A Study on Homogeneous Cooling Control of Multi-Channel Battery Modules Based on Reinforcement Learning

Hyeongkyu Jin[1], Yoonjeong Min[2], Taeyun Park[2], Jaejeung Kim[*]
[1] *Chungnam National University, Department of Bio-AI Convergence, Daejeon, South Korea*
[2] *SCS Inc., SW Development Team, Sejong, South Korea*
[*] *Chungnam National University, Department of Computer Science and Engineering, Daejeon, South Korea*
denevelove@gmail.com, {yjmin, typark}@scsai.net, jjkim@cnu.ac.kr

***Abstract*— The performance, lifespan, and safety of lithium-ion battery packs are critically dependent on thermal management. A key challenge is the temperature inhomogeneity among cells in multi-channel cooling structures, which amplifies electrothermal instability, leading to localized degradation and performance deterioration. Conventional control methods, such as PID or rule-based controls, struggle to address the complex and nonlinear dynamics of the system and are limited in their ability to ensure temperature uniformity as they typically rely on average or maximum temperature values. This study proposes an active and homogeneous cooling control strategy for multi-channel battery modules using Deep Reinforcement Learning (DRL). The proposed controller takes real-time state inputs from the battery module including multi-point temperatures, voltage, and current to execute continuous actions that individually adjust the coolant flow rate in each channel. A multi-objective reward function was designed to consider four critical goals simultaneously: Safety, Uniformity, Energy Efficiency, and Control Stability, enabling the agent to learn a policy that achieves these complex objectives. The efficacy of the proposed method was validated using the Twin-Delayed Deep Deterministic policy gradient (TD3) algorithm in a high-fidelity simulation environment. The results demonstrate that the DRL controller significantly reduces the maximum temperature deviation compared to conventional methods and optimizes the energy consumed for cooling, thereby enhancing the overall performance and durability of the battery system. This research presents a successful application of a data-driven, intelligent control technique to the complex problem of battery thermal management and is expected to contribute to the advancement of thermal management technologies for high-performance electric vehicles and energy storage systems.**

Keywords*— *Battery Thermal Management, Temperature Homogeneity, Reinforcement Learning, Deep Deterministic Policy Gradient (DDPG), Multi-objective Optimization, Intelligent Control

I. INTRODUCTION

As high-energy-density lithium-ion batteries become central components in electric vehicles (EVs) and energy storage systems (ESS), the importance of the Battery Thermal Management System (BTMS) has become more pronounced than ever. The optimal operating temperature range for batteries is generally known to be between 20°C and 40°C. Operating outside this range can lead to performance degradation, reduced lifespan, and, in severe cases, safety issues like thermal runaway [1, 2].

However, traditional thermal management research has primarily focused on maintaining the 'maximum temperature' of the battery pack within safe limits. This study argues that such an approach is insufficient for ensuring the long-term reliability of the battery system and posits that achieving 'temperature homogeneity' among the cells within the pack is a more fundamental challenge. Even minor temperature differences can cause variations in the electrochemical properties of cells, leading to imbalanced current distribution and, consequently, differences in heat generation. This positive feedback loop of 'electro-thermal instability' amplifies temperature deviations over time, accelerating the premature degradation of specific cells and reducing the lifespan and available capacity of the entire pack [2, 5]. According to research, a temperature difference of just 5°C can increase thermal aging by 25% and reduce pack capacity by up to 2% [2].

Commonly used air and liquid cooling systems can themselves be a source of the problem, as the cooling medium inevitably heats up as it absorbs heat, creating a temperature gradient between the inlet and outlet [4, 5]. Furthermore, Proportional-Integral-Differential (PID) controllers, which are widely used in industry, have a fundamental limitation in that they control the entire system based on a single value, such as the average temperature, thus failing to consider the spatial distribution of temperature [11].

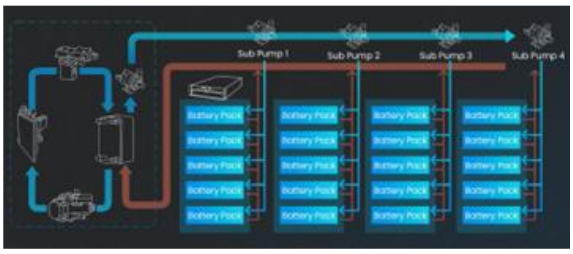

Fig. 1. BESS Liquid Cooling System schematic.

