2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)
October 8-12, 2022, Macau, China

WaveLearner: A Knowledge-Combined Reinforcement Learning to Understand Coordinated Traffic Signal Control along Urban Arteries

Tianyang Han[1], Suxing Lyu[2] and Takashi Oguchi[3]

Abstract— **With the development of detection and computation techniques, the use of reinforcement learning (RL) in traffic signal control problems is widely discussed. After formulating diverse isolated RL agents to control one intersection design, most existing studies tend to directly duplicate isolated agents for large-scale coordinated control problems. However, two questionable challenges are 1) the coordinated control commonly differs from isolated control owing to different objectives; 2) the coordination necessity varies under different traffic demand. Thus, a naive duplication or aggregation of isolated RL agents seems unsatisfactory. In this paper, we focus on the classical arterial control problem to investigate an appropriate coordination strategy. Inspired by green-wave control, a knowledge-combined RL controller is proposed that can predict the potential opportunity of creating non-stop traffic through an artery by matching an ego intersection's phase selection and upstream historical states. Relying on realistic detection, the potential coordination cases can be recorded and rewarded, which can enhance the controller to catch opportunities to create a green wave in further learning. A simulation experiment was conducted to systematically compare the existing coordinated RL methods. According to the results, a promised performance of the proposed method was observed under various traffic conditions.**

I. INTRODUCTION

Traffic signal control is considered the most common means of improving the efficiency and safety of road networks. An isolated intersection signal control can be understood by minimizing the delay of all coming vehicles [1]. By assuming an arriving pattern of coming vehicles based on investigation, the optimum control of an isolated intersection can be achieved theoretically. However, when extending the control method into multiple intersections along one street, the optimization problem becomes completely different. For the so-called arterial network with all signalized intersections, vehicles arbitrarily enter and leave. After passing the first intersection, the random arrival vehicles will be aggregated into near-saturated fleets forwarding to the downstream. Unless the coming vehicles are sufficiently sparse to resist such an aggregation effect, the demand pattern of downstream intersection will have a dramatic variation depending on its upstream intersections. With rapid urbanization, more cities have begun operating signalized arteries. Thus, one emerging research question is "How do traffic signals cooperate among intersections?" [2]. Previous studies have explored answers this question for decades.

In traffic engineering, the coordination problem for an artery begins with defining the traffic demand level. For light traffic, the core task of coordination is to create the green wave, which can allow traffic through the artery or critical path without stopping. Considering only arterial traffic or emergency vehicles with priority, the maximization of the probability of vehicles entering the non-stop traffic is equivalent to delay minimization. In theory, the probability of entering a green wave is modeled using the bandwidth. Bandwidth-based signal controls began with the MAXBAND program [3], which first formulated the bandwidth maximization problem for arterial traffic. Gartner et al. [4] generalized the problem into variable bandwidths for different links between intersections, and the method was called the multi-band approach. Latest bandwidth-based signal controls can be considered as extensions of the MAXBAND and multi-band approaches [5], [6], [7]. For heavy traffic, the objective should not be to create a green wave. An excessive throughput of any traffic movement might result in an over-saturation condition in the downstream. The spill-back queuing in the downstream will cause a blockage such that any control upstream becomes meaningless. In this case, all congested intersections should be aggregated in the perimeter control.

As a result, traffic engineering methods are hard to be applied for diverse traffic pattern. With the development of detection and computation techniques, reinforcement learning (RL) has been introduced for more generalized control. Over the past decades, researchers focused on two main directions [8][9]: 1) Exploring appropriate functional approximations for RL to solve optimum control economically; 2) proposing rational traffic state representation to make the traffic signal control problem understandable for machines. Additionally, very recent studies seem to reach an agreement of using neural networks as a general approximation and representation of traffic states with real-time queuing information [10], [11], [12], [13], [14], [15]. However, all these studies aimed to develop an isolated RL agent to control one intersection. When encountering the network control problem, the isolated RL agent is always duplicated or aggregated to cover the entire network. Wei et al. [16] summarized three coordination strategies in RL-based traffic signal control studies: centralized control, joint-action modeling, and decentralized control.

A centralized design will result in an explosive increase in parameters owing to combining complex network controls into one global agent [17]. A joint-action modeling or

[1]Tianyang Han is with Graduate School of Engineering, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan hanty@iis.u-tokyo.ac.jp

[2]Suxing Lyu is with Center for Spatial Information Science, University of Tokyo, 5-1-5 Kashiwa-no-ha, Kashiwa, Chiba, 277-8568, Japan

[3]Takashi Oguchi is with Institute of Industry Science, University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo, 153-8505, Japan

2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC) | 978-1-6654-6880-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ITSC55140.2022.9922269

local centralized strategy might be a solution to parameter explosions. However, all the intersections cooperate through a deterministic framework (almost adjacency) without considering the variation in traffic [18], [19]. The rationality of joint-action requires more prior considerations of groups and their priorities. Therefore, almost all previous studies could be included in the decentralized category. Although earlier studies always duplicated their agent to cover entire networks, some studies recently presented some sophisticated attempts. On one hand, graph learning has been introduced to learn the relationship between the ego intersection and its adjacency [20], [2]. By extending one intersection of traffic states, a more cooperative control scheme can be produced. On the other hand, the ego intersection and its downstream can be connected using the concept of "pressure" [21]. The pressure can explicitly represent the upstream demand and downstream supply, such that the blockage of congested downstream can be easily recognized by independent RL agents [22], [11], [12]. Compared with the solutions in traffic engineering, RL-based cooperation methods rarely consider arterial control but state their works as general solutions for "network control". Even with considering upstream-downstream interaction, such coordination cannot produce non-stop opportunity for mainstream traffic as the idea in traffic engineering. Wei et. al. [23] proposed a decentralized RL method for arterial control, nevertheless, instead of agents' collective behavior, this study focused on making agents transferable. Although this method was applied to arterial control, it is essentially the same as the aforementioned duplication strategy.

