IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 6, 2020, accepted January 19, 2020, date of publication January 23, 2020, date of current version January 31, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2968937

A Distributed Control Method for Urban Networks Using Multi-Agent Reinforcement Learning Based on Regional Mixed Strategy Nash-Equilibrium

ZHAOWEI QU, ZHAOTIAN PAN, YONGHENG CHEN, XIN WANG, AND HAITAO LI
School of Transportation, Jilin University, Changchun 130022, China

Corresponding author: Yongheng Chen (cyh@jlu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 51705196.

ABSTRACT Urban network traffic congestion can be caused by disturbances, such as fluctuation and disequilibrium of traffic demand. This paper designs a distributed control method for preventing disturbance-based urban network traffic congestion by integrating Multi-Agent Reinforcement Learning (MARL) and regional Mixed Strategy Nash-Equilibrium (MSNE). To enhance the disturbance-rejection performance of Urban Network Traffic Control (UNTC), a regional MSNE concept is integrated, which models the competitive relationship between each agent and its neighboring agents in order to improve the decision-making process of MARL. The learning rate is enhanced with a self-adaptive ability to avoid a local optimal dilemma; Jensen-Shannon (JS) divergence is utilized to define the learning rate of the modified MARL. A two-way rectangular grid network with nine intersections is modeled via a Cell Transmission Model (CTM). A probability distribution mechanism, which can update the turn ratio of each approach dynamically and discretely, is established to represent the segmented route-decision process of the vehicles. The effectiveness of the proposed control method is evaluated through simulations in the grid network. The results show the influence of major disturbances, such as fluctuation of vehicle arrival rate, fluctuation of traffic demand (e.g. a rapidly rising flow and extreme changes in origin-destination distribution), and disequilibrium of traffic demand (e.g. different arrival flows at each boundary of the urban network), on the performance of the suggested control method. The results can be used to improve the state of the art in order to reduce urban network traffic congestion due to these disturbances.

INDEX TERMS Urban network traffic control, distributed traffic signal control system, multi-agent reinforcement learning, mixed strategy Nash-equilibrium, numerical simulation.

I. INTRODUCTION

In urban networks, traffic congestion can occur for various reasons [1], including traffic incidents, constraints on network capacity or stochastic fluctuations in demand. Traffic signal control is an effective method that has been extensively studied to alleviate this congestion.

Existing signal control methods employed in Urban Network Traffic Control (UNTC) systems can be classified in several ways. The control type method is mainly categorized as fixed-timed, traffic responsive [2], [3], or predictive control strategies. These methods can be further categorized in terms of hierarchical structure to one of the following three approaches: centralized approaches, sectional centralized approaches or distributed control approaches. In terms of input mode, they can be divided into online and offline methods. From the viewpoint of artificial intelligence (AI), some control methods can be categorized as heuristic algorithms, expert systems, etc.

In recent years, various optimization algorithms have been employed to exploit the potential of traffic signal control. The concept of optimization algorithms is to consider the optimization of UNTC as a combinatorial problem in nature. Various meta-heuristic and intelligent algorithms, including

The associate editor coordinating the review of this manuscript and approving it for publication was Qilian Liang.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

Genetic Algorithms (GA) [4], Particle Swarm Optimization (PSO) [5], Ant Colony Optimization (ACO) [6] and Bee Colony Optimization (BCO) [7] have been applied to UNTC to solve these problems. However, the control schemes obtained by these heuristic algorithms can only be employed to a stable traffic state.

With the development of Machine Learning (ML) theory and Artificial Intelligence (AI), Reinforcement Learning (RL) has been applied to the field of traffic control, which has the powerful advantage of experiential learning. Multi-Agent Systems (MAS) effective solutions in terms of negotiation protocol [8], strategy diffusion [9] and network decision-making [10]. In an urban network, the network formed by each intersection node is similar to a Social Network (SN) which has been previously studied in the literature [8], [9], with a correlation between intersections in urban networks and similar agent relationships in the literature [9], [10]. Since the structure of interactions between control behaviors at the intersection is dependent on a hierarchical UNTC, it is appropriate to analyze the UNTC problem from a multi-agent perspective. This has promoted the generation of a Multi-Agent Reinforcement Learning (MARL) method that combines with the concepts of Game Theory (GT) and RL based on MAS, in order to accumulate historical experiences. MARL has been shown to be extraordinarily promising and has been recognized as a powerful tool to reduce the extent of network traffic congestion [11], [12]. The main advantages of MARL are that it can learn from its own experience and adapt to its environment [13]. Depending on their reward strategy attributes, existing traffic control systems based on agent-based RL can be divided into three categories: individual MARL, group MARL and global MARL. In urban networks, the global MARL method requires expensive communication resources and a high CPU performance to achieve the desired control effect, and it can be difficult to distinguish states due to its large dimensionality. This can be avoided with the individual MARL method which uses a distributed iterative process with individual goals, although it may take a long time for the results to converge to the same level as the global MARL. We have observed that there has been little research into the process of decision-making in MARL. At the kernel of MARL is the accumulation of historical experiences that are used for decision-making. This accumulation is designed to obtain the ability to make valid decisions through trial and error. Any method that can improve this decision-making by trial and error can enhance the learning efficiency of the individual MARL method. For the improved individual MARL method, any invalid historical experience accumulated by the unimproved individual MARL is ignored at the same iteration level. Although MARL incorporates GT concepts, most researchers have applied GT to the process of updating experiences or obtaining equilibrium between multi-agent system goals. Nash-equilibrium is a method that is commonly used to measure changes in market structures [14] and Mixed Strategy Nash-Equilibrium (MSNE) can be used to solve the uncertainty competition [15]. Since the structure of MAS is similar to the structure of the market to some extent, the concept of MSNE in GT can be considered to improve the decision-making process of MARL. Additionally, there has been little research on the learning rate of MARL. In previous studies, the learning rate of MARL has been usually defined as a constant or a time-varying attenuation resulting in applications using MARL requiring feedback-based learning and acclimatization with an offline pattern. However, using an attenuation to represent the learning rate leads to a high learning time cost when faced with new situations and the method is insensitive to fluctuations in state after convergence. Therefore, there are two significant components of MARL requiring further study: (a) the decision-making process, which is fundamental in increasing the rate of convergence in MARL; (b) the learning rate, which plays a critical role in the sensitivity of MARL.

In this article, we suggest a distributed urban network traffic signal control method based on concepts and mechanisms inspired by GT and MAS. We design a decentralized multi-agent architecture without hierarchy, where each agent represents a traffic signal controller assigned to each signalized intersection in the urban network. In this architecture, each agent communicates and competes with adjacent agents. Additionally, each agent accumulates experience and increases its adaptive capacity for disturbances in the urban network by using an improved MARL algorithm. We particularly focus on considering a global perspective to the decision-making process of MARL, and on enhancing the method's ability to quickly adapt to local disturbances within urban network area.

The main contributions of this study to the advancement of the state of the art are summarized as follows:

(1) The framework of MARL is modified to improve its management of disturbance-based traffic congestion in urban networks. Spatial and temporal stochastic disturbances are considered and the improved MARL framework is capable of disturbance detection within the local area and self-learning at high speed for disturbance mitigation.

(2) The notion of MSNE in GT is integrated into the decision-making process of MARL to enhance its ability to resist disturbances and prevent disturbance-based urban network traffic congestion. The modified decision-making process accelerates the convergence process of MARL and can reduce the MARL online learning time.

(3) The Jensen-Shannon (JS) divergence is introduced to define the learning rate of MARL to provide self-adaptive ability to manage new scenarios generated in the urban network. This method enhances the sensitivity of MARL after convergence and improves the accumulation of new experiences when faced with state transitions.

The remainder of this article is organized as follows. In Section II, we present a literature review with focus on various control methods, multi-agent systems and multi-agent reinforcement learning applied to traffic signal control. In Section III, the main concepts and mechanisms of basic GT and MARL are firstly introduced and a proposed distributed control method for urban networks and pseudo code of corresponding algorithms are then described in detail. In Section IV, a detailed description of the numerical simulation framework for performance assessment is provided. In Section V, a numerical simulation in terms of related parameters, application scenarios and contrast methods is given and the results are analysed. Finally, conclusions of this work are given in Section VI.

II. LITERATURE REVIEW

UNTC is one of the most difficult problems in the field of traffic control and has been widely explored over the last few decades. In earlier studies, the most representative achievement was TRANSYT, which was a signal control strategy proposed by Robertson [16] which has been applied with various derivative versions [17]–[19]. Since then, several classic control strategies have emerged, such as SCOOT [3], OPAC [20] and RHODES [21]. However, all of these systems employed the control mechanisms along arterial routes where there is major demand, but did not consider the important network-wide control effect. There are various modeling methods that have been employed in the UNTC field, such as Model-Predictive Control (MPC) [22]–[24] and the Max Pressure (MP) control [25]–[27]. As these techniques have developed further swarm intelligence techniques have appeared in UNTC during the last decade and some representative methods have already been mentioned in Section I. These control methods or systems are all optimization solutions in nature. Some of these methods can be regarded as expert systems, which are adaptations of historical offline optimal solutions, and the other methods can be classified as heuristic algorithms, which are optimal solutions that are searched online. In order to achieve a global optimal solution, most of these methods have a centralized or hierarchical architecture. The architecture has a centralized solution procedure but the complexity rises exponentially as the network range increases. This can be addressed using a flexible cluster balancing centralized method and decomposed solution procedure, as suggested in [28]. However, since all of the above approaches are open-loop control methods, the performance of these methods approaches a bottleneck as the complexity is increased. Therefore, these systems are suitable for stable traffic conditions but have weak resistance to disturbances.

As an alternative, reinforcement learning (RL) has shown strong potential for self-learning closed-loop optimal traffic signal control in a stochastic traffic environment [13], [29]. MAS are a sub-field of Artificial Intelligence (AI) that are widely employed in many fields [30]–[33], and provide principles for constructing complex multi-agent systems and mechanisms for coordinating the behavior of independent agents. Game Theory (GT) provides tools to model MAS as a multiplayer game and provides a rational strategy for each player in a game [34]. Various studies have used MAS architecture, including in combination with the fuzzy logic theory [35], embedded in the model predictive control [36] and integrated with a biological immune system [37]. The most recent research effort focuses on developing distributed approaches using multi-agent technology [38]–[39]. These provide an appropriate approach for the application of MAS in UNTC.

