Prior Knowledge-Augmented Self-Supervised Feature Learning for Few-shot Intelligent Fault Diagnosis of Machines

Tianci Zhang, Jinglong Chen, *Member, IEEE*, Shuilong He, and Zitong Zhou

***Abstract*—Data-driven intelligent diagnosis models expect to mine the health information of machines from massive monitoring data. However, the size of faulty monitoring data collected in engineering scenarios is limited, which leads to few-shot fault diagnosis as a valuable research point. Fortunately, it is possible to reduce the required amount of training data by integrating prior diagnosis knowledge into diagnosis models. Inspired by this, we present a prior knowledge-augmented self-supervised feature learning framework for few-shot fault diagnosis. In the framework, 24 signal feature indicators are built to form prior features set based on existing diagnosis knowledge. Besides, a convolutional auto-encoder is used to mine the general features, which are considered to potentially contain fault information that prior features do not possess. We design a self-supervised learning scheme for training the diagnosis model, which enables the model to learn both prior and general features served as proxy labels. As a result, the model is expected to mine richer features from limited monitoring data. The effectiveness of the proposed framework is verified using two mechanical fault simulation experiments. From the angle of prior diagnosis knowledge, the proposed framework provides a new perspective to the problem of few-shot intelligent diagnosis of machines.**

***Index Terms*—Intelligent fault diagnosis, few-shot learning, prior knowledge, self-supervised learning**

I. Introduction

MACHINE fault diagnosis plays a significant role in prognostics health management (PHM) by linking the machine monitoring data and its health status [1]. Deep learning provides a powerful tool for machine fault diagnosis, which enables diagnosis models to handle massive monitoring data and automatically output the health status of machines [2]–[4]. Due to the complex network structure and parameters, intelligent diagnosis models based on deep learning rely on large amounts of machine fault data for model training, and their diagnosis performance is usually positively correlated with the amount of training data [5]. However, in engineering scenarios, machines work in a normal condition for a long time and failures rarely occur. Although condition monitoring systems can continuously record machine working data, the size of the collected fault data is very limited. With insufficient fault data, the trained intelligent diagnosis model has weak generalization ability, which makes it difficult to achieve satisfactory application results in engineering scenarios.

Few-shot fault diagnosis aims to use a small number of fault data samples to train diagnosis models. In recent years, scholars have achieved many remarkable results in this problem [5]. Existing research addressed it from three main perspectives. The first one is to augment the limited fault data [6]–[8]. For example, Zhang et al. [7] proposed a fault data augmentation method based on generative adversarial networks, which can generate fault data with similar distribution to real fault data. The experimental results show that data augmentation helps to improve the diagnosis performance of the model in the case of small samples. The second one is to improve the utilization of fault samples or to mine more information from limited fault samples [9]–[11]. For example, Ren et al. [9] constructed a capsule auto-encoder-based diagnosis model for fault detection using few samples, in which various fault feature capsules are extracted from limited fault data samples. The third one is to build diagnosis models based on transfer learning [12]–[14]. For instance, He et al. [14] proposed an intelligent diagnosis scheme based on a deep transfer auto-encoder, which can learn important fault features from a large related vibration dataset and transfer the parameter knowledge to the target but small dataset. However, data augmentation-based approaches consume large computational resources and are often hard to provide additional fault information for model training. Methods by improving the fault samples utilization have high requirements on the structural design of the diagnosis models, which usually increase the complexity of the models and are difficult to achieve the optimal states. Transfer learning-based schemes require the construction of source domain datasets that are large enough and relevant to the target diagnosis task, which consumes much manual labor.

Recently, some latest studies in fields such as computer

This work was supported financially by the National Natural Science Foundation of China (No. 91960106, No. 51875436, No. U1933101), China Postdoctoral Science Foundation (No. 2020T130509, No. 2018M631145), Liuzhou Natural Science Foundation (No. 2021AAA0112), and Fundamental Research Funds for the Central Universities (No. XZY022020007, No. XZY022021006). (corresponding author: Jinglong Chen).

Tianci Zhang and Jinglong Chen are with the State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an, 710049, China. (e-mail: jlstrive2008@mail.xjtu.edu.cn).

Shuilong He is with School of Mechanical and Electrical Engineering, Guilin University of Electronic Technology, Guilin 541004, China. (e-mail: xiaofeilonghe@163.com).