From the perspective of traffic engineering, we argue that the above attempts were unsatisfactory owing to the following reasons: 1) The cooperative behavior is not interpretable and contributable with a general formula. Meanwhile, the objective of coordination will be completely different for varying traffic patterns. 2) The interaction between the upstream and downstream is simplified with simultaneous traffic states. However, the impact from upstream to downstream is not a pure spatial evolution but has a time offset caused by shockwave propagation. 3) The data condition is not realistic. The queue-based RL signal requires the precise queue number in the entering lane, which is still an open problem to be estimated using realistic detection.

To fill these gaps, this paper proposes a knowledge-combined green-wave rewarding mechanism in conjunction with our previous study [15] on the classical arterial control problem. Instead of precise queue state, using an estimated queue could be more feasible. Learning a realistic outcome from detection and queue estimation model can provide an extra robustness to the learning machine. Inspired by episodic learning in robotics, we developed a modularity in which the coordination cases can be recorded and rewarded through realistic detection, which can enhance the controller in determining an opportunity to create a green wave through further learning. Expected to learn the behavior of green wave control, the proposed method is called WaveLearner. The green-wave reward is set to be proportional to the estimated queuing vehicle number, by which the RL controller can behave diversely according to traffic patterns. For a light traffic, the green-wave reward is small owing to sparse vehicles, and the reward is more significant for a higher arterial traffic under saturation. When a blockage occurs in an artery, the rewarding will be invalid since there is no chance to create the non-stop traffic. In summary, our contributions are outlined as follows:

- The proposed WaveLearner is an interpretable combination of traffic engineering and RL. Through matching the upstream and downstream traffic, a green wave reward is provided for episodic training of the RL controller.
- The WaveLearner framework is based on our previous study [15] on combining queuing estimation methods with RL-based isolated traffic signal control. A decentralized mult-agent design inherited diverse benefits from traffic engineering methods with similar data requirement which indicates it promising practicability.
- Comprehensive experiments were conducted on two synthetic arterial networks. Simulating various traffic, we showed the necessity of introducing our green-wave rewarding mechanism compared with other coordination methods. The proposed WaveLearner was also demonstrated to efficiently maintain stable and superior performance as an extension of an isolated controller.

II. METHODOLOGY

A. Preliminary

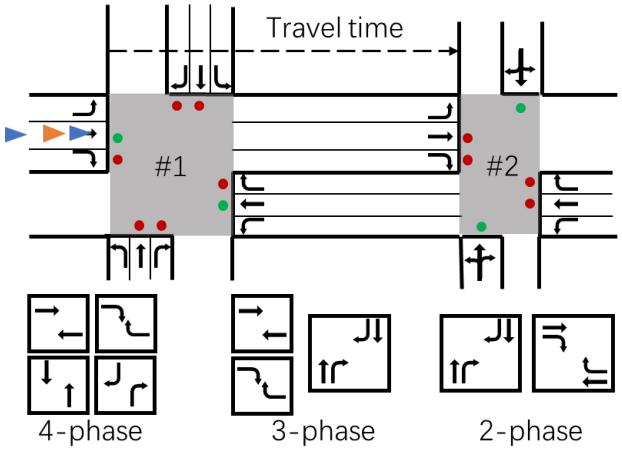

Fig. 1. Arterial networks with two signalized intersections

Isolated Traffic Signal Control. Considering one intersection as the left intersection #1 in Fig. 1, the traffic entering the intersection from one approach to another is called a **movement**. Note that in most practical cases, the left-turn movement for a right-hand-driving road is out of traffic light control with dedicated lane owing to no conflict. Therefore, for intersection #1, only eight movements are under control. The core task of signal control for an intersection is to temporally separate the conflicted movement by aggregating

them into **phases** and provide the permission sequentially for them. As shown in the bottom part of Fig. 1, four-, three-, or two-phase plans are common practical settings. Furthermore, a **signal plan** can be represented by a sequence of phases p and their duration T $\{(p_n, T_n)\} = (p_1, T_1), ..., (p_n, T_n)$. In traffic engineering, a cyclic fixed timing is commonly applied to repeat a fixed sequence of phases and their fixed duration. With the provided composition of phases in one cycle, the signal timing problem can be converted into determining the **cycle** length and **split** for each phase to optimize the efficiency of an isolated intersection [1].

Coordinated Traffic Signal Control. As shown in Fig. 1, a double-intersection system is the simplest coordination scenario. A coordinated control can be defined as the control to optimize the overall performance of multiple intersections. An intuitive concept called **green-wave control** has been proposed in traffic engineering. Considering only one fleet leaving intersection #1 for intersection #2, the coordinated control will provide permission for the fleet as it is arriving at the stop line of intersection #2. Thus, the fleet passing intersection #1 can go through intersection #2 without stopping. Such a non-stop traffic is so-called a green wave. Theoretically, after obtaining cycle length and splits, the parameter **offset** is introduced to optimize the switching timing in coordination.

