Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-17557-2

1243: MULTI-MODAL INFORMATION ANALYSIS AND APPLICATIONS BASED ON CHAT-GPT

Unknown fault detection method for rolling bearings based on image and signal series feature fusion enhancement

Di Niu[1] · Shusong Yu[1] · Jiali Xu[2] · Chenglong Wang[1] · Ruoxi Li[1] · Xiangqian Ding[1]

Received: 18 August 2023 / Revised: 10 September 2023 / Accepted: 16 October 2023

Abstract

This paper mainly studies the problem of finding new fault classes under different modes in the field of intelligent fault diagnosis, that is, in the case of some labeled faults, new classes are revealed in unlabeled fault samples. In this paper, we introduce a comprehensive multi-modal framework for novel fault discovery and explore the impact of different modalities on the task of identifying new fault classes. To enhance the robustness of feature representation in complex environments, We adopted the approach inspired by ChatGPT, wherein we conducted pre-training on a substantial amount of labeled data to learn general features, patterns, and representations of various fault types. During the pre-training process, we integrated multiple modalities of data to prevent the loss of information due to the absence of single-modal data, thereby enhancing the accuracy of clustering. Furthermore, we introduced a multi-modal fusion method based on saliency correlation to complementarily fuse the information from different modalities. This approach effectively eliminated data redundancy arising from diverse modalities. Adhering to the principle that improving the quality of pseudo-label generation during the new class discovery phase enhances the accuracy of clustering for new classes, we extend the multi-modal concept. We introduce a Discriminative Relationship Enhancement method that capitalizes on cross-validation of pseudo-label predictions from different modalities during the pseudo-label prediction phase. This augmentation enhances the precision of pseudo-labels during the new class discovery phase. We evaluated the effectiveness of our proposed framework on fault datasets CWRU and PU, achieving promising results.

Keywords Novel class discovery · Intelligent fault diagnosis · Multimodal · Convolutional neural network

✉ Shusong Yu
yushusong@ouc.edu.cn

✉ Jiali Xu
jlxu@qnlm.ac

Extended author information available on the last page of the article

Published online: 18 November 2023

1 Introduction

Rolling bearings play a crucial role as essential components in rotating machinery. The state of their health directly impacts the operational condition of mechanical equipment. In severe scenarios, their deterioration can even result in mechanical failures [1, 2]. Hence, the timeliness and accuracy of fault recognition are of paramount importance in industrial production. Intelligent fault detection refers to the use of artificial intelligence techniques to monitor, identify, and predict equipment faults. Unknown faults refer to fault scenarios that have not been previously recorded or analyzed. In complex working environments where multiple factors come into play, the variety of possible fault types can increase. Therefore, the detection of unknown faults can be even more precise compared to traditional intelligent fault detection. This precision enables the accurate identification of fault information and aids in making informed decisions to minimize losses, particularly in intricate operational contexts.

The more complex the situation, the greater the need for a substantial amount of data for analysis. Deep learning, a commonly used method in the era of artificial intelligence, has developed rapidly due to the vast amount of data available. It has achieved amazing results in computer vision, natural language processing, and other fields. However, while the abundance of data has led to performance improvements through supervised learning, it has also brought about challenges in data annotation and incurred high costs. To address these issues and adapt to scenarios where labeled data is scarce or unavailable in practical applications, an increasing number of researchers have shifted their focus towards utilizing techniques that involve a limited amount of labeled or entirely unlabeled data. As a result, four approaches have emerged: unsupervised learning, semi-supervised learning, transfer learning, and new class discovery. These methods aim to tackle the challenges posed by the scarcity of labeled data and enable the utilization of the extensive amount of unlabeled data for diverse tasks.

Unsupervised learning relies on neither labeled information nor annotations. It discovers relationships between samples by exploring the inherent structure or features of the data, thereby accomplishing tasks such as clustering and dimensionality reduction [3–5]. Due to the lack of labeled information and the need for more empirical parameter tuning, it may exhibit significant biases during prediction.Semi-supervised learning [6–9] solves the problems of weak generalization ability and imprecision in the supervised learning model by adding a large number of unlabeled samples to a small number of labeled samples for learning. The core of transfer learning lies in finding similarities between new and existing knowledge, using pre-existing knowledge to learn new knowledge. It aids the model in better convergence and generalization by leveraging pre-training and enhances learning efficiency for the target domain. However, transfer learning [10] is applicable under the basic premise that the data in the source domain and the target domain are different but of the same category. To address this limitation, researchers introduced the concept of new class discovery, aiming to classify previously unseen, unlabeled data into appropriate categories. Unsupervised clustering, in principle, can solve the new class discovery problem, but its effectiveness in practice is limited. The goal of the new class discovery method is to use the prior knowledge of known classes to identify new classes in unlabeled data, train a model by learning the potential commonality of known class knowledge in labeled data, and use the model to classify the new class data. This concept is similar to that of ChatGPT, both involving extensive pre-training to extract generic features for improved performance on specific tasks [11–13]. The new class discovery realizes the effect of efficient classification of both known and unknown class data and becomes the current state-of-the-art method to solve the classification of unknown data.

Currently, the known methods for new class discovery are primarily focused on the natural image domain, with limited achievements in other domains. However, as the country vigorously developing industrial intelligence, fault diagnosis technology has become particularly important in the industrial sector. Traditional fault diagnosis methods heavily rely on specialized skills and expertise, often demanding substantial human, material, and time resources, and their accuracy is limited. Furthermore, with the increasing complexity of modern automated and intelligent equipment that underpins industrial production and services, fault diagnosis has become increasingly challenging, presenting significant challenges to fault diagnosis technology. Despite the gradual replacement of traditional fault diagnosis techniques with many deep learning-based methods, the issue of unknown fault types that may arise in real fault diagnosis scenarios remains unresolved. In the context of industrial intelligence, there is a pressing need for new class discovery techniques that can effectively handle unknown fault types and contribute to more accurate and efficient fault diagnosis, reducing reliance on manual expertise and mitigating the challenges posed by complex and automated systems.

In order to solve the above problems, a method of unknown fault diagnosis for rolling bearings is proposed in this paper. Considering the different data characteristics of different data sources, in the face of today's more complex fault situations, we aim to make better use of the diversity of data. The central idea is based on multimodality [14, 15], which involves integrating fault features from various data formats to enhance fault diagnosis. In the subsequent experimental section, different data formats are individually tested to demonstrate the superiority of the multimodal approach. To apply multimodality to the discovery of new fault classes, a framework suitable for bearing fault diagnosis is designed, building upon the existing UNO model. The contributions to this article are shown below:

1. This paper introduces a new class discovery framework for fault detection that utilizes multimodal data information, effectively leveraging the complementary nature of different modalities and enriching feature information. It proposes a novel multimodal data representation fusion module based on saliency correlation to address redundancy issues in data fusion and fully exploit the complementarity between different modalities.
2. A novel multi-scale deep feature extractor is proposed that adopts different multi-scale feature extraction structures for different modalities. By combining multi-scale information from shallow features and deep semantic features, the model's feature extraction capability is enhanced, thereby improving the accuracy of clustering during the new class discovery stage.
3. A novel multimodal pseudo-label generation module is proposed. It initially calculates the weight hyperparameters for fusing different modalities by leveraging the difference in information entropy between modalities to form a fused multimodal discriminative relation vector. Additionally, it enhances the significance of inter-class discriminative relations in the fused vector by utilizing the probabilities of inter-class discriminative relations.

The remainder of this paper is structured as follows: Section 2 of this paper mainly describes the research related to this work. Details about the methodology of this article are provided in Section 3. In the fourth section, the experimental process, relevant data, and indicators to verify the validity of this model are introduced in detail. Section 5 summarizes the full text and looks forward to future work.

