



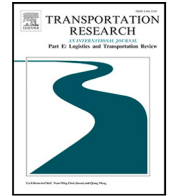
A multi-objective reinforcement learning-based velocity optimization approach for electric trucks considering battery degradation mitigation

Downloaded from: <https://research.chalmers.se>, 2024-12-20 13:29 UTC

Citation for the original published paper (version of record):

Jia, R., Gao, K., Cui, S. et al (2025). A multi-objective reinforcement learning-based velocity optimization approach for electric trucks considering battery degradation mitigation. *Transportation Research Part E: Logistics and Transportation Review*, 194. <http://dx.doi.org/10.1016/j.tre.2024.103885>

N.B. When citing this work, cite the original published paper.



A multi-objective reinforcement learning-based velocity optimization approach for electric trucks considering battery degradation mitigation

Ruo Jia ^a, Kun Gao ^{a,*}, Shaohua Cui ^a, Jing Chen ^b, Jelena Andric ^a

^a Department of Architecture and Civil Engineering, Chalmers University of Technology, Goteburg SE-412 96, Sweden

^b Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, Shanghai 201804, China

ARTICLE INFO

Keywords:

Battery-powered electric truck
Sustainable logistics
Velocity optimization
Deep reinforcement learning
Battery degradation mitigation

ABSTRACT

Electrification of commercial vehicles for more sustainable logistic systems has been promoted in the past decades. This study proposes a deep reinforcement learning method for velocity optimization and battery degradation minimization during operation for battery-powered electric trucks (BETs), aiming to achieve a safe, efficient, and comfortable driving control policy for BETs. To obtain an optimal solution considering both calendar and cyclic battery degradation, Deep Deterministic Policy Gradient and Twin Delayed Deep Deterministic Policy Gradient (TD3) approaches are integrated within a simulation environment. To optimize overall BET velocity performance, a trade-off among safety, efficiency, comfort, and battery degradation is incorporated into the reward function of reinforcement learning using Mixture of Experts (MoE) model. The results indicate that the proposed TD3-MoE model achieves safe, efficient, and comfortable car-following control while optimizing total battery degradation. Specifically, the model achieves reductions in total battery capacity loss ranging from 2.4% to 8.3% at different states of charge (SoC) of battery compared to human-driven scenarios. Moreover, despite calendar battery degradation being inevitable, the cyclic battery degradation is effectively mitigated by 27.7% to 29.6% compared to the same SoCs in human-driving data. Furthermore, the TD3-MoE model achieves significant energy consumption reductions, ranging from 35.3% to 39.8% compared to real car-following trajectories.

1. Introduction

Road transportation logistic is essential for global economic activity but also acts as a significant environmental polluter, presenting challenges for achieving a sustainable and low-carbon future (Shoman et al., 2023; Mulholland et al., 2018). Especially, heavy-duty trucks, although constituting a small fraction of road vehicles, have a disproportionately high environmental and economic impact. For instance, in Europe, while heavy-duty trucks represent less than 5% of road vehicles, they accounted for 15%–22% of CO₂ emissions from road transport in 2019, with an additional increase of 9% between 2014 and 2019 (Shoman et al., 2023). Similarly, in the United States, the truck industry transported 70% of all freight tonnage and generated approximately \$700.1 billion in gross freight revenues in 2017 (Osieczko et al., 2021). Despite comprising only 4% of the vehicle population, heavy-duty trucks consume 18% of the energy in the transportation sector (Shoman et al., 2023). With global freight activity expected to more than double by 2050, minimizing energy consumption in truck operations becomes increasingly urgent (Xu et al., 2023).

* Correspondence to: Sven Hultins Gata 6, Chalmers University of Technology, Gothenborg, SE-412 96, Sweden.

E-mail address: gkun@chalmers.se (K. Gao).

<https://doi.org/10.1016/j.tre.2024.103885>

Received 2 August 2024; Received in revised form 22 November 2024; Accepted 24 November 2024

Available online 14 December 2024

1366-5545/© 2024 The Authors.

Published by Elsevier Ltd.

This is an open access article under the CC BY license

(<http://creativecommons.org/licenses/by/4.0/>).

With the advancement of electrification, the deployment of battery-powered electric trucks (BETs) could significantly reduce greenhouse gas emissions from road freight logistic systems (Osieczko et al., 2021). BETs vehicles are capable of achieving zero tailpipe emissions and typically result in reduced maintenance costs (Shoman et al., 2023). However, the adoption of electric commercial vehicle such as trucks on a global scale is impeded by significant limitations including restricted operational range and high battery costs. Additionally, the inevitable capacity loss of batteries over time poses a significant drawback, particularly due to the costly nature of battery replacement (Zhang et al., 2022; Cui et al., 2023). Technically, the capacity of battery tends to decrease over time due to factors such as the dissolution of positive electrode materials and the formation of solid electrolyte interface layers (Zhang et al., 2022). On a broader scale, battery aging is exacerbated by extreme temperatures, inappropriate states of charge (SoC), and high/low temperatures, which are influenced by the driving dynamics of BETs and recognized as critical factors affecting battery longevity (Chung et al., 2020; Schimpe et al., 2018b; Lin et al., 2014; Schimpe et al., 2018a). These elements underscore that optimizing the driving strategy for BETs is crucial for minimizing greenhouse gas emissions, energy consumption and battery degradation, for sustainable implementation of BETs in logistic systems.

To address the operational challenges of BETs, with a dual aim of reducing battery degradation and increasing velocity control efficiency (Han et al., 2018; Li et al., 2024), eco-driving approaches offer velocity control to minimize energy consumption or emissions during velocity optimization. Eco-driving optimal control methods, which deliver direct control actions such as acceleration and steering control, are extensively employed in vehicle motion control, notably through model predictive control (MPC). MPC operates by continuously optimizing control actions over multiple future time horizons (prediction horizons), while only executing the action planned for the immediate next time horizon (control horizon) (Qiu et al., 2017; Han et al., 2018). However, the substantial computational demands of MPC limit its practicality for real-time implementation, particularly for BETs where constraints and complexity exponentially increase. Furthermore, effective implementation of MPC for eco-driving requires precise predictions of leading vehicles (LV) behaviors, which may not always be feasible in BET studies. In instances where predictions are inaccurate, alternative strategies such as car-following logic and desired speed ranges must be employed to ensure safety and efficiency (Yang et al., 2024).

To address computational complexity, deep reinforcement learning (DRL)-based methods have been developed for devising velocity control strategies using extensive experiential data from diverse environments, thereby circumventing the need for predefined rules or modeling of intricate systems (Han et al., 2023; Qu et al., 2023). Specifically, DRL empowers an agent to make decisions based on the state alone, with simulators managing the modeling and state transitions through the reward function (Du et al., 2022). Studies indicate that DRL-based velocity control often surpasses MPC in driving performance (Lin et al., 2020; Zhu et al., 2018). The integration of DRL has attracted considerable attention due to these advantages, with its efficacy and wide applicability demonstrated across various settings (Zhu et al., 2018; Du et al., 2022). Nonetheless, integrating battery considerations into velocity control, particularly the precise estimation of the influence on battery degradation, remains a significant challenge.

The literature reviewed reveals three critical gaps: (1) While the energy consumption of BETs is always considered in velocity controls, the impacts of velocity optimization on battery degradation require further exploration and modeling; (2) Existing optimization algorithms are too complex for real-time velocity control implementation, and addressing the multi-objective optimization of energy consumption, battery degradation, efficiency, comfort, and safety presents an extremely complex problem; (3) In the context of DRL approaches, efficiently considering battery degradation and balancing the weights of different optimization targets is crucial and currently under-explored. This study aims to bridge these gaps by proposing a velocity control methodology based on reinforcement learning (Twin Delayed Deep Deterministic Policy Gradient, TD3) that particularly considers battery degradation for velocity optimization of BETs. The battery degradation model is derived from battery cell-level analysis to model how velocity profiles during driving affect battery degradation, considering the charging and discharging properties. Besides battery degradation, the velocity optimization approach conducts a multi-objective optimization considering efficiency, comfort, and safety. To address the complexity and integration of multiple optimization objectives in a unified framework, a Mixture of Experts approach is embedded into TD3 for optimizing the weights of different objectives in the reward of learning process. The objective is to introduce a novel approach to velocity optimization for BETs that specifically addresses battery degradation during operational phases. This enhancement aims to boost both efficiency and sustainability of BETs within logistic systems.

The remainder of this paper is organized as follows: Section 2 presents a comprehensive literature review. Section 3 introduces the proposed RL-based velocity control model and describes in detail the experimental design. Section 5 discusses the experimental results and relevant findings. Finally, Section 6 provides conclusions and future research directions.

2. Literature review

2.1. Velocity optimization with reinforcement learning

Velocity control is vital in enhancing the performance of vehicles, with numerous studies highlighting its significance (Qu et al., 2020; Shi et al., 2023; Wang et al., 2024a). Velocity control approaches including rule-based and optimization-based techniques, such as classical car following models, dynamic programming, and model predictive control, are traditionally employed to optimize velocity planning and control based on driving trajectories or within a prediction horizon. However, these methods often face limitations in complex systems or changing scenarios, requiring substantial computational resources or struggling to achieve optimal solutions at short periods (Du et al., 2022).

In contrast, learning-based methods formulate velocity optimization strategies from extensive experiential data across diverse environments, eliminating the need for predefined rules or modeling of complex systems. Imitation learning utilizes driving

datasets of expert demonstrations in velocity optimization to train a policy for a represented range of scenarios. [Le Mero et al. \(2022\)](#). However, imitation learning-based policies rely on a great deal of high-quality expert demonstrations, meanwhile, the generalization performance of trained policies is quite limited. To deal with these issues, DRL is proposed and widely-used in learning-based velocity planning and control. DRL enables an agent to make decisions based on modelings in high-fidelity simulation and interactions with changing driving environments ([Du et al., 2022](#)). In this way, DRL can handle velocity optimization problems with complicated vehicle systems in dynamic driving environment. Evidence suggests that DRL-based velocity optimization can outperform model predictive control in terms of various driving performances ([Lin et al., 2020](#); [Zhu et al., 2018](#)). The adoption of DRL in intelligent transport systems has garnered significant attention due to these advantages, and its efficacy and broad applicability have been demonstrated in various contexts, such as safety and ride comfort ([Zhu et al., 2018](#); [Du et al., 2022](#)). Recent studies have explored velocity optimization under varying scenarios and constraints. [Qi et al. \(2019\)](#) introduced a DRL-based distributed velocity control strategy for connected vehicle under communication failures to stabilize traffic oscillations. [Wegener et al. \(2021\)](#) proposed a reinforcement learning approach for energy-saving potential in a signalized urban roads and multiple preceding vehicles environment. Furthermore, [Yang et al. \(2024\)](#) developed eco-driving strategies for mixed traffic scenarios with limited information availability using reinforcement learning.

Although DRL offers significant advantages, it is known for its instability, time-intensive nature, and context-specific requirements. Unresolved challenges in reward function design and parameter optimization hinder its full potential ([Ye et al., 2019](#)). Additionally, in the realm of multi-objective optimization, reinforcement learning typically combines optimization targets using a reward function. Common practice in many DRL studies involves directly summing objectives such as safety, efficiency, energy consumption, and ride comfort ([Du et al., 2022](#); [Yang et al., 2024](#)). This approach often relies on prior knowledge and experience, which lacks systematic and scientific rigor. Meanwhile, human knowledge and experiences are limited and hence cannot cover all possible situations. Even different objectives are treated as equally important in some studies ([Han et al., 2023](#)), the weights of different objectives in a reward function are adjusted to the same magnitude based on massive trials, not absolutely equal. Thus, how to learn weights of different objectives in reward functions reliably and automatically is an urgent problem of DRL-based velocity optimization with multiple objectives.