Therefore, this paper proposes Deep Reinforcement Learning (DRL) as a solution to effectively respond to the complex and nonlinear thermal dynamics of batteries and to intelligently control a multi-channel cooling system to maximize temperature homogeneity. As a model-free approach, reinforcement learning is ideal for this problem because it can learn an optimal control policy through data-driven interactions without requiring a perfect mathematical model of the system. In this study, we formulate the battery thermal management problem as a Markov Decision Process (MDP) and aim to develop an optimal controller by designing a multi-objective reward function that considers safety, uniformity, efficiency, and stability.

**corresponding author*

II. Reinforcement Learning Problem Formulation

To solve the homogeneous cooling problem for a battery module, we formalize it within the framework of a Markov Decision Process (MDP). An MDP is defined by an agent (the controller), an environment (the battery system), a set of states, a set of actions, and a reward function.

A. Sate Space Design

The state (st) is the set of information the agent observes from the system at a specific time t. For effective control, the state must comprehensively represent the thermal and electrical distribution of the system. The state vector is composed as follows:

- Multi-point Temperatures (T_{cells}): Cell surface temperatures measured from multiple sensors distributed throughout the module. This is a key element providing spatial information about the temperature distribution [15, 16].
- Coolant Inlet/Outlet Temperatures ($T_{coolant}$): Indicates the overall heat removal performance of the cooling system.
- Electrical State (I_{pack},V_{pack}): The total current and voltage of the pack. This provides crucial information for predicting the current load status and near-future heat generation [15].
- Internal State Estimates (SoC_{cells}, SoH_{cells}): The State of Charge (SoC) and State of Health (SoH) of each cell. These values, estimated by the Battery Management System (BMS), allow the agent to learn the changes in dynamics due to cell degradation [8].
- Ambient Temperature ($T_{ambient}$): A variable to account for the thermal load from the external environment [15].

B. Action Space Design

The action (at) is the decision the agent makes after observing the state st. To achieve the goal of localized temperature control, we design a multi-dimensional, continuous action space that allows for independent control of each cooling channel. For a 4-channel liquid cooling system, the action vector is as follows:

$$at=[flow\_ratech1, flow\_ratech2, flow\_ratech3, flow\_ratech4] \quad (1)$$

The flow rate for each channel is a continuous value within the range $[0, flowmax]$, enabling the agent to perform precise and differential control, such as directing more coolant flow to hotter areas.

C. Multi-objective Reward Function Design

The reward function (rt) is the most critical design element that guides the agent's learning process. To achieve the complex objectives required for battery thermal management, this study proposes a multi-objective reward function that combines four goals in the form of negative penalties. The agent learns to minimize the sum of these penalties (i.e., maximize the cumulative reward).

$$rt = r_{safety} + r_{uniformity} + r_{efficiency} + r_{stability} \quad (2)$$

- Safety Penalty (r_{safety}): Enforces that the battery cell temperature does not exceed a predefined safety limit (T_{limit}). A quadratic term is used to ensure the penalty increases exponentially as the temperature surpasses the limit.

$$r_{safety} = -w_s \cdot max(0,\ T_{max\_cells} - T_{limit})^2 \quad (3)$$

- Uniformity Penalty ($r_{uniformity}$): The core penalty for ensuring temperature homogeneity. It combines the max-min temperature difference to control outliers and the standard deviation (σ) to consider the overall temperature distribution.

$$r_{uniformity} = -w_{u1} \cdot (T_{max\_cells} - T_{min\_cells}) - w_{u2} \cdot \sigma(T_{cells}) \quad (4)$$

- Energy Efficiency Penalty ($r_{efficiency}$): Discourages excessive energy consumption by the cooling system (pumps, fans, etc.). This minimizes unnecessary cooling operations and increases overall system efficiency.

$$r_{efficiency} = -w_e \cdot P_{cooling} \quad (5)$$

- Control Stability Penalty ($r_{stability}$): A penalty considering real-world hardware application, which suppresses abrupt changes in the control values (actions). This reduces mechanical stress on actuators and ensures stable system operation.

$$r_{stability} = -w_{st} \cdot |a_t - a_{t-1}|^2 \quad (6)$$

The weights for each penalty term (w_s , w_{u1}, w_{u2}, w_e , w_{st}) are hyperparameters that must be carefully tuned according to the priority of the control objectives.