2 Related work

Unknown fault detection technology involves analyzing and processing known fault data to learn feature representations that can be applied in the process of novel fault discovery, even when the fault patterns are unknown. In the pre-training phase of unknown fault discovery, the capability to recognize and extract features from known fault classes is a crucial prerequisite for successful novel fault class discovery. Therefore, accurate diagnosis of fault classes is of paramount importance in our endeavor.

The accuracy of extracted features or representations is an important prerequisite for improving the performance of machine learning algorithms [16]. Moreover, models built upon these features have limited diagnostic capabilities and insufficient generalization, making it challenging to effectively handle complex situations [17]. The presence of non-linear processing units in hierarchical structures enables deep learning methods to create high-level representations of data [18]. With the rapid advancement of computer hardware and the availability of vast amounts of data, deep learning has gradually replaced machine learning as the mainstream method of intelligent fault diagnosis. Within the realm of deep learning-based fault diagnosis, there are generally four categories: fault diagnosis based on autoencoders, fault diagnosis utilizing Restricted Boltzmann Machines (RBM), fault diagnosis through Convolutional Neural Networks (CNN), and fault diagnosis employing transfer learning. Due to the rapid development of pattern recognition technology represented by CNN, it has shown a strong ability to extract defect features from noise and vibration signals for fault diagnosis. Therefore, this paper also proposes a multi-scale network structure based on high-level semantic features based on convolutional neural networks.

Next, a detailed exposition of the development process of CNN in intelligent fault diagnosis is presented: In 2016, CNN was first utilized for bearing fault recognition. In order to achieve a better balance between training speed and accuracy, adaptive CNN (ADCNN) is proposed to dynamically change the learning rate [19]. Initially, one-dimensional time-domain raw data is superimposed into a two-dimensional vector form similar to image representation and then passed to the convolution layer for feature extraction. Directly using traditional CNN for processing vibration signals would lead to lengthy training times and redundant computational costs. To address this, the one-dimensional convolutional network (Conv1D) was introduced, which possesses fewer parameters and operational requirements. It can directly extract abstract features from raw vibration signals, making it suitable for fault diagnosis applications.

Subsequently, methods such as the Fast Fourier Transform were applied to convert one-dimensional vibration signals into spectrograms. The bearing fault diagnosis problem is effectively transformed into an image classification problem [20]. To mitigate high-frequency noise interference, a Wide First Layer Deep Convolutional Neural Network (WDCNN) was introduced. It directly processes raw vibration data, utilizing significant kernel convolution in the first layer to achieve data denoising [21]. Similar to this, Pythagorean Space Pooling (PSPP) was introduced as the first layer of the CNN, leading to improved diagnostic accuracy at varying rotational speeds [22]. To capture distinct signal resolutions, an improved Multi-Scale Cascaded Convolutional Neural Network (MC-CNN) was proposed, leveraging different-scale filters to effectively enhance signal information [23]. Additionally, apart from CNN-based methods, there are three other deep learning-based approaches capable of addressing fault diagnosis. These approaches are briefly showcased in the subsequent Table 1.

Table 1 Introduction to other deep learning-based methods

	Model and method	Main role
Fault Diagnosis Method Based on Autoencoders	Autoencoder (AE) [24]	Improve feature extraction ability by learning features in high-level representation.
	Denoising autoencoder (DAE) [25]	During training in AE, noise is added to the raw data to learn a more robust representation.
	Stack Denoising encoder (SDA) [26]	This guarantees the effectiveness of feature extraction and is suitable for robust feature extraction from different noisy signals.
	Autoencoder based on Extreme Learning Machine (AE-ELM) [27]	It solves the problem of high computational load of deep AE model, and the learning speed is fast, the computational load is small, and the classification accuracy is high.
	Competitive Sparse Autoencoder (WTA-AE) [28]	Solve the problem of sparse AE, apply to the sparse rate of any target, and train fast.
	Depth Wavelet autoencoder based on Extreme Learning Machine (DWAE-ELM) [29]	As a nonlinear activation function, wavelet function can capture signal features effectively.
Fault diagnosis method based on constrained Boltzmann machine	DBM [30]	Undirected, each layer captures complex, higher-order correlations between the hidden neurons of the previous layer, potentially learning some complex internal representations, and can adjust its feedback regulation from top to bottom, with better robustness.
	DBN [31]	Oriented, using layer by layer greedy training, so the distribution of the front layer does not depend on the back layer. It has a good effect of unsupervised clustering, and the computation amount is reduced compared with DBM.
Transfer learning based approach	—	Some commonalities and similarities exist between different but similar domains to help learn and optimize new tasks. [32]

The goal of new class discovery (NCD) is to infer new object classes in unlabeled data by learning prior knowledge of labeled data containing different but related classes, mainly using single-stage and two-stage methods [33].

The two-stage methods first focus on the labeled dataset and then investigate the unlabeled dataset. They can be further categorized into learning similarity functions in the labeled data and learning latent space. CCN [34], KCL [34], and MCL [35] belong to the former category. CCN addresses the cross-task transfer learning problem related to NCD. KCL found that similarity information can be passed in addition to features, so the category information is simplified into pair-to-pair constraints, and the network is trained by using KL divergence to calculate the difference in distribution between data pairs. The network trained on the labeled dataset (D^l) is then applied to perform clustering on the unlabeled dataset (D^u). MCL improves upon KCL by optimizing the loss function and introducing a new strategy for learning multi-class classification through binary decision problems. DTC [36] is a two-stage method that learns the latent space from D^l. It uses deep transfer clustering to discover new visual categories in a two-stage task. It first trains a model on D^l and then applies the model to D^u, utilizing the knowledge learned in D^l to simultaneously learn new class representations and clustering in D^u. On the other hand, single-stage methods, unlike two-stage methods, leverage both D^l and D^u and jointly cluster them. Single-stage methods tend to obtain better latent representations and are less biased toward known classes.