2.2. Mixture of experts

Most reinforcement learning problems involving velocity optimization contend with multiple objectives, even though many algorithms designed for sequential decision-making focus primarily on optimizing a single objective. A common approach in deep learning is to aggregate all objectives into a unified additive reward function, usually through an iterative process of assigning numerical rewards or penalties based on the objectives ([Lin et al., 2023](#); [Fei et al., 2024](#)). However, this approach suffers from several drawbacks: it is semi-manual and somewhat arbitrary; it limits the decision-maker ability to make informed trade-offs; it reduces the explainability of the decision-making process; and it struggles to adapt when preferences between objectives change ([Hayes et al., 2022](#)). To address these issues, the Mixture of Experts (MoE) model has been proposed, which produces biased experts whose outputs are negatively correlated ([Zhou et al., 2022](#)). The model has recently played a crucial role in enhancing the training efficiency of large-scale language models ([Qu et al., 2023](#); [Yu et al., 2024](#)). MoE framework, originally proposed in the field of machine learning, utilizes specialized, negatively correlated experts to enhance model robustness and diversity ([Jacobs et al., 1991](#)). The MoE model operates by deploying a set of expert sub-networks, each designed to be selectively activated based on the input. This selective activation is governed by a gating network optimized to direct each input token to the most appropriate experts. MoE provides a powerful strategy for scaling model capacity within a fixed computational budget and it has been instrumental in enhancing the training efficiency ([Zhou et al., 2022](#)). This approach not only addresses computational constraints but also improves the model adaptability to varying objectives, establishing a sophisticated tool in complex decision-making systems. In recent years, MoE has been adapted for use in various fields including reinforcement learning applications such as water management and wind farm control ([Menezes et al., 2018](#)). For example, [Jin and Ma \(2019\)](#) employed a constrained Markov decision process to address multi-objective reinforcement learning challenges, aiming to identify Pareto optimal solutions. However, the application of MoE to multi-objective problems in velocity optimization remains limited. Furthermore, the integration of different components using MoE represents a critical aspect of our study, indicating a substantial gap in existing literature and underscoring the need for further exploration in this area.

2.3. Battery degradation

Battery degradation in electric vehicles such as BETs can be classified into two aging degradation processes, cyclic and calendar. Batteries sustain cyclic aging with each charge/discharge cycle, while calendar aging occurs regardless of these cycles. Both types of degradation are influenced by the state of battery and are exacerbated under unfavorable conditions, such as extremely high or low temperatures ($T \leq 5\text{ }^{\circ}\text{C}$ or $T \geq 35\text{ }^{\circ}\text{C}$) and high state of charge ($\text{SoC} \geq 70\%$). Therefore, it is crucial to maintain these conditions within optimal ranges to extend battery life ([Chung et al., 2020](#)). Despite the crucial role of battery degradation for electric trucks operation and significant research about charging optimization for reducing battery degradation, to the best of our knowledge, there is scarce research in developing reliable methods to consider battery degradation in velocity optimization, which is one of our main contributions in this study.

On the discharging side, most research has focused on the vehicle-to-grid system, which addresses uncertainties such as commuting behavior, charging preferences, and energy requirements ([Maeng et al., 2023](#)). However, few studies have explored

optimizing the discharging process alongside vehicle dynamics or have considered battery degradation for BETs. Verbruggen et al. (2019) employ a nested optimization approach to optimize battery degradation and lifetime, as well as electric machine size and other aspects in the design of electric trucks. Similarly, Zhang et al. (2022) and Wang et al. (2024b) propose a hybrid powertrain for electric wheel loaders that takes into account battery degradation and the recapture of braking energy during acceleration to extend battery life. However, these models primarily focus on the powertrain configurations of electric wheel loaders, which are not solely powered by electricity and still pose environmental challenges due to their hybrid nature. Moreover, charging on BETs not only replenishes their energy (Schimpe et al., 2018b) but also generates heat, influencing thermal dynamics (Lin et al., 2014). This process affects battery temperature, thereby enabling control over both SoC and temperature, which is crucial for minimizing battery aging and capacity loss.

3. Methodology

This section introduces the proposed deep reinforcement learning framework for velocity optimization in BET, which integrates considerations of battery degradation and energy optimization. As depicted in Fig. 1, the proposed framework considers three critical components of BET velocity optimization: the vehicle dynamics model, the battery model, and the powertrain model. The vehicle dynamics model is crucial as it pertains to driving performance characteristics such as efficiency, safety, and comfort, which are significant for the BET agent in DRL. The powertrain model comprises the traction, transmission, and electrical power models, all of which significantly influence the energy consumption of BET. Additionally, the battery model is another vital aspect of this research. An equivalent circuit model is proposed to represent the electrochemical behavior of the battery, reflecting the adaptive responses of the battery of BET under varying driving and environmental conditions. Furthermore, a degradation model is employed to quantify battery degradation that includes calendar and cycling battery aging. Building on these three components, the reward function is developed using a mixture of experts model to facilitate the optimization process, as explained in Section 4. Then the environment for this study is formulated based on the BET vehicle configuration and referenced trajectory data, as introduced in Section 5 to simulate real-world conditions more accurately.

3.1. Problem formulation

In the DRL system, a velocity optimization strategy is designed to learn the optimal velocity planning action a_t for BETs in a car-following scenario. This involves considering the BET state s_t and battery degradation. Reinforcement learning facilitates the optimization by allowing a DRL agent to interact with the state of environment (Yang et al., 2024). Specifically, this study focuses on a car-following scenario where the trajectory of the LV is considered. The rationale for emphasizing car-following scenarios is that free-driving cases, extensively explored in existing research, are relatively simpler to optimize for BETs. In this context, BETs learn an acceleration or deceleration action a_t based on the current state s_t at each time step t , and receive a reward r_t based on the objectives of BET velocity optimization. The state of BET is then determined by the action, progressing to the next state s_{t+1} . This sequence continues until a terminal state is reached, at which point the velocity optimization process is reset. Ultimately, the optimal velocity optimization policy is learned through analyzing the trajectories of all leading vehicles. The state for the n th BET includes the following vehicle speed $V_n(t)$, spacing to the LV $S_{n-1,n}(t)$, and the relative speed $\Delta V_{n-1,n}(t)$. Additionally, the state of charge of the battery is incorporated into the state dimensions as well (Du et al., 2022).

3.2. Powertrain model

The dynamic models consists of vehicle dynamics, electric motor, partially following the dynamic model from Verbruggen et al. (2019), we formulate the model as follows.

3.2.1. Vehicle traction model

The traction force, generated by the electric powertrain motor, reflects the required force to achieve the desired vehicle speed v (Verbruggen et al., 2019). Furthermore, the vehicle traction model, which accounts for rolling friction, aerodynamic drag, and gradient resistance, is described as follows:

$$F_r = m_v \left(c_r g \cos(\alpha) + g \sin(\alpha) + \frac{d}{dt} v \right) + \frac{1}{2} \rho A_f c_d v^2, \quad (1)$$

Where road grade α is set to be zero, and other detailed descriptions and values for these parameters are provided in Table 1.

The angular speed ω , and torque demand T_v at the wheels can be calculated by

$$\omega = \frac{v}{r_w}, \quad (2)$$

$$T_v = F_r r_w. \quad (3)$$

where r_w is the wheel radius.

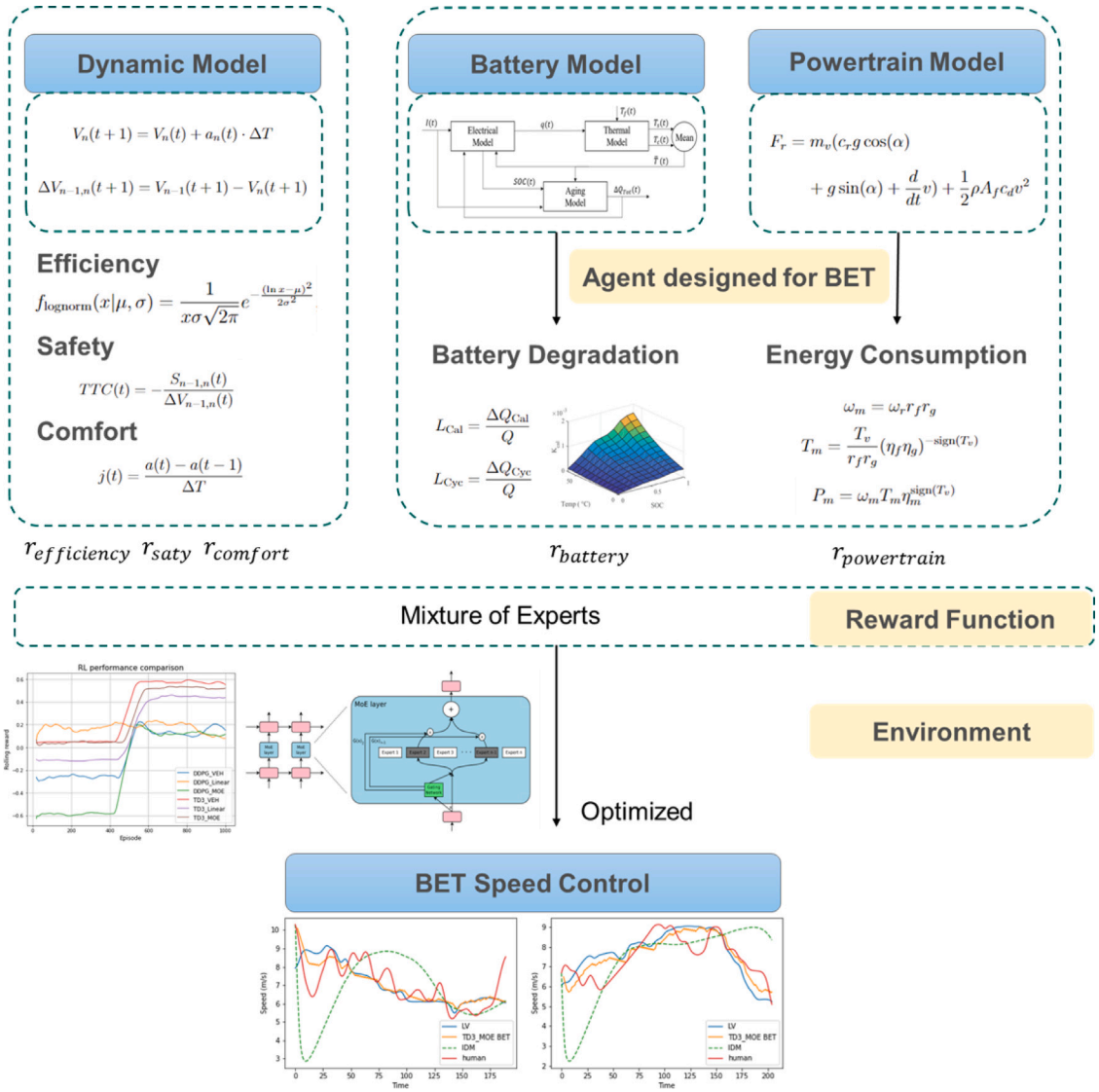


Fig. 1. Framework of BET velocity control.

Table 1
Summary of vehicle model parameters (Verbruggen et al., 2019).

Name	Symbol	Value	Unit
Wheel radius	r_w	0.492	m
Gravitational acceleration	g	9.81	m/s ²
Aerodynamic drag coefficient	c_d	0.73	-
Frontal area	A_f	0.75	m ²
Rolling friction coefficient	c_r	0.006	-
Air density	ρ	1.225	kg/m ³

3.2.2. Vehicle transmission model

The torque transmitted may experience some efficiency losses. To express the final drive transmission, the equations are formulated as follows:

$$\omega_m = \omega_r r_f r_g, \tag{4}$$

$$T_m = \frac{T_v}{r_f r_g} (\eta_f \eta_g)^{-\text{sign}(T_v)}. \tag{5}$$

where ω_m and T_m represent the angular speed and torque of the electric motor, respectively. Here, η_f and η_g are the fixed efficiencies of the final drive and the transmission, both set at 97%, applicable across all gears x_g . The transmission ratio r_g and the ratio of the final drive r_f are crucial; r_g is assumed to be a value reflecting gear engagement, and r_f is set to 1, indicating a 1:1 drive ratio.

3.2.3. Electric powertrain model

As for the energy consumption part, the electric powertrain consumption P_m is defined as

$$P_m = \omega_m T_m \eta_m^{\text{sign}(T_v)}, \quad (6)$$

where η_m is the efficiency of electric motor, depending on the torque ω_m and angular speed T_m . we assume the efficiency of motor to be constant as $\eta_m = 90\%$ in this study (Verbruggen et al., 2019).

3.3. Battery degradation model

This section introduces the electric-thermal-degradation model of the BET battery (Schimpe et al., 2018b; Lin et al., 2014). Additionally, the battery degradation of BET is modeled using a semi-empirical model that consist of calendar degradation and cyclic degradation (Schimpe et al., 2018a; Chung et al., 2020).

