Real-time Planning of Route, Speed, and Charging for Electric Delivery Vehicles: A Deep Reinforcement Learning Approach

Xiaowen Bi, Minyu Shen, Weihua Gu, Edward Chung, Yuhong Wang

***Abstract*—Motor vehicles typically exhibit a "speed-varying range" (SVR) characteristic. For battery-powered electric vehicles (BEVs), the range diminishes at higher speed. This characteristic greatly impacts BEV operation for demanding commercial uses like express delivery, given their limited range and long recharge times. In view of the above, this paper examines a new electric vehicle routing problem that explicitly models BEVs' SVR and considers the joint planning of BEV route, speed, and charging under stochastic traffic conditions. A deep reinforcement learning approach that exploits the interdependence among the above three decision aspects is then developed to generate real-time policies. Experiments on hypothetical and real-world instances showcase that the proposed approach can efficiently find high-quality policies that effectively accommodate BEVs' SVR.**

***Index Terms*—Electric Vehicle; Delivery Planning; Speed-varying Range; Uncertain Traffic Condition; Deep Reinforcement Learning**

I. Introduction

Electric vehicles (EVs) have been considered as a promising solution for sustainable road transportation. As compared to internal combustion engine vehicles, EVs have significantly less greenhouse gas and air pollutant emissions. They are also two- to four-times more energy efficient and contribute to reducing the societal reliance on fossil fuels [1].

The global EV market has been rapidly expanding. In 2022, the EV sales exceeded 10 million worldwide, accounting for 14 % of all new vehicles sold [2]. In particular, over 70% of the global stock are battery EVs (BEVs) [2]. Thanks to the steadily declining price of automotive battery (150 $/kWh at a pack level in 2022 [2]) and maturing BEV technologies, BEV models are now offered in most vehicle classes, including not only light-duty vehicles such as electric cars mostly for private use, but also medium- and heavy-duty vehicles like electric vans, buses, and trucks typically adopted in commercial applications. The expansion of BEV models has been driving the electrification of commercial vehicle fleets in major markets, and the penetration of medium- and heavy-duty BEVs is thus expecting a rapid growth. In the sector of distribution logistics, Amazon has been operating one of the largest fleets of more than 10000 Rivian electric delivery vans in the U.S. [3], and has initiated the electrification of its European delivery network with the rollout of 300 Rivian vans in German in 2023 [4]. The U.S. Postal Service ordered 9250 Ford E-Transits electric vans (to be delivered over the course of 2024) to electrify the last-mile deliveries and is also actively exploring other avenues, e.g., Canoo electric vans [5]. The decarbonization of public transit sector is also underway. In 2023, the Go-Ahead Group in UK signed a repeat order for 141 BYD-Alexander Dennis electric bus to continue its transition to zero-emission buses [6], and the New South Wales government in Australia launched a state program to drive towards zero emissions from public transport by 2050 [7].

Despite the commercial sectors' commitments to fleet electrification, the medium- and heavy-duty BEVs with higher gross weights suffer from more limited range (and longer charging time for the larger batteries) compared to electric cars. For medium-duty BEV models, the Amazon and Ford E-Transits electric vans have an estimated maximum range of 240 [3], [4] and 256 km [5], respectively. One of the latest double-decker electric bus model of BYD, C10MS, has a range of up to 256 km [8]. In comparison, the Tesla model 3 electric car can travel up to 629 km on a single charge [9], which is much longer than the heavier models. Given such range deficiencies, the commercial BEV models available nowadays could already be incompetent on paper for many real-world applications. For instance, both the delivery vans and transit buses in the U.S. need to travel more than 300 km per day on average [10], for which neither the aforementioned electric van nor bus models could make it through without extra (time-consuming) charging. In addition, the operating condition of commercial applications is generally more demanding (than private use), which would compromise the technical viability of BEV models even further. For instance, en-route charging could be difficult to arrange given a tight operational schedule, and the energy efficiency (and thus also the range) of BEV would be further degraded when the vehicle is heavily loaded

The work described in this paper was supported by grants from the Research Grants Council of Hong Kong (Project No. R5007-18), National Natural Science Foundation of China (Project No. 72201214), the Sichuan Science and Technology Program (Project No. 2023NSFSC1035), the Fundamental Research Funds for the Central Universities, China (Grant No. JBK23YJ01), and UIC Start-up Research Fund (Project No. UICR0700116-25). *(Corresponding author: Minyu Shen)*

X. Bi is with Guangdong Provincial/Zhuhai Key Laboratory of IRADS and Department of Statistics and Data Science, BNU-HKBU United International College, Zhuhai 519087, China (email: xiaowenbi@uic.edu.cn)

M. Shen is with the School of Management Science and Engineering, Southwestern University of Finance and Economics, 611130, Sichuan, China (email: shenminyu@swufe.edu.cn)

W. Gu and E. Chung are with the Department of Electrical Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong (email: weihua.gu@polyu.edu.hk, and edward.cs.chung@polyu.edu.hk).

Y. Wang is with the Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong (email: yuhong.wang@polyu.edu.hk).

or the air-conditioning consumption is high [11]. The core reason for BEV's range limitation lies in BEV batteries' energy density, which is about 165 $^{\text{kWh}}/_{\text{t}}$ – notably lower than diesel fuel's 12600 $^{\text{kWh}}/_{\text{t}}$ [12]. The development of heavier BEV models is thus challenging, as larger, heavier battery packs are required, which further reduce range. Major technological advances are needed to overcome these limitations. Hence, how to operate a commercial vehicle fleet that consists of medium- and heavy-duty BEVs remains a practical yet challenging problem.

In the literature, extensive engineering solutions have been proposed to address the challenges hindering the adoption of commercial BEVs. For instance, to facilitate the electrification of public transit systems, [13] proposed a scheduling framework for the electric bus fleets, taking the interactions between the transportation and power networks into consideration. [14] developed a strategy for the operation of electric buses in the regions where heterogeneous energy resources are integrated. In [15], the configuration of electrified public transit systems that implement battery swapping technology was optimized. The electric vans and trucks, on the other hand, are commonly adopted for delivery tasks, which can be studied by modeling as electric vehicle routing problems (EVRPs) or alike. EVRP incorporates the use of EVs and possibly the planning of en-route charging into the classic vehicle routing problem (VRP), which typically seek to determine the optimal routes for the delivery vehicles to traverse a given set of customers [16]. To date, many EVRP variants have been introduced, considering the time-of-use electricity price [17], battery swapping technology [18], vehicle-to-grid technology [19], traffic condition uncertainties [20], mixed fleet [21], [22], and classic VRP features, e.g., multiple depots [23], time window constraints [24], [25], [26], and the vehicle capacity constraint [27], [28], [29].

In particular, many studies highlighted the importance of accurately estimating the energy efficiency of commercial BEVs, which is highly sensitive to the operation conditions [30], [31], [32], [33]. In practice, the operation condition is governed by various factors, including not only "exogenous" ones like the wind speed, temperature, and road conditions [34], but also "endogenous" ones like the driving speed [35], [36], [11], which is "controllable" to a certain extent, given that either the human drivers or automated driving systems could adjust the vehicle speed as needed. Of all the factors, BEV speed is also one of the most influential, known as "speed-varying range" (SVR). That is, as the speed increases, the energy efficiency of BEV degrades and the effective range diminishes. In [37], the real-world range of various BEV models at highway speeds is examined. The statistics show that all of them have varying degrees of such a mileage shrinkage. For instance, the BMW iX1 xDrive30 SUV has a real-world range of only 274 km, far below the WLTP-rated [38] range of 440 km; the Mercedes EQV 300 electric van only manages to travel 273 km before running out of power, despite having a rated range of 363 km.

However, the intrinsically endogenous BEV speed is often addressed as an exogenous factor when modeling BEVs' SVR. For instance, [31] developed a probabilistic Bayesian learning model to estimate the energy consumption of BEV based on empirical speed profiles and then optimize the delivery routes accordingly. In [33], a robust optimization model was proposed to determine the delivery routes for the BEV fleets, where the vehicle speed is considered as one of the uncertainties. [39] proposed a robust scheduling strategy for electric buses under stochastic traffic conditions, where the bus speed is given by trip time distributions. Even when BEV speed is considered as an endogenous factor and either planned "offline" [40] or controlled "online" [30], very few studies have incorporated the requirements of commercial applications concurrently. In regard to BEV adoption for delivery tasks, [32] is one of the first papers that developed EVRP variants which model BEVs' SVR by addressing the vehicle speed as an endogenous factor. The conclusion was also insightful – a flexible speed choice can greatly reduce the delivery cost when operating BEVs. However, [32] is limited due to the following assumptions:

1) The BEV can be driven at any chosen speed.
2) The BEV is not allowed to be charged in the middle of a delivery assignment.

Evidently, 1) and 2) do not always hold in practice because: 1) whether the BEV can travel at a desired speed depends on the traffic conditions; 2) as discussed above, most commercial BEV models available on the market still have limited ranges, leaving them likely in need of en-route charging to complete the delivery. In addition, since the traffic condition is uncertain by nature, the outcome of speed planning & control is uncertain, and given BEVs' SVR, whether the BEVs need to be charged en-route is also uncertain – if yes, then how to arrange the BEV charging. Apparently, the removal of assumptions 1) and 2) would result in a much more complicated but practical problem. To the best of the authors' knowledge, there has been no solutions developed to it in the literature.

In light of the above, this paper aims to bridge the research gap by introducing a data-driven approach that determines the route, speed, and (en-route) charging for delivery BEV in real time, addressing the SVR of BEV. The main contributions are summarized as follows:

1) Based on the Markov decision process (MDP), a dynamic EVRP model which aims at minimizing the total delivery time is proposed, where:
   a) The SVR of BEV is modeled to "bridge" the vehicle speed and energy consumption.
   b) The traffic condition is uncertain and disclosed dynamically during the execution of delivery task [41].
   c) The decision is to plan the delivery for BEVs in terms of: (i) which path to take; (ii) how fast to drive; and (iii) when, where, and how long to charge.

   The model is referred to as D-EVRP-SVR-SP hereinafter, where the "D" and "SP" stand for the consideration of dynamic traffic condition and speed planning decision respectively.
2) A deep reinforcement learning (DRL) approach that explicitly exploits the domain knowledge of the interdependence among the three decisions is developed to solve the high-complexity D-EVRP-SVR-SP with real-time policy.

3) Numerical studies are performed using both hypothetical and real-world instances, demonstrating that:
   a) The proposed DRL approach can efficiently find quality D-EVRP-SVR-SP policies that outperform a variety of benchmarks.
   b) BEVs' SVR can be addressed by planning the speed proactively – by slowing down the BEV at appropriate times, the en-route charging time can be reduced so that it outweighs the added travel time, thereby reducing the total delivery time.

The rest of the paper is organized as follows. In Sect. II, the components and formulation of D-EVRP-SVR-SP model are introduced. Sect. III describes the proposed DRL approach. Numerical results and discussions are presented in Sect. IV. Sect. V concludes this paper.

II. System Model and Problem Formulation

A. Overview

The D-EVRP-SVR-SP model is illustrated in Fig. 1, where the figure at left shows the application scenario, and the curve at right demonstrates the modelling of BEVs' SVR. D-EVRP-SVR-SP considers a scenario where a set of customers scatters across an area. Each of them has a deterministic demand for goods. To satisfy their demands, a single BEV is dispatched from the depot to carry out the delivery. When the BEV does not have enough goods for any of the remaining customers, it needs to return to the depot for replenishment. The BEV can also visit the charging stations within the area to top up its battery if necessary. After all the customers are served, the BEV needs to return to the depot, which marks the end of the delivery.

The objective of D-EVRP-SVR-SP is to decide on the BEV route, speed, and charging, such that the total delivery time can be minimized. The delivery time comprises both BEVs' on-road travel time and en-route charging time. In particular, the travel time depends on the BEV speed, which is governed jointly by the speed decision and uncertain traffic condition that is disclosed in real time and dictates the range of candidate speeds. For instance, the BEV may not be able to speed up if the traffic is congested and slow (represented by the dotted arrows in Fig. 1).

It is worth noting that although the energy consumption of BEV is not directly optimized, D-EVRP-SVR-SP inherently characterizes the trade-off between traveling and (en-route) charging: sustaining a higher speed allows the BEV to traverse the customers faster, but as a result of BEVs' SVR, consume more energy at the same time, potentially necessitating costly en-route charging (i.e., deadhead trips and long charging time). Hence, the total delivery time is in fact an indirect indication of the BEV energy consumption.

Sects. II-B–II-D respectively introduce the components of D-EVRP-SVR-SP, modeling of D-EVRP-SVR-SP as an MDP, and statement of the underlying optimization problem.

B. Model Components

The components of D-EVRP-SVR-SP, including a road network and an energy dynamics model for BEV, are introduced.