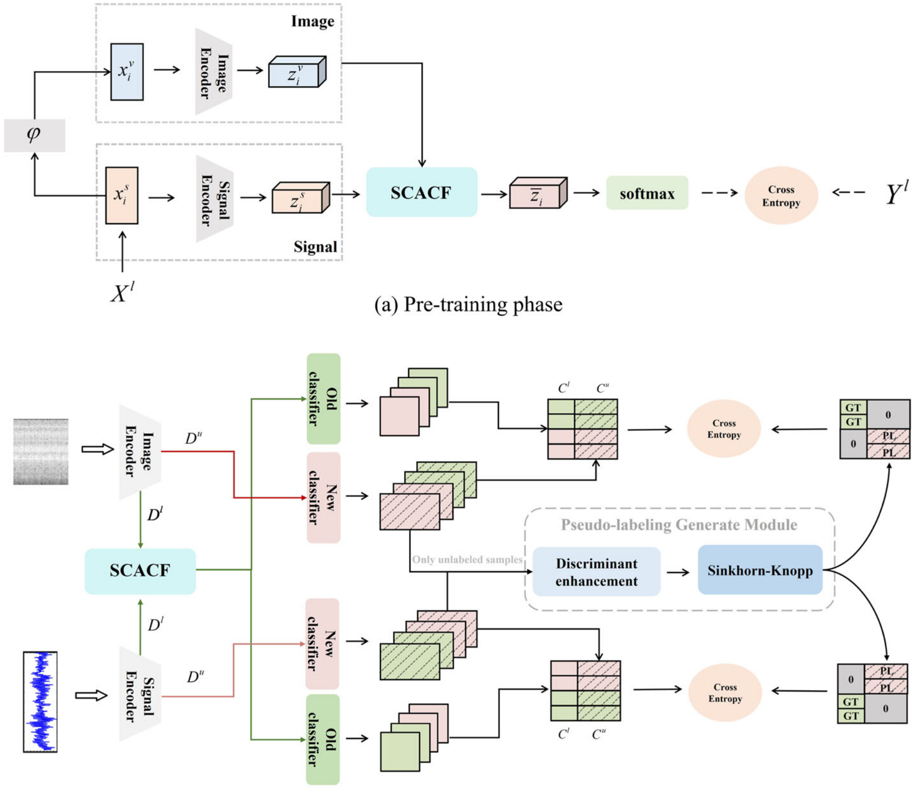
AutoNovel [37] is the first single-stage method to solve the NCD problem. It uses the RoNet [38] self-supervised learning (SSL) method to initialize the encoder and then adds classification and clustering networks on top of the encoder for joint learning on and. NCL [39] extends AutoNovel's loss by adding two contrastive learning terms to enhance the learning of discriminative representations. The first term involves supervised contrastive learning on labeled data with ground-truth labels, while the second term applies unsupervised contrastive learning to NCD on unlabeled data. RS [37, 40] addresses the issue that pre-training features only on labeled data can lead to biased features towards labeled data, which may be detrimental to clustering. Therefore, RS proposes to solve this problem by self-supervised pre-training on mixed labeled and unlabeled data, using the ranking statistical index as a pair of data similarity measures, and generating noisy pseudo-labels for training. UNO [41] breaks away from considering multiple objective functions for NCD and introduces a unified objective function for discovering new classes. UNO emphasizes the importance of high-quality pseudo-labels for unlabeled data and uses a unified cross-entropy loss for supervision, pioneering a new paradigm for NCD task learning. OpenMix [42] aims to enhance the robustness of pseudo-labels for unlabeled data by harnessing labeled data. It achieves this by generating novel training samples through a mixture of labeled and unlabeled samples, where pseudo-labels are established by combining the actual labels from labeled samples with cosine similarity scores calculated from unlabeled samples.

In contrast to the above methods, this approach proposes a multi-modal, multi-scale method. During the pre-training phase, a multi-scale deep encoder efficiently extracts features from different modalities, and a multi-modal information fusion mechanism is employed to merge the primary feature information from different modalities, improving the accuracy of novel class discovery.

3 Method

Given a mixed dataset $D = \{x_i, i = 1, \ldots, M\}$ containing both labeled and unlabeled data, our goal is to have these data automatically divided into $C = (C^l + C^u)$ different data categories. Additionally, we assume there is a labeled dataset | $D^l = \left\{\left(x_i^l, y_i^l\right), i = 1, \ldots, N\right\}$, where the class assignments $y_i^l = \{1, \ldots, C^l\}$ are known. The goal of automatic clustering of unlabeled data D^u is realized by learning the feature expression of data in labeled data set.

For this purpose, we propose a multi-modal-based method for discovering unknown faults. Figure 1 illustrates the overall architecture of the method, which consists of two stages: the pre-training stage and the new class discovery stage. First, the original vibration signals x_i^s are transformed into time-frequency domain images x_i^v using the Fast Fourier Transform φ, enabling the data augmentation to obtain the input data $x_i = (x_i^s, x_i^v)$. Then, the image deep encoder f_v and the signal deep encoderf_s encode the signal and image data into two feature vectors Z_i^s and Z_i^v, respectively. Subsequently, the data fusion is performed through the Saliency-Correlated Multi-Modal Representation Complementary Fusion Module, which retains salient features and data diversity from each modality. In the new class discovery stage, the method is used to identify and discover new classes, utilizing the mixed data as input



(a) Pre-training phase

(b) New class discovery phase

Fig. 1 Multi-modal fault diagnosis new class discovery architecture. The method includes two steps: pre-training and new class discovery. In the pre-training phase, the model is supervised trained on the labeled dataset using a multi-scale deep encoder to learn feature information from multi-modal samples, enhancing the pre-training model's generalization ability. In the new class discovery phase, the feature information of multi-modal data is extracted to improve the clustering accuracy of the unlabeled data. In this diagram, x_i^s represents labeled vibration signal data in the pre-training phase, x_i^v represents time-frequency domain image data, Y^l is the labels, φ represents the Fast Fourier Transform (FFT), "GT" stands for ground-truth label, and "PL" stands for pseudo label

data. In this stage, the present invention utilizes a pre-trained multi-scale deep encoder as the feature extractor. The labeled dataset D^l from the mixed data is used as the input to the encoder, and it goes through a SoftMax classification layer with C^l outputs to obtain the output F_{C^l}. The unlabeled dataset D^u is fed into the encoder and trained using a multi-layer perceptron. Finally, the SoftMax layer with C^u outputs is used for classification to obtain the output F_{C^u}. Subsequently, the two output features F_{C^l} and F_{C^u} are concatenated, and both labeled and pseudo-labeled data are used for training.

The end-to-end training of this model consists of two crucial parts: fully supervised training on the labeled dataset and unsupervised clustering training on the mixed dataset. The model used for unsupervised clustering is obtained through fully supervised training. Now, we will introduce the main components of the invention's framework.

3.1 Multi-modal learning and multi-modal representation complementary fusion module based on significance correlation (MMCF-SC)

In complex problems, single-modal learning may encounter difficulties because it provides limited information and is constrained by the data it receives. It may also suffer from noise, distortion, and other issues. In contrast, multi-modal learning can extract different forms of feature information across modalities and fuse information from multiple data sources. This enables the extraction of deeper and more enriched semantic features, making full use of the correlations and complementarity among different modalities. As a result, multi-modal learning achieves more efficient feature representation and learning compared to single-modal learning. In the field of fault diagnosis, x_i^s represents vibration signals, x_i^v represents time-frequency domain images, z_i^s and z_i^v are the feature vectors extracted from the signals and images, respectively, using encoders. For tasks that require analyzing signal features over time, z_i^s contains more desired feature information. On the other hand, tasks that focus on signal frequency features, z_i^v might be more relevant. Through multi-modal learning, feature vector $z_i = \left[z_i^v, z_i^s\right]$ is obtained, which encompasses all the major feature information from modalities z_i^s and z_i^v. This enables the model to handle both types of tasks mentioned above simultaneously, providing more detailed feature information than single-modal approaches and enhancing the model's performance in complex environments.

In the multimodal data generation stage, the fast Fourier algorithm is used to convert the vibration signal into time-frequency image.

In the multi-modal representation fusion phase, it is common to use concatenation or feature addition to integrate the data. While this approach preserves the original information from each modality and provides a more comprehensive and rich feature representation, it also introduces several issues: (a) High dimensionality: The feature dimension significantly increases, leading to complexity in computation and storage. (b) Feature weight balance: Different modalities have varying importance and expressive abilities, and simple addition or concatenation cannot handle the weight relationships between features. (c) Feature redundancy: Different modalities may contain redundant information, and direct concatenation or addition can result in feature redundancy, reducing feature discriminability and generalization. To address these issues, a multi-modal representation complementary fusion module based on significance association is proposed. This module improves the effectiveness and robustness of fusion results by considering the importance and relevance of features in the fusion process.

