Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira,
Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria
 Published by Research Publishing, Singapore.
doi: 10.3850/ESREL2026061419_esrel26-p26178-cd

Vison-based Predictive Energy-efficient Train Control via Deep Reinforcement Learning

Wenzhen Zhai
School of Automation and Intelligence, Beijing Jiaotong University, China. E-mail: 24110027@bjtu.edu.cn

Yiran Gu
School of Automation and Intelligence, Beijing Jiaotong University, China. E-mail: 24120197@bjtu.edu.cn

Fei Yan
School of Automation and Intelligence, Beijing Jiaotong University, China. E-mail: fyan@bjtu.edu.cn

Qiang Zhang
Beijing Al for Rail Technology Co, Ltd, China. E-mail: zhangqiang@aiforail.com

The Virtually Coupled Train Set (VCTS) represents a next-generation train operation paradigm that significantly enhances line capacity through cooperative control among multiple trains. However, existing coordination strategies primarily emphasize safety assurance while overlooking systematic energy optimization, and they heavily rely on inter-train communication, making them highly sensitive to latency and packet loss. To address these issues, this paper proposes a predictive energy-efficient control framework based on visual context perception and deep reinforcement learning (DRL), enabling the following train to achieve autonomous energy optimization under weak communication conditions. The proposed framework employs onboard cameras to capture forward track scenes and uses a variational autoencoder (VAE) to extract latent environmental features, constructing a compact state representation. A Soft Actor-Critic (SAC) algorithm is then applied to derive optimal control policies by jointly considering safety, energy consumption, and ride comfort, Simulation results demonstrate that, compared with the conventional model predictive control (MPC) method, the proposed approach effectively reduces energy consumption and improves average speed while maintaining formation stability.

Keywords: Virtually coupled train set, Deep reinforcement learning, Variational autoencoder, Soft Actor-Critic.

1. Introduction

The Virtually Coupled Train Set (VCTS) enhances line capacity through dynamic cooperative control among multiple trains Xun et al. (2019). However, existing VCTS control strategies suffer from two core flaws: excessive reliance on continuous vehicle-to-vehicle communication Song et al. (2022), making them vulnerable to delays, packet loss, or interruptions—and crucially, this reliance on communication also invalidates conventional hard safety constraints that depend on real-time data exchange; and a primary focus on safe tracking while neglecting systematic energy optimization, thus failing to meet the urgent demand for green development in modern rail transit.

To reduce communication reliance, vision-based sensing offers rich information at low cost. Yet existing vision applications are mostly independent modules whose discrete outputs cannot be deeply integrated into continuous train control loops Hu et al. (2021). How to extract compact, driving-relevant features from visual streams in real time for communication-free tracking remains a key challenge.

At the control methodology level, traditional model-driven methods like Model Predictive Control (MPC) achieve accurate tracking under good communication but rely heavily on precise models and future trajectory predictions Li et al. (2025), leading to rapid performance degradation under communication interruptions or model mismatches. Data-Driven Control (DDC), particularly Deep Reinforcement Learning (DRL), has

shown promise in single-train energy-efficient operation Wang et al. (2024). However, when applied to multi-train virtual coupling, most DRL studies assume ideal communication for obtaining preceding train states Felez et al. (2019), thus failing to eliminate communication dependence. In fact, DRL offers unique advantages over MPC, observer-based control, or supervised learning: it is model-free Arulkumaran et al. (2017), learns end-to-end from high-dimensional visual inputs Mnih et al. (2015), and relies only on the ego vehicle's observations—naturally aligning with autonomous tracking under weak communication conditions.