1) Road Network: D-EVRP-SVR-SP is defined over a complete graph $\mathcal{G} = (\mathcal{N}, \mathcal{E}, \mathcal{W})$, where $\mathcal{N}$ denotes the set of nodes (points-of-interest) that includes:

- *A depot (n_o)*: The depot is assumed to have infinite supplies of goods. The goods can be loaded onto the BEV at the depot in no time.
- *A set of customers ($\mathcal{N}_s$)*: Each customer $i \in \mathcal{N}_s$ has a deterministic demand h_i. The unloading time at a customer node is also assumed to be zero.
- *A set of charging stations ($\mathcal{N}_c$)*: All the charging stations are assumed to be equipped with level-3 chargers (e.g., 50 kW DC chargers) with the same specifications. The BEV can start charging at any station without waiting.

In addition, $\mathcal{E} = \{(i,j)|i,j \in \mathcal{N}, i \neq j\}$ denotes the set of edges that represent the road segments connecting the nodes. The last element $\mathcal{W} = \left\{(d_{ij}, \bar{v}_{ij}, \underline{v}_{ij})|(i,j) \in \mathcal{E}\right\}$ denotes the set of weights associated with the edges, where d_{ij} is the length of the edge between nodes i and j, and $\bar{v}_{ij}$, $\underline{v}_{ij}$ are the upper and lower speed limits[1]. To model the uncertain traffic conditions, $\bar{v}_{ij}$ and $\underline{v}_{ij}$ are assumed to be randomly generated when the BEV starts off on edge (i,j).

2) An Energy Dynamics model for the BEV: To characterize how the BEV's range varies with its speed, an energy dynamics model that bridges the BEV speed and state-of-charge (SoC) is adopted [11]. It should be noted that the proposed DRL approach is model-free and can thus be used in conjunction with any model proven to be accurate. The traction power of a BEV is given by:

$$P_{\text{tract}} = \left(\mu_r m g \cos\zeta + mg \sin\zeta + \frac{1}{2}\rho \mu_d f v^2 + m\dot{v}\right) \cdot v, \quad (1)$$

where μ_r is the rolling friction coefficient; m is the mass of the vehicle, consisting of curb weight m_g, and cargo weight m_c (with a maximum payload of $\bar{m}_c$); g is the gravitational acceleration; ζ is the road gradient; ρ is the air density; μ_d is the aerodynamic drag coefficient; f is the vehicle's frontal area; v and $\dot{v}$ denote the speed and acceleration of the BEV, respectively. Eq. (1) states that a running BEV must overcome the rolling, grade, aerodynamic and inertia resistances (which correspond to the four terms within the parentheses of Eq. (1) from left to right). Eq. (1) is a cubic function of v, indicating that more energy is needed to cover a trip traveled at a higher speed.

The energy required for the propulsion of a BEV is provided by its battery. Assuming the BEV is equipped with an electric motor that has an energy efficiency of η, the electric power to be drawn from the battery is estimated as:

$$P_{\text{batt}} = \frac{1}{\eta} P_{\text{tract}}. \quad (2)$$

The dynamical behavior of BEV's Lithium-ion (Li-ion) battery can be modeled using the equivalent electrical circuit methods [42]. Here, each electro-chemical cell of the battery pack is represented by a circuit consisting of a controlled open-circuit voltage source $U_{\text{oc}}(t)$ in series with a constant resistance

[1]Speed limits are considered as macroscopic attributes of the road network edges, setting aside the microscopic behaviours of individual vehicles including acceleration and deceleration.

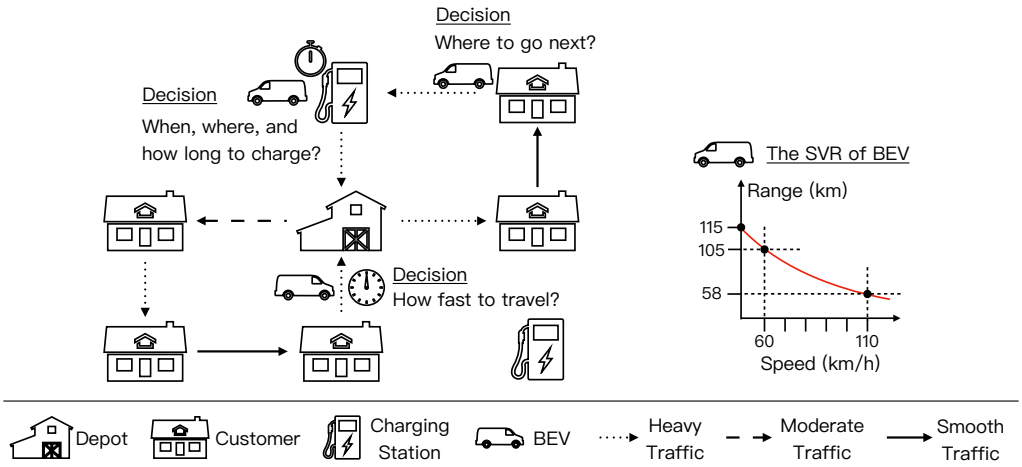


Fig. 1: Illustration of D-EVRP-SVR-SP, where the SVR curve is drawn based on the performance of Mitsubishi i-MiEV [11].

R_{oc}. An empirical relationship between $U_{\text{oc}}(t)$ and the BEV's $\text{SoC}(t)$ can then be established as:

$$U_{\text{oc}}(t) = x_1 - \frac{x_2}{\text{SoC}(t)} + x_3 \text{e}^{-x_4 Q_{\text{batt}}(1-\text{SoC}(t))}, \tag{3}$$

where x_1, x_2, x_3, and x_4 are model parameters to be estimated from the battery's discharging curves; and Q_{batt} is the capacity of the battery pack. Details are omitted here. Interested readers can refer to [42]. The $\text{SoC}(t)$ in the discharging and charging processes of the battery can be derived as follows:

- *Discharging process*: The discharging rate is given by:

$$\frac{\text{d}}{\text{d}t}\text{SoC}(t) = -\frac{I(t)}{Q_{\text{batt}}}, \tag{4}$$

where $I(t)$ is the circuit current at time t, obtained as:

$$I(t) = \frac{U_{\text{oc}}(t) - \sqrt{U_{\text{oc}}^2(t) - 4 \cdot R_{\text{oc}} \cdot P_{\text{cell}}}}{2 \cdot R_{\text{oc}}}, \tag{5}$$

where P_{cell} is the electrical power supplied by each cell. Assuming the cells are connected in a N_p-parallel-and-N_s-series structure (where N_p is the number of parallel series connections of cells and N_s is the number of cells in each serial connection), and the battery energy is evenly distributed across them, P_{cell} can then be calculated as $P_{\text{batt}}/(N_p \cdot N_s)$. Combining Eq. (1) – (5) with the speed profile and the battery's initial conditions will give the $\text{SoC}(t)$ at any time t when the BEV is traveling.

- *Charging process*: The cells of Li-ion battery are charged with firstly a constant current (CC) I_{cc} and then a constant voltage (CV) U_{cv}. During the CC phase, the charging curve is roughly linear. Thus, $\text{SoC}(t)$ is given by:

$$\text{SoC}(t) = \text{SoC}_0 - \frac{I_{\text{cc}} \cdot t}{Q_{\text{batt}}}, \forall t < t_s, \tag{6}$$

where SoC_0 is the initial SoC when the charging starts; t_s is the time when the terminal voltage, $U(t)$, reaches U_{cv}. In other words, $U(t) < U_{\text{cv}}$ if $t < t_s$, and $U(t) = U_{\text{cv}}$ if $t \geq t_s$. The $U(t)$ is supplied by the charger. Note that I_{cc} is negative, indicating that the current flows from $U(t)$ to the voltage source $U_{\text{oc}}(t)$. After $U(t)$ increases to U_{cv}, the charger switches to the CV phase and the charging rate gradually decreases as follows:

$$\frac{\text{d}}{\text{d}t}\text{SoC}(t) = \frac{U_{\text{cv}} - U_{\text{oc}}(t)}{Q_{\text{batt}} \cdot R_{\text{oc}}}, \forall t \geq t_s. \tag{7}$$

Combining Eq. (3), (6), and (7) gives the $\text{SoC}(t)$ at any time during a charging process.

C. D-EVRP-SVR-SP as an MDP

D-EVRP-SVR-SP is modeled as a T-step (where T is a random variable) finite-horizon MDP defined by a 4-tuple $\{\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}\}$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces; $\mathcal{T}$ is a set of state transition rules; and $\mathcal{R}$ is the reward function. Detailed definitions are given as follows.

1) State: $s_k = \left\{m_c^{(k)}, \text{SoC}^{(k)}, n^{(k)}, \mathcal{H}^{(k)}\right\} \in \mathcal{S}$ defines the state at step k when the BEV has completed the activity at its present location (i.e., cargo loaded at the depot, unloaded at a customer node, or refueled at a charging station), and is departing to continue the delivery. As mentioned before, $m_c^{(k)}$ and $\text{SoC}^{(k)}$ are the present cargo weight and battery SoC of the BEV, respectively; $n^{(k)}$ is the node where the BEV currently stays; and $\mathcal{H}^{(k)} = \left\{h_i^{(k)} | i \in \mathcal{N}_s\right\}$ is the set of remaining customer demands. Two types of terminal states (sub-scripted with T, i.e., $\mathcal{S}_T \subset \mathcal{S}$) are defined as:

- $\text{SoC}^{(k)} > 0$, $n^{(k)} = n_o$, and $h_i^{(k)} = 0, \forall i \in \mathcal{N}_s$, implying that the BEV has completed the delivery and returned to the depot.
- $\text{SoC}^{(k)} = 0$, suggesting that the BEV has run out of its power before completing the delivery.

2) Action: $a_k = \left(n^{(k+1)}, v^{(k,k+1)}, \tau_c^{(k+1)}\right) \in \mathcal{A}$ specifies the action which the BEV needs to perform at step k, where $n^{(k+1)} \in \mathcal{N} \setminus \left\{n^{(k)}\right\}$ is the next node to visit; $v^{(k,k+1)} \in \phi$ is the average speed to maintain[2] for the trip from $n^{(k)}$ to $n^{(k+1)}$, where ϕ is a set of discrete speeds; and $\tau_c^{(k+1)} \in \psi$ is the charging time at $n^{(k+1)}$ with ψ being a set of discrete charging times, which includes 0 indicating no charging.

[2]The acceleration $\dot{v}^{(k,k+1)}$ is assumed to be 0, i.e., no inertia resistance.

3) Transition: The MDP transits from state s_k to s_{k+1} by performing action a_k (i.e., $s_{k+1} = \mathcal{T}(s_k, a_k)$). The features of s_{k+1} are updated as follows:

- *Cargo weight ($m_c^{(k+1)}$)*:

$$m_c^{(k+1)} = \begin{cases} \bar{m}_c & \text{if } n^{(k+1)} = n_o \\ m_c^{(k)} - h_{n^{(k+1)}}^{(k)} & \text{if } n^{(k+1)} \in \mathcal{N}_s \\ m_c^{(k)} & \text{if } n^{(k+1)} \in \mathcal{N}_c. \end{cases} \tag{8}$$

- *BEV SoC ($\text{SoC}^{(k+1)}$)*:

$$\text{SoC}^{(k+1)} = \text{SoC}^{(k)} - \begin{cases} \text{SoC}_d + \text{SoC}_c & \text{if } n^{(k+1)} \in \mathcal{N}_c \\ \text{SoC}_d & \text{otherwise,} \end{cases} \tag{9}$$

where $\text{SoC}_d = \text{DP}\left(\text{SoC}^{(k)}, m_c^{(k)}, v_s^{(k,k+1)}, d^{(k,k+1)}\right)$ and $\text{SoC}_c = \text{CP}\left(\text{SoC}^{(k)}, \tau_c^{(k+1)}\right)$ are the SoC variations incurred by BEV discharging or charging during the transition, with $\text{DP}(\cdot)$ and $\text{CP}(\cdot)$ being the functions which respectively encapsulate all relevant calculations that model the discharging and charging processes as in Sect. II-B2. $d^{(k,k+1)}$ is the distance between $n^{(k)}$ and $n^{(k+1)}$. $v_s^{(k,k+1)}$ denotes the average speed of BEV when traveling from $n^{(k)}$ to $n^{(k+1)}$. The value of $v_s^{(k,k+1)}$ depends not only on the action $v^{(k,k+1)}$, but also the stochastic traffic conditions. Let $\bar{v}$ and $\underline{v}$ be the associated upper and lower speed limits, then $v_s^{(k,k+1)}$ is given by:

$$v_s^{(k,k+1)} = \begin{cases} \underline{v} & \text{if } v^{(k,k+1)} < \underline{v} \\ v^{(k,k+1)} & \text{if } \underline{v} \le v^{(k,k+1)} \le \bar{v} \\ \bar{v} & \text{if } v^{(k,k+1)} > \bar{v}. \end{cases} \tag{10}$$

The transitions are thus stochastic considering the uncertainties in speed limits. In this work, no domain knowledge regarding the transition probability or the traffic conditions is assumed.