The fusion module shown in Fig. 2 is divided into two parts: data saliency-related alignment and multi-modal data balanced complementary fusion. After the previous feature extraction step, we obtain the vibration signal feature vector $z_i^s \in \mathbb{R}^{m \times d}$ and the time-frequency domain image feature $z_i^v \in \mathbb{R}^{n \times d}$ (where m = n). In the data alignment part, we first calculate the original correlation matrix S_{origin} between the vibration signal and the time-frequency domain image by performing the transpose crossmultiplication of z_i^v and z_i^s:

$$S_{\text{origin}} = z_i^v \cdot \left(z_i^s\right)^T , \tag{1}$$

After obtaining the original correlation matrix E, we apply a sparsity mechanism to obtain the sparsity-aligned correlation matrix S_{mask} . In this sparsity process, we set the elements s_{ij} in the original correlation matrix E that are greater than or equal to 0.5 to 1, and the elements s_{ij} that are less than 0.5 to 0. This is done to emphasize the relevant vectors and mask out

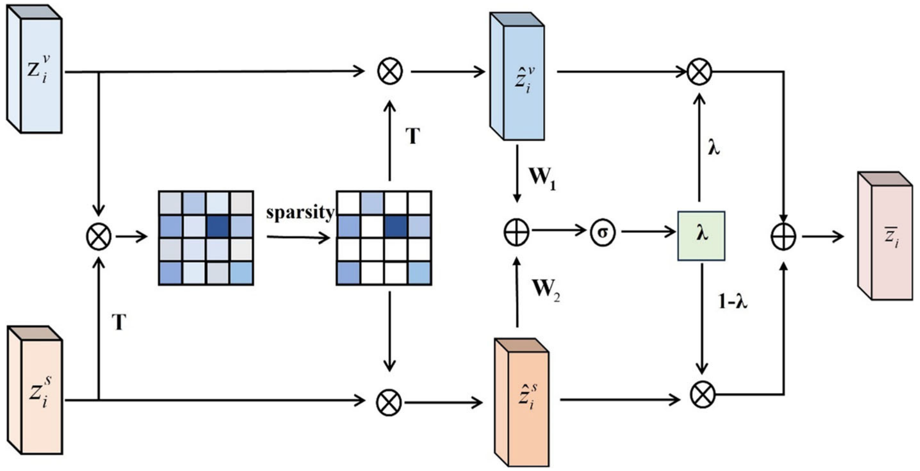


Fig. 2 Multi-modal representation complementary fusion module based on saliency correlation

irrelevant vectors in the correlation matrix.

$$s_{ij} = \begin{cases} 1, & \mathrm{s}_{ij} \geq 0.5 \\ 0, & \text{otherwise} \end{cases}, \tag{2}$$

By utilizing the aligned correlation matrix, we achieve the alignment of multi-modal data and obtain the aligned feature vectors:

$$\begin{aligned} \hat{z}_i^v &= z_i^v \cdot \left(s_{ij}\right)^T, \\ \hat{z}_i^s &= z_i^s \cdot s_{ij}, \end{aligned} \tag{3}$$

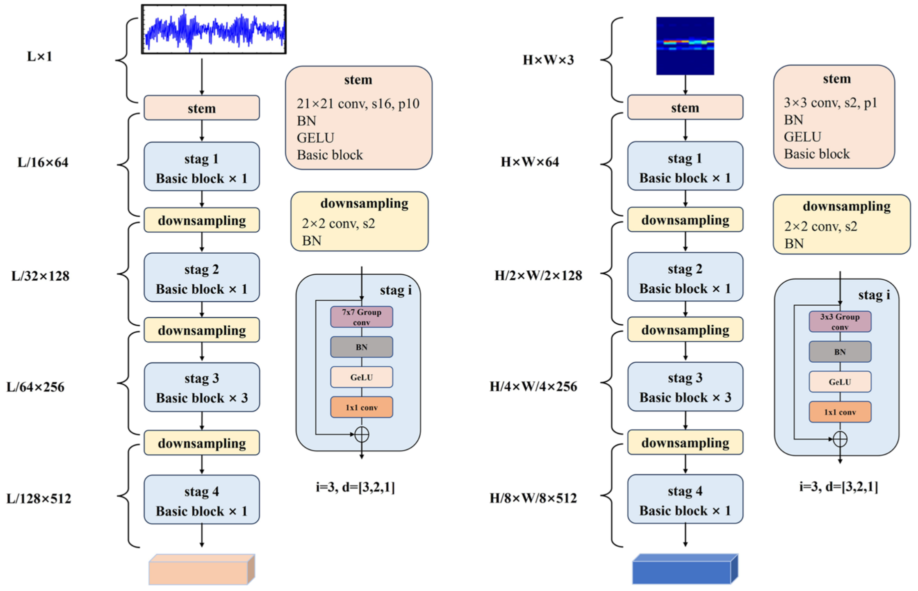
After achieving saliency-related alignment of multi-modal data, the aligned feature vectors are subjected to modality data balancing and complementary fusion. This is achieved by learning weight parameters that control the contribution of each modality's feature during the fusion process. For each modality's feature vector, a weight value is computed, indicating the importance of that modality's feature in the final fusion result. Modality features with higher importance are assigned larger weights, while those with lower importance are assigned smaller weights, thereby enhancing the quality and effectiveness of the fusion.

$$\begin{aligned} \lambda &= \text{Sigmoid}\left(\mathbf{W}_1 \hat{z}_i^v + \mathbf{W}_2 \hat{z}_i^s\right), \\ \bar{z}_i &= (1-\lambda)\hat{z}_i^s + \lambda \hat{z}_i^v, \end{aligned} \tag{4}$$

Where, in this process, W_1, W_2 represents the learned parameter matrix, and λ represents the corresponding weight coefficients. The $\overline{z_i}$ represents the final fused multi-modal feature vector after alignment, which contains a richer information representation. This improves the robustness and expressive power of the features, allowing the model to effectively leverage the complementary information from different modalities for enhanced performance in the subsequent tasks.

3.2 Multi-scale depth encoder

This invention extracts features from different modal samples at multiple scales in the phase of higher-level feature representation that possesses more semantic information. This effectively mitigates noise interference and information loss in complex environments. Inspired by the swin-transformer [43], both the image deep encoder and the signal deep encoder adopt a 1:1:3:1 deep encoding structure, as illustrated in Fig. 3. Taking signal data as an example, for vibration signal data $x_i^s = (x_{i,1}^s, x_{i,2}^s, \ldots, x_{i,n}^s)$, at each layer of the convolutional network, local structured operations are employed instead of traditional convolutional operations. This enhances the model's feature extraction capability, reduces computational complexity, and improves robustness. Furthermore, in the deep stage, by applying feature extraction with different scales k_1, k_2, k_3 to high-level features rich in semantic information, the feature representation is further refined, perceptual capability is enhanced, and semantic denoising of vibration signals is achieved. In the deeper stage with more semantic information, features are extracted at different semantic scales using dilation convolution with varying dilation factors $(3, 2, 1)$. These features are then mapped to the same feature space together with image features extracted by the image deep encoder, achieving a multi-modal complementary fusion based on salient correlations. The distinction between the image encoder and the signal encoder lies not only in the different convolutional approaches used but also primarily in how they process the data. In addition to the variation in convolution methods, the most significant difference is in their data processing. Due to the larger size of vibration signal samples, during the initial



(a) Signal depth encoder

(b) Image depth encoder

Fig. 3 Multi-scale high-level semantic feature deep encoder model diagram. Primarily composed of a Signal Deep Encoder f_s and an Image Deep Encoder f_v, wherein the Signal Deep Encoder employs a large convolutional kernel for feature extraction in the first convolutional layer. Each layer of both encoders undergoes local region partitioning, and varying scales of feature extraction are applied to high-level feature representations rich in semantic information

convolution, larger convolutional kernels are applied in the signal encoder. Furthermore, in each localized structural processing step, the convolutional kernels used in the signal encoder are also larger compared to those used in the image encoder. This larger kernel size in the signal encoder serves the purpose of noise reduction.

The multi-scale feature extraction for the data primarily manifests in the deep semantic features, while shallow-level features mainly capture the fundamental characteristics of the signal. In the context of complex signals and fault diagnosis, deep semantic features are more discriminative. This paper strengthens the multi-scale enhancement of high-level features with semantic information, effectively improving the model's feature extraction capability.

