

Condition-Aware Progressive Knowledge Distillation: Lightweight and Accurate Prediction of Lithium-Ion Battery Remaining Useful Life

Penghua Li[✉], *Member, IEEE*, Ao Chen[✉], *Student Member, IEEE*, Yunhong Che[✉], *Member, IEEE*, Yangming Zhang, Jingjing Zhou, Xinmin Zhang[✉], *Member, IEEE*, Guodong Wang[✉], *Member, IEEE*, Xiaosong Hu[✉], *Fellow, IEEE*, and Schahram Dustdar[✉], *Fellow, IEEE*

Abstract—This study presents a condition-aware progressive knowledge distillation (CPKD) framework for developing lightweight models capable of accurately predicting the remaining useful life (RUL) of lithium-ion batteries. Within this framework, a teacher model is constructed by embedding a conditional variational autoencoder (CVAE) into a Mamba backbone. Through high-order compression and reconstruction of local-global battery features and the adoption of a dependency-aware loss function, the latent space is regularized to capture the underlying structure of degradation trajectories, enabling high-precision RUL prediction. Subsequently, a lightweight student predictor, realized as a CVAE enriched with operating-condition descriptors, is configured as the target model. Unsupervised

clustering is employed in conjunction with a dual-path gating mechanism to activate an operating-condition monitor. While a multichannel topology constructed from depthwise separable convolutions (DSCs) is triggered under the synergistic action of the gating weights to extract discriminative features across varying operating regimes and to deepen the representation of complex degradation dynamics. To transfer domain knowledge, a stepwise parameter-freezing knowledge distillation (KD) strategy is devised whereby accuracy-controlled degradation knowledge is progressively distilled from the teacher, ensuring effective RUL prediction by the compact student. Validation indicates that the distilled lightweight model achieves average root-mean-square errors of 0.0163 on the MIT dataset and 0.0384 on a real-world EV dataset, while using only 0.167M parameters and requiring 13.88 ms per inference on embedded hardware. The code and models are available at <https://github.com/Lipenghua-CQ/CPKD>

Index Terms—Lightweight predicting model, lithium-ion batteries, progressive knowledge distillation, remaining useful life.

I. INTRODUCTION

LITHIUM-ION batteries are extensively deployed across a wide range of industrial sectors, including energy storage, transportation, consumer electronics, and mechanical manufacturing. In electric vehicle (EV) applications, the performance of lithium-ion batteries under complex operating conditions, such as fluctuating temperature and humidity, thermal cycling, mechanical vibrations, and dynamic load profiles, plays a critical role in determining the vehicle's power output and overall energy efficiency [1]. Identifying how to effectively extract degradation-relevant features from externally collected data, such as voltage, current, and temperature, under real-world operating conditions, without relying on internal electrochemical variables like solid electrolyte interface (SEI) film formation or concentration polarization [2], is crucial to accurately predict performance metrics of lithium-ion batteries, particularly the remaining useful life (RUL). This challenge is fundamental to the safe operation and predictive maintenance of lithium-ion batteries and remains a critical focus of both academia and industry [3], [4].

Currently, industry-standard approaches for predicting the battery RUL predominantly rely on features extracted from

Received 6 September 2025; revised 22 December 2025; accepted 11 April 2026. Date of publication 4 June 2026; date of current version 22 July 2026. This work was supported in part by China's National Key Program Project under Grant 2025YFE0103000, in part by China's National Natural Science Foundation under Grant 52272388, in part by Chongqing Talent Program "Young Top-Notch Talent" Project under Grant cstc2024ycjh-bgzxm0042, and in part by the Open Research Project of the State Key Laboratory of Industrial Control Technology under Grant ICT2025B64. (Corresponding authors: Penghua Li; Xiaosong Hu; Schahram Dustdar.)

Penghua Li is with the College of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China, and also with the School of Computer Science and Engineering, Chongqing University of Science and Technology, Chongqing 401331, China (e-mail: liph@cqust.edu.cn).

Ao Chen is with the College of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: udocen@163.com).

Yunhong Che is with the Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: yunhche@mit.edu).

Yangming Zhang is with the Key Laboratory of System Simulation for Complex Systems, Chaoyang, Beijing 100020, China (e-mail: zhangyang198810@163.com).

Jingjing Zhou is with China Automotive Engineering Research Institute Company Ltd., Chongqing 401122, China (e-mail: zhoujingjing@caeri.com.cn).

Xinmin Zhang is with the State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou, Zhejiang 310027, China (e-mail: xinminzhang@zju.edu.cn).

Guodong Wang is with Coovally Computing S.L., 08018 Barcelona, Spain (e-mail: gwang@coovally.ai).

Xiaosong Hu is with the Department of Automotive Engineering, Chongqing University, Chongqing 400044, China (e-mail: xiaosonghu@ieee.org).

Schahram Dustdar is with the Distributed Systems Group, Vienna University of Technology, 1040 Vienna, Austria, and also with ICREA, Universitat Pompeu Fabra, 08018 Barcelona, Spain (e-mail: dustdar@dsg.tuwien.ac.at).

Digital Object Identifier 10.1109/TTE.2026.3685654

2332-7782 © 2026 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.

See <https://www.ieee.org/publications/rights/index.html> for more information.

Authorized licensed use limited to: TU Wien Bibliothek. Downloaded on September 07, 2026 at 13:20:12 UTC from IEEE Xplore. Restrictions apply.

capacity degradation and internal resistance evolution [2], implemented through model-based and data-driven methodologies [3]. Model-based approaches predict RUL by constructing mathematical or physical representations of battery degradation processes, which are generally categorized into empirical, equivalent circuit, and electrochemical models [5]. In empirical models, such as the single exponential model [6], double exponential model [7], and Verhulst model [8], battery capacity degradation is described as a mathematical function of cycle number. In equivalent circuit models (ECMs), RUL is treated as a latent state variable of the battery. Electrochemical models require the identification of a broader set of aging-related parameters. However, their application to RUL prediction is limited by inherent complexity and limited dynamic accuracy [9]. Despite the strong physical interpretability of model-based approaches, their practical deployment is often hindered by the need for precise parameter calibration and their limited adaptability to complex and variable degradation patterns observed under real-world operating conditions.

Data-driven approaches, particularly those based on deep neural networks, have gained widespread acceptance in engineering applications for health management of onboard lithium-ion batteries. These methods aim to directly learn degradation behaviors from large-scale operational data, without requiring explicit understanding of the complex electrochemical and multiphysics mechanisms underlying battery aging processes [10]. Within the context of RUL prediction, time-series models such as 1-D convolutional neural networks (1D-CNNs) and long short-term memory (LSTM) networks have played an essential role [1]. Zhang et al. [11] combined CNN and LSTM architectures, where CNNs extract localized degradation patterns and LSTMs capture long-term temporal dependencies to achieve accurate and robust RUL predictions. In another study, Gu et al. [12] proposed a lightweight multitask learning framework tailored for limited data, which employs an architecture combining 1D-CNN and bidirectional gated recurrent unit (Bi-GRU) to jointly predict the capacity turning knee and RUL from partial charging voltage segments. To integrate multimodal battery information and enhance physical interpretability, recent studies have further advanced CNN-based architectures [4], [13]. For instance, Li et al. [4] introduced a lightweight CNN-ASTLSTM model in which the CNN branch extracts spatial features from voltage and current profiles. Alongside these neural network-based methods, a complementary line of research has explored physics-informed hybrid frameworks that couple electrochemical mechanisms with deep networks. For example, Zhang et al. [14] proposed a physics-informed hybrid multitask learning (PIHMTL) framework, which couples the partial differential equations-based electrochemical model with a deep generative neural network. Building on this, their subsequent study [15] proposed a physics-informed dual neural network (PIDNN) that leverages initial discharge features and capacity difference sequences to capture cell heterogeneity and degradation evolution.

More recently, Gu and Dao [16] introduced the Mamba architecture, a structured state space sequence (S4) model that unifies the strengths of recurrent, convolutional, and continuous-time modeling paradigms. This architecture has also been applied to RUL prediction tasks [17], [18], due to its ability to capture long-range dependencies across arbitrary time series while maintaining computational efficiency throughout both training and inference. Huang et al. [17] developed a lithium-ion battery RUL prediction model, named RUL-Mamba, which integrates the Mamba module, feature attention network, and gated residual network within an encoder-decoder architecture. In another study, Li et al. [18] developed a physics-informed Mamba neural network that integrates electrochemical prior knowledge to enhance the modeling of lithium-ion battery degradation. By embedding differential equations derived from fundamental physical laws into the Mamba architecture, the approach captures dynamic, transient, and quasi-steady-state behaviors through circuit parameters including internal resistance, capacitance, and open-circuit voltage. Although Mamba-based architectures have demonstrated promising results in battery prognosis, their high parameter complexity and computational cost remain major barriers to deployment in resource-constrained embedded systems.

To enable the deployment of deep learning models with high accuracy and robust generalization on computationally constrained vehicle embedded platforms, existing studies typically pursue two routes. One is to directly design lightweight neural network architectures [4], [12]. The other is to compress complex, high-accuracy models to meet onboard edge deployment requirements. Compared with the former, model compression offers greater flexibility in engineering practice and relies less on detailed AI implementation expertise and prior knowledge of battery mechanisms. Common compression approaches include parameter quantization, weight pruning, and knowledge distillation (KD). Parameter quantization compresses model size by reducing the bit-width of weight representations, thereby significantly lowering memory requirements while preserving performance [19]. Weight pruning simplifies model architecture by removing redundant parameters after training, leading to substantial reductions in computation load [20]. Compared to these methods, KD allows a compact student model to learn output representations from a larger and more complex teacher model [21], [22], which not only reduces memory footprint but also maintains high inference efficiency. Xu et al. [23] proposed a contrastive adversarial KD method that integrates adversarial adaptation and contrastive loss to mitigate fine-grained feature loss and architectural mismatch, achieving an 8%–12% reduction in root-mean-square error (RMSE) on the FD001 subset of the C-MAPSS dataset. A multisource adversarial online KD framework [24] was developed to enhance RUL prediction of rolling bearings across heterogeneous machines. The model integrates a dual knowledge transfer strategy, combining multilevel domain adaptation with dynamically weighted online distillation, within an ensemble learning architecture.

Despite recent progress in KD for RUL prediction, challenges persist in optimizing model architecture and improving distillation efficiency in real-world applications.

- 1) Teacher models are typically high-precision architectures to ensure effective knowledge transfer. Existing RUL prediction models often improve accuracy by reconstructing battery degradation features [25] or modeling dependencies between local and global degradation behaviors [26]. However, simple reconstruction may fail to capture complex nonlinear degradation mechanisms, while overemphasis on local patterns can obscure global trends and increase overfitting risk.
- 2) Student models are often trained under the assumption of distributional consistency between training and test data. This assumption is frequently violated in practice [11], as battery data collected under varying operational conditions exhibit temporal variability and distributional shifts. Without a condition-aware design, student models lack generalizability.
- 3) Existing co-training frameworks distill knowledge by either reconstructing features or globally updating student parameters [23], [24], often at the expense of efficiency and hardware cost.

To address the outlined challenges, this work introduces a condition-aware progressive KD (CPKD) framework for efficient and accurate RUL prediction. The main contributions are as follows.

- 1) A conditional variational autoencoder (CVAE)-based teacher model, CVAE-Mamba, is proposed to reconstruct hierarchical degradation features. This model integrates a Mamba module with local self-attention to capture localized degradation trends from sequential signals such as quasi-voltage (QV) curve variance and incremental capacity (IC) curve area. A global self-attention refinement (GSAR) module dynamically adjusts feature weights by modeling interdependencies, enabling the capture of complex correlations. A multilayer perceptron (MLP)-based CVAE encodes and decodes local and global characteristics into latent variables, providing high-order approximation and compact feature representation.
- 2) A condition-monitored student model, CM-CVAE, is developed for RUL prediction, incorporating a depthwise separable feature extractor with dual-channel attention. An adaptive condition monitor, combining unsupervised clustering and dual-path gating, computes intercondition distances to extract shared degradation characteristics while filtering irrelevant patterns. These gating signals direct a multichannel architecture with depthwise separable convolutions (DSCs) to enhance feature perception under diverse operating conditions. A 1D-CNN-based CVAE coupled with a dual-attention prediction module enables deep degradation representation and context-aware modeling of aging dynamics.
- 3) A staged parameter-freezing KD (SPF-KD) strategy is proposed for effective model compression. This progressive distillation approach enhances transfer fidelity and

maintains prediction accuracy in the lightweight student model.