B. Queue-based RL Controller

An RL controller for traffic signal can make a choice from predefined phases with a fixed frequency to solve both phase and duration in the problem stated above. A general RL is a autonomous control method to implement responsive actions (a) with observed (s) states. If the effect of actions can be measured as the reward, we can estimate the quality (q) of actions under states using the Bellman Equation 1 considering current benefits and potential future gains.

$$q(s, a) = r + \gamma \max q(s\prime, a\prime) \tag{1}$$

Where $q(s\prime, a\prime)$ is the quality of future action $a\prime$ under future states $s\prime$, $\gamma \leq 1$ is the discount factor representing the importance of future benefit. To employ RL-based methods, we should represent the signal control problem as a Markovian Process $\{\boldsymbol{s, a, r}\}$.

- **States** s: Queuing vehicles of each entering lane. With a realistic detection as an inductive loop or any cross-section detectors, the queuing number can be estimated and predicted using a simple input–output model [24].
- **Actions** a: Selecting the phase. In the arterial control problem, a three-phase signal composition is the common usage as shown in Fig. 1 to separate major and mix minor street traffic.
- **Rewards** r: The negative total number of waiting vehicles and green-wave rewarding. The justification of the former part was proved in previous study [10] that it is equivalent to the travel time minimization.

One real-time RL controller will take action a after observing the environment states s. Subsequently, the environment states will transit to the next state s' and feedback a reward r for further learning. The posterior s' and r brings difficulty for real-time continual learning that our previous study [15] introduced queue-estimation methods to predict the future queuing as s' and consequently calculated the total waiting vehicle number as reward r. The framework of the proposed RL agent is indicated in the blue rectangular part in Fig. 2.

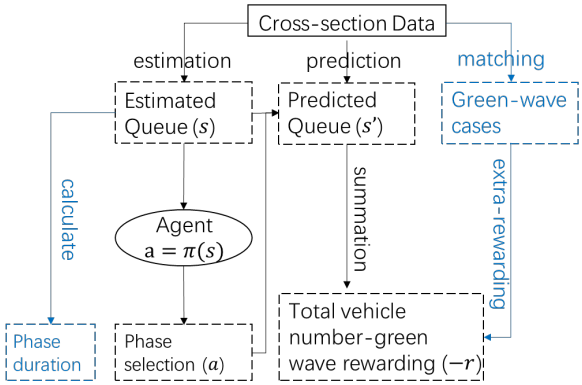

Fig. 2. Frameworks of the proposed decentralized agent for one intersection

On the top of previous work, we add two module to extend our isolated controller to arterial control problem, that marked in blue in the figure above.

C. Changeable Phase Duration

In most previous queue-based RL controllers, the phase duration is almost set as a pre-given fixed value. However, Wu et al. [12] compared different phase duration settings and observed a significant impact on control performance. Newell [25] determined that a given green time of signals should be tightly serve the queue formed before and during the current phase. Thus, the number of current and near-future queuing vehicles is sufficient to produce an optimum next phase and duration. Even without reliable future prediction, a current number of queuing vehicles can be the evidence for selecting a phase with fixed duration as the longest-queue-first (LQF) control that greedily provides permission for the phase with heavier current demand [26]. Principly, with s and s' in our methodology, it is conviced to produce an rational phase duration. Zhao et al. [27] proposed a modified control logic in addition to phase selection. After the phase selection by an RL, the phase duration is calculated using the number of vehicles. Their experiments indicated the significance of the changeable phase duration in reducing average travel time and creating green-wave control. The study also concluded that the changeable phase duration can hasten the convergence of the RL when compared with fixed-phase-duration methods. In this paper, we follow the concept of Zhao's. However, without communication among the connected vehicles, the phase duration cannot be explicitly calculated by waiting vehicles. For simplicity, we propose a linear phase duration $T(s, a)$ that is proportional to the estimated queuing number included in states s corresponding to the permitted movements indicated by the selected phase a (Eq. 2).

$$T(s,a) = \lceil h \times \max\{s_l\}/\tau \rceil \times \tau, \quad l \in L(a) \tag{2}$$

Where h is a proportional coefficient that represents the dispersion headway of queuing fleets physically, and s_l is the queuing number of lane l, and the lane l should be one of the entering lanes permitted by the selected phase a. Considering the decision making frequency τ, that the discrete time duration between s and s', a ceiling operation $\lceil \cdot \rceil$ should be introduced to ensure T as an integer multiplier of τ. Such changeable phase duration aims at dispersing the current queuing vehicles as much as possible. This treatment directly avoid the vehicle entering the entering vehicle waiting more than once.

D. Green-wave Rewarding

Inspired by episodic learning [28] that uses success and failure to train the RL, a direct green-wave rewarding is introduced as extra evidence to decide signal phase. In general, the proposed mechanism will provide reward for potential opportunity of creating green-wave to enhance the controller ability of coordination. In traffic engineering, we can defined a green wave by offset equal to travel time between two intersection. Here, the first step to catch potential opportunity of green wave is to derive travel time and offset.

As shown in Figure 3, we can estimate the travel time of one waiting vehicle from upstream to ego intersection by three parts: waiting for start-up, acceleration, and driving with desired speed.