Zitong Zhou is with ShaanXi Fast Gear Company Ltd., Xi'an 710119, China. (e-mail: zhouzitong@fastgroup.cn).

vision have shown that injecting prior knowledge into a network model can improve its performance when there are not sufficient training data [15], [16]. In intelligent fault diagnosis, although many remarkable results have been achieved, the diagnosis knowledge accumulated over time has not been paid enough attention to, and scholars prefer to let models process monitoring data and output results automatically. Traditional intelligent diagnosis models are more like end-to-end black-boxes, where researchers feed monitoring data into a model and expect the model to output the health status automatically [1]. The black-box nature of diagnosis models means that during the model training we cannot confirm whether the model has learned the knowledge related to fault identification. In particular, in the case of small training samples, models tend to learn some simple features to identify input data. In more extreme cases, features extracted by models may even be completely irrelevant to faults, because the data available for learning is very limited. In engineering scenarios, monitoring data is affected by environmental noise and working conditions, and data distribution is so complex that it is difficult to achieve accurate fault identification using only simple features learned by general models. Fortunately, prior diagnosis knowledge may help diagnosis models to learn more targeted and richer feature representations. Specifically, by injecting prior knowledge into diagnosis models, the models need not only to mine the black-box features of the input data but also to consider the discovery of prior knowledge from the input data. Under such conditions, although there are few available samples, the model can learn richer data features through the complexity of the training task, and thus the generalization ability can be improved.

In the field of fault diagnosis, domain experts have accumulated rich engineering experience and diagnosis knowledge in practice, such as fault mechanism of machines, fault features, diagnosis rules, and feature extraction algorithms, which form the prior knowledge in this field, and their effectiveness has been verified in numerous engineering cases [17]. For example, it is possible to determine whether a machine is abnormal by calculating signal feature indicators such as the peak and the mean value. This diagnosis knowledge is more targeted and reliable. Moreover, they are usually based on rigorous fault mechanism analysis, so they are also interpretable and can be understood by engineers. Recently, some scholars have tried to construct diagnosis models using prior diagnosis knowledge, and it has been shown that prior knowledge helps the models to learn more robust fault features and helps to improve the model interpretability [17]–[20]. However, it is worth noting that the current work has rarely explored the potential of prior diagnosis knowledge in reducing the amount of training data, which may be an important means to achieve accurate fault identification in the case of small samples.

To improve the performance of diagnosis models in the case of limited fault data, we try to enhance the training of diagnosis models with prior diagnosis knowledge. As mentioned earlier, the fault features of machines are an important part of the prior knowledge in this field. Among the machine fault features, the feature indicators of monitoring signals can reflect the health status of the machine. It is a common means for traditional fault diagnosis to determine the fault type of a machine by calculating the signal feature indicators. Inspired by [3] and [21], we first select 24 commonly used signal feature indicators as the prior feature set, including 12 time-domain indicators and 12 frequency-domain indicators whose validity has been verified in [3] and [21]. Besides, we use a deep convolutional auto-encoder (DCAE) to extract general features (24 dimensions/24D) of the signals, which may contain fault information that is not available in the prior feature indicators. Our purpose of adding general features is to take full advantage of the powerful data processing power of neural networks. After that, we train a deep convolutional neural network (DCNN) based fault identification model using a self-supervised learning strategy. Self-supervised learning is the latest popular paradigm for training neural networks without labeled data, which has attracted the attention of scholars in the field of computer vision, especially in the problem of image as well as video data processing [22]. Unlike traditional unsupervised learning, self-supervised learning artificially constructs proxy labels for the input data based on the properties of the data and then relies on the proxy labels to design supervised training tasks for neural networks. Neural networks are expected to learn feature representations from the input data by completing the designed supervised tasks. Recently, in intelligent fault diagnosis, scholars have started to initially explore the potential of self-supervised learning for unlabeled mechanical data processing [23], [24]. However, the idea of mining prior knowledge contained in mechanical data using self-supervised learning has not been tried yet, and this paper will be devoted to training diagnosis models to learn prior diagnosis knowledge from raw data with self-supervised training strategy.

In the structure proposed in this paper, the prior features and the general features are combined as the fusion features, which are served as the proxy labels. In the training process, we take a small amount of unlabeled data as the input of the fault identification model. The model is trained to output the proxy labels, i.e. fusion features, which means that the feature learning process of the model is guided by the fusion features rather than only by the fault labels. Then, we use a small amount of labeled data to fine-tune the model, and the fine-tuned model will be more adept at accurate fault identification. Through self-supervised feature learning, the model is expected to extract richer and more relevant fault features from fault data, thus the fault identification accuracy will be improved under small samples condition. In contrast to other feature engineering-based diagnostic models [20], we do not simply derive diagnosis results from signal feature indicators, but use prior features to enhance the training process of the end-to-end diagnosis model, and apply the trained end-to-end model to achieve fault identification. Finally, the authors need to point out that the purpose of this paper is to provide an idea or framework for training intelligent diagnosis models, rather than just showing some network modules or the diagnosis results of some cases. The network modules used in this paper, such as the DCAE and DCNN, can be easily replaced with other models like Long Short-Term Memory (LSTM) in practice according to the user's needs. Besides, the 24 prior features used in this

paper can likewise be replaced, added, or deleted according to the characteristics of the data to be analyzed.