TABLE 1. The representative MARL researches in the field of UNTC.

researcher	algorithm	action selection	learning rate	relations of agents
Thorpe [40]	SARSA	ε	constant	-
Abdulhai [41]	Q	Softmax	-	-
Jin [42]	Q/SARSA	ε /Softmax	-	-
Arel [43]	Q	ε	Boltzmann	cooperative (adjacent state)
Zhu [44]	R	ε -JTA	-	cooperative (adjacent action)
EI-Tantawy [34]	Q	ε -MARLIN	-	cooperative (adjacent action)
Bazzan [45]	Q	Softmax	constant	cooperative (global average)

The MARL method combines RL theory [46] with MAS and promotes the development of UNTC theory. Some representative research results from employing MARL in the field of UNTC are listed in TABLE 1. This table also compares the algorithm selection, including the process of action selection and the learning rate employed by researchers. The decentralized traffic control problem is an excellent application scenario for MARL due to the dynamicity and randomness of the traffic system [11], [47]. The advantages and disadvantages of MARL have been discussed in Section I, and can also be reviewed in [34]. There are various improved algorithms of MARL which accelerate the convergence process of multi-agent system to Nash-equilibrium, such as Minimax Q-Learning (Minimax-Q) [48], Nash Q-learning (Nash-Q) [49] and Correlated Q-Learning (CQ) [50]. Q-Learning (QL) algorithms consider the influence of the expected Q value (i.e. the final Q-function of equation (2) in Section III.B) on the convergence process of MARL. In addition, there are also some methods to accelerate MARL convergence, such as Motivated Reinforcement Learning (MRL) [51] and Equilibrium Transfer (ET) [52]. MRL can accelerate MARL convergence by employing a motivated mechanism to supplement the reward function. In UNTC, a similar effect can be achieved using a reasonable definition for the reward function. However, ET may be difficult to integrate with traffic control systems due to different historical situations at different intersections.

Framework analysis of MARL shows that this is a feasible approach to accelerate MARL convergence, as this approach shows a preference for the potential optimal action during the action selection process. The rationality of this approach has

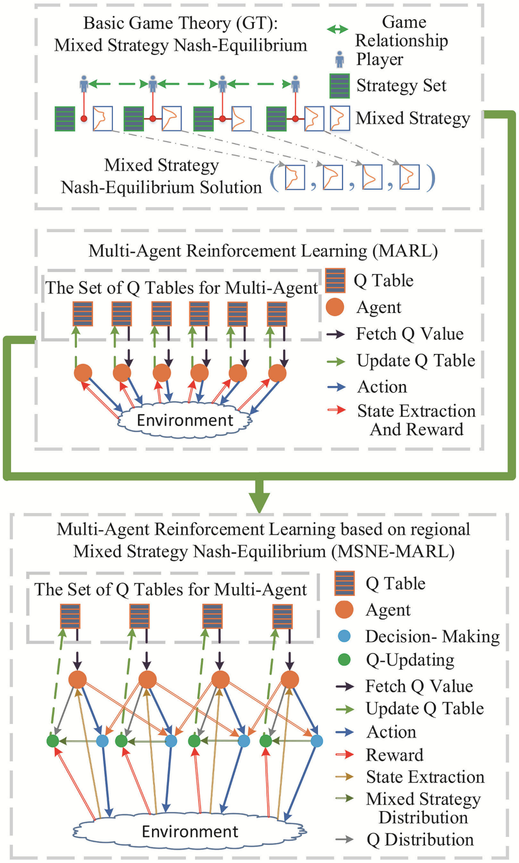

FIGURE 1. Overview of methods in Section III.

been discussed in Section I. However, the existing literature has a lack of research and ε greedy and softmax are typically used as the action selection strategies [46].

An effective algorithm should have the ability to jump out of the local search area and the learning rate has a strong influence on this aspect. However, there is also a lack of research into the MARL learning rate in the existing literature. In early studies, the MARL learning rate was usually set as a constant. In later studies, the learning rate was given a gradually decreasing form which considers both the rapid replacement of incorrect experiences in the early stage of learning and stability due to convergence in the later stage of learning with the iterative process revised by an embedding algorithm, such as Simulated Annealing (SA) [53]. The disadvantages of both of the above patterns and the necessity of improving the learning rate of MARL were fully discussed in Section I.

III. FRAMEWORK OF MULTI-AGENT REINFORCEMENT LEARNING (MARL) BASED ON REGIONAL MIXED STRATEGY NASH-EQUILIBRIUM (MSNE)

The framework described in this section is briefly illustrated in FIGURE 1 as three parts from top to bottom. The upper part of the figure is the Mixed Strategy Nash-Equilibrium (MSNE) schematic, which will be described in detail in Section III.A. The middle part shows Multi-Agent Reinforcement Learning (MARL), whose related concepts will be defined in Section III.B. After coupling MARL with MSNE, the lower part of the figure gives a brief schematic diagram of the framework proposed in this paper: Multi-Agent Reinforcement Learning based on regional Mixed Strategy Nash-Equilibrium (MSNE-MARL). The framework of MSNE-MARL will be described in detail in Section III.C. Two key improvements to the MARL framework will be described in Section III.C.4 and III.C.5. The corresponding algorithm is subsequently summarized in Section III.D. All parameters relating to GT, MARL and MSNE-MARL are listed in TABLE 2.

A. BASIC GAME THEORY (GT)

The concept of Game Theory (GT) was first proposed by Neumann and can be generally divided into two types: zero-sum games and general-sum games. In zero-sum games, the sum of benefits of each player is absolute. Zero-sum indicates a complete conflict of interest in which a player gains benefits at the expense of the others' benefits. In contrast, players have both compatible and conflicting interests in general-sum games. This situation provides a feasible precondition for the pursuit of higher overall rewards and has been widely studied by scholars.

In this section, we will first introduce some basic concepts of a game in strategic (or normal) form and briefly introduce the concept of Nash equilibrium [54], which is the baseline solution concept for general-sum games [49].

Definition 1: A game in strategic (or normal) form G is a tuple $\langle J, S, U\rangle$ [55]. *The elements of the tuple are described in TABLE 2.*

In a game with complete information, a pure-strategy indicates that only one specific strategy can be adopted for each given set of information. We can regard a mixed strategy as a generalization of the strategy selection.

Definition 2: A mixed strategy σ_i *is a probability distribution over the pure-strategy space* S_i [55]. *The space of player i's mixed strategies and the concept of relevant parameters are introduced in TABLE 2.*

According to **Definition 2**, the payoff of player i to profile σ is as follows:

$$\sum_{s\in S}\left(\prod_{i=1}^{I}\sigma_j\left(s_j\right)\right)u_i\left(s\right) \tag{1}$$

Note that player i's payoff to a mixed-strategy profile is a linear function of player i's mixing probability σ_i.

The concept of a Nash-equilibrium solution has attracted much research interest as it exists for a broad class of games.

Definition 3: A Nash-equilibrium is a profile of strategies such that each players' strategy is an optimal response to the other players' strategies. A mixed-strategy profile σ^* *is a Nash equilibrium for all players* $i \in J$, $u_i\left(\sigma_i^*, \sigma_{-i}^*\right) \geq u_i\left(s_i, \sigma_{-i}^*\right)$ *for all* $s_i \in S_i$ [55].

TABLE 2. Parameters relating to GT, MARL and MSNE-MARL.

Notations	Signification
i	a player of game /an agent of MAS /an intersection of urban network
I_i	The set of agents
I_{-i}	The competitive agents set of agent i
G	A game in strategic (or normal) form
J	The set of players, $\{1,\ldots,I\}$
S	The set of the pure-strategy space S_i
S_i	The pure-strategy space of player i
U	The set of the payoff functions u_i
$u_i(s)$	The von Neumann-Morgenstern utility of player i for each profile s of strategies
s	A strategy profile of all players, $\{s_1,\ldots,s_I\}$
s_i	The pure strategy adopted by player i, $s_i \in S_i$
σ_i	The mixed strategy adopted by player i, $\sigma_i \in \sigma$ and $\sigma_i \in \Sigma_i$
$\sigma_i(s_i)$	The probability that σ_i assigned to s_i, $\sigma_i(s_i) \in \sigma_i$
σ_i^*	The Nash-equilibrium mixed strategy of player i
σ^*	The mixed strategy Nash-equilibrium
Σ_i	The mixed strategy space of player i
Σ	The set of mixed strategy space Σ_i, $\sigma \in \Sigma$
X	The state space set, $X = \times_i X_i$
X_i	The state space of agent i
A	The action space set, $A = \times_i A_i$
A_i	The action space of agent i
R	The reward function of agents, $R: X \times A \to R$
Q_i	The Q function of agent i
γ	The discount factor, $\gamma \in [0,1)$
α	The learning rate, $\alpha \in [0,1)$
ε	The coefficient of greed
x_i	The state confronted by agent i, $x_i \in X_i$
x_i'	The future state confronted by agent i
a_i	The action executed by agent i, $a_i \in A_i$
a_i'	The future action executed by agent i
$Q_i^k(x_i,a_i)$	The Q value of agent i executing a_i under x_i after iterating k times
$Q_i^*(x_i,a_i)$	The convergent Q value of agent i executing a_i under x_i
$\lvert X_i \rvert$	The size of X_i
$link_{i,j}$	The directed link from intersection i to intersection j, $j \in I_{-i}$
y_{ij}	The number of vehicles in $link_{i,j}$ (unit: pcu)
$y_{ij}^{\max}$	The maximum number of vehicles in $link_{i,j}$ (unit: pcu)
$y_{ji}^{pre}(a_i,a_{-i},x_i)$	The expected number of vehicles in $link_{j,i}$ after the scenario that agent i executes action a_i under state x_i and its competitive agents execute actions a_{-i}
$P_i(a_i \mid x_i)$	The probability of agent i execute action a_i facing with x_i
P_i	The distribution of $Q_i(x_i,a_i)$ in action space A_i
$\hat{P}_i$	The optimum decision-making distribution of agent i

The existence of a Mixed Strategy Nash Equilibrium (MSNE) has been proven for some cases.

Theorem 4: Every finite strategic-form game has a mixed-strategy equilibrium[56].

It is worth noting that **Theorem 4** does not assert the existence of equilibrium with nondegenerate mixing.

B. MULTI-AGENT REINFORCEMENT LEARNING (MARL)

Multi-Agent Reinforcement Learning (MARL) can be regarded as an extension of Reinforcement Learning (RL) in Multi-Agent System (MAS) [57]. RL combines the two fields of supervised learning and Dynamic Programming (DP) [46], yielding a powerful Machine Learning (ML) system [11]. MARL can mainly be described by introducing the concept of Q-Learning (QL). QL, which stores value functions in the form of a Q factor, is an RL method applied widely.

Definition 5: The Q-learning basic elements of a multi-agent reinforcement learning is a tuple$\langle J, X, A, R\rangle$. The elements of the tuple are described in TABLE 2.

The updating procedure of QL can be simply defined. In this procedure, agent i starts with arbitrary initial values $Q_i^1(x_i, a_i)$ for all $x_i \in X_i$, $a_i \in A_i$ and updates its Q-values according to equation (2).

$$Q_i^{k+1}(x_i, a_i) = (1-\alpha)\, Q_i^k(x_i, a_i) + \alpha \left[r_i^k(x_i, a_i) + \gamma \max_{a_i' \in A_i(x_i')} Q_i^k(x_i', a_i') \right] \tag{2}$$

Lemma 6: Assuming that all states $x_i \in X_i$ and actions $a_i \in A_i$ have been visited infinite times and that the learning rate α satisfies certain constraints, equation (2) converges to $Q_i^*(x_i, a_i)$.

Lemma 6 has been proven by Watkins and Dayan in 1992. For specific details of the proof see [58].

Combining the above definitions, the Q-learning algorithm performed by each agent can be summarized in ALGORITHM 1 which is $O(n \times T)$.

