IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 10, 2020, accepted August 17, 2020, date of publication August 20, 2020, date of current version September 1, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3018267

A Distributed Assignment Method for Dynamic Traffic Assignment Using Heterogeneous-Adviser Based Multi-Agent Reinforcement Learning

ZHAOTIAN PAN, ZHAOWEI QU, YONGHENG CHEN, HAITAO LI, AND XIN WANG
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 The Dynamic Traffic Assignment (DTA) is one of the important measures to alleviate urban network traffic congestion. The congestions are usually caused by stochastic traffic demands, which are generally unassignable from time dimension in the real-world but are assumed to be assignable in existing DTA methods (i.e. real-time travel demands). In this paper, a distributed DTA method for preventing urban network traffic congestion caused by stochastic real-time travel demands by improving Multi-Agent Reinforcement Learning (MARL). A team structure, which consists of decision-makers and advisers, is designed to learn parallelly in realistic DTA tasks. To reduce the size of the solution space adaptively, the dynamic critical values advised by adviser agents are adopted as constraints for the strategy space of decision-makers (i.e. main agents). A collaborative heterogeneous-adviser mechanism is designed to avoid deviation of guidance. To enhance the adaptability of DTA to the changeable external environment, the mixed strategy concept is introduced to improve the decision-making process of main agents. The respective mapping mechanisms are designed to define adaptive learning rates to improve the sensitivity of MARL. The Sioux Falls (SF) network is established as a test platform via a Dynamic Network Loading (DNL). The effectiveness of the suggested DTA method is assessed through numerical simulations SF network. Under the influence of the scenario with stochastic real-time travel demands, the results show that the proposed method outperforms in terms of the throughput of the network and the individual average travel time among the overall network. Additionally, the ability of the proposed method in response to the external environment rapidly has also been demonstrated. Adopting the suggested method can improve the state of the art to assign stochastic real-time travel demands dynamically and to avoid potential traffic congestion fundamentally.

INDEX TERMS Dynamic traffic assignment, intelligent transportation system, multi-agent system, reinforcement learning, multi-agent reinforcement learning, numerical simulation.

I. INTRODUCTION

Traffic congestion of urban networks and its derivative effects are the problems faced by many cities [1]. The cause of traffic congestion is that existing traffic resources cannot be assigned to meet the rapidly increasing travel demand. Among existing solutions, Traffic Assignment (TA) provides administrations with a macro-perspective to eliminate or mitigate traffic congestion.

Modeling and solution techniques for TA have been extensively studied since Wardrop's Principles (W-Ps) [2]. Existing traffic assignment problems studied by researchers can be classified into two categories: (a) Static Traffic Assignment (STA) [3]; (b) Dynamic Traffic Assignment (DTA) [4].

STA is so macroscopic that it is difficult to cope with the real polytropic situation in Traffic Network (TN). Even if STA had been improved by some scholars considering travel time reliability [5], it still has limitations. Therefore, DTA has gradually gained the favor of many scholars.

The current DTA methods can be divided into several types depending on different perspectives.

According to the basic method types employed, DTA can be divided into two categories: (a) Analysis-based DTA (AB-DTA) [6], [7] (e.g. Mathematical Programming (MP) [10], [11], Variational Inequality (VI) [12], [13], Nonlinear Complementarity Problems (NCP) [14], Differential

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

Complementarity (DC) [16] and Differential Variational Inequality (DVI) [17], [18]); and (b) Simulation-based DTA (SB-DTA) [8], [9] (e.g. link performance functions [15], the point queue models [19], and Cell Transmission Model (CTM) [20] and the Lighthill-Whitham-Richards (LWR) model [21]).).

According to different goals, general DTA with different travel choice models can be roughly divided into three categories: (a) User Optimal DTA (UO-DTA) [24]; (b) Stochastic User Optimal DTA (SUO-DTA) [25]; and (c) System Optimal DTA (SO-DTA) [26]. Sometimes, the first two categories can be grouped into a category which is commonly referred to as User Equilibrium DTA (UE-DTA).

From the perspective of assignment schema, DTA can be classified into three categories: (a) Path-based DTA (PB-DTA) [27]; (b) Link-based DTA (LB-DTA); and (c) Intersection-based DTA (IB-DTA) [1], [28].

However, the above existing DTA methods both assign traffic demand generated in the network to achieve the expected objective (e.g. DUE, DSO, etc.) among the spatial-temporal scale. It is an ideal state that is hardly reached in the real world. In the realistic world, traffic demands are usually stochastic and instant. It means that this kind of demand needs to be dealt with immediately. In other words, these traffic demands are unassignable in the time dimension. Therefore, the DTA system should have the ability to assign traffic demands as they occur. Furthermore, the above DTA methods solve problems of TN from a top-level and centrical perspective, in which a lot of time resources are required for information transmission and solution calculation.

All of the above are challenging for the real-time performance of the DTA system. Hence, the real-time performance must be considered by engineers and scholars for applying DTA in a realistic TN. Therefore, it is necessary to improve the real-time performance of existing DTA methods.

Recently, Multi-Agent Systems (MAS) and Reinforcement Learning (RL) have been integrated and applied to the field of traffic management, such as traffic control [29], [30] and route planning [31], [32]. With the advantages of both MAS and RL, Multi-Agent Reinforcement Learning (MARL) was introduced for TA [33]. The problem can be solved by MARL in a distributed perspective due to that travelers decide which route to adopt and learn from their trips. MARL can provide a distributed structure to deal with the DTA problem. Hence, the portability and universality are guaranteed for applying MARL-based DTA methods to the complex and changeable traffic network. However, TA which was investigated and discussed in [33] is STA. Furthermore, to the authors' knowledge, it is a lack of research on the application of MARL in DTA. Therefore, it is necessary to investigate and modified MARL for DTA with a complex and dynamic environment.

Since the dimension and complexity faced by DTA exceed those of STA, it is necessary to suitably define the space of state and action in MARL for DTA. Furthermore, the equilibrium solution in DTA is time-varying, namely that the convergence point in DTA is time-varying. It requires MARL with capacity that keeps sensitivity to a dynamic environment and prepares to explore unfamiliar knowledge at any time. This capacity can't be established by employing an attenuated learning rate or attenuated greedy search strategy which is generally used in MARL. Therefore, there are significant components of MARL for DTA requiring further study: (a) the space of state and action, which affects the arithmetic speed of MARL; (b) the decision-making strategy, which balances exploration and exploitation; (c) the learning rate, which affects the sensitivity on the aspect of time-varying convergence point.

In this article, we suggest an SB-DTA framework based on concepts and mechanisms inspired by MARL, namely Heterogeneous-Adviser based Multi-Agent Reinforcement Learning (HAB-MARL). This framework is established on a multi-agent architecture with multiple teams. In this architecture, the agents can be divided into two categories: (a) main agents: learning strategy how to select the route for TN and (b) adviser agents (including two sub-types): updating recommended value (i.e. critical time and critical size) of the action space for main agents. For main agents, each agent independently updates its experience without considering the effects caused by other agents. For adviser agents, each agent provides recommendations on the space of action to the corresponding main agent and updates its experience from the external environment. Additionally, agents increase their adaptive capacity for DTA by employing different MARL algorithms. We particularly focus on considering flexible and directional guidance to expedite the convergence of MARL, and on enhancing the MARL's ability to quickly adapt to fluctuations in demand and supply for DTA.

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

(1) Aiming to improve MARL's capacity for DTA, multiple team architecture is established. In this architecture, two agent types are separated according to the learning task. The main agents are responsible for realizing DTA, which receives variable decision support provided by adviser agents.

(2) The variegated critical value from adviser agents is integrated to modify the space restriction which is used to reduce the computational complexity of MARL. The improved MARL framework is capable of adjusting the decision space adaptively, which can capture the direction of the optimal assignment for DTA. The convergence process of MARL can be accelerated.

(3) The mixed strategy concept is introduced to improve the decision-making process of MARL for main agents to update experience parallelly. It makes MARL explore potential solutions efficiently.

(4) Three independent mapping mechanisms are defined to model the learning rate of MARL in corresponding agents. These mechanisms promote MARL's self-adaptive ability facing with a new state, in which occurs quantitative transitions of demand and supply in TN. In other words, the sensitivity of MARL on the aspect of a dynamic external environment is enhanced.

The rest of this article is organized as follows. The related works on the aspect of DTA and MARL are introduced in Section II. The technical background and definition of the HAB-MARL method are given in Sections III and IV, respectively. Numerical simulation experiments and analyzation of results are exhibited in Section V. Finally, conclusions and future expectations of this work are given in Section VI.

II. RELATED WORKS

A. DYNAMIC TRAFFIC ASSIGNMENT (DTA)

The fundamental rules of DTA are established on W-Ps [2], which consists of two optimization principles: (a) Wardrop's first principle, namely User Equilibrium (UE); and (b) Wardrop's second principle, namely System Optimal (SO). Additionally, there are also some important components in DTA, which include travel cost function, route choice, etc.

Recently, the attention of scholars has been centered around the above components of DTA. For problem constraints and generalization, various methods were suggested for DTA solution, such as MP [10], [11], VI [12], [13], NCP [14], [15], etc.

On the aspect of algorithms, early studies on DTA were developed on the basic technology of STA, such as the application of the Frank-Wolfe (FW) [34]–[37]. To avoid the defect of FW in terms of convergence rate, the Column Generation (CG) [38] and Simplicial Decomposition (SD) [39] had been widely employed as alternative approaches. With the generalization of DTA, the Sub-gradient (SG) was adopted [40]. The above studies can be regarded as AB-DTA.

Although the uniqueness and existence of solutions are guaranteed and determined in advance, the limitation that traffic dynamics are difficult to be captured still restricts the application of the analytic approach. With the capacity of modeling dynamic characteristics of traffic flow, SB-DTA was presented to model DTA in a realistic environment.

In SB-DTA, scholars focus on what kind of traffic flow models can be select to reflect the real-word. This model is the basis for Dynamic Network Loading (DNL), which is a fundamental module structuring the architecture of DTA. Kuwahara et al [41] extended the point queue model to deal with physical queues and then proposed a corresponding solution algorithm for UO-DTA. The form widely used in SB-DTA is CTM. Lo et al solved UO-DTA by employing alternating direction method [42] or transforming NCP to an equivalent MP [43] on the CTM. Szeto and Lo [22] developed a CTM framework to solve route and departure time choice simultaneously under flexible demands scenarios for UO-DTA. Szeto and Lo [24] modeled SUO-DTA as a Fixed Point (FP) problem and employed CTM to capture the effect of the random evolution of traffic states.

In addition to being classified as AB-DTA and SB-DTA, DTA can also be divided according to the inflow form in which decision variables are assigned, namely PB-DTA, LB-DTA, and IB-DTA. In PB-DTA, the divergent or merged flow can be modeled due to its abundant information on path flow. Therefore, PB-DTA was studied in various studies, such as Chen and Feng [12], Lim and Heydecker [25], and Meng *et al.* [44]. Acquiring abundant information of PB-DTA relies on the enumeration of feasible paths, in which enumerative complexity increases exponentially with the growth of OD pairs and TN scale. Two approaches were usually to avoid or mitigate this disadvantage. One is to reduce the enumerative complexity of paths and accelerate computational speed via embedding the path generation algorithm in PB-DTA. It expands a sub-research in the field of DTA [45]. The other completely abnegates the requirement of path information to avoid enumeration fundamentally, which also known as LB-DTA. The studies in [14] and [15] previously mentioned can be regarded as researches on LB-DTA. To integrate the advantages of PB-DTA and LB-DTA, Long *et al.* proposed [28] and further generalized [1] IB-DTA. It transforms IB-DTA to FP that seeks the stabilized flow proportion at diverges and merges in TN.

However, as discussed in Section I, all traffic demands are counted and assigned among space-time scale in assumptions of the above DTA methods. This kind of assignment process is top-level and centralized actually. The corresponding algorithm finding a stable equilibrium point repeatedly. It is difficult and unrealistic to searching an evolutive equilibrium point caused by stochastic and real-time travel demands. Therefore, it is necessary to develop an effective approach that is self-adaptive to time-varying equilibrium immediately.

B. MULTI-AGENT REINFORCEMENT LEARNING (MARL)

In recent years, Artificial Intelligence (AI) technology has been greatly developed [46]–[48]. As a sub-field of AI, MAS has excellent portability. Therefore, MAS has been widely adopted to model systems with multi-participators, which are widespread in numerous fields including communication protocol [49], cooperative control [50], [51], fault-tolerant control [52], electrical power system [53], sensor deployment [54], transportation [55]–[57], etc. MAS researchers focus on structural mechanisms of complex systems. These mechanisms include internal learning mechanisms of each agent and externally interactive mechanisms among agents to other agents and the environment. For instance, MAS was employed to establish a real-world traffic simulation [58], [59]. With the development of Machine Learning (ML), RL with the advantage of model-free and self-learning has been proposed and exploited. MARL combines RL with MAS and promotes the development of theory and technology in transportation.

In transportation, MARL is extensively applied for many tasks, such as traffic control and route choice [33]. However, only the route choice in these tasks is associated with DTA weakly.