II. METHODS

The proposed CPKD framework, illustrated in Fig. 1, enables reliable prediction of the RUL of lithium-ion batteries under constrained conditions. It achieves this by feeding temporal features, such as the variance of Q - V curves and the integral area of IC curves [see Fig. 1(a)], into a teacher-student model architecture. The teacher model, CVAE-Mamba [see Fig. 1(b)], combines local and global self-attention mechanisms to capture degradation features and complex dependencies in the latent space. A reconstruction loss guides the model to extract intrinsic degradation patterns. The student model, CM-CVAE [see Fig. 1(c)], is lightweight and designed for efficiency, using unsupervised clustering and a dual-channel gating mechanism to recognize the operating state. A dynamically activated multichannel feature extractor further refines condition-sensitive degradation information. Based on this structure, a staged KD strategy, SPF-KD [see Fig. 1(d)], is proposed to gradually transfer degradation knowledge from the teacher model to the student model, leading to enhanced representational capacity and prediction accuracy while significantly reducing computational cost [see Fig. 1(e)].

A. CVAE-Mamba Teacher Model

The CVAE-Mamba teacher model comprises a locally attentive Mamba (LAM) module, a GSAR module, and two separate MLP-based CVAE feature reconstruction (MCFR) modules dedicated separately to degradation feature reconstruction and high-precision RUL prediction.

In the LAM module, which incorporates a localized self-attention mechanism, the input degradation sequence $X \in \mathbb{R}^{l \times d}$ is initially projected through an input embedding layer, followed by a hierarchical feature extraction process. It begins with a convolution-based local feature encoding operation f_{conv} , which captures spatially constrained patterns, and proceeds through a localized self-attention operation f_{atts} , enabling the extraction of early stage representations $H_{\text{atts}}^t \in \mathbb{R}^{l \times d_{\text{in}}}$. Here, l denotes the sequence length, d denotes the dimensionality of the input features, and d_{in} denotes the dimensionality of the embedded representations

$$\begin{aligned} H_{\text{atts}}^t &= f_{\text{atts}}(f_{\text{conv}}(X)) \\ &= W_{\text{atto1}}((\psi_e(X_t W_e + b_e)) W_c + b_c). \end{aligned} \quad (1)$$

Here, X_t denotes the input vector at the t th time step. W_e represents the parameter matrix of the embedding layer, while b_e corresponds to the bias vector of the embedding layer. The embedding layer's activation function is expressed as ψ_e . W_c signifies the parameter matrix of the 1D-CNN layer, and b_c denotes the bias term of the same layer. W_{atto1} indicates the parameter matrix of the self-attention layer.

To capture short- and long-term dependencies that evolve, the representations are subsequently processed by a state space model operator f_{ssm} , which produces temporally enriched features $H_{\text{ssm}}^t \in \mathbb{R}^{l \times d_{\text{in}}}$. To further enhance the contextual understanding of degradation behavior, a GSAR module is

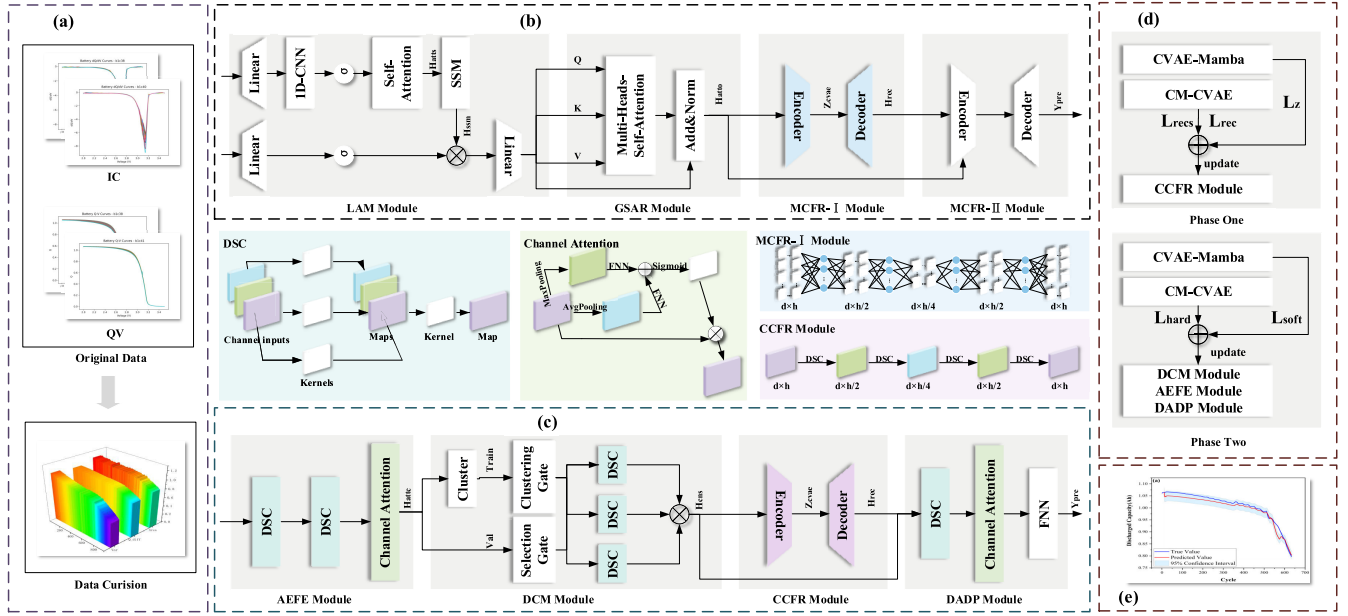


Fig. 1. Proposed CPKD framework for RUL prediction. (a)–(e) Input feature data module, the CVAE-Mamba teacher model, the CM-CVAE student model, the SPF-KD strategy, and the RUL prediction results, respectively.

introduced. This module applies a self-attention operator f_{atto} to the intermediate representation H_{ssm}^t , enabling the model to infer long-range interactions among local degradation features and capture their latent interdependencies with higher fidelity. The outcome is a globally informed representation $H_{atto}^t \in \mathbb{R}^{l \times d_{in}}$ that integrates both spatial locality and global coherence

$$\begin{aligned} H_{atto}^t &= f_{atto}(H_{ssm}^t) \\ &= \text{Concat} \left(\sigma \left(\frac{Q_{atto}^i K_{atto}^i}{\sqrt{d_h^i}} \right) V_{atto}^i \right) W_{atto}^o. \end{aligned} \quad (2)$$

Here, Q_{atto}^i , K_{atto}^i , and V_{atto}^i denote the query, key, and value matrices of the i th attention head obtained through linear transformations. The scaling factor $(d_h)^{1/2}$ is employed to prevent gradient vanishing. The variable h represents the number of attention heads, and σ corresponds to the softmax function. W_{atto}^o denotes the output projection matrix.

The refined features H_{atto}^t are then fed into an MCFR module, called the MCFR-I module, which is designed to reconstruct the degradation feature. The computational process of this module is defined as f_{cvae}^t . Through progressive encoding and decoding, this module recovers both localized patterns and global structure within the degradation signals, yielding a reconstructed output $\hat{H}_{rec}^t \in \mathbb{R}^{l \times d_{out}}$

$$\begin{aligned} \hat{H}_{rec}^t &= f_{cvae}^t(H_{atto}^t) \\ &= W_{dec}^t Z_{cvae}^t + b_{dec}^t \\ &= W_{dec}^t (W_{enc}^t ([H_{atto}^t \parallel H_{atts}^t]) + b_{enc}^t) + b_{dec}^t. \end{aligned} \quad (3)$$

The global degradation features H_{atto}^t and the local degradation features H_{atts}^t are concatenated and then compressed into a latent representation $Z_{cvae}^t \in \mathbb{R}^{l \times d/4}$. The encoder parameter matrix W_{enc}^t consist of three layers of linear weights:

$W_{e1}^t \in \mathbb{R}^{2d_{in} \times d_{in}}$, $W_{e2}^t \in \mathbb{R}^{d_{in} \times d_{in}/2}$, and $W_{e3}^t \in \mathbb{R}^{d_{in}/2 \times d_{in}/4}$. b_{enc}^t denotes the bias term of the encoder. The decoder parameter matrix W_{dec}^t also consist of three layers of linear weights: $W_{d1}^t \in \mathbb{R}^{d_{in}/4 \times d_{in}/2}$, $W_{d2}^t \in \mathbb{R}^{d_{in}/2 \times d_{in}}$, and $W_{d3}^t \in \mathbb{R}^{d_{in} \times d_{out}}$. b_{dec}^t denotes the bias term of the decoder. The nonlinear activation function used is still ReLU.

Subsequently, another MCFR module (MCFR-II module) trained for capacity prediction minimizes the discrepancy between the predicted and ground-truth capacities, producing a high-precision capacity prediction Y_{pre}^t from the teacher model. This operation is formalized as f_{pre}^t . Based on the operations of the four modules described above, the CVAE-Mamba teacher model can be formulated as follows:

$$Y_{pre}^t = f_{pre}^t (f_{cvae}^t (f_{atto}^t (f_{ssm}^t (f_{atts}^t (f_{conv}^t(X)))))). \quad (4)$$

B. CM-CVAE Student Model

The CM-CVAE student model comprises four key components designed to enable a deep understanding of the complex degradation processes in batteries. These components include an attention-enhanced feature extraction (AEFE) module, a degradation condition-monitored (DCM) module, a CNN-based CVAE feature reconstruction (CCFR) module, and a dual attention-driven prediction (DADP) module.

In the AEFE module, the input degradation features $X \in \mathbb{R}^{l \times d}$ are transposed and passed into the key feature extraction operation f_{sh} , yielding the key feature representation $H_{attc}^s \in \mathbb{R}^{d_{in} \times l}$.

The DCM module, driven by unsupervised clustering and a dual-path gating mechanism, receives the critical feature representation H_{attc}^s , which is modulated through the coordinated

operation f_{gate} of two gating units

$$G_i^j = \begin{cases} W_c & , j = c \\ \frac{\exp((W_s H_{attc}^s + b_s)_i)}{\sum_{n=1}^N \exp((W_s H_{attc}^s + b_s)_n)} & , j = s. \end{cases} \quad (5)$$

Here, G_i^j represents the weight of the i th channel, with $j \in \{c, s\}$ indicating parameters originating from distinct gating pathways. The matrix W_c denotes the clustering gating weights derived from K-means clustering, which regulate the participation level of different channels during the training process. The matrix W_s corresponds to the output-normalized gating weights that guide the selection of channel combinations during inference.

This modulated representation is then processed by the multichannel operator f_{cn} , constructed using DSCs, to generate the monitoring feature representation $H_{cn}^s \in \mathbb{R}^{d_{in} \times l}$

$$\begin{aligned} H_{cn}^s &= \sum_{i=1}^3 G_i^j f_{cn}(H_{attc}^s) \\ &= \sum_{i=1}^3 G_i^j (\sigma''((H_{attc}^s W_{DC,i}) W_{pc,i} + b_{cn})). \end{aligned} \quad (6)$$

Here, W_{DC} denotes the parameter matrix of the depthwise convolutional layer for the i th channel, and W_{pc} represents the parameter matrix of the 1-D pointwise convolutional layer for the i th channel. σ'' corresponds to layer normalization, with b_{cn} indicating the bias vector associated with the layer normalization.

The CCFR module, constructed with 1D-CNNs, strengthens the representational depth and contextual correlation of H_{cn}^s , producing the student reconstructed feature representation $\hat{H}_{rec}^s \in \mathbb{R}^{d_{out} \times l}$. The entire reconstruction process is represented as f_{cvae}^s

$$\begin{aligned} \hat{H}_{rec}^s &= f_{cvae}^s(H_{cn}^s) \\ &= W_{dec}^s Z_{cvae}^s + b_{dec}^s \\ &= W_{dec}^s (H_{cn}^s W_{enc}^s + b_{enc}^s) + b_{dec}^s. \end{aligned} \quad (7)$$

The feature H_{cn}^s is compressed and mapped to the student latent variable Z_{cvae}^s . The encoder parameters W_{enc}^s consist of three convolutional weight matrixes, specifically $W_{e1}^s \in \mathbb{R}^{d_{in} \times C \times 3}$, $W_{e2}^s \in \mathbb{R}^{d_{in}/2 \times d_{in} \times 3}$, and $W_{e3}^s \in \mathbb{R}^{d_{in}/4 \times d_{in}/2 \times 3}$. The encoder bias term is denoted by b_{enc}^s . The decoder parameters W_{dec}^s comprise three linear weight matrixes, namely $W_{d1}^s \in \mathbb{R}^{d_{in}/2 \times d_{in}/4 \times 3}$, $W_{d2}^s \in \mathbb{R}^{d_{in} \times d_{in}/2 \times 3}$, and $W_{d3}^s \in \mathbb{R}^{C \times d_{out} \times 3}$. The decoder bias term is represented by b_{dec}^s . All nonlinear activation functions employed are ReLU.