In recent years, energy-efficient control in virtual coupling scenarios has gradually gained attention. Tian et al. (2025) and Lv et al. (2023) explored the application of MPC and DDPG, respectively, for energy saving in virtual coupling. However, the former relies on reliable communication, while the latter focuses on offline speed profile optimization; neither addresses online multi-objective optimization under weak communication conditions. Moreover, existing methods typically use simple linear weighting for multi-objective rewards, lacking thorough consideration of trade-offs among energy consumption, safety, ride comfort, and efficiency Feng et al. (2019). Another key challenge for data-driven controllers in safety-critical railways is the "black-box" problem: DRL policies behave unpredictably outside training distributions, and generalization risks are hard to quantify Han et al. (2021), while traditional model-based safety analysis is inapplicable to deep neural network policies. Scenario optimization theory offers a statistically high-confidence upper bound on future violation probability independent of the unknown environment distribution Campi et al. (2021), yet its application in train operation control—especially combined with vision-based DRL for rigorous safety certification—remains largely unexplored.

To address these challenges, this paper proposes a predictive energy-saving control framework based on visual perception and DRL, enabling following trains to achieve autonomous, safe, and energy-efficient tracking under weak communication. The framework extracts latent environmental features from track images via a Variational AutoEncoder (VAE), adopts Soft Actor-Critic (SAC) with a composite reward balancing energy, safety, comfort, and efficiency, and introduces scenario optimization to provide a statistical safety certificate.

The main contributions of this paper are as follows: 1) A communication-vision collaborative autonomous control architecture that eliminates reliance on continuous vehicle-to-vehicle communication; 2) Integration of visual features into continuous control via VAE, enabling end-to-end learning from vision to actions; 3) A composite reward with provable safety certification using scenario optimization, providing a distribution-free risk bound; 4) Simulation results showing 29.2% energy reduction, 6.08% improved tracking accuracy, and a safety violation bound below 1.58% with 95% confidence.

The remainder of this paper is organized as follows: Section 2 formulates the VCTS energy-efficient tracking control problem and constructs the baseline framework integrating visual perception and deep reinforcement learning. Section 3 proposes the enhanced control framework incorporated with safety verification mechanisms based on scenario optimization theory. Section 4 presents the experimental setup, comparative results and in-depth performance analysis. Finally, Section 5 summarizes the full text and discusses future research directions.

2. Fundamental Framework

This section formalizes the energy-efficient tracking control problem for a following train in a VCTS and presents a predictive control framework based on visual perception and DRL. This framework forms the foundation of this study, aiming to address the core challenge of trains achieving multi-objective optimization autonomously under weak communication conditions.

2.1. *Problem Formulation*

Consider a basic VCTS unit comprising one leading train and one following train. The control ob-

jective for the following train is to autonomously regulate its traction and braking forces to achieve safe, energy-efficient, comfortable, and efficient tracking operation, without relying on continuous and reliable communication with the leading train. To systematically describe this problem, we model the control task as a Partially Observable Markov Decision Process (POMDP), with its key elements defined as follows:

- **State Space**: The true state of the environment at time t is denoted by Δ_t, which includes relative dynamic information between the leading and following trains, track topology and gradient, and external disturbances. This state is not fully observable to the following train's controller.
- **Observation Space**: The following train obtains a front track scene image I_t via an onboard visual sensor, which is a high-dimensional raw observation encoding partial state information.
- **Action Space**: The control output is the acceleration command for the following train, $a_t \in [a_{\min}, a_{\max}]$, where $a_{\min} < 0$ is the maximum braking deceleration and $a_{\max} > 0$ is the maximum traction acceleration.
- **State Transition and Observation Model**: The environmental state evolves with Markov properties and generates noisy visual observations affected by sensor limitations and environmental uncertainties.
- **Reward Function**: A multi-objective reward function is designed to simultaneously reflect energy consumption, tracking safety, ride comfort, and operational efficiency.
- **Optimization Objective**: The goal is to find an optimal policy π^* that generates control actions based on the historical sequence of observations, maximizing the expected cumulative reward to achieve long-term multi-objective cooperative optimization.

In this work, we do not assume that safety constraints can be enforced as hard inequalities, because the required real-time communication and precise dynamic models are unavailable under our target weak-communication scenario. Instead, we adopt a statistical safety framework: the policy is learned with a safety-informed reward, and its violation probability is subsequently bounded using scenario optimization.