Bazzan and Grunitzki [60] considered DTA with individual drivers as independent and autonomous agents, which employ QL to stepwise select suitable routes. According to the objective function of DTA, Grunitzki *et al.* [61] proposed two improved algorithms based on QL to maximize the utility

of agent and system respectively. Bazzan, and Klugl [59] believed that the improvement and deterioration of the whole system are affected by the combinations of actions among agents. Zolfpour-Arokhlo *et al.* [32] combined MARL with Q-value based Dynamic Programming (QVDP) to provide a priority route plan for vehicles. MARL in the above literature focuses on the definition of the reward function. By defining rational and precise reward functions, authors can achieve their objectives, such as minimizing individual travel time, maximizing system utility, etc.

Grunitzki *et al.* [33] suggested two MARL methods for TA and compared them with three classic TA methods. The concept of restricted search space was integrated into MARL aiming to expedite convergence. It is proven that the convergent process can be accelerated by reducing the action space felicitously. Nevertheless, the size of restricted action space in [33] is constant, which was obtained by the experiment. For DTA, in the actual situation, different drivers decide to travel from their origin to destination with distinguished spatial awareness. Moreover, the set of potential optimal routes is variational along with time due to that the situation in TN is time-dependent. Restricting the set of recommended paths adaptively can model the above situation approximatively.

As mentioned earlier, the equilibrium in DTA is evolutionary. Solving DTA is essentially an online closed-loop optimizing process. It requires that MARL has a high convergence rate and the ability to adjust the learning process adaptively. With respect to the state of the art, the motivated mechanism [62] and Equilibrium Transfer (ET) [63] are proposed to accelerate convergence. The motivated mechanism is introduced to supplement the reward function to accelerate MARL convergence. This method was defined as Motivated Reinforcement Learning (MRL). Nevertheless, MRL may lead to a deviation from the target due to that the reward function connects with the objective of DTA closely. Therefore, defining the reward function reasonably is more suitable for DTA. It has been shown in the above literature [32], [60], [61]. As for ET, its effectiveness is based on similar environments or historical experiences suffered by agents. This basic scenario is unfamiliar in DTA leading to poor applicability of ET. Additionally, the traditional assignment-loading reciprocating pattern of DTA impedes the further increase of computing speed. According to the framework of MARL, the feasible approach to accelerate MARL convergence is the assignment-loading parallel pattern, in which the assignment is divided into multi-unit as time goes on and executed synchronously along with loading. Moreover, the decision-making mechanism of MARL can be adjusted appropriately to improve the efficiency of exploration to accelerate MARL convergence.

In addition to the above discussion, the sensitivity of MARL is important to DTA. The time-dependent equilibrium in DTA requires MARL with the capacity of recognizing dynamic state and updating knowledge as soon as possible. Hence, frequently-used the attenuation learning rate, which lacks re-learning ability after convergence, is inapposite for MARL in DTA. It is necessary to explore the mapping mechanism which can self-adaptively regulate the learning rate according to the state of MAS and TN.

In conclusion, it is difficult to apply and popularize MARL in real-world DTA under the existing technological environment. To address the above difficulties singly, the present work proposes a Heterogeneous-Adviser based Multi-Agent Reinforcement Learning (HAB-MARL) architecture to deal with the DTA.

III. BASIC BACKGROUND TECHNOLOGY OF HETEROGENEOUS-ADVISER BASED MULTI-AGENT REINFORCEMENT LEARNING (HAB-MARL)

In this section, we introduce basic background technologies of Heterogeneous-Adviser based Multi-Agent Reinforcement Learning (HAB-MARL) architecture for DTA. In Section III.A, DTA will first be introduced as the application background of HAB-MARL. Then, a CTM-based DNL will subsequently be elucidated to establish numerical simulation in Section III.B. Finally, the common form and related concepts of MARL will be given in Section III.C. Additionally, all parameters relating to DTA, DNL, and MARL are listed in TABLE 1-3 respectively.

A. DYNAMIC TRAFFIC ASSIGNMENT (DTA) PROBLEM

In this section, we will first introduce some basic components of TA and briefly describe a Variational Inequality (VI), which is the familiar generalized solution form for DTA.

Definition 1: A traffic network G is a tuple $\langle V, E\rangle$. The elements of the tuple are described in TABLE 1.

In $G(V, E)$, flow is the control subject of DTA aiming to minimize travel costs. It can be defined in the form of VI.

Definition 2: For specific flow f_b^t, the corresponding travel cost function is $c_b\left(t, f_b^t\right)$. For DTA, the following generalized discrete VI exists.

$$\sum_{t\in T}\sum_{b\in B} c_b\left(t, f_b^*\right)\left(f_b^t - f_b^*\right) \geq 0, \quad \forall f_b^t \in \Omega_b \tag{1}$$

In formula (1), subscript b represents one element of $\{l, p\}$. If b refers to l, flows are adjusted in links of TN, in where DTA can be regard as LB-DTA. On the contrary, PB-DTA regulates flows on paths when b signifies p.

For travel cost function c_b, there are two frequently-used modalities: travel time and marginal travel time, which severally express implicit objectives of UE-DTA and SO-DTA. Additionally, it has more components in the travel cost function, such as pricing tactics, pollutant emission, capacity constraints, congestion delay, etc. This moment, c_b is usually referred to as the generalized cost function.

Moreover, the existence and uniqueness of the solution in VI for DTA have been proven in some cases.

Theorem 3: Formula (1) has a unique solution if and only if the travel cost function satisfies some specific mathematical conditions: continuity and monotonicity.

In **Theorem 3**, continuity and monotonicity be emphasized to guarantee that the equivalent mathematical programming

TABLE 1. Parameters relating to DTA.

Notations	Signification
G	The graph of traffic network (TN)
V	The set of nodes in TN
E	The set of edges in TN
$\boldsymbol{f}$	The vector of all flows assigned among TN in DTA
f_b^t	The flow of b at t
f_b^*	The optimal flow of b
T	The total simulation time
t	One moment in simulation
$c_b\left(t, f_b^t\right)$	The travel cost of b facing with f_b^t at t
B	The set of links and paths, $B=\{E,P\}$
b	The one element of $\{l,p\}$
l	The link, $l \in E$
P	The set of all paths in TN
p	The path, $p \in P$
Ω_b	Banach space of b: $\Omega_b=\{\Omega_l,\Omega_p\}$
	$\Omega_l=\{\boldsymbol{f}_l \mid \Lambda \boldsymbol{f}_p=\boldsymbol{q}, \Lambda \boldsymbol{f}_l=\boldsymbol{f}_p, \boldsymbol{f}_l \geq \boldsymbol{0}\}$
	$\Omega_p=\{\boldsymbol{f}_p \mid \Lambda \boldsymbol{f}_p=\boldsymbol{q}, \boldsymbol{f}_p \geq \boldsymbol{0}\}$

TABLE 2. Parameters relating to DNL.

Notations	Signification
t	One moment in simulation
l	The link, $l \in E$
E_{non}	The set of edges inside TN
E_{org}	The set of edges accessing origins in TN
E_{des}	The set of edges accessing destinations in TN
l_i^+	The link i in in the downstream of l, $l_i^+ \in \Gamma^+(l)$
l_i^-	The link i in in the upstream of l, $l_i^- \in \Gamma^-(l)$
x_l^t	The number of vehicles on link l at t
$y_{l_i^-,l}^t$	The transfer number of vehicles form l_i^- to l at t
u_l^t	The number of vehicles entering into link l at t
v_l^t	The number of vehicles exiting from link l at t
DE_l^t	The demands generated in link l at t
$SU_{l_i^+}^t$	The supplies provided by link l at t
ρ_l^t	The density of link l at t
$\vartheta_l^{de}\left(\rho_l^t\right)$	The function of demand with ρ_l^t
$\vartheta_l^{su}\left(\rho_l^t\right)$	The function of supply with ρ_l^t
χ_l^t	The capacity of link l
N_l^t	The total number of vehicles, which can exist in link l at t

function of VI holds strictly convex. The existence of the optimal solution in the finite solution space is further guaranteed. A similar process of proof has been fully discussed in [65].

B. DYNAMIC NETWORK LOADING (DNL) FOR DTA

Among numerous methods of DNL, a method based on CTM with the capacity of capturing traffic dynamics superiorly has been widely studied, which has various forms, such as [20], [66], [67]. For more details please refer to [20], [64] and [68]. In this section, a CTM-based approach is introduced to structure DNL for DTA.

In CTM-based DNL, the renewal process of traffic flow on a link of TN subjects to the following generalized regulations:

(Ordinary edge $l \in E_{non}$)

$$x_l^t = x_l^{t-1} + \sum_{l_i^- \in \Gamma^-(l)} y_{l_i^-,l}^{t-1} - \sum_{l_i^+ \in \Gamma^+(l)} y_{l,l_i^+}^{t-1} \tag{2}$$

(Origin edge $l \in E_{org}$)

$$x_l^t = x_l^{t-1} + u_l^t - \sum_{l_i^+ \in \Gamma^+(l)} y_{l,l_i^+}^{t-1} \tag{3}$$

(Destination edge $l \in E_{des}$)

$$x_l^t = x_l^{t-1} + \sum_{l_i^- \in \Gamma^-(l)} y_{l_i^-,l}^{t-1} - v_l^t \tag{4}$$

(Demand and supply of flow transmission)

$$DE_l^t = \min\left\{\vartheta_l^{de}\left(\rho_l^t\right), x_l^t, \chi_l^t\right\} \tag{5}$$

$$SU_{l_i^+}^t = \min\left\{\vartheta_{l_i^+}^{su}\left(\rho_{l_i^+}^t\right), \delta\left[N_{l_i^+}^t - x_{l_i^+}^t\right], \chi_{l_i^+}^t\right\} \tag{6}$$

(Connectors $y_{l,l_i^+}^t, l, l_i^+ \in E$)

$$y_{l,l_i^+}^t \leq \min\left\{DE_l^t, SU_{l_i^+}^t\right\} \tag{7}$$

$$s.t.$$

$$\sum_{l_i^- \in \Gamma^-(l)} y_{l_i^-,l}^t \leq SU_l^t \tag{8}$$

$$\sum_{l_i^+ \in \Gamma^+(l)} y_{l,l_i^+}^t \leq DE_l^t \tag{9}$$

The parameters in formulas (2)-(9) are described in TABLE 2. The formulas (2)-(4) demonstrate the discrete flow variation in link transmission. The potential demand and supply of links are represented by formulas (5) and (6) respectively. In formulas (5) and (6), three components on the right side represent density flow, current flow/allowance, and link capacity in turn. There is a unity of multiple possible link scenarios, such as the entrance of the network, and the exit of the network. The inequation (7) shows two states faced by connectors: ordinary state and branch state. In the ordinary state, inequation can be transformed into equality. For branch state, formula (7) expresses the possible scheme in connectors. The further strict constraints are sequentially given in formulas (8) and (9) for connectors at diverging and merging scenes.

With the advantages of simple structure and linearization, the above CTM-based DNL can be easily established for DTA in numerical simulation.

TABLE 3. Parameters relating to MARL.

Notations	Signification
i	An agent of MAS
k	The number of iteration step
Θ	The set of agents in QL
S	The state space set, $S=\times_i S_i$
S_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
$A_i(s_i)$	The action space of agent i facing with s_i
R	The reward function of agents, $R: S\times A\to R$
Q_i	The Q function of agent i
γ	The discount factor, $\gamma\in[0,1)$
α	The learning rate, $\alpha\in[0,1)$
s_i	The state confronted by agent i, $s_i\in S_i$
s_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
r_i	The reward function of agent i
$Q_i^k(s_i,a_i)$	The Q value of agent i executing a_i under s_i after iterating k times
$Q_i^*(s_i,a_i)$	The convergent Q value of agent i executing a_i under s_i
$\Pi(s^-,s^+,t_{critical})$	The whole action space of agent i with $t_{critical}$ among the transform from s^- to s^+
Π_{valid}	The valid action space of agent i with $t_{critical}$
$\Pi_{invalid}$	The invalid action space of agent i with $t_{critical}$
$\lvert\Pi_{precomputed}\rvert$	The size of restrained action space of main agent i precomputed in advance

C. MULTI-AGENT REINFORCEMENT LEARNING (MARL)

A classical model-free MARL method widely applied is Q-Learning (QL). In QL, the value functions are learned and stored in the form of Q factor.

Definition 4: For Q-learning, exiting a tuple $\langle\Theta, S, A, R\rangle$, the function updating process of Q-values subjects to the following equation:

$$Q_i^{k+1}(s_i,a_i) = (1-\alpha)\,Q_i^k(s_i,a_i) + \alpha\left[r_i^k(s_i,a_i)+\gamma\max_{a_i'\in A_i(s_i')} Q_i^k(s_i',a_i')\right] \quad (10)$$

The parameters in **Definition 4** have been described in TABLE 3. Equation (10) abstract the process in which agents acquire knowledge via selecting actions, interacting with the external environment.

Lemma 5 (Convergence): Exiting prerequisite that all states and actions have been visited infinitely and the learning rate is within definition, the equation (10) converges to $Q_i^*(s_i,a_i)$.

For more specific detail about proof of Lemma 5 see [69].