III. Deep Reinforcement Learning Algorithms

To select a suitable DRL algorithm for this problem with a continuous action space, we compare and analyze major off-policy actor-critic algorithms. Since high-fidelity battery simulations are computationally expensive, off-policy methods, which improve learning efficiency through data reuse, are overwhelmingly advantageous compared to on-policy methods (e.g., PPO).

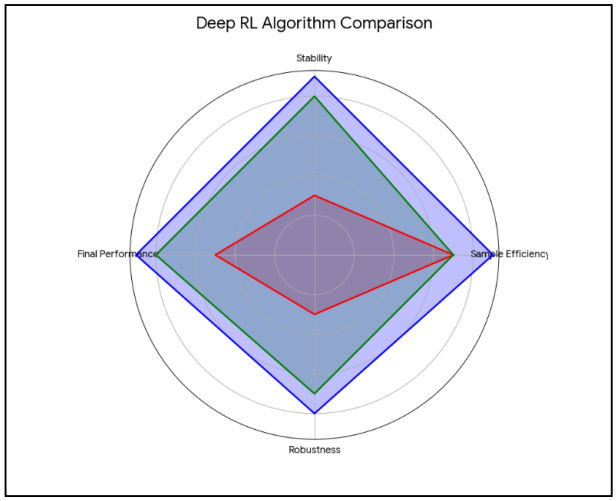

Fig. 2. Deep RL Algorithm Comparison.

A. DDPG (Deep Deterministic Policy Gradient)

An early actor-critic algorithm for continuous action spaces that learns a deterministic policy. However, it is often prone to unstable learning due to the overestimation bias of Q-values and sensitivity to hyperparameters [17].

B. TD3 (Twin-Delayed Deep Deterministic policy gradient)

An algorithm proposed to solve the overestimation bias problem of DDPG. It uses two independent critic networks to mitigate bias in Q-value estimation and delays policy updates to significantly improve learning stability. As it learns a deterministic policy, it is highly predictable and suitable for safety-critical systems [21].

C. SAC (Soft Actor-Critic)

An algorithm that aims to maximize not only the cumulative reward but also the entropy of the policy. Maximizing entropy encourages exploration by the agent, enabling more robust and efficient learning. Although it learns a stochastic policy, it generally exhibits high performance and stability [22].

In this study, the TD3 algorithm is adopted as the primary learning algorithm. TD3 offers improved stability and performance over DDPG. Furthermore, its deterministic policy is easier to predict and verify compared to the stochastic policy of SAC, making it more suitable for deployment in real industrial systems.

IV. Implementation and Verification Plan

A systematic 4-step roadmap is established for the development and deployment of the proposed reinforcement learning controller, progressively moving from simulation to actual hardware.

A. Step 1: High-Fidelity Virtual Environment Construction

Directly training on a real battery is impractical due to time, cost, and safety concerns. Therefore, building a high-fidelity simulation environment is a prerequisite. We will use MATLAB/Simulink's Simscape Battery and Simscape Fluids toolboxes to model the electro-thermal dynamics of the battery and the multi-channel cooling system. This Simulink environment will be interfaced with a Python reinforcement learning library (e.g., PyTorch) to serve as the training environment for the agent.

B. Step 2: Sim2Real Problem Solving and Virtual Verification

The policy trained in simulation may suffer performance degradation in the real world due to the "Sim2Real Gap". To address this, we apply the Domain Randomization technique. During simulation, physical parameters such as internal resistance, thermal conductivity, and sensor noise are randomly varied within a certain range during agent training. This forces the agent to learn a robust control policy that is resilient to various environmental changes [30].