2.2. *Visual State Representation*

To transform the high-dimensional, unstructured raw image I_t captured by the onboard camera on the following train into a compact state representation suitable for reinforcement learning decision-making, this framework employs a VAE as the core feature extractor. VAE is an unsupervised generative model that maps input data to a low-dimensional continuous latent space and can reconstruct the original input from the latent representation.

For an input image I_t, the VAE encoder maps it to a probability distribution in the latent space, outputting the mean μ_t and log variance $\log \sigma_t^2$. A latent variable z_t is sampled via the reparameterization trick:

$$z_t = \mu_t + \sigma_t \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{1}$$

where $\odot$ denotes element-wise multiplication. The decoder takes z_t as input to reconstruct the image $\hat{I}_t$. The VAE is trained by maximizing the Evidence Lower Bound (ELBO):

$$\begin{aligned} \mathcal{L}_{\text{VAE}} = \mathbb{E}_{q_\phi(z|I)} \left[\log p_\psi(I|z)\right] \\ -D_{\text{KL}} \left(q_\phi(z|I) \| p(z)\right) \end{aligned} \tag{2}$$

Here, the first term is the reconstruction loss, which forces the latent representation to retain sufficient original information; the second term is the Kullback-Leibler divergence (KL divergence), which regularizes the latent space z to approximate a standard normal distribution $p(z) = \mathcal{N}(0, I)$.

The VAE compresses a 128×128 grayscale image into a 128-dimensional vector z_t. It filters out texture details and background noise irrelevant to driving decisions while preserving critical environmental structural information for tracking control. The KL divergence term regularizes the latent space to approximate a standard normal distribution, learning disentangled and smooth latent features that support stable controller decision-making. Finally, the extracted visual feature z_t is

concatenated with the train's directly measurable self-state to form the complete observation state for the DRL agent at time t:

$$s_t = [z_t; v_t; a_{t-1}] \tag{3}$$

The VAE is implemented with convolutional encoder and decoder, trained for 50 epochs with a reconstruction MSE below 3×10^{-3} on the validation set.

2.3. *Policy Optimization with SAC*

To address the continuous control problem of multi-objective cooperative optimization under partially observable conditions, this paper adopts the SAC algorithm within the maximum entropy DRL framework. Unlike standard reinforcement learning that only maximizes cumulative reward, the maximum entropy framework additionally maximizes policy entropy. This design encourages exploration, avoids premature convergence to sub-optimal solutions, and learns stochastic policies among equivalent optimal actions to enhance robustness.

The state s_t obtained in Section 2.2 serves as the agent's observation. The action space is the normalized acceleration command $\hat{a}_t \in [-1, 1]$, which is linearly transformed into the actual acceleration a_t. The agent's goal is to learn a stochastic policy $\pi_\theta(a_t|s_t)$ that maximizes the weighted sum of the expected cumulative reward and the policy entropy:

$$J(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(s_t,a_t)\sim\rho_\pi} \left[\gamma^t \left(r(s_t, a_t) + \alpha H(\pi(\cdot|s_t))\right)\right] \tag{4}$$

where ρ_π is the state-action distribution under policy π, γ is the discount factor, $\alpha > 0$ is the temperature parameter regulating the importance of the entropy term, and H is the policy entropy.

The reward function r_t is critical for balancing multiple performance objectives. We design a composite reward function consisting of four core sub-items:

$$r_t = \omega_e \cdot r_{e,t} + \omega_s \cdot r_{s,t} + \omega_c \cdot r_{c,t} + \omega_v \cdot r_{v,t} \tag{5}$$

where $\omega_e, \omega_s, \omega_c, \omega_v$ are the weight coefficients for each objective, with the sub-functions defined as follows:

- **Energy-saving reward** $r_{e,t}$: Penalizes traction energy consumption to encourage economical driving.

$$r_{e,t} = -\left(a_t^+\right)^2, \quad a_t^+ = \max(0, a_t) \tag{6}$$

- **Safety reward** $r_{s,t}$: A dual penalty term that punishes deviations from the desired distance and imposes exponentially increasing penalties when approaching the minimum safe distance, forming a "safety barrier".