However, the complexity of applying the MARL method represented by QL in TA increases with the scale of TN. To deal with this limitation, the concept of critical time was introduced in [33] to divide the strategy space into two subsets: valid strategy space and invalid strategy space. The relationship of subsets is as follows:

$$\Pi_{valid}\cap\Pi_{invalid} = \varnothing \quad (11)$$

$$\Pi(s^-,s^+,t_{critical}) = \Pi_{valid}\cup\Pi_{invalid} \quad (12)$$

The parameters in formulas (11)-(12) have been given in TABLE 3. The (11) indicates that elements in the two subsets are totally different. The (11) illustrates that the combination of two subsets can represent the whole strategy space including all possible actions for TA. This improved method of QL has been named as Edge-based QL (EB-QL).

Additionally, another improved method named Route-based QL (RB-QL) was also presented in [33]. In RB-QL, agents receive precomputed recommendatory strategy space before the learning process begins, namely $A_i(s^-)\leftarrow\Pi_{precomputed}$. In contrast with the feasible strategy space, the recommendatory strategy space appears the enumerable feature due to that $\lvert\Pi_{precomputed}\rvert$ is small.

Both EB-QL and RB-QL essentially reduce MARL's complexity by cutting down the size of the strategy space. The above $t_{critical}$ and $\lvert\Pi_{precomputed}\rvert$ were configured as constant via mensurable experiment. The selecting of a critical time in EB-QL for agents is important. Smaller values lead to the loss of potential optimal solution, in which traffic flow only can be assigned into minority paths. On this account, exorbitant travel cost occurs on overloaded links. On the contrary, there are too many low-efficient strategies merged into a valid strategy set leading to the performance loss of learning when higher values are adopted. For RB-QL, the similar discussion about the size of $\lvert\Pi_{precomputed}\rvert$ had been presented in [33]. It is unsuitable for applying either EB-QL or RB-QL in DTA due to that $t_{critical}$ and $\lvert\Pi_{precomputed}\rvert$ are time-dependent with DTA evolution in TN. For the application of MARL in DTA, dynamic characters of DTA, such as time-vary demands, equilibrium transformation, the variation of travel cost, etc., must be considered.

IV. FRAMEWORK OF HETEROGENEOUS-ADVISER BASED MULTI-AGENT REINFORCEMENT LEARNING (HAB-MARL)

In this section, we establish Heterogeneous-Adviser based Multi-Agent Reinforcement Learning (HAB-MARL) architecture for DTA. The complete formulation of HAB-MARL will be described in detail in Section IV.A. The corresponding algorithm will be subsequently summarized in Section IV.B. All parameters relating to HAB-MARL are listed in TABLE 4.

A. HETEROGENEOUS-ADVISER BASED MULTI-AGENT REINFORCEMENT LEARNING (HAB-MARL)

In this section, MARL (described in Section III.C) theory is improved to design a distributed assignment method for DTA.

TABLE 4. Parameters relating to HAB-MARL.

Notations	Signification
G	The graph of traffic network (TN)
N	The set of nodes in TN
E	The set of edges in TN
i	An agent of MAS
k	The number of iteration step
f_b	The flow of b
$\Delta t_{critical}^{\max}$	Maximum variation value for adjusting recommendatory critical time
$t_i^{critical,k}$	Critical time received by main agent i at simulation episode k
$\Delta t_i^{critical,k}$	Corrected value of critical time provided by time adviser agent i at simulation episode k
c_b	The travel cost of b
$\Delta \overline{c}_i^{improve}$	The travel cost improved value of adviser agent i
$\Delta c_{\dagger_j}^{base}$	The basic travel cost improved value of main agent i for the j component of mixed strategy σ_i
$\Delta c_{\dagger_j}^{slack}$	The slack travel cost improved value of main agent i for the j component of mixed strategy σ_i
$\overline{c}_b^k$	The average travel cost of b at simulation episode k
$\overline{c}_i$	The average travel cost of main agent i for corresponding OD pair i
b	The one element of $\{l, p\}$
$b^{\dagger_j}$	The links or paths among OD pair i belong to the j component of mixed strategy σ_i
$b^{\bar{\dagger}}$	The links or paths among OD pair i not belong to mixed strategy σ_i
l	The link, $l \in E$
x_l	The number of vehicles on link l
δ	The reserve factor of surplus interspace in link l
S	The state space set, $S = \times_i S_i$
S_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
$A_i^{\dagger}(s_i)$	The restrained action space of main agent i
$A_i^{\dagger_{time}}(s_i)$	The restrained action space of main agent i, which refers to advise around critical time
$A_i^{\dagger_{size}}(s_i)$	The restrained action space of main agent i, which refers to advise around critical size
A_m	The action space of main agents
A_{at}	The action space of time adviser agents
A_{as}	The action space of time adviser agents
γ	The discount factor, $\gamma \in [0,1)$
α	The learning rate, $\alpha \in [0,1)$
ε	The coefficient of greed
s_i	The state confronted by agent i, $s_i \in S_i$
s_i'	The future state confronted by agent i
σ_i	The mixed strategy of main agent i
a_i	The action executed by agent i, $a_i \in A_i$
a_i'	The future action executed by agent i
$a_i^{\dagger_j}$	The j component of mixed strategy σ_i
r_i	The reward function of agent i

TABLE 4. *(Continued.)* Parameters relating to HAB-MARL.

$Q_i^k(s_i, a_i)$	The Q value of agent i executing a_i under s_i after iterating k times
$Q_i^*(s_i, a_i)$	The convergent Q value of agent i executing a_i under s_i
$\Delta\lvert\Pi\rvert_{critical}^{\max}$	Maximum variation value for adjusting recommendatory critical size
$\lvert\Pi\rvert_i^{critical,k}$	Critical size received by main agent i at simulation episode k
$\lvert\Pi\rvert_i^k$	Critical size executed by main agent i at simulation episode k
ζ_p	Logical value: Determines whether the path conforms to the optimization expectation or not
$\Delta\xi_i^{critical}$	The redundancy for critical size advised by size adviser agent i
$\varpi_i^{\dagger_j}$	The slack reword weight for mixed strategy σ_i
I	The set of agents in MARL
I_m	The set of main agents in HAB-MARL
I_{at}	The set of time adviser agents in HAB-MARL
I_{as}	The set of size adviser agents in HAB-MARL

We suggest a decentralized framework, where two categories of agents are associated with Origin-Destination (OD) pairs and strategy space support severally. In this architecture, there is a main agent group in which members decide how to assign corresponding flow into TN independently. The members of the other agent group are responsible for assisting the corresponding main agents. Communication occurs only among the auxiliary agents (advisers) and the main agents (deciders) for the transfer of support information. FIGURE 1 illustrates the suggested heterogeneous-adviser based multi-agent reinforcement learning (HAB-MARL) distributed assignment framework for DTA.

In FIGURE 1, it has four modules (i.e. A-D). Module A is the kernel of HAB-MARL including the external interactions and internal learning mechanisms among agents playing different roles in HAB-MARL. Due to that architecture of HAB-MARL is multi-group and decentralized, only one agent team for one OD pair in DTA has been provided as a basic standard team structure in module A. For DTA, this basic standard team structure of HAB-MARL can be replicated according to the size of the network. The details of module A will be described in the rest of this section. Module B, which has been introduced in Section III.B, can be employed to model traffic flow in the real world. In this paper, the exchange between modules A and B can be simply regarded as interaction with the external environment for HAB-MARL. It has two layers in module C, which represent the realistic and mathematical structure of road networks respectively. Module C is the indispensable foundation to establish module B. Module D contains the information about demand of network. In DTA, these demands require the services of module A, which are usually varied and can be defined according to scenarios (such as defined in Section V.A.3). In addition, the flow

of information of HAB-MARL in DTA tasks has also been abstracted among modules A-D.

In the following sub-sections, the various components of the HAB-MARL architecture shown in FIGURE 1 (A) are defined. To simplify pertinent expressions, the label of iteration series is omitted. Moreover, the differentiated agent groups in MAS have their specific Q-functions severally. It is necessary to define the components of the Q function for each agent in different groups.

1) ADVISER AGENT

For adviser agents, their interaction process, in which actions affect the external environment indirectly, is slightly different from that of main agents. Furthermore, the reward is perceived by the main agents from the external environment. However, to keep cognitive consistency about the outside world, the state of adviser agents must be consistent with that of the main agents.

a: TASK

The task of adviser agents is that critical-value is learned to restrict relevant action set provided for main agents. In this paper, the critical-value contains two forms: expectant critical time and expectant size of strategy set.

b: STATE SPACE

As mentioned, for both adviser agents and main agents, the state is perceived from the external environment. In DTA, the state is employed to describe the scene in TN. Additionally, the evolution of traffic in DTA is modeled in this paper as a network of density (see Section III.B). The state of each agent is the vector of states at all links (edges) in TN. Besides, it is worth noting that rationalize the size of state space to reduce computational complexity [30]. Fortunately, the state of TN only needs to be detected once to be accessible by all agents. Inspired by the concept of distributed expression in Deep Learning (DL) [70], we formally define the state of TN as follows.

Definition 6: For a traffic network $G(N, E)$, the state for each agent can be expressed as a vector (distributed expression): $\boldsymbol{s} = [s_1, \cdots, s_l, \cdots, s_{n_e}]$. The element of $\boldsymbol{s}$ represents the state of edge in TN, which can be defined by integrating formula (13) and formula (14). The related parameters are described in TABLE 4.

$$\psi_l = \frac{x_l}{x_l^{\max}} \tag{13}$$

$$s_l = \begin{cases} free, & \psi_l < \varphi_{free} \\ resistance, & \varphi_{free} \le \psi_l < \varphi_{jam} \\ jam, & \psi_l \ge \varphi_{jam} \end{cases} \tag{14}$$

In formula (13), ψ_l is evaluation index composed of vehicle proportions on link l. It is further defined by formula (14). In this paper, 0.5 and 0.8 are adopted as the values of φ_{free} and φ_{jam} respectively.

c: ACTION SPACE AND SELECTION

To reduce the complexity, we construct the action of each adviser agent using a similar form adopted in modeling state space.

Definition 7: The action of adviser agent i is $\boldsymbol{a}_i = [a_i^{\pm}, a_i^{value}]$. $a_i^{\pm}$ used to determine operation direction is an element of $\{-1, 1\}$. The operation amplitude can be defined by a_i^{value}, and $a_i^{value} \in [0, \kappa_i^{\max}]$.

In **Definition 7**, $\kappa_i^{\max}$ is one of $\left\{\Delta t_{critical}^{\max}, \Delta\,|\Pi|_{critical}^{\max}\right\}$ according that the task of agent i belongs to different categories (see Section IV.A.1.(a)).

To balance exploration (gaining of knowledge) and exploitation (usage of knowledge), adviser agents select their action according to ε greedy widely adopted in QL.

d: REWARD FUNCTION

The reward function is used to represent the effects of agent action, which is also used to establish a closed feedback loop. It needs to be defined according to the objective of agents. For adviser agents, the reward function is employed to assess the effect of recommendation.

Definition 8: The reward function of adviser agent i, $r_i(s_i, a_i)$ is a function of the difference in cumulative improvement of individual travel cost.

$$r_i(s_i, a_i) = \frac{1 + \exp^{\left(-\eta \cdot \log\left(\left|\Delta \bar{c}_i^{improve}\right| + 1\right)\right)}}{1 - \exp^{\left(-\eta \cdot \log\left(\left|\Delta \bar{c}_i^{improve}\right| + 1\right)\right)}} \tag{15}$$

Formula (15) is defined to ensure the range of the reward function among $[-1, 1]$ to avoid the increase of convergence time caused by sharp fluctuation of Q value. The formula (15) can be further defined by formulas (16)-(18) as follows:

$$\Delta \bar{c}_i^{improve} = \bar{c}_b^{k-1} - \bar{c}_b^{k} \tag{16}$$

$$\bar{c}_b^k = \frac{\sum\limits_{b \in B_i^k} c_b^k f_b^k}{\sum\limits_{b \in B_i^k} f_b^k} \tag{17}$$

$$\eta_i^{improve} = \begin{cases} 1, & \Delta \bar{c}_i^{improve} \ge 0 \\ -1, & \Delta \bar{c}_i^{improve} < 0 \end{cases} \tag{18}$$

Formula (16) represents the change of average travel cost on link/path b. The average travel cost of b is further defined in formula (17). $\eta_i^{improve}$ defined in formula (18) is a coefficient to ensure the plus or minus characteristic of formula (15). The parameters in formulas (15)-(18) have been described in TABLE 4.

e: OTHER RELEVANT PARAMETERS

The discount rate in QL reflects the preference of agents on long-and-short term reward. For adviser agents, they hope that the advice received by matching the main agent is perfect every time. Therefore, discount rates of adviser agents are set as 0.9 uniformly.