C. Step 3: Hardware-in-the-Loop (HIL) Verification

The trained control policy is deployed on an actual controller (ECU) and interfaced with a virtual battery model running on a real-time simulator for HIL verification. This step allows for thorough, risk-free testing of the controller's real-time performance, communication delays, and fault-handling capabilities before interacting with physical hardware [27].

D. Step 4: Prototype Testing and Deployment

The controller that passes HIL verification is applied to a physical battery prototype for final performance evaluation. Data is collected under various real-world driving scenarios and charging/discharging conditions, and the policy is fine-tuned as necessary to prepare for final deployment.

V. Comparison of System Control Response Characteristics and Cumulative Power Consumption in Simulation

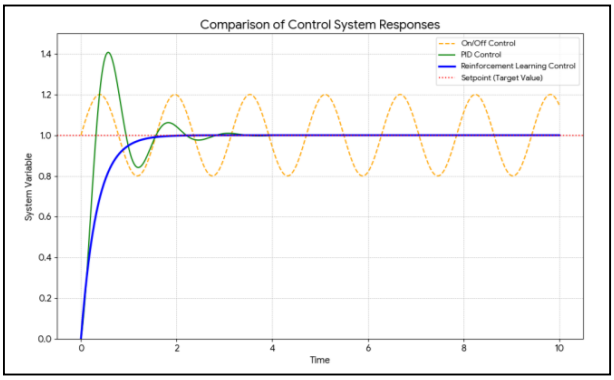

Fig. 3. Compariono of Control System Reponses.

This is a graph comparing the response characteristics of On/Off, PID, and Reinforcement Learning (RL) control methods. It illustrates how each control method regulates a system over time to reach a specific target value (Setpoint).

A. Legend

- Dotted Line (Setpoint): Target Value
- Orange Line : On/Off Control
- Green Line : PID Control
- Blue Line : Reinforcement Learning Control

B. Characteristics of Each Control Method

- On/Off Control: As the simplest control method, it causes the system variable to continuously fluctuate above and below the target value, a behavior known as oscillation. Precise control is difficult to achieve with this method.
- PID Control: A widely used method in industrial applications, the PID controller is much more stable and precise than On/Off control. It typically overshoots the target value initially before gradually stabilizing and converging at the setpoint.
- Reinforcement Learning (RL) Control: This method achieves the best performance among the three. It reaches the target value quickly and smoothly with no overshoot. This is because it learns the system's characteristics to find the optimal control strategy.

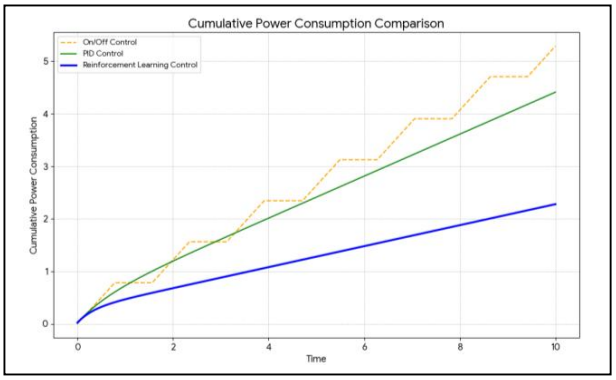

Fig. 4. Cumulative Power Consumption Comparison.

C. Analysis of Power Consumption by Control Method

- On/Off Control (Orange Line): This method continuously consumes unnecessary energy by simply switching the cooling system on and off. As seen in the graph, the cumulative consumption increases most steeply and results in the highest total, making it the least energy-efficient.
- PID Control (Green Line): Initially, this controller uses a relatively large amount of power to reach the target temperature, but power usage decreases after the system stabilizes. While much more efficient than On/Off control, it still requires continuous power to maintain a stable system.
- Reinforcement Learning Control (Blue Line): This method demonstrates the most optimized energy usage. It uses only the necessary amount of power to achieve the initial goal, and after reaching the target, it uses minimal power to maintain the state. The graph shows the slope becoming very gentle after the initial phase, ultimately achieving the best energy efficiency with the lowest total power consumption.