Subsequently, in the DADP module, the reconstructed representation $\hat{H}_{rec}^s$ is processed by the prediction operator f_{dp} to estimate the capacity, yielding the output Y_{pre}^s . Based on the operations of the four modules described above, the CM-CVAE student model can be formulated as follows:

$$Y_{pre}^s = f_{dp}(f_{cvae}^s(f_{cn}(f_{gate}(f_{sh}(X))))) \quad (8)$$

C. SPF-KD Strategy

To enable the student model to achieve the exceptional performance of the teacher model while maintaining a lightweight

structure, the proposed SPF-KD strategy leverages the teacher model's frozen parameters alongside its latent features, reconstructed features, and high-precision capacity predictions to guide a two-stage KD process. Specifically, the distillation process consists of feature distillation (FD) and KD. In FD, the intermediate representations of the student model are constrained to approximate, in the feature space, the latent states and reconstructed features of the teacher model that encode battery degradation information. KD acts on the output layer and drives the predictive distribution of the student model toward the target distribution produced by the teacher. Existing studies [23], [27], [28], [29] typically, adopt a sequential training scheme in which FD precedes KD, so that the student network first learns a representation space aligned with that of the teacher model and subsequently acquires its decision pattern in terms of both hard and soft labels, thereby improving the convergence of the distillation process and more fully exploiting the representational capacity of the student model. The effectiveness of this distillation order will be further validated in the ablation experiments presented below.

In the first distillation stage, the AEFE, DCM, and DADP modules of the student model are frozen, with only the parameters of the CCFR module participating in the paired learning between teacher and student. This freezing strategy is adopted because CCFR essentially acts as a regularization module whose primary role is to align the latent variables and reconstructed features of the student model with those of the teacher. It is therefore the key component in the FD stage. Through FD, the CCFR module learns a stable feature reconstruction prior constrained by the teacher under a fixed teacher feature distribution. In contrast, AEFE, DCM, and DADP modules together form the backbone of the student model, where AEFE and DCM modules jointly perform shared feature extraction and can be regarded as the main encoder, while the DADP module serves as the temporal prediction head.

Accordingly, this stage employs a joint loss function designed to compel the student model's CCFR module to learn the deep representation capability of the teacher model's MCFR module with respect to degradation features. Specifically, the joint loss $\mathcal{L}_{stage1}$ consists of three components: the latent feature alignment loss $\mathcal{L}_z$ between CCFR and MCFR-I module, the reconstruction feature discrepancy loss $\mathcal{L}_{rec}$ between CCFR and MCFR-I module, and the self-supervised loss $\mathcal{L}_{rec}$ representing the difference between the MCFR module's input features before and after reconstruction

$$\mathcal{L}_{stage1} = \mathcal{L}_z + \mathcal{L}_{rec} + \mathcal{L}_{rec}^s. \quad (9)$$

The latent feature alignment loss $\mathcal{L}_z$ is defined as the mean squared error between the teacher's latent representation Z_{cvae}^t and the student's latent representation Z_{cvae}^s

$$\mathcal{L}_z = \frac{1}{N} \sum_{i=1}^N \|Z_{cvae}^t - Z_{cvae}^s\|_2^2. \quad (10)$$

The reconstruction feature loss $\mathcal{L}_{rec}$ is expressed in the form of a perceptual loss, computed as the $L1$ distance between the

teacher's reconstructed features $\hat{H}_{rec}^t$ and those of the student $\hat{H}_{rec}^s$ across multiple intermediate layers

$$\mathcal{L}_{rec} = \sum \|\phi_l^t(\hat{H}_{rec}^t) - \phi_l^s(\hat{H}_{rec}^s)\|_1. \quad (11)$$

The self-supervised loss $\mathcal{L}_{rec}^s$ employs the $L1$ norm to quantify the elementwise discrepancy between the original input and the output reconstructed by the student's encoder-decoder pipeline, which reinforces the student's ability to preserve fine-grained input details

$$\mathcal{L}_{rec}^s = \|x - \hat{H}_{rec}^s\|_1. \quad (12)$$

In the second distillation stage, the CCFR module of the student model is frozen, while the remaining modules participate in the teacher-student paired learning process. The joint loss $\mathcal{L}_{stage2}$ in this stage consists of two components. The first is a hard label loss $\mathcal{L}_{hard}$, computed as the difference between the output of the student's DADP module and the corresponding ground-truth values. The second is a soft label loss $\mathcal{L}_{soft}$, which measures the discrepancy between the capacity predictions generated by the teacher's MCFR-II module and those produced by the student's DADP module

$$\mathcal{L}_{stage2} = \alpha \mathcal{L}_{hard} + (1 - \alpha) \mathcal{L}_{soft}. \quad (13)$$

The hard label loss $\mathcal{L}_{hard}$ employs label-smoothed cross-entropy loss to compute the distance between the ground-truth capacities and the predicted capacities in the student model

$$\mathcal{L}_{hard} = \frac{1}{N} \sum_{i=1}^N \|Y_i - Y_{pre_i}^s\|_2^2. \quad (14)$$

The soft target loss $\mathcal{L}_{soft}$ utilizes the Kullback-Leibler (KL) divergence to measure the distributional difference between the teacher and student capacity predictions, where τ denotes the distillation temperature

$$\mathcal{L}_{soft} = \tau^2 D_{KL}(Y_{pre}^t \| Y_{pre}^s). \quad (15)$$

III. DATA EXPERIMENT

A. Data Curation

To facilitate independent replication of our numerical experiments with the publicly released source code, we employ the MIT open battery dataset [30], which comprises 124 APR18650M1A lithium iron phosphate batteries. All batteries are cycled in a 30 °C environmental chamber under a protocol that combines fast charging and constant current/constant voltage (CC-CV) charging followed by constant current discharging. Specifically, each cycle begins with one- or two-step fast charging to 80% SOC, continues with CC-CV charging at 1 C to full charge, and ends with discharging at 4 A down to a voltage of 2 V. The dataset provides time series records of voltage, current, and temperature, together with the final cycle life labels for each battery.

We select 30 batteries (serial numbers #10–22, 25–40, 42), whose capacity, voltage, and degradation trajectories form a representative benchmark. This subset spans a broad range of cycle counts and degradation levels while avoiding systematic bias that may arise from adjacent serial numbers. Therefore, batteries #23, #24, and #41, etc., are omitted. Test batteries

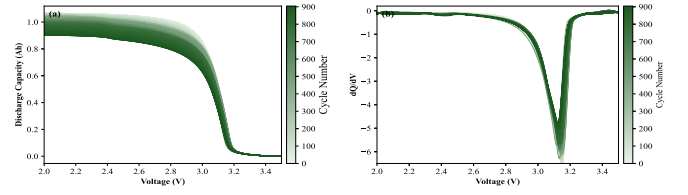


Fig. 2. Original samples from the MIT dataset that serve as the basis for computing the input features of the proposed CPKD framework. (a) and (b) QV curve and the IC curve of the battery #10, respectively.

(numbers #43–45 and #6–8) have no overlap with the training set, which eliminates data leakage. The small yet diverse test set also probes the model's stability under limited-data conditions.

We construct an input feature set that includes the variance of the Q - V curve and the area under the IC curve. Prior studies have shown that these descriptors correlate strongly with battery aging [31]. For example, deformation of the Q - V profile [see Fig. 2(a)] tracks the growth of internal resistance and the loss of electrochemical activity, so its variance rises as the battery ages. Shrinkage of the IC-curve area [see Fig. 2(b)] directly signals reduced active material and poorer reaction kinetics. Because these features encapsulate electrochemical, material, and structural degradation mechanisms, they provide both explanatory power and predictive value for end-of-life (EOL) prediction.

In addition, to evaluate the generalization of the proposed CPKD framework across battery systems, we construct a supplementary dataset (see Table I). Specifically, we randomly sample real-world operating data from seven EVs collected via the new-energy vehicle health-management platform of China Automotive Engineering Research Institute Company Ltd. (CAERI). The EV data upload follows the Chinese national standard GB/T 32960, with the time interval between consecutive messages no greater than 30 s. All seven EVs are equipped with ternary lithium-ion battery packs. The battery packs of the EVs #1075, #1096, #1106, and #1099 adopt a 3P90S configuration (three cells in parallel per group and 90 groups in series), with a rated voltage of 329 V, a nominal energy of 49 kWh, and a rated capacity of 160 Ah. The battery packs of the EVs #1095, #1098, and #1108 adopt a 3P92S configuration, with a rated voltage of 338 V, a nominal energy of 59 kWh, and a rated capacity of 170 Ah. For each EV, the raw data include 21 operating parameters (e.g., pack voltage, pack current, and temperature). To meet privacy protection and commercial confidentiality requirements and to remove fields unrelated to this task (e.g., latitude and longitude, vehicle brand, fault status, and sensor-probe position identifiers), we retain 13 parameters for modeling. The statistical characteristics of the retained parameters are summarized in Table I.

The online operational data from 7 EVs span up to 523 days and comprise 8 256 516 records (see Fig. 3). As the true capacity cannot be directly obtained under natural driving conditions, we perform offline capacity calibration for these EVs using CAERI's integrated charge, data acquisition, and inspection (ICDI) equipment developed in-house (see Fig. 4). After CAERI signs a testing cooperation agreement with

TABLE I
KEY PARAMETERS OF REAL-WORLD VEHICLE OPERATING DATA

Data fields	Unit	Character type	Valid value range	Ground truth range	Mean and standard deviation	Description
vehicle_status	-	int	1,2,3	1,2,3	-	1:Engine start 2:Engine off 3:Other states
charge_status	-	int	1,2,3,4	1,2,3,4	-	1:Parking charging 2:Driving charging 3:Not charging 4:Charging completed
speed	km/h	float	[0,220]	[0.0,147.5]	21.71±24.81	-
sum_mile_age	km	float	[0,999999.9]	[7712.0,155419.9]	65739.71±38812.10	-
sum_voltage	V	float	[0,1000]	[303.4,392.3]	354.51±16.88	-
sum_current	A	float	[-1000,1000]	[0.1,355.9]	19.53 ± 29.10	Discharge current
				[-157.6,-0.1]	-43.18 ± 39.72	Charging current
soc	-	int	[0,100]	[2,100]	67.99±20.96	The range from 1% to 100%
insulation_resistance	kΩ	int	[0,60000]	[12,1000]	999.70±13.26	-
max_cell_volt	mV	float	[0,15000]	[3378,4271]	3891.80±185.28	Maximum cell voltage
min_cell_volt	mV	float	[0,15000]	[3341.00,4259]	3875.64±182.59	Minimum cell voltage
max_temp	°C	int	[-40,210]	[3,42]	27.53±5.49	Minimum battery pack temperature
min_temp	°C	int	[-40,210]	[2,39]	23.81±5.20	Maximum battery pack temperature
test_capacity	Ah	float	[0,170]	[125.54,155.69]	139.76±8.07	-

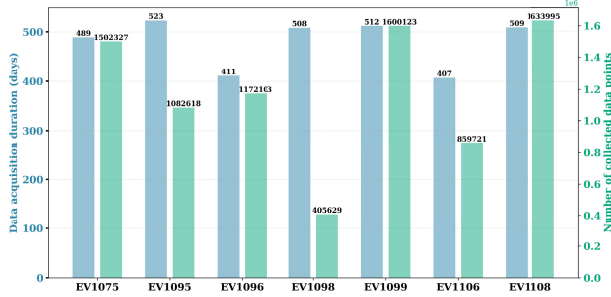


Fig. 3. Time span and record count of real-world vehicle operating data.



Fig. 4. Experimental setup for capacity ground-truth calibration. (a) Discharge test environment chamber and (b) charge test environment chamber.

the vehicle owner, each vehicle is transferred to the test facility when the SOC decreases to approximately 20%, where standardized charge and discharge tests are conducted under strictly controlled conditions.