The learning rate affects the cumulative effect of knowledge. It determines the degree of retention and replacement

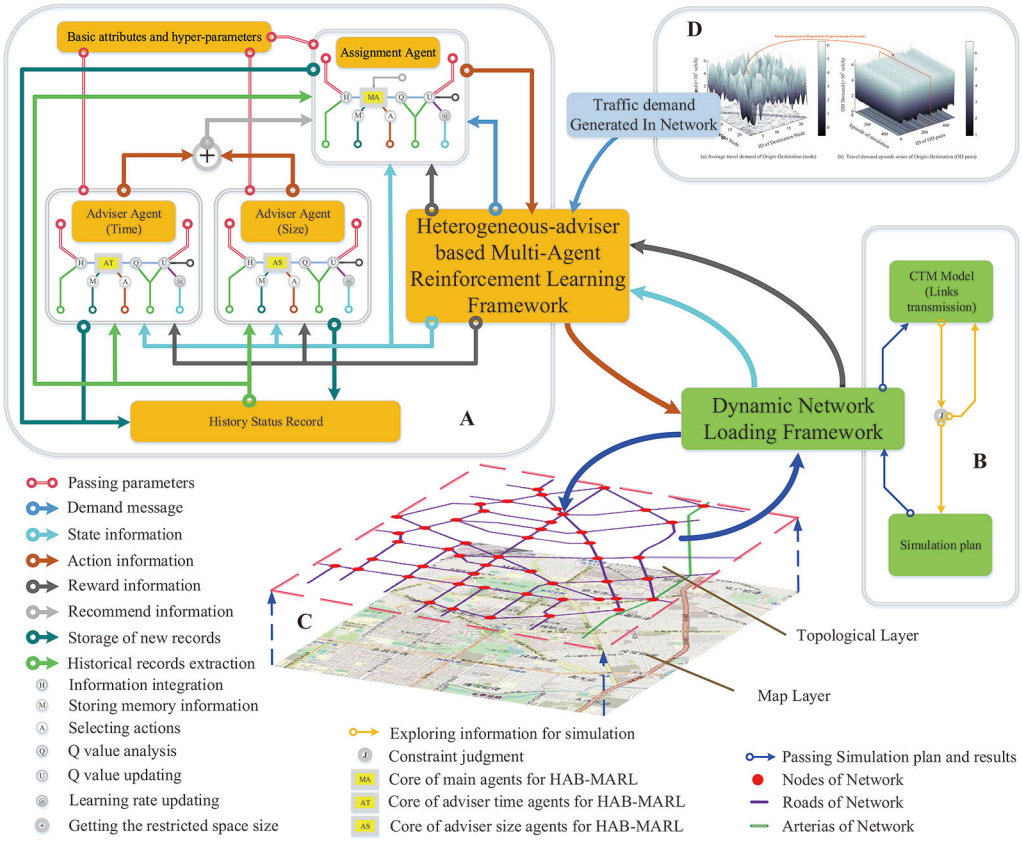


FIGURE 1. Diagrammatic drawing of HAB-MARL framework for DTA.

of historical experience. Additionally, considering the application in DTA and role in HAB-MARL, the learning rate of adviser agents needs to be adaptive to dynamic goals. Therefore, self-adaptive forms of learning rate should be defined for two adviser agent types according to differentiated task contents.

For critical time adviser agents, the best critical time should be the travel time corresponding to the optimal solution of DTA. However, in the process of moving towards the DTA optimal solution, the specific value of the optimal solution is unknown and uncertain. Hence, the critical value should be determined according to the current assignment effect, namely the corresponding actual travel time.

Definition 9: The learning rate of critical time adviser agent i, α_i can be expressed as a function of the difference between critical time and the actual average travel time:

$$\alpha_i = \frac{1 + \exp^{-\Delta \upsilon_i^{critical}}}{1 - \exp^{-\Delta \upsilon_i^{critical}}} \tag{19}$$

Formula (19) can be further expanded by introducing formulas (20) and (21).

$$\Delta \upsilon_i^{critical} = \left| t_{i,k}^{critical} - \bar{t}_i^k \right| \tag{20}$$

$$t_i^{critical,k} = t_i^{critical,k-1} + \Delta t_i^{critical,k} \tag{21}$$

For formula (19), the nonnegativity of its input value is guaranteed by formula (20). Furthermore, the value of formula (20) is far less than infinity. The above are the key elements to ensure that the learning rate can controlled within a reasonable range [0, 1).

Formula (20) represents the difference between the recommended value $t_{i,k}^{critical}$ and the desired value $\bar{t}_i^k$. The recommended value $t_{i,k}^{critical}$ is further defined in formula (21).

For strategy set size adviser agents, the homologous optimum expectant size of strategy set should be the number of paths adopted in DTA at the equilibrium point. Nevertheless, the traversal of paths is dynamic with tracking the time-varying equilibrium point. Under this circumstance, it is a feasible method to estimate the optimum expectant size by assessing the current path traversal situation. From the above, the learning rate of strategy set size adviser agents can be defined as follows.

Definition 10: The learning rate of strategy set size adviser agent i, α_i can be expressed as a function of the difference between expectant size and the estimated optimal size:

$$\alpha_i = \frac{1 + \exp^{-\log(\Delta \xi_i^{critical} + 1)}}{1 - \exp^{-\log(\Delta \xi_i^{critical} + 1)}} \tag{22}$$

Formula (22) can be further expanded by introducing formulas (23) and (24).

$$\Delta\xi_i^{critical} = \left| |\Pi|_i^{critical,k} - \left(|\Pi|_i^k - \sum_{p\in P_i^k} \zeta_p \right) \right| \tag{23}$$

$$\zeta_p = \begin{cases} 1, & p \in \Xi \\ 0, & p \notin \Xi \end{cases} \tag{24}$$

Formula (24) returns logic value of necessity judgment about paths traversal. The logic value can be set to 1 when the current traversal path p is inefficient (i.e. $p \in \Xi$). The set Ξ is a specific space that sub-set Ξ_i meets specified requirements:

$$\Xi_i = \Big\{ \tilde{p} | c_{\tilde{p}} - \min\left(c_p\right) \geq \max\left(c_l\right), \\ l \in \tilde{p}, l \notin \underset{\sim}{p}, \tilde{p} \in P_i^k, \underset{\sim}{p} \in \underset{\sim}{P}_i^k, P_i^k = \underset{\sim}{P}_i^k \cup \tilde{p} \Big\} \tag{25}$$

For adviser agent i, the corresponding set of inefficient paths is given in Ξ_i. Moreover, it is similar as formula (19) that formula (22) is constructed to hold the non-negative input value for formula (23). The range of α_i derived from formula (22) can be restrained in [0, 1).

The difference between the recommended value $|\Pi|_i^{critical,k}$ and the ideal value $\left(|\Pi|_i^k - \sum_{p\in P_i^k} \zeta_p \right)$ is given in formula (23). The ideal value $\left(|\Pi|_i^k - \sum_{p\in P_i^k} \zeta_p \right)$ is the number of effective path, which is executed in the current assignment. It is defined as difference value of the number of paths used actually $|\Pi|_i^k$ and the number of invalid paths $\sum_{p\in P_i^k} \zeta_p$.

In these ways, the sensitivity of adviser agents to dynamic differentiated targets is also preserved while the convergence is guaranteed. The parameters in this sub-section (i.e. **Definition 9**, **10,** and their relevant complementary formulas) can be seen in TABLE 4.

2) MAIN AGENT

a: TASK

For application in DTA, main agents are responsible for assign traffic flow on TN synchronously to avoid the emergence of traffic congestion in the local region. The process of assignment obeys W-Ps (see Section II.A).

b: STATE SPACE

The state space of the main agents is the same as that of adviser agents. The reason has been discussed in Section III.D.1.(b).

c: ACTION SPACE AND SELECTION

The action space of main agent i, $A_i^\dagger(s_i)$ is a sub-set of unabridged strategy space $A_i(s_i)$. It is acquired taking suggestions provided by corresponding adviser agents into account synthetically.

Definition 11: For main agent i, its action space $A_i^\dagger(s_i)$ is the intersection of two recommended strategy spaces: $A_i^\dagger(s_i) = A_i^{\dagger time}(s_i) \cap A_i^{\dagger size}(s_i)$.

In **Definition 11**, the recommended strategy spaces $A_i^{\dagger time}(s_i)$ and $A_i^{\dagger size}(s_i)$ are extracted from $A_i(s_i)$ on the basis of $t^{critical}$ and $|\Pi|_{critical}$ derived from advise agent respectively.

To chase the objective of DTA, it is necessary to select a proper way to decide how to carry out actions for main agents.

Definition 12: The action of main agent i is executed in a mixed strategy manner, namely σ_i, which is a probability distribution on $A_i^\dagger(s_i)$ computed based on partial history experience $\left\{ Q_i\left(s_i, a_i^{\dagger_j}\right) \right\}, a_i^{\dagger_j} \in A_i^\dagger(s_i)$.

In **Definition 12**, the mixed strategy manner σ_i can also be treated as a convex combination of all feasible $a_i^{\dagger_j}$, $a_i^{\dagger_j} \in A_i^\dagger(s_i)$. The relevant parameters have been listed in TABLE 4.

d: REWARD FUNCTION

In QL, the update of the value function is usually based on evaluating the effect of independent action. However, the main agent's action is in the form of a mixed strategy. The update process of Q-value for each sub-action of mixed strategy needs to be updated independently. It is necessary to define a special reward function modality for main agents.

Definition 13: For main agent i, the reward for its sub-action $a_i^{\dagger_j}$, $r_i\left(s_i, a_i^{\dagger_j}\right)$ can be defined as follows.

$$r_i\left(s_i, a_i^{\dagger_j}\right) = \frac{1 - \exp^{\left(-\eta_{\dagger_j} \cdot \log\left(\left|\Delta c_{\dagger_j}\right|+1\right)\right)}}{1 + \exp^{\left(-\eta_{\dagger_j} \cdot \log\left(\left|\Delta c_{\dagger_j}\right|+1\right)\right)}} \tag{26}$$

$$\Delta c_{\dagger_j} = \Delta c_{\dagger_j}^{base} + \varpi_i^{\dagger_j} \Delta c_{\dagger_j}^{siack} \tag{27}$$

$$\Delta c_{\dagger_j}^{base} = \bar{c}_{b^{\dagger_j}}^{k-1} - \bar{c}_{b^{\dagger_j}}^{k} \tag{28}$$

$$\Delta c_{\dagger}^{siack} = \bar{c}_{b^{\check{\dagger}}}^{k-1} - \bar{c}_{b^{\check{\dagger}}}^{k} \tag{29}$$

$$\varpi_i^{\dagger_j} = \frac{\exp^{Q_i\left(s_i, a_i^{\dagger_j}\right)}}{\sum\limits_{a_i^{\dagger_j} \in A_i(s_i)} \exp^{Q_i\left(s_i, a_i^{\dagger_j}\right)}} \tag{30}$$

$$\eta_{\dagger_j} = \begin{cases} 1, & \Delta c_{\dagger_j} \geq 0 \\ -1, & \Delta c_{\dagger_j} < 0 \end{cases} \tag{31}$$

Similar to formula (15), the range of formula (26) is restricted in $[-1, 1]$. Formula (27) shows that the travel cost input variable of a sub-action has two components: sub-action effect estimation $\Delta c_{\dagger_j}^{base}$ defined by formula (28) and main agent passive impact supplement estimation $\varpi_i^{\dagger_j} \Delta c_{\dagger_j}^{siack}$ defined by integrating equations (29) and (30). Formula (28) models the utility among $b^{\dagger_j}$ selected by sub-action $a_i^{\dagger_j}$. Formula (29) models the utility among $b^{\check{\dagger}}$ unselected by σ_i. It is the external utility provided by σ_i. The various mean values

in (28) and (29) can be defined in a similar way to (17). The weight of external utility assigned to sub-action $a_i^{\dagger_j}$ is defined in Formula (30). Similar to formula (18), formula (31) describes a coefficient $\eta_{\dagger_j}$ to ensure the plus or minus characteristic of formula (31).

Based on the above reward function, among main agents, their Q functions can be calculated with a series of rewards synchronously.

e: OTHER RELEVANT PARAMETERS

For the discount factor, a similar reason has been discussed in Section IV.A.1.(e). It is suitable to set the discount factor of the main agents as 0.9 uniformly.

Nevertheless, because of different mission objectives, the definition of the main agents' learning rate differs from that of adviser agents. Moreover, the sensitivity of MARL also needs to be considered among the main agents. In other words, each main agent should be equipped with a self-adaptive ability for dynamic and polytropic external factors and with learning ability for new sub-goal that derived from its task with new surroundings. Therefore, the learning rate can be defined as follows.

Definition 14: The learning rate of main agent i, α_i is a function of the variance of individual travel costs. It can be expressed as follows.

$$\alpha_i = \frac{1 + \exp^{-\mu_i}}{1 - \exp^{-\mu_i}} \tag{32}$$

In formula (32), the parameter in the right term can be further described via equation (33).

$$\mu_i = \sum_{b \in B_i} f_b \left(c_b - \bar{c}_i\right)^2 \tag{33}$$

The individual average travel cost $\bar{c}_i$ in formula (33) can be calculated in a manner similar to formula (17). Since that the variance of individual travel costs is non-negative and finiteness, the interval of the corresponding function (i.e. formula (33)) is also non-negative, [0, 1). In formula (33), $f_b \left(c_b - \bar{c}_i\right)^2$ represents the total deviation at b. The global deviation input is constructed by formula (33). Under this condition, it is feasible to adopt this definition as the learning rate of the main agent. Employing the above mapping mechanism to improve the learning rate can enhance the sensitivity of MARL among main agents. Additionally, the parameters used in this sub-section (i.e. the section of the main agent) have been listed in TABLE 4.

B. ALGORITHM OF HAB-MARL

According to the definitions in Section IV.A, it summarizes HAB-MARL algorithm performed by each agent in Algorithm 1.

