FairMove: A Data-Driven Vehicle Displacement System for Jointly Optimizing Profit Efficiency and Fairness of Electric For-Hire Vehicles

Guang Wang, Sihong He, Lin Jiang, Shuai Wang, Fei Miao, Fan Zhang, Zheng Dong, and Desheng Zhang

Abstract—With the worldwide mobility electrification initiative to reduce air pollution and energy security, more and more for-hire vehicles are being replaced with electric ones. A key difference between gas for-hire vehicles and electric for-hire vehicles (EFHV) is their energy replenishment mechanisms, i.e., refueling or charging, which is reflected in two aspects: (i) much longer charging processes vs. much shorter refueling processes and (ii) time-varying electricity prices vs. time-invariant gasoline prices during a day. The complicated charging issues (e.g., long charging time and dynamic charging pricing) potentially reduce the daily operation time and profits of EFHVs, and also cause overcrowded charging stations during some off-peak charging pricing periods. Motivated by a set of findings obtained from a data-driven investigation and field studies, in this paper, we design a fairness-aware vehicle displacement system called `FairMove` to jointly optimize the overall profit efficiency and profit fairness of EFHV drivers by considering both the passenger travel demand and vehicle charging demand. We first formulate the EFHV displacement problem as a Markov decision problem, and then we present a fairness-aware multi-agent actor-critic approach to tackle this problem. More importantly, we implement and evaluate `FairMove` with real-world streaming data from the Chinese city Shenzhen, including GPS data and transaction data from over 20,100 EFHVs, coupled with the data of 123 charging stations, which constitute, to our knowledge, the largest EFHV network in the world. Extensive experimental results show that our fairness-aware `FairMove` effectively improves the profit efficiency and profit fairness of the EFHV fleet by 26.9% and 54.8%, respectively. It also improves the charging station utilization fairness by 38.4%.

Index Terms—Data-driven, electric vehicle, for-hire vehicle, fairness, deep reinforcement learning.

1 Introduction

More and more countries and cities have started their electric vehicle initiatives [1], [2], [3], [4], [5] because of the ever-growing concern about air quality and energy security. It is reported that the worldwide sales of electric vehicles have nearly quadrupled since 2014, and about 50% of the vehicle sales will be electric vehicles by 2027 [6]. As one of the most common mobility modes, for-hire vehicles (FHVs), such as taxis and those operated by ride-hailing platforms [7] play a very important role in people's daily life, and they are also in the front line of vehicle electrification given their huge potential to reduce air pollution [3]. For example, the number of electric taxis (e-taxi) in the Chinese city Shenzhen has increased from 840 in 2015 to over 20,130 in 2020. For the EFHV fleet, one of the most important tasks for a management team to improve fleet efficiency and resultant fleet profit is **Vehicle Displacement**, i.e., making recommendations to individual vacant EFHVs (i.e., without passengers) to proactively go from one area to another area to balance two relationships: (i) the future passenger demand and vehicle supply and (ii) the future EFHV charging demand and charger supply.

In recent decades, a large number of works have been done to improve the efficiency of FHV fleets [8], [9], [10]. However, a majority of these works focused on conventional gas FHVs. For example, [9], [10] tried to optimize the efficiency of taxi drivers searching for customers, e.g., minimizing the average searching time. Different from the refueling process of gas FHVs, which usually takes about 3-5 minutes, the charging process of EFHVs typically lasts for half an hour to two hours even with fast chargers [11]. In addition, the electricity price is varying in different hours of the day, while the gasoline price is usually constant during a day. These long charging times and dynamic charging pricing lead to very different EFHV drivers' behaviors (where or when to charge) and incentives (whether to follow a recommendation). Although the energy level can be naively considered as a constraint of existing solutions to address the charging problem (e.g., if the energy level of an EFHV is lower than a threshold, the EFHV is set to be offline and removed from the system, it is similar to passenger searching), the key challenge is to decide which charging station the EFHV should go. The charging scheduling decisions are related to many factors like real-time traffic conditions, status of charging stations and charging prices, which should be considered for reducing the charging idle time and charging costs (e.g., avoiding overcrowded charging stations). Hence, the existing solutions for vehicle relocation or passenger searching [9], [10] are not suitable to the fairness-aware

- *G. Wang and L. Jiang are with the Department of Computer Science, Florida State University, Tallahassee, Florida, 32306, USA. (Email: guang@cs.fsu.edu)*
- *S. He and F. Miao are with the Department of Computer Science & Engineering, University of Connecticut, Storrs, CT, 06268, USA.*
- *S. Wang is with the School of Computer Science Engineering, Southeast University, Nanjing, Jiangsu, 211189, China.*
- *F. Zhang is with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong, 518055, China.*
- *Z. Dong is with the Department of Computer Science of Wayne State University, 5057 Woodward Ave., Room 14109.3 Detroit MI 48202, USA.*
- *D. Zhang is with the Department of Computer Science, Rutgers University, New Brunswick, NJ, 08901, USA.*

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2023.3326676

vehicle displacement problem well.

Recently, some works have been done to understand and improve the *charging efficiency of EFHVs* [5], [11], [12], but almost all of them mainly focus on optimizing charging idle time reduction instead of optimizing drivers' profits due to a lack of detailed transaction data from FHV fleets. The underlying assumption by the existing works is that optimizing charging idle time will prolong the EFHVs' operation time. However, we found that *charging idle time reduction does not necessarily indicate the prolonged time for serving passengers to maximize profit because some drivers may need to spend more time to seek passengers after charging in some regions with shorter charging waiting time but lower passenger demand* (as shown in Section 2.3).

In this paper, by working with an e-taxi agency, we utilize its detailed proprietary transaction data along with GPS data to improve the EFHV fleet's profit efficiency by designing a new vehicle displacement system called `FairMove`, which balances two relationships: *passenger demand vs. EFHV supply*, and *vehicle charging demand vs. charging station supply*. For the EFHV agency, its goal for vehicle displacement is to jointly optimize the overall profit efficiency of the EFHV fleet; whereas for the drivers, their incentive to follow vehicle displacement is to enable profit fairness among them. As a result, the key objective of our EFHV displacement is fleet-wide joint optimization of profit efficiency and profit fairness.

However, the EFHV displacement with this objective is challenging due to possible conflicting relationships (e.g., balancing future passenger demand and supply vs. balancing future EFHV charging demand and supply), and many confounding factors (e.g., individual drivers' charging behaviors like spatiotemporal charging preference, time-variant charging pricing, and individual-level fairness). To address these challenges, in this paper, we propose a deep reinforcement learning (DRL)-based approach (i.e., Fairness-Aware Multi-Agent Actor-Critic (FAMA2C for short) to learn the sophisticated EFHV displacement policy.

FAMA2C has three key advantages for the EFHV displacement compared to existing methods: (i) we integrate fairness into our algorithm for a fairness-aware RL algorithm, e.g., our reward function considers both the self-profit efficiency and the fairness, so each agent will not only maximize its own profit when it learns the policy but also cooperate with other agents to improve the overall profit fairness in a holistic view. To our best knowledge, this is the first RL algorithm with fairness consideration for fleet management of large-scale EFHVs. Compared to conventional actor-critic RL algorithms that typically target one single reward, our algorithm needs to balance weighted rewards with two possibly conflicting objectives. The weighted reward is also challenging to be optimized by naïve critic-actor RL algorithms. (ii) Our FAMA2C is adaptive to a highly dynamic and uncertain environment, which considers different practical factors and is fit for large-scale agent applications due to our centralized training and decentralized execution design. The centralized training process can motivate multiple agents to learn globally coordinated and cooperative policies. For example, the centralized critic is augmented with extra information such as actions and states of other agents (i.e., interacts with others), which means the critic explicitly uses extra information to ease the training process and adapts to the dynamic environment. During the execution process, each EFHV's local observation is taken as the input without requiring complete information in the training phase, and then the decentralized policy networks can output the actions with an efficiency and flexibility guarantee. (iii) We extend the standard advantage function with contextual information to address the high variability of the value function in our algorithm. (iv) FAMA2C maximizes the long-term reward of a sequence of decisions for profit efficiency and profit fairness improvement.

In particular, the key contributions of this paper include:

- We conduct an extensive data-driven analysis based on real-world multi-source data, from which we found some novel insights: (i) Charging time reduction does not necessarily indicate the prolonged time for serving passengers since some drivers may need to spend more time to seek passengers after charging in some regions with low passenger travel demand. (ii) The potential profits for serving passengers after charging in different stations may also be different, which is highly dynamic in both spatial and temporal dimensions. (iii) Prolonged charging time of EFHVs compared to the refueling processes of gas FHVs causes some real-world issues, e.g., intensive charging peaks and long charging wait time in some time slots induced by the time-varying charging pricing.
- Based on the data-driven insights, we design a new fairness-aware displacement system called `FairMove` to improve the overall profit efficiency and profit fairness for EFHV fleets by a FAMA2C approach. `FairMove` considers not only the operation behaviors of drivers and demand & supply but also the complicated charging processes (e.g., time-varying charging pricing, and intensive charging peaks). In addition, the time for seeking a passenger after charging and trip length are also considered for a more accurate revenue estimation. Finally, both the operating revenues and charging costs are fed to the `FairMove` system to make fair-profit-oriented decisions, which has the potential to make the system more sustainable.
- We implement and extensively evaluate our `FairMove` based on multi-source data from the Chinese city Shenzhen, including GPS records and transaction records from 20,130 EFHVs. The experimental results show our `FairMove` effectively increases the profit efficiency of the EFHV fleet by 26.9%, improves the profit fairness of EFHV drivers by 54.8%, and reduces the cruise time and idle time by 32.3% and 44.9% on average at the same time. It also improves the charging station utilization fairness by 38.4%.

A preliminary version of this work has been published in the research track of the 37th IEEE International Conference on Data Engineering (ICDE) as a full paper [13]. In this journal version, we have made a set of extensions and new contributions to enhance the conference version. (i) In the conference version, we mainly focus on the electric taxi fleet management, and we generalize it to profit and fairness joint optimization of electric for-hire vehicles in the journal

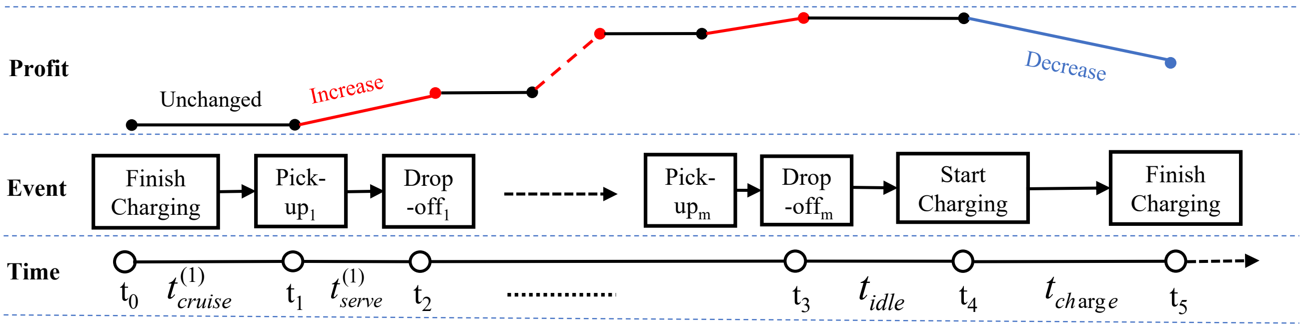


Fig. 1. Mobility decomposition of EFHVs.

version, which makes our system design more comprehensive. (ii) We extend the standard advantage function in the conference version with contextual information, which shows better performance to address the high variability of the value function. (iii) We show the framework and details of the designed FAMA2C algorithm. (iv) We define new metrics (e.g., Percentage Reduction of Charging Cost) and perform extensive additional experiments to further evaluate the performance of the proposed algorithm. (v) We add the convergence analysis of the displacement method to show the efficiency of our FairMove. (vi) We further study the system performance under different driver participation rates. (vii) We investigate the fairness of charging station utilization with different displacement strategies. (viii) We further utilize some off-policy evaluation methods to validate the learned policy against real-world data.

2 Data and Motivation

2.1 Data Description

All EFHVs in our dataset are the same vehicle model, i.e., BYD e6, whose battery capacity and maximum traveling distance are 80 kWh and 400 km, respectively [14]. There are five datasets used in our work, i.e., GPS data, transaction fare data, charging station data, urban partition data, and time-variant electricity rates data. The detailed information of the five datasets is shown as follows.

(i) GPS data includes vehicle IDs, real-time coordinates (i.e., longitudes and latitudes), time stamps, directions, speeds, and passenger indicator.

(ii) Transaction fare data includes vehicle IDs, the pickup and drop-off times, the pickup and drop-off coordinates (i.e., longitudes and latitudes), operating distances, cruising distances, and fares.

(iii) Charging station data includes station IDs, station names, coordinates (i.e., longitudes and latitudes), and the number of fast charging points in each station. There are 123 charging stations deployed in Shenzhen for EFHVs only in December 2019.