Discharge tests are conducted in the CAERI discharge laboratory [see Fig. 4(a)]. The facility integrates a chassis dynamometer and a high-precision coulomb metering system, enabling accurate EV discharge measurement under static and dynamic conditions in accordance with the Chinese national standard GB/T 18386.1. The EVs are conditioned at

25 °C ± 2 °C and 50% ± 10% relative humidity. Driving load is reproduced on the dynamometer, and the battery pack is discharged continuously in constant current or dynamic load mode to SOC ≈ 0%. The test ends when the EVs could no longer sustain 70 km/h or the instrument cluster indicates or alarms SOC = 0%. Charge tests are conducted in a standard environmental chamber in the charging laboratory [see Fig. 4(b)]. Active thermal control and air circulation suppress external disturbances and match the discharge test environment. During charging, the ICDI equipment samples temperature, voltage, current, and SOC at a sampling rate of at least 10 Hz. The protocol comprises three stages. The batteries are first charged to SOC ≈ 10% for initial parameter identification and online update of charging constraints, then charged at a 0.5 C in diagnostic mode for 70 min, and finally charged to full. After a complete charge and discharge calibration, the ICDI equipment computes the ground-truth capacity using coulomb counting via ampere-hour integration.

Note that the capacity calibration is voluntary. Therefore, records with capacity ground truth are uneven across vehicles. For example, the 1098# EV provides markedly fewer ground-truth samples than the others, indicating data sparsity (see Fig. 3). In addition, during real-world operation, charge and discharge counts and depths are strongly stochastic. As a result, our data preprocessing differs from that used for the MIT dataset. We do not segment battery sequences by charge and discharge cycles, nor compute IC and QV features using fixed time windows. Instead, we segment the data according to changes in capacity ground truth and compute IC and QV features indexed by driving mileage. To align with the MIT feature construction protocol, IC and QV features are generated using only total pack voltage, total pack current, and driving mileage.

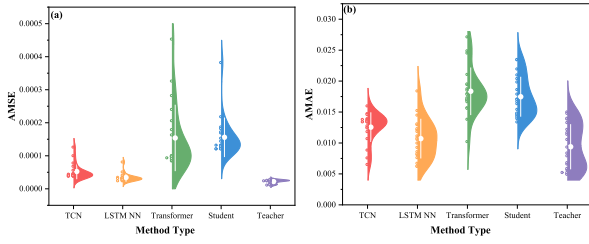


Fig. 5. Distribution of prediction performance for the SOTA models and proposed teacher–student models over 20 independent training runs. (a) and (b) Half-violin plots of the AMSE and AMAE for each model, respectively.

B. Benchmark Model Selection

In the evaluations, we compare the proposed KD framework against state-of-the-art (SOTA) predictors, namely the temporal convolutional network (TCN) [32], the LSTM NN [33], and the Transformer [34]. This analysis verifies both the high predictive accuracy of the teacher model and the learnability of the student model. Consistent with prevailing practice, mean square error (mse) is adopted as the training loss function, whereas RMSE and mean absolute error (MAE) are employed as evaluation metrics.

To ensure a fair comparison, every model is trained 20 times with distinct hyperparameter settings. The training data comprised the variance of the Q – V curve and the area under the IC curve for 30 batteries. After training, we average the mse and the MAE obtained for the 30 batteries, then carry out statistical analysis. Fig. 5 presents the resulting averaged mean squared error (AMSE) and averaged mean absolute error (AMAE) as half-violin plots. In each half-violin, the left panel shows individual error points. Their vertical positions represent the true error values, while a small horizontal jitter is added only to prevent overlap. The right panel depicts the probability density of the same errors, where vertical extent indicates the overall range and horizontal width reflects the degree of concentration. A simplified box inside the density marks the median together with the upper and lower quartiles.

The TCN in Fig. 5(a) illustrates the visualization. Under different hyperparameter configurations, the AMSE ranges from $3.81\text{E-}5$ to $1.26\text{E-}4$ (right-side distribution length) and is concentrated between $3.81\text{E-}5$ and $4.63\text{E-}5$ (right-side distribution width). The median is $4.21\text{E-}5$, and the first and third quartiles are $3.92\text{E-}5$ and $4.90\text{E-}5$. The 15 of the 20 points lie near $4.21\text{E-}5$, which matches the peak density on the right. Fig. 5(b) shows that the TCN MAE spans $6.52\text{E-}3$ to $1.60\text{E-}2$, concentrates between $1.24\text{E-}2$ and $1.46\text{E-}2$, has a median of $1.34\text{E-}2$, and quartiles of $1.24\text{E-}2$ and $1.46\text{E-}2$. Again, 15 points cluster near $1.36\text{E-}2$, confirming adequate training. Across all models, the error distributions are approximately normal. The teacher and TCN reach comparable maximum AMAEs, while the TCN and LSTM NN share similar maxima in AMSE, and the Transformer and student share another. These similarities indicate that the comparison is balanced. The teacher model exhibits the lowest medians in both AMSE and AMAE, confirming its superior predictive accuracy.

Based on the independent experimental analyses, the hyperparameter configuration that yielded the lowest prediction error

TABLE II
OPTIMAL HYPERPARAMETERS FOR SOTA AND PROPOSED MODELS

Methods	Hyperparameters						
	HLN	HL	LR	AH	CKS	CWMC	SSD
TCN[32]	64	8	0.009140	8	3	-	-
LSTM NN[33]	64	4	0.009320	-	-	-	-
Transformer[34]	64	12	0.001670	4	-	-	-
Student Model	32	12	0.000001	-	3	3	-
Teacher Model	32	19	0.001000	8	3	-	16

was selected as the optimal setting for each baseline model. The considered parameters include the input channels (IC), output channels (OC), input sliding window size (ISWS), hidden layer neurons (HLN), hidden layers (HL), learning rate (LR), attention heads (AH), convolutional kernel size (CKS), channels for working condition monitoring (CWMC), state space dimension (SSD), batch size (BS), epochs (EP), and optimizers (OP). The IC, OC, ISWS, BS, EP, and optimizer for each baseline model are set to 3, 1, 10, 32, 20, and Adam, respectively. All other hyperparameters are summarized in Table II.

C. Knowledge Distillation

After selecting the benchmark teacher–student pair, the proposed SPF-KD strategy is applied to distill the degradation knowledge from the teacher into the student, enabling accurate prediction of the RUL. To validate its effectiveness, four SOTA distillation methods, including S-KD [27], L1-KD [28], L2-KD [29], and KDNet [23], are implemented for comparison. S-KD transfers class-level probability distributions. L1-KD and L2-KD leverage intermediate feature representations, minimizing the L1 and L2 norms of teacher and student feature discrepancies, respectively. KDNet integrates soft labels, hard labels, and additional knowledge via specialized distillation modules. By contrast, SPF-KD progressively fuses soft labels, hard labels, latent feature maps, and reconstructed feature maps.

During the distillation process, two hyperparameters are tuned: the weight α in (13), which balances hard and soft losses, and the temperature τ in (15), which smooths the teacher’s output distribution (larger τ yields a flatter distribution). Guided by prior studies, α is explored in the set $\{0.1, 0.3, 0.5, 0.7\}$ and $\tau \in \{1, 3, 5, 7\}$. A grid search over all 16 combinations is executed on 30 independent battery-degradation training subsets. The AMSEs obtained from these experiments are visualized as scatter plots with color mapping in Fig. 6.

For example, in Fig. 6(a), the AMSE for S-KD ranges from brown ($9.21\text{E-}5$) to purple ($1.39\text{E-}4$). Across Fig. 6(a)–(e), the minimum AMSEs for S-KD, L1-KD, L2-KD, KDNet, and SPF-KD are $9.21\text{E-}5$, $9.56\text{E-}5$, $8.93\text{E-}5$, $8.82\text{E-}5$, and $7.80\text{E-}5$, respectively. Among the evaluated methods, S-KD, L1-KD, and L2-KD achieve their lowest AMSE at $\tau = 3$ and $\alpha = 0.5$. KDNet attains its minimum AMSE when $\tau = 5$ and $\alpha = 0.5$, whereas SPF-KD yields optimal performance at $\tau = 3$ and $\alpha = 0.3$. The maximum AMSE for the first three methods is approximately $1.35\text{E-}4$, while KDNet and SPF-KD

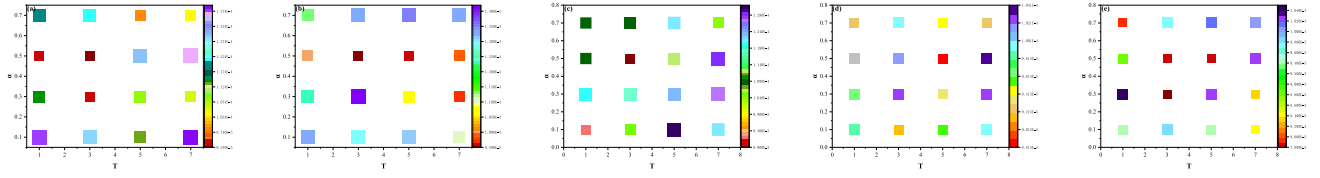


Fig. 6. Performance of the five KD methods under various hyperparameter settings of τ and α . (a)–(e) AMSE distributions for S-KD, L1-KD, L2-KD, KDNet, and SPF-KD, respectively.

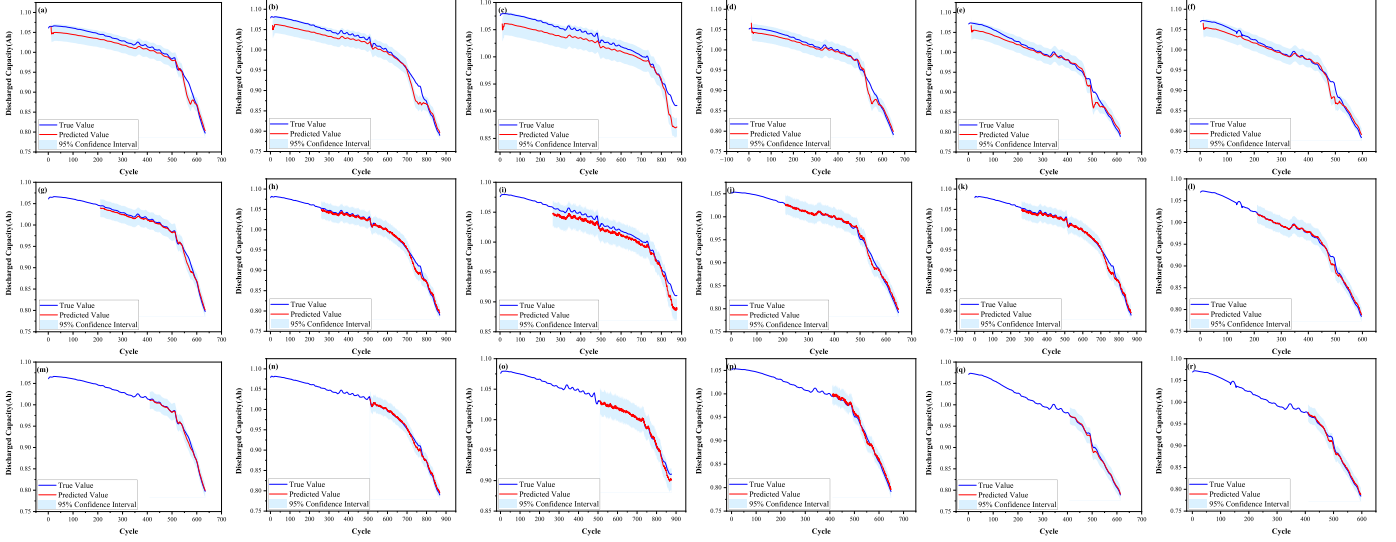


Fig. 7. Capacity prediction trajectories for six batteries (#6, #7, #8, #43, #44, #45) obtained using the student model are presented with predictions initiated at three stages. (a)–(f) Predictions starting from cycle 11. (g)–(l) Predictions initiated at the mid-life stage. (m)–(r) Predictions beginning at the late-life stage.

do not exceed $1.05\text{E-}4$, underscoring the benefit of feature-level distillation. Notably, overlapping AMSE ranges among different hyperparameter settings confirm that the comparisons are fair. The hyperparameter pair giving the lowest error for each method is adopted as its optimal configuration.

D. Model Performance Analysis

To validate the proposed framework, we evaluate the recursive-prediction performance of the selected SOTA neural networks and KD methods on six batteries (#6–8 and #43–45) with a sliding window of ten cycles. Given that the RUL is defined as the available time or number of cycles left before a battery's performance declines to its EOL threshold, typically identified as 80% of the rated capacity, the performance of RUL prediction ultimately hinges on the accuracy of capacity prediction. Building upon the conventional RUL definition (i.e., $RUL = T_{EOL} - T_{current}$), we adopt a more intuitive formulation, as proposed in [35], to facilitate interpretability

$$RUL = Q_{current}(c) - Q_{EOL}(c). \quad (16)$$

Here, $Q_{current}(c)$ denotes the actual capacity of the battery at the c th cycle, and $Q_{EOL}(c)$ represents the predefined capacity degradation threshold. Accordingly, we employ RMSE and mean absolute error MAE as evaluation metrics to assess the effectiveness of different prediction methods.