V. NUMERICAL SIMULATION EXPERIMENT

This section presents the results of the numerical simulation to evaluate the performance of the proposed distributed assignment method based on HAB-MARL. The traffic network topology, the contrast methods, and the corresponding experimental scenarios are described in Section V.A. Then, in Section V.B, the framework of the numerical simulation is given in pseudo-code. Finally, details of the results are discussed in Section V.C. The following simulation and evaluation are implemented employing an applicable evaluation testbed based on MATLAB. Moreover, the parameters, which were not mentioned above but will be adopted in this section, are listed in TABLE 5.

Algorithm 1 Pseudo-Code of HAB-MARL

Input: the discount rate γ_{HAB}; the set of state S;
the space of action A; the set of agents I_i;
simulation time T; greedy coefficient ε;
Initialize: the simulation time $t \leftarrow 0$; the step of simulation $k \leftarrow 0$
for $i \in I_m / I_{at} / I_{as}$ **do**
(main agents/critical time adviser agents/ critical size adviser agents)
if $s_i^k \in S_i, a_i^k \in A_i, S_i \in S, A_i \in A_m / A_{at} / A_{as}$ **do**
$Q_i^k \left(x_i^k, a_i^k\right) \leftarrow random$
end if
end for
Learning process:
While $t \leq T$ **do**
extracting the state: s_i^k
for $i \in I_{at} / I_{as}$ **do** (recommendation from adviser agents)
select: action a_i^k according to ε greedy
end for
for $i \in I_m$ **do** (calculating restricted strategy space of main agents)
calculate: the propositional strategy space $A_i^\dagger \left(s_i^k\right)$ (see **Definition 10**)
end for
for $i \in I_m$ **do** (main agents learning process)
select mixed strategy: σ_i^k according to $\left\{Q_i^k \left(s_i^k, a_i^{\dagger_j}\right)\right\}$
and $a_i^{\dagger_j} \in A_i^\dagger \left(s_i^k\right)$. (see **Definition 11**)
for $a_i^{\dagger_j} \in A_i^\dagger \left(s_i^k\right)$ **do**
calculate $r_i^k \left(s_i, a_i^{\dagger_j}\right)$
end for
update learning rate: α_i^k according to formula (32)
for $a_i^{\dagger_j} \in A_i^\dagger \left(s_i^k\right)$ **do**
update Q-value:

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

end for
end for
for $i \in I_{at} / I_{as}$ **do** (adviser agents learning process)
calculate: $r_i^k \left(s_i^k, a_i^k\right), a_i^k \in A_i, A_i \in A_{at} / A_{as}$
(consistent with recommendation from adviser agents)
update learning rate: α_i^k according to formulas (19) and (22)
update Q-value:

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

end for
$t = t + 1, k = k + 1$
end while

TABLE 5. The parameters of numerical simulation.

Notations	Signification
t_l	The travel time of link l
t_l^f	The free travel time of link l
f_l	The flow of l
χ_l^t	The capacity of link l
τ	The correction factor
β	The correction factor
$\Im$	The set of hyper-parameters $\Im=\left\{\gamma_{EB},\gamma_{RB},\gamma_{HAB},\alpha_{EB},\alpha_{RB},\varepsilon,t_{critical}^{EB},\left\|\Pi_{precomputed}^{RB}\right\|\right\}$
I_{HAB}	The set of agents in HAB-MARL
$I_{HAB\text{-}NAT}$	The set of agents in HAB-MARL-NAT
$I_{HAB-NAS}$	The set of agents in HAB-MARL-NAS
I_{RB}	The set of agents in RB-MARL
I_{EB}	The set of agents in EB-MARL
OTF_X	The overall throughput flow under using method X, $X\in\left\{\begin{matrix}\text{HAB-MARL-NAT,HAB-MARL-NAS,}\\\text{HAB-MARL,RB-MARL,EB-MARL}\end{matrix}\right\}$
ATT_X	The effective average individual travel time under using method X, $X\in\left\{\begin{matrix}\text{HAB-MARL-NAT,HAB-MARL-NAS,}\\\text{HAB-MARL,RB-MARL,EB-MARL}\end{matrix}\right\}$

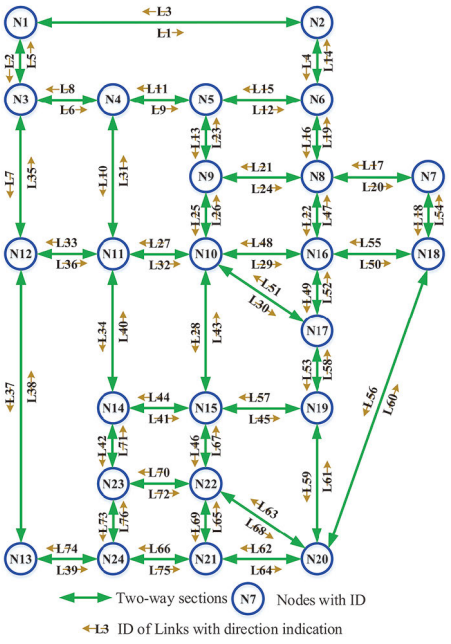


FIGURE 2. Sioux Falls network.

A. SIMULATION SETTING

1) TRAFFIC NETWORK

In this paper, we chose Sioux Falls (SF) network (see FIGURE 2) with 24 nodes and 76 edges widely used to test traffic assignment or route choice methods in the literature, such as [60], [33] and [68]. Compared to commonly adopted conceptual grid networks, which only has few OD pairs and one-way roads or paths, SF network describes a situation which is closer to the real world. In addition, compared with Anaheim network and Chicago network in [68], SF network has a relatively simple structure which is convenient for numerical calculation. The basic information of SF network is listed in Tables 8 and 9. Following the description in [71], the well-known Bureau of Public Roads (BPR) function (i.e. equation (34)) is employed to model the travel cost on each edge in SF network.

$$t_l\left(f_l\right)=t_l^f\left[1+\tau\left(\frac{f_l}{\chi_l}\right)^{\beta}\right] \tag{34}$$

The parameters in formula (34) have been described in TABLE 5. The values of coefficient τ and β are 0.15 and 4 respectively. The above calibration values are usually used in the BPR-type function that describes the travel cost of a general road, for instance, the travel cost in [71], [60], and [72].

2) CONTRAST METHODS

Considering that the traffic assignment process is impossible to be performed repeatedly in real traffic situations, many iterative methods requiring reassignment are impractical. Therefore, to evaluate and compare the proposed assignment approach based on HAB-MARL, we select two agent-based methods named EB-MARL and RB-MARL, which have been suggested in [33] and briefly mentioned in Section III.C. A more detailed description of EB-MARL and RB-MARL can be referred to [33].

For EB-MARL and RB-MARL, the recommended parameters have been verified optimal for TA in [33]. Considering that will be tested in SF network, the best combination of parameters for EB-MARL is $\alpha_{EB}=0.9$, $\gamma_{EB}=0.99$ and $t_{critical}^{EB}=9229$. Under the same experimental network, the recommended optimal parameters of RB-MARL are $\alpha_{RB}=0.9$, $\gamma_{RB}=0.99$, and $\left|\Pi_{precomputed}^{RB}\right|=10$.

We also construct two extra contrast methods via breaking the components of HAB-MARL. One, which is named HAB-MARL-NAT for short, has no critical time adviser. The other without critical size advisers can be called HAB-MARL-NAS. The settings of HAB-MARL-NAT and HAB-MARL-NAS refer to the corresponding components of HAB-MARL (see Section IV.A).

3) SCENARIO

To verify the performance of the proposed method and contrast methods, we designed a scenario with the time-dependent flow in SF network. In this scenario, the traffic demand of each OD-pair varies in the time dimension. To describe the traffic demand for each OD pair in this scenario succinctly and efficiently, the average of Poisson distribution is adopted to stand for corresponding fluctuant traffic demand. The scenario is visualized in FIGURE 3, the

FRAMEWORK 1 Pseudo-Code of the Numerical Simulation Framework

Input: the traffic network $G(N, E)$; the set of state S;
the space of action A; the set of agents I;
total simulation time T; the set of hyper-parameters $\Im$;
Initialize: the simulation time $t \leftarrow 0$; the step of simulation $k \leftarrow 0$
initialize the Q-value for all agent according **HAB-MARL /HAB-MARL-NAT/HAB-MARL-NAS /RB-MARL/EB-MARL**
initialize the condition of traffic network according Tables 8 and 9.
Simulation process:
While $t \leq T$ **do**
for $i \in I_{HAB}/I_{HAB-NAT}/I_{HAB-NAS}/I_{RB}/I_{EB}$ **do**
(DTA: dynamic traffic assignment)
search optimal traffic control actions a_i^k according **HAB-MARL /HAB-MARL-NAT/HAB-MARL-NAS/RB-MARL/EB-MARL**
end for
for $l \in E$ **do (DNL: dynamic network loading)**
update the condition S of traffic network according **DNL** and the above a_i^k
end for
for $i \in I_{HAB}/I_{HAB-NAT}/I_{HAB-NAS}/I_{RB}/I_{EB}$ **do (Process of learning)**
calculate: reward function (Evaluating the effects of agent actions)
if $i \in I_{HAB}/I_{HAB-NAT}/I_{HAB-NAS}$ **do**
update learning rate for each agent
end if
update the **Q-value** in **MARL** according **HAB-MARL /HAB-MARL-NAT/HAB-MARL-NAS/RB-MARL/EB-MARL**
end for
$t = t + 1, k = k + 1$
end while

accurate data (the statistical average value) of which are listed in Table 7.

In FIGURE 3, sub-graph (a) describes the stochasticity of demand (the statistical average value) among different OD pairs. The volatility of demand (the actual value) in time for each OD pairs has been shown in sub-graph (b).

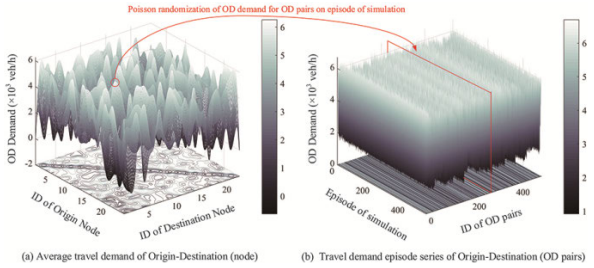


FIGURE 3. Configuration of demand among OD pairs.

B. FRAMEWORK OF NUMERICAL SIMULATION

For the sake of understanding, the framework of numerical simulation constructed in this paper will be described briefly in the form of pseudo-code (i.e. FRAMEWORK 1).

The parameters in FRAMEWORK 1 have been described in TABLE 4 and 5. This framework can be regarded as an all-purpose platform based on the DNL model.

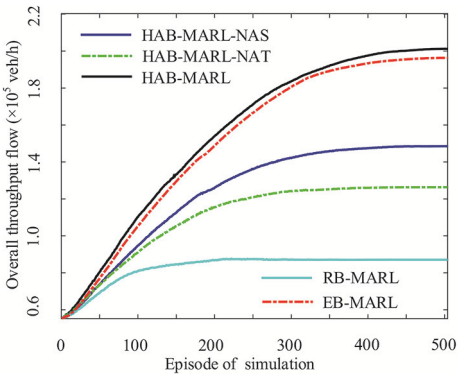


FIGURE 4. Overall throughput flow of the network.

C. ANALYSIS AND EVALUATION OF RESULTS

In this sub-section, the effectiveness and superiority of the method suggested by us can be demonstrated from three aspects: (a) throughput of the network, (b) the average travel cost of each vehicle, and (c) the situation faced by agents and the result of corresponding assignment.

1) THROUGHPUT

The throughput of the network reflects the service level of the network. The higher the throughput, the smoother the road network. Travelers can reach their destinations easily. The effect of each method on the throughput of the network is shown in FIGURE 4. In FIGURE 4, it is obvious that the suggested method performs better than other comparative approaches (i.e. relational expression (35)-(38) provided below). The increasing throughput indicates that HAB-MARL can improve the smoothness of the network continuously. The throughput improvement of HAB-MARL over other methods is recorded in TABLE 6. To simplify the description and analysis process, we use OTF_X to represent the throughput under using method X. In combination with Figure 4 and Table 6, The performance of each method follows the following relationship:

$$OTF_{HAB-MARL} > OTF_{EB-MARL} \quad (35)$$

$$OTF_{EB-MARL} > OTF_{HAB-MARL-NAS} \quad (36)$$

$$OTF_{HAB-MARL-NAS} > OTF_{HAB\_MARL-NAT} \quad (37)$$

$$OTF_{HAB\_MARL-NAT} > OTF_{RB-MARL} \quad (38)$$

Considering expressions (35)-(38), RB-MARL performs worst. It indicates that the fixed critical size (RB-MARL) will result in performance loss compared with the adaptive critical size (HAB-MARL and HAB-MARL-NAT). For RB-MARL, it can be thought of as the existence that flow assigned to impertinent paths or sectional appropriate paths with low utilization rate. These situations will lead to a decrease in the throughput of the network.

However, the opposite effect was observed in the similar comparison on the aspect of critical time (i.e. expression (36)). The situation can be attributed to the emergence of using inappropriate paths when the adaptive critical time

exceeds the fixed critical time. In other words, low efficiency is derived from that the adaptive critical time deviates from the optimal travel time.