3.3.1. Electrical model

Firstly, as for the electrical dynamics in the battery cells of BET, an equivalent circuit model is modeled, where the terminal voltage V_t is formulated as

$$V_t = V_{OCV}(SoC, \bar{T}) + I \cdot R_1(SoC, \bar{T}) + \text{sgn}(I) \cdot V_{hys}(SoC) \quad (7)$$

where V_{OCV} is the open-circuit voltage (OCV), which is modeled by a semi-empirical model depends on the SoC and temperature \bar{T}

$$V_{OCV} = V_{ocv,Ref}(SoC) + (\bar{T} - T_{ref}) \cdot \left(\frac{dV_{OCV}}{dT} \right)_{SoC, T_{ref}}, \quad (8)$$

Here, $V_{ocv,Ref}(SoC)$ is the open-circuit voltage related to SoC and temperature, and $I \cdot R_1(SoC, \bar{T})$ captures the voltage drop under current I over the battery ohmic resistance R_1 . $\text{sgn}(I)$ represent the direction of current I , and $V_{hys}(SoC)$ is another non-linear relationship between hysteresis voltage and SoC.

Fig. 2 suggests the non-linear relationship of SoC to open circuit voltage V_{OCV} , battery cell resistance to temperature, battery cell resistance and hysteresis voltage to SoC on a Sony 26650 LiFePO₄ battery which identified from Schimpe et al. (2018b). We estimate the equivalent circuit model for BET battery cell based on the calibration of these relationship.

3.3.2. Thermal model

The temperature \bar{T} of battery cell is commonly defined as the average of the surface temperature T_s and the core temperature T_c as shown in Eq. (9).

$$\bar{T} = \frac{T_s + T_c}{2}, \quad (9)$$

where \bar{T} facilitates the modeling of heat generation of battery cells and heat conduction between the core and the surface of the battery cells. This is regulated through heat convection at the battery surface in contact with the coolant. The heat generation at the battery core, denoted by q_c , and the thermal convection from the core can be modeled as

$$\frac{dT_c}{dt} = \frac{T_s - T_c}{R_c C_c} + \frac{q_c}{C_c}, \quad (10)$$

where R_c is the heat conduction resistance, and C_c is the core heat capacity. The heat generation q_c can be derived from electric model in Eq. (7)

$$\begin{aligned} q_c &= I(V_t - V_{OCV,Ref}) \\ &= I^2 R_1 + I \text{sgn}(I) \cdot V_{hys}(SoC) + I(\bar{T} - T_{ref}) \cdot \left(\frac{dV_{OCV}}{dT} \right)_{SoC, T_{ref}} \end{aligned} \quad (11)$$

The surface temperature will then be cooled from the coolant in thermal management systems with the temperature T_f based on the surface convection resistance R_s and surface heat capacity C_s

$$\frac{dT_s}{dt} = \frac{T_f - T_s}{R_s C_s} - \frac{T_s - T_c}{R_c C_s} \quad (12)$$

The parameters of electric-thermal-degradation model has been validated through experiments conducted on the 26650 LiFePO₄ battery from Lin et al. (2014) and Chung et al. (2020). More detailed definitions, values of parameters, and 95% confident intervals can be found in Table 2. In this study, the temperature of coolant is set to 25 °C.

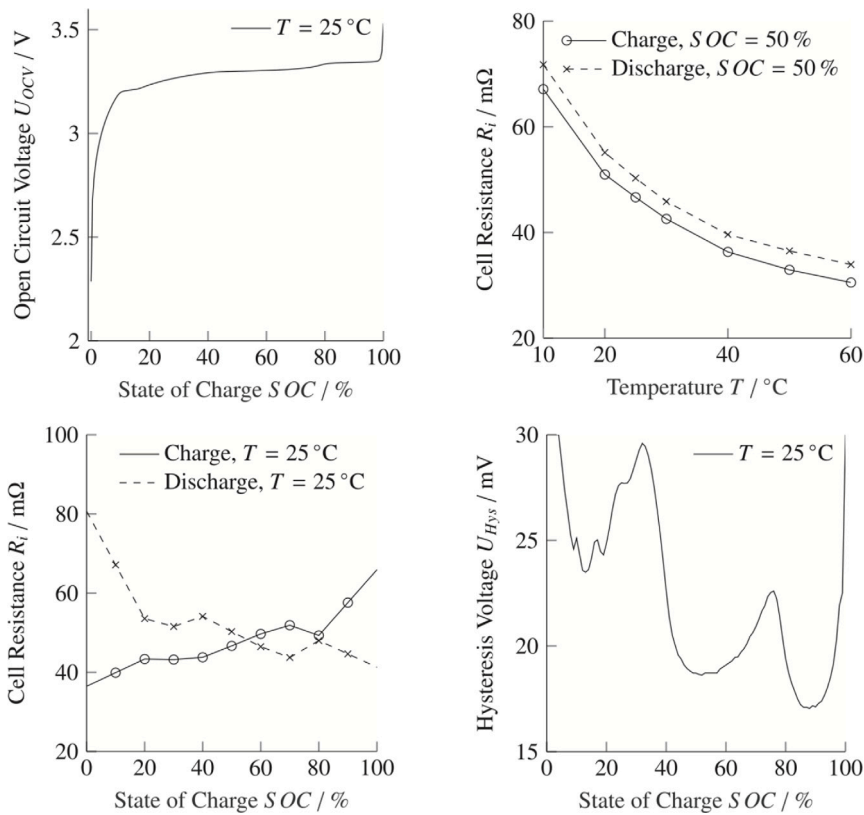


Fig. 2. Reference model (Schimpe et al., 2018b).

Table 2
Summary of thermal model parameters (Lin et al., 2014).

Parameter	Value	Unit	95% Confident interval
T_s	25	°C	–
C_s	4.5	J K ⁻¹	–
C_c	62.7	J K ⁻¹	59.1–66.4
R_s	3.19	KW ⁻¹	3.19–3.20
R_c	1.94	KW ⁻¹	1.86–2.01

3.3.3. Degradation model

We utilize the degradation model developed in Schimpe et al. (2018a), which encompasses calendar aging and cyclic aging. Calendar aging in battery is predominantly attributed to solid-electrolyte interface (SEI) layer. The degradation in capacity resulting from this growth exhibits a decrease over time because of the self-limiting nature. This dynamic is quantitatively described by the following integral model:

$$\Delta Q_{\text{Cal}} = \int [k_{\text{Cal}}(\text{SoC}, T) \cdot (2t^{0.5})^{-1}] dq, \quad (13)$$

where k_{Cal} is the degradation rate (stress factor), which is defined in Eq. (14). It incorporates Tafel and Arrhenius relationships to account for the impact of SoC-dependent effects and temperature dependence.

$$k_{\text{Cal}} = k_{\text{Cal,Ref}} \times \exp\left(-\frac{E_{a,\text{Cal}}}{R_g} \left(\frac{1}{T} - \frac{1}{T_{\text{Ref}}}\right)\right) \times \left(\exp\left[\frac{\gamma \cdot F}{R_g} \left(\frac{U_{a,\text{Ref}} - U_a(\text{SoC})}{T_{\text{Ref}}}\right)\right] + k_0\right), \quad (14)$$

where $E_{a,\text{Cal}}$ represents the activation energy associated with the degradation process, determined by fitting empirical data to the model at an SoC of 100%. The universal gas constant, R_g , used in the calculations is a constant parameter listed in Table 3. The second exponential term in Eq. (14) accounts for variations in the anode open circuit potential, U_a , which varies with the SoC. Additional parameters, γ and F , are detailed in Table 3.

In contrast to calendar aging, cyclic aging is influenced by different mechanisms (Chung et al., 2020), and it can be classified into several parts: the high-temperature cyclic aging degradation $\Delta Q_{\text{Cyc,HighT}}$ because of the SEI degradation, the low-temperature degradation $\Delta Q_{\text{Cyc,LowT}}$, and the low-temperature-high-SoC degradation $\Delta Q_{\text{Cyc,LowTHighSoC}}$ because of the lithium plating related to SoC and temperature difference.

Table 3
Summary of degradation model parameters (Schimpe et al., 2018a).

Model Parameter	Value	Environment
$k_{\text{Cal,Ref}}$	$3.69 \cdot 10^{-4} \cdot h^{-0.5}$	$T = 25 \text{ }^\circ\text{C}$, $SoC = 50\%$
$k_{\text{Cyc, High T, Ref}}$	$1.46 \cdot 10^{-4} \cdot Ah^{-0.5}$	$T = 25 \text{ }^\circ\text{C}$, $I = 1C$
$k_{\text{Cyc, Low T, Ref}}$	$4.01 \cdot 10^{-4} \cdot Ah^{-0.5}$	$T = 25 \text{ }^\circ\text{C}$, $I_{\text{Ch}} = 1C$
$k_{\text{Cyc,Low T High SoC,Ref}}$	$2.03 \cdot 10^{-6} \cdot Ah^{-1}$	$T = 25 \text{ }^\circ\text{C}$, $I_{\text{Ch}} = 1C$
$E_{a,\text{Cal}}$	$2.06 \cdot 10^4 \text{ J/mol}$	$SoC = 100\%$
$E_{a,\text{Cyc, High T}}$	$3.27 \cdot 10^4 \text{ J/mol}$	$I = 1C$
$E_{a,\text{Cyc, Low T}}$	$5.55 \cdot 10^4 \text{ J/mol}$	$I_{\text{Ch}} = 1C$
$E_{a,\text{Cyc,Low T High SoC}}$	$2.33 \cdot 10^5 \text{ J/mol}$	$I_{\text{Ch}} = 1C$
γ	$3.84 \cdot 10^{-1}$	
$\beta_{\text{Low T}}$	2.64 h	
$\beta_{\text{Low T High SoC}}$	7.84 h	
T_{Ref}	298.15 K	
$I_{\text{Ch,Ref}}$	3 A	
$U_{a,\text{Ref}}$	$1.23 \cdot 10^{-1} \text{ V}$	$SoC = 50\%$
k_0	$1.42 \cdot 10^{-1}$	

The $\Delta Q_{\text{Cyc,HighT}}$ focus on the effect of SEI formation during cycling in a high temperature status. Similar to ΔQ_{Cal} , the cyclic degradation in high temperature is modeled as

$$\Delta Q_{\text{Cyc,HighT}} = \int [k_{\text{Cyc,HighT}}(T) \cdot (2q^{0.5})^{-1}] dq \quad (15)$$

$$k_{\text{Cyc,HighT}} = k_{\text{Cyc,HighT,Ref}} \times \exp\left(-\frac{E_{a,\text{Cyc,HighT}}}{R_g} \left(\frac{1}{T} - \frac{1}{T_{\text{Ref}}}\right)\right) \quad (16)$$

where the related degradation factor $k_{\text{Cyc,HighT,Ref}}$ and activation energy $E_{a,\text{Cyc,HighT}}$ are detailed in Table 3.

The $\Delta Q_{\text{Cyc,LowT}}$ and $\Delta Q_{\text{Cyc,LowTHighSoC}}$ terms capture the effects of lithium plating under low temperature and high SoC, respectively. The main term $\Delta Q_{\text{Cyc,LowT}}$ is semi-empirically formulated as

$$\Delta Q_{\text{Cyc,LowT}} = \int [k_{\text{Cyc,LowT}}(T, I) \cdot (2q_{\text{chg}}^{0.5})^{-1}] dq_{\text{chg}}, \quad (17)$$

where $q_{\text{chg}} = \int |I_{\text{Ch}}(t)| dt$ is total charging current of battery cells.

$$k_{\text{Cyc,LowT}} = k_{\text{Cyc,LowT,Ref}} \times \exp\left(\frac{E_{a,\text{Cyc,LowT}}}{R_g} \left(\frac{1}{T} - \frac{1}{T_{\text{Ref}}}\right)\right) \times \exp\left(\beta_{\text{LowT}} \cdot \frac{I_{\text{Ch}} - I_{\text{Ch,Ref}}}{Q}\right), \quad (18)$$

where the cell capacity C_0 , reference current during charging $I_{\text{Ch,Ref}}$, and fitting parameter β_{LowT} are given in Table 3

Another degradation term under low temperature with high SoC is modeled as

$$\Delta Q_{\text{Cyc,LowTHighSoC}} = \int [k_{\text{Cyc,LowTHighSoC}} \cdot (2q_{\text{chg}}^{0.5})^{-1}] dq_{\text{chg}}. \quad (19)$$

And for the stress factor $k_{\text{Cyc,LowTHighSoC}}$, can be defined as

$$k_{\text{Cyc,LowTHighSoC}} = k_{\text{Cyc,LowTHighSoC,Ref}} \cdot \exp\left(-\frac{E_{a,\text{Cyc,LowTHighSoC}}}{R_g} \left(\frac{1}{T} - \frac{1}{T_{\text{Ref}}}\right)\right) \cdot \exp\left(\beta_{\text{LowTHighSoC}} \cdot \frac{I_{\text{Ch}} - I_{\text{Ch,Ref}}}{Q}\right) \cdot \left(\frac{\text{sgn}(SoC - SoC_{\text{Ref}}) + 1}{2}\right) \quad (20)$$

Therefore, the total capacity loss of battery cells ΔQ_{Total} is formulated as the sum of calendar loss, ΔQ_{Cal} , and cyclic losses. The cyclic losses are further broken down into $\Delta Q_{\text{Cyc}} = \Delta Q_{\text{Cyc,HighT}} + \Delta Q_{\text{Cyc,LowT}} + \Delta Q_{\text{Cyc,LowTHighSoC}}$.

$$\Delta Q_{\text{Total}} = \Delta Q_{\text{Cal}} + \Delta Q_{\text{Cyc}} \quad (21)$$

All the parameters are generated from the experiments conducted on a 26650 LiFePO4 battery in the literature (Chung et al., 2020; Schimpe et al., 2018a). The values of parameters for battery degradation models are summarized in Table 3.