(iv) Urban Partition Data describes the urban partition for the population census of the Chinese city Shenzhen, which is provided by the Shenzhen government. There are 491 regions, and each region has a region ID and longitudes & latitudes of its boundary.

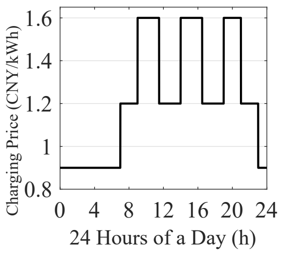


Fig. 2. Charging prices of EFHVs in Shenzhen.

(v) Charging Pricing Data. Many cities have time-variant charging pricing (similar to the time-variant electricity pricing), which breaks up 24 hours of a day into several intervals and charges a different price for each interval [15]. The rates in Shenzhen are divided into three types, i.e., off-peak prices (low rates), semi-peak prices (medium rates, also called flat rates), and peak prices (high rates), and the corresponding charging rates are 0.9, 1.2, and 1.6 CNY/kWh, respectively. The time-variant charging pricing in Shenzhen is shown in Fig. 2.

2.2 Mobility Decomposition of EFHVs

We decompose the mobility of EFHVs from three dimensions (i.e., time, event, and profit), as shown in Fig. 1, where t_0 to t_5 represents the activities of an EFHV during two consecutive charging events.

(i) At the time t_0, an EFHV finishes a charging event, and then it will cruise to find passengers to serve. At t_1, the EFHV picks the first passenger up and drops the passenger off at time t_2. We define the time for seeking a passenger as the **cruise time**, and the time for serving a passenger (onboard) as the **service time**. Specifically, we define the time duration $t_1 - t_0$ as the *first cruise time* $t^{(1)}_{cruise}$, and the time duration $t_2 - t_1$ as the *first service time* $t^{(1)}_{serve}$. During the *cruise time*, the EFHV neither has passengers on board nor charges, so the profit remains unchanged. During the *service time*, the EFHV's profit will increase with passengers on board. The profit is typically a function of time and distance.

(ii) After serving the first passenger, the EFHV will continue to cruise and serve the $2^{nd}, 3^{rd}, \ldots, m^{th}$ passenger, and the EFHV's profit keeps increasing during this period. After dropping the m^{th} passenger off, the energy level of this EFHV decreases to a threshold, so it will start to seek a charging station to charge at time t_3. We define time duration $t_3 - t_0$ as the **operation time** T_{op}, which equals to $T_{cruise} + T_{serve}$, where $T_{cruise} = \sum_{i=1}^{n} t^{(i)}_{cruise}$ and $T_{serve} = \sum_{i=1}^{n} t^{(i)}_{serve}$. The profit of the EFHV will increase during T_{serve} for serving passengers.

(iii) Due to some real-world issues (e.g., inadequate charging resources and intensive charging peaks), the EFHV may need to wait for a while to get an available charging point. Then at time t_4, there is an available charging point, so the driver will plug in the charger and charge the EFHV. We define time duration $t_4 - t_3$ as the **idle time** T_{idle} since the EFHV neither operates nor charges. The profit of the EFHV remains unchanged during the *idle time*.

(iv) After plugging in a charger, the EFHV will start to charge, and it finishes the charging event at time t_5. We

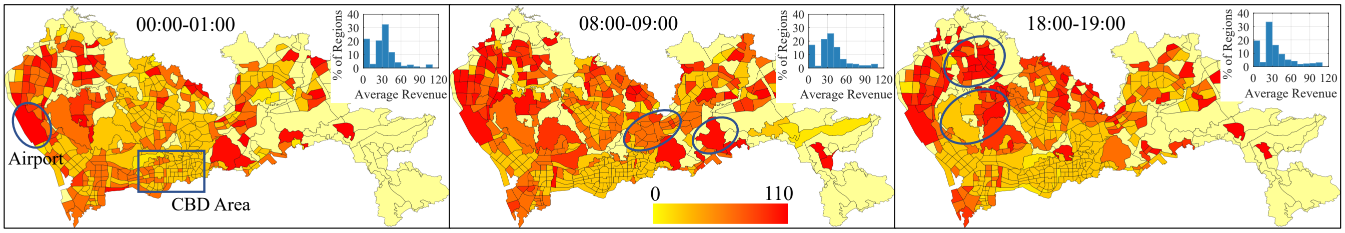


Fig. 3. Average per-trip revenue (CNY) in different regions during different hours of a day.

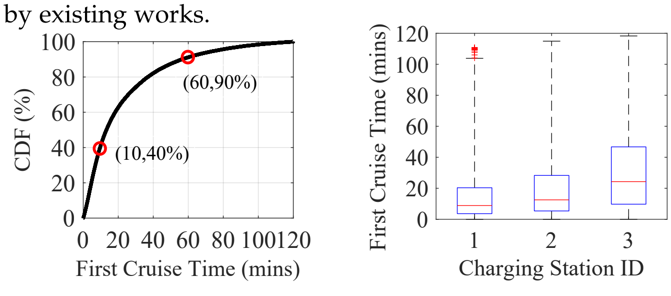
define the time duration with a charger plugin $t_5 - t_4$ as the **charge time** T_{charge}. During this time period, the profit of the EFHV will decrease due to energy replenishment.

(v) We define the time duration between two sequential charging events $t_5 - t_0$ as a **working cycle** T_{cycle} of an EFHV, which equals to $T_{op} + T_{idle} + T_{charge}$. Hence, during a long time period (e.g., one week), there will be a set of *working cycles* for each EFHV. In this paper, we focus on the long-term (e.g., weekly) profit fairness of EFHVs instead of the short-term profit, which also has the potential to achieve a higher overall profit efficiency for EFHV fleets.

2.3 Motivation By Data-Driven Findings

Based on our multi-source real-world data and the above definitions, we conduct in-depth data-driven analysis using one-month e-taxi data to show the uniqueness and motivation of our EFHV displacement design. In particular, we provide the following findings:

(i) *Idle time* reduction does not necessarily indicate the prolonged time for serving passengers since some EFHVs may need to spend more time seeking passengers after charging in some regions with low passenger travel demand. As shown in Fig. 4(a), we found 40% of EFHVs can find their first passengers after charging in 10 minutes, but there are still 10% of EFHVs need to cruise over an hour to find their first passenger after charging. In addition, the *first cruise time* $t^{(1)}_{cruise}$ is also different when charging in different stations. Fig. 4(b) shows the *first cruise time* of EFHVs after charging in three different charging stations. The three charging stations are located in different areas of the city, and there are different numbers of charging points in each station. We found that the *first cruise time* of the EFHVs has large differences after charging in different stations. Hence, the charging station selection not only impacts the *idle time* but also has influences on the *first cruise time* $t^{(1)}_{cruise}$. However, this finding has not been revealed and considered by existing works.



(a) Time distribution (b) Three charging stations

Fig. 4. First cruise time distribution and at three charging stations.

(ii) The potential revenue for serving passengers after charging may also be different at different time slots and stations, which is highly dynamic in both spatial and temporal dimensions. Fig. 3 shows a visualization of the average per-trip revenue in different regions during late night (00:00-01:00), morning rush hour (08:00-09:00), and evening rush hour (18:00-19:00). The dark red means higher average per-trip revenue (i.e., more long trips) in these regions, and light yellow means lower average per-trip revenue in these regions. We found the average per-trip revenue has a large gap between different regions across the city, ranging from several CNY to over 100 CNY. For example, the per-trip revenue in the airport region is always high, but it is very low in some suburban areas. In addition, we found the average trip length in a region may change during the day. We also quantify the average per-trip revenue in the 491 regions, which can be seen from the right upper corner of Fig. 3. We found that there are more regions with low prices per trip during the late night, but more regions with high prices per trip during rush hours.

Certainly, the passenger travel demand and supply in different regions are also different, so the probability to pick up a passenger is also different, which is usually considered by existing works. However, existing works rarely consider the revenue from a trip, which will also directly impact EFHVs' revenue. Hence, in this paper, we consider not only the demand and supply of EFHVs but also the potential revenue for serving passengers for displacement, which lays a foundation for EFHVs' revenue fairness.

(iii) Inequal profit efficiency of the EFHV fleet, which could be potentially improved by a centralized displacement system. As shown in Fig. 5, we found 20% of EFHVs' hourly profit efficiency is lower than 36, and there is also 20% of EFHVs' hourly profit efficiency is higher than 51, which means there is a huge profit gap between EFHVs, resulting in the profit of high-efficient drivers will be 42% higher than the low-efficient drivers. With people paying more attention to fairness and equity, such a large profit gap potentially hurts some drivers' daily life and makes them unsatisfactory. Hence, it is necessary to have a fairness-aware displacement system to improve EFHVs' profit fairness in the fleet without damaging the overall profit efficiency of the fleet.

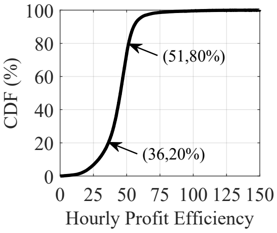


Fig. 5. Hourly profit efficiency.

In summary, based on our data-driven observations, we found the EFHV displacement problem is also different from the existing charging scheduling/recommendation since (i) the *idle time* reduction does not necessarily indicate more time for serving passengers. (ii) The *first cruise time* of the EFHVs has large differences after charging in different stations. (iii) Not only the probability of picking up passengers impacts EFHVs' revenue but also the trip length has a huge impact on it, which we could also consider improving the profit fairness for the EFHV fleet. (iv) It is necessary to design a fairness-aware displacement system for EFHV fleets to improve their profit efficiency and fairness, but it may not be achieved by existing solutions for conventional gas FHVs.

3 FairMove Displacement System Design

3.1 Key Idea of FairMove

The key idea of our FairMove is that we formulate the EFHV displacement problem as a large-scale sequential decision-making problem since the displacement decisions for EFHVs are sequential and highly repetitive, where each decision corresponds to scheduling an available (vacant) EFHV to a region or a charging station. There are multiple, possibly conflicting objectives in our displacement system, e.g., improving the profit efficiency of the EFHV fleet and reducing the profit unfairness between all EFHVs. In this work, we balance the profit efficiency and profit fairness by a weighted parameter to achieve the optimal displacement.

However, it is challenging to design an effective displacement strategy for EFHV fleets that can adapt to an environment involving dynamic demand & supply and complicated charging behaviors as shown in the above data-driven investigation. One major issue is that changes in a displacement decision will impact future demand & supply, and it is challenging for supervised learning approaches to capture and model these real-time changes. Inspired by successful applications in intellectually challenging decision-making problems (e.g., the game of Go [16], urban crowd-sensing [7], computation offloading [17], mobile edge computing [18]), in this paper, we try to target the EFHV displacement problem by deep reinforcement learning (DRL) based methods, which combine the advantages of Deep Neural Networks (DNNs) and Reinforcement Learning (RL) and has the capability of handling high-dimension data and highly dynamic environment features.

3.2 Problem Statement

Definition 1. (Profit Efficiency) Profit efficiency (PE) denotes the per unit time profit earned by an EFHV during its on-duty time in a period Γ (e.g., a week). The on-duty time of an EFHV includes a set of *working cycles*. Each *working cycle* consists of three components, i.e., operation time T_{op}, idle time T_{idle}, and charging time T_{charge}. The calculation method of the Profit efficiency of each EFHV can be represented as Equation 1.

$$PE = \frac{\text{Revenue} - \text{Costs}}{\sum_{k=1}^{z} T_{cycle}^{(k)}} = \frac{\sum_{i=1}^{m} R_{trip}^{(i)} - \sum_{j=1}^{n} C_{charge}^{(j)}}{\sum_{k=1}^{z}\left(T_{op}^{(k)} + T_{idle}^{(k)} + T_{charge}^{(k)}\right)} \quad (1)$$

where PE denotes the *Profit Efficiency* of an EFHV in a period Γ. *Revenue* and *Costs* denote the total revenue earned from serving passengers and the operation costs during a period Γ by the EFHV. m, n, and z denote the number of trips served by the EFHV, the number of charging events of the EFHV, and the number of working cycles of the EFHV during Γ, respectively. $R_{trip}^{(i)}$ is the revenue for serving i^{th} trip, $C_{charge}^{(j)}$ is the charging cost for j^{th} charging event. $T_{op}^{(k)}$, $T_{idle}^{(k)}$, $T_{charge}^{(k)}$ are the operation time T_{op}, idle time T_{idle}, and charge time T_{charge} of kth *working cycle*, respectively.