$$\begin{aligned} r_{s,t} = &-\lambda_1 \left(d_t - d_{\text{des}}\right)^2 \\ &-\lambda_2 \cdot \exp\left(-\lambda_3 \cdot (d_t - d_{\text{min}})\right) \end{aligned} \tag{7}$$

- **Comfort reward** $r_{c,t}$: Penalizes abrupt changes in acceleration, i.e., jerk, to improve ride comfort.

$$r_{c,t} = -\left(\frac{a_t - a_{t-1}}{\Delta t}\right)^2 = -j_t^2 \tag{8}$$

- **Efficiency reward** $r_{v,t}$: Encourages the train to operate close to the line speed limit v_{max}.

$$r_{v,t} = -\left|v_t - v_{\text{max}}\right| \tag{9}$$

Since the four reward components in Eq. (5) have different physical units and magnitudes, we normalize each component offline using its empirical mean and standard deviation collected from a set of heuristic baseline trajectories. The weights $\omega_e, \omega_s, \omega_c, \omega_v$ are then selected via a coarse grid search to balance the four objectives, ensuring that no single objective dominates the cumulative reward. This design enables the agent to learn a well-balanced control policy.

The SAC algorithm operates with five collaborative neural networks: a policy network outputting action distributions, two Q-value networks for state-action value estimation (using the minimum value to avoid overestimation), and two target Q-networks updated softly from the main networks to enhance training stability. The core optimization process alternates between two steps: first, updating the Q-network parameters by minimizing temporal difference error (policy evaluation); second, updating the policy network pa-

rameters by maximizing the weighted sum of expected Q-values and policy entropy (policy improvement). The temperature parameter α is adaptively adjusted based on a preset target entropy, balancing exploration and exploitation of the policy.

3. Enhanced Framework with Safety Verification

The framework proposed in Section 2 optimizes control performance via a data-driven approach, but it suffers from a critical flaw: the 'black-box' nature of DRL policies leads to unquantifiable and unguaranteed generalization risks. Specifically, the probability of violating safety constraints in unknown scenarios outside the training data distribution cannot be assessed. To address this issue, this section introduces scenario optimization theory to construct a 'learning-guarantee' dual-loop enhanced framework. Unlike traditional deterministic safety verification, which requires explicit models and is inapplicable to deep neural network policies, scenario optimization provides a statistical safety certificate—a high-confidence upper bound on the future violation probability that is independent of the unknown environment distribution. A Risk Assessment and Certification (RAC) Loop is added outside the DRL Learning Loop, aiming to provide the trained policy with a rigorous, data-driven safety certificate.

The proposed enhanced framework, shown in Figure 1, consists of two interwoven loops:

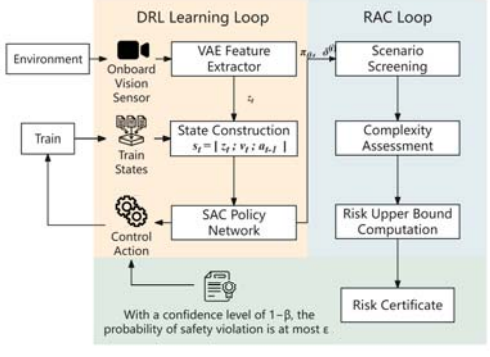

Fig. 1. The Proposed Framework

- **Inner Loop (DRL Learning Loop)**: This is the fundamental framework described in Section 2. It takes visual observations, extracts features via VAE, and optimizes policy parameters θ through the interaction between the SAC algorithm and the environment, ultimately outputting a high-performance policy π_θ to pursue optimal control performance.
- **Outer Loop (RAC Loop)**: This loop takes the policy π_θ and training interaction data as inputs. It employs scenario optimization theory to quantitatively evaluate the policy's generalization risk and outputs a high-confidence risk upper bound, serving as the safety guarantee for the learned policy.