3.3 Pseudo-label generation module

In the main direction of this invention, we have designed a pseudo-label generation module suitable for multi-modal fault discovery, as shown in Fig. 4, which consists of two main components: the Discriminative Enhancement Module and the Sinkhorn-Knopp [44] Module. The Discriminative Enhancement Module leverages the complementary characteristics of multiple modal information sources to optimize the discriminative vector logits from various perspectives, enhancing the saliency of inter-class discriminative relationships in the logits vector.

This means that the Discriminative Enhancement Module explores the differences between classes in multiple modal features, thereby making the generated pseudo-labels more discriminative and better able to distinguish between different class samples. By optimizing the logits, this module ensures more accurate and reliable generation of pseudo-labels, contributing to improved performance in the discovery of faults with unknown classes.

Vibration signal features z_i^s and time-frequency domain image features z_i^v are processed through the new class classification head to obtain their discriminative vectors l_i^s and l_i^v.

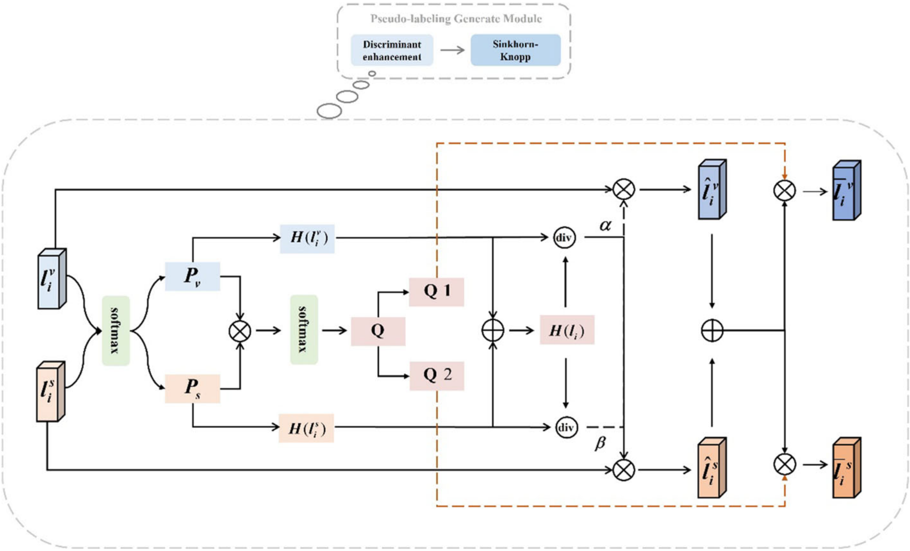


Fig. 4 Discriminative enhancement module (DE)

Then, their discriminative probability distributions are calculated separately:

$$\begin{aligned} P_v &= \text{Softmax}\left(l_i^v\right) = \{p_1, p_2, \ldots, p_{C^n}\}, \\ P_s &= \text{Softmax}\left(l_i^s\right) = \{p_1, p_2, \ldots p_{C^n}\}, \end{aligned} \tag{5}$$

Where, C^u represents the number of categories in the new class, we first consider the discriminative probability relationships of samples among different categories from multiple modalities and perspectives, aiming to enhance the saliency of the discriminative vectors:

$$Q = \text{Softmax}\left(P_v \cdot P_s\right). \tag{6}$$

For the discriminative relationship vectors, which are low-dimensional vectors, a smaller information entropy indicates better stability, robustness, and noise tolerance. Therefore, based on the information entropy of each modality, we determine the weights of different modalities for the fusion of their discriminative relationship vectors.

$$H\left(l_i^v\right) = -\sum_{i=1}^{C^u}\left(p_i^v\left(\log\left(p_i^v\right)\right)\right), \tag{7}$$

$$H\left(l_i^s\right) = -\sum_{i=1}^{C^u}\left(p_i^s\left(\log\left(p_i^s\right)\right)\right), \tag{8}$$

$$H\left(l_i\right) = H\left(l_i^v\right) + H\left(l_i^s\right), \tag{9}$$

$$\alpha = H\left(l_i^v\right)/H\left(l_i\right), \tag{10}$$

$$\beta = H\left(l_i^s\right)/H\left(l_i\right), \tag{11}$$

$$\hat{l}_i = \alpha l_i^s + \beta l_i^v, \tag{12}$$

where, $\alpha + \beta = 1$. Finally, we use the inter-class discriminative probability relationship as the inter-modal weights for the fused discriminative relation vector, optimizing to obtain the final logits vector with significantly enhanced discriminative relations:

$$\overline{l_i} = Q \cdot \hat{l}_i. \tag{13}$$

Formula (13) can be applied to each view individually, but it does not recommend predicting the consistency of the output of different views. Therefore, we perform a secondary transformation on Q by multiplying the value of the largest element in Q by a certain factor and reducing the values of the remaining elements by another factor. This is done to enhance the saliency of output predictions, resulting in distinct predictions for another view:

$$Q_{\text{transform}} = \begin{cases} 10 * Q_{\max}, & Q_{\max} = \max\left(Q_i\right) \\ Q_i/10, & \text{otherwise} \end{cases} \tag{14}$$

$$\tilde{l}_i = Q_{\text{tranform}} \cdot \hat{l}_i \tag{15}$$

In the pseudo-label assignment module, We adopt the approach from the UNO method; for the case where logits are equal to each other [45, 46], the Sinkhorn-Knopp algorithm is used. An entropy term is added to penalize situations where all logits are equal and encourage the unified assignment of pseudo-labels for all clusters C^u. Let $\mathbf{L} = \left[\bar{l}_i{}^1, \bar{l}_i{}^2, \ldots, \bar{l}_i^B\right]$ be the matrix computed for the new class heads with a size of B for the samples, and let $\tilde{\mathbf{Y}} = [y_1, y_2, \ldots, y_B]$ be the matrix of current batch's unknown pseudo-labels. The solution $\tilde{\mathbf{Y}}$ is obtained as follows:

$$\tilde{\mathbf{Y}} = \max_{\mathbf{Y}\in\Gamma} \text{Tr}(YL) + \varepsilon H(\mathbf{Y}), \tag{16}$$

Where H is the entropy function. Tr denotes the trace function, $\varepsilon > 0$ is a hyperparameter, and Γ is the transportation polytope, defined as:

$$\Gamma = \left\{ \mathbf{Y} \in \mathbb{R}_{+}^{C^u \times B} \mid \mathbf{Y}\mathbf{1}_B = \frac{1}{C^u} 1_{C^u}, \mathbf{Y}^T \mathbf{1}_{C^u} = \frac{1}{B} \mathbf{1}_B \right\}, \tag{17}$$

The generated pseudo-labels are represented by each row of y_i in $\tilde{\mathbf{Y}}$.

4 Experiment

4.1 Data sets

CWRU dataset [47] The CWRU dataset is a dataset developed by researchers from the Department of Mechanical and Aerospace Engineering at Case Western Reserve University in the United States. It is designed for studying the condition monitoring of mechanical bearings. The dataset includes vibration data from normal bearings and faulty bearings at the drive end. The experiment collected data on the normal bearing signal and the drive end fault signal at speeds of 12,000 samples per second and 48,000 samples per second, respectively. The data file is in Matlab format and contains fan and drive vibration data as well as motor speed. The points of failure were manufactured in 5 different sizes: 7 mils, 14 mils, 21 mils, 28 mils, and 40 mils (1 mil = 0.001 inch), and the load was divided into 0 HP, 1 HP, 2 HP, and 3 HP under each failure. The fan fault data collection speed is 12,000 samples per second. The outer raceway faults are located at different locations and tested at 3 oclock, 6 oclock, and 12 oclock, respectively. This paper is based on the driver fault data and the normal data set under 12k sampling. The faults under all loads of different diameters, such as 0.007 inch, 0.014 inch, and 0.028 inch at the 6 oclock position of the rolling body fault (ball), InnerRace fault, and OutRace fault, are divided into nine fault types, which form ten different fault categories with the normal type.