In this paper, since we define a *working cycle* as the time between two charging events, $z = n$ in Equation 1, and the operation time $T_{op}^{(k)}$ is equivalent to the sum of cruise time and serve time, i.e., $T_{cruise}^{(k)} + T_{serve}^{(k)}$. In addition, the charging costs is a function of the time-varying charging pricing and the charge time, so we describe the charge time of j^{th} charging event $T_{charge}^{(j)}$ as a three-dimensional vector $T_{charge}^{(j)} = \left[T_p^{(j)}, T_f^{(j)}, T_o^{(j)}\right]$, where $T_p^{(j)}$, $T_f^{(j)}$, and $T_o^{(j)}$ denotes the time in peak, flat, and off-peak charging pricing hours of the j^{th} charging event. Similarly, we also describe the time-varying charging pricing as a three-dimensional vector $\lambda = [\lambda_p, \lambda_f, \lambda_o]$, where λ_p, λ_f, λ_o denote the charging prices during peak, flat, and off-peak hours, respectively (as shown in Fig. 2). Hence, we convert Equation 1 into Equation 2 to calculate the *profit efficiency* of an EFHV.

$$PE = \frac{\sum_{i=1}^{m} R_{trip}^{(i)} - \sum_{j=1}^{n}\left(\lambda \cdot T_{charge}^{(j)}\right)}{\sum_{j=1}^{n}\left(T_{cruise}^{(j)} + T_{serve}^{(j)} + T_{idle}^{(j)} + T_{charge}^{(j)}\right)} \quad (2)$$

Definition 2. (Profit Fairness) It is typically challenging to define the fairness as different people may have different perceptions of fairness [19]. In addition, the fairness definition would also be different in different scenarios [20]. Hence, to better understand EFHV drivers' perceptions of fairness and define the profit fairness of EFHVs properly, our team has conducted a set of interviews with Shenzhen e-taxi drivers and asked them related questions. We found almost all e-taxi drivers thought it is fair when their profits are proportional to their working time. Motivated by this, in this paper, we define the *Profit Fairness PF* of an EFHV fleet as the variance of *profit efficiency* of all EFHVs in the fleet, which is denoted as Equation 3, so smaller PF means fairer for an EFHV fleet.

$$PF = \frac{1}{N}\sum_{k=1}^{N}\left(PE^{(k)} - \overline{PE}\right)^2 \quad (3)$$

where N is the total number of EFHVs in an EFHV fleet. $PE^{(k)}$ is the *profit efficiency* of the k^{th} EFHV. $\overline{PE}$ is the average *profit efficiency* of all EFHVs in the fleet, i.e., $\overline{PE} = \frac{1}{N}\sum_{k=1}^{N} PE^{(k)}$.

In this paper, we tackle the displacement problem for a large-scale available (i.e., vacant) EFHV fleet when considering both serving passengers and charging. The objectives of our displacement are three-fold: (1) Improving the overall *profit efficiency* PE of all EFHVs in the fleet during a period of Γ. (2) Enhancing the *profit fairness* of the EFHV fleet PF over Γ. (3) Tradeoff between the *profit efficiency* and *profit fairness*. A spatial-temporal illustration of the problem will be shown in Fig. 6. For the spatial partition, we utilize the urban partition data described in Section 2 to represent the map, which splits the Shenzhen city into 491 regions. Our partition is similar to the grid-based methods (e.g., square-grid [3], [5] and hexagonal-grid [21]), but our partition is more practical as it considers the geological structure of the city (e.g., a mountain or a lake will be partitioned in a single region). For the temporal partition, we split the duration of a day into T time slots. At each time slot, there are different numbers of passenger demands sporadically appearing in each region, and those passengers will be served by the available EFHVs in the same region. The role of the displacement system is to decide which region or charging station each vacant EFHV should go in each time slot to maximize the long-term *profit efficiency* and *profit fairness* of the EFHV fleet.

3.3 Problem Formulation

Formally, we model the EFHV displacement problem as a *multi-agent Markov decision process* $\mathcal{G}$ for N agents, which is defined by a five-tuple $\mathcal{G} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \beta)$, where $\mathcal{S}$ is the set of states; $\mathcal{A}$ is the joint action space; $\mathcal{P}$ is transition probability function; $\mathcal{R}$ is the reward function; and β is a discount factor. Markov decision process is typically used to model sequential decision-making problems.

In a Markov decision process, an agent observes the state from the environment and executes action based on the observed state. Then, the environment turns to the next state and feeds a reward for the action back to the agent. The agent evolves along with this interaction with the environment. The goal of a Markov decision process is to produce an optimal policy that maximizes the agent's expected cumulative rewards. The state-action value function of the agent can tell us how good a decision it makes at a particular location and time under given environment contexts with respect to the long-term objectives. By estimating the state-action value function, the value-based methods give the optimal action at each interaction step. The detailed definitions of the *multi-agent Markov decision process* $\mathcal{G}$ in our `FairMove` displacement system are shown as below.

Agent Set: We consider each available (i.e., vacant) EFHV as an agent, and EFHVs within the same spatial-temporal partition are homogeneous, i.e., EFHVs in the same region or charging station during the same time slot are considered as homogeneous agents (where agents have the same states), and the number of agents (available EFHVs) N_t is changing over time.

State $\mathcal{S}$: The state of an available EFHV k $s_t(k) \in \mathcal{S}(k)$ consists of a two-dimensional vector indicating its specific spatiotemporal status from both the local view and global view. And $k = 1, 2, ..., N_t$, $\mathcal{S}(k)$ is the state space for the available EFHV k. The joint state of all available EFHVs in the time slot t is denoted as $\mathbf{s}_t \in \mathcal{S} = \mathcal{S}(1) \times \mathcal{S}(2) \times \dots \mathcal{S}(N_t)$. We discrete one day into a set of T time slots. And we divide the city into a set of R regions and C charging stations, (i.e., $R \cup C$ = the whole city; $R \cap C = \emptyset$). We define a local-view state of an EFHV, $s_{t,lo} = [t, l] \in \mathcal{S}_{lo}$, where $t \in T$ is the time index (i.e., which time slot), and $l \in R \cup C$ is the location index (i.e., which region or charging station) where the EFHV is in. In this case, the finite local state space $\mathcal{S}_{lo}$ is a Cartesian product of the set of time slots and the set of regions + charging stations, i.e., $\mathcal{S}_{lo} = T \times (R \cup C)$ and the number of states is $|\mathcal{S}_{lo}| = |T| \times |(R \cup C)|$. The EFHVs in the same partition (region or charging station) in a time slot has the same local state. We also define a global-view state $s_{t,go}$, which is shared by all available EFHVs in the time slot t. The global-view state includes three different spatiotemporal features: (i) the number of available EFHVs in each region; (ii) the number of unoccupied charging points in each charging station; and (iii) the expected number of passengers in each region at the next time slot, which is predicted with historical and real-time data. In this work, we utilize the XGBoost model [22] to predict passenger demand, which seems simple but can achieve better performance compared to many state-of-the-art deep learning methods based on our experiments. The basic idea of our method is that the historical passenger demand information (e.g., the number of passengers in different time slots on different days) and the real-time traffic condition and meteorological data (e.g., weather conditions and temperatures) are considered as the input features of our prediction model. The global-view state $s_{t,go}$ will update in each time slot. Finally, the state of each available EFHV k during the time slot t can be represented as $s_t(k) = [s_{t,lo}(k), s_{t,go}(k)] \in \mathcal{S}(k)$ and EFHVs in the same partition and time slot have the same state.

Action $\mathcal{A}$: The action space of an EFHV k where $k = 1, 2, ..., N_t$, $\mathcal{A}(k)$ specifies where it is able to arrive at the next time slot. There are three types of actions in our EFHV displacement setting. (i) The first type of action is staying in the current region. (ii) The second type of action is displacing the EFHV to another adjacent region in the direction of the potential nearest passenger. We assume the travel conditions on road segments on the same day of different weeks typically are similar, so we utilize historical data to estimate the average speed of each road segment in different time slots on different days. This is a commonly adopted method in many existing papers [23], [24]. (iii) The third type of action is charging in a charging station. For the second type of action, each EFHV can go to its adjacent regions and EFHVs in different regions have a various number of neighbor regions, so they have different dimensions of action spaces. The EFHVs in the same region have the same action space. For the third type of action, we consider the nearest five charging stations for each EFHV to reduce the action space. The charging action is decided by the energy level of the EFHV k, which is estimated by the initial energy level and energy consumed for operating [20]. The energy consumption rate would be different if the EFHVs are made of various models with different characteristics, but the method is still applicable if we use separate energy consumption calculation formulas for them. At time slot t, the available EFHV k takes an action $a_t(k) \in \mathcal{A}(k)$, forming the joint action $\mathbf{a}_t \in \mathcal{A} = \mathcal{A}(1) \times \mathcal{A}(2) \cdots \times \mathcal{A}(N_t)$, which induces a transition in the environment according to the state transition function $\mathcal{P}(s_{t+1} | s_t, \mathbf{a}_t) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1)$.

Reward $\mathcal{R}$: Reward reflects the immediate feedback of the action in a specific state, but the maximizing immediate reward is not equivalent to the goal. However, reward usually determines the optimization goal of the displacement system, which usually utilizes rewards to guide the learning process. A typical measurement is to estimate the difference in the accumulated reward between with and without following the displacement system's decisions. We define three types of immediate rewards in our EFHV displacement scenario: (i) positive rewards for serving passengers; (ii) rewards for cruising; and (iii) negative rewards for charging. Note that both the positive rewards and negative rewards are nondeterministic as the positive rewards are mainly decided by the trip length, and the negative rewards are decided by the charging time t_{charge} and time-variant charging prices. We consider the EFHVs will always serve the nearest passengers, and the passengers in a region will always be served by the vacant and available EFHVs. We define the reward as 0 when the EFHV is cruising since there is no direct transaction. When the energy status of an EFHV decreases to a certain threshold η (e.g., 20%), the EFHV should go to charge. Even though the immediate reward for charging is negative, EFHVs cannot operate and serve passengers without energy, so running out of battery will cause no reward in the future. Hence, the charging action will also benefit the long-term positive reward for the EFHV, which means the impact of charging can be spread to its future states. Considering both the profit efficiency and fairness, the final reward of the EFHV k can be represented by Equation 4. $PE(k,t)$ is the profit efficiency of EFHV k in the time slot t (i.e., in regard to state $s_t(k)$ and action $a_t(k)$), $PF(t)$ is the profit fairness of all active EFHVs in the time slot t. Since the PF in Equation 3 indicates the unfairness of the system, so we have the minus here to maximize the profit fairness.

$$r(s_t(k), a_t(k)) = \alpha \cdot PE(k,t) + (1-\alpha) \cdot (-PF(t)) \tag{4}$$

To balance the profit efficiency and profit fairness, a real-valued parameter $\alpha \in [0,1]$ is leveraged to control how much we emphasize the profit efficiency and profit fairness of EFHVs in the fleet. As a boundary case, with $\alpha = 1$, we only explicitly maximize the profit efficiency for the fleet, while ignoring the level of unfairness among all EFHVs. With $\alpha = 0$, we only explicitly maximize the profit fairness for all EFHVs in the fleet, while ignoring the profit efficiency. Therefore, the Equation 4 can be converted to Equation 5.

$$\begin{aligned} r(k,t) = \alpha \cdot & \frac{\sum_{i=1}^{m} R_{trip}^{(i)}(k,t) - \sum_{j=1}^{n}\left(\lambda \cdot T_{charge}^{(j)}(k,t)\right)}{\sum_{j=1}^{n}\left(T_{cruise}^{(j)}(k,t) + T_{serve}^{(j)}(k,t) + T_{idle}^{(j)}(k,t) + T_{charge}^{(j)}(k,t)\right)} \\ & + (1-\alpha) \cdot \left(-\frac{1}{N_t}\sum_{h=1}^{N_t}\left(PE(h,t) - \overline{PE(t)}\right)^2\right) \end{aligned} \tag{5}$$

where we use $r(k,t)$ instead of $r(s_t(k), a_t(k))$ for short. Our reward function considers both self-profit efficiency and fairness, so every EFHV is not only maximizing its own profit when they learn the policy but also cooperating with each other to maximize the profit fairness, when every EFHV is trying to maximize its expected discounted accumulated rewards $\mathbb{E}\left[\sum_{i=0}^{\infty}\beta^i r(k,t+i)\right]$.

State transition function $\mathcal{P}$ is defined as a mapping $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1)$. $p(\mathrm{s}_{t+1}|\mathrm{s}_t, \mathrm{a}_t)$ denotes the probability of transition to s_{t+1} given a joint action a_t in the current state s_t.

Discount factor β essentially determines how much the reinforcement learning agents care about rewards in the distant future relative to those in the immediate future. The value of β is typically selected from $[0,1)$, so the final expected reward in the infinite horizon will be convergent and bounded to a finite number. If $\beta = 0$, the agent will be completely myopic and only learn about actions that produce an immediate reward without considering the future reward.