C. MSNE-MARL FRAMEWORK

In this section, MSNE theory (described in section III.A) and MARL theory (described in section III.B) are used to design a distributed traffic signal control method for an urban network. Since an urban network consists of numerous signalized intersections and links, we suggest a fully decentralized framework, where each agent is associated with a single signalized intersection. The architecture is horizontal (i.e. there is no hierarchy between agents, and all agents communicate and interact by sharing data on the traffic state of each of their relevant links). Regional games are achieved between neighboring agents based on the traffic state of segments connecting the intersections in the urban network. In this framework, each agent makes its own decisions autonomously based on the regional MSNE and by adopting data from neighboring

Algorithm 1 Pseudo-Code of Q-Learning Algorithm

```
Input: the learning rate $\alpha$; the discount rate $\gamma$;
       the coefficient of greed $\varepsilon$; the set of state $X$;
       the space of action $A$; the number of agents $n$
Initialize: the simulation time $t \leftarrow 0$; the step of simulation $k \leftarrow 0$
    for $i = 1, \ldots, n$ do
      if $x_i \in X_i, a_i \in A_i, X_i \in X, A_i \in A$ do
        $Q_i^k(x_i, a_i) \leftarrow random$
      end if
    end for
While $t \leq T$ do
  for $i = 1, \ldots, n$ do
    random seed: $seed \leftarrow random$
    If $seed \leq \varepsilon$ do
      select $a_i \in A_i(x_i)$ randomly
    else
      $a_i = \arg\max_{a_i \in A_i(x_i)} Q_i^k(x_i, a_i)$
    end if
    calculate: $r_i^k(x_i, a_i)$
    update the Q-value according the equation (2)
  end for
  $t = t + 1, k = k + 1$
end while
```

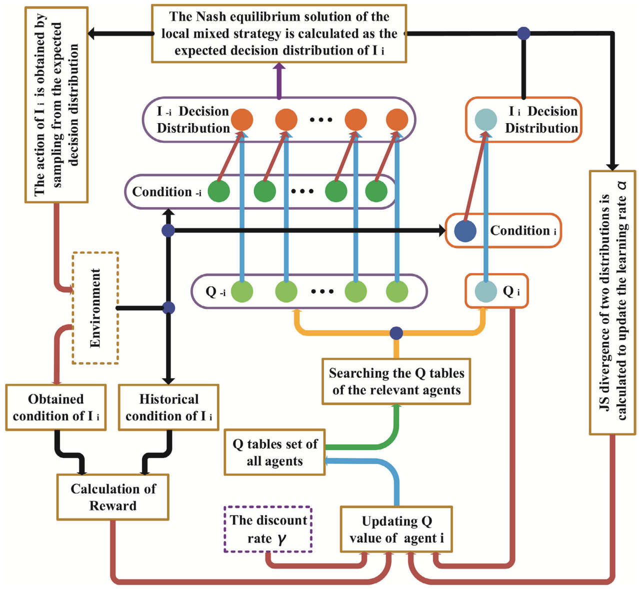
The action of I i is obtained by sampling from the expected decision distribution
The Nash equilibrium solution of the local mixed strategy is calculated as the expected decision distribution of I i
I -i Decision Distribution
I i Decision Distribution
Condition -i
Condition i
Q -i
Q i
Environment
JS divergence of two distributions is calculated to update the learning rate α
Searching the Q tables of the relevant agents
Obtained condition of I i
Historical condition of I i
Q tables set of all agents
Calculation of Reward
The discount rate γ
Updating Q value of agent i

FIGURE 2. Diagrammatic drawing of MSNE-MARL framework.

agents and collected from the intersection. FIGURE 2 illustrates the suggested Multi-Agent Reinforcement Learning based on regional Mixed Strategy Nash-Equilibrium (MSNE-MARL) distributed control framework.

In the following sub-sections, the various components of the MSNE-MARL framework shown in FIGURE 2 are defined. It should be mentioned that the iteration series k is omitted from the following formulae in order to make them easier to understand and simplify their expression.

1) STATE SPACE

A MAS must maintain n Q-functions, one for each agent in the system. In previous studies, the state space has been divided into two categories: the independent state space and the joint state space. Let $|X_i|$ be the size of agent i's state space X_i. Assuming that $|X_1| = \ldots = |X_n| = |X|$, the total number of entries in the independent state space and the joint state space are $n\,|X|$ and $|X|^n$ respectively. Therefore, in terms of state space complexity, the number of states in a MAS with an independent state space is linear. This is superior to a MAS with a joint state space, where the number of states grows exponentially with the number of agents. Therefore, it is appropriate to select an independent state space and a smaller value of $|X|$.

Definition 7: Assuming that any signal-controlled intersection has four approaches in an urban network, the state of agent i , can be expressed as: $x_i \leftarrow \left[x_i^e, x_i^w, x_i^s, x_i^n\right]$. Let dir represent one element of $\{e, w, s, n\}$. x_i^{dir}, the component of x_i, representing the state of dir approach at intersection i, which can be defined by integrating formula (3) and formula (4).

$$c_i^{dir} = \frac{y_{ji}}{y_{ji}^{\max}} \tag{3}$$

$$x_i^{dir} = \begin{cases} free, & c_i^{dir} < \varphi_{free} \\ resistance, & \varphi_{free} \leq c_i^{dir} < \varphi_{jam} \\ jam, & c_i^{dir} \geq \varphi_{jam} \end{cases} \tag{4}$$

In this paper, 0.5 and 0.8 are adopted as the values of φ_{free} and φ_{jam} respectively.

2) ACTION SPACE

The complexity of the action space is an important factor which should be considered before defining the action space. The complexity analysis process of the action space is similar to that described in section III.C.1. In addition, it is necessary to consider maximizing the coverage area of the action space in order for the same number of actions to be expressed with lower space complexity. Inspired by the concept of distributed expression in Deep Learning (DL) and the method described in [59], we construct the action of each agent using vectors. FIGURE 3 illustrates the dual-ring phase structure, which is the action structure employed by each agent to control the associated signalized intersection in the urban network.

In FIGURE 3, the action of agent i is $a_i \leftarrow \left[a_i^1, a_i^2, a_i^3\right]$. Element a_i^1 represents the direction of the road being controlled (i.e. Road A or B), and the tuple $\left\langle a_i^2, a_i^3\right\rangle$ represents the phase of the associated signalized intersection. The basic elements a_i^2 and a_i^3 represent two separate and non-conflicting streams of traffic flow in different driving directions (i.e. the fundamental phase in phase ring A and B).

3) REWARD FUNCTION

In QL, the Q value is the experience of QL and the updating process of the Q-function can be regarded as the process of accumulating experience. According to the definition in equation (2), full consideration should be given when selecting the range of the reward function. If the reward function

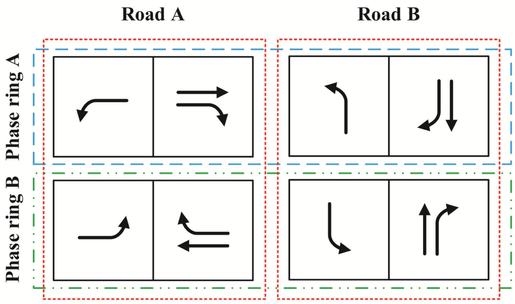

FIGURE 3. Dual-ring phase structure of the action.

has a range that is too wide, the difference in distribution of Q values between two iterations is too large and a long learning time is required to achieve convergence. If the reward function has a range that is two narrow, the difference in distribution of Q values between two iterations is too small and it will take a long time to accumulate experience as well as a long learning time to achieve convergence. It can be observed that the range selection of the reward function is related to the efficiency of QL. Additionally, to avoid excessive accumulation of the Q factor after a large number of iterations, both the reward and penalty forms should be considered when defining the reward function. Based on these considerations, the reward function will be formally defined.

Definition 8: The reward function of agent i, $r_i(x_i, a_i)$ *is a function of the difference in number of vehicles on all approaches of an associated intersection.*

$$r_i(x_i, a_i) = \frac{\sum_{j\in I_{-i}} \Delta y_{ji}}{\sum_{j\in I_{-i}} \left|\Delta y_{ji}\right|} \tag{5}$$

In formula (5), the summation of $\left|\Delta y_{ji}\right|$ is designed to ensure that the range of the reward function $r_i(x_i, a_i)$ converges within $[-1, 1]$.

4) ACTION SELECTION

The classical QL adopts ε greedy or softmax form as the action selection strategy. Action selection is a trial-and-error process to pick an appropriate strategy and accumulate experience. Therefore, we assume that relative correct actions can be selected during action selection and therefore a trial-and-error process of shorter duration that is more efficient than classical QL can be achieved. Based on the above assumption, we design a method for action selection using the MSNE concepts.

Considering its future application in urban networks, we anticipate that MAS can obtain a globally optimal state once it reaches final stability. Therefore, the payoff function of the agents needs to be properly defined to construct general-sum games.

Definition 9: In a general-sum game, agent i executes action a_i *in face of state* x_i*, when its competitive agents execute actions* a_{-i}*; the payoff function of agent i is as follows:*

$$u_i(a_i, a_{-i}, x_i) = \sum_{j\in I_{-i}} \varpi_{ji}\left(y_{ji}^{pre}(a_i, a_{-i}, x_i) - y_{ji}\right) \tag{6}$$

In formula (6), the weight ϖ_{ji} is designed to indicate the pressure to immediately improve the state in $link_{ji}$ and can be derived from formula (7):

$$\varpi_{ji} = \frac{y_{ji}}{\sum_{j\in I_{-i}} y_{ji}} \tag{7}$$

To obtain the MSNE, it is necessary to anticipate the strategy selection of other agents. Considering the process of the game to achieve the MSNE, we adopt the experience gained from QL to anticipate the equilibrium strategies of competitive agents.

Definition 10: Assuming that the state set which agent j is confronted with is x_j*, when agent i is affected by state* x_i*, the distribution of* $Q_j(x_j, a_j)$ *can be employed to represent components of the mixed strategy.*

$$P_j(a_j|x_j) = \frac{\exp^{Q_j(x_j,a_j)}}{\sum_{a_j\in A_j(x_j)} \exp^{Q_j(x_j,a_j)}} \tag{8}$$

Therefore, agent j's mixed strategy $\sigma_j \leftarrow P_j$ can be obtained. Furthermore, it is feasible to solve the MSNE of agent i based on the known anticipated mixed strategy adopted by competitive agents.

$$\sigma_i^* = \arg\max_{\sigma_i} u_i(\sigma_i, \sigma_{-i}, x_i) \tag{9}$$

Through random sampling within the probability distribution $\hat{P}_i \leftarrow \sigma_i^*$, agent i will execute the acquired action a_i.

The above decision mechanism can accelerate the convergence of MARL and improve the effect of UNTC. In addition, the introduction of a regional MSNE achieves disturbance detection within the local scope.

5) OTHER PARAMETERS

In QL, the value of the discount factor γ reflects each agent's preference for a long-term reward. The discount factor values may be different for different agent groups. To reduce the complexity of the suggested method and coordinate agents' attention to both long-term and short-term rewards, we adopt 0.5 as the value of the discount factor for each agent in this paper with considering the application of MSNE-MARL in UNTC.

The learning rate is related to the learning speed of the agents. A high learning rate leads to amnesic damage of the accumulated learning experience but a low learning rate leads to learning efficiency loss of the agents.

