Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Multisensor contrast neural network for remaining useful life prediction of rolling bearings under scarce labeled data*#

Binkun LIU[1,2,3,4], Zhenyi XU[†‡2,3,4], Yu KANG[1,3], Yang CAO[1,3], Yunbo ZHAO[†‡1,3]

[1]*Department of Automation, University of Science and Technology of China, Hefei 230027, China*

[2]*Key Laboratory of System Control and Information Processing, Ministry of Education, Shanghai 200240, China*

[3]*Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230088, China*

[4]*Jianghuai Advance Technology Center, Hefei 230000, China*

[†]E-mail: xuzhenyi@mail.ustc.edu.cn; ybzhao@ustc.edu.cn

Received Aug. 29, 2024; Revision accepted Dec. 30, 2024; Crosschecked May 12, 2025

Abstract: Predicting remaining useful life (RUL) of bearings under scarce labeled data is significant for intelligent manufacturing. Current approaches typically encounter the challenge that different degradation stages have similar behaviors in multisensor scenarios. Given that cross-sensor similarity improves the discrimination of degradation features, we propose a multisensor contrast method for RUL prediction under scarce RUL-labeled data, in which we use cross-sensor similarity to mine multisensor similar representations that indicate machine health condition from rich unlabeled sensor data in a co-occurrence space. Specifically, we use ResNet18 to span the features of different sensors into the co-occurrence space. We then obtain multisensor similar representations of abundant unlabeled data through alternate contrast based on cross-sensor similarity in the co-occurrence space. The multisensor similar representations indicate the machine degradation stage. Finally, we focus on finetuning these similar representations to achieve RUL prediction with limited labeled sensor data. The proposed method is evaluated on a publicly available bearing dataset, and the results show that the mean absolute percentage error is reduced by at least 0.058, and the score is improved by at least 0.122 compared with those of state-of-the-art methods.

Key words: Self-supervised; Remaining useful life prediction; Contrast learning
https://doi.org/10.1631/FITEE.2400753 **CLC number:** TP277; TP311

1 Introduction

Remaining useful life (RUL) prediction for bearings aims to forecast the time duration from current operation until failure of the bearings (Wen et al., 2019). As a critical component of intelligent machine health management (Tao et al., 2018; Souza et al., 2021; Wang WJ et al., 2022), RUL prediction for bearings can assist in reducing maintenance costs and preventing significant losses from accidents, thereby improving the competitiveness.

Traditional approaches for RUL prediction for bearings can be broadly classified as model-based and statistics-based methods. Model-based methods (Morales-Espejel and Gabelli, 2020) require extensive domain knowledge to build physical models that accurately reflect machine degradation. However, obtaining domain knowledge is challenging, and building accurate physical models is difficult due to the complex system structure and operating

‡ Corresponding authors

* Project supported by the Dreams Foundation of Jianghuai Advance Technology Center (No. 2023-ZM01Z002), the Open Project Program of Key Laboratory of Ministry of Education of System Control and Information Processing (No. SCIP20230109), and the Open Research Fund of Anhui Provincial Key Laboratory of Intelligent Low-Carbon Information Technology and Equipment

Electronic supplementary materials: The online version of this article (https://doi.org/10.1631/FITEE.2400753) contains supplementary materials, which are available to authorized users

ORCID: Zhenyi XU, https://orcid.org/0000-0002-5804-882X; Yunbo ZHAO, https://orcid.org/0000-0002-3684-5297

environment. On the other hand, statistics-based methods (Li Q et al., 2022) focus on building a stochastic model that describes the degradation process to predict RUL based on monitored machine degradation variables. Nevertheless, these methods have limited processing ability on low-quality data.

In recent years, using deep learning for RUL prediction for bearings has become a research trend due to its effectiveness in improving prediction accuracy. Data-driven approaches (Wang B et al., 2020; Wang X et al., 2021) use deep learning models to establish potential relationships between machine monitoring data and RUL labels or degradation labels. These approaches having powerful degradation feature extraction capabilities could effectively handle massive and complex structured data and reduce the need for domain knowledge.

Compared to model-based and statistical-based approaches for RUL prediction for bearings, data-driven approaches heavily rely on RUL-labeled data. As manufacturing levels continue to improve and machine reliability gradually increases, it is often challenging to obtain a sufficient amount of failure data with RUL labels over a short-term period or too expensive to obtain data. Therefore, data-driven approaches still face challenges in practical applications due to the scarcity of degradation data with RUL labels.

Benefiting from recent significant progress in addressing the lack of labeled data in areas such as computer vision (Korbar et al., 2018; Tian et al., 2020; Zhu and Pu, 2021; Wang YT et al., 2023), self-supervised techniques provide solutions for RUL prediction for bearings under scarce labeled data, but current efforts suffer from poor discrimination of degradation features. Typically, current methods (Ding et al., 2022; Krokotsch et al., 2022; Akrim et al., 2023; Kong et al., 2023) treat each sensor signal as a channel and mine the temporal autocorrelation of multisensors from massive unlabeled sensor data during pretraining. Temporal autocorrelation is then used as a representation and finetuned with the limited labeled data to achieve RUL prediction. However, methods using stacked channels may have the weakness in possessing similar sensor signals in different degradation states, which is not favorable for RUL prediction. We stack the vertical and horizontal acceleration time–frequency matrices of the bearings and calculate the cosine similarity of the stacked matrices at any two moments, as shown in Figs. 1a and 1c. The figures illustrate that there are plenty of red high-similarity regions, which indicate similar sensor signals at different degradation states. To improve the discrimination of degradation features, we employ multisensor similarity. As shown in Figs. 1b and 1d, we first take the dot product of the vertical and horizontal acceleration time–frequency matrices, and then calculate the cosine similarity of the dot product matrices at any two moments. This approach significantly reduces the similarity of sensors in different degradation states compared to Figs. 1a and 1c.

Hence, we propose a multisensor contrast neural network, in which cross-sensor similar features indicating machine health conditions are captured from a large amount of unlabeled sensor data. Specially, the sensor data are mapped to the time–frequency domain through wavelet transform. Then, we devise an alternate contrast process to extract similar features between sensors from a large amount of unlabeled data. The similar features indicate machine health conditions. In the alternate contrast stage, the multibranch ResNet18 is used as a feature extractor to span the features into a co-occurrence space between sensors. In the co-occurrence space, any sensor is selected as the main sensor, and the remaining ones are regarded as auxiliary sensors. The auxiliary sensor feature extractor uses momentum update to ensure the consistency of features. Then, we calculate the similarity between the main sensor feature and those of auxiliary sensors. By optimizing the noise contrastive estimation (NCE) loss, we enforce the maximum similarity between the main sensor feature and its corresponding auxiliary sensor features at the same moment, thereby extracting the similar feature between the main sensor and the auxiliary sensors. We repeat the above process until each sensor has served as the main sensor, leading to highly discriminative degradation features. Next, the model is finetuned by using the scarce data with RUL labels. In the finetuning stage, attention is paid to adjusting feature weights. Ultimately, RUL prediction under scarce labeled sensor data is achieved with the assistance of rich unlabeled sensor data.