The two loops form a closed cycle: the outer loop relies on the policy and data generated by the inner loop, and the risk certificate it outputs verifies the reliability of the inner loop's learning outcome.

Scenario optimization is a data-driven decision theory that enables decision-making based on finite independent and identically distributed scenarios and provides rigorous probabilistic guarantees for the generalization performance of solutions. Its core idea is as follows: for a decision $x \in X$, its performance under a future random scenario $\delta \in \Delta$ is measured by $f(x, \delta)$, where $f(x, \delta) \leq 0$ indicates satisfactory performance. The true scenario distribution P is unknown, and only N historical scenarios are available. The risk $V(x^*)$ of the optimal decision x^* is defined as the probability of unsatisfactory performance in future scenarios, which cannot be directly calculated due to the unknown P.

The core of the scenario approach is its "wait-and-judge" paradigm. Instead of conservatively estimating $V(x^*)$ in advance, it evaluates the risk by analyzing the complexity s^* of x^* after the decision is made. s^* refers to the minimum number of scenario samples required to uniquely determine x^*. Scenario theory proves that there exists a computable function $\varepsilon(s^*, N, \beta)$ such that $V(x^*) \leq \varepsilon(s^*, N, \beta)$ holds with a confidence level of at least $1 - \beta$.This risk upper bound only depends on s^*, N, and β and is independent of the

true distribution P, a characteristic called the separation principle that ensures the bound remains valid even if the prior model is incorrect.

We innovatively apply this theory to DRL policy certification by treating the trained policy π_θ as the decision x^* and taking safety constraint satisfaction as the performance evaluation target. A scenario δ is defined as an interaction trajectory segment, and a binary indicator function $J(\tau;\pi_\theta)$ is used to mark whether executing policy π_θ on trajectory τ violates safety constraints (1 for violation, 0 for compliance). The policy risk $V(\pi_\theta)$ is the probability of safety violations in future scenarios.

The complexity s^* of π_θ is defined as the minimum number of violation scenarios required to determine the policy's safety boundary, which can be approximated by analyzing support vectors or critical training samples that affect policy performance. Based on s^*, N, and the preset confidence parameter β, the risk upper bound ε can be calculated, providing a provable, distribution-free statistical safety guarantee for π_θ.

The complete workflow of the enhanced framework is streamlined as follows:

- Collect interaction data and train the initial DRL policy π_θ using the framework in Section 2.
- Screen trajectory segments with safety violations from the experience replay buffer to form a critical scenario set.
- Evaluate the complexity s^* of the current policy π_θ based on the critical scenario set.
- Calculate the risk upper bound ε using s^*, N, and β, and generate the safety certificate: with a confidence level of $1-\beta$, the probability of safety violation of policy π_θ in any future scenario does not exceed ε.

4. Experiments and Results

To comprehensively evaluate the effectiveness of the proposed vision-perception and DRL-based predictive energy-saving control framework, comparative experiments and analyses were conducted in a simulation environment, focusing on energy efficiency, safety, comfort, and comprehensive performance against the traditional MPC method for VCTS tracking control tasks. Experiments were conducted on an Intel Core i7 with NVIDIA RTX 2060 using Python 3.9, PyTorch 1.12, and Gymnasium. The simulation models a VCTS unit with one leading and one following train. Key parameters: $d_{\text{des}} = 50$ m, $d_{\min} = 10$ m, $v_{\max} = 30$ m/s, reward weights $\omega_e = 0.2$, $\omega_s = 0.5$, $\omega_c = 0.1$, $\omega_v = 0.2$. A VAE pre-trained on OSDaR23 Tagiew et al. (2023)) encodes 128×128 grayscale images into 128-dimensional features z_t, forming a 130-dimensional observation $s_t = [z_t; v_t; a_{t-1}]$.

