Label-Efficient Battery SOH Estimation via Domain-Aware Self-Supervised Learning

Ji Young Yun [*1] **Haechang Kim** [*1] **Jong Min Lee** [1]

[*]Equal contribution [1]*Department of Chemical and Biological Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, 08826, Republic of Korea*. Correspondence to: Jong Min Lee jongmin@snu.ac.kr.

1. Introduction

State-of-health (SOH) estimation for lithium-ion batteries is essential for the reliable operation of electric vehicles and energy storage systems [1, 2]. While SOH serves as a critical indicator in battery management systems, obtaining accurate SOH labels requires labor-intensive long-term cycling tests, making large-scale labeled data difficult to acquire. Consequently, practical battery health diagnostics often face a label-scarce environment where only a small fraction of operational data is labeled. This challenge necessitates the development of label-efficient learning strategies that can leverage abundant unlabeled voltage, current, and temperature profiles.

Self-supervised learning (SSL) has emerged as a promising solution, pretraining models on unlabeled data to capture intrinsic degradation patterns before fine-tuning on downstream tasks with minimal supervision. However, many existing SSL approaches for battery data are limited to reconstructing local signal patterns within individual cycles, often overlooking the global and cumulative nature of battery aging [3, 4]. To address this, we propose a domain-aware SSL framework that learns cycle-level representations by modeling inter-cycle relationships. By integrating a hybrid CNN-Transformer architecture with a novel difference-based pretext task, the proposed method effectively aligns latent features along the degradation trajectory, ensuring robust SOH prediction even under severe label scarcity. This capability has the potential to accelerate the deployment of intelligent BMS in real-world applications where extensive cycling data are rarely available.

2. Methodology

2.1 Data Preparation and Preprocessing

This study utilizes the battery aging dataset from Severson et al., which contains voltage, current, and temperature profiles collected from 124 lithium-iron phosphate cells across hundreds of charge–discharge cycles [5]. To ensure high-quality feature extraction, raw signals are denoised using a discrete wavelet transform (DWT), which suppresses high-frequency noise while preserving degradation-relevant information. Each cycle sequence is subsequently resampled via cubic interpolation to a fixed length of $L = 256$ and standardized to have zero mean and unit variance to facilitate stable training.

2.2 Learning Pipeline and Encoder Architecture

The framework follows a two-stage pipeline: (1) self-supervised pretraining on unlabeled cycles and (2) fine-tuning for SOH estimation. We employ a hybrid encoder where 1D-CNN layers capture local temporal features of the charge–discharge profile, and a Transformer encoder models long-range dependencies within the sequence. To aggregate these temporal features into a single cycle-level embedding, we utilize an attentive pooling mechanism that selectively emphasizes segments of the profile sensitive to degradation.

2.3 Domain-Aware SSL with Clipped-Quantile Sampling

The core of our pretraining strategy is a difference-based relative learning objective. Unlike reconstruction-based tasks, the model is trained on cycle pairs to predict the degree of degradation between them. To incorporate domain knowledge, we employ an aging-aware pretext label, defined as the integral of $\log(1 + c)$ over the cycle interval $[c_e, c_l]$, where c denotes the cycle index, and c_e and c_l represent the earlier and later cycle indices of the pair, respectively. This label design reflects the accelerated capacity fade characteristic of lithium-ion batteries, providing a more physically consistent supervision signal than simple cycle-count intervals.

To improve training stability, we introduce Clipped-Quantile Sampling (CQS). CQS refines the pair generation process by excluding cycle pairs with negligible differences or excessively large gaps that may break degradation continuity. The remaining candidates are stratified into K quantile bins, ensuring that the model learns from a balanced distribution of both short- and long-term aging patterns.

3. Results

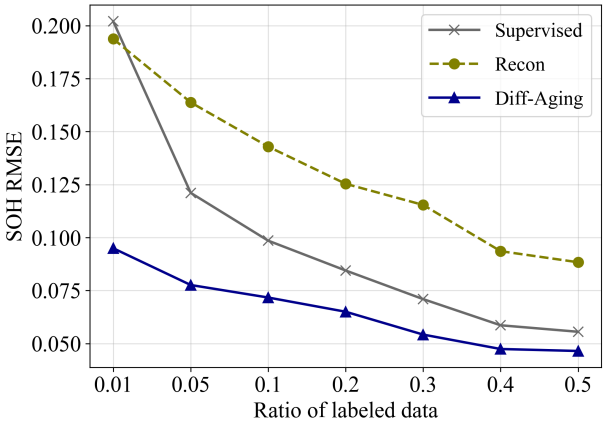

Fig. 1: SOH prediction performance across labeled data ratios.