To make the above definitions more clear, we show an example of the EFHV displacement process under the formulation of *Markov decision process* from spatial and temporal dimensions in Fig. 6. At time slot $t = 0$, EFHV 1 is displaced to stay at Region 333 r_{333} by action $\mathrm{a}_0(1)$, and another EFHV 2 is displaced from r_{230} to r_{333} by action $\mathrm{a}_0(2)$. An EFHV 3 is displaced to charge in charging station 1 CS_1, and there are available charging points in CS_1. At time slot $t = 1$, EFHV 1 and EFHV 2 arrive at r_{333}, and then they are considered as homogeneous for serving passengers in r_{333}. If there are two trips (e.g., passenger 1 valued for 20 and passenger 2 valued for 30 CNY) appearing in the region, these two passengers will be assigned to the two EFHVs equally random, e.g., the passenger 1 is served by the EFHV 1 with an immediate reward 20 and passenger 2 is served by the EFHV 2 with an immediate reward 30. This setting can reduce the complexity of the vehicle displacement algorithm, which is also used by existing research [25]. Since EFHV 3 starts to charge in CS_1, it will have a negative immediate reward, which is decided by how long it charges and the charging pricing. As shown in our reward function, if an EFHV's current profit efficiency is higher than the global average profit efficiency (e.g., EFHV 2), it will receive a weakened reward to reduce its advantage in competing for new orders. In contrast, previously inefficient vehicles (e.g., EFHV 3) will gain additional compensation in the probability to receive orders, which means those who have received unfair treatment in the previous decision-making process will be compensated in the future to ensure fairness. The fairness definition in our problem is based on a long time period (e.g., one day, one week, or one month) instead of a single time slot.

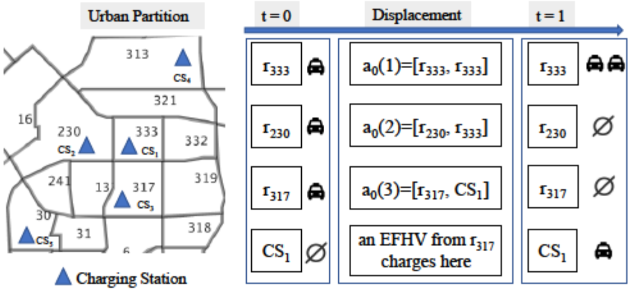


Fig. 6. Displacement process in spatiotemporal dimensions.

3.4 Fairness-Aware Multi-Agent Actor-Critic

In this paper, we propose a fairness-aware multi-agent actor-critic (FAMA2C) algorithm to solve the above-defined multi-agent problem for a large-scale EFHV displacement, which is a multi-agent policy gradient algorithm that iterates its policy to adapt to the dynamically evolving action space. The basic idea of the FAMA2C is that for each homogeneous agent (EFHVs within the same spatial-temporal partition), there are two networks, a policy network (i.e., Actor, which is utilized to output action) and a value network (i.e., Critic, which is leveraged to evaluate the performance of the policy network). EFHVs under different spatiotemporal conditions will use different policies. These policies are trained using different critics.

FAMA2C follows the paradigm of centralized training and decentralized execution. It means each agent learns a centralized critic, allowing agents to deal with the non-stationary environment, and each agent also has its own policy network. Besides, the centralized critic is augmented with extra information such as actions and states of other agents (i.e., interacts with others), which means the critic explicitly uses extra information to ease the training process and adapts to the dynamic environment. The information used during the training of the agents is not used in the execution process. When agents take actions using the well-trained policies, they only need local observations. Therefore, the execution process can be more efficient. We denote the total number of agents in the fleet as M. There are two key tasks for training the FAMA2C, i.e., (i) learning the parameters of the policy network θ_p^i and (ii) learning the parameters of the value network θ_v^i, where $i = 1, 2, ..., M$. For the notation convenience, we omit the superscript. Then both the Critic and Actor are parameterized with deep neural networks, and the parameters of the critic θ_v and actor θ_p are updated iteratively.

Since all agents' rewards contain the profit fairness, available EFHVs will cooperatively take actions for a fairness-aware optimal strategy. To improve the training efficiency, we adopt the offline training paradigm. Without loss of generality, the critic of a homogeneous agent is learned by minimizing the following loss function $\mathcal{L}(\theta_v)$ as shown in Equation 6.

$$\mathcal{L}(\theta_v) = \left(V_{\theta_v}(s_t) - V_{tg}(s_{t+1}; \theta_v^{'}, \pi)\right)^2 \tag{6}$$

Where θ_v denotes the parameters of the value network, and $V_{\theta_v}(s_t)$ is the estimated value of the value function under state s_t. $\theta_v^{'}$ denote the parameters of the target value network, and $V_{tg}(s_{t+1}(k); \theta_v^{'}, \pi)$ is the target value, which consists of the immediate reward and discounted estimated value function under the next state, as shown in Equation 7.

$$V_{tg}\left(s_{t+1}; \theta_v^{'}, \pi\right) = \sum_{a_t} \pi(a_t|s_t)\left(r_{t+1} + \beta V_{\theta_v^{'}}(s_{t+1})\right) \tag{7}$$

The parameter of the value network θ_v is updated by the gradient descent rule $\theta_v \leftarrow \theta_v + \lambda_1 \nabla_{\theta_v}\mathcal{L}(\theta_v)$, where λ_1 is the learning rate of the critic. The parameter of the policy network θ_p is updated using the gradient given by Equation 8.

$$\nabla_{\theta_p} J(\theta_p) = \nabla_{\theta_p} \log \pi_{\theta_p}(s_t, a_t)\left(r_{t+1} + \beta V_{\theta_v^{'}}(s_{t+1}) - V_{\theta_v}(s_t)\right) \tag{8}$$

Since the value function has high variability, we extend the standard advantage function [26] with contextual information to address it, which is given in Equation 9.

$$A(s_t, a_t, \mathrm{x}) = Q(s_t, a_t, \mathrm{x}) - V_{\theta_v}(s_t, \mathrm{x}) \tag{9}$$

where x denotes extra information, and we let the value function takes the actions of all agents and the global state information and outputs the estimated value function for each agent. $Q(s_t, a_t, \mathrm{x})$ is the state-action value function for action a_t in state s_t. $V_{\theta_v}(s_t, \mathrm{x})$ is the average value of that state, so this function tells us the improvement compared to the average the action taken at that state. $A(s_t, a_t, \mathrm{x}) > 0$ means the gradient is pushed in that direction, and $A(s_t, a_t, \mathrm{x}) < 0$ means the action does worse than the average value of that state.

Since
$$Q(s_t, a_t, \mathrm{x}) = r_{t+1} + \beta V_{\theta_v^{'}}(s_{t+1}, \mathrm{x}) \tag{10}$$

we combining Equation 9 with Equation 10 to obtain the estimation of the extended advantage function as shown in Equation 11, which is equivalent to the Temporal-Difference (TD) error [27] with extra information, so we can use an extra information-aware TD error as an estimation of the extended advantage function.

$$A(s_t, a_t, \mathrm{x}) = r_{t+1} + \beta V_{\theta_v^{'}}(s_{t+1}, \mathrm{x}) - V_{\theta_v}(s_t, \mathrm{x}) \tag{11}$$

Then all available EFHVs can cooperatively work for a fairness-aware optimal strategy. The details of the FAMA2C are shown in Algorithm 1.

ALGORITHM 1: FAMA2C for Displacement

```
Initialize the value network by randomly selecting
  policy network parameters $\theta_p^i$ and value network
  parameters $\theta_v^i$ where $i = 1, ..., M$.
for l =1 to the maximum iteration number do
    Reset the environment and obtain the initial joint
      states $\mathbf{s}_0$
    for each time slot $t \in [0, T)$ do
        for $i = 1$ to $M$ do
            Sample action $a_t(k)$ given $s_t(k)$ according
              to the policy $\pi_{\theta_p^i}$ for EFHV $k$, where
              $k \in N_i$, $N_i$ is the set contains all EFHVs
              in the same spatial-temporal partition
              and share the policy $\pi_{\theta_p^i}$;
        Execute the joint action $a_t(1) \times ... \times a_t(N_t)$
          and get the next state $s_t(1) \times ... \times s_t(N_t)$
        Store transitions
          $(s_t(k), a_t(k), r_{t+1}(k), s_{t+1}(k))_{k \in N_i}$ in the
          buffer $B_i$ for $i = 1, ..., M$.
    for j = 1 to a certain iteration number C do
        for $i = 1$ to $M$ do
            Sample a batch of transition from the
              buffer $B_i$
            Compute target value network $V_{tg}$ by
              Equation 7 and the advantage function
              $A(s_t, a_t, \mathrm{x})$ by Equation 9
            Update parameter of the value network $\theta_v^i$
              by minimizing the value loss function
              $\mathcal{L}(\theta_v^i)$ over the batch in Equation 6.
            Update the parameter of the policy
              network $\theta_p^i$ by $\theta_p^i \leftarrow \theta_p^i + \lambda_2 \nabla_{\theta_p^i} J(\theta_p^i)$,
              where $\nabla_{\theta_p^i} J(\theta_p^i)$ is calculated according
              to Equation 8.
```

3.5 Displacement Emulation

In this work, we design a data-driven emulation to verify the effectiveness of our system. In our emulation, the city is partitioned into urban regions as shown in Fig. 6. Based on historical passenger demand distributions and EFHVs' status from our real-world data, we train the model to learn optimal policies. As shown in Fig. 7, in each time slot (e.g., each round of displacement), there are the following steps: (1) Vehicle status and distribution update, which decides which EFHV is available and can be displaced, e.g., if an EFHV is vacant with enough energy, it is available for displacement, so the number of available EFHVs (agents) is changing over time. This setting is different from most of the previous studies related to multi-agent reinforcement learning, in which the number of agents is unchanged. It makes the design of the multi-agent learning algorithms more intractable; (2) Charging station status and traffic conditions update, which is learned from historical data at the same time. They provide information for EFHVs that need to charge and travel time to pick up passengers; (3) Passenger demand distribution generation, which provides information about which region that each EFHVs should be displaced to; (4) Computing states of EFHVs $\mathbf{s}_t$, which is computed as the input of displacement algorithm; (5) Generating policies using FAMA2C in Algorithm 1, which generate decisions for EFHVs that they should take; (6) Implementing displacement, then these available EFHVs will go to serve passengers or charge in the corresponding charging stations; (7) Computing rewards of EFHVs. Each EFHV will obtain a reward after performing an action. Then

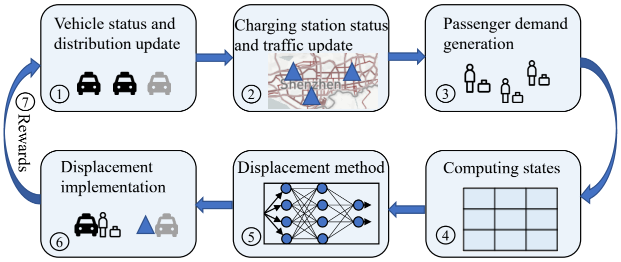

Vehicle status and distribution update
1
Charging station status and traffic update
2
Shenzhen
Passenger demand generation
3
Computing states
4
Displacement method
5
Displacement implementation
6
7
Rewards

Fig. 7. Displacement process in one time slot.

ALGORITHM 2: Displacement Emulation Process

```
Information of the historical passenger demand
  distributions and available EFHVs' distributions
for each time slot $t$ do
    (1) Updating the status and distribution of
      EFHVs based on real-world data. If the EFHV is
      serving passengers, it will be set as offline and
      cannot be displaced. If the EFHV is vacant, it
      will be set as online and can be displaced by the
      system.
    (2) Updating the status of charging stations and
      traffic.
    (3) Generating passenger demand distribution in
      this time slot by using real-world data.
    (4) Computing states $\mathbf{s}_t$ as input for the
      displacement algorithm.
    (5) Making the displacement decisions, i.e.,
      executing the joint action $\mathbf{a}_t$ according to the
      joint policies of the displacement algorithm,
      which decides which region or charging station
      each EFHV should go.
    (6) Assigning available EFHVs to serve
      passengers based on their distances, and
      scheduling low-energy EFHVs to charge.
    (7) Computing rewards of each EFHV.
```

In our deep reinforcement learning-based method, both the value function approximation networks and policy networks are three-layer networks, with 256, 128, and 64 nodes from the first to the last hidden layer. The activations of all hidden units are ReLu, while the output layers of the value function approximation networks and policy networks use Linear and Softmax activations, respectively.

4 Evaluation

4.1 Experimental Setup

In order to make our simulation closer to real-world scenarios, we make full use of our large-scale data for a systematic emulation. We first analyze our long-term large-scale GPS data and transaction data to learn real-world knowledge, e.g., dynamic passenger distribution, vehicle distribution, charging demand distribution, route selection distribution, and traffic conditions in different time slots of a day, etc. We then fuse the above real-world knowledge (e.g., demand distributions) as input to the emulator. We initialize each agent according to its location and time at the beginning of a day, and its actions afterward are determined by the emulator according to the fairness-aware displacement algorithm, so our design is more practical compared to other small-scale data-based designs.

Evaluation Data: To evaluate the effectiveness of our displacement system, one-month real-world data collected from an EFHV fleet in the Chinese city Shenzhen during December 2019 is utilized in this part. The data is from 20,130 e-taxis, which institute the largest full EFHV fleet in the world. The one-month EFHV data include 2.48 billion GPS records (247.9GB) and 23.2 million trip records (4084MB). In addition to EFHV data, the evaluation dataset also includes the metadata of 123 EFHV charging stations with over 5,000 charging points. The details of data formats have been introduced in Section 2.