An example of this data preprocessing is illustrated in Figure 2, which contrasts a raw track scene from the dataset with its corresponding preprocessed 128×128 grayscale input image. The SAC algorithm was adopted for training, with both policy and value networks designed as multilayer perceptrons with two hidden layers, the experience replay buffer capacity set to 100,000, and total training steps reaching 1,000,000 to ensure full policy convergence. The comparative MPC baseline adopted a simplified rolling optimization strategy with a prediction horizon of five, selecting optimal actions by enumerating finite candidate acceleration commands and simulating short-term future rewards. Each method was tested over five independent runs, and the average performance was recorded in Table 1.

Table 1. Performance Comparison between the Proposed Method and MPC

Metric	The Proposed Method	MPC
Reward	-1749.69	-1774.44
Average Tracking Distance (m)	45.59	57.45
Energy Consumption	6.85	9.68
Average Speed (m/s)	8.35	8.33
Comfort (RMS Jerk, m/s³)	0.365	0.159

Key metric analysis shows that the proposed method outperforms MPC in multiple core aspects: its slightly higher total reward verifies the global optimization capability of the multi-objective reward function-based strategy; the av-

Fig. 2. Example of visual input data preprocessing: (a) a raw track scene image from the OSDaR23 dataset; (b) the corresponding preprocessed 128×128 pixel grayscale input image.

erage tracking distance of 45.59 m is closer to the desired target, reducing the absolute tracking error by approximately 6.08% thanks to the predictive environmental perception capability of VAE-extracted visual features; traction energy consumption is reduced by 29.2%, attributed to the SAC algorithm's ability to learn smoother and more energy-efficient speed profiles under the maximum entropy framework. However, MPC achieves slightly better ride comfort, as its simplified linear control law generates more gradual action changes, while the SAC policy tolerates moderate acceleration fluctuations when balancing multiple objectives such as tracking precision and operational efficiency, indicating that adjusting the comfort weight ω_c in future work could further improve this performance.

To enable fair comparison across metrics with different units, each metric is normalized to a score in $[0, 1]$. For energy and comfort (RMS jerk), we apply $v' = 1/(1 + v)$ before normalization; tracking performance is evaluated by deviation from the desired 50 m. All other metrics are directly normalized using their empirical minima and maxima. Figure 3 comparison intuitively shows that the proposed method has clear advantages in total reward, distance tracking, and energy efficiency, is comparable to MPC in operational efficiency, and is slightly inferior only in ride comfort.

To verify the safety verification enhancement

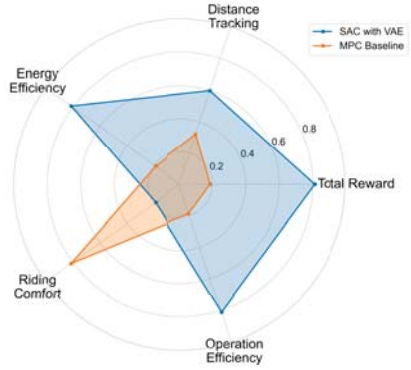

Fig. 3. Radar Chart Comparison of Comprehensive Performance between the Proposed Method and MPC

framework, scenario optimization theory was applied to conduct post-hoc risk assessment on the trained policy. A dataset with $N = 2000$ scenarios was constructed from independent training and testing interaction data, with a confidence parameter $\beta = 0.05$. The calculated upper bound of safety violation risk for the proposed method in unknown future scenarios is $\varepsilon(\hat{s}^* = 15, N = 2000, \beta = 0.05) \approx 1.58\%$, meaning we have 95% confidence that the probability of the policy violating safety constraints in any new scenario does not exceed 1.58%. Notably, this safety guarantee is statistical rather than deterministic: unlike traditional hard-constraint approaches that require accurate models and reliable communication to provide stronger guarantees, our frame-

work remains valid precisely under the weak-communication scenario where such prerequisites are not met. This rigorous, quantifiable risk certificate thus offers critical reliability support for deploying data-driven "black-box" control policies in safety-critical rail transit systems.