3.4. Deep reinforcement learning for BET velocity optimization

This section introduces the DRL algorithms applied to BET velocity optimization including the foundational Deep Q-learning Network (DQN), Deep Deterministic Policy Gradient (DDPG) and Twin Delayed DDPG (TD3). These algorithms are chosen for their effectiveness in handling tasks that demand continuous action spaces as detailed in Table 4. As outlined in Section 3.1, the objective for BET velocity optimization is to optimize continuous acceleration actions while considering factors such as battery degradation and energy consumption. To address this, the DDPG and TD3 algorithms are selected for this study.

Table 4
Comparison of DRL Algorithms: DQN, DDPG, and TD3.

Algorithm	Action	Structure	Policy	Key features
DQN	Discrete	Value-based	Off-policy, deterministic	Utilizes a deep neural network to approximate the Q-value function, introduces experience replay and fixed Q-targets to stabilize learning in complex environments.
DDPG	Continuous	Actor-Critic	Off-policy, deterministic	Uses a deterministic policy to operate in continuous action spaces; employs actor-critic architecture with a replay buffer and target networks.
TD3	Continuous	Actor-Critic	Off-policy, deterministic	Improves upon DDPG by using twin Q-networks to reduce overestimation bias, delayed policy updates, and target policy smoothing.

3.4.1. DDPG

We will start with introducing the DQN algorithm, which is a basic and value-based reinforcement learning method (Yang et al., 2024). It evaluates the value of an action in a given state using a value function, which is continually updated through learning experiences. However, DQN is limited to discrete action spaces and does not suit tasks requiring continuous actions, such as truck acceleration. To address continuous action domains, the Policy Gradient method and Actor-Critic architecture enhance decision-making by combining gradient ascent on expected rewards with a dual approach that integrates both policy and value functions to optimize actions (Yang et al., 2024).

DDPG is an extension of these concepts, merging the benefits of experience replay of DQN and delayed target network updates with the Actor-Critic structure of Deterministic Policy Gradient. This hybrid method enhances training efficiency and robustness. Experience replay involves storing state–action–reward transitions in a memory buffer, which is continuously updated. Learning occurs from a minibatch randomly selected from this buffer, mitigating the correlation issues associated with sequential or related experiences. Both the Actor and Critic employ a pair of networks: a target network that aids in convergence, and an evaluation network facilitating continuous learning improvements. This structure ensures that both components update their strategies based on reliable and decorrelates feedback from the stored experiences. The key steps in the update phase can be summarized as:

- Calculating target values for the sampled transitions using the target networks, which estimate future rewards.
- Minimizing the loss between these target values and the values predicted by the evaluation networks, thereby refining the Critic estimates.
- Updating the Actor policy using a policy gradient method, thereby improving the policy based on the Critic feedback.
- Softly updating the target networks to slowly track the learned networks, maintaining stability.

Algorithm 1 DDPG Algorithm for BET Velocity Optimization

- 1: Initialize Critic $Q(s, a|\theta^Q)$ and Actor $\mu(s|\theta^\mu)$ networks and their targets Q' , μ' .
 - 2: Initialize replay buffer R .
 - 3: **for** episode = 1, M **do**
 - 4: Initialize observation state s_0 : spacing, speed, relative speed, and battery degradation.
 - 5: **for** $t = 1, T$ **do**
 - 6: Select and execute action of acceleration $a_t = \mu(s_t|\theta^\mu) + N_t$, observe reward r_t , and new state s_{t+1} .
 - 7: Store (s_t, a_t, r_t, s_{t+1}) in R .
 - 8: Sample minibatch from R , compute $y_t = r_t + \zeta Q'(s_{t+1}, \mu'(s_{t+1}))$.
 - 9: Update Critic and Actor networks.
 - 10: Soft update target networks $\theta^{Q'}$, $\theta^{\mu'}$.
 - 11: **end for**
 - 12: **end for**
-

The DDPG algorithm commences by initializing the evaluation and target networks for both the Critic and Actor, essential for approximating the Q-function and policy function, respectively. An empty replay buffer is also set up to store past experiences in the Algorithm 1. Each training episode starts by gathering initial observations related to the dynamics of BET, such as spacing, speed, and battery SoC. Actions, specifically vehicle acceleration, are selected using the current policy augmented with noise to encourage exploration. This noise addition is crucial for ensuring that the policy does not converge prematurely to suboptimal actions. Transitions consisting of the current state, action, resultant reward, and subsequent state are recorded in the replay buffer. Training involves sampling from this buffer to break correlation between sequential updates, which is vital for stable learning. The algorithm iterates through these steps until convergence or the end of the allowed episodes, updating policies and value estimations to navigate the BET efficiently and safely.

3.4.2. TD3

TD3 algorithm enhances the foundational approaches of DDPG to address critical vulnerabilities, particularly in the context of hyperparameter sensitivity and Q-value overestimation, which can degrade policy performance (Fujimoto et al., 2018). TD3 incorporates three innovative modifications to mitigate these issues, making it highly effective for applications like autonomous velocity planning and control that require continuous action decisions. Compared to DDPG, TD3 addresses the drawbacks of DDPG from three specific aspects:

- **Clipped Double-Q Learning.** TD3 mitigates the overestimation of Q-values, a common failure in DDPG, by maintaining two separate Q-functions. It uses the minimum of these two Q-values to compute the target values in the Bellman update, which helps in providing a more conservative estimate of the Q-values.

$$y_t = r_t + \zeta \min_{i=1,2} Q'_i(s_{t+1}, \mu'(s_{t+1}|\theta^{\mu'}))|\theta^{Q'_i} \quad (22)$$

- **Delayed Policy Updates.** To prevent the policy from exploiting inaccuracies in the Q-function, TD3 updates the policy less frequently than the Q-functions. Specifically, the policy and target networks are updated once for every two updates of the Q-function, allowing the Q-function to provide more stable and reliable estimates before each policy update.

$$a'(s') = \text{clip}\left(\mu_{\text{targ}}(s') + \text{clip}(\epsilon, -c, c), a_{\text{Low}}, a_{\text{High}}\right), \quad \epsilon \sim \mathcal{N}(0, \sigma) \quad (23)$$

- **Target Policy Smoothing.** TD3 introduces noise to the target policy to prevent the exploitation of Q-function errors. This noise smooths out the action selection of policy, making it harder for the policy to focus excessively on regions where the Q-function might be inaccurately high.

$$\theta^{Q'_i} \leftarrow \tau \theta^{Q'_i} + (1 - \tau) \theta^{Q'_i}, \quad \theta^{\mu'} \leftarrow \tau \theta^{\mu'} + (1 - \tau) \theta^{\mu'} \quad (24)$$

Both DDPG and TD3 are effective reinforcement learning methods for continuous action space optimization. However, due to the enhanced stability and reliability of TD3, it tends to be more effective in solving complex, non-linear problems, while DDPG may offer advantages in simpler, more linear scenarios. In the context of BET velocity control, the non-linear empirical model introduces a high level of complexity, which increases the optimization demands. The stability and robustness of TD3 make it particularly well-suited for controlling BET speed in dynamic environments. The complete TD3 algorithm is summarized in Algorithm 2.

Algorithm 2 TD3 Algorithm for BET Velocity Control

- 1: Initialize two Critic networks $Q_1(s, a|\theta^{Q_1})$ and $Q_2(s, a|\theta^{Q_2})$, and Actor $\mu(s|\theta^\mu)$ networks.
 - 2: Initialize target networks Q'_1, Q'_2 and μ' with weights $\theta^{Q'_1} \leftarrow \theta^{Q_1}, \theta^{Q'_2} \leftarrow \theta^{Q_2}, \theta^{\mu'} \leftarrow \theta^\mu$.
 - 3: Initialize replay buffer \mathcal{R} .
 - 4: **for** episode = 1, M **do**
 - 5: Initialize observation state s_0 : spacing, speed, relative speed, and battery degradation.
 - 6: **for** $t = 1, T$ **do**
 - 7: Select and execute action $a_t = \mu(s_t|\theta^\mu) + N_t$ (where N_t is noise for exploration).
 - 8: Execute a_t , observe reward r_t and new state s_{t+1} .
 - 9: Store transition (s_t, a_t, r_t, s_{t+1}) in \mathcal{R} .
 - 10: **if** $t \bmod d == 0$ **then** update networks every d steps
 - 11: Sample minibatch from \mathcal{R} .
 - 12: Compute target actions: $a' = \mu'(s_{j+1}|\theta^{\mu'}) + \text{clip}(\mathcal{N}(0, \sigma), -c, c)$.
 - 13: Compute target values: $y_j = r_j + \zeta \min_{i=1,2} Q'_i(s_{j+1}, a'|\theta^{Q'_i})$.
 - 14: Update Q_1 and Q_2 by minimizing the loss:

$$L_i = \frac{1}{N} \sum (y_j - Q_i(s_j, a_j|\theta^{Q_i}))^2 \quad \text{for } i = 1, 2$$
 - 15: Update μ using the sampled policy gradient:

$$\nabla_{\theta^\mu} J \approx \frac{1}{N} \sum \nabla_a Q_1(s, a|\theta^{Q_1})|_{s=s_t, a=\mu(s_t)} \nabla_{\theta^\mu} \mu(s|\theta^\mu)|_{s_t}$$
 - 16: Soft update target networks:

$$\theta^{Q'_i} \leftarrow \tau \theta^{Q_i} + (1 - \tau) \theta^{Q'_i} \text{ for } i = 1, 2, \theta^{\mu'} \leftarrow \tau \theta^\mu + (1 - \tau) \theta^{\mu'}$$
 - 17: **end if**
 - 18: **end for**
 - 19: **end for**
-

3.5. Mixture of experts

In the multi-objective optimization problem inherent in BET velocity optimization, the reward function comprises multiple components with varying scales. While many studies opt to manually adjust the weight of each component using a linear reward

function (Du et al., 2022; Yang et al., 2024), this approach may not convincingly handle the complexities of different objectives. Particularly in BET, where battery degradation, vehicle dynamics, and energy consumption vary in unit magnitude, ensuring precise optimization is crucial. A dominant objective could bias the velocity optimization strategy.

To address this, we have incorporated a Mixture of Experts structure in our study. The MoE concept (Jordan and Jacobs, 1994) employs a Gaussian mixture model that significantly enhances model capacity with minimal computational overhead. Recent adaptations include the MoE layer, extending to deep neural networks, which has shown substantial success in various deep learning applications. Typically, an MoE layer includes several expert networks that share the same architecture and are governed by a single algorithm, with a gating function directing inputs to select experts (Chen et al., 2022).