- *Next node ($n^{(k+1)}$)*: $n^{(k+1)}$ is as given by the associated action.
- *Customer demands ($\mathcal{H}^{(k+1)}$)*: If $n^{(k+1)} \in \mathcal{N}_s$, the demand of $n^{(k+1)}$ is fulfilled and the associated element in $\mathcal{H}^{(k+1)}$ is updated accordingly, i.e., $h_{n^{(k+1)}}^{(k+1)} = 0$. Here, demand splitting is not considered, i.e., each customer can only be visited once.

4) Reward: The transition from state s_k to s_{k+1} incurs a step reward $r_k = \mathcal{R}(s_k, a_k, s_{k+1})$, which is defined as:

$$r_k = \frac{d^{(k,k+1)}}{v_s^{(k,k+1)}} + \begin{cases} \tau_c^{(k+1)} + \epsilon(s_{k+1}) & \text{if } n^{(k+1)} \in \mathcal{N}_c \\ \epsilon(s_{k+1}) & \text{otherwise,} \end{cases} \tag{11}$$

where $\epsilon(\cdot)$ is a penalty function of s_{k+1}, which returns a large positive constant $\bar{\epsilon}$ if BEV's battery is depleted at step $k+1$ (governed by the battery discharging process in Sect. II-B2), and 0 otherwise. Eq. (11) incentivizes the minimization of time cost (i.e., travel time and charging time (if applicable) of BEV) and constraint violation resulted by the state transition.

D. Problem Statement for D-EVRP-SVR-SP

As an MDP, D-EVRP-SVR-SP seeks a stochastic policy π that maps the state s_k to the action a_k (i.e., $\pi(a_k|s_k)$), such that the following objective J, the expectation of episodic return G, is maximized (i.e., minimizing the total delivery time and constraint violations):

$$\max_{\pi} \; J = \mathbf{E}[G] = \mathbf{E}\left[\sum_{k=0}^{T} \gamma^k r_k\right], \tag{12}$$

where $\gamma \in [0, 1]$ is the factor for discounting future rewards. A smaller γ favors a more "short-sighted" decision-making.

III. Methodology

The proposed DRL approach for solving D-EVRP-SVR-SP follows the actor-critic paradigm, in which the actor and critic learn a policy and a value function respectively. To collect transition samples for agent training, the actor interacts with an "environment" that implements the MDP of D-EVRP-SVR-SP. In the following subsections, the building blocks of our approach, including the state embedding method, the design of actor and critic, and training algorithm, are respectively introduced in detail.

A. State Embedding

In order to facilitate the agent training, the state s_k is at first processed with an embedding layer, where each state feature is normalized and mapped into a D_h-dimensional space. The embedded state $\tilde{s}_k$ is given by:

$$\tilde{s}_k = \Bigg[F_m\left(\frac{m_c^{(k)}}{\bar{m}_c}\right) \oplus F_s\left(\text{SoC}_c^{(k)}\right) \oplus F_n\left(n^{(k)}\right) \oplus F_h\left(\left\{\frac{h_i^{(k)}}{\sum_{j\in\mathcal{N}_s} h_j^{(k)}} | i \in \mathcal{N}_s\right\}\right)\Bigg], \tag{13}$$

where $F_m(\cdot)$, $F_s(\cdot)$, $F_n(\cdot)$[3], and $F_h(\cdot)$ are non-linear mappings (e.g., multi-layer perceptron with ReLU activation) that embed the respective features; $\oplus$ is the vector concatenation operator. To facilitate a better reasoning, a vector of distances from $n^{(k)}$ to all the other nodes is also embedded and concatenated to $\tilde{s}_k$. The output of embedding layer $\hat{s}_k$ is thus:

$$\hat{s}_k = \left[\tilde{s}_k \oplus F_d\left(\left\{\frac{d_{n^{(k)},i}}{\bar{d}} | i \in \mathcal{N}\backslash\{n^{(k)}\}\right\}\right)\right], \tag{14}$$

where $\bar{d} = \max_{(i,j)}(\{d_{ij}|(i,j) \in \mathcal{E}\})$ is the maximum distance between any two nodes; $F_d(\cdot)$ is a non-linear mapping.

B. Actor

1) Coupled Decision-making: In D-EVRP-SVR-SP, the action space has three intuitively coupled dimensions. Determining which node the BEV should be visiting next (i.e., $n^{(k+1)}$) designates an upper bound for the step reward because the BEV cannot travel faster than the maximum speed (see Eq. (10)). The speed choice for the BEV, $v^{(k,k+1)}$, depends on the selection of $n^{(k+1)}$, or more specifically, the time-varying traffic condition and the distance to $n^{(k+1)}$. In order to improve J, the BEV may either speed up to reduce the travel time

[3]Since the location of BEV is a discrete node index, it is firstly written as a one-hot vector, and then fed into $F_n(\cdot)$.

(i.e., receiving a high immediate reward), or proactively slow down to reduce the energy consumption and consequently the charging time (i.e., anticipating a high future reward). Actions $n^{(k+1)}$ and $v^{(k,k+1)}$ jointly update the BEV status, e.g., SoC. The updated status and the delivery progress indicated by the remaining customer demands are then key references to the charging time decision $\tau_c^{(k+1)}$ given $n^{(k+1)} \in \mathcal{N}_c$.

In light of the above, the actor is designed based on the architecture of modified MDP [43], such that the "coupled" relationship between the three decisions is utilized. Specifically, the actor is composed of three M-layer residual neural networks [44]. They are referred to as the node network (left), speed network (middle), and charging network (right). Each network's residual layer has D_r neurons. Let y_i be the input to the i-th residual layer, and $F(\cdot)$ be a non-linear mapping, y_{i+1} can be obtained as:

$$y_{i+1} = F(y_i) + y_i. \tag{15}$$

Considering the embedded state at step k, $\hat{s}_k$, the probability distributions over the three dimensions of action a_k (i.e., the node, speed, and charging time, denoted by P_{node}, P_{speed}, and P_{charge}) are given by the three networks and the follow-up (masked) softmax activation successively as:

$$P_{\text{node}}(\hat{s}_k) = \text{softmax}\left\{\text{RN}_n\left[\hat{s}_k\right] + \log\left[Z_n^{(k)}\right]\right\}, \tag{16a}$$

$$P_{\text{speed}}(\hat{s}_k, n^{(k+1)}) = \text{softmax}\left\{\text{RN}_s\left[\hat{s}_k \oplus F_n\left(n^{(k+1)}\right)\right]\right\}, \tag{16b}$$

$$P_{\text{charge}}(\hat{s}_k, n^{(k+1)}, v^{(k,k+1)}) = \text{softmax}\Bigg\{\text{RN}_c\Bigg[\hat{s}_k \oplus F_n\left(n^{(k+1)}\right) \oplus F_v\left(\frac{v^{(k,k+1)}}{\max \phi}\right)\Bigg] + \log\left[Z_c^{(k)}\right]\Bigg\}, \tag{16c}$$

where $\text{softmax}\{\cdot\}$ is the activation function; $\text{RN}_n[\cdot]$, $\text{RN}_s[\cdot]$, and $\text{RN}_c[\cdot]$ are the operations of processing the input with M stacked residual layers using the respective networks. Vectors $Z_n^{(k)} \in \{0,1\}^{|\mathcal{N}|}$ and $Z_c^{(k)} \in \{0,1\}^{|\psi|}$ are respectively referred to as the node mask and charging time mask, where $|\cdot|$ is the cardinality of a set. They are built based on the raw state features to filter out the node and charging time options that are invalid at step k. The "infeasible" options are indexed by 0, and thus assigned with a probability approaching 0. Details on how these two masks are built are introduced later in this subsection.

Based on P_{node}, P_{speed}, and P_{charge}, the action n^{k+1}, $v^{(k,k+1)}$, and $\tau_c^{(k+1)}$ can be determined one-by-one through either acting greedily (i.e., picking with the highest probability) for evaluation, or random sampling for training (e.g., $n^{(k+1)} \sim P_{\text{node}}(\hat{s}_k)$). In particular, n^{k+1} and $v^{(k,k+1)}$ are processed by non-linear mappings $F_n(\cdot)$ and $F_v(\cdot)$ (which maps scalar speed $v^{(k,k+1)}$ to a D_h-dimensional vector), and then concatenated to $\hat{s}_k$ as "intermediate" feature inputs to either the speed or charging network. Eq. (16b) and (16c) essentially state that the selection of $v^{(k,k+1)}$ and $\tau_c^{(k+1)}$ is conditional on the preceding decision(s), aiming to capture the underlying interdependence. To proceed to the next state, action a_k is obtained by concatenating the three decisions and feeding them into the D-EVRP-SVR-SP environment.

2) Masking Scheme: To accelerate the training process and ensure the feasibility of solutions, a node mask and a charging mask are built whenever the actor is about to make a move. The masking procedures are introduced as follows:

- *Node mask ($Z_n^{(k)}$)*: Let $z_i^{(k)}$ be the element of $Z_n^{(k)}$ corresponding to node i at step k, the value of which is determined according to the node type and state s_k:
  - (i) $i = n^{(k)}$: $z_i^{(k)} = 0$.
  - (ii) $i = n_o$ or $i \in \mathcal{N}_c$: $z_i^{(k)} = 0$ if $n^{(k)}$ and i are of the same type, and 1 otherwise.
  - (iii) $i \in \mathcal{N}_s$: $z_i^{(k)} = 0$ if $h_i^{(k)} = 0$, and 1 if $0 < h_i^{(k)} \leq m_c^{(k)}$.

  where $n^{(k)}$, $h_i^{(k)}$, and $m_c^{(k)}$ are given by s_k.
- *Charging time mask ($Z_c^{(k)}$)*: If $n^{(k)} \in \mathcal{N}_c$, the elements of $Z_c^{(k)}$ associated with non-zero charging time options (in set ψ) are set to 1. Otherwise, no charging is the only feasible option.

C. Critic

To facilitate the parameter optimization of the actor (such as by reducing the gradient variance), the critic evaluates the performance of the actor by learning the state-value function V throughout the training. Like the node/speed/charging network of the actor, critic is also an M-layer residual network that takes the embedded state $\hat{s}_k$ as input. Yet, the output is a scalar state value $V(s_k)$. It should be noted that the critic uses a different embedding layer than the actor, which implies that the critic interprets the state from another perspective.

D. Training Algorithm

Let θ and ω denote the trainable parameters of the actor and critic (including both the embedding layers and residual networks), the targeting stochastic policy π and state-value function V are essentially π_θ and V_ω, where the subscripts represent the parameterization.

Based on the policy gradient theorem [45], the actual gradient of objective J with respect to θ is estimated as:

$$\nabla_\theta J = \frac{1}{B}\sum_{b=1}^{B}\left[\sum_{k=0}^{T}\left(G_k - V_\omega(s_k)\right)\nabla_\theta \log\left(\pi_\theta\left(a_k|s_k\right)\right)\right], \tag{17}$$

where B is the number of D-EVRP-SVR-SP episodes (i.e., sequences of all transitions from initial to terminal state) that are simulated following π_θ; G_k denotes the total reward accumulated from step k until any terminal state is reached. With a_k being a three-dimensional action, the logarithm of $\pi_\theta(a_k|s_k)$ is:

$$\log\left(\pi_\theta\left(a_k|s_k\right)\right) = \log\left(p_n^{(k)}\right) + \log\left(p_s^{(k)}\right) + \log\left(p_c^{(k)}\right), \tag{18}$$

where $p_n^{(k)} \in P_{\text{node}}$, $p_s^{(k)} \in P_{\text{speed}}$, and $p_c^{(k)} \in P_{\text{charge}}$ denote the probabilities associated with each element of a_k. Note that the probability vectors are in fact also functions of θ (e.g., P_{node}

should be written as $P_{\text{node},\theta}$). The subscript θ is omitted here for readability.

To prevent the premature convergence of π_θ and encourage the actor's exploration, $\nabla_\theta J$ is further regularized as:

$$\nabla_\theta \hat{J}(\lambda) = \nabla_\theta J + \lambda \nabla_\theta E(\pi_\theta), \tag{19}$$

where λ controls the strength of exploration, the value of which decays exponentially as the training progresses (i.e., from λ_{ini} to λ_{end}, at a rate controlled by a hyper-parameter α); $E(\pi_\theta)$ is the entropy of π_θ given by:

$$E(\pi_\theta) = -\frac{1}{B}\sum_{b=1}^{B}\Bigg\{\sum_{k=0}^{T}[\langle P_{\text{node}}, \log(P_{\text{node}})\rangle + \langle P_{\text{speed}}, \log(P_{\text{speed}})\rangle + \langle P_{\text{charge}}, \log(P_{\text{charge}})\rangle]\Bigg\}, \tag{20}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of vectors.