VI. Reinforcement Learning Simulation for Control Value Calculation of a Simple Model

TABLE I. SIMULATION CONDITION

Operating Voltage / Heater capacity
350V rated / 3.3kW PTC heater
PID Tunning values
(Before heating) Kp : 125, Ki : 15, Kb: 0 / Target temperature : 8℃
(After heating) Kp : 130, Ki : 40, Kb: 0 / Target temperature : 30℃
Climate Chamber set
Ambient Temperature : 40℃, Humidity : 60%
DRL temperature control range
29℃ ~ 31℃
BTMS AI control range
Compressor speed : 1500 ~ 4000 rpm EXV : 432 ~ 480 (max: 480) CFAN : 30% ~ 100% PUMP : 90%
Calculated data
Superheat, Subcool, Specific heat, Efficiency coefficient
Data collecting method
Operating Voltage / Heater capacity
Every step collecting by Python demon ➢ Calculated reward ➢ Save to web server DB ➢ Every step value show on Dashboard
Monitored data
Compressor, EXV, Ambient temperature, Humidity, High pressure, Low pressure, Refrigerant high/low temperature, Inlet/ outlet temperature, Liquid mass flow
Control data
Compressor, EXV
Current learning reward data and additional data
Current learning reward data : outlet temperature Additional data : Efficiency coefficient, Superheat, Subcool
Learning reward score
Every step regulated value reward measure (0~1) Setting reward data rated every step score will be max 1 Total reward data recording

TABLE II. TESTED DATASET FOR REINFORCEMENT LEARNING

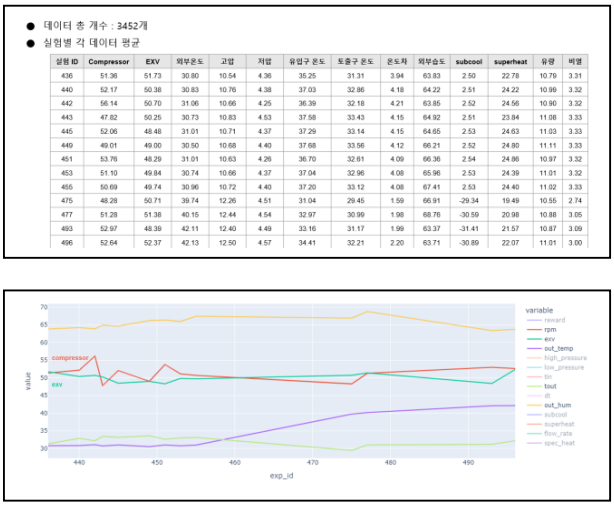

Fig. 5. Compressor, EXV, Outlet temperature average data.

In each experiment, the control values for the Compressor and EXV are uniformly distributed from 1% to 100%, causing the average control value to generally converge around 50%. Starting with Experiment ID 477, the external temperature was set to 40°C, which resulted in the observation of a slight upward trend in the discharge port temperature.

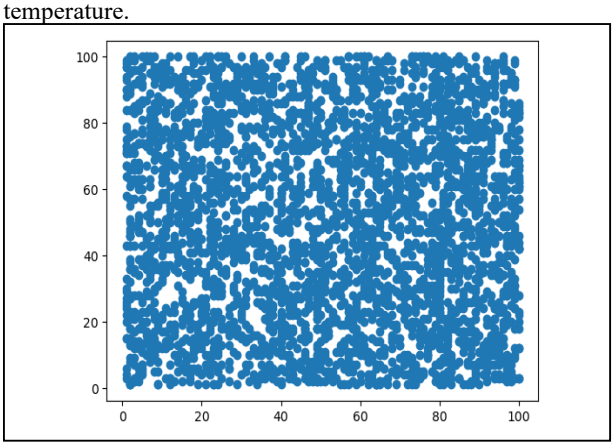

Fig. 6. Scatter Plot of Compressor and EXV Data from All Experiments.

The scatter plot of the Compressor and EXV data from the entire experiment shows that the combination of the two variables is evenly distributed across all ranges. This distribution characteristic indicates that data was collected under a wide variety of operating conditions. This provides a foundation for an in-depth analysis of the correlation and various data patterns between the Compressor and EXV.