Overall, the main contributions of the proposed method are summarized as follows:

1. A multisensor contrast neural network is proposed, which can use a large amount of unlabeled

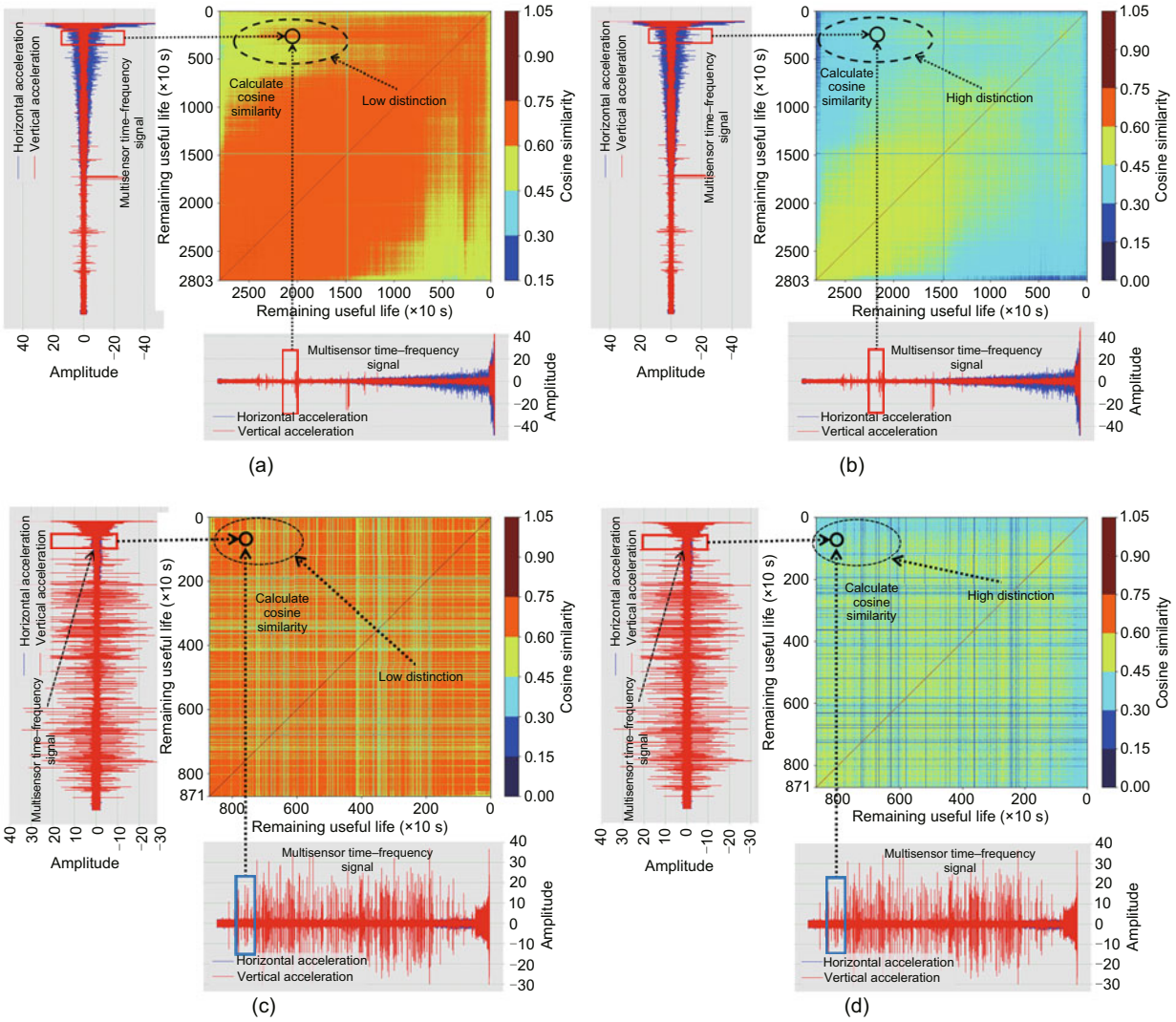

Fig. 1 Stacking multisensors causes different degradation states to behave similarly, and cross-sensor correlation improves the discrimination of degradation features: (a) the cosine similarity matrix between stacked signals at any two moments of bearing 1_1; (b) the correlation between the vertical and horizontal acceleration time–frequency signals calculated using the dot product, and the cosine similarity matrix between correlations at any two moments of bearing 1_1; (c) the cosine similarity matrix between stacked signals at any two moments of bearing 1_2; (d) the correlation between the vertical and horizontal acceleration time–frequency signals calculated using the dot product, and the cosine similarity matrix between correlations at any two moments of bearing 1_2. References to color refer to the online version of this figure

multisensor degradation data to model the degradation process and achieve RUL prediction when RUL labeled data are scarce. Compared with existing works, this method can extract more discriminative degradation features through cross-sensor contrast.

2. A cross-sensor alternate contrast process is devised to effectively mine highly discriminative degradation features from a large amount of unlabeled sensor data by maximizing the similarity of multisensor features at the same moment. Compared with existing works, the cross-sensor alternate contrast process can greatly improve the distinction of degradation features regarding RUL by mining potential shared degradation features among multisensors through alternate contrast.

3. The proposed method is evaluated on the Franche-Comté electronics mechanics thermal science and optics–sciences and technologies (FEMTO-ST) bearing dataset and the results show that the mean absolute percentage error (MAPE) is reduced

by at least 0.058, and the score is improved by at least 0.122 compared with those of state-of-the-art methods.

2 Related works

2.1 Works based on the FEMTO-ST bearing dataset

The FEMTO-ST bearing dataset is a dataset containing 17 sets of full-cycle bearing degradation data for three operating conditions and is mainly used to study RUL prediction for bearings. Recent research on the FEMTO-ST bearing dataset focuses on improving the accuracy of RUL prediction and the predicted RUL across working conditions.

Recent RUL prediction efforts have been excellent with sufficient data (Yang et al., 2023; Zou et al., 2023). Li Y et al. (2022) proposed a two-dimensional long short-term memory (2D-LSTM)-based fusion network for RUL prediction. The 2D-LSTM framework is used to extract deep time features of sensor data one by one and fuse multisensor features using an information fusion unit (IFU) to predict the RUL of bearings. Zuo et al. (2023) proposed a hybrid attention-based multiwavelet coefficient fusion method to evaluate the RUL of bearings. A hybrid attention-based convolutional LSTM network was used to self-adaptively extract features from the original signal after wavelet packet transformation to evaluate RUL. These efforts have demonstrated excellent performance.

Another research area of interest is to achieve accurate RUL prediction across operating conditions (Behera and Misra, 2023; Dong et al., 2023), and these approaches show excellent performance when relying on source domains with sufficient labeled RUL data. Deng et al. (2023) proposed a calibration-based hybrid transfer learning framework to improve data fidelity and model generality while demonstrating superiority in prediction accuracy and uncertainty quantification. To predict the RUL of bearings under invisible operating conditions, Ding et al. (2023) proposed an adversarial out-of-domain augmentation framework to generate pseudo-domains, thus increasing the diversity of available samples, and improving the generalization of inaccessible target domains.

2.2 Self-supervised learning

Deep neural network represented by convolutional neural network and the Transformer has made remarkable strides in the fields of computer vision and natural language processing, but it usually relies on sufficient labeled data. In some special fields, such as intelligent manufacturing and medicine, it is often very difficult to collect enough labeled data. Self-supervised learning can transform unsupervised learning based on unlabeled data into supervised learning based on labeled data by leveraging certain properties of unlabeled data to set pseudo-supervised tasks and learn features that are beneficial to the real task.

Self-supervised methods are mainly classified into generation-based methods and contrast-based methods. The generative methods are mainly concerned with the pseudo-supervisory task of data generation. The contrast-based approaches verify that multiple different input data channels correspond to each other. Tian et al. (2020) constructed a contrast pseudo-supervision task by maximizing the mutual information between different views of the same scene. Korbar et al. (2018) matched the visual and auditory elements of the video to achieve self-supervised learning

At present, there are relatively few research works (Melendez et al., 2019; Saeed et al., 2021; Zhang BM et al., 2021; Zhang WW et al., 2021) on the self-supervised RUL prediction. Ding et al. (2022) proposed a method to learn the multisensor self-sequence temporal correlation by contrasting the similarity between different sensor data augmentations. Krokotsch et al. (2022) considered the higher similarity of adjacent sensor time-series, and used the network to estimate the time difference between any two segments of a time-series segment. These methods all perform well but retain the difference between multisensors which may adversely affect RUL prediction.