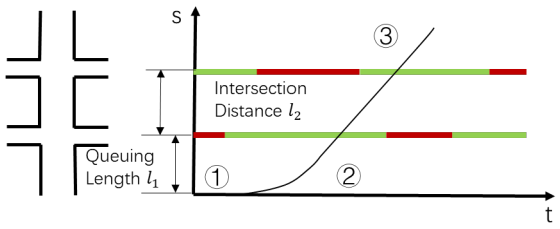

Fig. 3. Travel time of the vehicle in the green wave

At the beginning of the upstream green phase, the waiting vehicle will first wait for the vehicles ahead to start up, that shockwave back propagation from the stop-line to its position [29] in ① in Fig. 3. The dispersion wave speed can be calculated using the saturation flow q_S, jam density k_J and critical density k_C of this road section as $q_S/(k_C - k_J)$ [1]. Note that all the parameters above are related to the fundamental diagram of the road facility and independent of the time and traffic demand that the wave speed can be considered a constant. After the shock-wave reaches the waiting vehicle, the vehicle should startup and accelerate to the desire speed as in ② in Fig. 3. Finally, owing to the non-stop characteristics of the green wave, the vehicle can maintain the desired speed as it passes the ego intersection as in ③. Assuming a constant desired speed v_{Desire}, constant acceleration $acce$, and no overtaking behavior of all vehicles, the estimated travel time of any vehicle under green wave control $\hat{TT}$ can be calculated using only the queuing length ahead l_1.

$$\hat{TT}(l_1) = \frac{l_1(k_C - k_J)}{q_S} + \frac{v_{Desire}}{acce} + \frac{l_1 + l_2 - v_{Desire}^2/acce}{v_{Desire}} \tag{3}$$

To make sure all vehicle waiting currently included in consideration, we choose to estimate the travel time of the vehicle at the end of queuing. Thus, the argument l_1 in equation 3 could be replaced by

$$l_1 = s/k_J \tag{4}$$

Where s is the queuing number state as before. Then we can considering to match the estimated travel time and offset. Let's define the offset will be the time difference between green start of upstream and ego intersection. Since the offset is also a integer multiplier of τ, we can search the matching with τ length step at current time t. Naming upstream intersection of intersection i as $i-1$ with the coordination lane l_c, we can design a green-wave matching with historical data of upstream intersection as follow.

Algorithm 1 Matching the potential Green-wave

Input: historical states of upstream $\{s_{i-1,l_c}\}$, maximum iteration number N

Output: update reward $r_i(t)$ regarding green wave

```
for n = 1 : N do                                          ▷ search back
    if upstream action a_{i-1}(t − nτ) permit through then;
        calculate l̂_1(t − nτ) by s_{i-1,l_c}(t − nτ) (Eq. 4);
        calculate T̂T by l̂_1(t − nτ) (Eq. 3);
        if n × τ = T̂T then                               ▷ confirm green wave
            update r_i(t) (Eq. 5);                        ▷ reward green wave
            breaks;
```

Note that the N-step historical memory is request for the matching above which will bring a slight more requirement on data storage than our previous work. However, when considering the memory also required by replay of RL, a smaller N than the RL memory size makes no extra requirement. Finally, once the matching confirm the existence of green wave opportunity, we can reward the coordinated phase of ego intersections according to size of the upstream platoon using Eq. 5.

$$r_i(t) \leftarrow r_i(t) + \beta_{GW} \times s_{i-1,l_c}(t - n \times \tau) \tag{5}$$

Where β_{GW} is the coefficient of green-wave $\in [0,1]$. Essentially, the green wave rewarding does not directly make RL agents understand how to create a green wave, but it enhances the memory on the ego queuing when green is likely to occur. Therefore, the positive weight or the priority should not be larger than the current estimated queuing number.

[1] q_S i the maximum vehicle number pass within unit time and k_C is the number of vehicle occupied unit length when achieving q_S. k_J is the maximum number of vehicle stop within unit length.

III. Experiments

To demonstrate the effectiveness of the proposed method, we conducted a simulation experiment based on the Simulation of Urban MObility (SUMO).

A. Comparison Models

To demonstrate the control performance of the proposed WaveLearner, we introduce several traffic engineering methods and RL controllers are introduced as comparison models.

- **Webster**. A modification from [1]. For each intersection included in an artery, the traffic is controlled using the three-phase plan in Fig. 1 calibrated by an aggregated flow of synthetic scenario.
- **LQF**. A modification of LQF from [26] using our detection condition. The movements are aggregated with the same three phases in Webster's method, and green is always given to the phase with longest queue for every τ seconds.
- **RHS**. A modification of the RHS from [25]. In addition to the FIX control, for every τ seconds, the current queuing number is estimated for the current phase. If no vehicle is waiting, the signal is switched, or the phase duration will extend by τ seconds.
- **QLight**. A modification from [10], [13] that uses a lane-wise queuing number and negative summation of queuing number as the state and reward in reinforcement learning. To maintain fairness, the queuing number is estimated using an input–output model, and the action corresponds with the three-phase selection in Webster's method.
- **QueueLearner**. The queue-based RL controller in our previous study [15] is directly implemented for each intersection. The difference from QLight is that the future states are predicted using an input–output model with constant input assumption.

QueueLearner and QLight can be considered comprehensive approximators of LQF and Webster with or without knowledge-driven near future prediction. The proposed WaveLearner can be understand as an extension of QueueLearner with changeable phase duration form [27] and original green-wave rewarding mechanism in arterial control.