Data Management: Due to the large size of our EFHV data, it requires significant efforts for efficient management, querying, and processing. Hence, we performed a detailed cleaning process to filter out the error, duplicate, and incomplete GPS and transaction data on a high-performance cluster with Spark and Hadoop, which was equipped with 80 TB of memory and 20 nodes. We train and test our model on a desktop with 32GB memory, 1TB HDD storage, and Intel Xeon CPU E5-1660 v3, installed with the latest Windows 10 and Python coding environment.

Baseline Setting: We compare our FAMA2C-based `FairMove` to the following baselines.

- **GT** is the **G**round **T**ruth, which is obtained from our real-world data. Based on our data, we have historical passenger distributions. We also inferred the charging events of the EFHVs according to the method in [20], and we calculated their cruise time, idle time, profit efficiency, and profit fairness, etc.
- **SD2** is the **S**hortest **D**istance based **D**isplacement [28]. In this setting, the EFHVs are always displaced to serve their nearest passengers or charge in the nearest charging stations no matter what regions the passengers and charging stations are, and it does not have a learning process for a long-term reward. Even though it is a naive method, it is very easy to implement in complicated real-world scenarios. One potential drawback is that some charging stations will be overcrowded in some time slots with this displacement method.
- **TQL** is the standard **T**abular **Q**-**L**earning [29], which is a widely used method for single-agent scenario. It estimates the expected total discounted rewards of

state-action pairs by learning a Q-function table with ϵ-greedy policy.

- **DQN** (**D**eep **Q**-**N**etwork) [30] is a popular method in deep reinforcement learning and has been previously applied to multi-agent settings. DQN learns the action-value function Q^* corresponding to the optimal policy by minimizing the loss: $\mathcal{L}(\theta) = \mathbb{E}_{s,a,r,s'}\left[(Q^*(s,a|\theta) - y)^2\right]$, where $y = r + \beta \max_{a'} \bar{Q}^*(s', a')$. where $\bar{Q}$ is a target Q function whose parameters are periodically updated with the most recent θ, which helps stabilize learning.
- **TBA** is the **T**rip **B**andit **A**pproach [9], which is also a reinforcement learning-based method. It is proposed in SIGSPATIAL Cup 2019. It adopts the REINFORCE rule [31] to update the policy. In this setting, EFHVs only know their own states and cannot communicate with each other, so they are purely competitive, and EFHVs will also be displaced to serve their nearest passengers before orders expire. They will also be displaced to charge in the nearest charging stations if they need to charge.
- **Care** is a charging and relocation recommendation system for EFHV drivers based on multi-agent reinforcement learning [3]. It jointly schedules EFHVs' charging and relocation decisions to maximize his long-term cumulative reward. A key difference between this work with our FairMove is that this system does not consider the fairness between drivers, which potentially causes biases and the unsustainability of the recommendation system.

Parameter Setting: The batch size of all deep learning networks is set to be 3500, and we utilize AdamOptimizer with a learning rate of 0.001. We set 10 minutes as a time slot, which is widely adopted by existing works [29], so the one day is divided into $T = 144$ time slots. For the discount factor, we select $\beta = 0.9$ to guarantee convergence. We set the weighted factor $\alpha = 0.6$ for the following experiments, and we will show the reason in Section 4.2.6. All the experiments are repeated 10 times to ensure the robustness of the results.

Evaluation Metrics: Our FairMove aims to optimize the *profit efficiency* and *profit fairness* of an EFHV fleet at the same time. According to Equation 2, an implicit indicator of improving EFHVs' the *profit efficiency* is the reduction of their total cruise time T_{cruise}, idle time T_{idle}, or charging duration in electricity pricing peak hours (i.e., reduction of charging costs). Hence, we utilize the following metrics to measure the system performance, including (i) *Percentage Reduction of Cruise Time (PRCT)*, (ii) *Percentage Reduction of Idle Time (PRIT)*, (iii) *Percentage Reduction of Charging Costs (PRCC)*, (iv) *Percentage Increase of Profit Efficiency (PIPE)*, and (v) *Percentage Increase of Profit Fairness (PIPF)*. We also show the impact of the parameter α on the system performance. We also conduct convergence analysis and investigate the impact of driver participation rate on system performance.

$$PRCT(D) = \frac{\sum_{i=1}^{M} T_{cruise}^{(i)}(G) - \sum_{i=1}^{M} T_{cruise}^{(i)}(D)}{\sum_{i=1}^{M} T_{cruise}^{(i)}(G)} \times 100\% \tag{12}$$

$$PRIT(D) = \frac{\sum_{j=1}^{Z} T_{idle}^{(j)}(G) - \sum_{j=1}^{Z} T_{idle}^{(j)}(D)}{\sum_{j=1}^{Z} T_{idle}^{(j)}(G)} \times 100\% \tag{13}$$

$$PRCC(D) = \frac{\sum_{k=1}^{N} Cost_k(G) - \sum_{k=1}^{N} Cost_k(D)}{\sum_{k=1}^{N} Cost_k(G)} \times 100\% \tag{14}$$

where $T_{cruise}^{(i)}(D)$ is the *cruise time* for ith trip under the displacement strategy D, which could be SD2, TQL, DQN, or FairMove (based on FAMA2C); M is the total number of trips served by the EFHV fleet. $T_{cruise}^{(i)}(G)$ is the *cruise time* for ith trip of the Ground Truth; $T_{idle}^{(j)}(D)$ is *idle time* of the jth charging events under displacement strategy D; Z is the total number of charging events of the EFHV fleet; $T_{idle}^{(j)}(G)$ is the *idle time* of the jth charging events of the Ground Truth; $Cost_k(D)$ is total charging cost of kth EFHV under the displacement strategy D; $Cost_k(G)$ is the *charging cost* of kth EFHV of the Ground Truth; N is the total number of EFHVs in the fleet; The Ground Truth is obtained by merging the GPS data, transaction data, and the charging station data.

$$PIPE(D) = \frac{\sum_{k=1}^{N} PE_k(D) - \sum_{k=1}^{N} PE_k(G)}{\sum_{k=1}^{N} PE_k(G)} \times 100\% \tag{15}$$

$$PIPF(D) = \frac{PF(G) - PF(D)}{PF(G)} \times 100\% \tag{16}$$

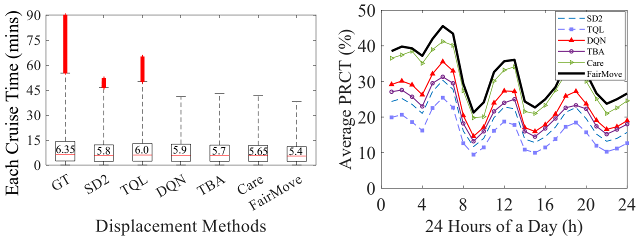
where $PE_k(D)$ is the *profit efficiency* of the EFHV k under the displacement strategy D, which can be SD2, TQL, DQN, or FairMove (based on FAMA2C); $PE_k(G)$ is the *profit efficiency* of the EFHV k without any external displacement; $PF(G)$ is the *profit fairness* of the Ground Truth; $PF(D)$ is profit fairness of the displacement strategy D.

4.2 Displacement Performance

4.2.1 Cruise Time Comparison

Since a key impact factor of profit efficiency of EFHVs is their *cruise time*, we compare our FairMove to other state-of-the-art baselines considering the cruise time reduction. Fig. 8(a) shows the cruise time distributions under different displacement methods. The medians of various methods in Fig. 8(a) are 6.35, 5.8, 6.0, 5.9, 5.7, 5.65, and 5.4, respectively. We found all methods reduce the cruise time for seeking passengers at different degrees compared to the ground truth due to their centralized management mode. The average value of the cruise time without other displacement methods is around 11.1 minutes, and it decreases to 7.5 minutes under our FairMove displacement. In addition to the decrease of the median value of the cruise time, its variance also becomes smaller with FairMove displacement, which could be induced by our fairness consideration. Fig. 8(b) shows the average PRCT for all trips during different hours of a day. We found FairMove achieves the best performance compared to other methods. Particularly, FairMove reduces over 40% of cruise time for EFHVs during 5:00-7:00, when there are few passengers and drivers need to cruise a longer time to find passengers without centralized displacement.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2023.3326676

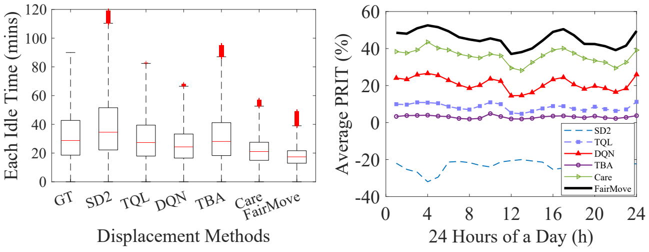


(a) Per-trip cruise time (b) PRCT distribution

Fig. 8. Cruise time under different methods.

In general, our FairMove achieves 32.3% of PRCT for each trip compared to the ground truth on average. The reason could be that deep reinforcement learning-based FairMove not only considers the short-term immediate benefits but also considers the long-term benefits. Care and DQN also achieve good performance with 28.8% and 23.6% of PRCT, followed by TBA and SD2 with 21.3% and 19.4% of PRCT compared to the ground truth. Since the cruise time reduction also potentially indicates the reduction of the average passenger waiting time, our FairMove can also potentially reduce passenger waiting time. Thus, some vehicles that used to move longer to pick up passengers can finish the order at a smaller cost and have a longer time to serve passengers. Therefore, the profit fairness is supposed to be increased when using FairMove.

4.2.2 Idle Time Comparison



(a) Per-charge idle time (b) PRIT distribution

Fig. 9. Idle time under different methods.

Since the idle time for charging will also impact the profit efficiency of the EFHVs, we also compare our FairMove to other state-of-the-art baselines considering the idle time reduction. Fig. 9(a) shows the idle time distribution for each charging event under different displacement methods. We found that our FairMove achieves the best performance, and 75% of the per-charge idle time is less than 22 minutes. However, SD2 prolongs the idle time since many EFHVs around charging stations will be displaced to the same charging stations, which causes long queuing in the overcrowded charging stations. Fig. 9(b) shows the average PRIT of all charging events during 24 hours of a day. We found our FairMove achieves the most PRIT during the high charging demand hours, e.g., 4:00-5:00 and 17:00-18:00, which potentially indicates our method can also benefit the charging issues for EFHVs, especially for addressing the intensive charging peaks. Therefore, FairMove can save drivers idle time for low-battery vehicles to serve more passengers, which may increase profit fairness by improving low-battery vehicles' revenue.

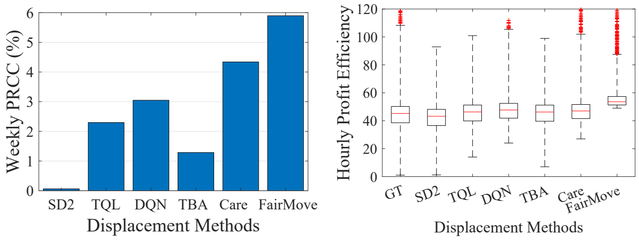
In general, our FairMove achieves 44.9% of PRIT for each charging event compared to the ground truth on average, as shown in Table 1. The reason would be that deep learning-based FairMove will choose the stations with the consideration of long-term benefits. Care also achieves good performance with 36.7% of PRIT. However, SD2 has a negative PRIT, which means it prolongs the idle time as many near EFHVs have been displaced to the same charging stations, resulting in long queuing in these stations. Although TBA may also cause some charging stations overcrowded, it achieves 3.1% of PRIT due to the long-term benefit consideration and potential cruise time reduction with the reinforcement learning method.

TABLE 1
Average percentage reduction of idle time (PRIT).

Methods	SD2	TQL	DQN	TBA	Care	FairMove
PRIT	-23.1%	8.4%	21%	3.1%	36.7%	44.9%

4.2.3 Charging Costs Comparison

Due to different displacement methods may impact EFHVs' operation patterns and time to charge, the charging costs and the number of charges may also be different. As shown in Fig. 10(a), we compare the weekly total charging cost of the EFHV fleet under different displacement methods with the metric PRCC. We found that the SD2 has a similar charging cost with ground truth, which potentially indicates drivers' heuristic charging behavior, i.e., charging in the nearest charging stations. TBA also achieves a small percentage of increase due to the reinforcement learning for long-term benefits. Other reinforcement learning methods, e.g., Care and DQN can also reduce the charging cost, but the performance of FairMove is superior to them. Since our method considers both vehicle dispatching and charging scheduling, it also reduces the cruise time for EFHVs besides scheduling them to charge in low-rate hours, so the number of charges will also change with different methods. In general, the reduction of the extra time for seeking passengers (e.g., proactively dispatch EFHVs to the future high passenger demand regions and avoid heavy traffic congestion) also reduces the energy and potentially saves charging costs for the EFHV fleet. In particular, our FairMove achieves the best performance by reducing 5.9% of the charging cost. Even though the percentage is not very large, it can reduce 4.561 million CNY for the Shenzhen EFHV fleet per month considering the large number of EFHVs.