On the other hand, the parameters of the critic ω can be updated with an aim of minimizing the mean squared error (MSE) of the state value estimation:

$$L_{\text{critic}} = \frac{1}{B}\sum_{b=1}^{B}\left[\sum_{k=0}^{T}\left(G_k - V_\omega(s_k)\right)^2\right]. \tag{21}$$

Based on above gradient estimations, a training algorithm is proposed as in Algorithm. 1 to optimize the agent parameters.

Algorithm 1: The training of DRL agent

input : I - total number of iteration; B - number of simulation episode; γ - discount factor; $\beta_{\text{actor}}, \beta_{\text{critic}}$ - agent learning rates; $\lambda_{\text{ini}}, \lambda_{\text{end}}, \alpha$ - exploration strength

output: Trained actor parameter θ^*

Initialize the agent parameters θ, ω randomly, and $\lambda \leftarrow \lambda_{\text{ini}}$;
for $i \leftarrow 1$ **to** I **do**
  Initialize an episode buffer $\mathcal{U} \leftarrow \emptyset$;
  for $b \leftarrow 1$ **to** B **do**
    Initialize an episode $u \leftarrow \emptyset$;
    for $k \leftarrow 1$ **to** T **do**
      For non-terminal state s_k, sample a node for BEV to visit: $n^{(k+1)} \sim P_{\text{node}}(\hat{s}_k)$; /* Eq. (16a) */
      Sample the BEV speed: $v^{(k,k+1)} \sim P_{\text{speed}}(\hat{s}_k, n^{(k+1)})$; /* Eq. (16b) */
      Sample the BEV charging time: $\tau_c^{(k+1)} \sim P_{\text{charge}}(\hat{s}_k, n^{(k+1)}, v^{(k,k+1)})$; /* Eq. (16c) */
      Take action $a_k \leftarrow \left(n^{(k+1)}, v^{(k,k+1)}, \tau_c^{(k+1)}\right)$ in D-EVRP-SVR-SP environment to receive the next state $s_{k+1} \leftarrow \mathcal{T}(s_k, a_k)$, and step reward $r_k \leftarrow \mathcal{R}(s_k, a_k, s_{k+1})$;
      Add transition $\{s_k, a_k, s_{k+1}, r_k\}$ to episode u;
    end
    Add episode u to buffer $\mathcal{U}$;
  end
  Based on the episodes in $\mathcal{U}$, update θ and ω as: $\theta \leftarrow \theta + \beta_{\text{actor}} \nabla_\theta \hat{J}(\lambda)$, and $\omega \leftarrow \omega + \beta_{\text{critic}} \nabla_\omega L_{\text{critic}}$; /* Eq. (19) and (21) */
  Perform entropy annealing as: $\lambda \leftarrow \lambda_{\text{end}} + (\lambda_{\text{ini}} - \lambda_{\text{end}}) \cdot \mathrm{e}^{-\frac{i}{\alpha}}$;
end

IV. Numerical Results and Discussions

This section aims to: 1) evaluate the performance of the proposed DRL approach; and 2) justify the significance of considering speed planning and SVR modeling in BEV-related operational problems. Both the D-EVRP-SVR-SP environments and the DRL model are implemented in Python (with Pytorch). All the experiments are performed on a Dell workstation with Intel Xeon Gold 6126 CPU (2.60 GHz × 24) and 64 GB DDR4 memory.

A. Settings

1) DRL: The actor and critic map each state or intermediate feature (e.g., BEV speed sampled for coupled decision-making) into a 32-dimensional vector space (i.e., $D_h = 32$). Despite the differences in input and output dimensions, all the agent networks (i.e., node networks of the actor and value network of the critic) have $M = 4$ residual layers, each of which has a size of $D_r = 256$. In each training iteration, $B = 32$ episodes of the problem instance are generated, and the model parameters are updated using the Adam optimizer at a rate of $\beta = 2e-4$. The training is performed for $I = 1500$ iterations in total. The exploration strength decays from $\lambda_{\text{ini}} = 0.05$ to $\lambda_{\text{end}} = 0.01$ within approximately 550 iterations, which is given by $\alpha = 150$. Note that the above hyper-parameters are selected for achieving the best overall performance across various setups, as identified through extensive grid searches.

2) D-EVRP-SVR-SP Environment: The following two classes of D-EVRP-SVR-SP instances are considered for numerical studies:

- *Hypothetical instances*: 120 hypothetical D-EVRP-SVR-SP instances with three different settings (i.e., 40 instances for each) are randomly generated. For the convenience of discussion, they are named as "C3-S6-SoC0.4", "C4-S10-SoC0.6", and "C5-S14-SoC0.8" to indicate the difference. For example, in a "C5-S14-SoC0.8" instance there are 5 charging stations and 14 customers (i.e., 20 nodes in total including the depot), and the BEV departs the depot with 80% SoC. The coordinates of the nodes are sampled uniformly within a square area of $[0, 50000]$ (m) $\times$ $[0, 50000]$ (m). The different initial SoC represents real-world operational scenarios where the

BEV may not start the delivery with a fully-charged battery, e.g., overnight or cross-shift deliveries, insufficient charging infrastructures for the entire fleet.

- *A real-world instance*: A D-EVRP-SVR-SP instance in real-world scale is built that encapsulates 13 main service points of JD Logistics [46] (as customer nodes) and 7 public charging stations in Beijing, China. The locations of the service points and charging stations are indicated in latitude and longitude via Google Maps. Google Distance Matrix API [47] is then adopted to obtain the distance adjacency matrix. Please refer to Appendix. A for more details on the instance. Following the naming of hypothetical instances, it is hereinafter referred to as "JD-C7-S13-SoC1.0".

The following settings are fixed for all instances:

- For each customer, a continuous number is uniformly sampled from 0.1 to 0.4 to indicate the demand as a percentage of the BEV's maximum payload.
- The speed and charging time set are $\phi = \{5, 10, ..., 30\}$ ($^{\text{m}}/_{\text{s}}$) and $\psi = \{300, 600, ..., 1800\}$ (s), respectively. When the BEV departs a node, two distinct speeds will be sampled from ϕ as the upper and lower speed limits, respectively. It should be noted that the top speeds in set ϕ (e.g., $25, 30$ $^{\text{m}}/_{\text{s}}$) are valid for JD-C7-S13-SoC1.0 instance, given the fact that the nodes are mostly located outside the Fifth Ring Road of Beijing, where the speed limit is up to 100 $^{\text{km}}/_{\text{h}}$(≈ 28 $^{\text{m}}/_{\text{s}}$) [48]. The roads in the network are assumed to be flat, i.e., gradient $\zeta = 0$ (rad).
- Function $\epsilon(\cdot)$ returns a constant $\bar{\epsilon} = 50000$ to penalize actions that result in constraint violation (i.e., BEV running out of power en-route). The episodic return G is discounted by $\gamma = 0.99$.
- Nissan E-NV200 electric van is adopted for the delivery, and the charging stations are all installed with 50 kW DC fast chargers. Based on the battery characteristics (e.g., discharging profile) provided by the supplier Automotive Energy Supply Corporation [49] and vehicle specifications provided by Nissan [50], the parameters associated with its energy dynamics are derived as listed in Tab. I. In particular, one can refer to [42] for more details on the modelling of the battery behaviour (see Eq. (3)). In Appendix. B, our energy dynamics model for E-NV200 is further validated.

TABLE I: Parameters associated with the energy dynamics of E-NV200.

Parameters (Notations)	Values (Units)
Curb weight (m_g)	1500 (kg)
Maximum payload ($\bar{m}_c$)	700 (kg)
Rolling friction coefficient (μ_r)	0.01
Aerodynamic drag coefficient (μ_d)	0.4
Gravitational acceleration (g)	9.81 ($^{\text{m}}/_{\text{s}^2}$)
Vehicle's frontal area (f)	1.75 (m^2)
Air density (ρ)	1.055 ($^{\text{kg}}/_{\text{m}^3}$)
Electric motor's energy efficiency (η)	0.9
Capacity of battery pack (Q_{batt})	32.5 (Ah)
Resistance (R)	1.033e-3 (Ω)
Number of parallel series connections of cells (N_p)	2
Number of cells in each serial connection (N_s)	96
Constant current (I_{cc})	-62.5 (A)
Constant voltage (U_{cv})	4.015 (V)

B. Algorithmic Performance on Hypothetical Instances

In order to comprehensively evaluate the algorithmic performance of proposed DRL approach, numerical comparative studies based on the large variety of hypothetical D-EVRP-SVR-SP instances are performed in this subsection.

1) Convergence Performance: For demonstrative purpose, the proposed DRL approach is hereinafter termed as "DRL-CPD", where "CPD" stands for coupled decision-making. The assessment is performed by comparing DRL-CPD with two DRL-based alternatives:

- *DRL with "plain" actor (DRL-PLN):* DRL-PLN is the same as DRL-CPD except that the former's actor is represented with a single residual network (instead of three) with $M = 6$, $D_r = 256$. The input to the actor is still the embedded state, while the output has a dimension of $|\mathcal{N}| + |\phi| + |\psi|$, meaning that the probability distributions associated with the three-dimensional decisions are generated concurrently.
- *DRL with attention-based actor (DRL-ATTN):* DRL-ATTN extends the attention-based model presented in [51] for the node network. Since the model has showcased promising results when tackling VRP variants [52], DRL-ATTN could be a rival to DRL-CPD. It is introduced as in Appendix. C.

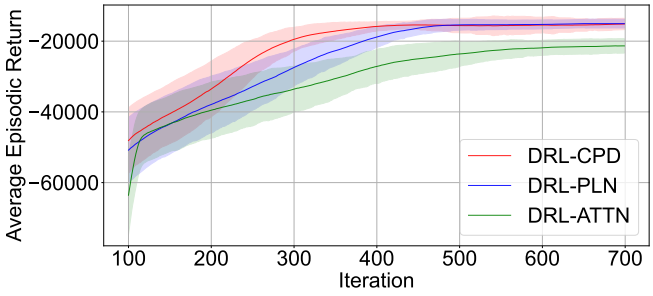


Fig. 2: Learning curves of DRL-based methods.

Fig. 2 presents typical learning curves of DRL-CPD, DRL-PLN, and DRL-ATTN for solving a C5-S14-SoC08 instance. Compared to DRL-CPD and DRL-PLN, DRL-ATTN learned more slowly, and converged to a policy that falls behind in average episodic return ($B = 32$), accounting for the total delivery time and penalty as given by Eq. (11) and (12). The possible reason is as follows. The original design of DRL-ATTN [51] aimed at producing policies that can generalize well to problem instances with unseen settings, i.e., the trained policy acted like a meta algorithm. However, our goal is to find quality policy for specific D-EVRP-SVR-SP instances, guiding the BEV delivery under uncertain traffic conditions. The increased model complexity by incorporating attention

mechanism seems to only make DRL-ATTN more difficult to train.

Although the convergent policies of DRL-CPD and DRL-PLN appear to have comparable average returns, DRL-CPD costs nearly 50 less training iterations, or equivalently 1600 episodes of interactions with the D-EVRP-SVR-SP environment. This observation shows that the structure of coupled decision-making endows DRL-CPD a better "understanding" on the interdependence of decisions, and hence an improved overall sample efficiency.

The training of DRL-CPD policies for C3-S6-SoC0.4, C4-S10-SoC0.6, C5-S14-SoC0.8 instances for $I = 1500$ iterations takes around 20, 40, and 50 minutes, respectively. Once the policy is trained offline, it could be deployed online to suggest real-time route, speed, and charging decisions within a blink of an eye (i.e., within a second).

2) Quality of Solution Policies: The DRL-CPD policies are compared against DRL-PLN, DRL-ATTN policies, and two baseline policies, including: 1) a heuristic policy devised based on the optimal solutions of a branch-and-price method (BPH); and 2) a DRL policy trained using DRL-CPD with the speed network disabled (SND), i.e., no speed planning:

- *BPH:* Considering the stochastic nature and large solution space of D-EVRP-SVR-SP, developing an optimal policy using conventional optimization approaches directly would be very difficult. A branch-and-price method for capacitated VRP (CVRP) [53] is thus implemented to find the optimal *static* routes for the BEV to traverse the customer nodes. The BPH policy combines the optimal routes with a set of parameterized decision rules to solve D-EVRP-SVR-SP. Please refer to Appendix. D for more implementation details.
- *SND:* The SND policy is trained under the same training diagram and settings as DRL-CPD, except that the speed network is disabled and a deterministic speed of 30 $\mathrm{m/s}$ (which is the maximum possible speed in DRL-CPD as defined in Sect. IV-A2) is assigned whenever an action needs to be taken.