3 Overview

3.1 Preliminary

Definition 1 (RUL) The RUL corresponds to the duration of machine operation from the start moment t to the failure moment T. This is formally

described as follows:

$$T - t|T > t. \tag{1}$$

Definition 2 (RUL prediction) Estimation of machine RUL at the beginning of prediction is based on effective information, such as machine health condition. This is formally described as follows:

$$T - t|T > t, \boldsymbol{Z}(t), \tag{2}$$

where $\boldsymbol{Z}(t)$ represents the valid information such as the machine health condition. In what follows, we use $\boldsymbol{Z}$ instead of $\boldsymbol{Z}(t)$ for notational simplicity. Machine health condition is usually constructed from sensor data to reflect the degree of machine degradation.

3.2 Problem formulation

Model $E(\cdot)$ is constructed based on the unlabeled RUL sensor dataset $\boldsymbol{X}^{\mathrm{u}} \in \mathbb{R}^{N^{\mathrm{u}} \times C \times M}$ and the labeled RUL sensor dataset $\boldsymbol{X}^{\mathrm{l}} \in \mathbb{R}^{N^{\mathrm{l}} \times C \times M}$ to build machine health condition $\boldsymbol{Z}$, where N^{u} and N^{l} are the amounts of data in the cases of unlabeled RUL and labeled RUL, respectively ($N^{\mathrm{u}} \gg N^{\mathrm{l}}$). Additionally, C is the number of sensors, M is the length of the time-series, and $\mathbb{R}$ is the set of real numbers. Then, a prediction model $f(\cdot)$ is constructed to establish the association between the machine health condition $\boldsymbol{Z}$ and RUL. The ultimate goal is to predict the corresponding RUL based on the sensor data $\boldsymbol{x}(t) \in \mathbb{R}^{C \times M}$ at the starting moment t.

$$\boldsymbol{Z}(t) = E(\boldsymbol{x}(t)|\boldsymbol{X}^{\mathrm{u}}, \boldsymbol{X}^{\mathrm{l}}), \tag{3}$$

$$\mathrm{RUL} = f(\boldsymbol{Z}(t)). \tag{4}$$

4 Method

4.1 Framework

Fig. 2 shows the framework of multisensor contrast neural network with attention. It includes mainly three parts: data preprocessing, alternate contrast, and finetuning. The data preprocessing step transforms the original signal into a time–frequency domain by wavelet transform to extract the features of the time and frequency domains at the same time. Alternative contrast is composed of feature extractor and feature contrast, and alternately captures similar features between multisensors. The feature extractor spans the nonlinear features extracted from the time–frequency domain without RUL labels into a co-occurrence space to learn similar features between sensors. In feature contrast, one sensor is regarded as the main sensor, and the others are regarded as the auxiliary sensors. The auxiliary sensor feature extractors use momentum updates to ensure the consistency of features. Then, the similarity between the main sensor feature and the auxiliary sensor features is calculated in the co-occurrence space, and the features at the same moment for different sensors with the maximum similarity are obtained by optimizing the NCE loss. We repeat the above process until each sensor serves as the main sensor to obtain the similarity between different sensors. Finetuning reuses the parameters of the feature extractor and initializes the predictor, thereby reducing the need for labeled data in the finetuning stage. Feature fusion makes full use of features from different sensors, and the attention mechanism adjusts feature weights to achieve RUL prediction.

4.2 Data preprocessing

Considering that in the machine degradation process, not only the amplitude but also the frequency will gradually change. The frequency and time features are revealed through data preprocessing. The methods of transforming the original signal from the time domain to the time–frequency domain mainly include short-time Fourier transform (STFT) and wavelet transform. Since the sensor signal is usually nonstationary and the window of the STFT is fixed, the high-frequency signal is suitable for STFT with a small window and a low-frequency signal is suitable for STFT with a large window; thus, STFT cannot meet the needs of nonstationary signal changes. The wavelet transform replaces the infinitely long trigonometric function basis in the Fourier transform with a finitely long decaying wavelet basis, so it can show the corresponding time when the different frequency components appear. The wavelet transform formula is as follows:

$$\boldsymbol{I}_{i,t} = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} \boldsymbol{x}_i(t) \psi \left(\frac{t-\beta}{a} \right) \mathrm{d}t, \tag{5}$$

where $\boldsymbol{x}_i(t)$ represents the original signal of the i^{th} sensor with t as the starting time and $t+L$ as the ending time. L is the length of the input time-series, a is the scale parameter, β is the translation parameter,

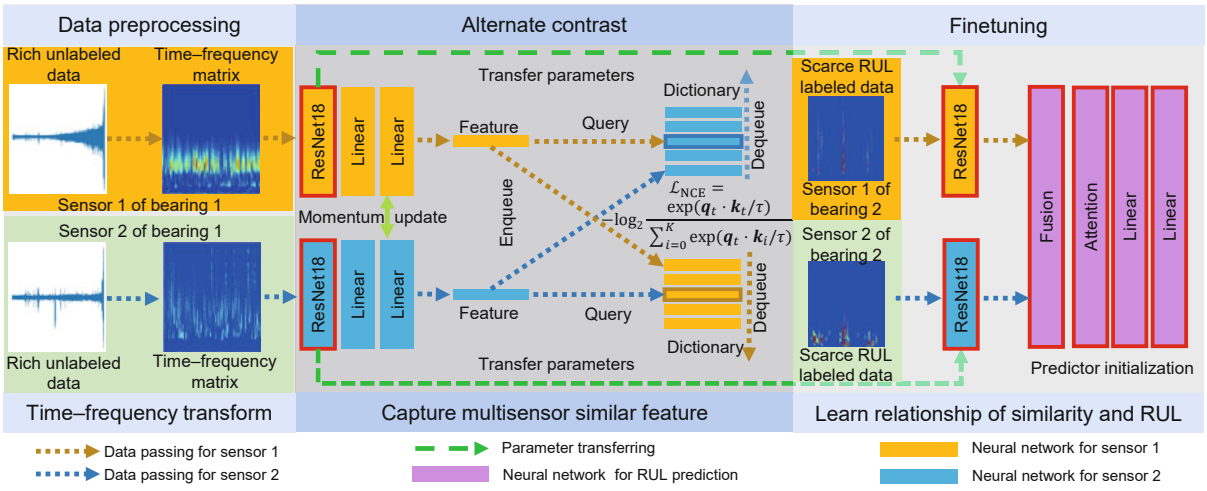

Fig. 2 The structure of the proposed framework. Dictionary: select any sensor alternately as the main sensor, and the others as auxiliary sensors; a dictionary is formed from the corresponding auxiliary sensor features; Enqueue: add the auxiliary sensor features of each batch to the corresponding dictionary; Dequeue: the earliest enqueued feature in the dictionary is deleted; Query: calculate the similarity between the main sensor feature and dictionary, and thus obtain the most similar auxiliary sensor feature; Momentum update: ignore the gradient of the auxiliary sensor network and use momentum to update slowly

and $\psi\left(\frac{t-\beta}{a}\right)$ represents the mother wavelet function. Here we choose the Gaussian derivative wavelet as the mother wavelet function. Through wavelet transform, the sensor raw signal $\boldsymbol{x}_i(t)$ of the sensor is transformed into a time–frequency matrix $\boldsymbol{I}_{i,t}$.

Bilinear interpolation $\varphi(\cdot)$ reduces the dimension of the time–frequency matrix $\boldsymbol{I}_{i,t}$ to speed up feature extraction. The equation is as follows:

$$\bar{\boldsymbol{I}}_{i,t} = \varphi(\boldsymbol{I}_{i,t}), \tag{6}$$

where $\bar{\boldsymbol{I}}_{i,t}$ is the time–frequency matrix after wavelet transformation and dimension reduction of the original signal of the i^{th} sensor with t as the start time and $t+L$ as the ending time.