PU dataset [48] The data set is the 6203 bearing data set obtained from Paderborn University, which includes both human-induced and real damage cases. A total of 32 different bearing experiments were carried out: 12 bearings were used as artificially damaged bearings, 14 bearings were damaged from accelerated life tests, and 6 bearings were in a healthy state. The dataset utilizes electrical discharge machining (EDM), drilling, and manual electric engraving to perform artificial damage and generate real damage samples through accelerated life tests. According to different damage combinations, multiple losses, and damage degrees, the PU data set is divided into 20 different fault types.

4.2 Experimental indicators

ACC We use the average clustering accuracy as the primary evaluation metric. It is defined as follows:

$$ACC = \max_{\text{pred} \in P} \frac{1}{N} 1 \left\{ y_i = \text{pred} \left(\hat{y}_i \right) \right\}, \tag{18}$$

Where, y_i and $\hat{y}_i$ represent the ground-truth lapels and clustering predictions of the unlabeled data samples in $x_i^u \in D^u$, respectively. P denotes all possible permutations calculated using the Hungarian algorithm [49].

NMI (Normalized Mutual Information) Mutual Information is a useful measure of information in information theory, which can be viewed as the amount of information contained in one random variable about another random variable. The mutual information I(X; Y) is the relative entropy of the joint distribution p(x, y) and the product of the marginal distributions $p(x)$ and $p(y)$. It is defined by the following formula:

$$I(X;Y) = \sum_{x}\sum_{y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{19}$$

NMI normalizes the mutual information and is calculated by the formula:

$$NM(X;Y) = 2\frac{I(X;Y)}{H(x)+H(Y)}, \tag{20}$$

Where H(x) and H(y) are information entropy.

ARI The Adjusted Rand Index (ARI) is an external evaluation metric used to assess the performance of clustering algorithms. It measures the similarity between the clustering results and the true labels. ARI is obtained by improving upon the Rand Index (RI), which takes values in the range of [0, 1]. A higher ARI value indicates a better alignment between the clustering results and the ground truth.

$$RI = \frac{TP+TN}{TP+FP+TN+FN}, \tag{21}$$

The value of ARI ranges from -1 to 1. The closer the value is to 1, the better the clustering result is; the closer the value is to 0, the more random the clustering result is; and the closer the value is to -1, the worse the clustering result is. The formula is as follows:

$$ARI = \frac{RI - E(RI)}{\max(RI) - E(RI)} \tag{22}$$

4.3 Contrast test

In order to verify the superiority of the framework proposed in this paper, two sets of experiments were carried out on the CWRU data set and the PU data set, respectively, with 10 kinds of models. The number of categories of new and old classes was set at 4 and 6, respectively, on the CWRU data set. Set the number of new class categories to 10 and the number of old class categories to 10 on the PU dataset. The experimental results are shown in Table 2. For a clearer visual display, we converted the tabular data into the graphical data shown in Fig. 5. In the comparative experiments for novel class discovery, we conducted a pre-training phase of 20 epochs and a discovery training phase of 100 epochs for all experiments. We used a learning rate of 0.4 throughout the training process. Both the UNO method and our proposed method utilized 20 heads.

In the above table, the first four methods belong to unsupervised learning approaches, while the remaining six are all related to novel class discovery methods. These six novel class methods were originally designed for the field of natural image analysis. However, we have reimagined and adapted their underlying concepts to the domain of fault diagnosis to discover unknown faults. Due to the differences in data sources, some models exhibit

Table 2 New class discovery and comparison of unsupervised clustering methods based on CWRU and PU

Method(Param)	Number of new classes=4 (CWRU)			Number of new classes=10 (PU)		
	ACC	NMI	ARI	ACC	NMI	ARI
K-means (-)	36.82±0.50	15.62±0.10	8.34±0.15	8.38±0.15	13.35±0.15	2.21±0.15
Affinity Propagation (-)	0.01±0.10	28.55±0.15	0.28±0.15	0.0005±0.00	48.25±0.50	0.88±0.01
DBSCAN (-)	0.33±0.10	39.10±0.10	0.00±0.00	0.01±0.01	32.40±0.50	0.00±0.00
HDBSCAN (-)	16.00±0.01	33.43±0.50	14.35±0.50	20.00±0.25	4.22±0.25	0.15±0.15
DTC (7.34M)	99.96±0.20	99.96±0.20	99.96±0.20	63.71±2.15	69.55±2.45	48.49±2.75
MCL (0.11M)	79.00±1.80	83.91±1.80	70.96±1.80	44.96±1.25	54.16±1.55	30.41±2.55
KCL (0.11M)	75.00±0.50	85.71±1.50	71.42±0.50	53.83±1.25	62.02±1.55	41.54±2.55
NCL (11.03M)	95.83±1.20	90.48±2.20	89.74±2.20	64.98±1.75	66.20±1.55	48.42±2.55
UNO (19.5M)	99.98±0.10	99.98±0.10	99.98±0.10	65.10±1.25	60.04±2.15	45.10±2.55
ours (56.6M)	**99.99±0.10**	**99.99±0.10**	**99.99±0.10**	**77.10±1.80**	**80.47±2.10**	**66.53±2.70**

The boldface in the table is to highlight the effect achieved by our model, and also to highlight the best effect

relatively less impressive performance on fault signals compared to their performance on natural images.

As depicted in the charts, while unsupervised algorithms are capable of clustering unlabeled data without labeled information, their performance is comparatively inferior. On the other hand, novel class discovery methods yield more favorable outcomes. Among these novel class discovery methods, our proposed model demonstrates superior performance on both the CWRU dataset and the PU dataset. In the case of the CWRU dataset, which is relatively simpler with lower fault recognition difficulty, most models exhibit satisfactory performance, leading to limited enhancement in our model's performance. Conversely, the PU dataset comprises a wide range of fault types and complex scenarios, making fault recognition more challenging. Consequently, our model shows significant improvement in such complex scenarios, indicating that it outperforms existing alternatives in handling intricate situations.

In conclusion, when facing complex scenarios, our model demonstrates superior performance compared to existing models, as evidenced by the results obtained on both the CWRU dataset and the PU dataset.

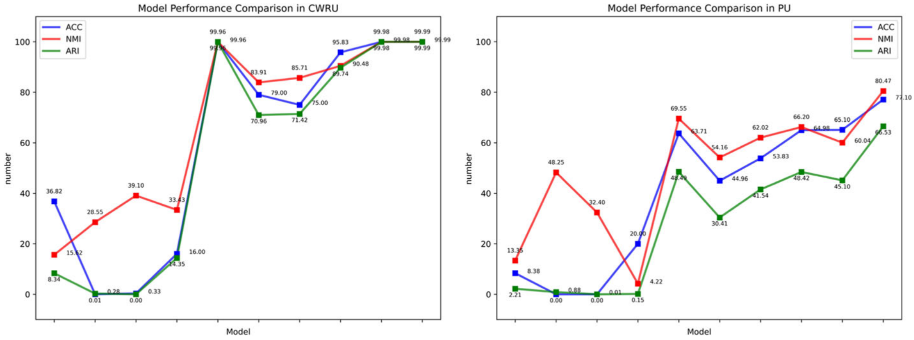


Fig. 5 Line graph of new class discovery and unsupervised clustering methods based on CWRU and PU

Table 3 Ablation study1: Performance comparison of independent functional modules on the PU dataset

Method	MMCF-SC	DE	Model	ACC	NMI	ARI
ISF2EM	✓	✗	✗	71.60	75.98	61.61
	✓	✗	✓	71.80	74.31	59.79
	✓	✓	✗	72.90	75.30	60.30
	✗	✓	✓	74.90	75.92	61.87
	✗	✓	✗	67.00	71.60	54.04
	✗	✗	✓	71.10	77.37	60.79
	✗	✗	✗	69.60	72.91	57.87
	✓	✓	✓	**77.10**	**80.47**	**66.53**

The boldface in the table is to highlight the effect achieved by our model, and also to highlight the best effect

4.4 Ablation experiment

To validate the superiority of the proposed functional modules, we conducted ablation experiments by running 60 epochs on the PU dataset with a new class count of 10 and an old class count of 10. Our validation process focused on two aspects. Firstly, we examined the impact of the presence or absence of each module on the overall model architecture. Secondly, we substituted equivalent methods for each module based on the specific problem they addressed, aiming to compare the superiority of these modules against alternative methods.