The learning rate value employed by previous studies can be generally divided into two forms: (a) a constant value; (b) an attenuated value. Algorithms that use form (a) as the learning rate prevailingly have a long learning time and low efficiency due to the accumulation of incorrect experience during the early stage of learning. Algorithms with form (b) as the learning rate can not only overcome this problem, but also improve the stability of the convergence in the later stage of learning. However, regardless of the form of the learning rate, the algorithm learning process is performed under known situations which have fewer stochastic fluctuations. Since the learning is offline in nature, it is difficult to quickly respond to sudden and obvious disturbances when the algorithm is running online.

Based on the above analysis, a flexible form of the learning rate should be established.

Definition 11: The distribution of $Q_i(x_i, a_i)$ in action space A_i is P_i. The optimum decision-making distribution of agent i is $\hat{P}_i$. Agent i's learning rate α_i can be defined by the Jensen-Shannon (JS) divergence of $\hat{P}_i$ and P_i as follows:

$$\alpha_i = JS\left(\hat{P}_i||P_i\right) \tag{10}$$

The JS divergence in formula (10) can be expressed by formula (11):

$$JS\left(\hat{P}_i||P_i\right)=\frac{1}{2}KL\left(\hat{P}_i||\frac{\hat{P}_i+P_i}{2}\right)+\frac{1}{2}KL\left(P_i||\frac{\hat{P}_i+P_i}{2}\right) \tag{11}$$

The first term on the right side of formula (11) is the Kullback–Leibler (KL) divergence, which can be expressed by formula (12):

$$KL\left(\hat{P}_i||\frac{\hat{P}_i+P_i}{2}\right) = \sum_{a_i\in A_i}\hat{P}_i\log\left(\frac{2\hat{P}_i}{\hat{P}_i+P_i}\right) \tag{12}$$

The second term on the right side of formula (11) can also be expressed in a similar form using formula (12).

JS divergence has two advantages: (a) symmetry: $JS\left(\hat{P}_i||P_i\right) = JS\left(P_i||\hat{P}_i\right)$; (b) fixed range of values between [0, 1].

This symmetry avoids any asymmetric influences due to contrasting position or the order of two probability distributions, so that any difference between two probability distributions can be measured consistently. Since the value of the JS divergence has a fixed range, it is feasible to adopt this value as the learning rate of agents.

The sensitivity of MARL is enhanced by employing the above method to improve the learning rate.

Algorithm 2 Pseudo-Code of MSNE-MARL

```
Input: the discount rate $\gamma$; the set of state $X$; the space of action $A$;
        the set of agents $I_i$
Initialize: the simulation time $t \leftarrow 0$; the step of simulation $k \leftarrow 0$
        for $i \in I_i$ do
          if $x_i^k \in X_i, a_i^k \in A_i, X_i \in X, A_i \in A$ do
            $Q_i^k\left(x_i^k, a_i^k\right) \leftarrow random$
          end if
        end for
While $t \le T$ do
  for $i \in I_{in}$ do
    search the competitors set of agent $i$: $I_{-i}$
    initially define the mixed strategy set of agents $I_{-i}$: $\sigma_{-i}^* = \{\}$
    for $j \in I_{-i}$
      calculate $P_j$ according $Q_j^k\left(x_j^k, a_j^k\right), a_j^k \in A_j$
      add $P_j\left(a_j^k|x_j^k\right), a_j^k \in A_j$ to $\sigma_{-i}^*$
    end for
    calculate MSNE to get: $\sigma_i^*$
    random seed: $seed \leftarrow random$
    select $a_i^k \in A_i$ according $seed$ and $\sigma_i^*$
    calculate: $r_i^k\left(x_i^k, a_i^k\right)$
    extract $\hat{P}_i \in \sigma_i^*$ and calculate $P_i$ according $Q_i^k\left(x_i^k, a_i^k\right)$, $a_i^k \in A_i$
    update learning rate: $\alpha_i^k = JS\left(\hat{P}_i||P_i\right)$
    update the Q-value according:
      $Q_i^{k+1}\left(x_i^{k+1}, a_i^{k+1}\right)$
        $= (1-\alpha)\,Q_i^k\left(x_i^k, a_i^k\right)$
        $+\alpha_i^k\left[r_i^k\left(x_i^k, a_i^k\right) + \gamma \max_{a_i'\in A(x_i')} Q_i^k\left(x_i', a_i'\right)\right]$
  end for
  $t = t+1, k = k+1$
end while
```

D. ALGORITHM OF MSNE-MARL

Based on the definitions in Section III.C, the MSNE-MARL algorithm performed by each agent can be summarized in ALGORITHM 2. The ALGORITHM 2 is $O\left(n \times |n_{-i}| \times T\right)$, where $|n_{-i}|$ denotes the maximum number of agents in I_{-i}.

IV. NUMERICAL SIMULATION FRAMEWORK

The numerical simulation framework is illustrated in FIGURE 4 and the parameters of the numerical simulation framework are listed in TABLE 3. In FIGURE 4, Parts A and B are the external inputs of the numerical simulation framework: Part A is the time series set of the probability distribution of the vehicular destination based on their origin, Part B is the time series set including the input flow of each origin. Parts C and D describe the dynamic traffic system architecture in the urban network and are both at the core of this numerical simulation framework. Part E is the control module of the framework. The content of the control module has been defined and described in detail in Section III. Part F is the output of the framework containing two modules: the record module and the evaluation module.

The mesoscopic traffic flow model in part D will be introduced in Section IV.A. The method used in part C to model the turning ratio at the intersection will be described in Section IV.B. The components of part F are discussed in Section IV.C. Finally, a summary of the entire numerical

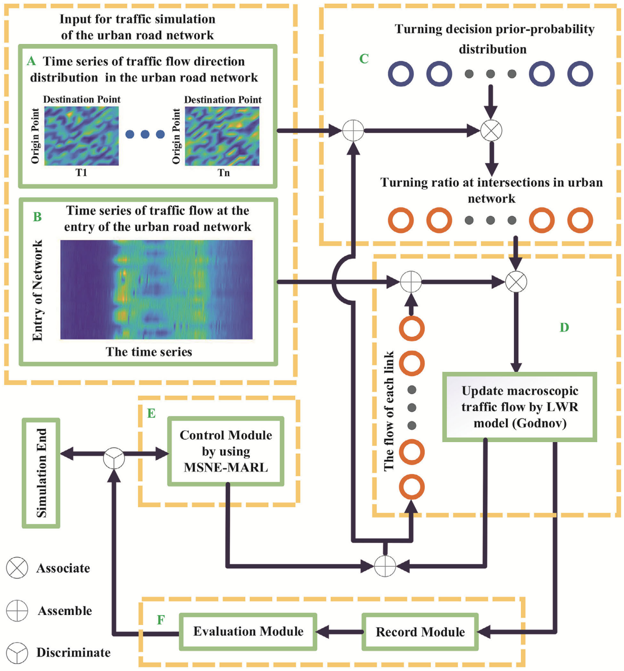
Input for traffic simulation of the urban road network
A Time series of traffic flow direction distribution in the urban road network
Destination Point
Origin Point
T1
Destination Point
Origin Point
Tn
B Time series of traffic flow at the entry of the urban road network
Entry of Network
The time series
C
Turning decision prior-probability distribution
Turning ratio at intersections in urban network
D
The flow of each link
Update macroscopic traffic flow by LWR model (Godnov)
E
Control Module by using MSNE-MARL
Simulation End
F
Evaluation Module
Record Module
Associate
Assemble
Discriminate

FIGURE 4. Overview of the numerical simulation framework.

simulation framework in the form of pseudo-code is given in Section IV.D.

A. MESOSCOPIC TRAFFIC FLOW MODEL

The traffic flow model is the core of the numerical simulation framework. A mesoscopic approach to model traffic flow not only has a high efficiency compared with the microscopic approach, but also has a high evaluation accuracy compared with the macroscopic approach.

To construct the foundation of the numerical simulation framework, we introduced a full-fledged mesoscopic traffic flow model known as the Cell Transmission Model (CTM), which has been proposed by Daganzo [60] and widely employed in traffic simulation [61]–[63]. The CTM employs a finite difference method and Godunov scheme [64] to simplify the solution scheme of the Lighthill-Whitham-Richards (LWR) model [65]. Therefore, CTM can also be regarded as an LWR model in discrete form.

The general discretization form of the LWR equation can be written as follows:

$$\rho_m^{k+1} = \rho_m^k - \frac{\Delta t}{\Delta d}\left(q_{e,m+1}^k - q_{e,m}^k\right) \tag{13}$$

In formula (13), $q_{e,m}^k$ can be defined by the CTM model by introducing a cellular transmission mechanism:

$$q_{e,m}^k = \min\left(q_{s,m-1}^k, q_{r,m}^k, q_{\max}\right) \tag{14}$$

To overcome the unclosed characteristic of the LWR, the components in formula (14) can be further defined by referring to the triangular fundamental diagram. The rationality of this operation has been discussed in [66].

TABLE 3. The parameters of numerical simulation framework.

Notations	Signification
$q_{e,m}^k$	The flow entering m segment after iterating k times
$q_{s,m-1}^k$	The demand flow generated by segment $m-1$, which is expected to send into segment m
$q_{r,m}^k$	The supply flow provided by segment m ,which is the upper limit of capacity that can receive the flow leaving from segment $m-1$
$q_{\max}$	The maximum flow, which is the upper limit of capacity of the section of road
$\xi_{j,i,g}$	The turning from $link_{j,i}$ to $link_{i,g}$.
ξ_d	The destination of vehicles
$P_{link_{j,i}}\left(\xi_{j,i,g} \mid \xi_d\right)$	The turning decision probability distribution of vehicles on $link_{j,i}$
$P_{link_{j,i}}^{\xi_{j,i,g}}$	The turning ratio of $\xi_{j,i,g}$
$P_{link_{j,i}}\left(\xi_d\right)$	The ratio of vehicular destination ξ_d on $link_{j,i}$
$P_{link_{j,i}}\left(\xi_d \mid \xi_{j,i,g}\right)$	The probability of destination ξ_d in turning $\xi_{j,i,g}$
T_{ATD}	The average travel delay value (unit: s/Km)
T_{AT}	The average travel time (unit: s/Km)
T_{AF}	The average free travel time (unit: s/Km) (Also be regard as the minimum travel time per kilometer)
$d_{link_{i,j}}$	The length of $link_{i,j}$ (unit: Km)
$t_{link_{i,j}}$	The travel time of $link_{i,j}$ (unit: s)
$t_{link_{i,j}}^f$	The free travel time of $link_{i,j}$ (unit: s)
$t_{interval}$	The interval time of two contiguous control decisions (unit: s)

B. MODELING TURNING RATIO FOR TRAFFIC FLOW AT INTERSECTIONS IN AN URBAN NETWORK

Existing numerical simulation frameworks generally configure the turning ratio as a constant. However, this is inconsistent with the actual traffic situation making it difficult to reflect the dynamics of a real traffic system and instead providing an excessively anamorphic and utopian traffic environment for control methods built on this type of framework. Additionally, the application of a constant turning ratio also causes a ''clock problem'' in the numerical simulation, where some vehicles travel in circles in the urban network without leaving. The ''clock problem'' can cause the number of deviations to accumulate as the simulation time increases. This accumulation of deviations has the consequence that the traffic control method is evaluated as being not objective enough. Additionally, static traffic allocation is usually employed to determine how vehicles pick applicable routes to their destination within the urban network, but in an actual traffic system, the arrival of vehicles is stochastic and dynamic. It is difficult to reflect the stochastic characteristics of vehicle arrival and the fluctuation of turning ratio at each approach of intersections in an numerical simulation by adopting macroscopic and static traffic allocation.