4.3 Alternate contrast

Since only a large amount of unlabeled sensor data is available during the alternate contrast phase, it is necessary to use the characteristics of the sensor data as training labels. For multisensor systems, a good machine health condition is invariant across sensors. In our work, it is desired to obtain degenerate features that are invariant across sensors, which indicate machine health conditions. Therefore, the alternate contrast is performed using cross-sensor similarity to maximize the similarity of multisensor features at the same moment while suppressing the similarities of multisensor features at different moments to obtain highly discriminative cross-sensor invariant degradation features.

4.3.1 Main sensor selection

A multibranch network is constructed to extract the features of each sensor and span a co-occurrence space. In this paper, the ResNet18 backbone network E as the feature extractor and the fully connected layer C as the predictor are constructed as a multibranch network. We select the i^{th} sensor V_i as the main sensor V_{m}, the corresponding backbone network f_i is denoted as f_{m}, and the time–frequency matrix $\bar{\boldsymbol{I}}_{i,t}$ is denoted as $\bar{\boldsymbol{I}}_{\text{m},t}$. The other sensors V_j $(i \neq j)$ are regarded as auxiliary sensors V_{a_j}, the corresponding backbone network f_j is recorded as f_{a_j}, and the time–frequency matrix $\bar{\boldsymbol{I}}_{j,t}$ is denoted as $\bar{\boldsymbol{I}}_{\text{a}_j,k}$.

4.3.2 Feature dictionary

The feature $\boldsymbol{q}_t$ of the main sensor is the machine health condition recorded by the main sensor at time t. The feature dictionary D_j consists of the features $\{\boldsymbol{d}_{j,0}, \boldsymbol{d}_{j,1}, \cdots, \boldsymbol{d}_{j,K}\}$ of the j^{th} auxiliary sensor. The meaning of D_j is the machine health condition recorded at $K+1$ different moments by the j^{th} auxiliary sensor. It is formulated as follows:

$$\boldsymbol{q}_t = f_{\text{m}}(\bar{\boldsymbol{I}}_{\text{m},t}) = C_{\text{m}}(E_{\text{m}}(\bar{\boldsymbol{I}}_{\text{m},t})), \tag{7}$$

$$\boldsymbol{d}_{j,k} = f_{\mathrm{a}_j}(\bar{\boldsymbol{I}}_{\mathrm{a}_j,k}) = C_{\mathrm{a}_j}(E_{\mathrm{a}_j}(\bar{\boldsymbol{I}}_{\mathrm{a}_j,k})), \tag{8}$$

where f_{m}, C_{m}, and E_{m} refer to the backbone network, predictor, and feature extractor corresponding to the main sensor respectively, and f_{a_j}, C_{a_j}, and E_{a_j} refer to the backbone network, predictor, and feature extractor corresponding to the j^{th} auxiliary sensor respectively.

4.3.3 Feature similarity calculation

First, the degradation feature $\boldsymbol{q}_t$ of the main sensor at moment t and the feature dictionary $D_j = \{\boldsymbol{d}_{j,0}, \boldsymbol{d}_{j,1}, \cdots, \boldsymbol{d}_{j,K}\}$ corresponding to the j^{th} auxiliary sensor are given.

Then, the similarity $\mathrm{Sim}_{t,k}^{j}$ between the main sensor feature $\boldsymbol{q}_t$ and the degradation feature $\boldsymbol{d}_{j,k}$ at moment k ($k = 0, 1, \cdots, K$) in the dictionary D_j is shown in Eq. (9):

$$\mathrm{Sim}_{t,k}^{j} = \boldsymbol{q}_t \cdot \boldsymbol{d}_{j,k} = \sum_{n=1}^{N^{\mathrm{F}}} \boldsymbol{q}_t^n \boldsymbol{d}_{j,k}^n, \tag{9}$$

where "$\cdot$" denotes the dot product. N^{F} denotes the number of elements in the degenerate features $\boldsymbol{q}_t$ and $\boldsymbol{d}_{j,k}$. $\boldsymbol{q}_t^n \boldsymbol{d}_{j,k}^n$ denotes the multiplication of the n^{th} element in the degenerate feature $\boldsymbol{q}_t$ with the n^{th} element in the degenerate feature $\boldsymbol{d}_{j,k}$.

Finally, following the above steps, the similarity $\left\{\mathrm{Sim}_{t,0}^{j}, \mathrm{Sim}_{t,1}^{j}, \cdots, \mathrm{Sim}_{t,K}^{j}\right\}$ between the main sensor degradation feature $\boldsymbol{q}_t$ and all the degradation features $\boldsymbol{d}_{j,0}, \boldsymbol{d}_{j,1}, \cdots, \boldsymbol{d}_{j,K}$ in the feature dictionary D_j is calculated.

4.3.4 Loss function

The feature similarities at different moments are normalized using softmax. Let the multisensor feature similarity $\mathrm{Sim}_{t,t}^{j}$ with the matching label of the same moment t be 1 and the multisensor feature similarity with the matching labels of different moments be 0. Then the normalized feature similarities are fed into the cross-entropy. The cross-entropy maximizes the similarity of multisensor features at the same moment while suppressing the similarity of sensor features at different moments. In summary, to maximize the similarity of sensor features at the same moment to obtain the degradation features that are invariant across sensors, the NCE loss function $\mathcal{L}_{\mathrm{NCE}}$ is constructed by measuring the sensor similarity by dot product as follows:

$$\mathcal{L}_{\mathrm{NCE}} = -\sum_{j} \log_2 \frac{\exp\left(\frac{\mathrm{Sim}_{t,t}^{j}}{\tau}\right)}{\exp\left(\frac{\mathrm{Sim}_{t,t}^{j}}{\tau}\right) + \sum\limits_{k=0,k\neq t}^{K} \exp\left(\frac{\mathrm{Sim}_{t,k}^{j}}{\tau}\right)} + \mu_{\mathrm{c}} \left\|\theta_{\mathrm{m}}\right\|_2, \tag{10}$$

where τ is the temperature hyperparameter, μ_{c} is the regularization weight, and $\|\theta_{\mathrm{m}}\|_2$ represents the L_2 regularization of the main sensor network f_{m}.

The feature $\boldsymbol{q}_t$ of the main sensor looks for the most similar sample among the $K+1$ samples of auxiliary sensors, and this process is similar to the $K+1$ classification of feature $\boldsymbol{q}_t$.

4.3.5 Feature dictionary update by enqueuing and dequeuing

Since the feature $\boldsymbol{q}_t$ is constantly changing during the training process, the feature dictionary needs to be dynamically maintained. Longer dictionaries will use auxiliary sensor data. Since the length of the feature dictionary is much larger than the batch size, it is impractical to recalculate all the features in the feature dictionary each time. Therefore, updating is accomplished by queuing in a first-in-first-out mode. In each batch, auxiliary sensor features are enqueued to the corresponding feature dictionary, and the oldest auxiliary sensor feature is dequeued.

4.3.6 Updating auxiliary sensor network parameters

Because the auxiliary sensor backbone network is being updated, the parameters corresponding to these oldest features may be significantly different from the current parameters. Through the above operations, the dynamic update of the feature dictionary is gradually achieved, while ensuring that the dictionary length can be much larger than the batch size. It will reduce consistency among features in the dictionary by the rapidly changing parameters of the auxiliary sensor network. Therefore, the gradients of the auxiliary sensor networks are ignored and the momentum update is used for parameters update (He et al., 2020).

$$\theta_j \leftarrow m\theta_j + (1-m)\theta_{\mathrm{m}}, \tag{11}$$

where m is the momentum update factor, θ_{m} represents the main sensor network parameters, and θ_j is the j^{th} auxiliary sensor network parameters.

4.3.7 Alternate mechanism

The above processes are alternated every F_{ex} epochs until each sensor has served as the main sensor to ensure that the network of each sensor can learn good representation. Note that F_{ex} is the hyperparameter.