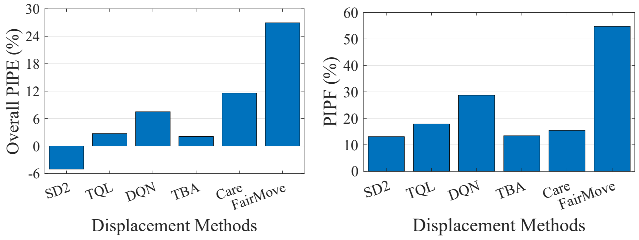


(a) Weekly PRCC (b) Hourly profit efficiency

Fig. 10. Weekly PRCC and profit efficiency.

4.2.4 Profit Efficiency Comparison

Since one of the most important objectives of the paper is to improve the profit efficiency of the EFHV fleet, we compare FairMove to baselines considering their profit efficiency changes. Fig. 10(b) shows the hourly profit efficiency of each EFHV under different displacement methods. We found the hourly profit efficiency varies from 0 to 120 without displacement, and the average value is 44.03. The profit efficiency of SD2 has a slight decrease due to the prolonged idle time. Both `Care` and DQN increase the hourly profit efficiency for EFHVs on average, but our `FairMove` achieves the best performance, with a mean value of 55.89. In addition, the variance between the EFHVs becomes smaller since we consider the fairness between them.



(a) PIPE comparison (b) PIPF comparison

Fig. 11. PIPE and PIPF under different methods.

In Fig. 11(a), we show the overall PIPE in one month of different displacement methods. We found our `FairMove` increases the profit efficiency for the EFHV fleet by 26.9%, followed by `Care` with an 11.6% of increase. However, SD2 reduces the profit efficiency for the EFHV fleet by 5% due to the prolonged idle time.

4.2.5 Profit Fairness Comparison

Another key objective of the paper is to improve the profit fairness for the EFHV fleets. From 11(b), we found our `FairMove` achieves the best performance with 54.8% of PIPF. The reason may be that we formulate the problem with fairness as a part of the objective function, and solve it by deep reinforcement learning methods, so it not only improves the profit efficiency but also improves the profit fairness for the EFHV fleet. SD2 and TBA achieve similar improvements in the profit fairness of the EFHV fleet by 13%. Due to the fairness consideration, TQL and DQN also improve the fairness efficiency by 28.7% and 17.9%, respectively. Even though `Care` also increases the profit fairness of the EFHV fleet, its performance is worse than our `FairMove` since it does not consider the fairness between drivers.

4.2.6 Performance Under Different Weighted Factor α

In this subsection, we conduct a sensitivity analysis to study the impacts of the reward-weighted factor α for the training of the proposed multi-agent deep reinforcement learning method. As mentioned, α measures the tradeoff between the overall profit efficiency of the fleet and the fairness between individual EFHVs. The higher the α, the more emphasis on the overall profit efficiency. The lower the α, the more emphasis on fairness. We compare the performance of the `FairMove` with different weighted factor α (from 0 to 1 with a step of 0.01). The average reward r of the proposed FAMA2C is shown in Fig. 12, which shows that setting the parameter α around 0.6 leads to the best system performance. Since maximizing fairness alone may harm the overall profit efficiency of the EFHV fleet, this finding is reasonable. This is also the reason why we select $\alpha = 0.6$ for the above comparisons.

Since our method is in a centralized training and decentralized execution fashion, it obtains a decision for each EFHV within one second after the training process based on our implementation, and the decision-making time for EFHVs is at a similar level due to the similar sizes of their action spaces, which is typically fast enough for the real-world displacement requirement.

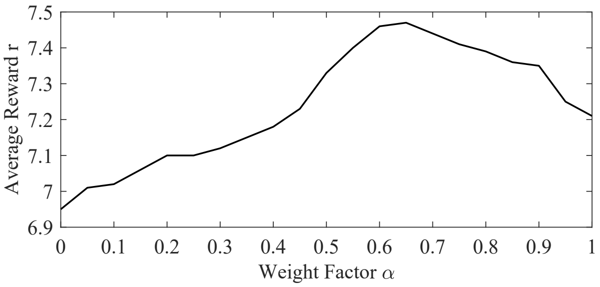


Fig. 12. System performance under different weighted factor α.

4.2.7 Convergence Behavior

It should be noted that providing a rigorous theoretical convergence proof for deep multi-agent reinforcement learning algorithms, especially for AC-based methods, is highly nontrivial and several major technical challenges naturally arise, such as communication cost for interactions [32], [33], [34]. Considering that single-agent AC algorithms' convergence is known to be fragile [35], the rigorous convergence analysis of multi-agent AC algorithms becomes much harder due to complex mutual agent interactions. Even so, we provide the following analysis to show the complexity of our proposed algorithm from three aspects in the revision of the paper, i.e., interactions, neural network sizes, and experiments on real-world data.

(i) In our multi-agent environment, each EFHV functions as an independent agent, capable of interacting with the environment and collecting valuable information including state, action, and reward. We collect exploration information from various agents and centralize them in the replay buffer to facilitate unified learning. This abundance of agents ensures a comprehensive exploration of the environment and brings a wide range of high-quality action execution solutions. As a result, the convergence process of our algorithm is naturally accelerated with low time complexity. (ii) The complexity of the proposed AC-based algorithms is also related to the neural network sizes. In our algorithm, the two neural networks have three hidden layers, with 256, 128, and 64 nodes from the first to the last hidden layer. The update strategies of these two networks in this size are efficient with low time complexity [36], [37]. (iii) Since this paper mainly focuses on addressing practical problems, i.e., jointly optimizing profit efficiency and fairness of electric for-hire vehicles, we use substantial and comprehensive experiments to demonstrate the efficiency, effectiveness, and convergence of the proposed FAMA2C algorithm. We study the convergence behavior of our FAMA2C in terms of the reward of agents. Fig. 13 shows the evolution of the mean and deviation of the reward of agents, which indicates the convergence behavior of the proposed FAMA2C in the training process. We found after about 800 iterations, the rewards of agents will start to be stationary. The mean reward becomes larger and the variance becomes smaller with the training process. Since we carefully define the actions and states space, which leads to a small decision-making space for each agent, our method is effective and efficient for the EFHV displacement.

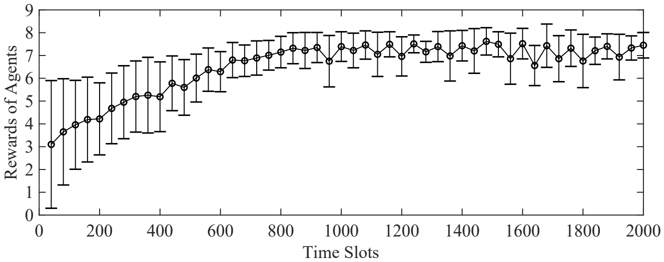
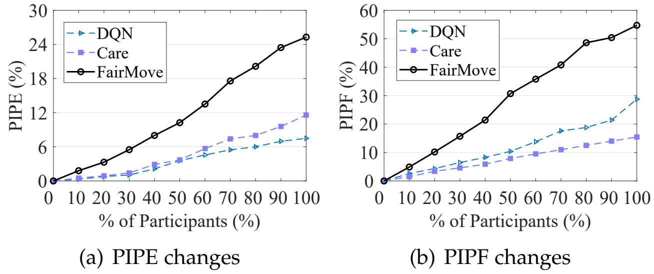


Fig. 13. Mean and variance of rewards of all agents.

4.2.8 Impact of Driver Participation Rate

Even though our fairness and efficiency consideration can potentially offer an incentive for drivers to participate in our displacement, in real-world scenarios, it is possible that some drivers may not follow our displacement decisions due to their own preferences. Hence, we further investigate the system performance under different participation rates p and show how it may affect the system as a whole. In our simulator, this probability p indicates how likely the drivers would accept our displacement decisions. We vary p from 0 to 1, where $p = 0$ means no drivers follow our decisions, which is equivalent to the Ground Truth (i.e., no displacement) and $p = 1$ means that all drivers would accept all the displacement proposals.



(a) PIPE changes (b) PIPF changes

Fig. 14. PIPE and PIPF under different participation rates.

We compare our FairMove with the two advanced methods, i.e., DQN-based displacement and Care in terms of PIPE and PIPF. As shown in Fig. 14(a) and Fig. 14(b), we found that with more drivers participating in the displacement system, both the profit efficiency and profit fairness will increase. Care achieves higher profit efficiency that DQN-based displacement but its profit fairness is lower. Our FairMove can achieve the best performance for both profit efficiency and fairness. Even with half drivers following our displacement decisions, our FairMove can improve the profit efficiency by 10.2% and profit fairness by 30.7% for the EFHV fleet, which indicates our FairMove has the potential to put into practice.

4.2.9 Fairness of Charging Station Utilization

Since charging stations in a city may belong to different operators (e.g., there are more than 10 EFHV charging station operators in Shenzhen), different vehicle displacement methods may also cause unfairness between charging station operators. In this part, we study the impact of displacement methods on fairness between charging station operators. We envision that the occupation time of different charging stations should be similar to achieve fairness between different operators.

We define the daily *Charging Station Utilization (CSU)* of a station s_i to quantify the average daily occupation time of each charging point in the station, which is denoted as:

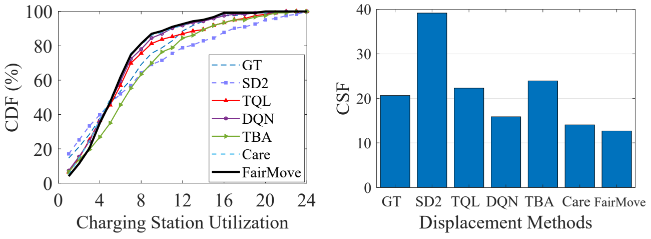
$$CSU(s_i) = \frac{\sum_{j=1}^{|s_i|} T_{charge}\left(s_i^j\right)}{|s_i|} \tag{17}$$

where $T_{charge}(s_i^j)$ is the daily occupation time of j^{th} charging point s_i^j in the station s_i; $|s_i|$ is the number charging points in the station s_i.

Hence, the fairness between charging stations can be described as follows:

$$CSF = \frac{1}{M}\sum_{i=1}^{M}\left(CSU(s_i) - \overline{CSU}\right)^2 \tag{18}$$

where M is the number of EFHV charging stations in the city and CSF is the fairness of the EFHV charging network. A smaller value of CSF means fairer between different charging stations.



(a) Charging station utilization (b) Charging station fairness

Fig. 15. Charging station utilization and fairness.

As shown in Fig. 15(a), we found many charging stations have low utilization without any displacement methods, e.g., the CSU of 21% of stations is lower than 2, which means the average occupation time is shorter than 2 hours. This low CSU potentially causes charging resource waste for these stations. In addition, we found that the CSU of over 17% of stations is higher than 12, which potentially causes crowded charging stations. This high CSU discrepancy potentially results in unfairness between charging stations. We found the SD_2 displacement method makes the situation worse, leading to the CSU of 25.2% of stations lower than 2 and 25.2% of stations higher than 12. With our FairMove, 61% of charging stations have the CSU between 4 to 8, which potentially makes them efficient for charging. As shown in Fig. 15(b), our FairMove achieves the highest fairness for charging stations. In general, FairMove increases the CSF by 12.7, which is 38.4% of improvement compared to the ground truth of 20.6.

4.2.10 Off-Policy Evaluation

Since the environment where the learned policy will be deployed may deviate from the past environment where the ground truth data were collected and should be considered, in order to avoid deploying a bad policy, it is imperative that policies learned by RL algorithms are thoroughly evaluated prior to real-world deployment. Off-policy evaluation (OPE) [38] is the task of predicting the online performance of a policy using only pre-collected historical data (collected from an existing deployed policy or set of policies), it uses episodes generated via a behavioral policy π_b (data-generating policy) to evaluate the value function of the target policy π_e whose value function we seek to estimate.

In recent years, OPE becomes a very hot topic in the reinforcement learning community and more and more researchers pay attention to it. However, there is little research on applying OPE methods for fleet management or even multi-agent deep reinforcement learning. Hence, in this paper, we try to utilize OPE methods to validate the learned policy against the real-world data for better real-world deployment, but it is extremely challenging to apply OPE methods to validate the learned policy even though using massive data because our large-scale fleet management environment is very complicated, which includes different practical factors, e.g., the number of active agents, number of available charging points in each charging stations, future passenger demand, etc. We tried different OPE methods, including Approximate Model [39], Doubly Robust [40], and MAGIC [41]. Approximate Model(AM) is one the most commonly used model-based direct OPE method, which directly estimates the value functions of the evaluation policy with regression-based techniques. The transition dynamics, reward function and termination condition are directly estimated from historical data. Doubly Robust(DR) is an unbiased estimator of the value function of the evaluation policy that achieves promising empirical and theoretical results by leveraging an approximate model of an MDP to decrease the variance of the unbiased estimates produced by ordinary importance sampling. MAGIC directly optimizes mean squared error, which combines a purely model-based estimator with weighted doubly robust and has shown good performance [41].

