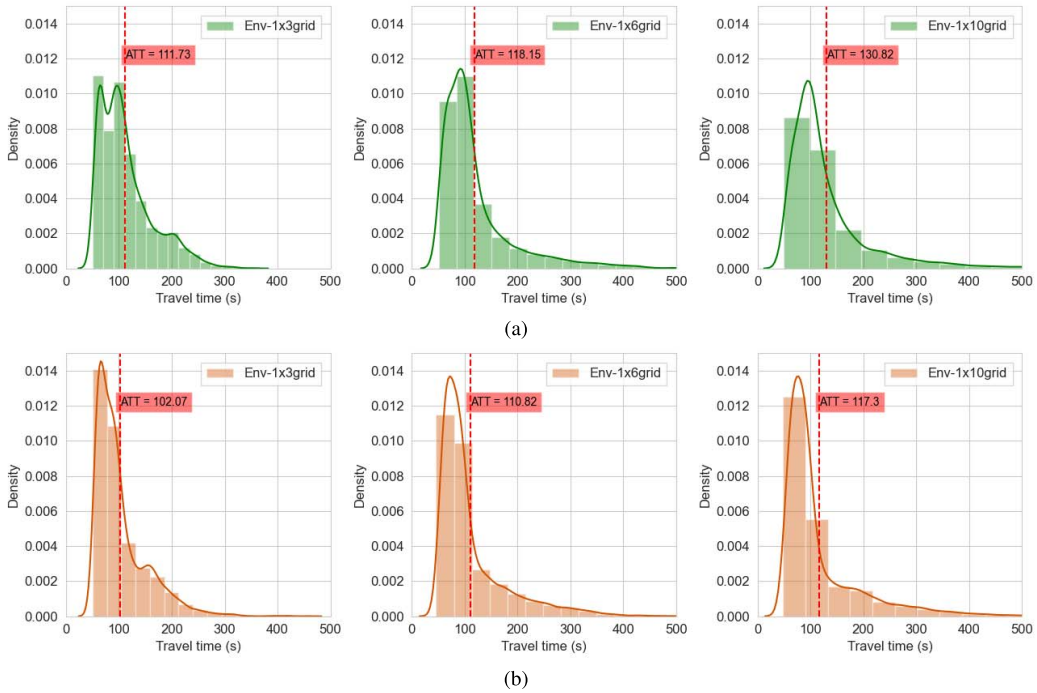


Fig. 10. The distribution of travel time for vehicles on three datasets: (a) AM-band method, (b) MASAC method.

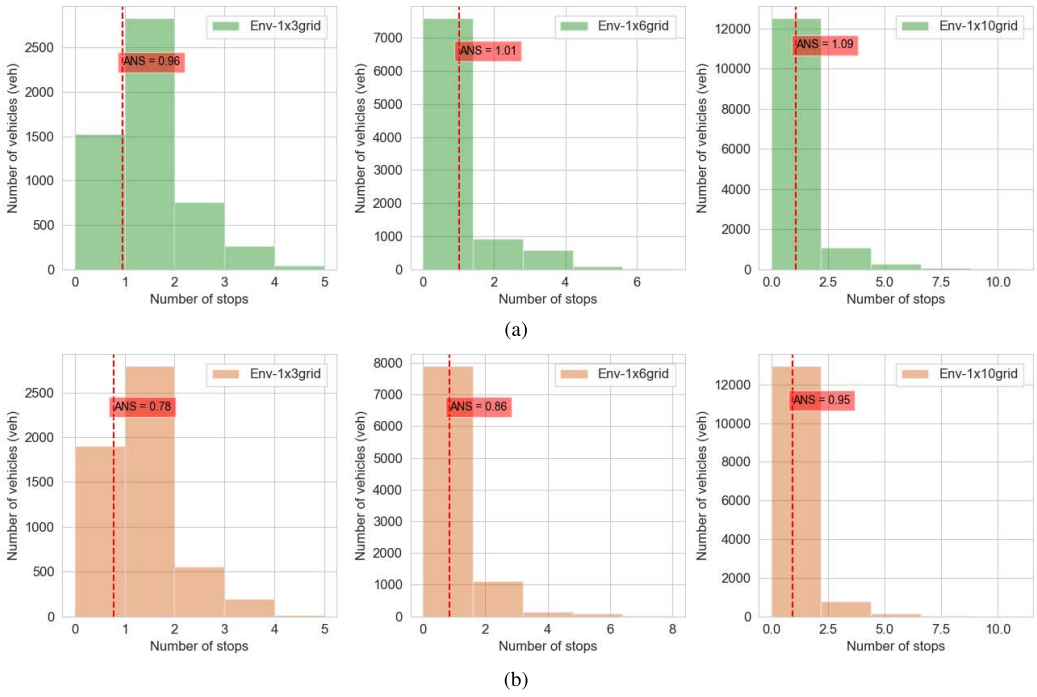


Fig. 11. The distribution of the number of stops for vehicles on three datasets: (a) AM-band method, (b) MASAC method.