B. Experiment Setting

For an experimental validation of the proposed methods, two synthetic arterial networks were constructed in the simulation. **SHORT** was a double-intersection system with four 500-m approach. The arterial was along with east–west direction. The lane for each intersection was aligned as in #2 intersection in Fig. 1. All the vehicle entered the network from one marginal point and left from another marginal point. All the driving behavior modeling was set as the default of SUMO [2]. The short arterial network was constructed as the least unit of coordination control. Thus, the different methods of coordination in the RL context are discussed using this network. **LONG** was an extension of **SHORT**; the number of intersection was expanded to 5, and each intersection was aligned as in #1 intersection in Fig. 1. The five-intersection arterial network was constructed to reproduce a real-world scenario. The corresponding experiments covered both RL and traffic engineering methods.

Another part of the simulation was the traffic demand. In this study, a systematic traffic demand generation process was introduced based on gamma-distributed headway [31]. Using the SHORT network as an example, the network traffic could be divided by 30 origin–destination (OD) pairs according to our assumption. Since no route choice was involved in this scenario, we considered the traffic generated with the same OD i had the headway subjected to the same Gamma distribution Γ as below.

$$headway_i \sim \Gamma(\alpha_i, \lambda_i) \tag{6}$$

Where α and λ are the shape and scale parameters of the gamma distribution. Considering the SHORT network was a part of an urban corridor, the traffic through all east–west links could be considered the arterial traffic. To artificially create an artery, we set the arterial traffic with $\frac{\alpha}{5}$. Subsequently, the traffic was generated with only two different gamma distributions. The mean value of headway was expressed as $\alpha \times \lambda$ and the variation coefficient is λ. By adjusting α and λ, the different randomness of arriving vehicles could be realized in our simulation.

Finally, to measure the performance of control methods, the multi-entry-exit detector[3] in SUMO was introduced. For overall performance evaluation the entries were set 250 m upstream on entering approaches of and the exits were set 250 downstream on exiting approaches. The entire arterial network was aggregated as an area using the setting above. The following metrics were included in this study.

- **meanTravelTime** (TT). The average time vehicles require from entering to exiting the area.
- **meanSpeed** (TS).The average speed of vehicles travel through the area.
- **meanHaltsPerVehicle** (halts). The average number of halts per vehicle passing the area.
- **vehicleSum** (vSum). The vehicle count passing the area.

C. Result Analysis

As the design aforementioned, two groups of experiments are included in our discussion: 1) using **SHORT** networks to demonstrate difference among RL coordination strategies; 2) using **LONG** networks to simulate control in the realistic artery and comparing the performance of our method with both traffic engineering methods and the RL controller in our previous study.

The first comparison was to examine different strategies to coordinate RL agents, including centralized, local centralized, and decentralized control strategies. In this study, a directly rewarding green-wave was proposed as a supplement

[2]The reader could be refer to [30] to understand the characteristics of SUMO traffic flow.

[3]https://sumo.dlr.de/docs/Simulation/Output/Multi-Entry-Exit_Detectors_(E3).html

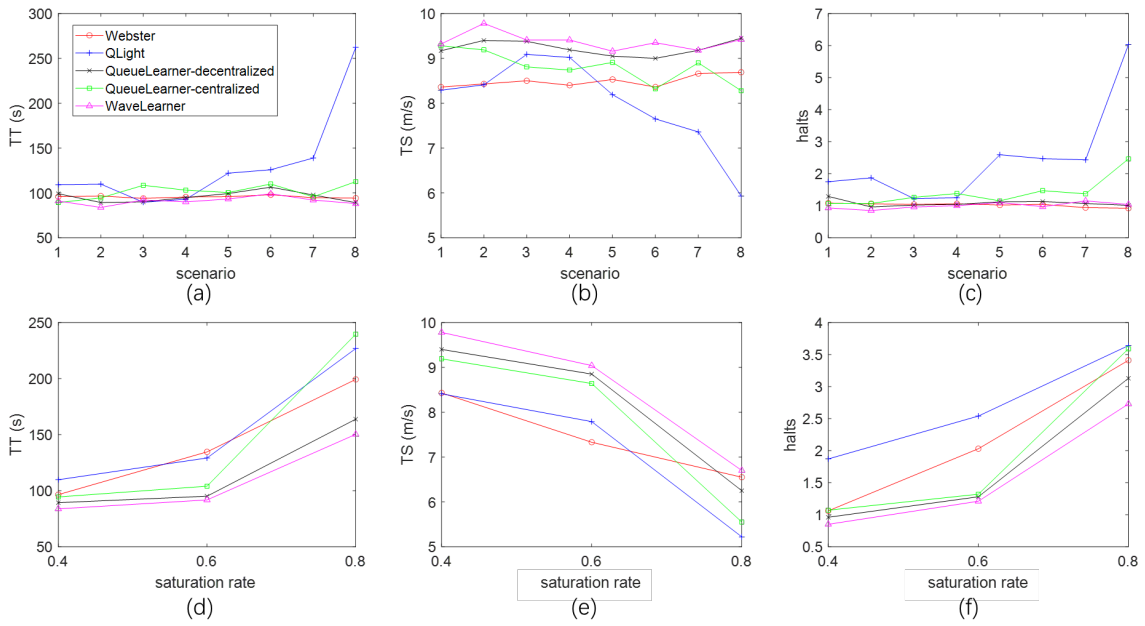

Fig. 4. Comparison of different coordinated RL controllers using the SHORT network: (a-c) test through scenarios of different randomness with the same average arriving rate of $\alpha\lambda = 200$ and the variation coefficient $\lambda = \{20, 25, 40, 50, 100, 200\}$ from scenarios 1-6, a time-variate average headway in scenario 7-8 but maintaining the same hourly average. (d-f) Test through scenarios of different demand levels for a saturation rate of 0.4 case is with the same average headway $\alpha = 8; \lambda = 25$ as in (a-c); the demand larger demand was generated by $\alpha = 6$ and $4; \lambda = 25$