The MoE layer in our context consists of M expert networks, f_1, \dots, f_M , each paired with a linear gating network. The gating function $h(x; \Theta)$ aggregates input x across P dimensions, parameterized by $\Theta = [\theta_1, \dots, \theta_M] \in \mathbb{R}^{d \times M}$, and outputs an M -dimensional vector. The final output F of the MoE layer is computed as

$$F(x; \Theta, \mathbf{W}) = \sum_{m \in T_x} \pi_m(x; \Theta) f_m(x; \mathbf{W}), \quad (25)$$

where $T_x \subseteq [M]$ represents selected indices, and $\pi_m(x; \Theta)$ denotes the routing gate values,

$$\pi_m(x; \Theta) = \frac{\exp(h_m(x; \Theta))}{\sum_{m'=1}^M \exp(h_{m'}(x; \Theta))}. \quad (26)$$

Our implementation of MoE utilizes nonlinear neural networks for experts, critical for the success observed in our models. Specifically, for the m th expert, we employ a convolutional neural network

$$f_m(x; \mathbf{W}) = \sum_{j=1}^J \left[\left(\sum_{p=1}^P (w_{m,j} \cdot x^{(p)}) \right) \right], \quad (27)$$

with $w_{m,j} \in \mathbb{R}^d$ representing the weight vector of the j th filter in the m th expert, and J being the number of filters. The weights $\mathbf{W}_m = [w_{m,1}, \dots, w_{m,J}]$ define the weight matrix of the m th expert, collectively denoted as $\mathbf{W} = [\mathbf{W}_m]_{m=1}^M$. In this approach, the weights of each component in the reward function are dynamically adjusted during the reinforcement learning training process. This adjustment effectively addresses the challenge of balancing disparate objectives such as battery degradation and vehicle dynamics, which is crucial for managing the complexities inherent in BET velocity optimization strategies.

4. Features for reward function

The optimization objectives are structured around several primary aspects: battery degradation score r_b , energy consumption r_p , driving efficiency r_e , safety r_s , and comfort r_c , as formulated in Eq. (28):

$$r = w_b r_b + w_e r_e + w_s r_s + w_c r_c \quad (28)$$

where, w_b represent the weighted coefficients optimized using the MoE layer for the reward of battery degradation, which vary discretely from 0.1 to 10. The weights for velocity control reward r_e , r_s , and r_c to generally set to 1 to balance the effect of multiple objectives in the total reward, which widely adopted in other studies (Yang et al., 2024).

4.1. Battery degradation and energy consumption

One of our primary objectives is to mitigate battery degradation during operation. Battery degradation, denoted as ΔQ_{Tot} , comprises both calendar and cyclic degradation components as described in Eq. (21).

$$\Delta Q_{\text{Tot}} = \Delta Q_{\text{Cal}} + \Delta Q_{\text{Cyc}} \quad (29)$$

In this way, the calendar loss and cyclic loss of BET can be formulated as

$$L_{\text{Cal}} = \frac{\Delta Q_{\text{Cal}}}{Q} \quad (30)$$

$$L_{\text{Cyc}} = \frac{\Delta Q_{\text{Cyc}}}{Q} \quad (31)$$

$$L_{\text{Total}} = L_{\text{Cal}} + L_{\text{Cyc}} \quad (32)$$

where, $Q = 3C$ signifies the capacity of the battery cell employed in our simulations. Additionally, the score of battery degradation in the reward function, r_b , is defined by

$$r_b = L_{\text{Total}} \times k \quad (33)$$

where k is a scaling factor reflecting the simulated number of cycles. Considering the subtle nature of degradation measurements in the context of road experiments (Chung et al., 2020; Schimpe et al., 2018a), where the calendar and cyclic degradation during each simulation cycle are on the order of 10^{-8} to 10^{-7} Ah, these values are too small to interpret meaningfully on their own. Therefore, we applied the scaling factor k to adjust the degradation values to a magnitude comparable to other parameters in the model. This study models battery degradation over 10,000 operational cycles, thereby setting k at 10,000. It is important to note that the value of k does not affect the total reward function due to the self-adjusting mechanism of the weighted MoE layer.

4.2. Driving efficiency

Efficient driving is characterized by maintaining a short headway that falls within a safe range, defined as the time interval between the LV and the following vehicle (FV). In our research, the efficiency component of the reward function adopts the lognormal distribution model to calculate time headway values, a methodology aligned with prevalent practices in the field and partially based on [Zhu et al. \(2020\)](#). We use empirical NGSIM data to estimate the mean ($\mu = 0.4226$) and log standard deviation ($\sigma = 0.4365$) of time headway. This approach has been shown to enhance model stability and performance ([Pu et al., 2020](#)). Accordingly, our reinforcement learning agents are optimized to maintain a time headway of approximately 1.26 s, reflecting the optimal feature value identified in previous studies, where the probability density function is defined as

$$f_{\text{lognorm}}(x|\mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}, \quad x > 0 \quad (34)$$

where, x represents the time headway in this study. Then the score of headway feature r_e in the reward function is constructed as

$$r_e = f_{\text{lognorm}}(\text{headway}|\mu = 0.4226, \sigma = 0.4365) \quad (35)$$

4.3. Safety

Safety should be one of the most important elements of BET velocity planning which is evaluated by time to collision (TTC). This metric estimates the time remaining before two vehicles potentially collide, and is formulated as

$$TTC(t) = -\frac{S_{n-1,n}(t)}{\Delta V_{n-1,n}(t)} \quad (36)$$

Following the TTC calibrated in NGSIM data ([Zhu et al., 2020](#)), the normalized score of TTC feature in reward function is constructed as

$$r_s = \begin{cases} \log\left(\frac{TTC}{4}\right) & \text{if } 0 < TTC \leq 4 \\ -1 & \text{otherwise} \end{cases} \quad (37)$$

4.4. Driving comfort

The jerk of velocity profile $j(t)$, defined as the rate of change of acceleration, is a critical measure for assessing driving comfort due to its significant impact on passenger comfort and vehicle stability ([Du et al., 2022](#)).

$$j(t) = \frac{a(t) - a(t-1)}{\Delta T} \quad (38)$$

The jerk feature is evaluated as the score of comfort in the reward function r_c , is computed by

$$r_c = -\frac{j(t)^2}{1600} \quad (39)$$

This formulation implies that smaller jerk values, which correspond to smoother driving, result in higher comfort levels. The squared jerk is normalized by dividing by 1600, a base value derived from the following rationale

- The time step ΔT is 0.1 s.
- The acceleration is constrained between -2 m/s^2 and 2 m/s^2 for BET configuration.
- Consequently, the maximum possible jerk value is $\frac{2-(-2)}{0.1} = 40 \text{ m/s}^3$; squaring this value yields 1600.

Additionally, minimizing longitudinal acceleration enhances driving comfort by reducing speed variability, underscoring the importance of both longitudinal jerk and acceleration in the evaluation of longitudinal ride comfort ([Du et al., 2022](#)). However, it is noted that the scales of longitudinal jerk and acceleration differ significantly.

5. Experiment

5.1. BET vehicle configuration

To meet the voltage and power specifications for electric trucks, we incorporate a battery pack consisting of 88,000 LiFePO₄ cells for BET configuration which includes 220 cells arranged in series to achieve over 700 V and 400 groups in parallel, with each cell having a capacity of 3.0 Ah, resulting in a total nominal storage capacity of 845 kWh for the entire battery pack. The vehicle configuration refers to [Shoman et al. \(2023\)](#) that suggests BETs should be equipped with a battery capacity sufficient for a 435 km range (approximately 750 kWh), and [Zhang et al. \(2022\)](#) that utilized a battery pack LiFePO₄ cells providing 84.5 kWh within electric wheel loaders. Additionally, as detailed in Section 3.3, the initial temperature is set at 25 degrees Celsius at the beginning

Table 5
Model parameters of the IDM model used throughout this paper.

Parameter	Typical value
Desired velocity v_0	120 km/h
Safe time headway T	1.6 s
Maximum acceleration a	0.73 m/s ²
Desired deceleration b	1.67 m/s ²
Acceleration exponent δ	4
Jam distance s_0	2 m
Jam Distance s_1	0 m

of running considering the effects of battery thermal management. In the BET experiment training, a random SoC ranging from 35% to 95% is used to account for battery degradation, enhancing the model robustness by simulating diverse battery conditions and allowing the learning algorithm to explore a broader operational range. To simplify the model, it is assumed that all necessary information regarding the battery components of BET is accurately detected and integrated without any biases. For instance, both the environmental temperature and the temperature of the coolant interacting with the battery cells are set to be constant.

5.2. Reference trajectory data

The trajectory data are generated from Next Generation Simulation (NGSIM) project (Zhu et al., 2020). We focus on a truck following scenario where the LV is taken from NGSIM data. The velocity optimization and control produced by the DDPG and TD3 algorithms are compared with empirical observations from the trajectory data, especially those LV and FV data on trucks. The used trajectory data were collected on April 13, 2005, in the Bay area of Emeryville, CA. The dataset encompasses an aggregate of 45 min, segmented into three 15-minute intervals. These intervals capture various traffic conditions, ranging from the buildup of congestion to a fully congested state during peak traffic periods. Detailed vehicle location data, sampled at a rate of 0.1 s, were used to enhance model accuracy. Furthermore, the reconstructed NGSIM I-80 data were employed to ensure high data quality (Zhu et al., 2020).

5.3. Training configuration

Utilizing the NGSIM I-80 trajectory dataset, which comprises 1,341 pairs, we allocate 70% (938 pairs) for training and 30% (403 pairs) for testing. The training involves over 1,000 learning episodes, with each episode consisting of randomly selected car-following events from this dataset. The architecture for the actor and critic layers in the DDPG and TD3 models is standardized, each featuring a three-layer actor and critic neural network with dimensions of 64, 128, and 64, respectively, and concludes with a softmax activation layer. The MoE model consists of a three-layer GRU neural network with gate control. The design of DDPG, TD3, MoE, and their variants are detailed as follows:

- **TD3 Veh:** Applying the TD3 algorithm solely for BET vehicle velocity optimization, ignoring battery degradation with $r_b = 0$. The complete reward function is $r = r_e + r_s + r_c$.
- **TD3 Linear:** Using the TD3 algorithm for BET velocity optimization and battery degradation optimization, with equal weights in the reward function, namely $r = r_b + r_e + r_s + r_c$.
- **TD3 MoE:** Implementing the TD3 algorithm for BET velocity optimization and battery degradation optimization. The weight for battery degradation, w_b , is optimized through a MoE layer during training.
- **DDPG Veh:** Employing the DDPG algorithm solely for BET velocity optimization, without considering battery degradation, setting $r_b = 0$. The complete reward function is $r = r_e + r_s + r_c$.
- **DDPG Linear:** Utilizing the DDPG algorithm for BET velocity optimization and battery degradation optimization, with all weights in the reward function set to 1, such that $r = r_b + r_e + r_s + r_c$.
- **DDPG MoE:** Implements the DDPG algorithm for BET velocity and battery degradation optimization. The weight of battery degradation, w_b , is optimized through a MoE layer during training.

The training utilizes Pytorch 1.12.1 on an Nvidia A100 GPU and a 16-core CPU with 64 GB of memory.

5.4. Benchmark models

As for baseline model, we use intelligent driver model (IDM) as the benchmark (Treiber et al., 2000). The acceleration in IDM is a continuous function incorporating different driving modes, which is given by

$$a_{\text{IDM}}(s, v, \Delta v) = \frac{dv}{dt} = a \left[1 - \left(\frac{v}{v_0} \right)^\delta - \left(\frac{s^*(v, \Delta v)}{s} \right)^2 \right], \quad (40)$$

$$s^*(v, \Delta v) = s_0 + vT + \frac{v\Delta v}{2\sqrt{ab}}. \quad (41)$$

The model parameters are referred to Treiber et al. (2000) and summarized in Table 5.

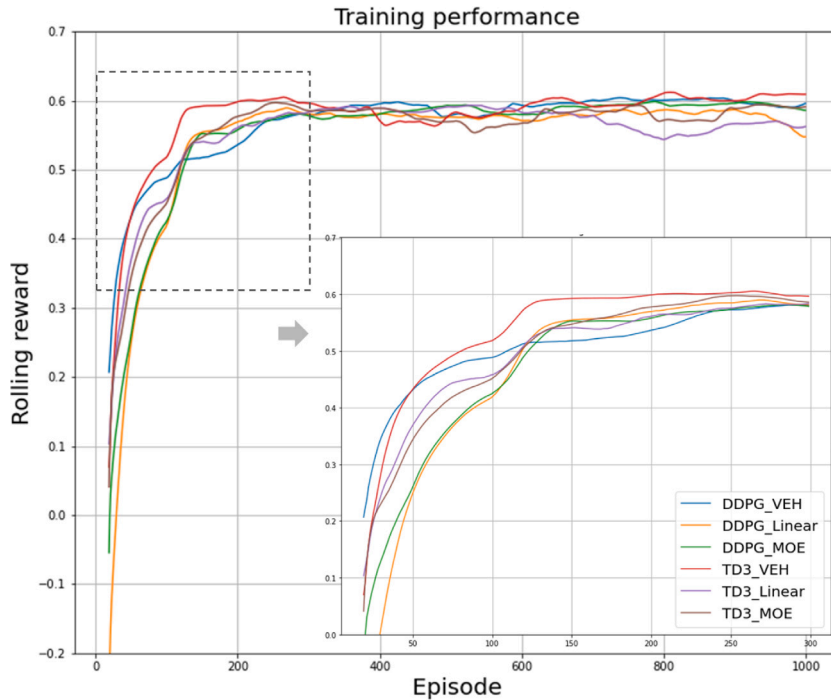


Fig. 3. Training reward comparison between DDPG, DDPG with MoE, TD3, TD3 with MoE.