Fig. 12 presents the number of search solution spaces of various DRL algorithms on three datasets during the training process. It is obvious that by applying SAC and ALight+ACritic module, the DRL algorithm can search the largest solution spaces. Besides, the result indicates that the execution performance of a DRL algorithm is positively correlated with the search solution spaces size.

To test the transferability of various DRL algorithms, we implement the best model trained in the peak hour to a non-peak hour of Env-1$\times$ 6 grid and apply the best model trained in the non-peak hour to peak hour of Env-1 $\times$ 10 grid. Table VII summarizes the execution performance of various DRL methods in the new traffic scenarios. The experimental results indicate that the best trained MASAC model is able

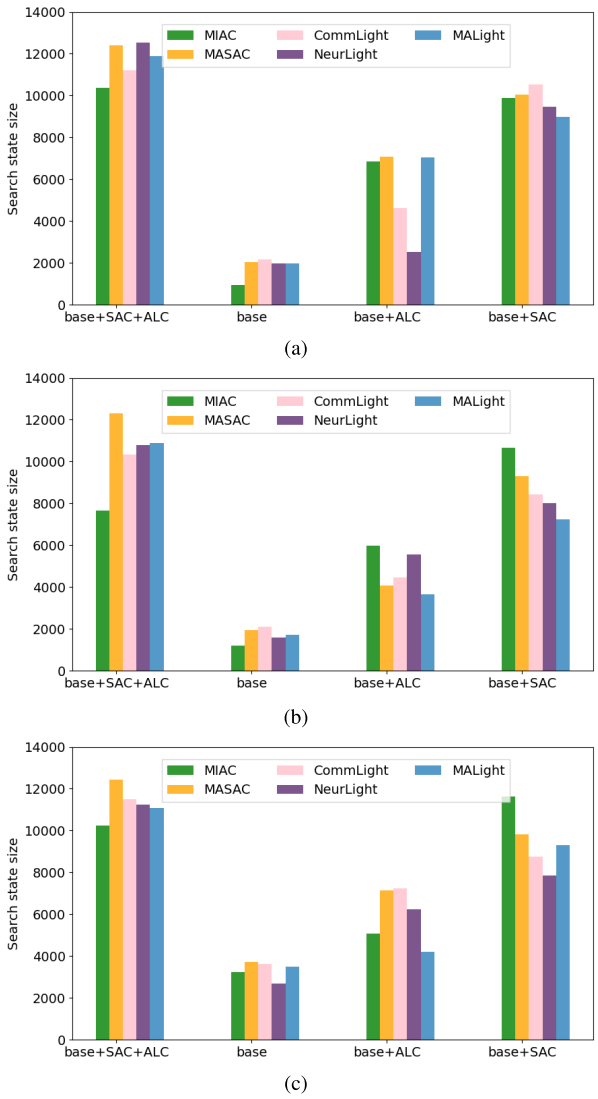


Fig. 12. Search state size comparison of various DRL algorithms (ALC: ALgiht + ACritic): (a) Env-1 × 3 grid, (b) Env-1 × 6 grid, (c) Env-1 × 10 grid.

to transfer to other periods and can outperform the AM-band method.

VI. Conclusion and Discussion

In this paper, we propose the MASAC model to master arterial traffic control. We conduct hundreds of experiments to give answers to the three questions raised at the beginning and draw the following experimental findings:

(1) Centralized training technique is important. A decentralized training method may lead to slow and unstable training results. The communication module may not be useful when an algorithm applied the centralized training technique. Moreover, the attention mechanism can notably help the ALight network and ACritic network extract useful information from large state-action spaces.

(2) The SAC algorithm contributes most to master the proper signal timing patterns for arterial traffic control. By combing the SAC and attention mechanism, most MARL algorithms generally achieve better performances, because these algorithms can search for larger solution spaces.