4.4.1 Ablation study

ISF2EM represents the image and signal series feature fusion enhancement method proposed in our work, MMCF-SC represents our proposed multi-modal complementary fusion module based on significance correlation to fusion multi-modal data, DE represents our discriminant relationship enhancement module in the pseudo-label generation stage, and model represents our proposed feature extractor architecture. It can be seen from the data in Table 3 that the proposed method has an obvious effect on performance improvement, and the evaluation indicators ACC, NMI, and ARI have improved by about 7.5

4.4.2 Contrast ablation

Feature extractor ablation In Table 4, our proposed feature extractor showed improvements over commonly used feature extractors such as ResNet18 and VGG in terms of evaluation

Table 4 Ablation study2: Performance of different feature extractors in PU data sets

	ACC	NMI	ARI
ResNet18	72.90	75.30	60.30
VGG	72.30	74.83	59.22
ours	**77.10**	**80.47**	**66.53**

The boldface in the table is to highlight the effect achieved by our model, and also to highlight the best effect

Table 5 Ablation study3: The effect of multimodal fusion in the pre-training stage

	ACC	NMI	ARI
Option 1	70.90	76.00	59.90
Option 2	69.90	73.82	57.51
Option 3	67.50	78.01	62.13
Ours	**77.10**	**80.47**	**66.53**

The boldface in the table is to highlight the effect achieved by our model, and also to highlight the best effect

metrics ACC, NMI, and ARI. Specifically, compared to ResNet, our feature extractor achieved an improvement of 4.2%, 5.17%, and 6.23% for ACC, NMI, and ARI, respectively. Compared to VGG, our feature extractor achieved improvements of 4.8%, 5.64%, and 7.31% for ACC, NMI, and ARI, respectively. This improvement can primarily be attributed to our consideration of semantic information within the deep network architecture as well as the balanced incorporation of both global and local information. As a result, our approach yields superior performance.

Multimodal fusion ablation In Table 5, “option1” represents the direct addition of signal and image data, “option2” indicates denoising alignment between image and signal data, and “option3” signifies direct adaptive fusion without alignment between image and signal data. From Fig. 6, it is evident that our proposed multimodal fusion approach based on saliency-driven cross-modal complementarity outperforms other multimodal fusion techniques.

Multimodal frame contrast ablation To validate the effectiveness of our proposed multimodal approach, we conducted two sets of experiments within our new class discovery framework, replacing the input data with vibration signals and time-frequency domain image data, respectively. The experimental results are presented in Table 6. Since time-frequency domain images are obtained from vibration signals through the fast Fourier transform, their purpose is to compensate for the lack of time-frequency information in the original signals.

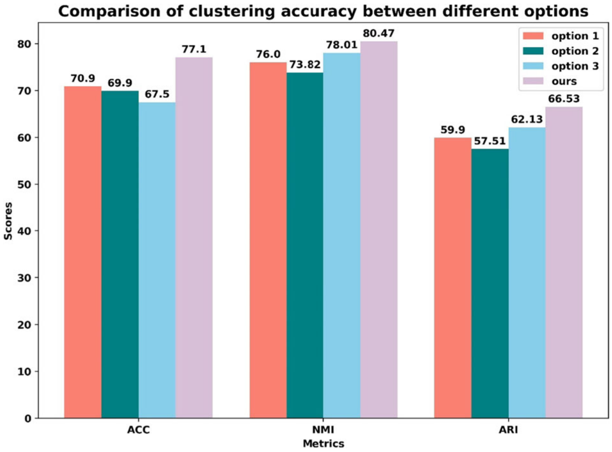


Fig. 6 Comparison of multimodal fusion methods

Table 6 Ablation study4: Comparison of the effects of multimode and single mode

	ACC	NMI	ARI
Only signal	73.30	72.72	59.36
Only image	63.20	60.70	44.15
Ours	**77.10**	**80.47**	**66.53**

The boldface in the table is to highlight the effect achieved by our model, and also to highlight the best effect

Consequently, using only time-frequency domain images yields inferior results compared to using only vibration signals.

However, when combining both approaches, the performance improvement is even more significant. It results in a 3.8% increase in ACC, 7.75% in NMI, and 7.17% in ARI compared to using only vibration signals. Moreover, compared to using only time-frequency image data, the combined approach yields a substantial enhancement, with a 13.9% increase in ACC, 19.77% in NMI, and 22.38% in ARI. The visual representation in Fig. 7 further highlights the evident advantages of this combined approach, providing strong evidence that our notion of utilizing multimodal information to compensate for the limitations of single-modal features is indeed valid.

5 Conclusion

In this paper, we propose a data diversity-based multi-modal fusion framework for new class discovery. This method learns recognition features from different data formats of the same fault type, leveraging the enriched properties of multi-modal feature information to improve clustering accuracy during the pre-training phase. Furthermore, in the new class discovery phase, we utilized the relative independence of discriminative relationships between single-modal information to guide each other, reducing the inter-class prediction fluctuations

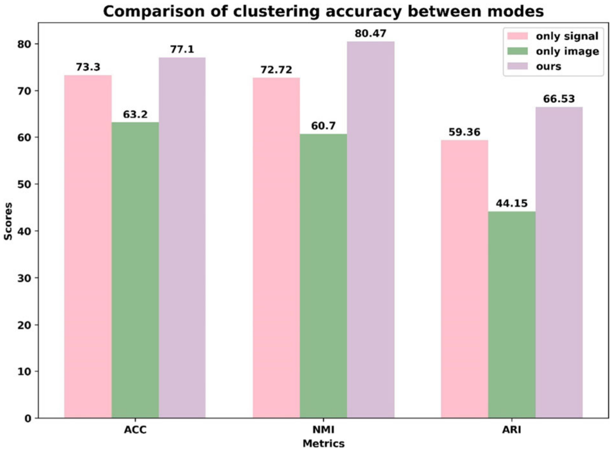


Fig. 7 Comparison of clustering accuracy among different modes

among new classes and further enhancing the model's advantages. On both the CWRU and PU datasets, our approach outperformed the latest NCD method significantly for different numbers of new classes, with the advantage becoming more pronounced in more complex scenarios. However, the way we interacted between modalities was straightforward, and we plan to conduct further research on multi-modal interaction methods to improve the model's performance.

Acknowledgements This work was supported in part by the National Key Research and Development Program of China (2020YFB1711700), the National Natural Science Foundation of China (62072418, 62172376), the Major Scientific and Technological Innovation Project of Shandong(2019JZZY020705), the Fundamental Research Funds for the Central Universities (202042008), and the Key Research and Development Program of Qingdao Science and Technology Plan (21-1-2-18-xx).

Data Availability The authors declare that all data used in the manuscript's experiments are stated in the manuscript.

Declarations

Conflicts of interest The authors declare that there is no conflict of interest.