[Comparison of Compressor and EXV for RL Simulation and PID]

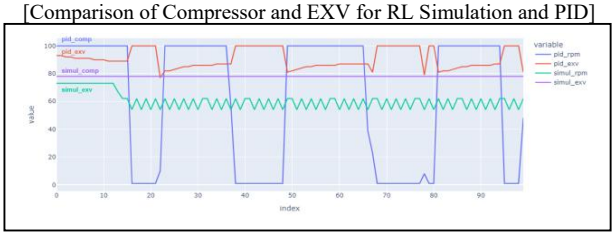

Fig. 7. BESS Liquid Cooling System schematic.

The simulation results confirm that the control of the Compressor and EXV is relatively more stable, with a smaller fluctuation range, compared to the PID method. This indicates that the control algorithm in the simulation environment operates in a more refined manner. However, this result serves as a comparative analysis between the simulation and the actual PID environment. Further quantitative comparison with experimental data under the current chamber conditions is necessary.

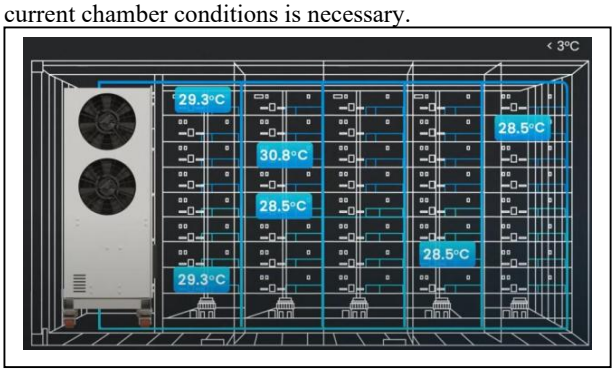

Fig. 8. Temperature difference within 3 degrees, BESS homogeneous cooling control.

CONCLUSION

This study proposed an intelligent thermal management strategy based on deep reinforcement learning to address the temperature homogeneity problem in multi-channel battery modules. To overcome the limitations of conventional control methods, the problem was formulated as an MDP, considering the spatial and nonlinear characteristics of the battery system. A multi-objective reward function encompassing safety, uniformity, efficiency, and stability was designed. TD3, an algorithm known for its stability and predictability, was chosen as the primary learning algorithm, and a systematic implementation roadmap from simulation to hardware was presented. The proposed reinforcement learning controller can actively adapt to changing load and environmental conditions in real-time to optimally distribute the coolant flow rate in each channel, thereby maximizing the temperature uniformity of the entire battery pack. This will directly contribute to preventing localized degradation and improving the overall lifespan and safety of the battery. Future research could include extending this work to Multi-Agent Reinforcement Learning (MARL) for controlling an entire pack system composed of multiple modules, and incorporating Explainable AI (XAI) techniques to ensure the reliability of the control policy. The data-driven, intelligent control approach presented in this study is expected to serve as an important foundation for the development of advanced thermal management technologies for next-generation electric vehicles and energy storage systems.

ACKNOWLEDGEMENT

This work was supported by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2022-II221200, Convergence security core talent training business(Chungnam National University))

REFERENCES

[1] T. M. Bandhauer, S. Garimella, and T. F. Fuller, "A critical review of thermal issues in lithium-ion batteries," Journal of The Electrochemical Society, vol. 158, no. 3, p. R1, 2011.

[2] K. A. Severson et al., "Data-driven prediction of battery cycle life before capacity degradation," Nature Energy, vol. 4, no. 5, pp. 383-391, 2019.

[3] C. Forgez, D. V. Do, G. Friedrich, M. Morcrette, and C. Delacourt, "Thermal modeling of a cylindrical LiFePO4/graphite lithium-ion battery," Journal of Power Sources, vol. 195, no. 9, pp. 2961-2968, 2010.

[4] Y. Ji, Y. Zhang, and C.-Y. Wang, "Li-ion cell operation at low temperatures," Journal of The Electrochemical Society, vol. 160, no. 4, p. A636, 2013.