5. Conclusion

This study proposes a vision-based deep reinforcement learning framework for predictive energy-efficient control of VCTS under weak communication conditions. Leveraging VAE for compact environmental feature extraction and SAC algorithm for multi-objective optimization, the framework enables autonomous, safe and economical train tracking, outperforming conventional MPC with 29.2% lower energy consumption and improved tracking accuracy. Moreover, scenario optimization provides a rigorous safety certificate: the probability of safety violations in unseen scenarios is below 1.58% with 95% confidence. This work establishes a reliable "learning-guarantee" control paradigm for communication-resilient autonomous railways. Future research will focus on reward function trade-off optimization, real-vehicle generalization validation, and the exploration of online risk-adaptive mechanisms.

Acknowledgement

This work was supported by the National Natural Science Foundation of China Regional Innovation Joint Fund Key Project (U22A2053), the Guangxi Key R&D Project (Gui scientificAB 23075209), and the State Key Laboratory of Advanced Rail Autonomous Operation, Beijing Jiaotong University (Project No. RAO2026K03).

References

Arulkumaran, K., M. P. Deisenroth, M. Brundage, and A. A. Bharath (2017). A brief survey of deep reinforcement learning. *arXiv preprint arXiv:1708.05866*.

Campi, M. C., A. Carè, and S. Garatti (2021). The scenario approach: A tool at the service of data-driven decision making. *Annual Reviews in Control 52*, 1–17.

Felez, J., Y. Kim, and F. Borrelli (2019). A model predictive control approach for virtual coupling in railways. *IEEE Transactions on Intelligent Transportation Systems 20*(7), 2728–2739.

Feng, S., Y. Zhang, S. E. Li, Z. Cao, H. X. Liu, and L. Li (2019). String stability for vehicular platoon control: Definitions and analysis methods. *Annual Reviews in Control 47*, 81–97.

Han, M., Y. Tian, L. Zhang, J. Wang, and W. Pan (2021). Reinforcement learning control of constrained dynamic systems with uniformly ultimate boundedness stability guarantee. *Automatica 129*, 109689.

Hu, X., Y. Cao, Y. Sun, and T. Tang (2021). Railway automatic switch stationary contacts wear detection under few-shot occasions. *IEEE Transactions on Intelligent Transportation Systems 23*(9), 14893–14907.

Li, J., D. Tian, J. Zhou, X. Duan, J. Zhang, Z. Sheng, D. Zhao, and D. Cao (2025). A mpc-based distributed bidirectional control strategy for virtual coupling with unreliable train-to-train communications. *IEEE Internet of Things Journal*.

Lv, W., H. Liu, Y. Lang, and X. Luo (2023). Optimization of energy-saving operation strategy for virtual coupling based on ddpg. In *2023 China Automation Congress (CAC)*.

Mnih, V., K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, and G. Ostrovski (2015). Human-level control through deep reinforcement learning. *Nature 518*(7540), 529–533.

Song, X., D. Wei, Y. Chen, and W. Zhu (2022). Bidirectional asymmetric delay feedback for cooperative adaptive cruise control of vehicle platoons with unreliable communication. *Asian Journal of Control 24*(6), 3066–3079.

Tagiew, R., P. Klasek, R. Tilly, M. Köppel, P. Denzler, P. Neumaier, T. Klockau, M. Boekhoff, and K. Schwalbe (2023). Osdar23: Open sensor data for rail 2023. In *2023 8th International Conference on Robotics and Automation Engineering (ICRAE)*.

Tian, S., F. Yan, W.-L. Shang, A. Majumdar, H. Chen, M. Chen, M. Zeinab, and Y. Tian (2025). Energy consumption analysis of trains based on multi-mode virtual coupling operation control strategies. *Applied Energy 377*, 124684.

Wang, D., J. Wu, Y. Wei, X. Chang, and H. Yin (2024). Energy-saving operation in urban rail transit: A deep reinforcement learning approach with speed optimization. *Travel Behaviour and Society 36*, 100796.

Xun, J., J. Yin, R. Liu, F. Liu, Y. Zhou, and T. Tang (2019). Cooperative control of high-speed trains for headway regulation: A self-triggered model predictive control based approach. *Transportation Research Part C: Emerging Technologies 102*, 106–120.