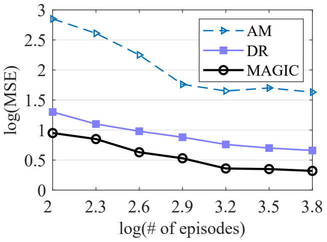


Fig. 16. Comparison of OPE.

We utilize the mean squared error (MSE) over the episode to evaluate each method, $MSE = \mathbb{E}\left[\left(\hat{V}(\pi_e) - V(\pi_e)\right)^2\right]$, which is a standard metric for OPE, where $\hat{V}(\pi_e)$ is calculated based on the behavioral policy π_b; $V(\pi_e)$ is the value function of the target policy π_e. As shown in Fig. 16, we found MAGIC achieves the best performance (i.e., smallest MSE with the increase of episodes) while the AM achieves the worst performance. Even though the MSE would be small when more data is used, there is still a difference between the value function of the target policy and the value with the real-world data. One possible reason would be the complicated environment of our fleet management environment, which makes the OPE methods challenging to simulate all factors.

5 Discussions

5.1 Lessons Learned

- **Data-Driven Findings.** Based on our data-driven investigation, we obtain some new findings. (i) Idle time reduction does not necessarily indicate the prolonged time for serving passengers since some drivers may need to spend more time seeking passengers after charging in some regions with low passenger travel demand (Fig. 4(a) and Fig. 4(b)), which is rarely considered by existing charging recommendation works. (ii) The potential revenue for serving passengers after charging in different stations may be also different, which is highly dynamic in both spatial and temporal dimensions (Fig. 3). (iii) There are intensive charging peaks during shifts and low charging pricing durations, which potentially indicates the charging price is a key factor impacting their charging behaviors.
- **Real-World Insights From Field Studies.** During the project, we have been conducting a series of field studies in Shenzhen, where we interviewed 12 e-taxi drivers. All drivers think they care more about profits. The three most important factors for them to consider when finding a charging station are charging price, queuing time, and distance, even though they mentioned they also charge in stations that are near to their homes. Hence, the charging costs and queuing time should be considered in a displacement system. Moreover, we found almost all e-taxi drivers think it is fair when their profits are proportional to their working time, which motivates the profit fairness definition of this paper (Equation 3).
- **Fairness-Aware Displacement for EFHV Fleets.** We found there is a huge profit efficiency gap between EFHVs in the Shenzhen EFHV fleet (Fig. 5), which could be improved to be fairer by an effective displacement system. Our multi-agent reinforcement learning method shows good performance for EFHV displacement, which has the potential to improve profit efficiency and profit fairness at the same time.

5.2 Potential Implications and Future Work

Implementation in Different Cities: In this paper, we only leverage the data from the Chinese city Shenzhen to conduct experiments to verify our `FairMove`, and we admit it is hard to justify the universality of the proposed method without data from other cities due to the data-driven characteristics of our method. We are also in the process of obtaining EFHV data from other cities to investigate if our displacement system is applicable to other cities. However, since Shenzhen is the only city that has such a large-scale and all EFHV fleet in the world currently, it is challenging to find another large-scale EFHV fleet for a parallel study. Even so, we argue that our method has the potential to generalize to other cities if we can have access to EFHV data from other cities. The reason is that our method relies on only the drivers' mobility, charging, and profit patterns instead of city features.

Drivers' Willingness to Use Our System. Since different drivers may have different priorities, it is possible that some drivers may not follow the displacement despite the fact that they know the displacement decisions can benefit them. Even though our fairness consideration can potentially offer an incentive for drivers to participate in our displacement, we also investigate the system performance by taking drivers' willingness into consideration. As shown in Fig. 14(a) and Fig. 14(b), we show the system performance under different participation rates and show how it may affect the system as a whole.

Impact of EV Types. In our work, all EVs are the same type, i.e., BYD e6, and they have the same battery consumption model and charging model, so they share the same solution. For different types of EVs, they have different battery capacities, consumption rates, and charging rates, which will impact their charge time. Although the charging idle time will be different if EVs have different charging rates and battery capacities, our solution considers the status (cruise time, idle time, and charge time) of individual EVs, so different EV models will not impact our solution since we will input the cruise time, idle time, and charge time of individual EV to our model for decision making. If there are different EV models, we only need to specify their battery capacities, consumption rates, and charging rates to calculate the cruise time, idle time, and charge time of individual EVs to feed into our optimization objective.

Fairness of Different Driver Groups. Since drivers may have different performance, which is decided by many factors like EFHV driving years, accidents, and reputation, it is also reasonable to divide all drivers into different groups by their performance levels and quantify their fairness within the same group. Even though we did not divide the drivers into different groups, we found the government and EFHV companies have comprehensively evaluated each driver's performance based on multiple factors and label it on the EFHV, which is normally represented by a five-star rating. Hence, we can directly merge it into our displacement system for five groups and achieve fairness in each group.

Dynamic Thresholds for Charging Decisions and Dynamic Temporal Partition. In this work, we only utilize a lower bound of the energy level to decide when displacing EFHVs to charge, which is adopted by many existing EFHV charging works [12], [42]. In practice, the system performance may be enhanced if we set different and dynamic thresholds for the EFHVs even though the problem will be also more complicated. Besides, we only utilize a fixed temporal partition, but a non-regular partition may cause higher performance with some extra computational costs. We will study the impacts of dynamic thresholds on system performance and computational complexity in future work.

6 Related Work

In this section, we review four streams of literature relevant to this paper and show the uniqueness of our work.

6.1 Traditional Gas FHV Dispatching

In the last decade, with the wide development of mobile sensors and advanced communication technologies, a large number of works have been done to improve the service efficiency of FHV fleets based on real-world data, e.g., GPS data and transaction data [43], [44]. Ding et al. [7] presented a multi-agent RL framework to help FHVs make distributed routing decisions for urban crowdsensing based on taxi GPS and transaction data. Xu et al. [44] built a dispatching system in which the predicted demands and destinations were used for the taxi reallocation towards the future supply-demand balance in the city. However, all these works failed to consider the fairness between FHV drivers when they make dispatching decisions. Even though [45] proposed a route assignment mechanism for fair taxi route recommendations, it focused on the conventional gas FHVs, which has different operation patterns and energy replenishment mechanisms with EFHVs. In addition, the complicated charging process has not been considered, which makes it challenging to be reapplied for EFHV displacement.

6.2 EFHV Charging Scheduling

With the rapid vehicle electrification process, more and more research [3], [20], [46], [47], [48] focuses on EFHV charging issues. Among all these works, e-taxi charging scheduling is one of the most popular topics. Wang et al. [49] designed a real-time charging scheduling system called tCharge to reduce the queuing time of e-taxi drivers for available charging points. Dong et al. [12] developed a real-time charging scheduling framework for EFHV fleets to reduce the queuing time of EFHVs. However, most of these works only focused on the charging issues of EFHVs without considering the potential revenue loss related to charging. In addition, they neglected the fairness between EFHVs, which may potentially cause drivers not to follow their scheduling decisions and make the system unsustainable in the long run.

Recently, there are some works [20], [50] trying to seek fairness-aware scheduling for EFHVs. Yang et al. [50] proposed a charging coordination solution for EFHVs to reduce their queuing time in charging stations. Wang et al. [20] designed a fairness-aware Pareto efficient charging recommendation system called FairCharge to minimize the total charging idle time (traveling time + queuing time) in charging stations combined with fairness constraints. However, all these works only considered the charging processes of EFHVs while neglecting their overall revenue, which is a key concern of FHV drivers. We found that charging idle time reduction does not mean prolonged time for serving passengers and higher profits because some drivers may need to spend longer time to seek passengers after charging in some stations with shorter charging waiting time.

6.3 Deep Reinforcement Learning

There is an increasing number of studies that utilize deep reinforcement learning to solve sequential decision-making problems (e.g., taxi dispatching [3], computation offloading [17], [51], and navigation [52]) due to its excellent performance. Ding et al. [7] proposed a multi-agent reinforcement learning framework for urban crowdsensing with for-hire vehicles. Shi et al. [17] developed a deep reinforcement learning-based computation offloading scheme to motivate vehicles to share their computing resources. There are also some works focusing on electric vehicle dispatching with deep reinforcement learning. Wang et al. [3] designed a charging and relocation recommendation system for e-taxi drivers with deep reinforcement learning. However, the fairness between drivers has not been considered in these works, which potentially causes the system unsustainable.

6.4 Fairness in Transportation

Fairness in transportation has attracted much interest in the research community due to its importance. Researchers have focused on fairness in different transportation modes such as ride-hailing [25], [53], [54], taxi [13], [20], bikesharing [55], and e-scooter sharing [56]. Sun et al. [25] exploit joint order dispatching and driver repositioning to optimize the long-term fairness in a ride-hailing system. Shi et al. [53] consider the dependency between current and future assignments to improve the performance of fair task assignments in ride-hailing. Yan et al. [55] designed a fairness-aware spatiotemporal model for predicting new mobility resource demand. He et al. [56] proposed a socially-equitable flow prediction system for dockless e-scooter sharing. However, there is no existing work on jointly optimizing the profit efficiency and fairness of EFHVs.

6.5 Uniqueness of Our Work

In summary, to our best knowledge, our `FairMove` is the first displacement system for EFHVs to improve the overall profit efficiency of EFHV fleets with fairness consideration, which is motivated by data-driven findings based on real-world data. `FairMove` considers not only the passenger demand but also the complicated charging issues of EFHVs (e.g., unique charging behaviors and time-varying electricity pricing). Moreover, `FairMove` also emphasizes the fairness between EFHVs. Our multi-agent deep reinforcement learning method FAMA2C algorithm with centralized training and de-centralized execution also shows good performance for vehicle displacement.

7 Conclusion

In this paper, we design the first data-driven fairness-aware displacement system called `FairMove` based on multi-source data, which aims to jointly optimize the overall profit efficiency and profit fairness of the entire EFHV fleet. We first conduct a data-driven investigation, from which we found some new findings about the uniqueness of the EFHV displacement problem to motivate our work. We then formulate the EFHV displacement as a sequential decision problem and then propose a fairness-aware multi-agent actor-critic approach to tackle this problem. `FairMove` considers not only dynamic passenger demand & supply in both temporal and spatial dimensions but also considers the complicated charging problems (e.g., time-variant electricity pricing) and per-trip profit. We implement and evaluate our `FairMove` based on a real-world dataset obtained from a large-scale EFHV fleet including over 20,100 vehicles. Extensive experimental results show that our fairness-aware `FairMove` effectively improves the profit efficiency and profit fairness by 26.9% and 54.8%, respectively. It also improves the charging station utilization fairness by 38.4%.

References

[1] G. Fan, Z. Yang, H. Jin, X. Gan, and X. Wang, "Enabling optimal control under demand elasticity for electric vehicle charging systems," *IEEE Transactions on Mobile Computing*, 2020.

[2] C. Wang, Y. Song, G. Fan, H. Jin, L. Su, F. Zhang, and X. Wang, "Optimizing cross-line dispatching for minimum electric bus fleet," *IEEE Transactions on Mobile Computing*, 2021.

[3] E. Wang, R. Ding, Z. Yang, H. Jin, C. Miao, L. Su, F. Zhang, C. Qiao, and X. Wang, "Joint charging and relocation recommendation for e-taxi drivers via multi-agent mean field hierarchical reinforcement learning," *IEEE Transactions on Mobile Computing*, 2020.

[4] G. Wang, X. Chen, F. Zhang, Y. Wang, and D. Zhang, "Experience: Understanding long-term evolving patterns of shared electric vehicle networks," in *The 25th Annual International Conference on Mobile Computing and Networking*, 2019, pp. 1–12.

[5] Y. Li, J. Luo, C.-Y. Chow, K.-L. Chan, Y. Ding, and F. Zhang, "Growing the charging station network for electric vehicles with trajectory data analytics," in *Data Engineering (ICDE), 2015 IEEE 31st International Conference on*. IEEE, 2015, pp. 1376–1387.

[6] V. I. Roland Irle, Jos Pontes, "The electric vehicle world sales database," http://www.ev-volumes.com/, 2018.

[7] R. Ding, Z. Yang, Y. Wei, H. Jin, and X. Wang, "Multi-agent reinforcement learning for urban crowd sensing with for-hire vehicles," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 2021, pp. 1–10.

[8] X. Zhou, H. Rong, C. Yang, Q. Zhang, A. V. Khezerlou, H. Zheng, M. Z. Shafiq, and A. X. Liu, "Optimizing taxi driver profit efficiency: A spatial network-based markov decision process approach," *IEEE Transactions on Big Data*, 2018.

[9] F. Borutta, S. Schmoll, and S. Friedl, "Optimizing the spatio-temporal resource search problem with reinforcement learning (gis cup)," in *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2019, pp. 628–631.