Based on Bayesian Learning (BL) [67] concepts, we establish a driving direction selection mechanism with prior knowledge to dynamically update the turning ratio of each approach at the intersections in the urban network.

Definition 12: Assume that drivers only select non-detour route strategies in the urban network. Taking into account that each vehicle may have a different destination, the turning decision probability distribution can be defined in the form of a prior-probability distribution (i.e. $P_{link_{j,i}}\left(\xi_{j,i,g}|\xi_d\right)$ *in TABLE 3).*

The turning ratio can then be obtained as follows:

$$P_{link_{j,i}}^{\xi_{j,i,g}} = \sum_{\xi_d} P_{link_{j,i}}\left(\xi_{j,i,g}|\xi_d\right) P_{link_{j,i}}\left(\xi_d\right) \tag{15}$$

The probability of destination ξ_d in turning $\xi_{j,i,g}$ can be deduced as follows:

$$P_{link_{j,i}}\left(\xi_d|\xi_{j,i,g}\right) = \frac{P_{link_{j,i}}\left(\xi_{j,i,g}|\xi_d\right) P_{link_{j,i}}\left(\xi_d\right)}{P_{link_{j,i}}^{\xi_{j,i,g}}} \tag{16}$$

Several conservation constraints should be observed as follows:

$$s.t. \sum_{\xi_d} P_{link_{j,i}}\left(\xi_d\right) = 1 \tag{17}$$

$$\sum_{\xi_{j,i,g}} P_{link_{j,i}}\left(\xi_{j,i,g}|\xi_d\right) = 1 \tag{18}$$

$$\sum_{\xi_{j,i,g}} P_{link_{j,i}}^{\xi_{j,i,g}} = 1 \tag{19}$$

This method can overcome fundamental difficulties due to the "clock problem" fundamentally and its multiple decision process can be regarded as dynamic route selection.

C. RECORD AND EVALUATION OF THE NUMERICAL SIMULATION FRAMEWORK

The record module of the numerical simulation framework is responsible for recording the data generated by the numerical simulation process (e.g. the flow in each link, the control command at each intersection Ω, etc.).

For the evaluation module, the Performance Index (PI) used to evaluate the effect of UNTC should be further explained. The main target of traffic control at a network level should be considered in selection of the PI. Therefore, we employed the average travel delay value as the PI to assess the traffic control effect in an urban network. The average travel delay value is defined as follows:

$$T_{ATD} = T_{AT} - T_{AF} \tag{20}$$

$$T_{AT} = \frac{\sum_i \sum_j t_{link_{i,j}}}{\sum_i \sum_j d_{link_{i,j}}} \tag{21}$$

$$T_{AF} = \frac{\sum_i \sum_j t^f_{link_{i,j}}}{\sum_i \sum_j d_{link_{i,j}}} \tag{22}$$

Therefore, the relevant PI can be immediately extracted by the numerical simulation framework using the LWR (Godunov) model. It should be emphasized that queue delays are not only caused by queuing at intersections, but also due to congestion delays caused by excessive vehicles on the roads within the urban network. The average travel delay value not only integrates both of these scenarios, but also eliminates the potential impact of possible inconsistencies in road length by calculating the ratio of time to distance. Therefore, it is appropriate to adopt the average travel delay value as the PI.

D. PSEUDO-CODE OF THE NUMERICAL SIMULATION FRAMEWORK

Combining the above definitions, the numerical simulation framework is summarized in FRAMEWORK 1. Therefore, the entire numerical simulation framework has now been fully constructed.

V. NUMERICAL SIMULATION EXPERIMENT

This section presents the results of the numerical simulation to evaluate the performance of the proposed distributed

Framework 1 Pseudo-Code of the Numerical Simulation Framework

```
Input: the set of links $L$; the set of state $X$; the space of action $A$;
       the set of intersections (agents) $I_i$
Initialize: the simulation time $t \leftarrow 0$; the step of simulation $k \leftarrow 0$
       initialize the Q-value for all agent according MSNE-MARL
       initialize the condition of urban network
While $t \leq T$ do
  if mod $(t, t_{interval}) == 1$ do
    for $i \in I_i$ do
      search optimal traffic control actions $a_i^k$ according MSNE-MARL
    end for
    update the record of control command $\Omega$ according $A$
  end if
  for $i \in I_i$ do
    calculate the throughput of intersection(agent) $i$
  end for
  for $link_{i,j} \in L$ do
    update the condition $X$ according LWR (Godunov)
  end for
  if mod $(t, t_{interval}) == 0$ do
    for $i \in I_i$ do
      calculate $r_i^k\left(x_i^k, a_i^k\right)$ according the change of links' state around
      the intersection(agent) $i$
    end for
    update the Q-value in MSNE-MARL according MSNE-MARL
  end if
  $t = t + 1, k = k + 1$
end while
```

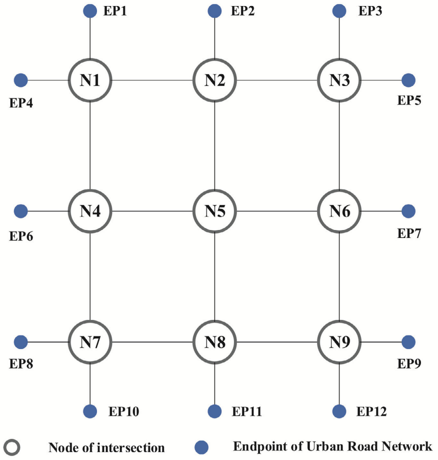

FIGURE 5. Sketch of the 3×3 grid network.

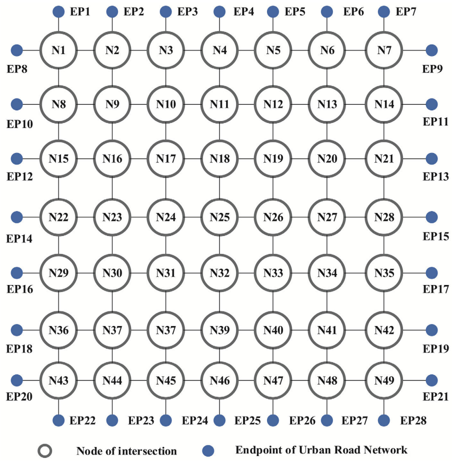

FIGURE 6. Sketch of the 7×7 grid network.

control method based on MSNE-MARL. The network topology, the contrast methods and the experimental scenarios are described in section V.A. Then, the optimal learning rate of the control method named II-MARL is obtained in section V.B. Ablation study of MSNE-MARL is implemented in section V.C. The method proposed in this paper is compared with two control methods for four different scenarios (i.e. scenarios 1-4) in 3×3 grid network, and the corresponding results are presented in section V.D. Furtherly, the suggested method is also compared with two control methods in 7×7 grid network with one scenario (i.e. scenario 5), and the corresponding results are presented in section V.E. The analysis and evaluation are implemented using an applicable and scientific evaluation platform built with MATLAB according to the numerical simulation framework described in section IV.

A. CONFIGURATION OF NUMERICAL SIMULATION EXPERIMENT

The numerical simulation experiment is described in terms of three aspects: network topology, contrasting methods and scenarios.

1) NETWORK TOPOLOGY

Two grid networks of the intersections are employed to assess the performance of MSNE-MARL as shown in FIGURE 5 and 6. The 3×3 grid network consists of 9 nodes, 12 endpoints and their 48 connecting lines. The 7×7 grid network consists of 49 nodes, 28 endpoints and their 224 connecting lines. In FIGURE 5 and 6, the numbered gray hollow circles and the blue solid circles represent the intersections and access points of the urban network respectively. The solid black lines indicate the two-way roads. Therefore, each intersection contains four approaches and each link is 1000 m in length.

2) CONTRASTING METHODS

The performance of MSNE-MARL is assessed with reference to two control methods: (a) a Fixed-Time Control (FTC) method; (b) an Independent Individual Multi-Agent Reinforcement Learning (II-MARL) control method.

a: FTC

For each signal controller of FTC, the same sequence of phases is repeated for a fixed duration, which is always arranged in the same order and represents a fixed cycle. In this paper, the cycle time employed in [36] and [68] is considered and a phase duration for the signal controller with a total cycle time of 120 s (25 s for each phase with a 5 s interval) is adopted in this paper.

b: II-MARL

There are various applications in the literature using the MARL control method. However, the control effect may be affected by different definitions in the literature on some aspects, including state division, action selection, reward function and learning rate. In order to avoid deviations due to these conditions, we set the definitions and parameter for II-AMRL as defined below:

(1) The decision-making process adopted for II-MARL is Boltzmann exploration (softmax);

(2) The learning rate of II-MARL is a constant equal to 0.001 (see Section V.B);

(3) The discount factor of II-MARL is a constant equal to 0.5.

The other definitions of II-MARL are the same as the corresponding definitions in MSNE-MARL.

3) SCENARIOS

Four different scenarios in 3×3 grid network were designed from a scientific and objective perspective to analyze the influences of disequilibrium distribution and stochastic fluctuation in traffic demand on the control effects of the three control methods. Then, a scenario in 7×7 grid network was designed to analyze the influence of network size on the control effects of the three control methods.

Stochastic fluctuation in traffic demand can be divided into two categories: (a) a rush in traffic demand at the origin of

TABLE 4. Configuration for scenario 1.

	Initialization		Simulation				
Endpoint ID	Input (pcu/h)	The distribution of OD	Input (pcu/h)				The distribution of OD
EP1	800	uniform distribution	800				uniform distribution
EP2	800		800				
EP3	800		800				
EP4	800		800				
EP5	800		800				
EP6	800		800				
EP7	800		800				
EP8	800		800				
EP9	800		800				
EP10	800		**linear increase from 800 to 2400**	2400	**linear decrease from 2400 to 800**	800	
EP11	800		800				
EP12	800		800				
Duration Time (s)	3600		900	1800	900	3600	
Aggregate Time (s)	3600		7200				
Distribution of arrival rate	uniform distribution		uniform distribution				

the urban network; (b) changes in origin-destination distribution in the urban network. For disequilibrium distribution in traffic demand, there are two different arrival rates that can have an influence on the control effect. The arrival rate can have a uniform distribution or a Poisson distribution. The configurations of the four different scenarios are provided in TABLES 4-7.

In TABLES 4-7, scenario 1 describes stochastic fluctuation in traffic demand due to a rush in traffic demand at the origin of the network; scenario 2 describes stochastic fluctuation in traffic demand due to changes in origin-destination distribution; scenario 3 represents a Poisson arrival rate and scenario 4 represents a uniform distribution of arrival rates with disequilibrium traffic flow.

To analyze the effect of different network scales, a scenario which represents a Poisson arrival rate with disequilibrium traffic flow is provided in TABLE 8.