to decentralized control. Thus, the first experiment aimed to examine the rationality and necessity of the proposed methods. To fair comparison, we used the **SHORT** networks to examine WaveLearner without a changeable phase duration, similar to the decentralized and centralized QueueLeaner. The joint action method is equivalent to centralized control with only two intersections. As references, the QLight and Webster's method were used to re-examine the conclusion in [15]. Using TT, TS, and halt, the evaluation of controller's performance under scenarios with different arriving vehicles randomness and demand levels are shown in Fig. 4.

As shown in Fig. 4 (a-c), the conclusion in [15] was verified that without theoretical traffic states transition modeling, a pure model-free RL, QLight, exhibited poor performance. A better performance of such model-free scheme might be achieved after enough pre-training process as previous study [11]. In comparison with centralized and decentralized QueueLearner, a green-wave rewarding resulted in a more stable performance through all scenarios. Compared with the Webster method pre-calibrated by generated traffic demand, the proposed WaveLearner achieved almost the same travel time and halts and a better travel speed. Additionally, with a proportionally higher demand scenario in Fig. 4 (d-f), the advantage of WaveLearner over Webster was significant owing to the delay minimization being inefficient for near-saturated traffic and invalid for over-saturated traffic. Therefore, we can conclude that the proposed green-wave rewarding coordination is more effective than decentralized or centralized control. Thus, we validated two common phenomena: 1) the QLight without sufficient pre-training consistently has difficulty responding to various traffic scenarios; 2) the comparison also revealed consistency with previous studies that decentralized control always outperforms centralized or joint-action control. Therefore, for convenience, in further experiments (Table I), we did not compare the QLight and centralized QueueLearner. The QueueLearner appearing later was only decentralized.

Subsequently, to verify the control performance of our methods with realistic longer arterial networks and demonstrate how they can maintain superiority than simply duplicating the isolated agent, we conducted the experiment using the **LONG** network. In the earlier experiments, the randomness of vehicle arriving seemed less significant than demand level. In the second group of experiments, only different saturation rate was discussed in scenarios design. To consider both overestimation and underestimation of demand case, we pre-calibrated the Webster and corresponding RHS methods using the scenario with a 0.75 saturation rate ($\alpha = 6; \lambda = 50$). Subsequently, we tested all the comparison models with saturation rates of 0.5 ($\alpha = 8; \lambda = 50$) and 1 ($\alpha = 4; \lambda = 50$) saturation rate.Since an 1 saturation rate scenario was included, we also focused on the throughput of network measured using vSum in addition to TT, TS, and halt. The experiment results are shown in Table I.

According to the result, the proposed WaveLearner occupied most of best performances. Additionally, the obvious poorer performance of decentralized QueueLearner proved the necessity of the enhancement in this study. QueueLearner performed worst in heaviest traffic demand scenario. The spilling back queuing status invalidated the queue-based traffic representation, which further resulted in difficulty for RL's decision making and continual learning. Moreover,

TABLE I

MODEL COMPARISON USING LONG NETWORK (BEST WITH BOLD)

saturation rate	methods	TT	TS	halts	vSum
0.5	WaveLearner	**161.26**	**9.92**	1.47	**1123**
	QueueLearner	215.2	8.44	1.32	1070
	Webster	183.08	8.7	1.36	1121
	LQF	222.5	8.78	**1.17**	1089
	RHS	164.97	9.73	1.56	1120
0.75	WaveLearner	**184.22**	**9.01**	1.4	1499
	QueueLearner	253.35	7.78	2.57	1488
	Webster	189.8	8.65	1.46	**1512**
	LQF	261.37	8.59	2.13	1476
	RHS	194.16	8.95	**1.24**	1492
1	WaveLearner	**197.93**	8.54	**1.65**	**2296**
	QueueLearner	351.08	6.96	3.81	2055
	Webster	248.02	7.91	2.04	2190
	LQF	203.01	**8.65**	2.53	2290
	RHS	217.9	8.22	1.77	2295

the green-wave rewarding mechanism of WaveLearner can predict the large coming vehicle from upstream intersection, which can be considered as extra evidence of control even with current detection invalid. To further clarify the efficiency of the proposed WaveLearner that maintain various inherited advantages from traffic engineering methods, we still require to interpret the observation about other methods and our failure in Table I.