In addition to the above analyses, advice on critical time contributes more to performance improvement than that on critical size. The reason for the significant difference in performance is the extent difference on the exploitation of potential feasible path set with facing high load demand in some sections of the network. The fact that the potential path is not utilized effectively indicates that the constraint of critical size on the strategy space of MARL is tighter than that of critical time.

It is also verified to some extent from the fact that HAB-MARL is superior to HAB-MARL-NAS. The exploration of bad paths is reduced by supplementary constraints from critical size. The loss of performance is thus avoided.

About throughput, the performance of EB-MARL is second only but close to that of HAB-MARL. The fixed critical time of EB-MARL is the optimal travel time resulting in that the infeasible paths are excluded. Nevertheless, the fixed critical time is still not suitable for dynamic systems because of the temporal variability of the optimal travel time. Besides, the optimal travel time is only for the specified scenario, which needs to be re-optimized once the scenario changes. It is enough to explain the poor portability of EB-MARL.

In summary, HAB-MARL, which is the method proposed by us, is effective and excellent.

TABLE 6. Evaluation results of HAB-MARL.

contrast Methods	The improvement of HAB-MARL (%)	
	Throughput	Average travel time
HAB-MARL-NAT	59.278	34.674
HAB-MARL-NAS	35.515	26.755
EB-MARL	2.5084	7.1733
RB-MARL	131.06	46.634

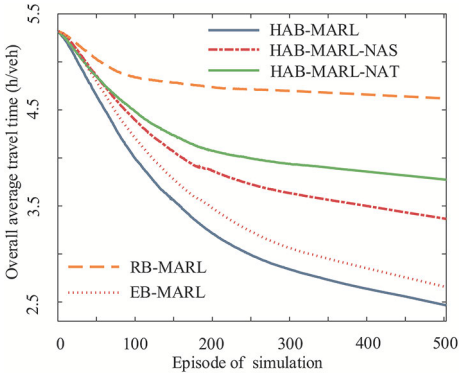


FIGURE 5. Individual average travel time among the overall network.

2) TRAVEL COST

In this paper, the cost of individual travel in the network can be expressed as average travel time (ATT). The lower ATT, the more efficient the trip, namely travelers can reach their destinations sooner. In FIGURE 5, it displays the effect of each approach on the travel cost of individual traveler in the network. The ATT improvement of HAB-MARL over other approaches has been listed in TABLE 6. Similar to throughput analysis, we also adopt ATT_X to represent the individual travel cost under using method X. Moreover, the following similar relational expressions can be acquired:

$$ATT_{HAB-MARL} < ATT_{EB-MARL} \quad (39)$$

$$ATT_{EB-MARL} < ATT_{HAB-MARL-NAS} \quad (40)$$

$$ATT_{HAB-MARL-NAS} < ATT_{HAB-MARL-NAT} \quad (41)$$

$$ATT_{HAB-MARL-NAT} < ATT_{RB-MARL} \quad (42)$$

The expressions show that the performance of methods tested in this paper on the aspect of individual travel time is consistent with that on the aspect of throughput. The reasons for the difference in the degree of reducing individual average travel time can be divided into two categories: (a) the excessive utilization of non-optimal paths (such as RB-MARL, HAB-MARL-NAS), and (b) the idle spatiotemporal resources of feasible paths (such as EB-MARL, RB-MARL, HAB-MARL-NAT). The similar detailed analysis process has been stated in the throughput analysis (see Section V.C.1). Analyzing the influence of each method on the improvement degree of ATT can be executed by referring to the analysis process in the previous sub-section. Hence, we will not provide a detailed discussion here. To sum up, HAB-MARL performs well in terms of the individual average travel time in the whole network.

In addition to the above analysis, the influence of the suggested method on individual average travel time among each OD pair is further shown in FIGURE 6.

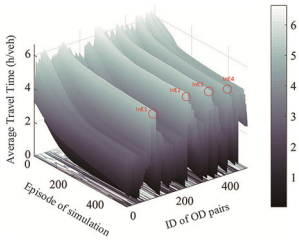


FIGURE 6. Individual average travel time among each OD pair (HAB-MARL).

In FIGURE 6, we chose 4 OD pairs (red circles) with inefficient improvement on the aspect of ATT seemingly and name them from IE1 to IE4 (red font). However, the actual improvement effects are as follows: (a) IE1 (N6-N13): 12.72% (from 5.7954 h/veh to 5.2153 h/veh), (b) IE2 (N13-N1): 16.52% (from 6.5426 h/veh to 5.4609 h/veh), (c) IE3 (N17-N12): 14.37% (from 6.0682 h/veh to 5.1964 h/veh), and (d) IE4 (N21-N6): 21.48% (from 6.468 h/veh to 5.0787 h/veh). Obviously, the worst improvement was over 10%. Moreover, IE1 has the worst improvement among all OD pairs

statistically. It indicates that HAB-MARL improves the individual average travel time for each OD pair effectually.

3) SITUATION WITH THE CORRESPONDING ASSIGNMENT RESULT

To illustrate the properties of DTA based on HAB-MARL, we select five arbitrary OD pairs in SF network. For these OD pairs, the effective travel time evolutions of each path as well as the corresponding situations of the assignment are shown in FIGURE 7-11.

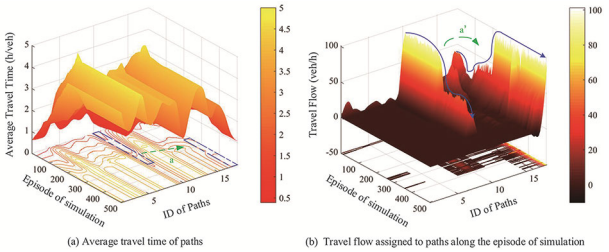


FIGURE 7. **Average travel time and flow assignment of each path among N1-N22.**

The evolution process of effective travel time for each route is plotted in FIGURE 7(a). Areas with the lowest effective travel times have been highlighted with a blue dotted line. In FIGURE 7(b), the path flow trend is outlined in the blue curve. In FIGURE 7, with the change of the lowest location area of the effective travel time ('a' with green in FIGURE 7(a), the fluctuation range of the effective travel time controlled within 0.4583 h/veh to 1.0833 h/veh), the flow also shifted correspondingly between paths ('a'' with green in FIGURE 7(b), the flow assigned from path 10 (i.e. N1→N3→N4→N5→N9→N10→N15→N22) to path 18 (i.e. N3→N12→N13→N24→N23→N22)). The transfer process indicated by the green virtual arrow is the specific manifestation of equilibrium in DTA. It indicates that DTA with HAB-MARL can deal well with flow assignment under fluctuant effective travel time.

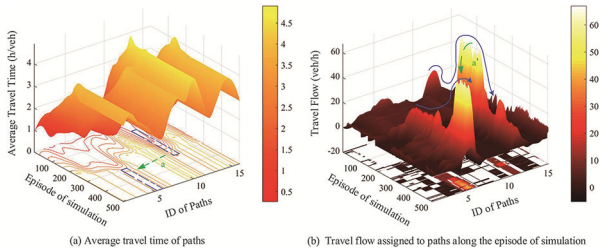


FIGURE 8. **Average travel time and flow assignment of each path among N4-N22.**

The 'a' (the fluctuation range of the effective travel time is controlled within 0.3194 h/veh to 2.2917 h/veh) and 'a'' (the flow assigned from path 12 (i.e. N4→N5→N9→N10→N15→N22) to path 6 (i.e. N4→N3→N12→N13→N24→N23→N22)) in FIGURE 8(a)-(b) verify that HAB-MARL implements flow migration between paths efficaciously once again. Additionally, the fluctuation, which blue curve on path 12 of N4-N22 (i.e. N4→N5→N9→N10→N15→N22) drops briefly in early-stage (FIGURE 8(b), from 17.67 veh/h to 3.9812 veh/h to 62.4286 veh/h), reflects the responsiveness of HAB-MARL for the complex and changeable external environment.

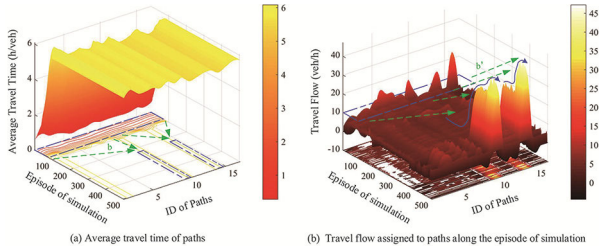


FIGURE 9. **Average travel time and flow assignment of each path among N6-N14.**

For N6-N14, the effective travel times of each path begin with low value at the early stage (between 0.4167 h/veh and 0.7917 h/veh). However, they both increase rapidly in the following process (between 5.4167 h/veh and 6.0833 h/veh). This process is reflected via the areas marked by blue dotted lines in FIGURE 9(a). In this situation, the effective travel time at the later stage is inconspicuously low among paths (the deviation between maximum and minimum is 11.03%). In FIGURE 9(b), with the support of HAB-MARL, DTA shifts the distribution pattern of flow from early uniform distribution among all paths to later centralized distribution among a few paths (path 10 (i.e. N6→N8→N7→N18→N20→N19→N15→N14) with effective travel time 5.4861 h/veh and path 14 (i.e. N6→N8→N9→N10→N15→N14) with effective travel time 5.4444 h/veh). Combining with FIGURE 9(a) and 9(b), it can be found that HAB-MARL realizes the process of assigning flow to the feasible paths (i.e. paths with the lowest effective travel time) in simulation (i.e. the process 'b' and 'b'' with green in FIGURE 9(a) and 9(b)). This phenomenon conforms to the notion of DTA. It shows that HAB-MARL can implement DTA effectively.

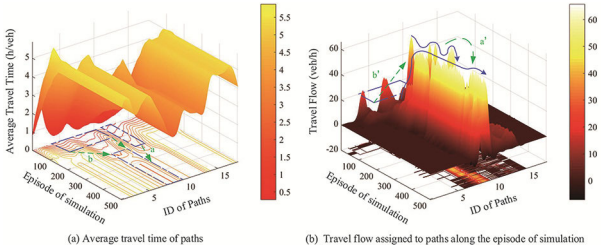


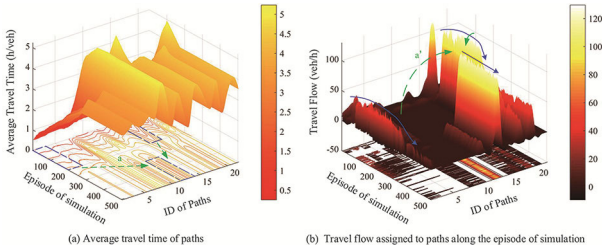
FIGURE 10. **Average travel time and flow assignment of each path among N11-N20.**

In FIGURE 10(a) and 10(b), processes marked by 'a' (the fluctuation range of the effective travel time controlled within 0.2778 h/veh to 2.6111 h/veh) and 'a'' (from path 11 (i.e. N11→N10→N16→N18→N20) to path 9 (i.e.

TABLE 7. Average value of stochastic travel demand for each origin-destination (OD) in the Sioux-falls (SF) network.