B. OPTIMAL LEARNING RATE OF II-MARL

To search the optimal learning rate of II-MARL, the impact of learning rate is discussed in this section. The learning rate of II-MARL ranges from 10^{-1} to 10^{-5}. The influence of different learning rates on II-MARL control effect is analyzed based on the 3×3 grid network and scenario 3. The control effect curve is drawn in FIGURE 7.

As can be seen in FIGURE 7, the optimal learning rate of II-MARL is 0.001 (10^{-3}). Unless otherwise specified, the learning rate used by II-MARL in the following experiments is 0.001.

C. ABLATION STUDY OF MSNE-MARL

To analyze the control effect improvement of MSNE-MARL under the improved decision-making process, ablation study of MSNE-MARL is necessary.

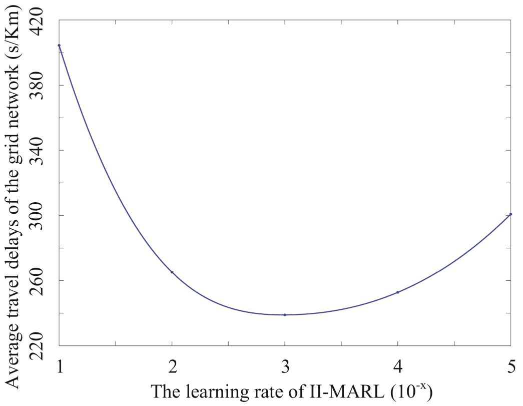

FIGURE 7. Average travel delay curve controlled by II-MARL with different learning rates.

According to the description in section V.A, the II-MARL can be regarded as an MSNE-MARL with default learning rate, default decision-making process. In addition, since the learning rate improvement of MSNE-MARL is based on the improved decision-making process, it is impossible to set a MSNE-MARL with adaptive learning rate and default decision-making process. Therefore, II-MARL with default learning rates (0.1 and 0.001), ablative MSNE-MARL with default learning rates (0.1 and 0.001) and MSNE-MARL are tested under scenarios 3 and 4 in 3×3 grid network for 7200 seconds. The average travel delays of grid network over the duration time controlled by different methods under different scenarios are listed in TABLE 9.

It can be found from TABLE 9 that MSNE-MARL performs best in scenario 3 and 4. Compared with II-MARL with default learning rate 0.001, MSNE-MARL improves the control effect by 21.26% while ablative MSNE-MARL with default learning rate 0.001 improves the control effect by only 18.42% in scenario 3. In scenario 4, MSNE-MARL and abla-

TABLE 5. Configuration for scenario 2.

Endpoint ID	Initialization		Simulation	
	Input (pcu/h)	The distribution of OD	Input (pcu/h)	The distribution of OD
EP1	1000	uniform distribution	1000	uniform distribution
EP2	1000		1000	
EP3	1000		1000	
EP4	1000		1000	
EP5	1000		1000	
EP6	1000		1000	
EP7	1000		1000	
EP8	1000		1000	
EP9	1000		1000	
EP10	1000		1000	
EP11	1000		1000	
EP12	1000		1000	**Main direction: EP12-EP1 Other direction: uniform distribution**
Aggregate Time (s)	3600		7200	
Distribution of arrival rate	uniform distribution		uniform distribution	

TABLE 6. Configuration for scenario 3.

Endpoint ID	Initialization		Simulation	
	Input (pcu/h)	The distribution of OD	**Input** (pcu/h)	The distribution of OD
EP1	1000	uniform distribution	1198	uniform distribution
EP2	1000		2263	
EP3	1000		1335	
EP4	1000		258	
EP5	1000		1319	
EP6	1000		1078	
EP7	1000		1968	
EP8	1000		1928	
EP9	1000		424	
EP10	1000		2363	
EP11	1000		800	
EP12	1000		2230	
Aggregate Time (s)	3600		7200	
Distribution of arrival rate	uniform distribution		**Poisson distribution**	

TABLE 7. Configuration for scenario 4.

Endpoint ID	Initialization		Simulation	
	Input (pcu/h)	The distribution of OD	**Input** (pcu/h)	The distribution of OD
EP1	1000	uniform distribution	1198	uniform distribution
EP2	1000		2263	
EP3	1000		1335	
EP4	1000		258	
EP5	1000		1329	
EP6	1000		1078	
EP7	1000		1968	
EP8	1000		1928	
EP9	1000		424	
EP10	1000		2373	
EP11	1000		800	
EP12	1000		2230	
Aggregate Time (s)	3600		7200	
Distribution of arrival rate	uniform distribution		**uniform distribution**	

TABLE 8. Configuration for scenario 5.

Endpoint ID	Initialization		Simulation	
	Input (pcu/h)	The distribution of OD	**Input** (pcu/h)	The distribution of OD
EP1	1000	uniform distribution	1992	uniform distribution
EP2	1000		2193	
EP3	1000		479	
EP4	1000		2209	
EP5	1000		1591	
EP6	1000		415	
EP7	1000		813	
EP8	1000		1403	
EP9	1000		2307	
EP10	1000		2323	
EP11	1000		547	
EP12	1000		2335	
EP13	1000		2306	
EP14	1000		1268	
EP15	1000		1961	
EP16	1000		512	
EP17	1000		1128	
EP18	1000		2215	
EP19	1000		1943	
EP20	1000		2311	
EP21	1000		1643	
EP22	1000		279	
EP23	1000		2068	
EP24	1000		2255	
EP25	1000		1693	
EP26	1000		1867	
EP27	1000		1835	
EP28	1000		1063	
Aggregate Time (s)	3600		3600	
Distribution of arrival rate	uniform distribution		**Poisson distribution**	

tive MSNE-MARL with default learning rate 0.001 improve the control effect by 29.53% and 26.76% respectively based on the benchmark which is II-MARL with default learning rate 0.001. From the above statement, it can be concluded that the improved decision-making process has a major impact on the performance improvement of MSNE-MARL. In addition, although the improved adaptive learning rate has a weak influence on the performance of MSNE-MARL, it still explores the potential performance of MSNE-MARL. In scenarios 3 and 4, comparing and analyzing the effects of different default learning rates on II-MARL and MSNE-MARL, it shows that the variation of the learning rate has a weaker effect on MSNE-MARL than II-MARL. It also indirectly verifies that adaptive learning rate performs minor on the performance improvement of MSNE-MARL.

D. EVALUATION RESULTS UNDER THE 3×3 GRID NETWORK

The average travel delay is used as the performance index to assess the control effects of the three control

TABLE 9. Evaluation results of ablation study.

Test Methods (The learning rate)	The average travel delay of grid network (s) Scenario 3	 Scenario4
II-MARL(0.001)	217.4597	238.9469
II-MARL(0.1)	404.4374	436.0783
MSNE-MARL(0.001)	177.3975	174.9978
MSNE-MARL(0.1)	180.6042	179.0649
MSNE-MARL	**171.2179**	**168.3874**

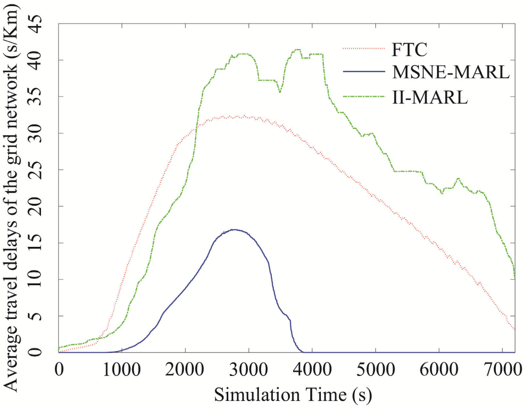

FIGURE 8. Average travel delay curves for scenario 1.

methods under the four different scenarios (i.e. scenarios 1-4) in 3×3 grid network.

1) SCENARIO 1

As can be seen in FIGURE 8, the initial average travel delay values using all three control methods are 0. Although some travel time delay of FTC appears at the start, it does not exceed 1s/Km until after 519s into the simulation. The red dashed curve for FTC climbs moderately from 517s to 2930s and approaches a peak of 32.43s/Km. After a few seconds at the peak, the average travel delay value of FTC begin to decline slowly. This suggests that the decrease in average travel delay value has a distinct hysteretic nature. The early stage trend of the green dash-dot curve for II-MARL is similar to the red dotted curve for FTC. Although the average travel delay values of II-MARL in early stage exceed that of FTC within 0.5-0.8s/km, its trend has been held until 1013s. The phenomenon indicates that II-MARL can suppress traffic congestion to some extent. However, due to a lack of global information, the average travel delay value of II-MARL increases rapidly from 1013s to 3623s and approaches a peak of 41.425s/Km at 3755s. The short-term drop in the middle of the green dash-dot curve indicates that II-MARL has experiential learning ability. The green dash-dot curve also shows a slow downward trend after the peak which is similar but higher than the red dotted curve in late term. For MSNE-MARL, it can be seen from the blue solid curve that compared with the other methods, the travel delay of MSNE-MARL appears later at 740s and disappears earlier at 3866s, and reaches a lower peak value of 16.821s/Km. It can also be seen that the value of the average travel delay increases slowly but subsequently decreases rapidly. In terms of the cumulative improvement effect over 7200s of the simulation, MSNE-MARL has an improvement compared with FTC and II-MARL of 81.86% and 85.51% respectively.

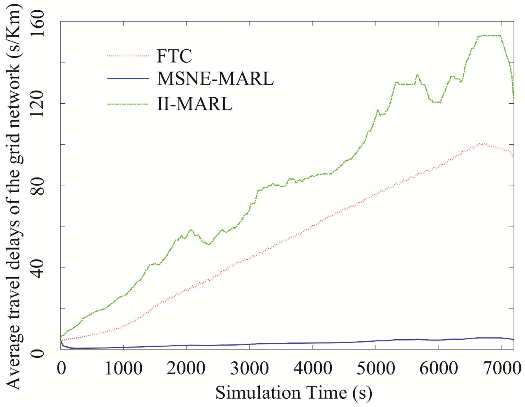

FIGURE 9. Average travel delay curves for scenario 2.

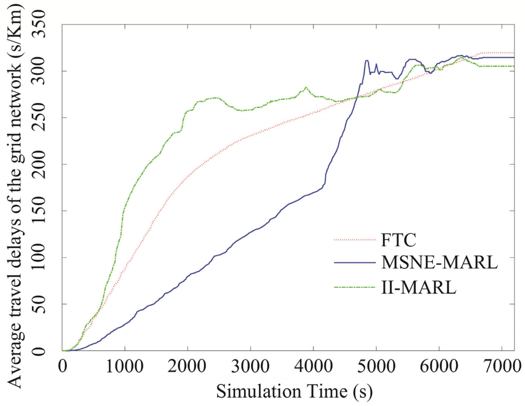

FIGURE 10. Average travel delay curves for scenario 3.

2) SCENARIO 2

In FIGURE 9 for scenario 2, as the distribution adjustment of the OD ratio is not recovered during the simulation, the average travel delay values of FTC and II-MARL keep increasing for 6500s. it can be observed by the red dotted curve for FTC that the average travel delay of FTC gradually converges in the range of 90.05-96.355s/Km after 6500s. at this time, the green dash-dot curve shows that the average travel delay value of II-MARL declines rapidly. this phenomenon can be regarded as being due to its powerful experiential learning ability. however, for MSNE-MARL, the maximum delay is not more than 5s/km from the beginning to the end of the simulation indicating that MSNE-MARL has the best control effect for scenario 2.