6. Results

6.1. Training efficiency

Fig. 3 displays the rolling reward for six proposed reinforcement learning algorithms, illustrating their performance throughout the training phase. Although the peak rewards attained by the six algorithms are relatively similar, underscoring their effectiveness in velocity optimization, subtle differences are observed between episodes 800 to 1000. Among these, TD3-VEH and DDPG-VEH achieve the highest rolling rewards, demonstrating that reinforcement learning algorithms can achieve significant control effectiveness in terms of efficiency, safety, and comfort without considering battery degradation. When a linear score of battery degradation is incorporated into the reward function, DDPG-Linear and TD3-Linear show the least effective results among the six methods. This indicates that a simplistic linear integration of battery degradation may compromise the overall optimization objectives including efficiency, safety and comfort. However, this loss is mitigated when introducing the MOE approach. Both DDPG-MoE and TD3-MoE achieve more favorable optimization outcomes compared to their linear counterparts, highlighting the efficacy of the proposed method. In the initial episodes (from 0 to 300 episodes), the TD3 method demonstrates a more rapid convergence than the DDPG method. This observation is supported by the fact that the TD3-VEH curve is consistently lower than that of DDPG-VEH (0 to 50 episodes), which can be attributed to the variability in the initial SoC during training, leading to fluctuating initial rolling rewards. However, after the initial 50 episodes, TD3-VEH consistently outperforms DDPG-VEH, illustrating its superiority in terms of training efficiency. Furthermore, effectiveness of TD3 can be observed in the performance of TD3-Linear and TD3-MoE compared to their DDPG counterparts. Both TD3 variants perform better than the corresponding DDPG variants, especially in scenarios where the reward function incorporates the complexities of battery degradation. This suggests that TD3 is more effective in addressing complex problems, particularly those involving intricate reward structures that account for battery degradation.

To further explore the performance of the proposed methods in terms of velocity and battery degradation optimization, a detailed comparative analysis of model performances is presented in Table 6. This table summarizes key metrics calculated during training, including average training reward scores (r_s, r_e, r_c, r_b), BET vehicle dynamics such as average headway, jerk, and TTC per simulation step, as well as the average battery degradation and energy consumption after the rolling reward getting stable. It is important to note that battery degradation is calculated for each individual battery cell. However, when expressed as a relative percentage, this measure also reflects the overall degradation of the entire battery pack for the BET in this study. Moreover, energy consumption is normalized based on the overall powertrain energy from Eq. (6) and divided by the travel distance during simulation to make it comparable, resulting in a normalized unit of kWh/km. The total energy consumption is calculated for the entire BET, offering a comprehensive view of its energy efficiency.

Table 6
Summary of model performance metrics.

Model	r_s	r_e	r_c	r_b	Hdw (s)	Jerk (m/s^3)	TTC (s)	L_{Total} (%)	Energy (kWh/km)
DDPG-VEH	0	0.6546	-0.0007	-0.0030	1.4776	0.0297	25.1168	2.2067	2.2153
DDPG-Linear	0	0.6560	-0.0004	-0.0024	1.6977	0.0348	23.3014	2.2033	1.5247
DDPG-MoE	0	0.6552	-0.0003	-0.0023	1.5469	0.0355	24.9655	2.1934	1.5230
TD3-VEH	0	0.6551	-0.0002	-0.0018	1.4079	0.0326	25.0124	2.2039	1.6073
TD3-Linear	0	0.6558	-0.0003	-0.0022	1.6086	0.0434	24.7638	2.1937	1.4332
TD3-MoE	0	0.6534	-0.0016	-0.0017	1.5242	0.0344	24.5878	2.1872	1.4897

Regarding safety, all six methods recorded no collisions ($r_s = 0$), demonstrating their robustness in safety-related objective. Moreover, r_c and r_b , which represent the reward components for comfort and battery degradation respectively, show that these values are relatively close across all models. Compared to the efficiency metrics, the scores for comfort and battery degradation are marginal. This suggests that assigning equal weights ($w = 1$) to all components may lead to an underestimate on objectives that are less significant in scale. This observation highlights the necessity for a more refined approach in assigning weights to different performance aspects in our models. Furthermore, by applying a mixture of experts for re-weighting, DDPG-MoE and TD3-MoE achieved the lowest scores for battery degradation. The adaptive weighting in MOE allows for better adjustment of the relative importance of battery degradation in the total reward, enhancing the precision in evaluating its impact on overall performance.

DDPG-VEH and TD3-VEH models prioritize velocity optimization without explicitly addressing battery degradation. These models demonstrate relatively low headway times of 1.4776 s and 1.4079 s, respectively, which can increase traffic efficiency. Additionally, they achieve better safety performance with better TTC values compared to other methods. However, these models also exhibit higher levels of battery degradation (2.2067% for DDPG-VEH and 2.2039% for TD3-VEH) and energy consumption (2.2153 kWh/km for DDPG-VEH and 1.6073 kWh/km for TD3-VEH) compared to other methods.

Incorporating considerations of battery degradation into the reward mechanism reveals a significant trade-off. Models with more aggressive acceleration profiles, such as DDPG-Linear (0.0348 m/s^3) and TD3-Linear (0.0434 m/s^3), also show a reduction in TTC (23.3014 s for DDPG-Linear and 24.7639 s for TD3-Linear), indicating an improvement in overall vehicular dynamics. Furthermore, these adjustments lead to enhanced total battery degradation and energy consumption metrics, as demonstrated by DDPG-Linear and TD3-Linear.

Models utilizing mixture of experts approaches, DDPG-MoE and TD3-MoE, display improvements in battery degradation compared to other configurations. These models achieve smoother acceleration and significantly lower energy consumption by tailoring MoE within the total reward function. Remarkably, TD3-MoE and DDPG-MoE exhibit the lowest total battery degradation rates at 2.1872%, substantially enhancing battery longevity. Moreover, TD3-MoE records lower energy consumption at 1.4897 kWh/km, while DDPG-MoE achieves 1.5230 kWh/km. Although TD-MoE shows higher energy consumption compared to TD3-Linear, this observation does not undermine the effectiveness of the MoE approach. The subtlety of the optimal battery degradation reward weight (w_b) suggests that reductions in battery degradation are not immediately apparent in L_{Total} . Given the inherent randomness in training simulations, these variations are within acceptable limits, highlighting the nuanced effectiveness of the MoE strategy in dynamic reward adjustment.

6.2. Model performance in testing

We further test proposed methods in the test dataset to check the BET velocity optimization performance without any inferior in training dataset. We choose TD3-Linear and TD3-MoE which are the relatively best two methods as shown in Fig. 3. Unlike the random variable SoC used from 35% SoC to 95% SoC during training, we set the initial SoC of BET to be 75% in this analysis to ensure consistent and controlled comparisons of performance.

6.2.1. Demonstrations with sampled events

To illustrate the BET velocity optimization capabilities of the TD3 model, four car-following scenarios were selected and compared in this section. These scenarios were chosen to discuss the velocity optimization strategies across different driving conditions, including deceleration, acceleration, speed fluctuations, and scenarios involving relatively high speeds. Fig. 4 displays the observed leading vehicle speed alongside the actual human driving car-following speed, and the corresponding outputs generated by the TD3-VEH, TD3-MoE, and IDM models. A deceleration trajectory is shown in Fig. 4(a), with the initial speed of the controlled BET is based on the initial speed of the real car-following truck. The human driving trajectory exhibits some sudden stops and starts from the LV, whereas both TD3-VEH and TD3-MoE avoid these abrupt accelerations and deceleration, providing a smoother trajectory. Notably, TD3-MoE more closely matches the trajectory of LV. Fig. 4(b) illustrates a low-speed acceleration scenario. Initially, the policies trained by TD3-VEH and TD3-MoE attempt to reduce headway and closely follow the speed of LV within a safe distance. However, the human driving trajectory and the IDM method exhibit larger variations and delayed responses to speed of LV changes. Fig. 4(c) presents a longer trajectory where the LV exhibits significant speed variations. Here, the trained TD3-VEH and TD3-MoE models follow trajectories that lie between those of the LV and human driving. Fig. 4(d) shows a relatively high-speed acceleration scenario. The TD3-MoE demonstrates the smoothest trajectory, effectively meeting safety and comfort requirements. While the TD3-VEH also shows larger variations compared to TD3-MoE, but its trajectory generally lies between those of the LV and human driving, illustrating its effective adaptation to this driving context.

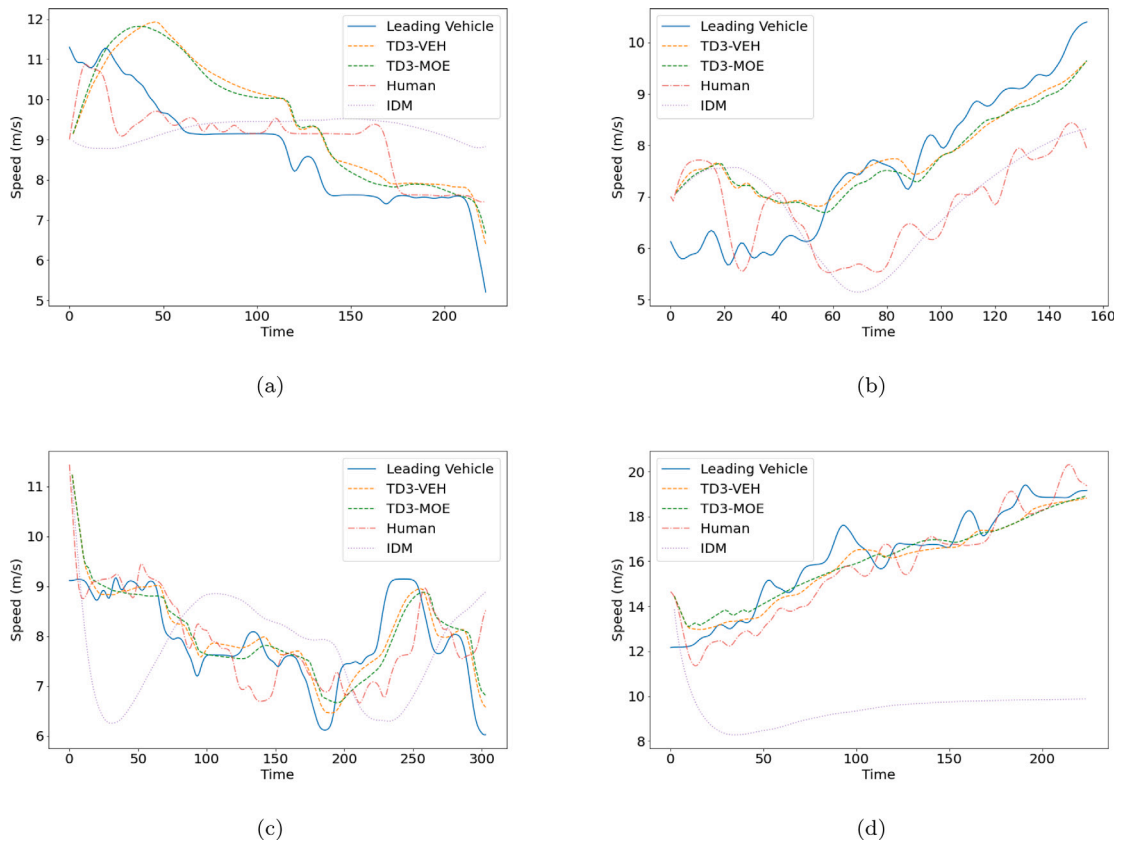


Fig. 4. Four selected car-following scenarios for algorithm comparison.

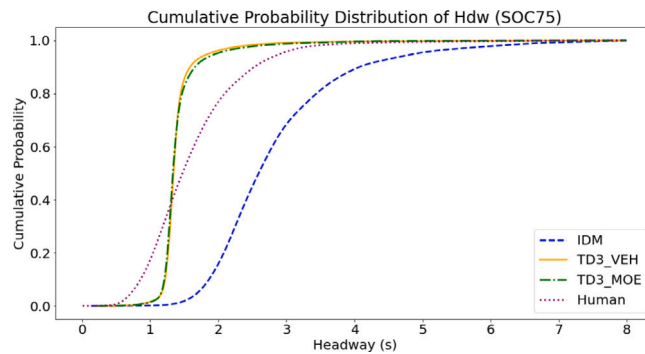


Fig. 5. Cumulative distribution of headway of BET (SoC=75%).