[10] J.-S. Kim, D. Pfoser, and A. Züfle, "Distance-aware competitive spatiotemporal searching using spatiotemporal resource matrix factorization (gis cup)," in *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2019, pp. 624–627.

[11] G. Wang, F. Zhang, H. Sun, Y. Wang, and D. Zhang, "Understanding the long-term evolution of electric taxi networks: A longitudinal measurement study on mobility and charging patterns," *ACM Transactions on Intelligent Systems and Technology (TIST)*.

[12] Z. Dong, C. Liu, Y. Li, J. Bao, Y. Gu, and T. He, "Rec: Predictable charging scheduling for electric taxi fleets," in *Real-Time Systems Symposium (RTSS), 2017 IEEE*. IEEE, 2017, pp. 287–296.

[13] G. Wang, S. Zhong, S. Wang, F. Miao, Z. Dong, and D. Zhang, "Data-driven fairness-aware vehicle displacement for large-scale electric taxi fleets," in *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 2021, pp. 1200–1211.

[14] Wikipedia, "Byd e6," https://en.wikipedia.org/wiki/BYD_e6, 2020.

[15] G. electric vehicle, "Shenzhen will promote over 10,000 electric taxis. how to deal with the complicated charging issues?" https://www.gg-ev.com/asdisp2-65b095fb-23169-.html, 2020.

[16] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton *et al.*, "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[17] J. Shi, J. Du, Y. Shen, J. Wang, J. Yuan, and Z. Han, "Drl-based v2v computation offloading for blockchain-enabled vehicular networks," *IEEE Transactions on Mobile Computing*, 2022.

[18] S. Wang, Y. Guo, N. Zhang, P. Yang, A. Zhou, and X. Shen, "Delay-aware microservice coordination in mobile edge computing: A reinforcement learning approach," *IEEE Transactions on Mobile Computing*, vol. 20, no. 3, pp. 939–951, 2019.

[19] M. Fleurbaey and F. Maniquet, *A theory of fairness and social welfare*. Cambridge University Press, 2011, vol. 48.

[20] G. Wang, Y. Zhang, Z. Fang, S. Wang, F. Zhang, and D. Zhang, "Faircharge: A data-driven fairness-aware charging recommendation system for large-scale electric taxi fleets," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 1, pp. 1–25, 2020.

[21] E. Liang, K. Wen, W. H. Lam, A. Sumalee, and R. Zhong, "An integrated reinforcement learning and centralized programming approach for online taxi dispatching," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[22] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[23] G. Wang, X. Xie, F. Zhang, Y. Liu, and D. Zhang, "bcharge: Data-driven real-time charging scheduling for large-scale electric bus fleets," in *2018 IEEE Real-Time Systems Symposium (RTSS)*. IEEE, 2018, pp. 45–55.

[24] D. Zhang, Y. Li, F. Zhang, M. Lu, Y. Liu, and T. He, "coride: Carpool service with a win-win fare model for large-scale taxicab networks," in *Proceedings of the 11th ACM conference on embedded networked sensor systems*, 2013, pp. 1–14.

[25] J. Sun, H. Jin, Z. Yang, L. Su, and X. Wang, "Optimizing long-term efficiency and fairness in ride-hailing via joint order dispatching and driver repositioning," in *Proceedings of the 28th ACM SIGKDD*

Conference on Knowledge Discovery and Data Mining, 2022, pp. 3950–3960.

[26] Y. Li, "Deep reinforcement learning: An overview," *arXiv preprint arXiv:1701.07274*, 2017.

[27] H. Van Seijen, A. R. Mahmood, P. M. Pilarski, M. C. Machado, and R. S. Sutton, "True online temporal-difference learning," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 5057–5096, 2016.

[28] F. Stollar, J. McCann, and R. Chatley, "Fleet management in on-demand transportation networks: Using a greedy approach," *Imperial College London*, 2018.

[29] I. Jindal, Z. T. Qin, X. Chen, M. Nokleby, and J. Ye, "Optimizing taxi carpool policies via reinforcement learning and spatio-temporal mining," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 1417–1426.

[30] J. Foerster, I. A. Assael, N. De Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," in *Advances in neural infor. processing systems*, 2016, pp. 2137–2145.

[31] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[32] F. Hairi, J. Liu, and S. Lu, "Finite-time convergence and sample complexity of multi-agent actor-critic reinforcement learning with average reward," in *International Conference on Learning Representations*, 2021.

[33] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar, "Fully decentralized multi-agent reinforcement learning with networked agents," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5872–5881.

[34] Z. Chen, Y. Zhou, R.-R. Chen, and S. Zou, "Sample and communication-efficient decentralized actor-critic algorithms with finite-time analysis," in *International Conference on Machine Learning*. PMLR, 2022, pp. 3794–3834.

[35] Z. Yang, Y. Chen, M. Hong, and Z. Wang, "Provably global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost," *Advances in neural information processing systems*, vol. 32, 2019.

[36] H. Kumar, A. Koppel, and A. Ribeiro, "On the sample complexity of actor-critic method for reinforcement learning with function approximation," *Machine Learning*, pp. 1–35, 2023.

[37] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," *Advances in neural information processing systems*, vol. 12, 1999.

[38] C. Voloshin, H. M. Le, N. Jiang, and Y. Yue, "Empirical study of off-policy policy evaluation for reinforcement learning," *arXiv preprint arXiv:1911.06854*, 2019.

[39] C. Paduraru, "Off-policy evaluation in markov decision processes," Ph.D. dissertation, McGill University, 2013.

[40] N. Jiang and L. Li, "Doubly robust off-policy value evaluation for reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2016, pp. 652–661.

[41] P. Thomas and E. Brunskill, "Data-efficient off-policy policy evaluation for reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2016, pp. 2139–2148.

[42] Y. Yuan, D. Zhang, F. Miao, J. Chen, T. He, and S. Lin, "pˆ 2charging: Proactive partial charging for electric taxi systems," in *2019 IEEE 39th ICDCS*. IEEE, 2019, pp. 688–699.

[43] F. Miao, S. He, L. Pepin, S. Han, A. Hendawi, M. E. Khalefa, J. A. Stankovic, and G. Pappas, "Data-driven distributionally robust optimization for vehicle balancing of mobility-on-demand systems," *ACM Trans. Cyber-Phys. Syst.*, vol. 5, no. 2, Jan. 2021. [Online]. Available: https://doi.org/10.1145/3418287

[44] J. Xu, R. Rahmatizadeh, L. Bölöni, and D. Turgut, "Taxi dispatch planning via demand and destination modeling," in *2018 IEEE 43rd Conference on Local Computer Networks (LCN)*. IEEE, 2018, pp. 377–384.

[45] S. Qian, J. Cao, F. L. Mouël, I. Sahel, and M. Li, "Scram: A sharing considered route assignment mechanism for fair taxi route recommendations," in *Proceedings of the 21th ACM SIGKDD*, 2015, pp. 955–964.

[46] L. Yan, H. Shen, Z. Li, A. Sarker, J. A. Stankovic, C. Qiu, J. Zhao, and C. Xu, "Employing opportunistic charging for electric taxicabs to reduce idle time," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, p. 47, 2018.

[47] G. Wang, W. Li, J. Zhang, Y. Ge, Z. Fu, F. Zhang, Y. Wang, and D. Zhang, "sharedcharging: Data-driven shared charging for large-scale heterogeneous electric vehicle fleets," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–25, 2019.

[48] S. Schoenberg and F. Dressler, "Reducing waiting times at charging stations with adaptive electric vehicle route planning," *IEEE Transactions on Intelligent Vehicles*, 2022.

[49] G. Wang, F. Zhang, and D. Zhang, "tcharge-a fleet-oriented real-time charging scheduling system for electric taxi fleets," in *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*, 2019, pp. 440–441.

[50] Z. Yang, T. Guo, P. You, Y. Hou, and S. J. Qin, "Distributed approach for temporal–spatial charging coordination of plug-in electric taxi fleet," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 6, pp. 3185–3195, 2018.

[51] L. Huang, S. Bi, and Y.-J. A. Zhang, "Deep reinforcement learning for online computation offloading in wireless powered mobile-edge computing networks," *IEEE Transactions on Mobile Computing*, vol. 19, no. 11, pp. 2581–2593, 2019.

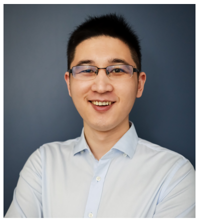
[52] C. H. Liu, X. Ma, X. Gao, and J. Tang, "Distributed energy-efficient multi-uav navigation for long-term communication coverage by deep reinforcement learning," *IEEE Transactions on Mobile Computing*, vol. 19, no. 6, pp. 1274–1285, 2019.

[53] D. Shi, Y. Tong, Z. Zhou, B. Song, W. Lv, and Q. Yang, "Learning to assign: Towards fair task assignment in large-scale ride hailing," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 3549–3557.

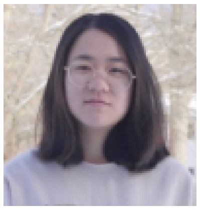
[54] V. Nanda, P. Xu, K. A. Sankararaman, J. Dickerson, and A. Srinivasan, "Balancing the tradeoff between profit and fairness in rideshare platforms during high-demand hours," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 02, 2020, pp. 2210–2217.

[55] A. Yan and B. Howe, "Fairness-aware demand prediction for new mobility," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 1079–1087.

[56] S. He and K. G. Shin, "Socially-equitable interactive graph information fusion-based prediction for urban dockless e-scooter sharing," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 3269–3279.



Guang Wang is an assistant professor at the Department of Computer Science, Florida State University. Before that, he was a Postdoctoral Research Associate at Massachusetts Institute of Technology. He obtained his Ph.D. degree in Computer Science, Rutgers University, New Brunswick, NJ, USA. He is interested in mobile computing, cyber-physical systems, big data analytics, and machine learning. His technical contributions have led to more than 50 peer-reviewed publications in premium conferences and journals, e.g., MobiCom, IMWUT/UbiComp, RTSS, KDD, ICDE, WWW, AAAI, IEEE TMC, TKDE, TITS, TVT, TBD, ACM TIST, and TOSN.

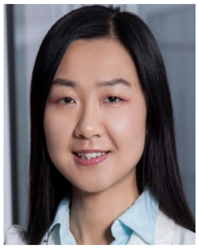


Sihong He is currently a PhD student in Computer Science and Engineering at University of Connecticut, Storrs, CT, USA. She received her bachelor's degree from Southern University of Science and Technology, Shenzhen, China, in 2017. After that, she obtained her master's degree in Statistics from the Donald Bren School of Information and Computer Sciences, the University of California, Irvine, CA, USA in 2019. Her current research interests include data-driven optimization, robust optimization, reinforcement learning, machine learning, etc.

Lin Jiang is currently a PhD student in Computer Science at Florida State University. He received his bachalor's degree from Chien-Shiung Wu Honored College, Southeast University, China, in 2020. After that, he obtained his master's degree in cyber science from the same university, in 2023. He is interested in big data analytics and cyber-physical systems, especially the decision-making algorithms applied in the novel applications such as smart cities.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2023.3326676

Shuai Wang received the B.S. and M.S. degrees from the Huazhong University of Science and Technology, China, and the Ph.D. degree from the Department of Computer Science and Engineering, University of Minnesota, in 2017. He is currently a Professor with the School of Computer Science and Engineering, Southeast University. His research interests include the Internet of Things, cyber physical systems, data science, and wireless networks and sensors.



Fei Miao is an Assistant Professor in the Department of Computer Science and Engineering at the University of Connecticut. She received the Ph.D. degree from the University of Pennsylvania in 2016. Her research interests include data-driven real-time optimization and control of cyber-physical systems under model uncertainties, and resilient and secure control cyber-physical systems. She was a Best Paper Award Finalist at the 6th ACM/IEEE International Conference on Cyber-Physical Systems in 2015

Fan Zhang received the Ph.D. degree in communication and information system from the Huazhong University of Science and Technology in 2007. He was a Post-Doctoral Fellow with the University of New Mexico and with the University of Nebraska-Lincoln from 2009 to 2011. He is currently a Professor with the SIAT, Chinese Academy of Sciences, China. His research topics include intelligent transportation systems, cyber-physical systems, and urban computing.

Zheng Dong is an assistant professor in the Department of Computer Science at Wayne State University. He received the PhD degree from the Department of Computer Science at the University of Texas at Dallas in 2019. His research interests include real-time cyber physical systems and mobile edge computing. He received the Outstanding Paper Award at the 38th IEEE RTSS. He is a member of the IEEE.

Desheng Zhang is an assistant professor at the Department of Computer Science at Rutgers University. Desheng is broadly concentrated on bridging cyber-physical systems and big urban data by technical integration of communication, computation and control in data-intensive urban systems. He is focused on the life cycle of big data-driven urban systems, from multi-source data collection to streaming-data processing, heterogeneous-data management, model abstraction, visualization, privacy, service design and deployment in complex urban setting.