[5] S. Panchal et al., "A review on the use of artificial neural networks for the thermal management of batteries," Journal of Power Sources, vol. 320, pp. 248-269, 2016.

[6] X. Feng et al., "The catastrophic failure of lithium-ion batteries: A review," Energy Storage Materials, vol. 10, pp. 246-267, 2018.

[7] T. D. Hatchard, D. D. MacNeil, A. A. Stevens, and J. R. Dahn, "Comparison of the thermal stability of various electrode materials in the presence of electrolyte," Journal of The Electrochemical Society, vol. 148, no. 7, p. A755, 2001.

[8] W. He et al., "State of health estimation of lithium-ion batteries: A review of the main methods," Journal of Power Sources, vol. 283, pp. 526-538, 2015.

[9] L. H. Saw, Y. Ye, and A. A. O. Tay, "Computational fluid dynamics and thermal analysis of Lithium-ion battery pack with air cooling," Journal of Power Sources, vol. 239, pp. 623-632, 2013.

[10] J. Kim, J. Oh, and H. Lee, "A comprehensive review of the battery thermal management system for electric vehicles," Applied Thermal Engineering, vol. 149, pp. 192-212, 2019.

[11] Z. Wei, B. Quan, and T. Zhao, "A review on reinforcement learning for building energy management," Energy and Buildings, vol. 229, p. 110500, 2020.

[12] R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction. MIT press, 2018.

[13] V. Mnih et al., "Human-level control through deep reinforcement learning," Nature, vol. 518, no. 7540, pp. 529-533, 2015.

[14] Y. Zhang, W. Wang, and Z. Chen, "A review on reinforcement learning-based approaches for building energy management," Renewable and Sustainable Energy Reviews, vol. 134, p. 110332, 2020.

[15] L. Lian, Y. Zhang, Z. Wang, and J. Zhang, "Deep reinforcement learning for joint optimization of battery charging and thermal management in electric vehicles," Energy, vol. 227, p. 120481, 2021.

[16] T. T. Nguyen, T. M. N. Nguyen, and J.-W. Choi, "Deep reinforcement learning-based thermal management for lithium-ion batteries in electric vehicles," IEEE Access, vol. 8, pp. 199343-199355, 2020.

[17] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," arXiv preprint arXiv:1509.02971, 2015.

[18] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a

stochastic actor," in International conference on machine learning, 2018, pp. 1861-1870.

[19] C. Liu, X. Cao, and P. X. Liu, "Reinforcement learning-based multiobjective optimization for building energy management," IEEE Transactions on Industrial Informatics, vol. 16, no. 6, pp. 3966-3976, 2019.

[20] F. Mocanu, E. Mocanu, P. H. Nguyen, A. Liotta, and M. Gibescu, "On-line building energy optimization using deep reinforcement learning," IEEE Transactions on Smart Grid, vol. 10, no. 4, pp. 3698-3708, 2018.

[21] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in International conference on machine learning, 2018, pp. 1587-1596.

[22] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor," in International Conference on Machine Learning, 2018.

[23] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347, 2017.

[24] A. I. OpenAI, "Solving rubik's cube with a robot hand," arXiv preprint arXiv:1910.07113, 2019.

[25] A. K. Tan, "Deep reinforcement learning for robotics: A survey," in 2018 IEEE International Conference on Robotics and Automation (ICRA), 2018, pp. 1-8.

[26] A. A. Pesaran, "Battery thermal management in EVs and HEVs: Issues and solutions," Battery Man, vol. 43, no. 5, pp. 34-49, 2001.

[27] MathWorks, "Simscape Battery," MathWorks Documentation, 2023.

[28] Y. Chen, S. E. Li, and K. Li, "Co-simulation of Simulink and Python for reinforcement learning-based control," in 2019 Chinese Control Conference (CCC), 2019, pp. 8366-8371.

[29] F. Sadeghi, M. F. Dehghan, and S. A. Gadsden, "A survey of sim-to-real transfer techniques for robotics," Annual Reviews in Control, vol. 52, pp. 473-490, 2021.

[30] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017, pp. 23-30.

[31] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)," IEEE Access, vol. 6, pp. 52138-52160, 2018.