4.4 Finetuning

At the alternate contrast stage, the model has extracted cross-sensor invariant degradation features. To establish an association between cross-sensor invariant degradation features and RUL, the model is finetuned using scarce sensor data with RUL labels. The parameters of the feature extractor are reused and the predictor C_{F} is initialized.

4.4.1 Feature fusion

Features from different sensors are fused to fully use them. The fusion method selects concatenation, and the formula is as follows:

$$\boldsymbol{F}_t = \text{Cat}\left(\left[E_1\left(\bar{\boldsymbol{I}}_{1,t}\right), E_2\left(\bar{\boldsymbol{I}}_{2,t}\right), \cdots, E_i\left(\bar{\boldsymbol{I}}_{i,t}\right)\right]\right), \quad (12)$$

where Cat is the feature concatenation operator, and F_t is the concatenated feature. Specifically, it means that the degenerate features $E_1\left(\bar{\boldsymbol{I}}_{1,t}\right)$, $E_2\left(\bar{\boldsymbol{I}}_{2,t}\right)$, $\cdots$, $E_i\left(\bar{\boldsymbol{I}}_{i,t}\right)$ are concatenated according to the last dimension. $E_i(\cdot)$ refers to the feature extractor corresponding to the i^{th} sensor.

4.4.2 Attention mechanism

Spatial attention $\boldsymbol{A}_{\text{s}}$ is devised to adjust feature weights. The corresponding equations are as follows:

$$\boldsymbol{A}_{\text{s}} = \text{Softmax}\left(\boldsymbol{W}_{\text{i}}\boldsymbol{\alpha}_{\text{s}} + \boldsymbol{b}_{\text{i}}\right), \quad (13)$$

$$\boldsymbol{\alpha}_{\text{s}} = \text{Tanh}\left(\boldsymbol{W}_{\text{s}}\boldsymbol{F}_t + \boldsymbol{b}_{\text{s}}\right), \quad (14)$$

where $\boldsymbol{\alpha}_{\text{s}}$ is the nonlinear feature, $\boldsymbol{W}_{\text{s}}$ and $\boldsymbol{W}_{\text{i}}$ are trainable weight parameters, and $\boldsymbol{b}_{\text{s}}$ and $\boldsymbol{b}_{\text{i}}$ are trainable bias parameters. Tanh and Softmax are both nonlinear activation functions.

$$\bar{\boldsymbol{F}}_t = \boldsymbol{A}_{\text{s}} * \boldsymbol{F}_t, \quad (15)$$

where "$*$" is the element-wise multiplication, and $\bar{\boldsymbol{F}}_t$ is the feature after attention reconstruction.

4.4.3 RUL prediction

The predictor C_{F} performs regression prediction on the reconstructed feature $\bar{\boldsymbol{F}}_t$.

$$\text{RUL}_{\text{p}} = C_{\text{F}}\left(\bar{\boldsymbol{F}}_t\right), \quad (16)$$

where $\text{RUL}_{\text{p}} \in \mathbb{R}^+$ is the RUL predicted by the model. The loss function $\mathcal{L}_{\text{F}}$ at the finetuning stage is formulated as follows:

$$\mathcal{L}_{\text{F}} = \left(\text{RUL}_{\text{p}} - y_t\right)^2 + \mu_{\text{f}}\|\theta\|_2, \quad (17)$$

where y_t is the corresponding RUL label, $\|\theta\|_2$ represents L_2 regularization, and μ_{f} is the regularization coefficient.

5 Experiments and results

5.1 Data and setup

5.1.1 Description of datasets

The FEMTO-ST bearing dataset (Nectoux et al., 2012) contains accelerated degradation data of bearings. The bearing operating state was recorded every 10 s using vertical and horizontal acceleration sensors with a sampling frequency of 25.6 kHz. Each sampling lasted 0.1 s. This dataset contains a total of 17 bearing degradation datasets under three operating conditions. The analysis of this dataset is listed in Section I of supplementary materials.

5.1.2 Dataset setup

In real-world scenarios, factories usually collect only limited RUL-labeled data and a huge amount of unlabeled degradation data due to expensive collection costs and other reasons. Due to the limited research on RUL prediction in this scenario, there is currently a lack of publicly available datasets that are suitable for this scenario. For the research and comparison with the baseline methods, we make reasonable modifications to the original FEMTO-ST bearing dataset that contains a large amount of RUL data. The modifications to the dataset are aligned with real-world scenarios and can serve as a reference for subsequent similar research.

The training data for the alternate contrast phase and the finetuning phase of the experiments are shown in Table 1. The unlabeled dataset is constructed with a large amount of unlabeled data in the alternate contrast phase, and the RUL labeled

Table 1 Experimental dataset setup

Working condition	Alternate contrast phase (Training without RUL labels)	Finetuning phase (Training with RUL labels)	Test data
Condition 1	bearing 1_2, bearing 2_1, bearing 2_2, bearing 3_1, bearing 3_2	50% of bearing 1_1 data	bearing 1_3, bearing 1_4, bearing 1_5, bearing 1_6, bearing 1_7
Condition 2	bearing 1_1, bearing 1_2, bearing 2_2, bearing 3_1, bearing 3_2	50% of bearing 2_1 data	bearing 2_3, bearing 2_4, bearing 2_5, bearing 2_6, bearing 2_7
Condition 3	bearing 1_1, bearing 1_2, bearing 2_1, bearing 2_2, bearing 3_2	50% of bearing 3_1 data	bearing 3_3

dataset is constructed using a few RUL data in the finetuning phase. In the alternate contrast phase, the model is pre-trained using only the vibration acceleration data of the bearings. During the alternate contrast phase, any RUL labels are not involved in the training of the model. Then, in the finetuning phase, the model is finetuned using a small number of vibration accelerations with RUL labels. The labels are related to the training phase and not to the sensors themselves.

Taking Condition 1 as an example. The dataset used during the alternate contrast phase includes degraded data from bearing 1_2, bearing 2_1, bearing 2_2, bearing 3_1, and bearing 3_2. During the entire alternate contrast process, there are no RUL labels involved in the training. During the finetuning phase, the model is finetuned using only the RUL-labeled vibration acceleration data for the last 50% of the bearing 1_1. Bearing 1_3, bearing 1_4, bearing 1_5, bearing 1_6, and bearing 1_7 constitue the test sets.

5.1.3 Hyperparameters

In the alternate contrast stage, the regularization factor μ_{c} is 0.0001, the learning rate is 0.0001, the number of epochs is 800, and the main sensors alternate every 100 epochs. Stochastic gradient descent (SGD) is selected as the optimization algorithm, and the corresponding momentum is 0.9. The batch size is 128, the dictionary size $K+1$ is 3201, the momentum update factor m is 0.999, the temperature coefficient τ is 0.07, and the feature dimension is 128.

In the finetuning stage, the regularization factor μ_{f} takes a value of 0.01, the learning rate is 0.0001, and the number of epochs is 200. SGD is selected as the optimization algorithm, the corresponding momentum is 0.9, and the batch size is 128.

5.1.4 Metrics

The evaluation metrics are MAPE and the score. They are formulated as follows:

$$\mathrm{MAPE}=\frac{1}{11}\sum_{i=1}^{11}\frac{|\mathrm{ActRUL}_i-\widehat{\mathrm{RUL}}_i|}{\mathrm{ActRUL}_i}, \tag{18}$$

$$\mathrm{Er}_i=\frac{\mathrm{ActRUL}_i-\widehat{\mathrm{RUL}}_i}{\mathrm{ActRUL}_i}\times 100\%, \tag{19}$$

$$\mathrm{score}\ =\frac{1}{11}\sum_{i=1}^{11}A_i, \tag{20}$$

$$A_i=\begin{cases}\exp^{-\ln(0.5)(\mathrm{Er}_i/5)}, & \text{if } \mathrm{Er}_i\le 0,\\ \exp^{+\ln(0.5)(\mathrm{Er}_i/20)}, & \text{if } \mathrm{Er}_i>0,\end{cases} \tag{21}$$

where A_i is the metric that measures the prediction accuracy of the model for the i^{th} bearing, ActRUL_i and $\widehat{\mathrm{RUL}}_i$ are the true and predicted RUL values of the i^{th} bearing, and Er_i is an official intermediate metric used to evaluate the accuracy of RUL prediction. The score reflects the average performance of the model in predicting the final RUL of the 11 bearings. The higher the score is, the better the model performance will be. Further, the lower the MAPE, the higher the model performance.