References

1. Du W, Tao J, Li Y, Liu C (2014) Wavelet leaders multifractal features based fault diagnosis of rotating mechanism. Mech Syst Signal Process 43(1–2):57–75
2. Chen Z, Li W (2017) Multisensor feature fusion for bearing fault diagnosis using sparse autoencoder and deep belief network. IEEE Trans Instrum Meas 66(7):1693–1702
3. Arthur D, Vassilvitskii S (2006) k-means++: The advantages of careful seeding (Tech. Rep.). Stanford Infolab 8090:778. https://ilpubs.stanford.edu
4. Chang L, Wang L, Meng G, Xiang S, Pan C (2017) Deep adaptive image clusterin. In: Proceedings of the IEEE international conference on computer vision pp 5879-5887
5. Ghasedi Dizaji K, Herandi A, Deng C, Cai W, Huang H (2017) Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. In: Proceedings of the IEEE international conference on computer vision pp 5736-5745
6. Chapelle O, Scholkopf B, Zien A (2006) Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. IEEE Trans Neural Netw 20(3):542
7. Basu S (2002) Semi-supervised clustering by seeding. In: Proc. ICML-2002
8. Callut J, Françoisse K, Saerens M, Dupont P (2008) Semi-supervised classification from discriminative random walks. In: Joint european conference on machine learning and knowledge discovery in databases. Springer, pp 162-177
9. Mehrkanoon S, Alzate C, Mall R, Langone R, Suykens JA (2014) Multiclass semisupervised learning based upon kernel spectral clustering. IEEE Trans Neural Netw Learn Syst 26(4):720–733
10. Pan S, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Discov Data Eng 22(10). IEEE press
11. Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training
12. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners. OpenAI blog 1(8):9
13. Brown T et al (2020) Language models are few-shot learners. Adv Neural Inf Process Syst 33:1877–1901
14. Nie W, Jiao C, Qu L, Liu AA (2023) CPG3D: Cross-modal Priors Guided 3D Object Reconstruction. IEEE Trans Multimed
15. Nie W, Bao Y, Zhao Y (2023) Long dialogue emotion detection based on commonsense knowledge graph guidance. IEEE Trans Multimed
16. Bengio Y, Courville A, Vincent P (2013) Representation learning: A review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35(8):1798–1828

17. Tang J, Deng C, Huang G-B (2015) Extreme learning machine for multilayer perceptron. IEEE Trans Neural Netw Learn Syst 27(4):809–821
18. Guo Y, Wu Z, Ji Y (2017) A hybrid deep representation learning model for time series classification and prediction. In: 2017 3rd International conference on big data computing and communications (BIGCOM). IEEE, pp 226-231
19. Janssens O et al (2016) Convolutional neural network based fault detection for rotating machinery. J Sound Vib 377:331–345
20. Huijie Z, Ting R, Xinqing W, You Z, Husheng F (2015) Fault diagnosis of hydraulic pump based on stacked autoencoders. In: 2015 12th IEEE international conference on electronic measurement & instruments (ICEMI), vol. 1, pp 58-62. IEEE
21. Zhang W, Peng G, Li C, Chen Y, Zhang Z (2017) A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. Sensors 17(2):425
22. Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Commun ACM 60(6):84–90
23. Huang W, Cheng J, Yang Y, Guo G (2019) An improved deep convolutional neural network with multi-scale information for bearing fault diagnosis. Neurocomputing 359:77–92
24. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323(6088):533–536
25. Vincent P, Larochelle H, Bengio Y, Manzagol P-A (2008) Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine learning, pp 1096-1103
26. Hinton GE, Osindero S, Teh Y-W (2006) A fast learning algorithm for deep belief nets. Neural Comput 18(7):1527–1554
27. Mao W, He J, Li Y, Yan Y (2017) Bearing fault diagnosis with auto-encoder extreme learning machine: A comparative study. Proc Institut Mech Eng, Part C: J Mech Eng Sci 231(8):1560–1578
28. Makhzani A, Frey BJ (2015) Winner-take-all autoencoders. Adv Neural Inf Process Syst 28
29. Haidong S, Hongkai J, Xingqiu L, Shuaipeng W (2018) Intelligent fault diagnosis of rolling bearing using deep wavelet auto-encoder with extreme learning machine. Knowl-Based Syst 140:1–14
30. Salakhutdinov R, Hinton G (2009) Deep boltzmann machines," in Artificial intelligence and statistics, pp 448-455. PMLR
31. Hinton GE (2009) Deep belief networks. Scholarpedia 4(5):5947
32. Lu W, Liang B, Cheng Y, Meng D, Yang J, Zhang T (2016) Deep model based domain adaptation for fault diagnosis. IEEE Trans Ind Electron 64(3):2296–2305
33. Troisemaine C, Lemaire V, Gosselin S, Reiffers-Masson S, Flocon-Cholet S, Vaton S (2023) Novel class discovery: an introduction and key concepts. arXiv:2302.12028
34. Hsu Y-C, Lv Z, Kira Z (2017) Learning to cluster in order to transfer across domains and tasks. arXiv:1711.10125
35. Hsu Y-C, Lv Z, Schlosser J, Odom P, Kira Z (2019) Multi-class classification without multi-class labels. arXiv:1901.00544
36. Han K, Vedaldi A, Zisserman A (2019) Learning to discover novel visual categories via deep transfer clustering. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 8401-8409
37. Han K, Rebuffi S-A, Ehrhardt S, Vedaldi A, Zisserman A (2021) Autonovel: Automatically discovering and learning novel visual categories. IEEE Trans Pattern Anal Mach Intell 44(10):6767–6781
38. Gidaris S, Singh P, Komodakis N (2018) Unsupervised representation learning by predicting image rotations. arXiv:1803.07728
39. Z. Zhong, E. Fini E, Roy S, Luo Z, Ricci E, Sebe N (2021) Neighborhood contrastive learning for novel class discovery. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10867-10875
40. Han K, Rebuffi S-A, Ehrhardt S, Vedaldi A, Zisserman A (2020) Automatically discovering and learning new visual categories with ranking statistics. arXiv:2002.05714
41. E. Fini, E. Sangineto, S. Lathuiliere, Z. Zhong, M. Nabi, and E. Ricci (2021) A unified objective for novel class discovery. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9284-9292
42. Zhong Z, Zhu L, Luo Z, Li S, Yang Y, Sebe N (2021) Openmix: Reviving known knowledge for discovering novel visual categories in an open world. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9462-9470
43. Z. Liu et al. (2021) Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF international conference on computer vision, pp 10012-10022
44. (2013) Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems 26

45. Caron M, Misra I, Mairal J, Goyal P, Bojanowski P, Joulin A (2020) Unsupervised learning of visual features by contrasting cluster assignments. Adv Neural Inf Process Syst 33:9912–9924
46. Asano YM, Rupprecht C, Vedaldi A (2019) Self-labelling via simultaneous clustering and representation learning. arXiv:1911.05371
47. Bearing Data Center accessed on Mar. 2017. [Online]
48. Paderborn University. Mechanical Fault Diagnosis Dataset. [Online]
49. Kuhn HW (2005) The Hungarian method for the assignment problem. Nav Res Logist (NRL) 52(1):7–21

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Di Niu[1] · Shusong Yu[1] · Jiali Xu[2] · Chenglong Wang[1] · Ruoxi Li[1] · Xiangqian Ding[1]

Di Niu
niudi@stu.ouc.edu.cn

Chenglong Wang
wangchenglong@stu.ouc.edu.cn

Ruoxi Li
liruoxi@stu.ouc.edu.cn

Xiangqian Ding
dingxq@ouc.edu.cn

[1] Faculty of Information Science and Engineering, Ocean University of China, Qingdao, Shandong, China

[2] QNLM: Laoshan Laboratory, Qingdao, China