6.2.2. Vehicle dynamics

Vehicle dynamic testing is conducted on the test dataset as illustrated in Fig. 5, Fig. 6 and Fig. 10. The TD3-Linear and TD3-MoE, the two best-performing methods, are selected for this analysis. We set a consistent initial SoC of 75% for these tests. Time headway for testing is measured at each time step, with the cumulative distributions presented in Fig. 5. The real NGSIM dataset, which reflects human car-following driving behavior, exhibited a broad range of time headways, spanning from 0s to 6s. This range included potentially dangerous headways under 1 s and inefficient headways exceeding 3 s. In contrast, the TD3-VEH and TD3-MoE models maintained efficient and safe time headways, which are evident from their more centered distribution and reduced average headway in Fig. 5. These models more close to real-world human driving patterns compared to the IDM model, effectively following the leading vehicle with enhanced safety and efficiency.

Fig. 6 displays the cumulative distributions of TTC values for the TD3-VEH and TD3-MoE models, along with comparisons to human drivers and the IDM. The results indicate that both the TD3-VEH and TD3-MoE models yield higher TTC values compared

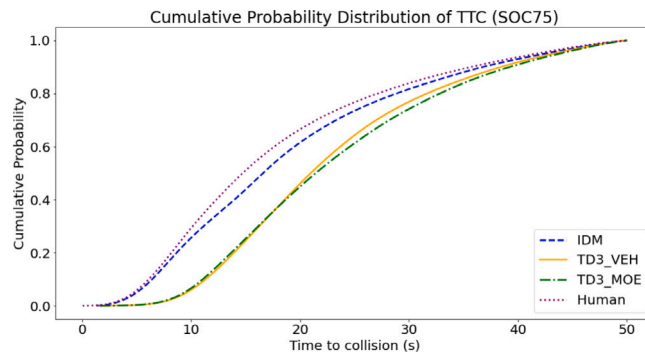


Fig. 6. Cumulative distribution of time to collision of BET (SoC=75%).

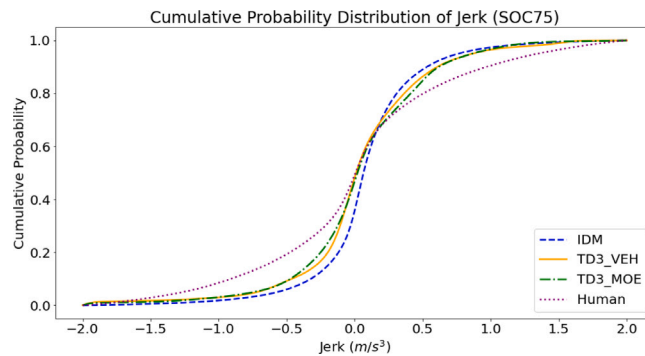


Fig. 7. Cumulative distribution of Jerk of BET (SoC=75%).

to those observed in human drivers and the IDM algorithm, suggesting safer car-following behavior. Furthermore, no collisions are recorded with either model during testing phases. The TTC distribution for both models is quite similar, with minor differences observed in the 20 to 50 s range, where the TD3-MoE model consistently shows slightly higher TTC values, indicating a higher safety margin.

Fig. 7 displays the cumulative distributions of jerk. It is evident that the TD3-VEH and TD3-MoE models produce trajectories with lower jerk values, with distributions more concentrated around the center. As lower jerk values indicate smoother and more comfortable driving experiences, it is reasonable to conclude that the TD3 models manage truck velocity more effectively than human drivers, and they also slightly surpass the performance of the IDM.

6.3. Battery degradation under different initial soc

In Section 3.3, we discussed the modeling of calendar loss L_{Cal} and cyclic loss L_{Cyc} in BET battery across different initial SoC. SoC is a critical factor in BET velocity optimization, significantly impacting battery degradation and energy consumption through a non-linear relationship. To explore this, we initialize the SoC of battery cells at 35%, 55%, 75%, and 95% to evaluate the efficacy of the proposed TD3-MoE method under varying initial SoC conditions. The detailed results of battery degradation are illustrated in Table 7, which categorizes degradation into calendar loss L_{Cal} , cyclic loss L_{Cyc} , and total degradation loss L_{Total} . It is important to note that while human and IDM driving policies directly relate to real-world vehicle following dynamics, variations in SoC may influence battery degradation values as well.

Table 7 reveals that higher SoC levels accelerate battery degradation regardless of velocity optimization algorithm, predominantly due to calendar loss, which is 2 to 10 times greater than the cyclic loss observed. The differences in calendar loss across various methods are trivial, but the variations in cyclic loss affected by velocity optimization contribute significantly to the difference in overall battery degradation. The total degradation L_{Cyc} is optimized by approximately 30% to 40% across different initial SoC levels compared to human driving data. Specifically, for TD3-MoE, the cyclic battery degradation L_{Cyc} is reduced by 27.7% at SoC=35%, 29.6% at SoC=55%, 29.8% at SoC=75%, and 29.5% at SoC=95% compared to the same initial SoC in human driving data, significantly mitigating cyclic loss. Furthermore, the MoE design significantly enhances the reduction in total battery degradation. L_{Total} is reduced by 8.3% at SoC=35%, 7.6% at SoC=55%, 5.2% at SoC=75%, and 2.4% at SoC=95% compared to a human-driven car following behavior. As for the energy consumption, TD3-MoE help to reduce 35.6% at SoC=35%, 39.8% at SoC=55%, 38.5% at SoC=75%, and 35.3% at SoC=95%.

Table 7
Performance metrics across different models and SoC levels.

Model	SoC	Hdw (s)	Jerk (m/s^3)	TTC (s)	L_{Cal} (%)	L_{Cyc} (%)	L_{Total} (%)	Energy (kWh/km)
TD3-VEH	35%	1.4589	0.0235	25.1496	1.2406	0.4265	1.6671	1.5380
TD3-MoE		1.5083	0.0174	23.8236	1.2413	0.4083	1.6496	1.4855
Human		1.6581	0.0191	19.0823	1.2335	0.5646	1.7981	2.3082
IDM		3.0074	0.0880	19.6172	1.2375	0.6529	1.8905	1.5340
TD3-VEH	55%	1.4749	0.0240	24.8181	1.5507	0.4109	1.9616	1.5191
TD3-MoE		1.4941	0.0253	23.8456	1.5511	0.3898	1.9409	1.3906
Human		1.6581	0.0191	19.0823	1.5489	0.5537	2.1026	2.3082
IDM		3.0074	0.0880	19.6172	1.5499	0.6471	2.1969	1.5340
TD3-VEH	75%	1.4468	0.0267	23.8439	2.0331	0.3490	2.3821	1.4724
TD3-MoE		1.5358	0.0257	24.1842	2.0347	0.3335	2.3682	1.4204
Human		1.6581	0.0191	19.0823	2.0234	0.4748	2.4982	2.3082
IDM		3.0074	0.0880	19.6172	2.0284	0.5535	2.5819	1.5340
TD3-VEH	95%	1.4085	0.0235	24.1879	2.4777	0.1662	2.6438	1.4614
TD3-MoE		1.5158	0.0241	24.3272	2.4778	0.1612	2.6390	1.4941
Human		1.6581	0.0191	19.0823	2.4762	0.2286	2.7049	2.3082
IDM		3.0074	0.0880	19.6172	2.4770	0.2584	2.7354	1.5340

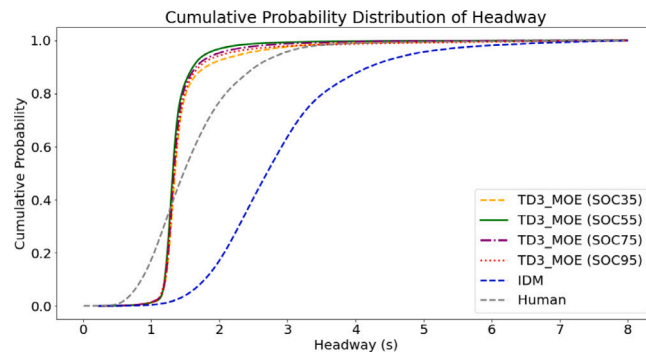


Fig. 8. Cumulative distribution of BET headway at different SoC.

TD3-MoE and TD3-VEH have similar metrics, yet TD3-MoE shows some improvements in reducing battery degradation, primarily due to better cyclic degradation results (4.3% better at SoC=35%, 5.1% at SoC=55%, 4.4% at SoC=75%, and 3.0% at SoC=95%). Since calendar loss is inevitable, improvements in cyclic loss may have more substantial benefits in long-term BET velocity optimization scenarios. The differences between TD3-MoE and TD3-VEH mainly arise from energy consumption perspectives, with improvements in energy consumption of 3.4% at SoC=35%, 8.5% at SoC=55%, 3.5% at SoC=75%, and a reduction -2.2% at SoC=95% by TD3-MoE. This negative value of -2.2% at SoC=95% suggests that in this state, TD3-MoE focuses more on other scores in the reward function, with the battery degradation score being less than $r_b = 1$ compared to the TD3-Linear model. It is reasonable to assume that at high SoC, a slight trade-off from energy consumption is made to improve driving efficiency (7.0% larger headway), safety (0.6% larger TTC), and battery degradation (3.0% reduction).

Fig. 8 shows that the TD3-MoE with an initial SoC of 55% achieves the lowest headway, suggesting enhanced performance in traffic flow efficiency. However, it is crucial to note that while our velocity optimization methods exhibit optimal performance at an SoC of 55%, this does not necessarily mean that the strategy at this SoC level is superior to those at other initial SoC. Indeed, the performance differences among various SoC levels for TD3-MoE are trivial, with all configurations surpassing both human driving and the IDM model. The similar conclusion is supported by the cumulative distribution of TTC and Jerk depicted in Fig. 9 and Fig. 10. Notably, with an initial SoC of 35%, the trained TD3-MoE tends to produce more aggressive control responses, such as larger headways, shorter TTCs, and more pronounced jerk. This can be interpreted as a consequence of battery degradation constraints, where the reinforcement learning algorithm compensates by reducing control precision to allow larger safety margins.

7. Discussion and conclusion

This study proposes a TD3-MoE reinforcement learning method to optimize BET velocity during car-following scenarios considering multiple objectives including safe, efficient and comfortable driving while minimizing battery degradation. Real-world driving data from NGSIM dataset are used to train and test the model. The proposed method is validated by comparing results with several benchmark models including DDPG, real car-following data, and the IDM with variants of considering battery degradation in velocity optimization or not. Results indicate that the proposed model for BET significantly outperforms human drivers and

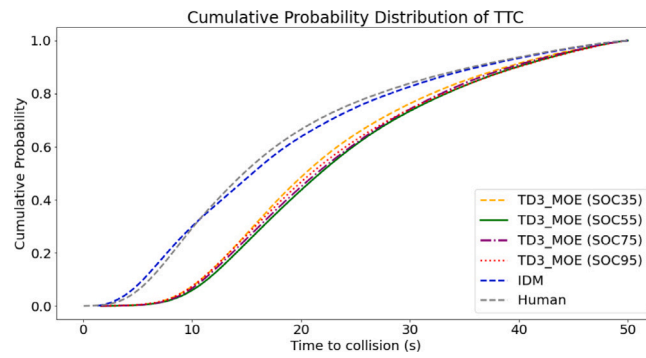


Fig. 9. Cumulative distribution of TTC of BET at different SoC.