Origin	Destination	Stochastic travel demand (average value) (veh/h) Origin-Destination																						
	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11	N12	N13	N14	N15	N16	N17	N18	N19	N20	N21	N22	N23	N24
N1	0	5129	5543	1964	5579	4286	1829	2664	3893	5786	5821	2107	5843	5786	3614	5064	2036	3321	5593	5021	5793	4393	1543	5286
N2	5679	0	4500	4864	4800	3186	4393	2164	4629	1529	2657	1593	1829	5164	4579	2836	5750	0	1536	3400	0	3136	0	0
N3	4900	5036	0	2243	3636	3429	4350	4643	4850	2650	4507	4393	2129	1929	3671	5793	2943	0	0	0	0	4071	2407	0
N4	4836	2550	3707	0	4593	5479	5793	3900	2014	2064	2564	5250	2550	5129	2500	5657	2993	2286	2536	4214	3557	3000	5200	4071
N5	3907	5610	2693	4864	0	4850	3129	3993	1729	1629	3821	4964	5679	1979	3993	3536	1436	0	2929	2129	5036	2814	3814	0
N6	2143	4150	2593	4386	4550	0	4821	3450	1764	2436	5579	2079	5179	3857	5964	1743	3414	1871	5807	1400	4943	5136	5379	1771
N7	3221	2579	5057	3364	5571	2214	0	2593	2050	2007	5379	4043	3907	2050	5300	4243	2993	3743	3229	1729	2486	1950	2229	2486
N8	3300	1607	5529	5729	3636	3629	2936	0	5521	3079	1893	4971	3171	2493	3236	1821	1986	5714	5779	4029	1657	2457	3007	5157
N9	1450	1579	2157	4364	4743	4357	3457	3893	0	2743	4807	2250	4536	2221	3071	4257	4971	1750	5657	4950	3621	3386	3436	2786
N10	3721	3729	5143	5036	4343	3121	5114	3829	2993	0	5700	5407	3907	4243	4079	2336	2764	3543	2443	5264	2279	2421	2164	2429
N11	3386	2814	5629	3357	2229	5543	5886	3400	1893	2564	0	3257	4114	2586	4150	4650	2400	1921	2743	2843	3329	3714	1771	2586
N12	5064	1514	5650	4736	3629	4043	2471	3493	5807	3893	3779	0	2443	3629	4250	4507	3200	3071	5921	1550	5450	5579	5043	1836
N13	2586	2921	4507	2007	4700	1871	4386	3650	4964	4671	5536	5479	0	2914	4593	2293	1521	4800	3679	3586	5543	4186	4221	5336
N14	5086	4036	2221	2486	5457	1514	3636	2150	5879	4657	3679	3550	1657	0	4514	1579	1707	3779	1821	5143	5143	4700	2071	4414
N15	3764	5857	4364	5064	3464	3371	5179	1764	1993	2179	3179	5207	5071	1657	0	3214	3800	3300	4400	4271	2721	3364	1450	5907
N16	2150	1871	3093	2293	3636	2943	5757	5614	1621	4771	2614	3321	3900	5714	3300	0	5900	2764	4607	4443	3857	4593	4443	2200
N17	1971	5979	2164	1529	3964	5436	4457	2257	3079	3500	5893	2100	5314	4343	3107	2257	0	3350	3600	1936	4093	2421	3150	4064
N18	2536	0	0	2714	0	4221	2600	5171	5900	4736	2964	4064	1879	5550	5429	5143	2579	0	4114	1486	3336	2821	2121	0
N19	2200	3321	0	1814	4136	3543	4579	4600	4314	1536	1693	2850	3821	4393	3257	5150	4686	5836	0	3821	2879	1864	4193	4964
N20	3329	1800	0	2607	2086	2671	3407	3807	3486	5407	3764	5721	4314	5786	2486	4493	2707	4471	4579	0	1693	2550	2407	4450
N21	5264	0	0	2964	4971	4486	1414	4150	3157	5593	1386	3507	3329	3500	4921	2864	4993	3550	1543	2186	0	4700	3557	2086
N22	2950	4171	2264	4779	2500	5600	2614	4900	2250	2700	1800	4029	4521	3893	3336	4343	4357	4500	4307	5729	2343	0	4643	2464
N23	1929	0	4171	3450	3493	4421	4921	2993	4421	3293	5250	5214	2557	4200	4057	3864	5379	2600	2843	1929	5700	4350	0	3586
N24	4321	0	0	3886	0	4357	3879	4693	3786	5950	2386	1864	1886	1671	3243	3443	30644	0	4893	4271	4929	5671	5857	0

N11→N10→N15→N22→N20)) are similar to those in FIGURE 7 and FIGURE 8. Moreover, processes singed by 'b' (the fluctuation range of the effective travel time controlled within 0.3889 h/veh to 2.6111 h/veh) and 'b' (converge the flow assigned paths 7 (i.e. N11→N10→N9→N8→N7→N18→N20), 8 (i.e. N11→N10→N15→N19→N20), 9 (i.e. N11→N10→N15→N22→N20), 10 (i.e. N11→N10→N16→N17→N19→N20) to path 9 (i.e. N11→N10→N15→N22→N20)) in FIGURE 10(a) and (b) are similar to those in FIGURE 9. The above processes in FIGURE 10 fully demonstrate the effectiveness of HAB-MARL in DTA.



(a) Average travel time of paths (b) Travel flow assigned to paths along the episode of simulation

FIGURE 11. Average travel time and flow assignment of each path among N20-N12.

The situation in FIGURE 11 (the fluctuation range of the effective travel time controlled within 0.2917 h/veh to 2.4861 h/veh; the flow assigned from path 1 (i.e. N20→N18→N7→N8→N6→N2→N1→N3→N12) and 20 (i.e. N20→N22→N21→N24→N13→N12) to path 13 (i.e. N20→N19→N15→N10→N11→N12)) can be categorized together with those in FIGURE 7 and 8. Additionally, the emergence of this phenomenon in FIGURE 11 indicates that HAB-MARL can recognize a small change in path effective travel time and adjust the assignment action immediately. It is further evidence that HAB-MARL can accomplish DTA tasks dynamically and efficiently.

In this sub-section, all of the above five different situations confirmed that HAB-MARL strictly complied with W-Ps when completing DTA tasks.

VI. CONCLUSION

The dynamic traffic demand, which is unassignable on the time scale, is the general state of the modern urban network. This objective and actual state usually not be considered in the assumptions of existing ideal DTA methods. For DTA with this state, a decentralized method based on MARL has been presented in this article. A flat hierarchical architecture of agent-teams is employed, where each agent-team is associated with an OD pair and interacts with the environment. To restrict strategy space of MARL in response to dynamic equilibrium objective adaptively and rapidly, the basic structure of agent-team has been designed as a combination of one main agent (corresponding TA) and two adviser agents (corresponding critical time and critical size respectively). To achieve DTA, the decision mechanism of main agents is defined as a mixed strategy, which is a concept in Game Theory (GT). In this way, the experience of main agents can be updated in batches. To maintain the ability to re-learn at any time for tacking fickle equilibrium point, learning rates have been designed specially according to different tasks of different agents. The global search ability of MARL on DTA tasks has been enhanced. With this improvement, the DTA system can rapidly and effectively respond to dynamic traffic demands which are unassignable in temporal-dimension. Thus, DTA can achieve its desired objectives, namely minimizing the individual average travel time in an urban network. Meanwhile, the traffic congestion or potential congestion

TABLE 8. Attributes of Sioux-Falls network.

ID of Link	ID of head node	ID of tail node	Free travel time (s)	Length (m)	Capacity (veh/h)
1	2	1	360	9656	25900
2	3	1	240	6437	23403
3	1	2	360	9656	25900
4	6	2	300	8046	4958.2
5	1	3	240	6437	23403
6	4	3	240	6437	17111
7	12	3	240	6437	23403
8	3	4	240	6437	17111
9	5	4	120	3218	17783
10	11	4	360	9656	4908.8
11	4	5	120	3218	17783
12	6	5	240	6437	4948
13	9	5	300	8046	10000
14	2	6	300	8046	4958.2
15	5	6	240	6437	4948
16	8	6	120	3218	4898.6
17	8	7	180	4828	7841.8
18	18	7	120	3218	23403
19	6	8	120	3218	4898.6
20	7	8	180	4828	7841.8
21	9	8	600	16093	5050.2
22	16	8	300	8046	5045.8
23	5	9	300	8046	10000
24	8	9	600	16093	5050.2
25	10	9	180	4828	13916
26	9	10	180	4828	13916
27	11	10	300	8046	10000
28	15	10	360	9656	13512
29	16	10	240	6437	4854.9
30	17	10	480	12875	4993.5
31	4	11	360	9656	4908.8
32	10	11	300	8046	10000
33	12	11	360	9656	4908.8
34	14	11	240	6437	4876.5
35	3	12	240	6437	23403
36	11	12	360	9656	4908.8
37	13	12	180	4828	25900
38	12	13	180	4828	25900
39	24	13	240	6437	5091.3
40	11	14	240	6437	4876.5
41	15	14	300	8046	5127.5
42	23	14	240	6437	4924.8
43	10	15	360	9656	13512
44	14	15	300	8046	5127.5
45	19	15	180	4828	14565
46	22	15	180	4828	9599.2
47	8	16	300	8046	5045.8

TABLE 9. Attributes of Sioux-Falls network.

ID of Link	ID of head node	ID of tail node	Free travel time (s)	Length (m)	Capacity (veh/h)
48	10	16	240	6437	4854.9
49	17	16	120	3218	5229.9
50	18	16	180	4828	19680
51	10	17	480	12875	4993.5
52	16	17	120	3218	5229.9
53	19	17	120	3218	4824
54	7	18	120	3218	23403
55	16	18	180	4828	19680
56	20	18	240	6437	23403
57	15	19	180	4828	14565
58	17	19	120	3218	4824
59	20	19	240	6437	5002.6
60	18	20	240	6437	23403
61	19	20	240	6437	5002.6
62	21	20	360	9656	5059.9
63	22	20	300	8046	5075.7
64	20	21	360	9656	5059.9
65	22	21	120	3218	5229.9
66	24	21	180	4828	4885.4
67	15	22	180	4828	9599.2
68	20	22	300	8046	5075.7
69	21	22	120	3218	5229.9
70	23	22	240	6437	5000
71	14	23	240	6437	4924.8
72	22	23	240	6437	5000
73	24	23	120	3218	5078.5
74	13	24	240	6437	5091.3
75	21	24	180	4828	4885.4
76	23	24	120	3218	5078.5

risk caused by stochastic traffic demand can be mitigated or eliminated from the origin.

To assess the proposed method (i.e. HAB-MARL), a comparative analysis of performance is designed. In SF network with fluctuant and unbalanced OD demands, HAB-MARL is benchmarked against two existing methods (i.e. EB-MARL and RB-MARL) and two ablative methods (i.e. HAB-MARL-NAT and HAB-MARL-NAS) to evaluate the performance of HAB-MARL in terms of the network overall throughput and the network individual average travel time. By visualizing the path travel distribution and effective individual travel time on the perspective of OD pairs, the ability of HAB-MARL to fleetly track variable equilibrium points in different OD pairs has been further analyzed. The results show that HAB-MARL outperforms the other contrast approaches in terms of the network throughput and the network individual average travel time. Although the performance of the EB-MARL method with a preset optimal value is close to that of HAB-MARL. Although the performance of EB-MARL with a preset optimal value is close to that of HAB-MARL. EB-MARL lacks HAB-MARL's ability to self-learn without optimal preconditions, which brings superior portability for HAB-MARL. Moreover, HAB-MARL is demonstrated that can be timely and excellent in adjusting flow distribution among paths in response to rise and fall of effective individual travel time.

The method proposed in this paper optimizes the algorithm framework for the scene under the dynamic traffic demand which is unassailable on time dimension. It enhances the adaptive learning ability of MARL for DTA tasks from the perspective of algorithmic theory. Hence, HAB-MARL is basic and can be a compatible framework. For example, existing path generation methods can be embedded in the process of searching HAB-MARL's strategy space to make DTA system more efficient. Exploiting possible extension patterns is a part of our future work.

Additionally, the framework of HAB-MARL can also be reorganized for other similar tasks, which need to be achieved via employing MARL with an action space restricted by cooperative constraints. It means that the architecture of HAB-MARL has universality and plasticity.

APPENDIX

In the appendix, the configuration of travel demands for each Origin-Destination (OD) and the basic attributes of the Sioux-falls (SF) network are provided in Tables 7–9, respectively.

REFERENCES

[1] J. Long, W. Y. Szeto, H.-J. Huang, and Z. Gao, "An intersection-movement-based stochastic dynamic user optimal route choice model for assessing network performance," *Transp. Res. B, Methodol.*, vol. 74, pp. 182–217, Apr. 2015.