(3) MASAC method yields better performance than previous MARL algorithms and notably outperformed the multiband-based method in mastering the arterial traffic signal control problem.

These findings provide references on how to design model structures for other MARL problems. For future work, we will investigate how to design an elaborate communication mechanism to boost the performance of the MASAC model. Moreover, our MASAC model can be applied to traffic signal control in the case of larger-scale intersections by replacing the road network and traffic flow files. We will explore how to combine the MASAC model and knowledge transfer so that we can train our MARL models in large-scale road networks with acceptable training time. Constrained by the page length limit, we will write a dedicated paper for regional traffic signal control in the near future.

Appendix A

A. Training Algorithms

Algorithm 1 presents the detailed training process for our proposed MASAC model.

B. Algorithm and Baselines Details

We list the implementations of algorithm and baselines below.

1) MIAC:

$$\begin{aligned}\text{Actor, } \pi_{i,t}\left(o_{i,t}\right) &= \text{softmax}\left(fc\left(fc\left(o_{i,t}\right)\right)\right),\\ \text{Critic, } Q_{i,t}\left(o_{i,t}\right) &= \text{relu}\left(fc\left(fc\left(o_{i,t}\right)\right)\right).\end{aligned}$$

2) MASAC:

$$\begin{aligned}\text{Actor, } \pi_{i,t}\left(o_{i,t}\right) &= \text{softmax}\left(Attend\left(embed1\left(o_{i,t}\right)\right)\right),\\ \text{Critic, } Q_{i,t}\left(S_t\right) &= \text{relu}\left(Attend\left(embed2\left(S_t\right)\right)\right).\end{aligned}$$

3) CommLight:

$$\begin{aligned}\text{Actor, } \pi_{i,t}\left(S_t\right) &= \text{softmax}\\ &\quad\left(fc\left(\text{concat}\left(fc\left(o_{i,t}\right), fc\left(o_{\mathcal{N}_i,t}\right)\right)\right)\right),\\ \text{Critic, } Q_{i,t}\left(S_t\right) &= \text{relu}\left(fc\left(fc\left(S_t\right)\right)\right).\end{aligned}$$

4) NeurLight:

$$\begin{aligned}&\text{Actor, } \pi_{i,t}\left(S_t\right)\\ &\quad= \text{softmax}\left(fc\left(\text{concat}\left(fc\left(o_{i,t}\right), \text{mean}\left(fc\left(o_{\mathcal{N}_i,t}\right)\right)\right.\right.\right.\\ &\qquad\left.\left.\left.\text{mean}\left(fc\left(\pi_{\mathcal{N}_{it-1}}\right)\right)\right)\right)\right),\\ &\text{Critic, } Q_{i,t}\left(S_t\right)\\ &\quad= \text{relu}\left(fc\left(fc\left(S_t\right)\right)\right).\end{aligned}$$

5) MALight:

$$\begin{aligned}\text{Actor, } \pi_{i,t}\left(S_t\right) &= \text{softmax}\\ &\quad\left(fc\left(\text{Attend}\left(fc\left(o_{i,t}\right), fc\left(o_{\mathcal{N}_i,t}\right)\right)\right)\right),\\ \text{Critic, } Q_{i,t}\left(S_t\right) &= \text{relu}\left(fc\left(fc\left(S_t\right)\right)\right).\end{aligned}$$

Algorithm 1: Multi-Agent Attention-Based Soft Actor-Critic (MASAC) Algorithm

Input: θ_i for the ALight network, $w_{i,1}$, $w_{i,2}$ for ACritic networks, temperature paramter α_i, buffer $\mathcal{B}$, time horizon T, max epochs M, agent set $\mathcal{V}$, update frequency f