The capacity-fade trajectories predicted by the student model distilled through the SPF-KD strategy are presented

for all six test batteries in Fig. 7(a)–(r). In the most data-scarce scenario, with predictions initiated at cycle 11 [see Fig. 7(a)–(f)], the student model effectively captures degradation trends and achieves reasonable accuracy. Due to the recursive sliding-window prediction scheme, cumulative errors induce deviations beyond the 95% confidence interval at later stages. However, the condition-monitoring module coupled with SPF-KD constrains these deviations, realigning the true capacity within the confidence bounds, as validated by subsequent ablation studies. Predictive accuracy improves notably with increased historical data. Fig. 7(g)–(r) correspond to predictions initiated at one-third and two-thirds of the batteries' lifetimes, respectively. For batteries #43–45 and #6, prediction windows start at cycles 200 and 400, while for batteries #7 and #8, they start at cycles 250 and 500. Apart from battery #7 in Fig. 7(h), all measured capacities remain within the 95% confidence intervals, and by cycle 400 [see Fig. 7(n)] battery #7 is also included.

To thoroughly evaluate the performance of SOTA models, teacher models, and student models trained with various KD strategies at different prediction start points, Fig. 8 displays radar plots of RMSE and MAE distributions across six battery datasets. Notably, the results of the TCN, LSTM NN, and Transformer are included to demonstrate the feasibility of the teacher model and the baseline student model. The performance of student models derived with S-KD, L1-KD, L2-KD, and KDNet serves to verify the effectiveness of the proposed SPF-KD strategy.

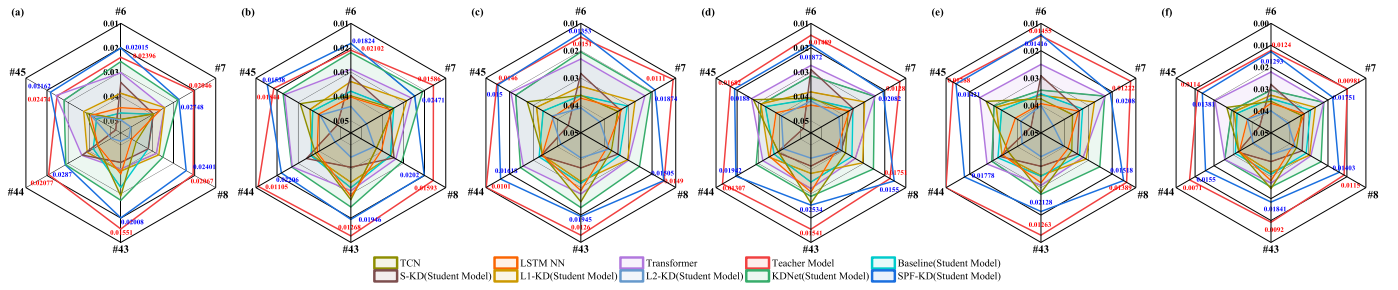


Fig. 8. Radar plot comparison of RMSE and MAE distributions for SOTA, teacher, and distilled student models at different prediction start points. (a)–(c) RMSE distributions and (d)–(f) MAE distributions, respectively.

TABLE III
RMSES OF DIFFERENT MODELS ON MIT DATASET

Methods	Battery						ARMSE
	#6	#7	#8	#43	#44	#45	
TCN[32]	0.0363	0.0344	0.0371	0.0248	0.0339	0.0274	0.0323
LSTM NN[33]	0.0374	0.0328	0.0390	0.0288	0.0338	0.0384	0.0350
Transformer[34]	0.0231	0.0230	0.0299	0.0277	0.0257	0.0206	0.0250
Teacher Model	0.0151	0.0111	0.0149	0.0126	0.0096	0.0146	0.0130
Candidate benchmark	0.0364	0.0304	0.0336	0.0315	0.0324	0.0341	0.0331
S-Model							
S-KD[27]	0.0283	0.0334	0.0344	0.0371	0.0331	0.0447	0.0352
L1-KD[28]	0.0330	0.0272	0.0303	0.0322	0.0337	0.0313	0.0313
L2-KD[29]	0.0409	0.0420	0.0382	0.0408	0.0416	0.0378	0.0402
KDNet[23]	0.0305	0.0303	0.0253	0.0287	0.0246	0.0256	0.0275
SPF-KD	0.0135	0.0187	0.0150	0.0194	0.0162	0.0150	0.0163

Each axis in the radar plots corresponds to one battery, with the scale for RMSE or MAE decreasing uniformly by 0.01 from the center outward. Each colored polyline represents the distribution of RMSE or MAE across the six datasets for a specific model. A line closer to the periphery indicates lower error and thus better performance. For instance, in Fig. 8(a), the teacher model exhibits the lowest RMSE on most datasets, with its line residing near the outermost region except for battery #6 and battery #45, indicating superior overall performance. It also shows that the student model trained with SPF-KD achieves the second-best performance in aggregate. Importantly, as shown in Fig. 8(a)–(c), the polylines for all compared models tend to shift outward as the prediction start point moves further along the cycling trajectory, demonstrating a consistent trend of decreasing RMSE with later prediction initiation. This trend aligns with the error evolution observed in Fig. 7(g)–(r). Particularly, the student model trained with SPF-KD attains performance comparable to the teacher model on battery #8 and battery #45, and even outperforms it on battery #6. Similar conclusions can be drawn from the MAE distributions shown in Fig. 8(d)–(f).

To provide a more detailed assessment of the predictive accuracy of each model, the specific values of RMSE corresponding to the radar plots in Fig. 8(c) and (f) are presented in Table III. The teacher model consistently achieves lower RMSE and MAE than TCN, LSTM, and Transformer across all six batteries, demonstrating its superior predictive accuracy. The baseline student model yields an ARMSE of $3.31\text{E-}2$. Although this error is higher than those of the three reference SOTA prediction models, both metrics remain within the 10^{-2} order of magnitude, indicating that the baseline student model possesses adequate learning capacity. Furthermore, the student

model distilled using the proposed SPF-KD strategy achieves lower ARMSE than all comparison models, including TCN, LSTM, Transformer, and student models trained with S-KD, L1-KD, L2-KD, and KDNet. Notably, this student model achieves an ARMSE of 0.0163, corresponding to 99.66% of the teacher model's ARMSE. These results confirm that the proposed KD strategy enables the student model to effectively absorb accuracy-aware degradation knowledge from the teacher model, leading to high-fidelity prediction performance.

To further assess the generalization capability of the CPKD framework across different battery chemistries, we performed supplementary cross-domain validation using operating data from traction batteries in real vehicles. Note that the student model based on neural networks inherently relies on the training data to learn the mapping between inputs and outputs, and its generalization performance is highly dependent on the operating conditions and data distribution represented there. Due to pronounced differences in battery type, load characteristics, temperature, and usage or aging patterns, the MIT dataset collected under tightly controlled laboratory conditions shows a substantial distribution shift from the real vehicle operating data obtained in complex environments. Under such circumstances, directly deploying the current student model to another battery chemistry or testing protocol is unlikely to yield reliable predictions. A more appropriate strategy is to retrain the model end-to-end using data from the target application scenario, and thereby obtain a new student model which is tailored to that scenario.

On the basis of these considerations, we retrain and evaluate the teacher and student models in the proposed CPKD framework in an end-to-end manner on the real-world vehicle operating dataset. For the dataset partitioning, we use records

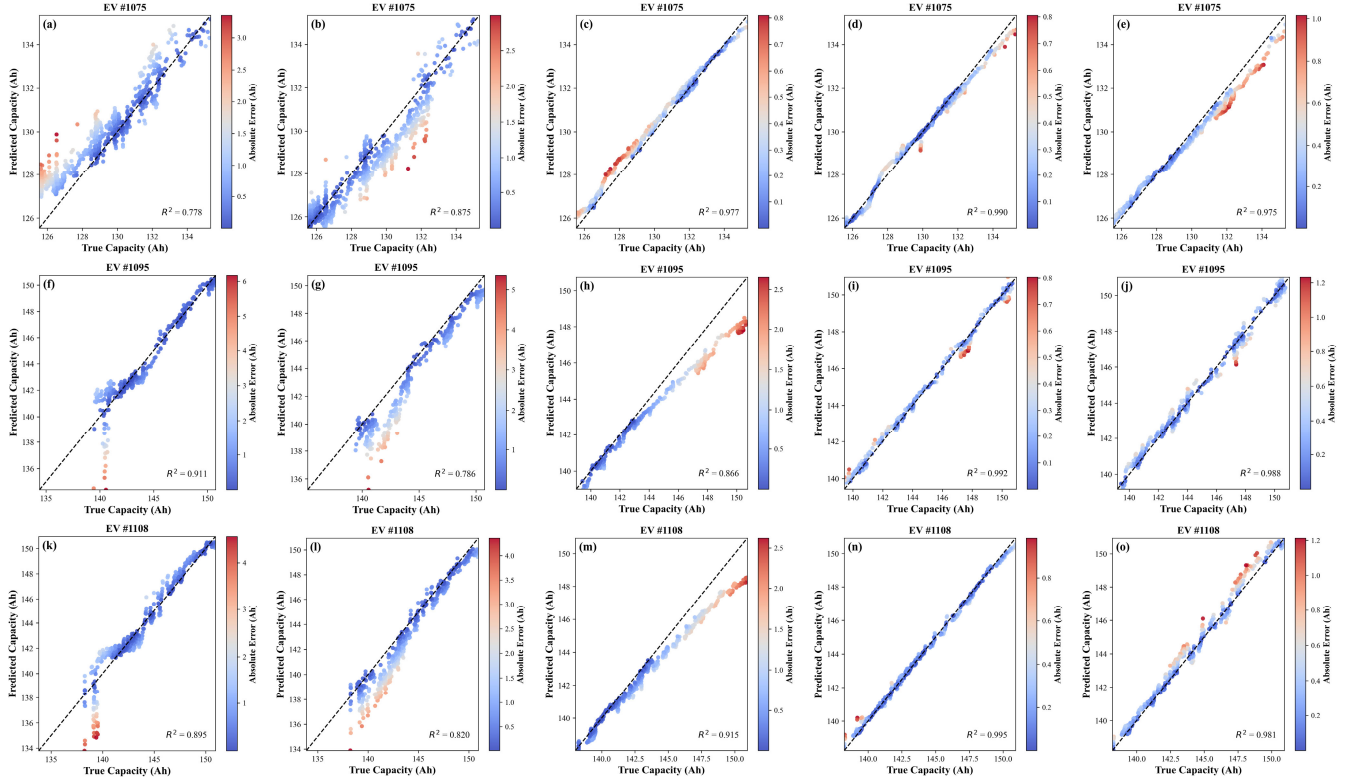


Fig. 9. Joint distributions between the ground-truth and predicted capacities, together with the predictive AEs, for different models on EV operating data. (a)–(e) EV#1075 and, in order, the LSTM NN, TCN, Transformer, the teacher model, and the student model distilled via SPF-KD. (f)–(j) EV#1095 and (k)–(o) EV#1108, with the same model order.

from EVs #1096, #1098, #1099, and #1106 for training, and records from EVs #1075, #1095, and #1108 for testing. Both the training and test sets cover two battery pack series parallel configurations, 3P92S and 3P90S. For feature extraction, we follow the feature extraction procedure previously used on the MIT dataset, adapted to the characteristics of real-world vehicle operating data, and obtain the area of the IC curve and the variance of the QV curve. Specifically, we segment the raw data according to changes in the ground-truth capacity and construct the IC and QV curves using driving mileage as the index. We then compute the two features with the same mileage index and feed them into the CPKD framework for training and testing.

To ensure a relatively fair comparison, we adopt the following experimental protocol. Across the SOTA models (TCN, LSTM NN, and Transformer), the teacher and student models, and the distillation variants (S-KD, L1-KD, L2-KD, KDNet, and SPF-KD), we strictly follow the architectures and hyperparameter configurations in Table II, adjusting only the LR. We then perform ten independent end-to-end runs on the training set to identify the optimal LR for each model, which is $8.0\text{E-}3$ for TCN, $7.0\text{E-}3$ for LSTM NN, $1.2\text{E-}3$ for Transformer, $1.0\text{E-}3$ for the teacher, and $1.0\text{E-}6$ for the student. For distillation hyperparameters, we use the optimal configuration obtained by grid search in the MIT dataset experiments. In SPF-KD, τ is set to 3 and α is set to 0.3, while the other distillation methods use the corresponding optimal parameter combinations reported above.