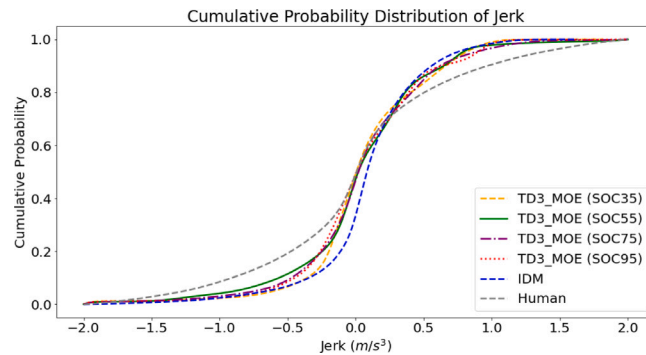


Fig. 10. Cumulative distribution of Jerk of BET at different SoC.

baseline models, demonstrating its capability to drive safely, efficiently, comfortably, with less battery degradation and better energy efficiency. The findings of this study can be summarized as follows:

- The convergence of the proposed TD3 algorithm, integrated with a mixture of experts model, shows superior performance and faster convergence compared to its counterparts. This outcome suggests that the algorithm configuration is well-suited for rapid adaptation with complex objectives.
- The TD3-MoE model consistently displays lower headway values and higher TTC values, with smoother trajectories compared to TD3-VEH, human driving data, and the IDM algorithm. The MoE structure helps to achieve better trade-off between velocity optimization and battery degradation rather than simply using constant weights in the reward function. Velocity optimization using TD3-MoE within safety bounds, enhances traffic flow efficiency and reduces battery degradation compared to other methods.
- As the SoC decreases, the proposed TD3-MoE becomes more effective at optimizing total battery degradation, achieving reductions of L_{Total} by 8.3% at SoC=35%, 7.6% at SoC of 55%, 5.2% at SoC of 75%, and 2.4% at SoC of 95% compared to a human driver. The main role of the TD3-MoE policy is to minimize cyclic loss to achieve this goal.
- The TD3-MoE model with an initial SoC of 55% achieves better energy consumption reduction compared to human drivers, with reductions of 39.8% at SoC of 55%. Moreover, reductions of energy consumption can reach 35.6% at SoC of 35%, 38.5% at SoC of 75%, and 35.3% at SoC of 95% as well.

However, this study could be further enhanced by designing better advanced reward mechanisms and by incorporating weight comparisons within the MoE model. For instance, the results show that, particularly concerning energy consumption, TD3-MoE increases total battery degradation at SoC of 95% by 2.2%. This shortfall could be addressed by incorporating a more complex MoE that separately handles calendar degradation and cyclic degradation with different weight to achieve a more balanced policy. Additionally, regenerative braking of BETs was not fully considered in our model. As a complex but widely applied strategy, this direction merits deeper exploration in subsequent studies. Furthermore, battery temperature can vary due to many factors, and maintaining a uniform temperature requires additional energy consumption. Moreover, road conditions (e.g., grade, surface smoothness) and traffic flow information were not considered in this study due to the lack of available data. Future work could incorporate these factors to further enhance real-world applicability.

CRediT authorship contribution statement

Ruo Jia: Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis. **Kun Gao:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Shaohua Cui:** Writing – original draft, Validation, Methodology, Investigation. **Jing Chen:** Writing – original draft, Methodology, Investigation. **Jelena Andric:** Writing – review & editing, Validation, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was funded and carried out as part of the e-MATS project (P2023-0029) funded by JPI Urban Europe and supported by Area of Advance Transport at Chalmers University of Technology, Sweden. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the sponsors' views.

Data availability

Data will be made available on request.

References

- Chen, Z., Deng, Y., Wu, Y., Gu, Q., Li, Y., 2022. Towards understanding the mixture-of-experts layer in deep learning. *Adv. Neural Inf. Process. Syst.* 35, 23049–23062.
- Chung, C.-H., Jangra, S., Lai, Q., Lin, X., 2020. Optimization of electric vehicle charging for battery maintenance and degradation management. *IEEE Trans. Transp. Electr.* 6 (3), 958–969.
- Cui, S., Gao, K., Yu, B., Ma, Z., Najafi, A., 2023. Joint optimal vehicle and recharging scheduling for mixed bus fleets under limited chargers. *Transp. Res. E* (ISSN: 1366-5545) 180, 103335.
- Du, Y., Chen, J., Zhao, C., Liu, C., Liao, F., Chan, C.-Y., 2022. Comfortable and energy-efficient speed control of autonomous vehicles on rough pavements using deep reinforcement learning. *Transp. Res. C* 134, 103489.
- Fei, Y., Shi, P., Liu, Y., Wang, L., 2024. Critical roles of control engineering in the development of intelligent and connected vehicles. *J. Intell. Connect. Veh.* 7 (2), 79–85.
- Fujimoto, S., Hoof, H., Meger, D., 2018. Addressing function approximation error in actor-critic methods. In: *International Conference on Machine Learning*. PMLR, pp. 1587–1596.
- Han, J., Sciarretta, A., Ojeda, L.L., De Nunzio, G., Thibault, L., 2018. Safe-and eco-driving control for connected and automated electric vehicles using analytical state-constrained optimal solution. *IEEE Trans. Intell. Veh.* 3 (2), 163–172.
- Han, Y., Wang, M., Leclercq, L., 2023. Leveraging reinforcement learning for dynamic traffic control: A survey and challenges for field implementation. *Commun. Transp. Res.* 3, 100104.
- Hayes, C.F., Rădulescu, R., Bargiacchi, E., Källström, J., Macfarlane, M., Reymond, M., Verstraeten, T., Zintgraf, L.M., Dazeley, R., Heintz, F., et al., 2022. A practical guide to multi-objective reinforcement learning and planning. *Auton. Agents Multi-Agent Syst.* 36 (1), 26.
- Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E., 1991. Adaptive mixtures of local experts. *Neural Comput.* 3 (1), 79–87.
- Jin, J., Ma, X., 2019. A multi-objective agent-based control approach with application in intelligent traffic signal system. *IEEE Trans. Intell. Transp. Syst.* 20 (10), 3900–3912.
- Jordan, M.I., Jacobs, R.A., 1994. Hierarchical mixtures of experts and the EM algorithm. *Neural Comput.* 6 (2), 181–214.
- Le Mero, L., Yi, D., Dianati, M., Mouzakitis, A., 2022. A survey on imitation learning techniques for end-to-end autonomous vehicles. *IEEE Trans. Intell. Transp. Syst.* 23 (9), 14128–14147.
- Li, W., Rios-Torres, J., Wang, B., Khattak, Z.H., 2024. Experimental assessment of communication delay's impact on connected automated vehicle speed volatility and energy consumption. *Commun. Transp. Res.* 4, 100136.
- Lin, W., Hu, X., Wang, J., 2023. Multi-level objective control of AVs at a saturated signalized intersection with multi-agent deep reinforcement learning approach. *J. Intell. Connect. Veh.* 6 (4), 250–263.
- Lin, Y., McPhee, J., Azad, N.L., 2020. Comparison of deep reinforcement learning and model predictive control for adaptive cruise control. *IEEE Trans. Intell. Veh.* 6 (2), 221–231.
- Lin, X., Perez, H.E., Mohan, S., Siegel, J.B., Stefanopoulou, A.G., Ding, Y., Castanier, M.P., 2014. A lumped-parameter electro-thermal model for cylindrical batteries. *J. Power Sources* 257, 1–11.
- Maeng, J., Min, D., Kang, Y., 2023. Intelligent charging and discharging of electric vehicles in a vehicle-to-grid system using a reinforcement learning-based approach. *Sustain. Energy Grid. Netw.* 36, 101224.
- Menezes, E.J.N., Araújo, A.M., Da Silva, N.S.B., 2018. A review on wind turbine control and its associated methods. *J. Clean. Prod.* 174, 945–953.
- Mulholland, E., Teter, J., Cazzola, P., McDonald, Z., Gallachóir, B.P.Ó., 2018. The long haul towards decarbonising road freight—A global assessment to 2050. *Appl. Energy* 216, 678–693.
- Osieczko, K., Zimon, D., Płaczek, E., Prokopiuk, I., 2021. Factors that influence the expansion of electric delivery vehicles and trucks in EU countries. *J. Environ. Manag.* 296, 113177.
- Pu, Z., Li, Z., Jiang, Y., Wang, Y., 2020. Full Bayesian before-after analysis of safety effects of variable speed limit system. *IEEE Trans. Intell. Transp. Syst.* 22 (2), 964–976.
- Qi, X., Luo, Y., Wu, G., Boriboonsomsin, K., Barth, M., 2019. Deep reinforcement learning enabled self-learning control for energy efficient driving. *Transp. Res. C* 99, 67–81.

- Qiu, L., Qian, L., Zomorodi, H., Pisu, P., 2017. Global optimal energy management control strategies for connected four-wheel-drive hybrid electric vehicles. *IET Intell. Transp. Syst.* 11 (5), 264–272.
- Qu, X., Lin, H., Liu, Y., 2023. Envisioning the future of transportation: Inspiration of ChatGPT and large models. *Commun. Transp. Res.* 3, 100103.
- Qu, X., Yu, Y., Zhou, M., Lin, C.-T., Wang, X., 2020. Jointly dampening traffic oscillations and improving energy consumption with electric, connected and automated vehicles: A reinforcement learning based approach. *Appl. Energy* 257, 114030.
- Schimpe, M., von Kuepach, M.E., Naumann, M., Hesse, H.C., Smith, K., Jossen, A., 2018a. Comprehensive modeling of temperature-dependent degradation mechanisms in lithium iron phosphate batteries. *J. Electrochem. Soc.* 165 (2), A181.
- Schimpe, M., Naumann, M., Truong, N., Hesse, H.C., Santhanagopalan, S., Saxon, A., Jossen, A., 2018b. Energy efficiency evaluation of a stationary lithium-ion battery container storage system via electro-thermal modeling and detailed component analysis. *Appl. Energy* 210, 211–229.
- Shi, H., Zhou, Y., Wu, K., Chen, S., Ran, B., Nie, Q., 2023. Physics-informed deep reinforcement learning-based integrated two-dimensional car-following control strategy for connected automated vehicles. *Knowl.-Based Syst.* 269, 110485.
- Shoman, W., Yeh, S., Sprei, F., Plötz, P., Speth, D., 2023. Battery electric long-haul trucks in europe: Public charging, energy, and power requirements. *Transp. Res. D* 121, 103825.
- Treiber, M., Hennecke, A., Helbing, D., 2000. Congested traffic states in empirical observations and microscopic simulations. *Phys. Rev. E* 62 (2), 1805.
- Verbruggen, F.J., Rangarajan, V., Hofman, T., 2019. Powertrain design optimization for a battery electric heavy-duty truck. In: 2019 American Control Conference. ACC, IEEE, pp. 1488–1493.
- Wang, S., Gao, K., Zhang, L., Liu, Y., Chen, L., 2024a. Probabilistic prediction of longitudinal trajectory considering driving heterogeneity with interpretability. *IEEE Intell. Transp. Syst. Mag.* 2–18.
- Wang, F., Zhang, Q., Wen, Q., Xu, B., 2024b. Improving productivity of a battery powered electric wheel loader with electric-hydraulic hybrid drive solution. *J. Clean. Prod.* 440, 140776.
- Wegener, M., Koch, L., Eisenbarth, M., Andert, J., 2021. Automated eco-driving in urban scenarios using deep reinforcement learning. *Transp. Res. C* 126, 102967.
- Xu, W., Liu, Q., Chen, M., Zeng, H., 2023. Ride the tide of traffic conditions: Opportunistic driving improves energy efficiency of timely truck transportation. *IEEE Trans. Intell. Transp. Syst.*
- Yang, Z., Zheng, Z., Kim, J., Rakha, H., 2024. Eco-driving strategies using reinforcement learning for mixed traffic in the vicinity of signalized intersections. *Transp. Res. C* 165, 104683.
- Ye, Y., Zhang, X., Sun, J., 2019. Automated vehicle's behavior decision making using deep reinforcement learning and high-fidelity simulation environment. *Transp. Res. C* 107, 155–170.
- Yu, B., Gao, K., Cheng, Z., Chen, Y., Yue, L., 2024. A human-like visual perception system for autonomous vehicles using a neuron-triggered hybrid unsupervised deep learning method. *IEEE Trans. Intell. Transp. Syst.* 25 (7), 8171–8180.
- Zhang, H., Wang, F., Xu, B., Fiebig, W., 2022. Extending battery lifetime for electric wheel loaders with electric-hydraulic hybrid powertrain. *Energy* 261, 125190.
- Zhou, Y., Lei, T., Liu, H., Du, N., Huang, Y., Zhao, V., Dai, A.M., Le, Q.V., Laudon, J., et al., 2022. Mixture-of-experts with expert choice routing. *Adv. Neural Inf. Process. Syst.* 35, 7103–7114.
- Zhu, M., Wang, Y., Pu, Z., Hu, J., Wang, X., Ke, R., 2020. Safe, efficient, and comfortable velocity control based on reinforcement learning for autonomous driving. *Transp. Res. C* 117, 102662.
- Zhu, M., Wang, X., Wang, Y., 2018. Human-like autonomous car-following model with deep reinforcement learning. *Transp. Res. C* 97, 348–368.