Result: $\{\theta_i, w_i, \alpha_i\}, v_i \in \mathcal{V}$

Set parameters for target ACritic networks $w_{\text{tar},i,1} \leftarrow w_{i,1}, w_{\text{tar},i,2} \leftarrow w_{i,2}$;
initialization;
for *epoch* $= 1$ **to** M **do**
  for $t = 1$ **to** T **do**
    for $i = 1$ **to** $\|\mathcal{V}\|$ **do**
      observe $o_{i,t}$, S_t;;
      calculate $a_{i,t}$ based on Eq. (9);;
    end
    for $i = 1$ **to** $\|\mathcal{V}\|$ **do**
      execute $a_{i,t}$;;
    end
    store $\{S_t, \{o_{i,t}\}, \{a_{i,t}\}, \{r_{i,t}\}, S_{t+1}, \{o_{i,t+1}\}\}$, $v_i \in \mathcal{V}$ in $\mathcal{B}$
  end
  if $i \,\%\, f == 0$ **then**
    for $i = 1$ **to** $\|\mathcal{V}\|$ **do**
      update ACritic parameters $w_{i,1}$, $w_{i,2}$ based on Eq. (17);;
      update ALight parameters θ_i based on Eq. (18);;
      update temperature parameters α_i based on Eq. (19);;
      update target ACritic parameters $w_{\text{tar},i,1}$, $w_{\text{tar},i,2}$ based on Eq. (20);;
    end
  end
end

TABLE VIII
HYPERPARAMETER SETTINGS FOR MARL ALGORITHMS

Hyperparameters	Value
lr (α)	3e-3
gamma (γ)	0.99
episodes	100
time step (Δt)	10s
horizon (T)	3600s
pre time	60s
memory capacity	1000
batch size	64
τ	0.01
update frequency (f)	20
target entropy ($\bar{H}$)	-log4

6) STLight:

$$\text{Critic}, Q_{i,t}(S_t) = \text{relu}(Attend(fc(S_t))).$$

C. Settings of Hyperparameters

We summarize the hyperparameter settings of various MARL algorithms in Table VIII.

REFERENCES

[1] C. Li, X. Ma, L. Xia, Q. Zhao, and J. Yang, "Fairness control of traffic light via deep reinforcement learning," in *Proc. IEEE 16th Int. Conf. Autom. Sci. Eng. (CASE)*, Aug. 2020, pp. 652–658.

[2] L. Li, Y. Lv, and F.-Y. Wang, "Traffic signal timing via deep reinforcement learning," *IEEE/CAA J. Autom. Sinica*, vol. 3, no. 3, pp. 247–254, Jul. 2016.

[3] A. Oroojlooy, M. Nazari, D. Hajinezhad, and J. Silva, "Attendlight: Universal attention-based reinforcement learning model for traffic signal control," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 4079–4090.

[4] H. Wei, G. Zheng, H. Yao, and Z. Li, "IntelliLight: A reinforcement learning approach for intelligent traffic light control," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 2496–2505.

[5] G. Zheng et al., "Learning phase competition for traffic signal control," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, Nov. 2019, pp. 1963–1972.

[6] H. Wei et al., "PressLight: Learning max pressure control to coordinate traffic signals in arterial network," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 1290–1298.

[7] C. Chen et al., "Toward a thousand lights: Decentralized deep reinforcement learning for large-scale traffic signal control," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 4, 2020, pp. 3414–3421.

[8] T. Chu, S. Chinchali, and S. Katti, "Multi-agent reinforcement learning for networked system control," 2020, *arXiv:2004.01339*. [Online]. Available: https://arxiv.org/abs/2004.01339

[9] T. Chu, J. Wang, L. Codecà, and Z. Li, "Multi-agent deep reinforcement learning for large-scale traffic signal control," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 1086–1095, Mar. 2019.

[10] Z. Li, H. Yu, G. Zhang, S. Dong, and C.-Z. Xu, "Network-wide traffic signal control optimization using a multi-agent deep reinforcement learning," *Transp. Res. C, Emerg. Technol.*, vol. 125, Apr. 2021, Art. no. 103059.

[11] X. Wang, L. Ke, Z. Qiao, and X. Chai, "Large-scale traffic signal control using a novel multiagent reinforcement learning," *IEEE Trans. Cybern.*, vol. 51, no. 1, pp. 174–187, Jan. 2021.

[12] Y. Wang, T. Xu, X. Niu, C. Tan, E. Chen, and H. Xiong, "STMARL: A spatio–temporal multi-agent reinforcement learning approach for cooperative traffic light control," *IEEE Trans. Mobile Comput.*, vol. 21, no. 6, pp. 2228–2242, Jun. 2020.

[13] H. Wei et al., "CoLight: Learning network-level cooperation for traffic signal control," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, Nov. 2019, pp. 1913–1922.

[14] Z. Yu et al., "MaCAR: Urban traffic light control via active multi-agent communication and action rectification," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 2491–2497.

[15] X. Zang, H. Yao, G. Zheng, N. Xu, K. Xu, and Z. Li, "Metalight: Value-based meta-reinforcement learning for traffic signal control," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 1, 2020, pp. 1153–1160.