5.1.5 Baselines

Due to our goal of RUL prediction under scarce labeled data, we choose four self-supervised methods to better show the method performance.

Self-supervised pretraining via contrast learning (SSPCL) (Ding et al., 2022): SSPCL is based on data augmentation. The most similar variant to the current variant is searched among all moments of data variants augmented by other data in the pretraining phase.

Self-supervised learning (SSL) (Krokotsch et al., 2022): An SSL model is proposed for RUL prediction. The time interval between any two time-series segments is estimated in the pretraining phase, thus learning the temporal correlation of the signals.

Deep self-supervised learning (DeepSSL) (Akrim et al., 2023): A DeepSSL model is proposed to overcome the lack of labeled data for RUL prediction. The encoding and decoding architectures are designed using the gated recurrent unit (GRU) to perform temporal prediction of the sensor signals in the pretraining phase.

Unlabeled sample learning (USL) (Kong et al., 2023): A contrastive learning framework is proposed for RUL prediction. First, an unlabeled sample augmentation is developed to extend the sample set. Then, an USL architecture is proposed to learn degradation information from unlabeled samples to improve the performance of general deep learning models in RUL prediction.

5.1.6 Experimental fairness

Our problem setting involves massive unlabeled sensor data and scarce sensor data with RUL labels. The self-supervised method has training data that are consistent with our method in two phases. Thus, all comparisons of our method with the self-supervised baselines are fair.

5.2 Results

5.2.1 Comparison with self-supervised baselines

Our proposed method and all self-supervised baseline methods are validated on 11 bearings using a large amount of unlabeled bearing data and 50% of the labeled degradation data. Table 2 presents the true values of RUL, the predicted values of RUL, MAPE, and the score (the mean value of A_i). Our method achieves an optimal score of 0.738 and MAPE of 0.108, while the suboptimal score is 0.616 and the suboptimal MAPE is 0.166. Our proposed multisensor contrast neural network outperforms SSPCL, SSL, DeepSSL, and USL in terms of MAPE and score, with at least a 0.058 decrease in MAPE, and a 0.122 increase in score. These results demonstrate the effectiveness of our alternate contrast, which maximizes the similarity of features between multisensors at the same moment,

Table 2 Experimental results for comparison with self-supervised methods with 50% labeled data and sufficient unlabeled data

Metric	Method	B 1_3	B 1_4	B 1_5	B 1_6	B 1_7	B 2_3	B 2_4	B 2_5	B 2_6	B 2_7	B 3_3	Mean
True RUL (s)		5730	339	1610	1460	7570	7530	1390	3090	1290	580	820	
Predicted RUL (s)	SSL	5402	282	1426	**1576**	5044	3139	**1446**	2803	1072	524	749	
	SSPCL	3484	287	1348	1334	3907	**3709**	1167	2634	1156	523	729	
	DeepSSL	5367	297	1479	1487	6502	2778	1202	2518	1107	443	756	
	USL	5445	279	1300	1344	6953	2344	1269	2573	1103	492	762	
	Our method	**5607**	**314**	**1505**	1483	**7000**	2670	1418	**2807**	**1204**	**543**	**783**	
A_i	SSL	0.820	0.558	0.673	0.332	0.315	0.133	0.572	0.725	0.557	0.716	0.741	0.558
	SSPCL	0.257	0.588	0.569	0.741	0.187	**0.172**	0.573	0.600	0.698	0.711	0.681	0.525
	DeepSSL	0.803	0.651	0.754	0.774	0.613	0.112	0.626	0.526	0.612	0.441	0.763	0.607
	USL	0.842	0.542	0.513	0.759	0.754	0.092	0.740	0.560	0.605	0.591	0.783	0.616
	Our method	**0.928**	**0.774**	**0.798**	**0.804**	**0.770**	0.107	**0.756**	**0.728**	**0.794**	**0.802**	**0.855**	**0.738**
MAPE	SSL	0.057	0.168	0.114	0.079	0.334	0.583	0.040	0.093	0.169	0.097	0.087	0.166
	SSPCL	0.392	0.153	0.163	0.086	0.484	**0.507**	0.160	0.148	0.104	0.098	0.111	0.219
	DeepSSL	0.063	0.124	0.081	0.018	0.141	0.631	0.135	0.185	0.142	0.236	0.078	0.167
	USL	0.050	0.177	0.193	0.079	0.082	0.689	0.087	0.167	0.145	0.152	0.071	0.172
	Our method	**0.021**	**0.074**	**0.065**	**0.016**	**0.075**	0.645	**0.020**	**0.092**	**0.067**	**0.064**	**0.045**	**0.108**

B denotes bearing; e.g., B 1_3 denotes bearing 1_3. The mean value of A_i is the official metric score of the FEMTO-ST dataset. The higher the score is, the better the model performance will be. Best results are in bold

emphasizing the discriminative power of similar features across sensors. Self-supervised RUL prediction algorithms, represented by SSL, SSPCL, DeepSSL, and USL, typically construct pretraining tasks by stacking sensor signals and exploiting the temporal correlation of sequences. However, these algorithms may be inefficient when the multisensor signals at different degradation stages are relatively similar. Therefore, our proposed method is needed to improve the discrimination of the features and achieve better performance by constantly alternating contrast to maximize the similarity of features among multisensor at the same moment.

We notice that compared with other bearings, the RUL prediction effect for bearing 2_3 is unsatisfactory. However, all the comparison baselines struggle to achieve good performance. This is possibly due to the uniqueness of bearing 2_3 itself, which makes it difficult to predict effectively given the scarcity of labeled RUL data.

Fig. 3 shows the RUL prediction for bearing 1_3. As can be observed from Fig. 3, in the early stage of bearing degradation, RUL prediction is close to horizontal, and our method does not perform very well. Whereas at the end of bearing degradation, our method can fit RUL change better. This is because during model training, we only use labeled data from the second half of the bearing degradation, so our method will perform better at the second half of the stage.

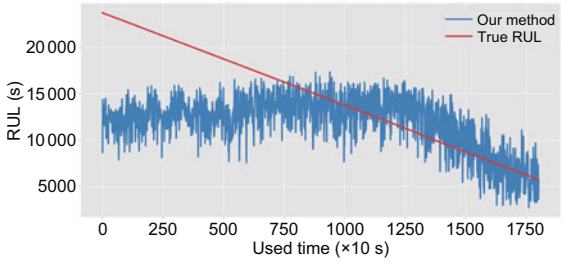

Fig. 3 RUL prediction for bearing 1_3. The horizontal axis is the used time and the vertical axis is the predicted RUL

Fig. 4 shows the visualization of the bearing 1_3 fusion feature after t-distributed stochastic neighbor embedding (t-SNE). Blue and green colors represent the early stage of degradation, yellow and orange indicate the middle stage of degradation, and red shows the end stage of degradation. The degradation trajectory is visible, and the transition from the current degradation stage to the next degradation stage can be very smooth. This proves that our method captures the degradation pattern of the bearing.