Because the truth capacity in the battery pack data collected from 7 EVs does not decay to 80% of the rated capacity, the full life degradation trajectory of the packs cannot be assessed. Under this constraint, the test phase follows the protocol used in the MIT dataset by placing the prediction start in the late life stage. Specifically, the first 80% of data from each EV is used as the test input, namely the first 584 samples of EV#1075, the first 292 samples of EV#1095, and the first 384 samples of EV#1108, while the remaining data serve as the reference ground truth, namely the last 146 samples of EV#1075, the last 73 samples of EV#1095, and the last 97 samples of EV#1108. It should be emphasized that the samples defined here are conceptually distinct from the data records described above. Each sample comprises QV and IC features, both calculated as functions of mileage. Because the data are segmented by capacity variation, the mileage segments associated with different vehicles are not identical. Further details are provided in the publicly available dataset released with this study. The prediction results of the compared models, the proposed teacher model, and the lightweight student models obtained with different distillation methods are shown in Fig. 9.

As shown in Fig. 9(d), (i), and (n), the teacher model achieves R^2 values of 0.9903, 0.9918, and 0.9955 on EV#1075, EV#1095, and EV#1108, respectively. The predicted and ground-truth capacity pairs cluster tightly around the diagonal reference line, and most points exhibit minimal deviation, indicated by blue colors corresponding to low absolute error, demonstrating high prediction accuracy. In contrast, the

TABLE IV

RMSES OF DIFFERENT MODELS ON THE REAL-WORLD EV DATASET

Methods	Battery Packs			ARMSE
	EV#1075	EV#1095	EV#1108	
TCN[32]	0.0769	0.0623	0.0842	0.0759
LSTM NN[33]	0.0834	0.0676	0.0913	0.0823
Transformer[34]	0.0595	0.0482	0.0652	0.0587
Teacher Model	0.0309	0.0250	0.0338	0.0305
Candidate benchmark	0.0787	0.0638	0.0862	0.0777
S-KD[27]	0.0837	0.0678	0.0916	0.0826
L1-KD[28]	0.0745	0.0603	0.0815	0.0735
L2-KD[29]	0.0958	0.0776	0.1048	0.0945
KDNet[23]	0.0655	0.0531	0.0717	0.0646
SPF-KD	0.0398	0.0301	0.0429	0.0384

TABLE V

PREDICTION RMSES OF THE STUDENT MODEL WITH DIFFERENT INPUT FEATURES UNDER EV OPERATING CONDITIONS

Input Data	Battery Packs			ARMSE
	#1075	#1095	#1108	
I, V	0.0396	0.1938	0.1599	0.1104
IC_area	0.0489	0.0677	0.0812	0.0615
QV_var	0.0663	0.0753	0.1110	0.0648
QV_var, IC_area	0.0398	0.0301	0.0429	0.0384

student model shows a moderately broader scatter distribution with more orange and red points, yet it still attains R^2 values of 0.778, 0.911, and 0.8955 on the three EV test sets. Overall, it outperforms LSTM NN and TCN, and it is only inferior to the Transformer on EV#1075. These results indicate that the student model derived from the CPKD framework effectively learns and inherits the predictive capability of the teacher model.

To further compare different models and distillation methods from the perspective of quantitative error, Table IV reports the RMSE on the three EV battery pack test sets. As seen in Table IV, the teacher model delivers markedly better performance than all compared models, achieving the lowest ARMSE of 0.0305. Meanwhile, the candidate student model without distillation still shows substantial room for improvement, with an ARMSE of 0.0777, which remains far worse than the teacher model. Among the compared distillation strategies, SPF-KD performs best, yielding ARMSE values of 0.0398 for #1075, 0.0301 for #1095, and 0.0429 for #1108, with an ARMSE of 0.0384. In particular, on EV#1095, SPF-KD reduces the ARMSE to 0.0301, which approaches the teacher model result of 0.0250, indicating that this strategy transfers the key knowledge of the teacher model to the lightweight student model with high efficiency.

To assess the robustness of the proposed CPKD framework with respect to different input features, we conducted three additional comparative experiments across input feature sets on the final selected student model using the EV dataset, as summarized in Table V. We focus on the student model because it is obtained via distillation from the teacher model and directly inherits its performance. Therefore, the robustness of the student model can serve as a proxy for the robustness of the teacher model, thereby reflecting the overall robustness of the CPKD framework. The evaluated feature

sets include charge–discharge current I and charge–discharge voltage V , IC features only as IC_area, QV features only as QV_var, and the combination of IC_area and QV_var. The results show that IC_area and QV_var yield comparable ARMSE when used individually, and both outperform the configuration using I and V . When QV_var and IC_area are used jointly, the ARMSE reaches the minimum value of 0.0384, further supporting the conclusion reported in the literature [31].

E. Ablation Study

To rigorously assess the efficacy of the proposed CPKD framework in capturing battery degradation dynamics under lightweight constraints, we conduct a series of ablation studies on the MIT dataset by systematically deactivating or replacing individual modules and evaluating the resulting model variants. Specifically, the teacher side ablation experiments are designed to assess the effectiveness of the CVAE composed of two MCFR modules. In essence, the MCFR module alone is structurally a standard variational autoencoder (VAE). The first VAE module provides the conditional component for feature reconstruction and is called the MCFR-I module for clarity. The second VAE module, namely the MCFR-II module, acts as the prediction head. Although the encoder and decoder in a VAE can be instantiated with MLP, LSTM, or Transformer backbones, we adopt MLPs as the backbone in practice. This design is supported by controlled comparisons in which all MLPs are replaced with Transformers, as well as by ablation results obtained by removing the MCFR-I module. For alignment, the complete teacher model is denoted as CVAE Mamba@MLP and is used as the reference baseline for teacher ablations. To ensure a fair ablation study, MCFR-I, MCFR-II, and the Transformer adopted an identical network topology, with an encoder of $[HL = 3, HLN_1 = 32, HLN_2 = 16, HLN_3 = 8]$ and a decoder of $[HL = 3, HLN_1 = 8, HLN_2 = 16, HLN_3 = 32]$. Student model ablations are designed to quantify the contribution of each component to three core capabilities, including inference of operating conditions associated with degradation, extraction of degradation-relevant features, and modeling of the long-short-term temporal dependencies intrinsic to the degradation trajectory. For consistency, the complete student model is denoted as CM-CVAE@1. When distilled using the proposed SPF-KD strategy, it yields the baseline lightweight variant CM-CVAE@1 + SPF-KD@1, which serves as the reference point for subsequent comparisons.

- 1) *Effect of the CVAE Backbone*: Replacing the backbone of all encoders and decoders in the two MCFR modules with a Transformer (CVAE Mamba@T) substantially increases FLOPs, parameters, and inference latency, while ARMSE (see Table VI) shows no significant reduction relative to the selected teacher model (CVAE Mamba@MLP). This is attributable to the $\mathcal{O}(\mathcal{L}^2)$ complexity of Transformer self-attention and the $\mathcal{O}(\mathcal{L}d^2)$ complexity of the feedforward network, together with an $\mathcal{O}(d^2)$ parameter scaling, which collectively inflate FLOPs and model size. Here, $\mathcal{L}$ denotes the number of tokens, and d denotes the model dimension. Furthermore, the teacher model already integrates Mamba

TABLE VI
ABLATION STUDY RESULTS OF KEY MODULES
IN THE CPKD FRAMEWORK

Models	Variants	Performance Index			ARMSE
		FLOPs(M)	Paras(M)	PC-GPU(ms)	
Teacher	CVAE-Mamba@T	0.905	1.835	70.21	0.0189
	CVAE-Mamba@VAE	0.437	1.149	44.87	0.0394
	CVAE-Mamba@MLP	0.443	1.153	44.97	0.0130
	DCM@0+SPF-KD@0	0.084	0.091	4.55	0.0498
	DCM@C+SPF-KD@0	0.158	0.167	5.60	0.0386
Student	DCM@G+SPF-KD@0	0.157	0.167	5.58	0.0635
	CCFR@0+SPF-KD@0	0.105	0.114	5.23	0.0404
	CM-CVAE@1+SPFKD@0	0.159	0.167	5.61	0.0331
	CM-CVAE@1+SPFKD@①	0.159	0.167	5.63	0.0387
	CM-CVAE@1+SPFKD@1	0.159	0.167	5.61	0.0163

with multihead self-attention to capture long-range dependencies in battery operating condition data. Therefore, further Transformerization of CVAE offers limited marginal benefit in ARMSE and can even degrade ARMSE due to increased training difficulty and elevated overfitting risk.

- 2) *Effect of the Conditional Module in CVAE*: Removing MCFR-I while retaining only MCFR-II yields an architecture equivalent to a single VAE, denoted as CVAE Mamba@VAE. The results indicate that using only a single VAE reduces the first layer connectivity of the CVAE, yielding modest decreases in FLOPs, parameter count, and inference time (see Table VI). However, this ablation removes the additional discriminative information provided by the other VAE in the CVAE, weakening the ability of the latent variables to distinguish different operating conditions. This is equivalent to forcing the teacher model to rely on MCFR-II to explain variations originally captured by MCFR-I, leading to a pronounced increase in ARMSE.
- 3) *Effect of the DCM Module*: Excluding the DCM module (DCM@0) while retaining AEFE, CCFR, and DADP under SPF-KD@0 results in higher inference speed but significantly reduced accuracy (see Table VI). This highlights the DCM module's role in interpreting operating conditions and adaptively emphasizing salient features, thereby enhancing representational capacity and prediction accuracy.
- 4) *Effect of the Clustering Component in the DCM Module*: Under the nondistilled setting (SPFKD@0), removing the cluster component from the DCM module leaves the student model's FLOPs, parameters, and inference latency largely unchanged, but increases ARMSE relative to the complete student model (CM CVAE@1 + SPFKD@0). This degradation arises because, without clustering, the operating condition information implicitly embedded in randomly sampled training data is indiscriminately routed to the gating component, which cannot effectively impose operating condition priors to guide condition monitoring channels toward targeted learning. As a result, condition-related supervisory signals are diluted, and the uncertainty of learning condition representations increases, raising the risk of insufficient

learning for specific operating conditions and consequently impairing generalization.

- 5) *Effect of the Gating Component in the DCM Module*: Removing the gating component from the DCM module without distillation (SPFKD@0) keeps the student model's FLOPs, parameters, and inference latency unchanged, but markedly increases ARMSE compared with the complete student model (CM-CVAE@1 + SPFKD@0). This occurs because when the gating component is deactivated, the clustering component defaults to forwarding only the top-ranked operating condition information to the condition monitoring channel, thereby suppressing representations from other operating conditions. Consequently, the student model overfits to a subset of operating conditions, substantially degrading cross-condition generalization and even causing ARMSE to exceed that of the variant without the DCM module (DCM@0).
- 6) *Effect of the CCFR Module*: Removing CCFR (CCFR@0) while preserving AEFE, DCM, and DADP under SPF-KD@0 decreases model complexity but also degrades performance. This suggests that CCFR is critical for aligning reconstructed features with their true degradation patterns, thereby supporting accurate prognostics.
- 7) *Effect of the SPF-KD Strategy*: Eliminating SPF-KD from the full model (CM-CVAE@1 + SPF-KD@0) leaves the computational cost unchanged, but causes a notable drop in accuracy (see Table VI). This confirms that SPF-KD effectively transfers the teacher model's ability to model both short- and long-term degradation dynamics, which is essential for maintaining high predictive fidelity.
- 8) *Effect of the SPF-KD Sequence*: After swapping the order of FD and KD in SPF-KD (CM-CVAE@1 + SPF-KD@①), the computational cost and memory footprint remain unchanged. The prediction accuracy, however, decreases markedly, and the GPU inference time increases slightly. This modification is analogous to imposing large changes on the intermediate layers of an already converged model, which can easily disrupt the learned decision boundaries and entails a considerable risk of overall performance degradation.