[2] J. G. Wardrop, "Road paper. Some theoretical aspects of road traffic research," *Proc. Inst. Civil Engineers*, vol. 1, no. 3, pp. 325–362, May 1952.
[3] L. Wu, Z. Huang, J. Wu, Z. Gao, and D. Qin, "A day-to-day stochastic traffic flow assignment model based on mixed regulation," *IEEE Access*, vol. 8, pp. 12815–12823, 2020.
[4] L. Zhang, J. Liu, B. Yu, and G. Chen, "A dynamic traffic assignment method based on connected transportation system," *IEEE Access*, vol. 7, pp. 65679–65692, 2019.
[5] A. A. Prakash, R. Seshadri, and K. K. Srinivasan, "A consistent reliability-based user-equilibrium problem with risk-averse users and endogenous travel time correlations: Formulation and solution algorithm," *Transp. Res. B, Methodol.*, vol. 114, no. 1, pp. 171–198, Aug. 2018.
[6] S. Peeta and A. K. Ziliaskopoulos, "Foundations of dynamic traffic assignment: The past, the present and the future," *Netw. Spatial Econ.*, vol. 1, nos. 3–4, pp. 233–265, Sep. 2001.
[7] Y. Jiang, S. C. Wong, H. W. Ho, P. Zhang, R. Liu, and A. Sumalee, "A dynamic traffic assignment model for a continuum transportation system," *Transp. Res. B, Methodol.*, vol. 45, no. 2, pp. 343–363, Feb. 2011.
[8] M. Florian, M. Mahut, and N. Tremblay, "Application of a simulation-based dynamic traffic assignment model," *Eur. J. Oper. Res.*, vol. 189, no. 3, pp. 1381–1392, Sep. 2008.
[9] Y. Tian and Y.-C. Chiu, "A variable time-discretization strategies-based, time-dependent shortest path algorithm for dynamic traffic assignment," *J. Intell. Transp. Syst.*, vol. 18, no. 4, pp. 339–351, Oct. 2014.
[10] Y. Nie, "A cell-based Merchant–Nemhauser model for the system optimum dynamic traffic assignment problem," *Transp. Res. Part B, Methodol.*, vol. 45, no. 2, pp. 329–342, Feb. 2011.
[11] S. T. Waller, D. Fajardo, M. Duell, and V. Dixit, "Linear programming formulation for strategic dynamic traffic assignment," *Netw Spat. Econ.*, vol. 13, no. 4, pp. 427–433. 2013.
[12] H.-K. Chen and G. Feng, "Heuristics for the stochastic/dynamic user-optimal route choice problem," *Eur. J. Oper. Res.*, vol. 126, no. 1, pp. 13–30, Oct. 2000.
[13] S. Han, "Dynamic traffic modelling and dynamic stochastic user equilibrium assignment for general road networks," *Transp. Res. B, Methodol.*, vol. 37, no. 3, pp. 225–249, Mar. 2003.
[14] B.-W. Wie, R. L. Tobin, and M. Carey, "The existence, uniqueness and computation of an arc-based dynamic network user equilibrium formulation," *Transp. Res. B, Methodol.*, vol. 36, no. 10, pp. 897–918, Dec. 2002.
[15] X. Ban, H. X. Liu, M. C. Ferris, and B. Ran, "A link-node complementarity model and solution algorithm for dynamic user equilibria with exact flow propagations," *Transp. Res. B, Methodol.*, vol. 42, no. 9, pp. 823–842, Nov. 2008.
[16] X. Ban, J.-S. Pang, H. X. Liu, and R. Ma, "Modeling and solving continuous-time instantaneous dynamic user equilibria: A differential complementarity systems approach," *Transp. Res. B, Methodol.*, vol. 46, no. 3, pp. 389–408, Mar. 2012.
[17] T. L. Friesz, D. Bernstein, Z. Suo, and R. L. Tobin, "Dynamic network user equilibrium with state-dependent time lags," *Netw Spat. Econ.*, vol. 1, nos. 3–4, pp. 319–347. 2001.
[18] K. Han, T. L. Friesz, and T. Yao, "Existence of simultaneous route and departure choice dynamic user equilibrium," *Transp. Res. B, Methodol.*, vol. 53, no. 1, pp. 17–30, Jul. 2013.
[19] X. Nie and H. M. Zhang, "Delay-function-based link models: Their properties and computational issues," *Transp. Res. B, Methodol.*, vol. 39, no. 8, pp. 729–751, Sep. 2005.
[20] C. F. Daganzo, "The cell transmission model, part II: Network traffic," *Transp. Res. B, Methodol.*, vol. 29, no. 2, pp. 79–93, Apr. 1995.
[21] P. I. Richards, "Shock waves on the highway," *Oper. Res.*, vol. 4, no. 1, pp. 42–51, Feb. 1956.
[22] W. Y. Szeto and H. K. Lo, "A cell-based simultaneous route and departure time choice model with elastic demand," *Transp. Res. B, Methodol.*, vol. 38, no. 7, pp. 593–612, Aug. 2004.
[23] W. Y. Szeto, Y. Jiang, and A. Sumalee, "A cell-based model for multi-class doubly stochastic dynamic traffic assignment," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 26, no. 8, pp. 595–611, Nov. 2011.
[24] M. Blumberg-Nitzani and H. Bar-Gera, "The effect of signalised intersections on dynamic traffic assignment solution stability," *Transportmetrica A, Transp. Sci.*, vol. 10, no. 7, pp. 622–646, Aug. 2014.
[25] Y. Lim and B. Heydecker, "Dynamic departure time and stochastic user equilibrium assignment," *Transp. Res. B, Methodol.*, vol. 39, no. 2, pp. 97–118, Feb. 2005.
[26] W. Shen and H. M. Zhang, "System optimal dynamic traffic assignment: Properties and solution procedures in the case of a many-to-one network," *Transp. Res. B, Methodol.*, vol. 65, no. 1, pp. 1–17, Jul. 2014.
[27] S. Fiorenzo-Catalano, R. V. Nes, and P. H. L. Bovy, "Choice set generation for multi-modal travel analysis," *Eur. J. Transport. Infrastruct. Res.*, vol. 4, no. 2, pp. 31–49, 2004.
[28] J. Long, H.-J. Huang, Z. Gao, and W. Y. Szeto, "An intersection-movement-based dynamic user optimal route choice problem," *Oper. Res.*, vol. 61, no. 5, pp. 1134–1147, Oct. 2013.
[29] 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.
[30] Z. Qu, Z. Pan, Y. Chen, X. Wang, and H. Li, "A distributed control method for urban networks using multi-agent reinforcement learning based on regional mixed strategy Nash-equilibrium," *IEEE Access*, vol. 8, pp. 19750–19766, 2020.
[31] Y. Shoham and K. Leyton-Brown, *Multiagent Systems: Algorithmic, Game Theoretic and Logical Foundations*. Cambridge, U.K.: CU Press, 2008.
[32] M. Zolfpour-Arokhlo, A. Selamat, S. Z. Mohd Hashim, and H. Afkhami, "Modeling of route planning system based on Q value-based dynamic programming with multi-agent reinforcement learning algorithms," *Eng. Appl. Artif. Intell.*, vol. 29, pp. 163–177, Mar. 2014.
[33] R. Grunitzki and A. L. C. Bazzan, "Comparing two multiagent reinforcement learning approaches for the traffic assignment problem," in *Proc. Brazilian Conf. Intell. Syst. (BRACIS)*, Uberlândia, Brazil, Oct. 2017, pp. 139–144.
[34] B. N. Janson, "Dynamic traffic assignment for urban road networks," *Transp. Res. B, Methodol.*, vol. 25, nos. 2–3, pp. 143–161, Apr. 1991.
[35] R. Jayakrishnan, W. K. Tsai, and A. Chen, "A dynamic traffic assignment model with traffic-flow relationships," *Transp. Res. C, Emerg. Technol.*, vol. 3, no. 1, pp. 51–72, Feb. 1995.
[36] B. Ran and D. E. Boyce, "A link-based variational inequality formulation of ideal dynamic user-optimal route choice problem," *Transp. Res. C, Emerg. Technol.*, vol. 4, no. 1, pp. 1–12, Feb. 1996.
[37] H.-K. Chen and C.-F. Hsueh, "A model and an algorithm for the dynamic user-optimal route choice problem," *Transp. Res. B, Methodol.*, vol. 32, no. 3, pp. 219–234, Apr. 1998.
[38] S. Han, "A route-based solution algorithm for dynamic user equilibrium assignments," *Transp. Res. B, Methodol.*, vol. 41, no. 10, pp. 1094–1113, Dec. 2007.
[39] T. Larsson and M. Patriksson, "Simplicial decomposition with disaggregated representation for the traffic assignment problem," *Transp. Sci.*, vol. 26, no. 1, pp. 4–17, Feb. 1992.
[40] P. Zhang and S. Qian, "Path-based system optimal dynamic traffic assignment: A subgradient approach," *Transp. Res. B, Methodol.*, vol. 134, no. 1, pp. 41–63, Apr. 2020.
[41] M. Kuwahara and T. Akamatsu, "Dynamic user optimal assignment with physical queues for a many-to-many OD pattern," *Transp. Res. B, Methodol.*, vol. 35, no. 5, pp. 461–479, Jun. 2001.
[42] H. K. Lo and W. Y. Szeto, "A cell-based variational inequality formulation of the dynamic user optimal assignment problem," *Transp. Res. B, Methodol.*, vol. 36, no. 5, pp. 421–443, Jun. 2002.
[43] H. K. Lo and W. Y. Szeto, "A cell-based dynamic traffic assignment model: Formulation and properties," *Math. Comput. Model.*, vol. 35, nos. 7–8, pp. 849–865, Apr. 2002.
[44] Q. Meng and H. L. Khoo, "A computational model for the probit-based dynamic stochastic user optimal traffic assignment problem," *J. Adv. Transp.*, vol. 46, no. 1, pp. 80–94, Jan. 2012.
[45] M. Owais and A. Alshehri, "Pareto optimal path generation algorithm in stochastic transportation networks," *IEEE Access*, vol. 8, pp. 58970–58981, 2020.
[46] G. Xiao, Q. Cheng, and C. Zhang, "Detecting travel modes using rule-based classification system and Gaussian process classifier," *IEEE Access*, vol. 7, pp. 116741–116752, 2019.
[47] G. Xiao, R. Wang, C. Zhang, and A. Ni, "Demand prediction for a public bike sharing program based on spatio-temporal graph convolutional networks," *Multimedia Tools Appl.*, early access, Mar. 2020, doi: 10.1007/s11042-020-08803-y.
[48] Q. Zhaowei, L. Haitao, L. Zhihui, and Z. Tao, "Short-term traffic flow forecasting method with MB-LSTM hybrid network," *IEEE Trans. Intell. Transport. Syst.*, early access, Jul. 29, 2020, doi: 10.1109/TITS.2020.3009725.

[49] L. He, J. Zhang, Y. Hou, X. Liang, and P. Bai, "Time-varying formation control for second-order discrete-time multi-agent systems with directed topology and communication delay," *IEEE Access*, vol. 7, pp. 65379–65389, 2019.
[50] B. Wang, W. Chen, and B. Zhang, "Semi-global robust tracking consensus for multi-agent uncertain systems with input saturation via metamorphic low-gain feedback," *Automatica*, vol. 103, no. 1, pp. 363–373, May 2019.
[51] B. Wang, W. Chen, J. Wang, B. Zhang, and P. Shi, "Semi-global tracking cooperative control for multi-agent systems with input saturation: A multiple saturation levels framework," *IEEE Trans. Autom. Control*, early access, May 6, 2020, doi: 10.1109/TAC.2020.2991695.
[52] C. Deng and C. Wen, "Distributed resilient observer-based fault-tolerant control for heterogeneous multi-agent systems under actuator faults and DoS attacks," *IEEE Trans. Control Netw. Syst.*, early access, Feb. 10, 2020, doi: 10.1109/TCNS.2020.2972601.
[53] W. Du, G. Yang, C. Pan, and P. Xi, "A heterogeneous multi-agent system model with navigational feedback for load demand management of a zonal medium voltage DC shipboard power system," *IEEE Access*, vol. 7, pp. 148073–148083, 2019.
[54] S. Izumi, S.-I. Azuma, and T. Sugie, "Analysis and design of multi-agent systems in spatial frequency domain: Application to distributed spatial filtering in sensor networks," *IEEE Access*, vol. 8, pp. 34909–34918, 2020.
[55] S. Zheng and H. Liu, "Improved multi-agent deep deterministic policy gradient for path planning-based crowd simulation," *IEEE Access*, vol. 7, pp. 147755–147770, 2019.
[56] 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.
[57] B. Chen and H. H. Cheng, "A review of the applications of agent technology in traffic and transportation systems," *IEEE Trans. Intell. Transport. Syst.*, vol. 11, no. 2, pp. 485–497, Jun. 2010.
[58] M. Balmer, N. Cetin, K. Nagel, and B. Raney, "Towards truly agent-based traffic and mobility simulations," in *Proc. AAMAS*, New York, NY, USA, 2004, pp. 60–67.
[59] A. L. C. Bazzan, and F. Klugl, *Multi.Agent Systems for Tfc and Transportation Engineering*, Hershey, PA, USA: IGI Press, 2009.
[60] A. L. C. Bazzan and R. Grunitzki, "A multiagent reinforcement learning approach to en-route trip building," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Vancouver, BC, Canada, Jul. 2016, pp. 5288–5295.
[61] R. Grunitzki, G. D. O. Ramos, and A. L. C. Bazzan, "Individual versus difference rewards on reinforcement learning for route choice," in *Proc. Brazilian Conf. Intell. Syst. (BRACIS)*, São Carlos, Brazil, Oct. 2014, pp. 253–258.
[62] T. Kohonen, "Things you haven't heard about the self-organizing map," in *Proc. IEEE Int. Conf. Neural Netw.*, San Francisco, CA, USA, 1993, pp. 1147–1156.
[63] 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.
[64] F. Zhu and S. V. Ukkusuri, "A cell based dynamic system optimum model with non-holding back flows," *Transp. Res. C, Emerg. Technol.*, vol. 36, pp. 367–380, Nov. 2013.
[65] T. L. Friesz, D. Bernstein, T. E. Smith, R. L. Tobin, and B. W. Wie, "A variational inequality formulation of the dynamic network user equilibrium problem," *Oper. Res.*, vol. 41, no. 1, pp. 179–191, Feb. 1993.
[66] G. F. Newell, "A simplified theory of kinematic waves in highway traffic, part I: General theory," *Transp. Res. B, Methodol.*, vol. 27, no. 4, pp. 281–287, Aug. 1993.
[67] C. Osorio, G. Flötteröd, and M. Bierlaire, "Dynamic network loading: A stochastic differentiable model that derives link state distributions," *Transp. Res. B, Methodol.*, vol. 45, no. 9, pp. 1410–1423, Nov. 2011.
[68] K. Han, G. Eve, and T. L. Friesz, "Computing dynamic user equilibria on large-scale networks with software implementation," *Netw. Spatial Econ.*, vol. 19, no. 3, pp. 869–902, Sep. 2019.
[69] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, 1992.
[70] I. Goodfellow, Y. Bengio, X. Zhang, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[71] L. J. LeBlanc, E. K. Morlok, and W. P. Pierskalla, "An efficient approach to solving the road network equilibrium traffic assignment problem," *Transp. Res.*, vol. 9, no. 5, pp. 309–318, Oct. 1975.
[72] B. Hou, S. Zhao, and H. Liu, "A combined modal split and traffic assignment model with capacity constraints for siting remote park-and-ride facilities," *IEEE Access*, vol. 8, pp. 80502–80517, 2020.

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, dynamic traffic assignment, route planning, application of game theory, and multiagent systems.

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 main current research interests include traffic control theory, traffic organization, traffic big data, and intelligent transportation systems.

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.

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.

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 and traffic data analysis and its applications related approach.