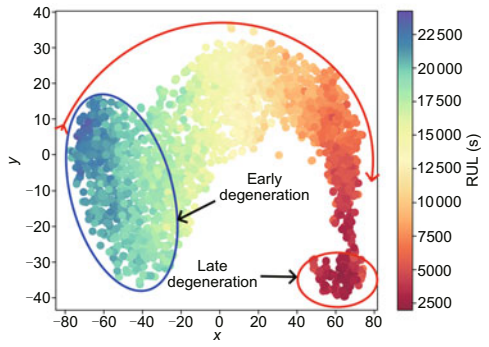

Fig. 4 Degenerate trajectories of bearing 1_3 in feature space. References to color refer to the online version of this figure

5.2.2 Ablation studies

To better explain the superiority of our method, ablation experiments are designed to construct the following two variants:

Variant-NP: To verify the effectiveness of the alternate contrast phase in the framework, the alternate contrast phase is skipped and the model is trained directly using a few data with RUL labels, and the remainder of the framework is retained unchanged.

Variant-Loss: To verify the effectiveness of the NCE loss function in the framework, the NCE loss function is replaced with the mean squared error (MSE) loss function for calculating the feature differences between different sensors at the same moment, and the remainder of the framework remains unchanged.

Table 3 presents a comparative analysis of predicted RUL, MAPE, and the score of our proposed method and its variants. We conduct these experiments to verify the effectiveness of the alternate contrast and NCE loss of our proposed method. To verify the effectiveness of alternate contrast, our proposed method is compared with Variant-NP, which does not have the alternate contrast mechanism. Compared with Variant-NP, MAPE of our proposed method is reduced by 0.184, and the score is improved by 0.302. Alternate contrast based on cross-sensor similarity can effectively reduce the need for labeled RUL data by capturing multisensor similar

Table 3 Ablation results

Metric	Method	B 1_3	B 1_4	B 1_5	B 1_6	B 1_7	B 2_3	B 2_4	B 2_5	B 2_6	B 2_7	B 3_3	Mean
True RUL (s)		5730	339	1610	1460	7570	7530	1390	3090	1290	580	820	
Predicted RUL (s)	Variant-NP	4889	192	1110	770	6924	2559	1172	1445	1110	499	706	
	Variant-Loss	4896	211	1201	1186	6200	2556	1205	1982	1114	454	681	
	Our method	**5607**	**314**	**1505**	**1483**	**7000**	**2670**	**1418**	**2807**	**1204**	**543**	**783**	
A_i	Variant-NP	0.601	0.222	0.341	0.194	0.744	0.101	0.581	0.158	0.617	0.616	0.618	0.436
	Variant-Loss	0.604	0.270	0.415	0.522	0.534	0.101	0.630	0.289	0.623	0.471	0.556	0.456
	Our method	**0.928**	**0.774**	**0.798**	**0.804**	**0.770**	**0.107**	**0.756**	**0.728**	**0.794**	**0.802**	**0.855**	**0.738**
MAPE	Variant-NP	0.147	0.434	0.311	0.473	0.085	0.660	0.157	0.532	0.140	0.140	0.139	0.293
	Variant-Loss	0.146	0.378	0.254	0.188	0.181	0.661	0.133	0.359	0.136	0.217	0.170	0.257
	Our method	**0.021**	**0.074**	**0.065**	**0.016**	**0.075**	**0.645**	**0.020**	**0.092**	**0.067**	**0.064**	**0.045**	**0.108**

B denotes bearing; e.g., B 1_3 denotes bearing 1_3. Best results are in bold

features. It also demonstrates that our model does not rely on finetuning to achieve good results.

To verify the effectiveness of NCE loss, our proposed method is compared with Variant-Loss, which uses MSE loss instead of NCE loss. Compared with Variant-Loss, the MAPE of our proposed method is reduced by 0.149, and the score is improved by 0.282. The MSE loss captures multisensor similarity by closing the distance between multisensor features at the same moment. NCE loss ensures that cross-sensor invariant highly discriminative degradation features are extracted by drawing on the idea of classification to extract multisensor similarity at the same moment, while suppressing multisensor feature similarities at different moments.

5.2.3 Hyperparametric sensitivity analysis

Fig. 5 shows the effects of the dictionary size on the model performance. As the dictionary size increases, the prediction error decreases. This is because the larger the dictionary size in the alternate contrast phase is, the more difficult it will be to match the main sensor features to the auxiliary sensor features at the same moment, which will enhance the model's ability to extract similar features across sensors.

To further analyze the performance of the proposed method, we compare the performance of the proposed method with the supervised baselines, as well as analyze the effect of the amount of labeled data on the performance of the proposed method. These experimental results are located in Section 2 of supplementary materials.

6 Conclusions

In this paper, for RUL prediction, we propose a multisensor contrast method that uses abundant unlabeled sensor data to assist a small amount of sensor data with RUL labels. Typically, current methods stack multisensor signals and mine the multisensor temporal autocorrelation from a large amount of unlabeled sensor data during pretraining, but suffer from poor discrimination of degradation features. Our approach uses an alternate contrast process to capture similar features (machine health conditions) among multisensors, and can effectively improve the discrimination of degradation features. The attention mechanism is used for finetuning to establish an association between degradation features and RUL. We fully evaluate our approach using the open FEMTO-ST bearing dataset, where the test dataset contains 11 sets of bearing degradation data under three different operating conditions. The proposed model outperforms other state-of-the-art baselines on test data, showing that, for RUL prediction, our proposed model can use rich unlabeled sensor data to assist a few sensor data with RUL labels.

Our method has some limitations. For example, our model considers only the bearing functioning under a single operating condition and ignores the effect of variable operating conditions on the RUL. From the application perspective, our proposed multisensor contrast framework can greatly improve the prediction accuracy using only a small amount of labeled degradation data and thus can be widely used for the prediction of a bearing's RUL, thereby reducing the failure maintenance cost of bearings. In

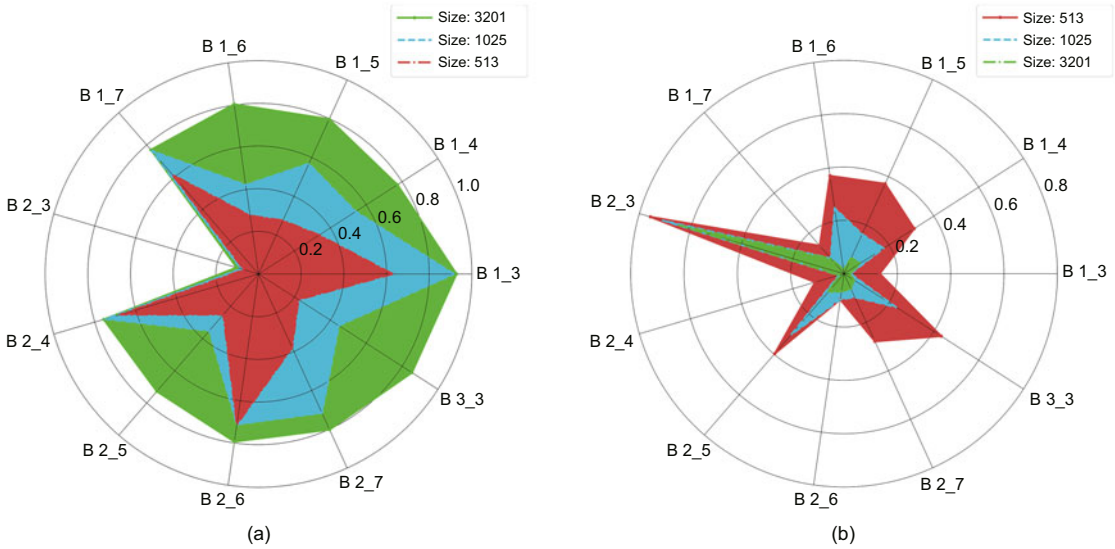

Fig. 5 Metrics of our method under different dictionary sizes: (a) score of our method under different dictionary sizes; (b) MAPE of our method under different dictionary sizes. B denotes bearing

the future, we are committed to analyzing and modeling the effects of dynamically changing operating conditions on the bearing's RUL, so that it can be adapted to a variety of complex industrial production environments.