Fig. 1 compares the downstream SOH prediction performance of the proposed SSL framework against supervised and reconstruction-based baselines. The proposed method consistently achieves the lowest root mean squared error (RMSE) across all label ratios. Notably, in the 1% label regime, it attains an RMSE of 0.0950, demonstrating superior label efficiency. This suggests that learning inter-cycle degradation differences is more effective for SOH estimation than conventional signal reconstruction

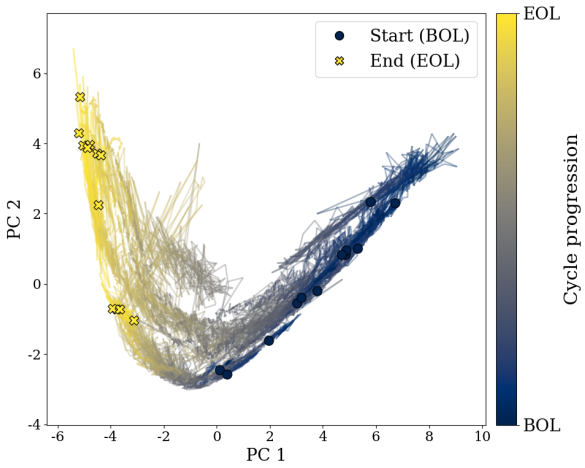

Fig. 2: PCA visualization of the latent space learned by the proposed method.

The structure of the learned latent space is visualized via Principal Component Analysis (PCA) in Fig. 2. The representations form a continuous and monotonic manifold from the beginning of life (BOL) to the end of life (EOL). This suggests that the pretraining objective successfully encodes the sequential progression of battery aging into the latent space, and the resulting separation of degradation stages is consistent with the improved downstream SOH regression performance.

To further assess physical interpretability, we analyze the attention weights of the pooling mechanism across aging stages. In the healthy phase near BOL, the model primarily attends to the voltage relaxation period, a stage reflecting electrochemical equilibrium after high-rate charging. As degradation accelerates beyond the knee-point, attention shifts toward the charging onset and the discharge-induced Ohmic drop, which are associated with increased internal resistance and polarization effects. These stage-specific transitions demonstrate that the model dynamically tracks key electrochemical aging indicators without explicit downstream supervision.

Fig. 3 shows the sensitivity of SOH estimation to the CQS bin count (K). While increasing K generally helps represent diverse aging scales, the performance peaks at $K = 12$. At this configuration, the SOH RMSE in the 1% label setting is reduced by 33.1% (from 0.1420 to 0.0950) compared to unconstrained sampling. However, a further increase to $K = 16$ results in diminished gains, likely due to overly fine-grained stratification, which reduces the number of samples per bin and destabilizes training.

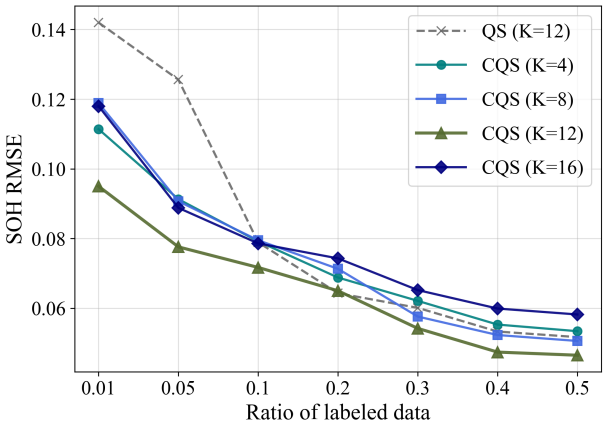

Fig. 3: Sensitivity analysis of CQS hyperparameters on downstream SOH estimation performance.

4. Conclusion

We presented a self-supervised learning framework for label-efficient SOH estimation that explicitly models degradation dynamics. By combining a CNN-Transformer encoder with a difference-based pretext task and CQS, the proposed method learns physically consistent representations from unlabeled operational data. Experimental results demonstrate consistent improvements over traditional baselines, particularly in data-scarce scenarios. This approach provides a practical foundation for reliable battery health monitoring in applications where labeled data are difficult to obtain.

References

[1] Mohammad Waseem, G Sree Lakshmi, Mumtaz Ahmad, and Mohd Suhaib. Energy storage technology and its impact in electric vehicle: Current progress and future outlook. *Next Energy*, 6:100202, 2025.

[2] FM Nizam Uddin Khan, Mohammad G Rasul, ASM Sayem, and Nirmal K Mandal. Design and optimization of lithium-ion battery as an efficient energy storage device for electric vehicles: A comprehensive review. *Journal of Energy Storage*, 71:108033, 2023.

[3] Yunhong Che, Yusheng Zheng, Xin Sui, and Remus Teodorescu. Boosting battery state of health estimation based on self-supervised learning. *Journal of Energy Chemistry*, 84:335–346, 2023.

[4] Tianyu Wang, Zhongjing Ma, Suli Zou, Zhan Chen, and Peng Wang. Lithium-ion battery state-of-health estimation: A self-supervised framework incorporating weak labels. *Applied Energy*, 355:122332, 2024.

[5] Kristen A Severson, Peter M Attia, Norman Jin, Nicholas Perkins, Benben Jiang, Zi Yang,

Michael H Chen, Muratahan Aykol, Patrick K Herring, Dimitrios Fraggedakis, et al. Data-driven prediction of battery cycle life before capacity degradation. *Nature Energy*, 4(5):383–391, 2019.