- At the middle-level demand, Webster had only a slight disadvantage and RHS achieving significantly fewer average halts owing to the correct calibration. This indicated that with proper calibration, even simple methods from decades ago can be effective. Therefore, instead of on the high performance of machine learning, the focus of this study was to demonstrate the fast and automatized learning ability of an RL controller in dynamic environment. The slight superiority of the proposed WaveLearner results only from the randomness of vehicle arriving behavior in 0.75 saturation case, and the lower and higher demand scenarios reflect how sensitive a pre-given control logic will be in real world.
- At the low-level demand, the LQF had a diametrical performance with a high-level demand. With a sparse traffic, a LQF controller is more similar to vehicle actuated control that provide green if vehicle exist. Furthermore, the τ-seconds fixed interval is more likely sufficient to disperse all the queuing vehicle, which can reduce average halts significantly. In contrast, with a near-saturation or over-saturation traffic, the LQF always attempts to maintain balance among phases and movements. It can effectively avoid queuing spillover. In this case, although the number of halts is considerably high, the vehicles in queuing still continue to move frequently and can maintain higher travel speeds. The performance LQF reflect the motivation of almost previous RL signal controller study to select phase selection strategy. However, with higher demand in arterial networks, the changeable phase duration setting is critical to resist limitation of LQF-like phase selection strategy.
- RHS could be considered a dynamic expansion of Webster, which greedily relies on the current queue number detected. At the low and middle level demand, the method is effective in that it always serves the entire fleet. However, as the demand increases, the method will suffer from a severe performance drop owing to the complex interaction between upstream and downstream traffic. The phase switching strategy in several RL traffic signal controls might suffer from the same limitation. Owing to the combination of phase selection and phase duration setting, our WaveLearner can maintain a low halt and high throughput with better travel time and travel speed performance than RHS.

D. Sensitivity analysis of Green-wave demand

Owing to lack of knowledge to tune the hyper-parameter β_{GW}, 0.5 was used naively in experiments above. In this section, we further conduct a sensitivity analysis of β_{GW} using SHORT network under different saturation rate as shown in Fig. 4. The $\beta_{GW} = \{0.2, 0.4, 0.6, 0.8, 1.0\}$ were used to run the simulation, the travel time of major street sections, minor street sections and all sections (Fig. 5).

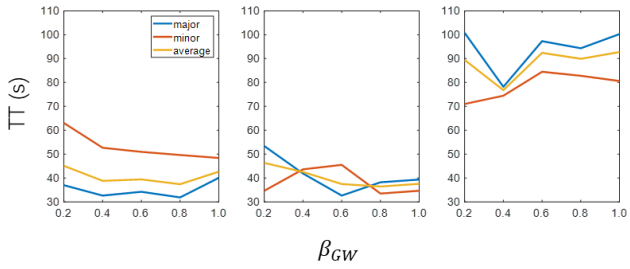

Fig. 5. Sensitivity analysis of demand of different demand levels: demand increasing from left to right

At the low demand level, the major street always had lower travel time owing to more lanes to separate the traffic. An increase in β_{GW} was beneficial for minor street traffic and the best overall performance was observed at $\beta_{GW} = 0.8$. The middle demand level indicated a significant function of β_{GW} compared with major and minor streets. To maintain a good performance of major street sections, β_{GW} should not be extremely large or small. The overall best performance was achieved at $\beta_{GW} = 0.6$. The high demand level indicated that a smaller $\beta_{GW} = 0.4$ is more suitable for both major streets and overall performance. Accordingly, we suggest the principle to tune the β_{GW}: 1) a lower value for a higher demand, 2) a value near to 0 or 1 is not good option for all demand levels.

IV. CONCLUSIONS

This paper has presented a coordinated RL controller for the arterial traffic signal control problem. Through the combination of traffic engineering knowledge, a green-wave

rewarding mechanism is proposed and observed to be effective for variant arterial traffic. Our research has a threefold contribution to both related researches and industries:

- A green-wave rewarding mechanism is proposed to combine green-wave control in traffic engineering and episodic training in RL. It is considerable as an innovative method of coordination of RL traffic signal control.
- The control performance of the proposed method is systematically examined using synthetic networks and traffic demand data. Through the experiments, we demonstrated the robustness of the proposed method and proved some conclusions in related studies.
- For potential practical applications, the tuning sensitivity of green-wave rewarding is discussed. The preliminary principles of tuning is provided.

The limitation of this study is that we considered the composition of an arterial networks as a priori. In a real-world congested road network, the relationship of major and minor streets is not always clear. In many cases, enhancing the major street efficiency does not imply an improvement in overall efficiency. In other words, for a general urban road network, the propose method can be applied only after clearly defining where and when to apply arterial control. Thus, the next step of this study should be a further generalization in defining the arterial or critical path from general networks. Moreover, we expected more empirical data included in our future works for further exploration.

REFERENCES

[1] F. V. Webster, "Traffic signal settings," Tech. Rep., 1958.

[2] H. Wei, N. Xu, H. Zhang, G. Zheng, X. Zang, C. Chen, W. Zhang, Y. Zhu, K. Xu, and Z. Li, "Colight: Learning network-level cooperation for traffic signal control," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 1913–1922.

[3] J. D. Little, M. D. Kelson, and N. H. Gartner, "Maxband: A versatile program for setting signals on arteries and triangular networks," 1981.

[4] N. H. Gartner, S. F. Assman, F. Lasaga, and D. L. Hou, "A multi-band approach to arterial traffic signal optimization," *Transportation Research Part B: Methodological*, vol. 25, no. 1, pp. 55–74, 1991.

[5] T. Arsava, Y. Xie, N. H. Gartner, and J. Mwakalonge, "Arterial traffic signal coordination utilizing vehicular traffic origin-destination information," in *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, 2014, pp. 2132–2137.

[6] C. Zhang, Y. Xie, N. H. Gartner, C. Stamatiadis, and T. Arsava, "Am-band: an asymmetrical multi-band model for arterial traffic signal coordination," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 515–531, 2015.

[7] W. Ma, L. Zou, K. An, N. H. Gartner, and M. Wang, "A partition-enabled multi-mode band approach to arterial traffic signal optimization," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 1, pp. 313–322, 2019.