Contributors

Binkun LIU designed the research. Zhenyi XU processed the data. Binkun LIU drafted the paper. Zhenyi XU helped organize the paper. Zhenyi XU and Yunbo ZHAO helped with data control and project management. Yunbo ZHAO, Yang CAO, and Yu KANG revised and finalized the paper. Zhenyi XU and Yunbo ZHAO provided the funding acquisition.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

References

Akrim A, Gogu C, Vingerhoeds R, et al., 2023. Self-supervised learning for data scarcity in a fatigue damage prognostic problem. *Eng Appl Artif Intell*, 120:105837. https://doi.org/10.1016/j.engappai.2023.105837

Behera S, Misra R, 2023. A multi-model data-fusion based deep transfer learning for improved remaining useful life estimation for IIOT based systems. *Eng Appl Artif Intell*, 119:105712. https://doi.org/10.1016/j.engappai.2022.105712

Deng YF, Du SC, Wang D, et al., 2023. A calibration-based hybrid transfer learning framework for RUL prediction of rolling bearing across different machines. *IEEE Trans Instrum Meas*, 72:1-15. https://doi.org/10.1109/TIM.2023.3260283

Ding YF, Zhuang JC, Ding P, et al., 2022. Self-supervised pretraining via contrast learning for intelligent incipient fault detection of bearings. *Reliab Eng Syst Saf*, 218:108126. https://doi.org/10.1016/j.ress.2021.108126

Ding YF, Jia MP, Cao YD, et al., 2023. Domain generalization via adversarial out-domain augmentation for remaining useful life prediction of bearings under unseen conditions. *Knowl-Based Syst*, 261:110199. https://doi.org/10.1016/j.knosys.2022.110199

Dong SJ, Xiao JF, Hu XL, et al., 2023. Deep transfer learning based on Bi-LSTM and attention for remaining useful life prediction of rolling bearing. *Reliab Eng Syst Saf*, 230:108914. https://doi.org/10.1016/j.ress.2022.108914

He KM, Fan HQ, Wu YX, et al., 2020. Momentum contrast for unsupervised visual representation learning. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.9729-9738. https://doi.org/10.1109/CVPR42600.2020.00975

Kong ZQ, Jin XH, Xu ZG, et al., 2023. A contrastive learning framework enhanced by unlabeled samples for remaining useful life prediction. *Reliab Eng Syst Saf*, 234:109163. https://doi.org/10.1016/j.ress.2023.109163

Korbar B, Tran D, Torresani L, 2018. Cooperative learning of audio and video models from self-supervised synchronization. Proc 32nd Int Conf on Neural Information Processing Systems, p.7774-7785.

Krokotsch T, Knaak M, Gühmann C, et al., 2022. Improving semi-supervised learning for remaining useful lifetime estimation through self-supervision. *Int J Progn Health*

Manage, 13(1):853. https://doi.org/10.36001/ijphm.2022.v13i1.3096

Li Q, Yan CF, Chen GY, et al., 2022. Remaining useful life prediction of rolling bearings based on risk assessment and degradation state coefficient. *ISA Trans*, 129:413-428. https://doi.org/10.1016/j.isatra.2022.01.031

Li Y, Wang HJ, Li JW, et al., 2022. A 2-D long short-term memory fusion networks for bearing remaining useful life prediction. *IEEE Sens J*, 22(22):21806-21815. https://doi.org/10.1109/JSEN.2022.3202606

Melendez I, Doelling R, Bringmann O, 2019. Self-supervised multi-stage estimation of remaining useful life for electric drive units. Proc IEEE Int Conf on Big Data, p.4402-4411. https://doi.org/10.1109/BigData47090.2019.9005535

Morales-Espejel GE, Gabelli A, 2020. A model for rolling bearing life with surface and subsurface survival: surface thermal effects. *Wear*, 460-461:203446. https://doi.org/10.1016/j.wear.2020.203446

Nectoux P, Gouriveau R, Medjaher K, et al., 2012. PRONOSTIA: an experimental platform for bearings accelerated degradation tests. Proc IEEE Int Conf on Prognostics and Health Management, p.1-8.

Saeed A, Salim FD, Ozcelebi T, et al., 2021. Federated self-supervised learning of multisensor representations for embedded intelligence. *IEEE Int Things J*, 8(2):1030-1040. https://doi.org/10.1109/JIOT.2020.3009358

Souza JS, Bezerril MC, Silva MA, et al., 2021. Motor speed estimation and failure detection of a small UAV using density of maxima. *Front Inform Technol Electron Eng*, 22(7):1002-1009. https://doi.org/10.1631/FITEE.2000149

Tao F, Qi QL, Liu A, et al., 2018. Data-driven smart manufacturing. *J Manuf Syst*, 48:157-169. https://doi.org/10.1016/j.jmsy.2018.01.006

Tian YL, Krishnan D, Isola P, 2020. Contrastive multiview coding. Proc 16$^{\text{th}}$ European Conf on Computer Vision, p.776-794. https://doi.org/10.1007/978-3-030-58621-8_45

Wang B, Lei YG, Li NP, et al., 2020. Multiscale convolutional attention network for predicting remaining useful life of machinery. *IEEE Trans Ind Electron*, 68(8):7496-7504. https://doi.org/10.1109/TIE.2020.3003649

Wang WJ, Wang Y, Wang J, et al., 2022. Ensemble enhanced active learning mixture discriminant analysis model and its application for semi-supervised fault classification. *Front Inform Technol Electron Eng*, 23(12):1814-1827. https://doi.org/10.1631/FITEE.2200053

Wang X, Wang TY, Ming AB, et al., 2021. Cross-operating condition degradation knowledge learning for remaining useful life estimation of bearings. *IEEE Trans Instrum Meas*, 70:3520911. https://doi.org/10.1109/TIM.2021.3091461

Wang YT, Cai F, Pan ZQ, et al., 2023. Self-supervised graph learning with target-adaptive masking for session-based recommendation. *Front Inform Technol Electron Eng*, 24(1):73-87. https://doi.org/10.1631/FITEE.2200137

Wen YX, Wu JG, Zhou Q, et al., 2019. Multiple-change-point modeling and exact Bayesian inference of degradation signal for prognostic improvement. *IEEE Trans Autom Sci Eng*, 16(2):613-628. https://doi.org/10.1109/TASE.2018.2844204

Yang L, Liao YH, Duan RK, et al., 2023. A bidirectional recursive gated dual attention unit based RUL prediction approach. *Eng Appl Artif Intell*, 120:105885. https://doi.org/10.1016/j.engappai.2023.105885

Zhang BM, Mao YF, Chen X, et al., 2021. Self-supervised learning advance fault diagnosis of rotating machinery. Proc 2$^{\text{nd}}$ Int Conf on Neural Computing for Advanced Applications, p.319-332. https://doi.org/10.1007/978-981-16-5188-5_23

Zhang WW, Chen DJ, Kong Y, 2021. Self-supervised joint learning fault diagnosis method based on three-channel vibration images. *Sensors*, 21(14):4774. https://doi.org/10.3390/s21144774

Zhu SH, Pu J, 2021. A self-supervised method for treatment recommendation in sepsis. *Front Inform Technol Electron Eng*, 22(7):926-939. https://doi.org/10.1631/FITEE.2000127

Zou WH, Lu ZQ, Hu ZY, et al., 2023. Remaining useful life estimation of bearing using deep multiscale window-based transformer. *IEEE Trans Instrum Meas*, 72:3514211. https://doi.org/10.1109/TIM.2023.3268453

Zuo T, Zhang K, Zheng Q, et al., 2023. A hybrid attention-based multi-wavelet coefficient fusion method in RUL prognosis of rolling bearings. *Reliab Eng Syst Saf*, 237:109337. https://doi.org/10.1016/j.ress.2023.109337

List of supplementary materials

1. Analysis of dataset

2. Extended experimental analysis

Table S1 Experimental results compared to supervised methods with 50% labeled data and sufficient unlabeled data

Fig. S1 Results for bearing 1_1

Fig. S2 Metrics of our method under different percentages of labeled data