[16] T. Tan, F. Bao, Y. Deng, A. Jin, Q. Dai, and J. Wang, "Cooperative deep reinforcement learning for large-scale traffic grid signal control," *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2687–2700, Jun. 2019.

[17] Y. Lin, X. Dai, L. Li, and F.-Y. Wang, "An efficient deep reinforcement learning model for urban traffic control," 2018, *arXiv:1808.01876*.

[18] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6382–6393.

[19] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 2, 2018, pp. 2974–2982.

[20] J. Foerster, I. A. Assael, N. de Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 2145–2153.

[21] J. Foerster et al., "Stabilising experience replay for deep multi-agent reinforcement learning," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, Aug. 2017, pp. 1146–1155.

[22] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. ICML*, vol. 80, Jul. 2018, pp. 1861–1870.

[23] N. H. Gartner, S. F. Assman, F. Lasaga, and D. L. Hou, "A multi-band approach to arterial traffic signal optimization," *Transp. Res. B, Methodol.*, vol. 25, no. 1, pp. 55–74, 1991.

[24] C. Zhang, Y. Xie, N. H. Gartner, C. Stamatiadis, and T. Arsava, "AM-band: An asymmetrical multi-band model for arterial traffic signal coordination," *Transp. Res. C, Emerg. Technol.*, vol. 58, pp. 515–531, Sep. 2015.
[25] T. Haarnoja et al., "Soft actor-critic algorithms and applications," 2018, *arXiv:1812.05905*.
[26] S. Sukhbaatar et al., "Learning multiagent communication with backpropagation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 2244–2252.
[27] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[28] V. Mnih et al., "Playing atari with deep reinforcement learning," 2013, *arXiv:1312.5602*.
[29] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
[30] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," 2015, *arXiv:1509.02971*. [Online]. Available: https://arxiv.org/abs/1509.02971
[31] V. Mnih et al., "Asynchronous methods for deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1928–1937.
[32] D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein, "The complexity of decentralized control of Markov decision processes," *Math. Oper. Res.*, vol. 27, no. 4, pp. 819–840, Aug. 2000.
[33] P. Christodoulou, "Soft actor-critic for discrete action settings," 2019, *arXiv:1910.07207*.
[34] D. Krajzewicz, J. Erdmann, M. Behrisch, and L. Bieker, "Recent development and applications of sumo-simulation of urban mobility," *Int. J. Adv. Syst. Meas.*, vol. 5, nos. 3–4, pp. 1–11, 2012.
[35] H. Ge, Y. Song, C. Wu, J. Ren, and G. Tan, "Cooperative deep Q-learning with Q-value transfer for multi-intersection signal control," *IEEE Access*, vol. 7, pp. 40797–40809, 2019.

Feng Mao received the B.S.E.E. and M.S. degrees from Sun Yat-sen University, Guangzhou, China, in 2016 and 2019, respectively. He is currently pursuing the Ph.D. degree with the Department of Automation, Tsinghua University, Beijing. His research interests include AI, traffic data mining, and intelligent transportation systems.

Zhiheng Li (Member, IEEE) received the Ph.D. degree in control science and engineering from Tsinghua University, Beijing, China, in 2009. He is currently an Associate Professor with the Department of Automation, Tsinghua University, and the Graduate School at Shenzhen, Tsinghua University, Shenzhen, China. His research interests include traffic operation, advanced traffic management systems, urban traffic planning, and intelligent transportation systems. He serves as an Associate Editor for the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS.

Yilun Lin received the Ph.D. degree in control science and engineering from the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, China, in 2019. He is currently a Research Scientist with the Shanghai AI Laboratory. His research interests include social computing, smart city, intelligent transportation systems, deep learning, reinforcement learning, and federated learning.

Li Li (Fellow, IEEE) is currently a Professor with the Department of Automation, Tsinghua University, Beijing, China, where he was involved in artificial intelligence, intelligent control and sensing, intelligent transportation systems, and intelligent vehicles. He has published over 150 SCI-indexed journal articles as the first author or the corresponding author. He was a member of the Editorial Advisory Board for *Transportation Research Part C: Emerging Technologies* and a member of the Editorial Board of *Transport Reviews* and *ACTA Automatica Sinica*. He serves as an Associate Editor for IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS and IEEE TRANSACTIONS ON INTELLIGENT VEHICLES.