For the 120 instances of cases C3-S6-SoC0.4, C4-S10-SoC0.6, and C5-S14-SoC0.8, the DRL-CPD, DRL-PLN, DRL-ATTN, and SND models are respectively trained, and the BPH hyper-parameters are searched. Each obtained policy is applied to roll out the associated D-EVRP-SVR-SP instance for 30 episodes, where the traffic conditions (i.e., speed limits) are randomly generated. The resulting delivery times are averaged out to serve as the performance indicator. The mean delivery times and improvements by the DRL-CPD policies (over the others) across all instances under different settings are reported in Tab. II. Several observations are worth noting:

- The convergent policies of DRL-CPD outperform those of DRL-ATTN by roughly 22% on average (which is consistent with Fig. 2), and achieve relative improvements of up to 4.2% (C3-S6-SoC0.4) over those of DRL-PLN. This observation manifests that the proposed structure of coupled decision-making also helps in finding better policies.
- The DRL-CPD policies reduce the mean delivery times by up to 6.3% (C5-S14-SoC0.8) compared to the BPH policies. Although BPH policies guide the BEV mostly based on the optimal routes generated by branch-and-price, they still trail because DRL-CPD manages the three decisions in a more comprehensive way. Note-worthily, the grid search of BPH takes 32400 episodes of rollouts (as specified in D), which is much more than DRL-CPD's 19200 – the training converges within 600 iterations as shown in Fig. 2).
- DRL-CPD outperforms SND by up to 3.2% (C3-S6-SoC0.4), which shows that in the context of EVRP, maintaining a high speed does not necessarily save the delivery time due to BEVs' SVR. The possible reason is that, by slowing down the BEV, the time spent on en-route charging can be reduced along with the energy consumption, and the saving in charging time and the charging-associated detours outweighs the added travel time. To validate our speculations, the analysis of DRl-CPD, BPH, and SND policies are furnished next.

Remark. *The trainings of all DRL agents are based on the same hyper-parameters settings in Sect. IV-A1. Tailoring the setting for each individual D-EVRP-SVR-SP instance (e.g., via further grid search) may yield a better policy, but this is beyond the scope of this work.*

3) Analysis of Solution Policies: In Tabs. III, IV, and V, the performance of DRL-CPD, BPH, and SND policies is respectively compared for one selected instance under each setting (i.e., S3-C6-SoC0.4, S4-C10-SoC0.6, and S5-C14-SoC0.8)[4]. The comparison is based on their average performance over 50 episodes of rollouts.

The results show that the DRL-CPD policies outperform the other two for minimizing the overall delivery time in all three instances, since they enable the delivery to be completed with a shorter en-route charging time. For instance, in the C5-S14-SoC0.8 instance, the DRL-CPD policy averages a charging time of 2160 seconds which is shorter than BPH's 2656.3 and SND's 2376 seconds. Despite scheduling the least amount of charging time, the DRL-CPD policies are in fact also fairly reliable, with no BEV depleting its power en-route (i.e., no infeasible episodes). In contrast, the BEV is more prone to be under-charged under the BPH policy particularly when the problem scale is large (e.g., 2 infeasible episodes for BPH policy in C5-S14-SoC0.8 instance).

Further examination of the tables shows that, the DRL-CPD policies behave very differently across different instances in terms of how to reduce the total delivery time. In particular, we consider that the DRL-CPD policies of S3-C6-SoC0.4 and S5-C14-SoC0.8 instances have similar characteristics which are different from that of S4-C10-SoC0.6 instance. Under the first two policies, though the BEV is not required to sustain the same maximum speed as with SND policy, a fairly high speed needs to be maintained to ensure the delivery won't be delayed too much on the road. Taking the S3-C6-SoC0.4 instance as an example, the BEV averages a travel speed of 20.5 and 21.4 $\mathrm{m/s}$ by following the DRL-CPD and SND

[4]Note that Tab. II compares the average policy performance across multiple instances under each setting.

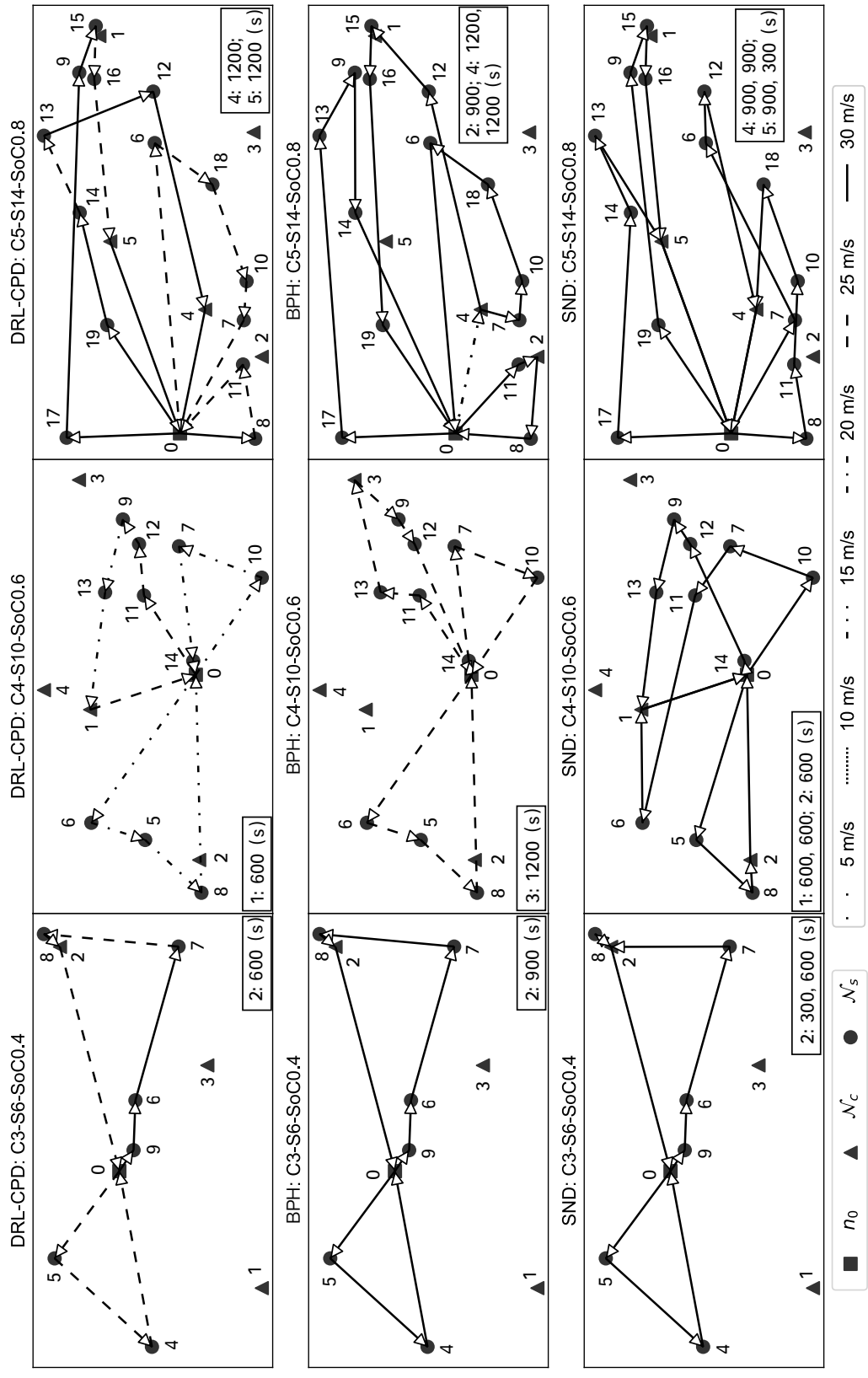


Fig. 3: The comparison of DRL-CPD, BPH, and SND policies on solving hypothetical instances.

TABLE II: Comparing DRL-CPD with DRL-based alternatives and baseline policies.

	Mean Total Delivery Time (second)					Improvements by DRL-CPD			
	DRL-CPD	-PLN	-ATTN	BPH	SND	-PLN	-ATTN	BPH	SND
C3-S6-SoC0.4	10438	10894	15195	10817	10784	4.2%	31.3%	3.5%	3.2%
C4-S10-SoC0.6	14695	14741	18855	15119	14764	0.3%	22.1%	2.8%	0.5%
C5-S14-SoC0.8	18503	18766	21783	19745	18796	1.4%	15.1%	6.3%	1.6%

TABLE III: The performance of policies for C3-S6-SoC0.4 instance.

Policies	Delivery Time (second)			#Infeasible	Energy	Energy	Travel	Travel
	Overall	Travel	Charging	Episodes	Consumption (Wh)	Efficiency ($^{\text{Wh}}/_{\text{km}}$)	Distance (km)	Speed ($^{\text{m}}/_{\text{s}}$)
DRL-CPD	**6271.2**	5671.2	**600.0**	**0**	**13240.2**	**113.9**	116.2	20.5
BPH	6666.9	5748.9	918.0	**0**	13435.0	115.9	**116.0**	20.2
SND	6335.5	**5435.5**	900.0	**0**	14323.5	123.0	116.4	21.4

TABLE IV: The performance of policies for C4-S10-SoC0.6 instance.

Policies	Delivery Time (second)			#Infeasible	Energy	Energy	Travel	Travel
	Overall	Travel	Charging	Episodes	Consumption (Wh)	Efficiency ($^{\text{Wh}}/_{\text{km}}$)	Distance (km)	Speed ($^{\text{m}}/_{\text{s}}$)
DRL-CPD	**10075.9**	9475.9	**600.0**	**0**	**18861.3**	**104.9**	**179.9**	19.0
BPH	10472.5	**9272.5**	1200.0	**0**	20510.0	111.5	183.9	19.8
SND	11896.5	10096.5	1800.0	**0**	26007.1	121.5	214.1	21.2

TABLE V: The performance of policies for C5-S14-SoC0.8 instance.

Policies	Delivery Time (second)			#Infeasible	Energy	Energy	Travel	Travel
	Overall	Travel	Charging	Episodes	Consumption (Wh)	Efficiency ($^{\text{Wh}}/_{\text{km}}$)	Distance (km)	Speed ($^{\text{m}}/_{\text{s}}$)
DRL-CPD	**19406.6**	17246.6	**2160.0**	**0**	**43416.2**	121.2	358.2	20.8
BPH	22512.6	19856.3	2656.3	2	43808.9	**111.4**	393.4	19.8
SND	19502.1	**17126.1**	2376.0	**0**	43709.7	122.2	**357.6**	20.9

policies, respectively. For a similar route planning outcome with a travel distance of approximately 116 km, the DRL-CPD policy achieves an average travel time of 5671.2 seconds which is slightly longer than SND's 5435.5 seconds. However, by introducing a flexible and sensible speed planning (i.e., slowing down the BEV proactively), the DRL-CPD policy reduces the total energy consumption from (SND's) 14323.5 to 13240.2 Wh, thereby lowering the charging time of BEV from 900 to 600 seconds. That is, the decrease in charging time outweighs the increase in travel time and ultimately contributes to less overall delivery time, which validates our speculation in Sect. IV-B2. The BPH policy, on the other hand, exhibits a rather "crude" decision-making process. As a result, the speed and charging decisions are relatively conservative (i.e., with the longest charging time of 918 seconds and lowest travel speed of 20.2 $^{\text{m}}/_{\text{s}}$), in order to ensure the completion of delivery as far as possible. The BPH policies are hence easily outperformed by DRL-CPD policies.

While for the S4-C10-SoC0.6 instance, the DRL-CPD policy takes a different strategy, which is to slow down the BEV as much as possible in exchange for the shortest possible en-route charging time. Under the DRL-CPD policy, the average speed of BEV is 19 $^{\text{m}}/_{\text{s}}$, which is even lower than the BPH's 19.8 $^{\text{m}}/_{\text{s}}$. The lowest speed delivers the best energy efficiency of 104.9 $^{\text{Wh}}/_{\text{km}}$, which, along with the shortest travel distance (i.e., 179.9 km), results in the least overall energy consumption (i.e., 18861.3 $^{\text{Wh}}/_{\text{km}}$) and the shortest charging time of 600 seconds for the DRL-CPD policy. In comparison, the BPH and SND policies cost the BEV a much longer charging time (of 1200 and 1800 seconds), and hence the overall delivery time. The observation reaffirms our speculation that with a proper planning particularly over the BEV speed, the time cost of en-route charging can indeed be reduced to offset the potential increase in travel time due to slower speed, thereby improving the overall delivery efficiency.