3) SCENARIO 3

In FIGURE 10, the green dash-dot curve for II-MARL shows a ladder-like characteristics, which indicates that II-MARL has faced two remarkably different traffic states in the simulation. the red dotted curve for FTC slows down

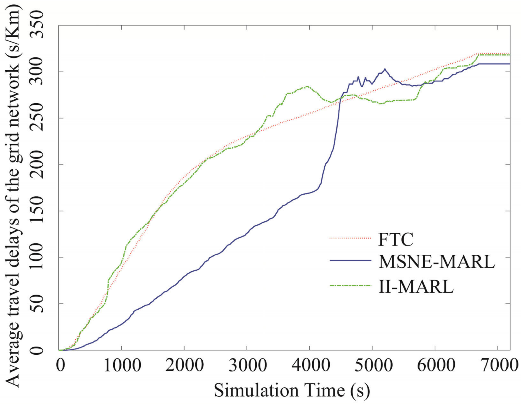

FIGURE 11. **Average travel delay curves for scenario 4.**

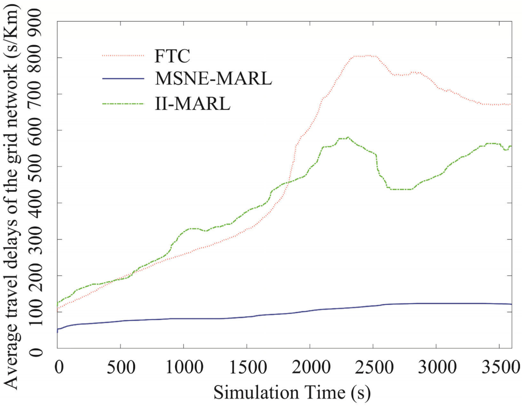

FIGURE 12. **Average travel delay curves for scenario 5.**

after 2000s, suggesting that the average travel delay value of FTC raises steadily until it stabilizes near its convergence value. the blue solid curve for MSNE-MARL has a slower trend than the red dotted curve until it reaches 4180s when it starts to raise rapidly between 4180s and 4835s and intersects with the red dotted curve at 4638s. although there is some fluctuations after 4835s, the convergence of the blue solid curve is similar to the red dotted curve. MSNE-MARL has a cumulative improvement effect over 7200s that is 21.33% and 28.34% higher than FTC and II-MARL respectively.

4) SCENARIO 4

The trends and forms of the curves in FIGURE 11 are similar to those in FIGURE 10. By comparing the blue solid curves for MSNE-MARL in FIGURE 10 and FIGURE 11, it can be seen that the near convergence fluctuation of curve for MSNE-MARL in scenario 4 is smaller than in scenario 3. MSNE-MARL has a cumulative improvement over the 7200s of the simulation of 1.65% higher in scenario 4 than in scenario 3. Similarly, the effect of II-MARL increases by 8.99% compared with scenario 3. This result can be explained by the fact that II-MARL only has a regional learning ability. Its lack of global learning ability leads to an accumulation of delays due to global variations. This is evidence of MSNE-MARL's global adaptability. However, FTC has the same cumulative control effects in both scenarios indicating that FTC is not sensitive to distribution variations in arrival rate.

E. EVALUATION RESULTS UNDER THE 7×7 GRID NETWORK

The average travel delay is used as the performance index to assess the control effects of the three control methods under the scenario 5 in 7×7 grid network.

1) SCENARIO 5

In FIGURE 12, the red dotted curve for FTC can be divided into three stages: (1) slow increase stage; (2) rapid ascent stage and (3) slow descent stage. The first-stage of the red dotted curve shows that the traffic congestion pressure is small in the early stage with increasing the size of network. As traffic flows gather towards the network center, the traffic congestion pressure increases rapidly causing the occurrence of the phenomenon which is presented by the second-stage red dotted curve. The third-stage red dotted curve confirms that reducing formed traffic congestion is a slow process.

The green dash-dot curve for II-MARL is similar to the FTC curve in the early stage. However, the peak of II-MARL curve is 221.91s smaller than that of FTC curve. It fully shows that II-MARL has a better control effect on traffic congestion than FTC when the network size increases. The green dash-dot curve presents an undulating state in the middle and later stage. The short-lived decline of II-MARL curve indicates that II-MARL has excellent learning ability to reduce the traffic congestion rapidly. In addition, the defect of II-MARL in the perception of local surroundings is verified again. This defect is the reason causing the again increase of green dash-dot curve. It can be regarded as the result of the shift of congestion centers which has happened in the large scale network.

For MSNE-MARL, it can be seen from the blue solid curve that the trend of MSNE-MARL curve appears flat. All travel delays of MSNE-MARL over 3600s are smaller than those of II-MARL and FTC. Moreover, the peak travel delay of MSNE-MARL is 123.26s/km. These performances fully indicate that MSNE-MARL has an excellent performance in restraining traffic congestion. MSNE-MARL has a cumulative improvement effect over 3600s that is 79.54% and 75.19% higher than FTC and II-MARL respectively. Compared with the cumulative improvement of MSNE-MARL from 1s to 3600s in scenario 3, cumulative improvement of MSNE-MARL in scenario 5 indicates that the improvement of MSNE-MARL is more obvious with increasing the size of network.

VI. CONCLUSION

In this article, a distributed control method has been presented for UNTC based on the principle of MARL and GT to prevent disturbance-based traffic congestion in an urban network. A flat hierarchical architecture of agents is employed, where each agent is associated with a signalized intersection and competes with adjacent agents. To accelerate the convergence of MARL, a novel decision mechanism is designed for each agent by integrating the concept of regional MSNE. We have particularly focused on improving the learning rate, as this can affect the global search ability of MSNE-MARL and the effectiveness of MARL in facing stochastic fluctuations. With this improvement, UNTC system can respond to disturbances rapidly and effectively to prevent disturbance-based traffic congestion from emerging in an urban network and thus UNTC can achieve its desired objectives, which are to minimize the average travel delay in an urban network.

To assess the proposed method, a detailed comparative analysis of performance is provided by benchmarking the suggested MSNE-MARL method against two control strategies (i.e. FTC and II-MARL). MSNE-MARL and its contrasting methods were tested for five different traffic situation simulations. The results of the comparative assessment show that MSNE-MARL outperforms the other methods in terms of average travel delay in the grid network for the first two scenarios and scenario 5. For scenarios 3 and 4, although MSNE-MARL has a similar final effect to FTC and II-MARL, MSNE-MARL shows excellent performance for the early and medium term. Moreover, MSNE-MARL demonstrates superior control performance in large size network in scenario 5.

Many definitions for MSNE-MARL have been proposed to establish a distributed control method for preventing disturbance-based urban network traffic congestion. The overall structure of MSNE-MARL framework has universality and portability. The framework of MSNE-MARL can be applied to other similar scenarios with a MAS architecture.

REFERENCES

[1] C. Wright and P. Roberg, "The conceptual structure of traffic jams," *Transp. Policy*, vol. 5, no. 1, pp. 23–35, Jan. 1998.

[2] M. Papageorgiou, C. Diakaki, V. Dinopoulou, A. Kotsialos, and Y. Wang, "Review of road traffic control strategies," *Proc. IEEE*, vol. 91, no. 12, pp. 2041–2042, Dec. 2003.

[3] D. Robertson and R. Bretherton, "Optimizing networks of traffic signals in real time-the SCOOT method," *IEEE Trans. Veh. Technol.*, vol. 40, no. 1, pp. 11–15, Feb. 1991.

[4] A. Stevanovic, J. Stevanovic, and C. Kergaye, "Optimization of traffic signal timings based on surrogate measures of safety," *Transp. Res. C, Emerg. Technol.*, vol. 32, pp. 159–178, Jul. 2013.

[5] J. Chen and L. Xu, "Road-junction traffic signal timing optimization by an adaptive particle swarm algorithm," in *Proc. 9th ICARCV*, Singapore, vols. 1–7, Dec. 2006, pp. 1–7.

[6] X. Zong, S. Xiong, and Z. Fang, "A conflict-congestion model for pedestrian-vehicle mixed evacuation based on discrete particle swarm optimization algorithm," *Comput. Oper. Res.*, vol. 44, pp. 1–12, Apr. 2014.

[7] A. Jovanović, M. Nikolić, and D. Teodorović, "Area-wide urban traffic control: A bee colony optimization approach," *Transp. Res. C, Emerg. Technol.*, vol. 77, pp. 329–350, Apr. 2017.

[8] Y. Jiang and J. C. Jiang, "Diffusion in social networks: A multiagent perspective," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 45, no. 2, pp. 198–213, Feb. 2015.

[9] Y. Jiang, "Concurrent collective strategy diffusion of multiagents: The spatial model and case study," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 39, no. 4, pp. 448–458, Jul. 2009.

[10] Y. Jiang, J. Hu, and D. Lin, "Decision making of networked multiagent systems for interaction structures," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 41, no. 6, pp. 1107–1121, Nov. 2011.

[11] S. El-Tantawy and B. Abdulhai, "Towards multi-agent reinforcement learning for integrated network of optimal traffic controllers (MARLIN-OTC)," *Transp. Lett.*, vol. 2, no. 2, pp. 89–110, Apr. 2010.

[12] T. Tan, TS. Chu, B. Peng, and J. Wang, "Large-scale traffic grid signal control using decentralized fuzzy reinforcement learning," in *Proc. IntelliSys*, London, U.K., 2016, pp. 652–662.

[13] B. Abdulhai and L. Kattan, "Reinforcement learning: Introduction to theory and potential for transport applications," *Can. J. Civil Eng.*, vol. 30, no. 6, pp. 981–991, Dec. 2003.

[14] C.-Y. Lee, "Nash-profit efficiency: A measure of changes in market structures," *Eur. J. Oper. Res.*, vol. 255, no. 2, pp. 659–663, Dec. 2016.

[15] C.-Y. Lee, "Mixed-strategy Nash equilibrium in data envelopment analysis," *Eur. J. Oper. Res.*, vol. 266, no. 3, pp. 1013–1024, 2018.

[16] D. I. Robertson, "TANSYT method for area traffic control," *Traffic Eng. Control.*, vol. 11, no. 6, pp. 276–281, 1969.

[17] S.-W. Chiou, "Optimization of area traffic control for equilibrium network flows," *Transp. Sci.*, vol. 33, no. 3, pp. 279–289, Aug. 1999.

[18] S. Wong, W. Wong, C. Leung, and C. Tong, "Group-based optimization of a time-dependent TRANSYT traffic model for area traffic control," *Transp. Res. B, Methodol.*, vol. 36, no. 4, pp. 291–312, May 2002.

[19] H. Ceylan and M. G. H. Bell, "Traffic signal timing optimisation based on genetic algorithm approach, including drivers' routing(Article)," *Transp. Res. B, Methodol.*, vol. 38, no. no. 4, pp. 329–342, 2004.

[20] N. Gartner, F. Pooran, and C. Andrews, "Implementation of the OPAC adaptive control strategy in a traffic signal network," in *Proc. IEEE Intell. Transp. Syst. (ITSC)*, Oakland, CA, USA, Nov. 2002, pp. 195–200.