Furthermore, comparison between the CCFR@0 + SPF-KD@0 and CM-CVAE@1 + SPF-KD@0 configurations shows that incorporating the CCFR module reduces the prediction error from 0.0404 to 0.0331, indicating a substantial improvement in student model accuracy. In fact, the CCFR module compresses the 32-D global monitoring features from the DCM module into 8-D local latent variables through a 32-channel encoder. These latent variables contain rich battery degradation information and strongly support the predictions of the student model. To quantify their contributions, we employ Shapley Additive exPlanations (SHAP) and compute the weighted average marginal contribution (AMC) of each channel feature using the following equation, thereby clarifying how global monitoring features and local latent features

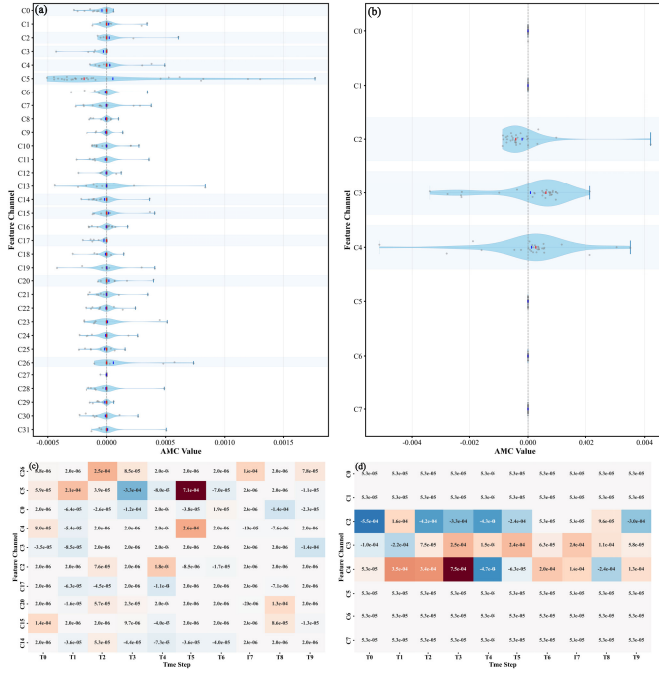


Fig. 10. SHAP interpretation of global monitoring features and local latent variables. (a) and (b) ACM violin plots for 32-channel global features and eight-channel local latent variables. (c) and (d) Channelwise contribution heatmaps for the top-10 channels of global features and all local latent variables.

affect capacity prediction:

$$\phi_{ij} = \sum_{S \subseteq F \setminus \{j\}} \omega(S) [f(S \cup \{j\}, x_i) - f(S, x_i)] \quad (17)$$

where ϕ_{ij} denotes the AMC of feature j to the i th capacity prediction over all possible feature combinations. The variable x_i denotes the test data used in the i th experiment with a fixed time window. The set F represents all features, and S is a subset of F that does not contain feature j . The function $f(S, x_i)$ denotes the capacity prediction when only the feature subset S is provided to the student model. The Shapley weight is $\omega(S) = |S|! (|F| - |S| - 1)! / |F|!$, which represents the probability that, among all permutations of the features, the subset S is exactly the set of features that appear before feature j . In this expression, $|S|!$ is the number of possible permutations of the features within S , $(|F| - |S| - 1)!$ is the number of permutations of the remaining features that follow feature j , and $|F|!$ is the total number of permutations of all $|F|$ features.

Specifically, based on the complete test trajectories starting from the late life prediction onset shown in Fig. 7(m)–(r) (totaling 1579 samples), we randomly selected 30 test records (180 samples) from each channel to assess the interpretability of the global monitoring features and the local latent variables. Using (17), we compute the AMC of each channel for both feature types, and obtain the channelwise SHAP distributions shown in Fig. 10(a) and (b) using SHAP violin plots. As shown in Fig. 10(a), the SHAP values of most channels are tightly concentrated around zero, indicating substantial redundancy with near-zero contribution in the global monitoring space.

A small number of channels exhibit broader distributions with longer tails. Among them, C5 shows the largest spread and the most pronounced positive tail, indicating that in some samples this channel increases the prediction relative to the prediction baseline, defined as the expected model output over the background set ($\phi_{ij} > 0$). Meanwhile, C5 also includes samples with negative contributions ($\phi_{ij} < 0$), suggesting that the same channel can exert opposite effects under different operating conditions. Similarly, nine channels, including C15 and C26, exhibit larger absolute AMC values and markedly wider ranges than the remaining channels, forming a subset with stronger contributions to the model prediction. In contrast, the local latent variables in Fig. 10(b) show stronger sparsity. Except for C2, C3, and C4, the remaining latent channels, such as C0, C1, C5, C6, and C7, lie close to zero with negligible contributions, indicating that the CCFR module further concentrates operating condition information into fewer local latent dimensions.

To further examine the temporal contributions of each channel for both feature types, we plotted heat maps over channels and time steps across 30 runs with a prediction window of 10 time steps, as shown in Fig. 10(c) and (d). For readability, Fig. 10(c) presents only the Top-10 channels ranked in Fig. 10(a). Fig. 10(c) shows that the importance of the global monitoring features is strongly selective in time rather than uniformly distributed across all time steps. For example, C5 exhibits its strongest positive mean contribution at T5, approximately 0.0007 while showing a clear negative contribution at T3, approximately -0.0003 . C4 also shows a strong positive contribution at T5, approximately 0.0003, whereas C26 exhibits moderate positive contributions at positions such as T2 and T7, approximately 0.0002. Overall, the model is mainly driven by a few key channels at a few critical time steps. Fig. 10(d) further reveals the dominant latent channels and their temporal structure. Most notably, C4 reaches the global maximum positive mean SHAP at T3, approximately 0.0008, and maintains positive contributions of approximately 0.0004 and 0.0003 at T1 and T2, respectively. It then switches to a pronounced negative contribution at T4, approximately -0.0005 , indicating a dynamic sign reversal at a critical time point. By contrast, C2 is predominantly negative in the early time steps, with mean contributions of approximately -0.0005 at T0 and -0.0004 at both T2 and T4. It exhibits only weak positive contributions at a few instants, for example, approximately 0.0002 at T1. C3 mostly exhibits moderate positive contributions, with a small number of negative instances at T0 and T1. These results indicate that the latent variables not only compress information, but also encode the temporal segments most sensitive to prediction more selectively within a small number of latent channels.

In summary, the analysis of Fig. 10 indicates that the CCFR module compresses the global monitoring features and aggregates degradation-related information, yielding a sparser and more concentrated distribution of channel importance while preserving temporal selectivity at key time steps. This capability plays a critical role in high-accuracy capacity prediction.

TABLE VII
COMPARE THE MODEL COMPLEXITY IN THE RUL PREDICTION TASK

Methods	Performance Index					
	FLOPs(M)	Paras(M)	EP-CPU(ms)	EP-GPU(ms)	Memory(MB)	Power(W)
TCN	0.242	0.247	92.73±14.53	21.31±0.96	304.5	3.49
LSTM NN	0.207	0.177	73.53±13.44	15.76±0.97	288.2	3.27
Transformer	0.496	1.847	244.98±34.57	61.49±1.97	478.3	5.97
Teacher Model	0.443	1.153	203.49±29.61	44.97±1.89	397.7	5.11
SPF-KD	0.159	0.167	63.21±9.69	13.88±0.86	270.1	2.85

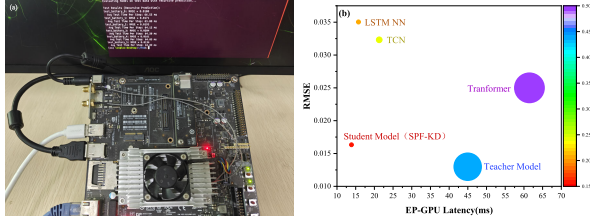


Fig. 11. Model lightweight deployment: (a) and (b) embedded platform setup and performance bubble chart postdeployment, respectively.

F. Lightweight Deployment of the Model

Deploying neural network-based battery RUL algorithms with strong generalization on onboard battery management systems (BMSs) is an enduring industrial demand, yet remains challenging under the hardware constraints of mass production. Current EV BMSs typically comprise a battery management unit (BMU), a cell monitoring unit (CMU), and a battery disconnect unit (BDU). Their computing platform is centered on an automotive functional safety BMU and lacks AI accelerators, such as GPUs or NPUs. For example, the Infineon AURIX TC3x provides, on the BMU side, about 300 MHz per core with up to six cores, no more than 16 MB of Flash, and slightly above 6 MB of on-chip RAM, which is insufficient for the compute and memory footprint required for deep model inference.

Two routes have emerged to mitigate the onboard compute bottleneck. The first relies on vehicle and cloud collaboration, uploading BMS data to the cloud for AI-based evaluation to deliver AI-enabled BMS functionality. The second, aligned with next-generation electrical and electronic architectures dominated by central computing and domain controllers, migrates core BMS computation to GPU or NPU platforms. Peak throughput of central computing platforms can reach about 254 TOPS, yet the compute budget available for AI battery health management is about 2 TOPS once competing workloads such as automated driving, human-vehicle interaction, and vehicle-body control are considered. Under these constraints, we use an NVIDIA Jetson TX2 to emulate a next-generation onboard BMS runtime environment and deploy a lightweight RUL prediction model on this platform. The TX2 integrates a 256-core Pascal GPU, a 1.92 GHz quad-core Arm Cortex A57 CPU, and 8 GB LPDDR4 memory, providing a peak AI throughput of about 1.26 TFLOPS, as shown in Fig. 11(a).

For fair benchmarking, the teacher model and SOTA baselines with the highest accuracy are evaluated on the same hardware. Table VII reports inference latencies on EP-CPU

and EP-GPU. It also summarizes FLOPs and parameter counts (computed on a standard PC), as well as the memory footprint and power consumption (measured on the TX2). As shown in Table VII, the student model yields a memory footprint of 270.1 MB and a power draw of 2.85 W, both the lowest among all methods. To visualize overall performance, Table VII data are presented in a bubble plot [see Fig. 11(b)], where bubble size and color encode parameter count and FLOPs, respectively. The proposed model, depicted as a small red bubble in the lower left quadrant, demonstrates the most favorable tradeoff between efficiency and accuracy.

IV. CONCLUSION

This study proposes a CPKD framework aiming at efficient compression and lightweight generation of high-accuracy RUL prediction models for lithium-ion batteries. Specifically, the CVAE-Mamba teacher model performs high-order compression and reconstruction of local and global degradation features. By incorporating a loss function that captures feature dependency information, the model is guided to learn the intrinsic structure and evolutionary patterns of the degradation process in the latent space, allowing for precise RUL prediction. The CM-CVAE student model monitors battery operating conditions through unsupervised clustering and a dual-path gating mechanism. By collaboratively adjusting two gating weights to activate a multichannel feature extractor composed of DSCs, the model enhances its perception and extraction of multicondition features, enabling accurate modeling and understanding of complex degradation processes. The SPF-KD strategy directs the student model to learn degradation features, reconstruction information, and predictive capabilities from the teacher model, achieving high-fidelity knowledge transfer to the lightweight student model. Numerical experiments demonstrate that the proposed distillation framework significantly improves inference speed on edge devices (EP-GPU Latency = 13.88 ms) while maintaining prediction accuracy (ARMSE = 0.0163). This work provides an efficient and practical solution for battery RUL prediction.

Although the proposed CPKD framework delivers strong predictive performance across multiple datasets, its effectiveness depends on the availability of sufficient source domain data. In practical deployment, where target domain data are often scarce and distribution shifts are common, the distillation performance may deteriorate. Future work may incorporate domain adaptation and meta-learning to enable rapid finetuning of the student model with limited target domain samples. In parallel, federated learning could be explored to aggregate battery data from multiple sources under privacy-preserving

settings, thereby enhancing the model's cross-domain transfer capability across different battery types and operating scenarios.