Fig. 3 visualizes the DRL-CPD, BPH and SND policies for the respective instances in Tabs. III, IV, and V. As captioned, the shape of the markers indicates the type of nodes. The line style of edges indicates the average speed that BEV needs to sustain (according to the underlying policy). The charging times and locations are given by the legend inside each sub-figure (e.g., in the form of "location: time"). Fig. 3 shows that the three policies can make sensible but distinct decisions with regard to the delivery route, speed, and charging of BEV. To demonstrate the impact of BEVs' SVR, let's take a closer look at the policies for the C4-S10-SoC0.6 instance (i.e., the sub-figures in the middle column). Under the SND policy, the BEV visits the following nodes in turn as its second and third round trips: $0 \rightarrow 10 \rightarrow 7 \rightarrow 11 \rightarrow 6 \rightarrow 1 \rightarrow 0 \rightarrow 5 \rightarrow 8 \rightarrow 2 \rightarrow 0$. It is apparent that after serving node 6, the BEV could have visited node 0 for replenishment and later stopped by node 2 for charging, thereby avoiding a detour to

charging station 1. However, the BEV has a SoC of only 27% when departing node 6, which is risky for it to visit node 0 and 2 consecutively at 30[5] m/s. A detour must hence be made to node 1. Such charging detours, together with the prolonged charging time incurred by high-speed travel, jointly undermine the overall performance of SND. In comparison, the DRL-CPD policy explicitly addresses the SVR of BEV, and implements a flexible speed planning, rendering the BEV to always have sufficient SoC to skip unnecessary detours (see the DRL-CPD policy in the first sub-figure of the middle column).

4) Policy Adaptability to Varying Traffic Conditions: One of the major advantages of DRL-based policies over conventional ones is the ability to generalize over unseen states and make informed decisions. In the context of D-EVRP-SVR-SP, the environmental uncertainty comes from the varying traffic conditions. To demonstrate the "responsiveness" of DRL-CPD policy, in Tab. VI we present its performance for one selected C5-S14-SoC0.8 instance under different traffic conditions, or more specifically, speed limit samples[6]. The results show that the episodic performance of the DRL-CPD policy may deviate significantly from the average performance (over 50 episodes) depending on speed limits. As one would intuitively expect, the DRL-CPD policy minimizes the overall delivery time when the traffic condition is "smooth" and the BEV is allowed to travel at any preferable speed (i.e., achieving a delivery time of 17426.2 seconds given $\bar{v}_{ij} = 30$ and $\underline{v}_{ij} = 5$ m/s). It is worth noting that even for this "best-case scenario", the actual BEV speed is 27.2 m/s, lower than the maximum possible speed of 30 m/s. Such an observation again demonstrates that sustaining a high speed in a BEV does not necessarily shorten the overall delivery time. When effective speed limits are introduced, the BEV is steered to travel as fast as possible (e.g., the BEV sustains a travel speed of $\bar{v}_{ij} = 20$ m/s under the "moderate" traffic condition), because the degradation of traffic condition has already resulted in considerably longer travel time, which outweighs the reduction in en-route charging time due to low-speed driving.

In Tab. VII, we present the decision outcomes of the DRL-CPD policy for the same set of episodes of the C5-S14-SoC0.8 instance in Tab. VI. The results show that the DRL-CPD policy can adapt well to the changes in traffic conditions. When the traffic is "smooth", the BEV needs to be charged 5 times at 4 different charging stations (including node $4, 3, 2$ and 5) to cope with the high overall energy consumption (of 56371.2 kW) due to high-speed travel (i.e., at 27.2 m/s). As the road network becomes congested and the BEV speed is limited, the DRL-CPD policy adapts accordingly in terms of the planning of BEV route and charging. For instance, the DRL-CPD policy mainly adjusts the first and third round trip[7] when the traffic condition changes from "smooth" to "heavy": for the first trip, the visit to charging station 4 is cancelled; while for the third trip, the two visits to charging station 5 is replaced with one visit to station 2, and part of the node visiting sequence is reversed (i.e., from $13 \to 18 \to 9$ to $9 \to 18 \to 13$).

[5]The actual speed is subject to the varying traffic condition

[6]Note that for each episodic performance, the same set of speed limits is set for all the roads to make up a fair comparison.

[7]We consider departing & returning to the depot 0 as one round trip.

C. The Real-world Case Study

To examine the practical significance of proposed approach under a real-world setting, we compare the DRL-CPD, BPH, and SND policies for D-EVRP-SVR-SP instance JD-C7-S13-SoC1.0 in Tab. VIII. For 50 episodes of rollouts, the DRL-CPD policy averages an overall delivery time of 19259.1 seconds, which outperforms BPH's 21325.3 seconds and SND's 19978.4 seconds. Similar to the policies for the C3-S6-SoC0.4 and C5-S14-SoC0.8 instance (see Sect. IV-B3), the DRL-CPD policy for JD-C7-S13-SoC1.0 instance steers the BEV to travel at a speed of 20.8 m/s, higher than BPH's 18.6 m/s but lower than SND's 21.1 m/s.

In terms of speed choices, BPH is apparently overly "conservative" which, although facilitating the least overall energy consumption (39438.9 kW) and therefore the shortest en-route charging time (1768.1 seconds), also led to the longest travel time (19557.2 seconds) among the three. As a result, BPH is again outperformed by DRL-based policies significantly.

While the SND policy pays a price for attaining the highest travel speed, with the longest en-route charging time (of 2268 seconds) and a travel time (of 17710.4 seconds) slightly longer than that of DRL-CPD policy, possibly due to the extra detours to the charging stations. Although the DRL-CPD policy only reduces the delivery time by 3.6% compared to SND policy, the real-world impact can be significant, particularly for large-scale, time-sensitive, marginally profitable applications such as trucking in the U.S. In 2022, trucks transported 11.46 billion tons of freight, accounting for 80.7% of the nation's freight expenditure [54]. The 14.33 million registered trucks traveled a combined 331.27 billion miles with an average speed of 55 miles per hour [55], [56]. The travel time thus totals around 6.02 billion hours. Due to BEV's range limitation, electrifying the truck fleet would necessitate charging en-route especially for interstate corridors [57]. A 3% of improvement translates to a reduction of about 180.6 million hours of delivery time and potential cost saving of 4.5 to 36.1 billion USD (value of travel time ranges from 25 to 200 USD per hour [57]).

In Fig. 4, the three policies for JD-C7-S13-SoC1.0 instance are visualized for a randomly picked episode. The illustration shows that all the policies can make sensible decisions and the underlying characteristics align with the policy performance in Tab. VIII. In Fig. 5, the DRL-CPD policy is visualized using Google Maps [58]. Fig. 5 shows that the DRL-CPD policy can indeed make well-rounded and reasonable decisions in a practical sense, and thus has the potential to be deployed for real-world applications.

The comparison using JD case again highlights the importance of coping with BEV's SVR when deploying BEV for time-sensitive applications such as logistics delivery, and the advantages of proposed DRL-CPD policies under a real-world setting.

Remark. *The DRL-CPD policy considers four charging stops because: 1) it learns to limit charging time per stop to avoid the inefficient CV phase to improve delivery efficiency; 2) there is currently no explicit limit on the number of charging stops during training – node 7's convenient location near nodes 15, 16, and 17 justifies the decisions. In the future works, penalties*

TABLE VI: The adaptability of DRL-CPD policy for C5-S14-SoC0.8 instance to varying traffic conditions.

		Delivery Time (second)			#Infeasible	Energy	Energy	Travel	Travel
		Overall	Travel	Charging	Episodes	Consumption (Wh)	Efficiency (Wh/km)	Distance (km)	Speed (m/s)
Average Performance over 50 Episodes		21835.6	19339.6	2496.0	0	45352.4	116.4	389.6	20.1
Episodic Performance Given Speed Limits (m/s)	$\bar{v}_{ij} = 30$ $\underline{v}_{ij} = 5$ (Smooth)	17426.2	14426.2	3000.0	N/A	56371.2	143.5	392.8	27.2
	$\bar{v}_{ij} = 10$ $\underline{v}_{ij} = 5$ (Heavy)	40211.9	38711.9	1500.0	N/A	26435.4	68.3	387.1	10.0
	$\bar{v}_{ij} = 20$ $\underline{v}_{ij} = 10$ (Moderate)	21513.2	19413.2	2100.0	N/A	39777.9	102.5	388.3	20.0

TABLE VII: The decision outcomes of DRL-CPD policy for C5-S14-SoC0.8 instance under different traffic conditions.

Traffic Sample in Speed Limits (m/s)		Decisions
$\bar{v}_{ij} = 30$ $\underline{v}_{ij} = 5$ (Smooth)	Route	$0 \to 8 \to 12 \to 17 \to 14 \to 15 \to \mathbf{4} \to 0 \to \mathbf{3} \to 7 \to 16 \to 11 \to \mathbf{2} \to 0 \to 6 \to \mathbf{5} \to 13 \to 18 \to 9 \to \mathbf{5} \to 0 \to 10 \to 19 \to 0$
	(Desired) Speeds (m/s)	$25 \to 25 \to 25 \to 25 \to 30 \to 30 \to 30 \to 25 \to 25 \to 25 \to 25 \to 25 \to 30 \to 30 \to 25 \to 25 \to 25 \to 25 \to 25 \to 30 \to 30 \to 25 \to 30$
	Charging Times (s)	$\mathbf{4}: 600 \to \mathbf{3}: 600 \to \mathbf{2}: 600 \to \mathbf{5}: 600 \to \mathbf{5}: 600$
$\bar{v}_{ij} = 10$ $\underline{v}_{ij} = 5$ (Heavy)	Route	$0 \to 8 \to 12 \to 17 \to 14 \to 15 \to 0 \to \mathbf{3} \to 7 \to 16 \to 11 \to \mathbf{2} \to 0 \to 6 \to 9 \to 18 \to 13 \to \mathbf{2} \to 0 \to 10 \to 19 \to 0$
	(Desired) Speeds (m/s)	$25 \to 25 \to 25 \to 25 \to 30 \to 30 \to 25 \to 25 \to 25 \to 25 \to 25 \to 30 \to 30 \to 25 \to 25 \to 25 \to 25 \to 30 \to 30 \to 25 \to 30$
	Charging Times (s)	$\mathbf{3}: 600 \to \mathbf{2}: 300 \to \mathbf{2}: 600$
$\bar{v}_{ij} = 20$ $\underline{v}_{ij} = 10$ (Moderate)	Route	$0 \to 8 \to 12 \to 17 \to 14 \to 15 \to \mathbf{4} \to 0 \to \mathbf{3} \to 7 \to 16 \to 11 \to \mathbf{2} \to 0 \to 6 \to 9 \to 18 \to 13 \to \mathbf{2} \to 0 \to 10 \to 19 \to 0$
	(Desired) Speeds (m/s)	$25 \to 25 \to 25 \to 25 \to 30 \to 30 \to 30 \to 25 \to 25 \to 25 \to 25 \to 25 \to 30 \to 30 \to 25 \to 25 \to 25 \to 25 \to 30 \to 30 \to 25 \to 30$
	Charging Times (s)	$\mathbf{4}: 300 \to \mathbf{3}: 600 \to \mathbf{2}: 600 \to \mathbf{2}: 600$

TABLE VIII: Comparative performance analysis of policies for JD-C7-S13-SoC1.0 instance over 50 episodes.

Policies	Delivery Time (second)			#Infeasible	Energy	Energy	Travel	Travel
	Overall	Travel	Charging	Episodes	Consumption (Wh)	Efficiency (Wh/km)	Distance (km)	Speed (m/s)
DRL-CPD	**19259.1**	**17369.1**	1890.0	**0**	43390.5	119.8	**362.1**	20.8
BPH	21325.3	19557.2	**1768.1**	3	**39438.9**	**108.2**	364.6	18.6
SND	19978.4	17710.4	2268.0	**0**	45626.4	122.1	373.8	21.1

for repeated visits to charging nodes or charging delays will be considered to improve policy feasibility.

V. Conclusion

This paper proposed a new EVRP model, D-EVRP-SVR-SP, which explicitly models the BEVs' SVR and jointly plans the route, speed, and charging for the BEV under stochastic traffic conditions. D-EVRP-SVR-SP seeks to minimize the expected total delivery time, characterising the trade-off between BEV traveling and charging given the impact of SVR (i.e., higher speeds means faster traverse but potentially more en-route charging). A DRL approach, namely DRL-CPD, is developed to solve D-EVRP-SVR-SP with real-time policy. The actor of DRL-CPD is specifically designed in a "coupled" structure to exploit the interdependence among the three decisions.