[21] P. Mirchandani and L. Head, "A real-time traffic signal control system: Architecture, algorithms, and analysis," *Transp. Res. C, Emerg. Technol.*, vol. 9, no. 6, pp. 415–432, Dec. 2001.

[22] J. Maciejowski, *Predictive Control: With Constraints*, London, U.K.: Prentice-Hall, 2002.

[23] K. Aboudolas, M. Papageorgiou, A. Kouvelas, and E. Kosmatopoulos, "A rolling-horizon quadratic-programming approach to the signal control problem in large-scale congested urban road networks," *Transp. Res. C, Emerg. Technol.*, vol. 18, no. 5, pp. 680–694, Oct. 2010.

[24] A. Jamshidnejad, I. Papamichail, M. Papageorgiou, and B. De Schutter, "A mesoscopic integrated urban traffic flow-emission model," *Transp. Res. C, Emerg. Technol.*, vol. 75, pp. 45–83, Feb. 2017.

[25] J. G. Dai and W. Lin, "Maximum pressure policies in stochastic processing networks," *Oper. Res.*, vol. 53, no. 2, pp. 197–218, Apr. 2005.

[26] X.-F. Xie, S. F. Smith, L. Lu, and G. J. Barlow, "Schedule-driven intersection control," *Transp. Res. C, Emerg. Technol.*, vol. 24, pp. 168–189, Oct. 2012.

[27] P. Varaiya, "Max pressure control of a network of signalized intersections," *Transp. Res. C, Emerg. Technol.*, vol. 36, pp. 177–195, Nov. 2013.

[28] M. Rinaldi, W. Himpe, and C. M. Tampère, "A sensitivity-based approach for adaptive decomposition of anticipatory network traffic control," *Transp. Res. C, Emerg. Technol.*, vol. 66, pp. 150–175, May 2016.

[29] S. El-Tantawy and B. Abdulhai, "An agent-based learning towards decentralized and coordinated traffic signal control," in *Proc. 13th IEEE ITSC*, Madeira Island, Portugal, Sep. 2010, pp. 665–670.

[30] M. M. Al-Tarabily, R. F. Abdel-Kader, G. Abdel Azeem, and M. I. Marie, "Optimizing dynamic multi-agent performance in E-learning environment," *IEEE Access*, vol. 6, pp. 35631–35645, 2018.

[31] Y. Quan, W. Chen, Z. Wu, and L. Peng, "Observer-based distributed fault detection and isolation for heterogeneous discrete-time multiagent systems with disturbances," *IEEE Access*, vol. 4, pp. 4652–4658, 2016.

[32] W. Housseyni, O. Mosbahi, M. Khalgui, Z. Li, L. Yin, and M. Chetto, "Multiagent architecture for distributed adaptive scheduling of reconfigurable real-time tasks with energy harvesting constraints," *IEEE Access*, vol. 6, pp. 2068–2084, 2018.

[33] C. Dou, D. Yue, Q.-L. Han, and J. M. Guerrero, "Multi-agent system-based event-triggered hybrid control scheme for energy Internet," *IEEE Access*, vol. 5, pp. 3263–3272, 2017.

[34] S. El-Tantawy, B. Abdulhai, and H. Abdelgawad, "Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (MARLIN-ATSC): Methodology and large-scale application on downtown Toronto," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 3, pp. 1140–1150, Sep. 2013.

[35] I. Kosonen, "Multi-agent fuzzy signal control based on real-time simulation," *Transp. Res. C, Emerg. Technol.*, vol. 11, no. 5, pp. 389–403, Oct. 2003.

[36] S. Darmoul, S. Elkosantini, A. Louati, and L. Ben Said, "Multi-agent immune networks to control interrupted flow at signalized intersections," *Transp. Res. C, Emerg. Technol.*, vol. 82, pp. 290–313, Sep. 2017.

[37] L. B. De Oliveira and E. Camponogara, "Multi-agent model predictive control of signaling split in urban traffic networks," *Transp. Res. C, Emerg. Technol.*, vol. 18, no. 1, pp. 120–139, Feb. 2010.

[38] A. L. C. Bazzan and F. Klügl, "A review on agent-based technology for traffic and transportation," *Knowl. Eng. Rev.*, vol. 29, no. 3, pp. 375–403, Jun. 2014.

[39] B. Chen and H. H. Cheng, "A review of the applications of agent technology in traffic and transportation systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 2, pp. 485–497, Jun. 2010.

[40] T. Thorpe, "Vehicle traffic light control using SARSA," M.S. thesis, Dept. of Comput. Sci, Colorado State Univ., Fort Collins, CO, USA, 1997.

[41] B. Abdulhai and G. J. Karakoulas, "Reinforcement learning for true adaptive traffic signal control," *J. Transp. Eng.*, vol. 129, no. 3, pp. 278–285, 2003.

[42] J. Jin and X. Ma, "Adaptive group-based signal control by reinforcement learning," *Transp. Res. Procedia*, vol. 10, no. 1, pp. 207–216, 2015.

[43] I. Arel, C. Liu, T. Urbanik, and A. Kohls, "Reinforcement learning-based multi-agent system for network traffic signal control," *IET Intell. Transp. Syst.*, vol. 4, no. 2, p. 128, 2010.

[44] F. Zhu, H. A. Aziz, X. Qian, and S. V. Ukkusuri, "A junction-tree based learning algorithm to optimize network wide traffic control: A coordinated multi-agent framework," *Transp. Res. C, Emerg. Technol.*, vol. 58, pp. 487–501, Sep. 2015.

[45] A. L. Bazzan, D. De Oliveira, and B. C. Da Silva, "Learning in groups of traffic signals," *Eng. Appl. Artif. Intell.*, vol. 23, no. 4, pp. 560–568, Jun. 2010.

[46] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 1998.

[47] A. L. C. Bazzan, "Opportunities for multiagent systems and multiagent reinforcement learning in traffic control," *Auton. Agent Multi-Agent Syst.*, vol. 18, no. 3, pp. 342–375, Jun. 2009.

[48] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Proc. ICML*, New Brunswick, NJ, USA, 1994, pp. 157–163.

[49] J. Hu and M. P. Wellman, "Nash Q-learning for general-sum stochastic games," *J. Mach. Learn. Res.*, vol. 4, pp. 1039–1069, Nov. 2003.

[50] A. Greenwald, M. Zinkevich, and P. Kaelbing, "Correlated Q-learning," in *Proc. ICML*, Washington, DC, USA, 2003, pp. 242–249.

[51] T. Kohonen, "Things you haven't heard about the self-organizing map," in *Proc. ICNN*, San Francisco, CA, USA, 1993, pp. 1147–1156.

[52] Z. Du, C. Wang, Y. Sun, and G. Wu, "Context-aware indoor VLC/RF heterogeneous network selection: Reinforcement learning with knowledge transfer," *IEEE Access*, vol. 6, pp. 33275–33284, 2018.

[53] W.-K. Wong, H.-Y. Chen, C.-Y. Hsu, and T.-K. Chao, "Reinforcement learning of robotic motion with genetic programming, simulated annealing and self-organizing map," in *Proc. Int. Conf. Technol. Appl. Artif. Intell.*, Washington, DC, USA, Nov. 2011, pp. 292–298.

[54] J. Nash, "Non-cooperative games," *Ann. Math.*, vol. 54, no. 2, pp. 286–295, 1951.

[55] D. Fudenberg and J. Tirole, *Game Theory*. Cambridge, MA, USA: MIT Press, 1991.

[56] J. F. Nash, "Equilibrium points in N-person games," *Proc. Nat. Acad. Sci. USA*, vol. 36, no. 1, pp. 48–49, Jan. 1950.

[57] C. Wang, Z. Shi, L. Chang, *Multi-Agent Systems and Their Applications*, vol. 2. Beijing, China: Tsinghua Univ. Press, 2003.

[58] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, 1992.

[59] *Traffic Controller Assemblies With NTCIP Requirements Version 03.07, TS 2-2016*, NEMA, Rosslyn, VA, USA, 2016.

[60] C. F. Daganzo, "The cell transmission model, part II: Network traffic," *Transp. Res. B, Methodol.*, vol. 29, no. 2, pp. 79–93, Apr. 1995.

[61] C. Xu, Z. Li, Z. Pu, Y. Guo, and P. Liu, "Procedure for determining the deployment locations of variable speed limit signs to reduce crash risks at freeway recurrent bottlenecks," *IEEE Access*, vol. 7, pp. 47856–47863, 2019.

[62] Y. Xu, Y. Wang, J. He, M. Su, and P. Ni, "Resilience-oriented distribution system restoration considering mobile emergency resource dispatch in transportation system," *IEEE Access*, vol. 7, pp. 73899–73912, 2019.

[63] H. Yuan, R. Wang, X. Zhang, Y. Hu, F. Zhang, T. Zhu, and H. Liu, "Evacuation strategy optimization study based on system theory," *IEEE Access*, vol. 7, pp. 111232–111244, 2019.

[64] S. K. Godunov, "A difference scheme for numerical solution of discontinuous solution of hydrodynamic equations," *Mat. Sbornik*, vol. 47, no. 3, pp. 271–306, 1959.

[65] P. I. Richards, "Shock waves on the highway," *Oper. Res.*, vol. 4, no. 1, pp. 42–51, Feb. 1956.

[66] L. Li, R. Jiang, B. Jia, *Modern Traffic Flow Theory and Application: Volume 1, Freeway Traffic Flow*. Beijing, China: Tsinghua Univ. Press, 2011.

[67] I. Pan and D. Bester, "Fuzzy Bayesian learning," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 3, pp. 1719–1731, Jun. 2018.

[68] L.-W. Chen, P. Sharma, and Y.-C. Tseng, "Dynamic traffic control with fairness and throughput optimization using vehicular communications," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 504–512, Sep. 2013.

ZHAOWEI QU received the Ph.D. degree in traffic information engineering and control from Jilin University, Changchun, China. He is currently a Professor and a Ph.D. Supervisor with the School of Transportation, Jilin University. His current research interests mainly include traffic control theory, traffic organization, traffic big data, and intelligent transportation systems.

ZHAOTIAN PAN received the B.Sc. degree in traffic engineering from Jilin University, Changchun, China, in 2016, where he is currently pursuing the Ph.D. degree in traffic information engineering and control. His research interests include intelligent transportation systems, traffic flow theory, traffic simulation, and the application of game theory and multiagent systems in the above fields.

YONGHENG CHEN received the B.S. degree in civil engineering and the Ph.D. degree in traffic information engineering and control from Jilin University, Changchun, China. He is currently an Associate Professor with the School of Transportation, Jilin University. His research interests include traffic control, traffic flow theory, and traffic organization.

XIN WANG received the B.Sc. degree in traffic engineering from Jilin University, Changchun, China, in 2016, where she is currently pursuing the Ph.D. degree in traffic information engineering and control. Her research interests include traffic organization, traffic data analysis, and its applications related approach.

HAITAO LI received the B.Sc. degree in traffic engineering from Jilin University, Changchun, China, in 2016, where he is currently pursuing the Ph.D. degree in traffic information engineering and control. His research interests include pattern recognition, traffic flow prediction, and traffic information forecasting.