REFERENCES

- [1] P. Li, X. Wu, R. Grosu, J. Hou, M. Ilolev, and S. Xiang, "Applying neural network to health estimation and lifetime prediction of lithium-ion batteries," *IEEE Trans. Transport. Electric.*, vol. 11, no. 1, pp. 4224–4248, Feb. 2025.
- [2] R. Xiong, Y. Pan, W. Shen, H. Li, and F. Sun, "Lithium-ion battery aging mechanisms and diagnosis method for automotive applications: Recent advances and perspectives," *Renew. Sustain. Energy Rev.*, vol. 131, Oct. 2020, Art. no. 110048.
- [3] X. Hu, L. Xu, X. Lin, and M. Pecht, "Battery lifetime prognostics," *Joule*, vol. 4, no. 2, pp. 310–346, Feb. 2020.
- [4] P. Li et al., "An end-to-end neural network framework for state-of-health estimation and remaining useful life prediction of electric vehicle lithium batteries," *Renew. Sustain. Energy Rev.*, vol. 156, Mar. 2022, Art. no. 111843.
- [5] R. K. Singleton, E. G. Strangas, and S. Aviyente, "Extended Kalman filtering for remaining-useful-life estimation of bearings," *IEEE Trans. Ind. Electron.*, vol. 62, no. 3, pp. 1781–1790, Mar. 2015.
- [6] N. Li, Y. Lei, J. Lin, and S. X. Ding, "An improved exponential model for predicting remaining useful life of rolling element bearings," *IEEE Trans. Ind. Electron.*, vol. 62, no. 12, pp. 7762–7773, Dec. 2015.
- [7] M. Catelani, L. Ciani, R. Fantacci, G. Patrizi, and B. Picano, "Remaining useful life estimation for prognostics of lithium-ion batteries based on recurrent neural network," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.
- [8] Q. Zhu, X. Yu, Q. Wu, Y. Xu, F. Chen, and R. Huang, "Semi-empirical degradation model of lithium-ion battery with high energy density," *Energy Storage Sci. Technol.*, vol. 11, no. 7, p. 2324, 2022.
- [9] J. S. Edge et al., "Lithium ion battery degradation: What you need to know," *Phys. Chem. Chem. Phys.*, vol. 23, no. 14, pp. 8200–8221, 2021.
- [10] P. Li et al., "State-of-health estimation and remaining useful life prediction for the lithium-ion battery based on a variant long short term memory neural network," *J. Power Sources*, vol. 459, May 2020, Art. no. 228069.
- [11] Y. Zhang, R. Xiong, H. He, and M. G. Pecht, "Long short-term memory recurrent neural network for remaining useful life prediction of lithium-ion batteries," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 5695–5705, Jul. 2018.
- [12] P. Gu, Y. Zhang, B. Duan, C. Zhang, Y. Kang, and C. Li, "A lightweight multitask learning framework for battery degradation prediction with limited data," *IEEE Trans. Transport. Electric.*, vol. 11, no. 5, pp. 12547–12558, Oct. 2025.
- [13] G. Ma, Y. Zhang, C. Cheng, B. Zhou, P. Hu, and Y. Yuan, "Remaining useful life prediction of lithium-ion batteries based on false nearest neighbors and a hybrid neural network," *Appl. Energy*, vol. 253, Nov. 2019, Art. no. 113626.
- [14] S. Zhang, Z. Liu, Y. Xu, and H. Su, "A physics-informed hybrid multitask learning for lithium-ion battery full-life aging estimation at early lifetime," *IEEE Trans. Ind. Informat.*, vol. 21, no. 1, pp. 415–424, Jan. 2025.
- [15] S. Zhang, Z. Liu, Y. Xu, G. Chen, and H. Su, "An electrochemical aging-informed data-driven approach for health estimation of lithium-ion batteries with parameter inconsistency," *IEEE Trans. Power Electron.*, vol. 40, no. 5, pp. 7354–7369, May 2025.
- [16] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2023, *arXiv:2312.00752*.
- [17] J. Huang, L. Liu, H. Zhao, T. Li, and B. Li, "RUL-Mamba: Mamba-based remaining useful life prediction for lithium-ion batteries," *J. Energy Storage*, vol. 120, Jun. 2025, Art. no. 116376.
- [18] P. Li, Z. Ao, J. Hou, S. Xiang, and Z. Wang, "Physics-informed mamba neural network with potential knowledge for state-of-charge estimation of lithium-ion batteries," *J. Energy Storage*, vol. 123, Jul. 2025, Art. no. 116546.
- [19] J. Wu, C. Leng, Y. Wang, Q. Hu, and J. Cheng, "Quantized convolutional neural networks for mobile devices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4820–4828.
- [20] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," 2015, *arXiv:1510.00149*.
- [21] M. R. U. Saputra, P. Gusmao, Y. Almalioglu, A. Markham, and N. Trigoni, "Distilling knowledge from a deep pose regressor network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 263–272.
- [22] L. Wang and K.-J. Yoon, "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3048–3068, Jun. 2022.
- [23] Q. Xu, Z. Chen, M. Ragab, C. Wang, M. Wu, and X. Li, "Contrastive adversarial knowledge distillation for deep model compression in time-series regression tasks," *Neurocomputing*, vol. 485, pp. 242–251, May 2022.
- [24] K. Liu and Y. Li, "Remaining useful life prediction across machines using multi-source adversarial online knowledge distillation," *Eng. Appl. Artif. Intell.*, vol. 130, Apr. 2024, Art. no. 107726.
- [25] Z. Niu et al., "Knowledge distillation-based domain-invariant representation learning for domain generalization," *IEEE Trans. Multimedia*, vol. 26, pp. 245–255, 2024.
- [26] F. Xiao and L. Wu, "Personalized federated lithium-ion battery capacity prediction via cluster and fusion modules," *IEEE Trans. Transport. Electric.*, vol. 10, no. 3, pp. 6434–6448, Sep. 2024.
- [27] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
- [28] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, "Learning efficient object detection models with knowledge distillation," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 30, 2017, pp. 742–751.
- [29] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," 2014, *arXiv:1412.6550*.
- [30] K. A. Severson et al., "Data-driven prediction of battery cycle life before capacity degradation," *Nature Energy*, vol. 4, no. 5, pp. 383–391, Mar. 2019.
- [31] N. He, Q. Wang, Z. Lu, Y. Chai, and F. Yang, "Early prediction of battery lifetime based on graphical features and convolutional neural networks," *Appl. Energy*, vol. 353, Jan. 2024, Art. no. 122048.
- [32] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018, *arXiv:1803.01271*.
- [33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
- [34] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 6000–6010.
- [35] X. Wu et al., "LDNet-RUL: Lightweight deformable neural network for remaining useful life prognostics of lithium-ion batteries," *IEEE Trans. Power Electron.*, vol. 40, no. 9, pp. 3514–3528, Sep. 2025.



Penghua Li (Member, IEEE) received the B.S. and Ph.D. degrees from Chongqing University (CQU), Chongqing, China, in 2008 and 2012, respectively.

From 2013 to 2025, he was with Chongqing University of Posts and Telecommunications (CQUPT), Chongqing, as a Lecturer in 2013, an Associate Professor from 2014 to 2019, and a Professor from 2020 to 2025. He is currently a Professor with Chongqing University of Science and Technology (CQST), Chongqing, where he is also the Vice Dean of the School of Computer Science and Engineering. He

has received Chongqing Distinguished Young Scholars Fund. He has authored over 40 articles in IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, IEEE TRANSACTIONS ON POWER ELECTRONICS, IEEE TRANSACTIONS ON TRANSPORTATION ELECTRIFICATION, and IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, including nine ESI top 1% highly cited articles, contributed to two national standards, and holds over 30 Chinese invention patents. His research interests include neural network theory and its innovative applications in intelligent systems.

Dr. Li received eight provincial/ministerial awards, including the First Prize of the Chongqing Science and Technology Progress Award. He serves on the Technical Committee on Fault Diagnosis and Safety of Technical Processes, Chinese Association of Automation, and has chaired special sessions and forums in major control and automation conferences. He was selected for the Chongqing Talents Program (Young Top-Notch Talent) and Chongqing Innovation Young S&T Talent Program. He was also listed among the World's Top 2% Scientists.



Ao Chen (Student Member, IEEE) received the B.S. degree in Internet of Things engineering from Chongqing University of Posts and Telecommunications, Chongqing, China, in 2023, where she is currently pursuing the M.S. degree with the School of Automation.

Her research interests include battery fault diagnosis, time series forecasting, deep learning, and knowledge distillation.



Yunhong Che (Member, IEEE) received the B.E. degree in automotive engineering from Chongqing University, Chongqing, China, in 2019, where he is currently pursuing the M.S. degree in state estimation and life prediction of li-ion batteries with the Vehicle Power System Laboratory, Department of Automotive Engineering.



Yangming Zhang received the Ph.D. degree in management science and engineering from Air Force Engineering University, Xi'an, China, in 2018.

From 2018 to 2024, he was an Assistant Researcher with the Key Laboratory of Complex System Simulation, Beijing, China. He is currently the Deputy Director and an Associate Researcher at the Research Department, Key Laboratory of Complex System Simulation. He received Zhiqiang Fund support from the Academy of Military Sciences and also the Professional Technical Talent Position

Allowance. He has published over 20 academic articles in various journals and holds more than ten national invention patents and defense patents. His research interests include intelligent game and decision-making, and complex system modeling and simulation.

Dr. Zhang received the Provincial/Ministerial Level Second Prize for Scientific and Technological Progress in 2022 and 2025.



Jingjing Zhou received the Ph.D. degree from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 2019.

From 2013 to 2015, he worked at Contemporary Ampere Technology Company Ltd. (CATL), Fujian, China, as a Research and Development Engineer. Since 2019, he has been with China Automotive Engineering Research Institute Corporation (CAERI), Chongqing, China, as a Senior Expert. He has authored 16 SCI articles, contributed to two national standards, one local standard, and six

industry standards, and holds over 20 Chinese invention patents. His research interests include machine learning, large models, and their innovative applications in the prognostics and health management (PHM) of new energy vehicles.

Dr. Zhou serves as a Council Member for China Association for Inspection and Testing and a member for the Youth Science and Technology Work Committee. He has received five provincial/ministerial awards. He was selected for Chongqing Talent Program ("Hongyan Plan") and the New Chongqing Youth Innovation Talent Program.



Xinmin Zhang (Member, IEEE) received the Ph.D. degree in system science from Kyoto University, Kyoto, Japan, in 2019.

From April 2019 to December 2019, he was a Post-Doctoral Research Fellow with the Department of Systems Science, Kyoto University. He is currently an Associate Professor with the College of Control Science and Engineering, Zhejiang University, Hangzhou, China. His research interests include process control, process data analysis, machine learning, and industrial sensing technology

with applications to industrial processes.



Guodong Wang (Member, IEEE) received the Ph.D. degree in computer science from Vienna University of Technology, Vienna, Austria, in January 2019.

He was a Machine Learning Researcher with the Institute of Computer Engineering, Vienna University of Technology, dealing with theoretical research of machine learning and its industrial applications. He is currently the Managing Director of the Sino-Austria Research Institute for Intelligent Industries (SINOAI), Nanjing, China. He is the Vice Chief Engineer of Shanghai Institute of Computing Tech-

nology, Shanghai, China. He is a Guest Professor with Hangzhou Dianzi University, Hangzhou, China. His research interests include learning representation, deep neural networks, automated machine learning, cyber-physical systems, and the application of machine learning techniques on industrial data analytics.

Dr. Wang is a member of ACM SIGBED China.



Xiaosong Hu (Fellow, IEEE) received the Ph.D. degree in automotive engineering from Beijing Institute of Technology, Beijing, China, in 2012. He did scientific research and completed the Ph.D. dissertation at the Automotive Research Center, University of Michigan, Ann Arbor, MI, USA, from 2010 to 2012.

He was a Post-Doctoral Researcher with the Department of Civil and Environmental Engineering, University of California, Berkeley, Berkeley, CA, USA, from 2014 to 2015; and also at Swedish

Hybrid Vehicle Center and the Department of Signals and Systems, Chalmers University of Technology, Gothenburg, Sweden, from 2012 to 2014. He was also a Visiting Post-Doctoral Researcher with the Institute for Dynamic Systems and Control, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland, in 2014. He is currently a Professor with the State Key Laboratory of Mechanical Transmissions and the Department of Automotive Engineering, Chongqing University, Chongqing, China. His research interests include modeling and control of alternative powertrains and energy storage systems.

Dr. Hu was a recipient of several prestigious awards/honors, including the Emerging Sustainability Leaders Award in 2016, the EU Marie Curie Fellowship in 2015, the ASME DSCD Energy Systems Best Paper Award in 2015, and Beijing Best Ph.D. Dissertation Award in 2013.



Schahram Dustdar (Fellow, IEEE) is a Full Professor of Computer Science (Informatics) with a focus on Internet Technologies heading the Distributed Systems Group at the TU Wien.

Dr. Dustdar was founding co-Editor-in-Chief of *ACM Transactions on Internet of Things* (ACM TloT). He is Editor-in-Chief of *Computing* (Springer). He is an Associate Editor of *IEEE TRANSACTIONS ON SERVICES COMPUTING*, *IEEE TRANSACTIONS ON CLOUD COMPUTING*, *ACM Computing Surveys*, *ACM Transactions on the Web*,

and *ACM Transactions on Internet Technology*, as well as on the editorial board of *IEEE INTERNET COMPUTING* and *IEEE COMPUTER*. He is recipient of multiple awards: TCI Distinguished Service Award (2021), IEEE TCSVC Outstanding Leadership Award (2018), IEEE TCSC Award for Excellence in Scalable Computing (2019), ACM Distinguished Scientist (2009), ACM Distinguished Speaker (2021), IBM Faculty Award (2012). He is an elected member of the Academia Europaea: The Academy of Europe, where the chairman of the Informatics Section for multiple years. He is an IEEE Fellow (2016), an Asia-Pacific Artificial Intelligence Association (AAIA), a President (2021), and a Fellow (2021). He is an EAI Fellow (2021) and an I2CICC Fellow (2021). He is a Member of the IEEE Computer Society Fellow Evaluating Committee (2022 and 2023).