The comparison of DRL-CPD with two DRL-based alternatives shows that the actor's "coupled" structure can effectively accelerate the convergence to high-quality policies, and DRL-CPD as an instance-specific approach outperforms the popular attention-based ones that emphasize generalization capability. Comparing DRL-CPD against two baseline policies shows that always driving BEV at the top speed does not necessarily save the total delivery time, since the greater energy consumed by high-speed travel (given BEVs' SVR) potentially entails more charging activities (which may also require detours) and longer charging times. The DRL-CPD policies learn to intentionally

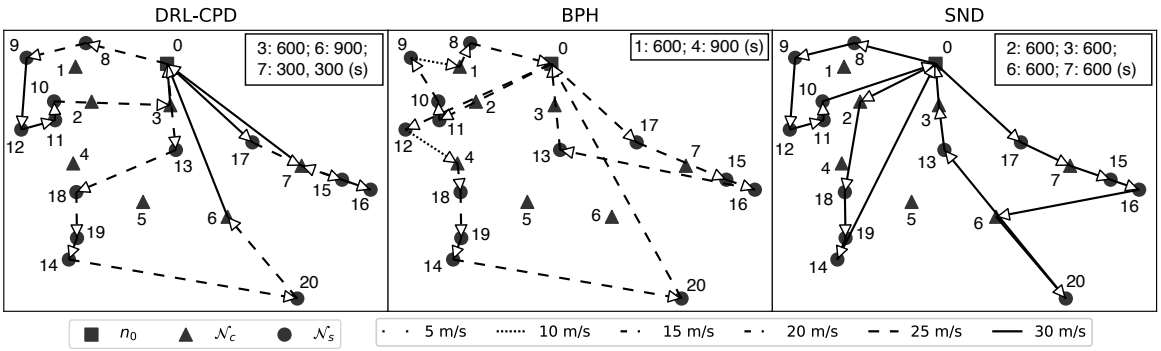


Fig. 4: The illustration of DRL-CPD, BPH, and SND policies for instance JD-C7-S13-SoC1.0.

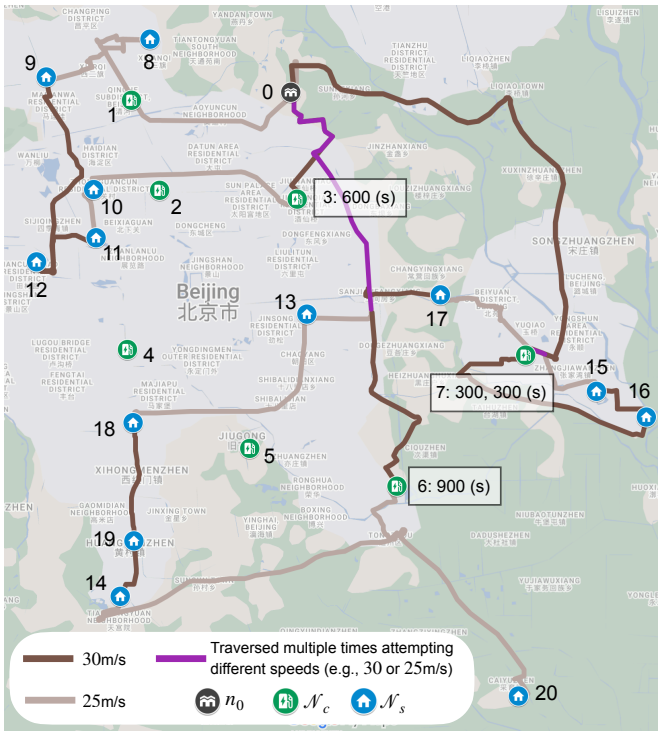


Fig. 5: The DRL-CPD policy for JD logistics instance on real-world map.

slow down the BEV at appropriate times, so that the time saved from en-route charging outweighs the added cruise time due to lower speeds. In practice, the proposed approach is applicable to any operational scenarios where BEV's limited range necessitates en-route charging and addressing the SVR of BEV is beneficial, e.g., interstate trucking in the U.S., self-driving BEV for 7/24 delivery.

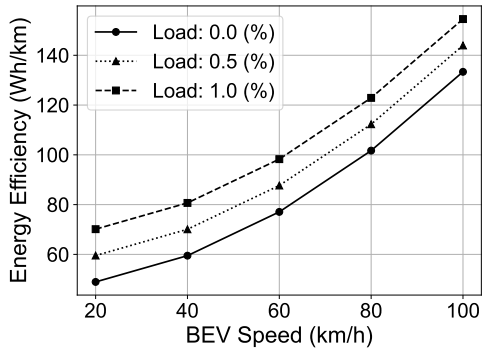
In the future works, the D-EVRP-SVR-SP model could be further extended to examine the impacts of (BEVs') SVR on BEV adoption in delivery tasks with a greater level of realism: a more comprehensive energy dynamics model considering the impacts of other exogenous factors (e.g., weather, road gradient), unexpected delays (e.g., queueing for en-route charging), multiple vehicles with heterogeneous characteristics, and time window constraints. Additional modeling features come with increased model complexities, demanding more sophisticated solution methods, e.g., multi-agent DRL to model the cooperation among BEVs, advanced DRL agent to introduce stronger generalization capability. Other extensions of the model may involve the economic or environmental objectives that exploit more benefits of speed planning, e.g., reducing the electricity consumption and associated emissions, and mitigating battery degradation through indirect control of discharging rate.

Appendix A
Specifications of JD-C7-S13-SoC1.0 Instance

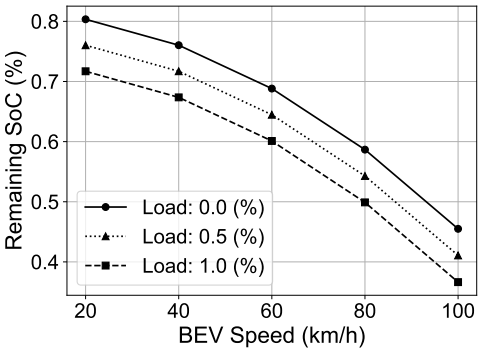
TABLE A.1: The geographic locations of nodes in JD-C7-S13-SoC1.0 instance.

ID	Location	Latitude	Longitude
0	JD Logistics Distribution Centre	40.04105	116.47808
1	Yanqingyuan Charging Station	40.03614	116.34183
2	Cuiwei Shopping Mall Charging Station	39.97718	116.36580
3	Scenery Line Charging Station	39.97122	116.48299
4	The Blues International Charging Station	39.87304	116.33841
5	Qingxinyuan Charging Station	39.80817	116.44239
6	Star Charging Area	39.78313	116.56819
7	Ruidu Park Charging Station	39.86908	116.67835
8	JD Service Point (Longyue Road)	40.07598	116.35731
9	JD Service Point (Dezheng Road)	40.05119	116.26901
10	JD Logistics Haidian Station	39.97759	116.30985
11	JD Service Point (Wanshousi Road)	39.94593	116.31145
12	JD Service Point (Banbidian 2nd Street)	39.93013	116.26104
13	JD Service Point (Guangqu Road)	39.89580	116.49121
14	JD Service Point (Linxiao Road)	39.71100	116.33190
15	JD Logistics Huanhu Distribution Station	39.84562	116.73859
16	JD Service Point (Jingtang Road)	39.82878	116.78109
17	JD Freight Service	39.90856	116.60518
18	JD Service Point (Youan Road)	39.82473	116.34301
19	JD Logistics Xinghua Distribution Station	39.74721	116.34402
20	JD Logistics Gu'an Large-Sized Object Operation Center	39.64524	116.67197

The geographic locations of the nodes in JD-C7-S13-SoC1.0 instance are listed in Tab. A.1. For each origin-destination pair, one can request the distance in between from Google Distance Matrix API under the transportation mode of "driving" and



(a)



(b)

Fig. B.1: The energy dynamics of E-NV200 for traveling 100 km at different average speed and under different cargo load (i.e., $m_c/\bar{m}_c$ in %).

build up a 20×20 adjacency matrix. Due to page limit, it is excluded here but we're happy to share upon request.

Appendix B
Validation of The Energy Dynamics Model for Nissan E-NV200

To validate our energy dynamics model, we simulate traveling the E-NV200 for 100 km under different operating conditions, and used our model to estimate the energy efficiency and remaining SoC of the vehicle. The results are shown in Fig. B.1. One can observe that both the energy efficiency and remaining SoC degrades non-linearly as the increase in either vehicle speed or cargo load, which is consistent with the results in the existing studies [11]. It is hence well-founded to capture and exploit the SVR of BEV based on our model.

Appendix C
Implementation Details of DRL-ATTN

Under the framework in [51], the node network of DRL-ATTN produces the visiting probability of nodes by feeding the embedded state (where each feature is mapped into a 32-dimensional vector space) and the travel history restored using a recurrent neural network (RNN) decoder (with a hidden state size of 128) into an attention mechanism. In addition to the cargo weight of the BEV and the remaining demands of customers, the BEV SoC is also considered a "dynamic element" to be processed by the attention layer. The speed and charging networks of DRL-ATTN are residual networks with a structure of $[128, 64, 32]$ (i.e., number of hidden units in each layer) and ReLU activations. The context vectors computed by the attention-based node network and the preceding decision(s) are taken as inputs to generate the probability distributions of the associated actions in a "coupled" manner as described in Sect. III-B1). It should be noted that the above setting delivers the best results of all tested ones. In future works, an attention-based model dedicated for D-EVRP-SVR-SP is expected.

Appendix D
Implementation Details of BPH Policy

For any given D-EVRP-SVR-SP instance, BPH policy can be set up via the following two steps:

- Formulate a CVRP instance by detaching all the components of the D-EVRP-SVR-SP instance in regard to electrification, e.g., charging stations. Then solve this instance using branch-and-price for optimal routes [53].
- Perform a grid search to determine the values of the following hyper-parameters:
  - (i) $\text{SoC}_{\text{min}} \in \{10, 20, 30\}$ (%): whenever the SoC falls below SoC_{min}, the BEV shall visit the nearest charging station.
  - (ii) $\text{SoC}_{\text{max}} \in \{60, 70, 80\}$ (%) is the maximum SoC the BEV can charge up to, i.e., $\text{SoC}_c \leq \text{SoC}_{\text{max}}$. Like DRL-based methods, BPH also picks discrete charging times from set ψ.
  - (iii) $v_d \in \phi$ (m/s) is the speed at which the BEV traverses the customers and depot.
  - (iv) $v_c \in \phi$ (m/s) is the BEV's speed when moving towards charging stations.

  For each value combination, 100 episodes of the original D-EVRP-SVR-SP instance are rolled out, in which the BEV follows the CVRP route (found in the last step) to perform the delivery, and chooses the timing and duration for en-route charging, as well as the driving speed, based on the given values. The combination with the best average performance (i.e., minimizing the total delivery time) across all 100 episodes will be selected.

References

[1] International Energy Agency, "Global EV Outlook 2022." [Online]. Available: https://www.iea.org/reports/global-ev-outlook-2022
[2] ——, "Global EV Outlook 2023." [Online]. Available: https://www.iea.org/reports/global-ev-outlook-2023
[3] AXIOS, "Amazon Reaches 10,000 Rivian Electric Delivery Vans in U.S." [Online]. Available: https://www.axios.com/2023/10/17/amazon-rivian-electrification-10000-climate
[4] Auto Futures, "Amazon Rolls Out First Electric Delivery Vans From Rivian in Europe." [Online]. Available: https://www.autofutures.tv/topics/amazon-rolls-out-first-electric-delivery-vans-from-rivian-in-europe/s/44d0358f-7788-4c1a-8c93-555a1ff9a389
[5] Car and Driver, "USPS Buying Six of EV Startup Canoo's Pod-Like Delivery Vans." [Online]. Available: https://www.caranddriver.com/news/a46522207/canoo-electric-delivery-vans-usps/
[6] Sustainable Bus, "Go-Ahead London to add nearly 300 electric buses from BYD–Alexander Dennis in 2023." [Online]. Available: https://www.sustainable-bus.com/electric-bus/go-ahead-london-byd-alexander-dennis-300-2023/

[7] Transport for NSW, "Zero Emission Buses." [Online]. Available: https://www.transport.nsw.gov.au/projects/current-projects/zero-emission-buses

[8] U.S. Department of Energy, "Alternative Fuels Data Center." [Online]. Available: https://afdc.energy.gov/vehicles/search/results?vehicle_type=heavy&category_id=5&fuel_id=41

[9] Tesla, "Tesla Model 3." [Online]. Available: https://www.tesla.com/en_hk/model3

[10] The National Renewable Energy Laboratory, "Commercial Fleet Vehicle Operating Data." [Online]. Available: https://www.nrel.gov/transportation/fleettest-fleet-dna.html

[11] G. Wager, J. Whale, and T. Braunl, "Driving electric vehicles at highway speeds: The effect of higher driving speeds on energy consumption and driving range for electric vehicles in Australia," *Renew. Sustain. Energy Rev.*, vol. 63, pp. 158–165, 2016.