[8] P. Mannion, J. Duggan, and E. Howley, "An experimental review of reinforcement learning algorithms for adaptive traffic signal control," *Autonomic road transport support systems*, pp. 47–66, 2016.

[9] K.-L. A. Yau, J. Qadir, H. L. Khoo, M. H. Ling, and P. Komisarczuk, "A survey on reinforcement learning models and algorithms for traffic signal control," *ACM Computing Surveys (CSUR)*, vol. 50, no. 3, pp. 1–38, 2017.

[10] G. Zheng, X. Zang, N. Xu, H. Wei, Z. Yu, V. Gayah, K. Xu, and Z. Li, "Diagnosing reinforcement learning for traffic signal control," *arXiv preprint arXiv:1905.04716*, 2019.

[11] C. Chen, H. Wei, N. Xu, G. Zheng, M. Yang, Y. Xiong, K. Xu, and Z. Li, "Toward a thousand lights: Decentralized deep reinforcement learning for large-scale traffic signal control," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 3414–3421.

[12] Q. Wu, L. Zhang, J. Shen, L. Lü, B. Du, and J. Wu, "Efficient pressure: Improving efficiency for signalized intersections," *arXiv preprint arXiv:2112.02336*, 2021.

[13] Z. Wang, H. Zhu, M. He, Y. Zhou, X. Luo, and N. Zhang, "Gan and multi-agent drl based decentralized traffic light signal control," *IEEE Transactions on Vehicular Technology*, pp. 1–1, 2021.

[14] L. Zhang, Q. Wu, and J. Deng, "Knowledge intensive state design for traffic signal control," *arXiv preprint arXiv:2201.00006*, 2021.

[15] T. Han, M. Ito, K. Shirahata, and T. Oguchi, "A study on possibility of predictive deep reinforcement learners for isolated intersection signal control," *SEISAN KENKYU*, vol. 73, no. 2, pp. 107–112, 2021.

[16] H. Wei, G. Zheng, V. Gayah, and Z. Li, "A survey on traffic signal control methods," *arXiv preprint arXiv:1904.08117*, 2019.

[17] L. Prashanth and S. Bhatnagar, "Reinforcement learning with average cost for adaptive control of traffic lights at intersections," in *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2011, pp. 1640–1645.

[18] L. Kuyer, S. Whiteson, B. Bakker, and N. Vlassis, "Multiagent reinforcement learning for urban traffic control using coordination graphs," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2008, pp. 656–671.

[19] E. Van der Pol and F. A. Oliehoek, "Coordinated deep reinforcement learners for traffic light control," *Proceedings of Learning, Inference and Control of Multi-Agent Systems (at NIPS 2016)*, 2016.

[20] T. Nishi, K. Otaki, K. Hayakawa, and T. Yoshimura, "Traffic signal control based on reinforcement learning with graph convolutional neural nets," in *2018 21st International conference on intelligent transportation systems (ITSC)*. IEEE, 2018, pp. 877–883.

[21] P. Varaiya, "Max pressure control of a network of signalized intersections," *Transportation Research Part C: Emerging Technologies*, vol. 36, pp. 177–195, 2013.

[22] H. Wei, C. Chen, G. Zheng, K. Wu, V. Gayah, K. Xu, and Z. Li, "Presslight: Learning max pressure control to coordinate traffic signals in arterial network," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 1290–1298.

[23] H. Wei, C. Chen, K. Wu, G. Zheng, Z. Yu, V. Gayah, and Z. Li, "Deep reinforcement learning for traffic signal control along arterials," 2019.

[24] A. Sharma, D. M. Bullock, and J. A. Bonneson, "Input–output and hybrid techniques for real-time prediction of delay and maximum queue length at signalized intersections," *Transportation Research Record*, vol. 2035, no. 1, pp. 69–80, 2007.

[25] G. F. Newell, "The rolling horizon scheme of traffic signal control," *Transportation Research Part A: Policy and Practice*, vol. 32, no. 1, pp. 39–44, 1998.

[26] R. Wunderlich, C. Liu, I. Elhanany, and T. Urbanik, "A novel signal-scheduling algorithm with quality-of-service provisioning for an isolated intersection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 9, no. 3, pp. 536–547, 2008.

[27] W. Zhao, Y. Ye, J. Ding, T. Wang, T. Wei, and M. Chen, "Ipdalight: Intensity-and phase duration-aware traffic signal control based on reinforcement learning," *Journal of Systems Architecture*, vol. 123, p. 102374, 2022.

[28] H. Zhu, T. Han, W. K. Alhajyaseen, M. Iryo-Asano, and H. Nakamura, "Can automated driving prevent crashes with distracted pedestrians? an exploration of motion planning at unsignalized mid-block crosswalks," *Accident Analysis & Prevention*, vol. 173, p. 106711, 2022.

[29] H. X. Liu and W. Ma, "A virtual vehicle probe model for time-dependent travel time estimation on signalized arterials," *Transportation Research Part C: Emerging Technologies*, vol. 17, no. 1, pp. 11–26, 2009.

[30] S. Krauß, P. Wagner, and C. Gawron, "Metastable states in a microscopic model of traffic flow," *Physical Review E*, vol. 55, no. 5, p. 5597, 1997.

[31] G. Zhang, Y. Wang, H. Wei, and Y. Chen, "Examining headway distribution models with urban freeway loop event data," *Transportation Research Record*, vol. 1999, no. 1, pp. 141–149, 2007.