[12] Wikipedia, "Electric Vehicle Battery." [Online]. Available: https://en.wikipedia.org/wiki/Electric_vehicle_battery

[13] L. Kong, H. Zhang, W. Li, H. Bai, and N. Dai, "Spatial-temporal scheduling of electric bus fleet in power-transportation coupled network," *IEEE Trans. Transp. Electrif.*, vol. 9, no. 2, pp. 2969–2982, 2022.

[14] H. Fan, D. Wang, Z. Yu, and L. Du, "Bi-level optimal scheduling of electric bus fleets in regional integrated electricity-gas-heat energy systems," *IEEE Trans. Transp. Electrif.*, vol. 9, no. 2, pp. 2792–2807, 2022.

[15] A. Ayad, N. A. El-Taweel, and H. E. Farag, "Optimal design of battery swapping-based electrified public bus transit systems," *IEEE Trans. Transp. Electrif.*, vol. 7, no. 4, pp. 2390–2401, 2021.

[16] H. Wang, W. Li, Z. Zhao, Z. Wang, M. Li, and D. Li, "Intelligent distribution of fresh agricultural products in smart city," *IEEE Trans. Ind. Inf.*, vol. 18, no. 2, pp. 1220–1230, 2021.

[17] G. Ferro, M. Paolucci, and M. Robba, "Optimal charging and routing of electric vehicles with power constraints and time-of-use energy prices," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 14 436–14 447, 2020.

[18] R. Raeesi and K. G. Zografos, "The electric vehicle routing problem with time windows and synchronised mobile battery swapping," *Transp. Res. Part B Methodol.*, vol. 140, pp. 101–129, 2020.

[19] A. Abdulaal, M. H. Cintuglu, S. Asfour, and O. A. Mohammed, "Solving the multivariant EV routing problem incorporating V2G and G2V options," *IEEE Trans. Transp. Electrif.*, vol. 3, no. 1, pp. 238–248, 2017.

[20] X. Bi and W. K. S. Tang, "Logistical planning for electric vehicles under time-dependent stochastic traffic," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3771–3781, 2019.

[21] F. Y. Vincent, P. Jodiawan, and A. Gunawan, "An adaptive large neighborhood search for the green mixed fleet vehicle routing problem with realistic energy consumption and partial recharges," *Appl. Soft Comput.*, vol. 105, p. 107251, 2021.

[22] M. S. Shojaei, F. Fakhrmoosavi, A. Zockaie, M. Ghamami, A. Mittal, and J. Fishelson, "Sustainable transportation networks incorporating green modes for urban freight delivery," *J. Transp. Eng. Part A. Syst.*, vol. 148, no. 6, p. 04022028, 2022.

[23] M. E. H. Sadati and B. Çatay, "A hybrid variable neighborhood search approach for the multi-depot green vehicle routing problem," *Transp. Res. Part E Logist. Transp. Rev.*, vol. 149, p. 102293, 2021.

[24] B. Lin, B. Ghaddar, and J. Nathwani, "Deep reinforcement learning for the electric vehicle routing problem with time windows," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 11 528–11 538, 2021.

[25] X. Bi, R. Wang, and Q. Jia, "On the speed-varying range of electric vehicles in time-windowed routing problems with en-route partial recharging," *IEEE Trans. Consum. Electron.*, vol. 70, no. 1, pp. 3650–3657, 2023.

[26] Z. Liu, X. Zuo, M. Zhou, W. Guan, and Y. Al-Turki, "Electric vehicle routing problem with variable vehicle speed and soft time windows for perishable product delivery," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 6, pp. 6178–6190, 2023.

[27] Y.-H. Jia, Y. Mei, and M. Zhang, "Confidence-based ant colony optimization for capacitated electric vehicle routing problem with comparison of different encoding schemes," *IEEE Trans. Evol. Comput.*, vol. 26, no. 6, pp. 1394–1408, 2022.

[28] Y. Chen, J. Xue, Y. Zhou, and Q. Wu, "An efficient threshold acceptance-based multi-layer search algorithm for capacitated electric vehicle routing problem," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 6, pp. 5867–5879, 2024.

[29] C. Wang, F. Qin, X. Xiang, H. Jiang, and X. Zhang, "A dual-population-based co-evolutionary algorithm for capacitated electric vehicle routing problems," *IEEE Trans. Transp. Electrif.*, vol. 10, no. 2, pp. 2663–2676, 2024.

[30] Y. Zhang, X. Qu, and L. Tong, "Optimal eco-driving control of autonomous and electric trucks in adaptation to highway topography: Energy minimization and battery life extension," *IEEE Trans. Transp. Electrif.*, vol. 8, no. 2, pp. 2149–2163, 2022.

[31] R. Basso, B. Kulcsár, and I. Sanchez-Diaz, "Electric vehicle routing problem with machine learning for energy prediction," *Transp. Res. Part B Methodol.*, vol. 145, pp. 24–55, 2021.

[32] Y. Xiao, X. Zuo, I. Kaku, S. Zhou, and X. Pan, "Development of energy consumption optimization model for the electric vehicle routing problem with time windows," *J. Cleaner Prod.*, vol. 225, pp. 647–663, 2019.

[33] S. Pelletier, O. Jabali, and G. Laporte, "The electric vehicle routing problem with energy consumption uncertainty," *Transp. Res. Part B Methodol.*, vol. 126, pp. 225–255, 2019.

[34] Z. Yi and P. H. Bauer, "Effects of environmental factors on electric vehicle energy consumption: A sensitivity analysis," *IET Electr. Syst. Transp.*, vol. 7, no. 1, pp. 3–13, 2017.

[35] Y. Hua, M. Sevegnani, D. Yi, A. Birnie, and S. Mcaslan, "Fine-grained RNN with transfer learning for energy consumption estimation on EVs," *IEEE Trans. Ind. Inf.*, vol. 18, no. 11, pp. 8182–8190, 2022.

[36] H. Lu, C. Shao, B. Hu, K. Xie, C. Li, and Y. Sun, "En-route electric vehicles charging navigation considering the traffic-flow-dependent energy consumption," *IEEE Trans. Ind. Inf.*, vol. 18, no. 11, pp. 8160–8171, 2021.

[37] ArenaEV, "Electric Vehicle Range at Highway Speed." [Online]. Available: https://www.arenaev.com/57_electric_cars_range_tested_at_highway_speeds__who_wins-news-1904.php

[38] Wikipedia, "Worldwide Harmonised Light Vehicles Test Procedure." [Online]. Available: https://en.wikipedia.org/wiki/Worldwide_Harmonised_Light_Vehicles_Test_Procedure

[39] X. Tang, X. Lin, and F. He, "Robust scheduling strategies of electric buses under stochastic traffic conditions," *Transp. Res. Part C Emerging Technol.*, vol. 105, pp. 163–182, 2019.

[40] K. Jin, X. Li, W. Wang, X. Hua, and W. Long, "Energy-optimal speed control for connected electric buses considering passenger load," *J. Cleaner Prod.*, vol. 385, p. 135773, 2023.

[41] V. Pillac, M. Gendreau, C. Guéret, and A. L. Medaglia, "A review of dynamic vehicle routing problems," *Eur. J. Oper. Res.*, vol. 225, no. 1, pp. 1–11, 2013.

[42] S. Pelletier, O. Jabali, G. Laporte, and M. Veneroni, "Battery degradation and behaviour for electric vehicles: Review and numerical analyses of several models," *Transp. Res. Part B Methodol.*, vol. 103, pp. 158–187, 2017.

[43] Y. Zhang, Q. H. Vuong, K. Song, X.-Y. Gong, and K. W. Ross, "Efficient entropy for policy gradient with multidimensional action space," *arXiv preprint arXiv:1806.00589*, 2018.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[45] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT press, 2018.

[46] "JD Logistics." [Online]. Available: https://corporate.jd.com/ourBusiness#jdLogistics

[47] "Google Distance Matrix API." [Online]. Available: https://developers.google.com/maps/documentation/distance-matrix/overview

[48] Wikipedia, "Beijing's 5th Ring Road." [Online]. Available: https://en.wikipedia.org/wiki/5th_Ring_Road

[49] Automotive Energy Supply Corporation, "Characteristics of E-NV200's Battery Cell." [Online]. Available: https://www.qnovo.com/blogs/inside-the-battery-of-a-nissan-leaf

[50] Nissan, "E-NV200 E-Brochure." [Online]. Available: https://www.nissan-cdn.net/content/dam/Nissan/ireland/Brochures/e-NV200%20E-Brochure.pdf

[51] M. Nazari, A. Oroojlooy, L. Snyder, and M. Takác, "Reinforcement learning for solving the vehicle routing problem," *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

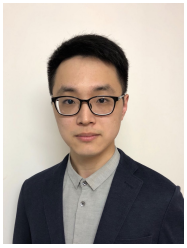
[52] J. Li, Y. Ma, R. Gao, Z. Cao, A. Lim, W. Song, and J. Zhang, "Deep reinforcement learning for solving the heterogeneous capacitated vehicle routing problem," *IEEE Trans. Cybern.*, vol. 52, no. 12, pp. 13 572–13 585, 2022.

[53] D. Feillet, "A tutorial on column generation and branch-and-price for vehicle routing problems," *4OR*, vol. 8, no. 4, pp. 407–424, 2010.

[54] American Trucking Associations, "Truck Freight Tonnage and Revenues Rise in 2022." [Online]. Available: https://www.trucking.org/news-insights/truck-freight-tonnage-and-revenues-rise-2022-according-report

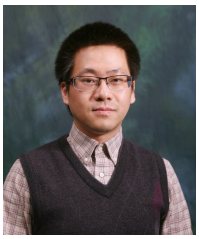
[55] ——, "Economics and Industry Data." [Online]. Available: https://www.trucking.org/economics-and-industry-data

[56] Federal Highway Administration, "Average Truck Speeds on Selected Interstate Highways." [Online]. Available: https://ops.fhwa.dot.gov/freight/freight_analysis/nat_freight_stats/docs/10factsfigures/table3_8.htm
[57] ——, "Measuring Travel Time in Freight-Significant Corridors." [Online]. Available: https://ops.fhwa.dot.gov/freight/documents/travel_time_flyer.pdf
[58] "Google My Maps." [Online]. Available: https://www.google.com/intl/nl/maps/about/mymaps/



Xiaowen Bi received his B.S. degree from Shandong University in 2014, then completed his M.S. and Ph.D. at City University of Hong Kong in 2015 and 2020. He was a postdoctoral fellow at The Hong Kong Polytechnic University and City University of Hong Kong. He is currently an Assistant Professor in Guangdong Provincial/Zhuhai Key Laboratory of IRADS and Department of Statistics and Data Science at BNU-HKBU United International College. His research focuses on computational intelligence, complex networks, and their applications in advancing intelligent transportation systems.

Minyu Shen received his Ph.D. (2020) in electrical engineering from The Hong Kong Polytechnic University. He is a lecturer at the School of Management Science and Engineering, Southwestern University of Finance and Economics, China. He has published several papers in Transportation Science, Transportation Research Part B, C. His research interests include public transit, day-to-day traffic dynamics, and applying reinforcement learning in the transportation field.



Weihua Gu is an associate professor at Hong Kong Polytechnic University (PolyU). He received a Ph.D., a M.Sc., and a M.A. from University of California, Berkeley, USA. He also received a B.S. and a M.Eng. from Tsinghua University, China. His research interests include public transit systems, multimodal urban transport, and queueing models.

Edward Chung is a Professor of Intelligent Transport Systems (ITS) at the Department of Electrical Electronic Engineering of The Hong Kong Polytechnic University (PolyU). With a background as both an engineer and an academic researcher, Professor Chung has worked on numerous national and international projects. He has held various positions, including Senior Research Scientist at the Australian Road Research Board, Manager of Infrastructure Analysis and Modelling at the Victorian Department of Infrastructure, Australia, Visiting Professor at the Centre for Collaborative Research, University of Tokyo, and Head of the ITS Group at LAVOC, EPFL, Switzerland. Before joining PolyU, Professor Chung was a professor at the Queensland University of Technology (QUT) and held the position of Director of the Smart Transport Research Centre. He received his Bachelor's degree and PhD from Monash University.

Yuhong Wang received his M.Eng. and B.Eng. degree at Tongji University, Shanghai, China in 1996, and the Ph.D. and MSc in civil engineering at the University of Kentucky, the USA in 2003 and 2001, respectively. He is currently a Professor with the Department of Civil and Environmental Engineering, Hong Kong Polytechnic University, Hong Kong. His research interest focuses on the new generation of urban infrastructure, which includes how to make future cities cleaner and more environmentally friendly, more resistant to floods, better serve urban residents, smarter, and how to promote biodiversity in the urban environment.