The contributions of this paper are summarized as follows.

1) We present a prior diagnosis knowledge-augmented self-supervised feature learning scheme for machine fault identification using limited faulty monitoring data, which may provide a new perspective for few-shot fault diagnosis from the angle of prior knowledge.
2) A self-supervised learning framework is designed to train the fault identification model with the prior and the general features served as the proxy labels in the pre-training stage, which makes the model is no longer a complete black-box, and the fault features it learns are interpretable to some extent.
3) Two mechanical fault simulation experiments are carried out to verify the effectiveness of the proposed model. Based on the experimental data, we analyze the learned fault features from the qualitative and quantitative perspectives using feature visualization technology and Pearson correlation coefficients.

The rest of this paper is organized as follows. Section II presents related works on prior knowledge and self-supervised learning. Section III describes the proposed model in detail. Section IV verifies and discusses the proposed model based on two mechanical fault simulation experiments. Section V makes a conclusion for this paper.

II. Related Works

This paper attempts to apply prior knowledge to enhance the training of intelligent diagnosis models through a self-supervised learning strategy. Therefore, this section will introduce the related works about intelligent diagnosis with prior knowledge and self-supervised representation learning.

A. Intelligent diagnosis with prior knowledge

Data-driven intelligent diagnosis models expect to find health information of machines from massive historical monitoring data. However, these intelligent models are trained automatically according to the input data, engineers cannot intervene in their training process, so these models are often incomprehensible and uninterpretable. In engineering scenarios, an uninterpretable model is usually not trusted because it may go wrong under certain unexpected circumstances [25]. While in practice, domain experts give diagnosis results for machines based on extensive engineering experience and background knowledge of diagnosis, rather than singularly on monitoring data. The experience as well as the knowledge that experts rely on, such as fault mechanism of machines, fault features, diagnosis rules, and feature extraction algorithms, can be called the prior knowledge in the field of fault diagnosis. Prior diagnosis knowledge is usually based on fault simulation and mechanism analysis, which means that they are rigorous and reliable. More importantly, they can be understood by engineers and therefore more easily trusted in engineering scenarios [17].

Some scholars have tried to integrate existing diagnosis knowledge into their intelligent models to improve the diagnosis performance and interpretability of the models. For example, A multi-task convolutional neural network was applied in [17] for fault diagnosis and localization, the fault characteristics frequency was fused into the first convolution layer. Zheng et al. [18] proposed a deep domain generalization network with prior diagnosis knowledge for bearing fault diagnosis, in which the signals were preprocessed under the guidance of prior knowledge. Yu et al. [19] presented a knowledge-guided deep belief network-based diagnosis model, in which the classification rules were inserted into the model as prior knowledge. In [20], the empirical fault features selected by engineers are combined with deep learning-based data features, and the fused features were fed into the XGBoost classifier for classification directly.

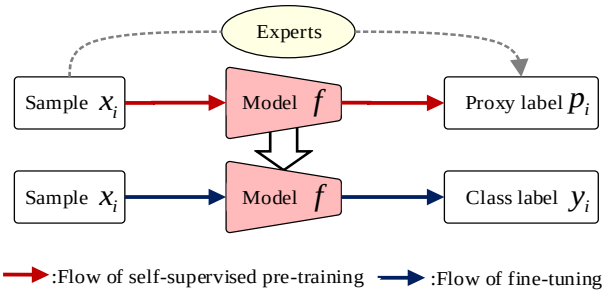

Fig. 1. Illustration of self-supervised learning.

Existing works have demonstrated that prior knowledge helps intelligent diagnosis models to learn robust fault features. In addition, the interpretability of models can also be improved. However, little work has been done to use prior knowledge to improve the diagnosis performance of models under limited fault data conditions. With only a small number of fault samples available for model training, it is difficult to learn discriminative and representative fault features. In this paper, we will enhance the training process of the model using prior diagnosis knowledge, which is expected to enable the model to mine richer fault information from a small amount of data, thus improving the generalization ability of the diagnosis models.

B. Self-supervised representation learning

Self-supervised learning is an emerging neural network training paradigm in recent years, which is dedicated to mining high-level semantic features from unlabeled data [22]. As shown in Fig. 1, it consists of two steps, self-supervised pre-training and fine-tuning, which can also be referred to as upstream and downstream tasks, respectively. Given a training dataset $D=\{(x_1,y_1),...,(x_n,y_n)\}$, where (x_i,y_i) is formed by a sample x_i and its corresponding class label y_i. For the model f to be trained, the training objective in the pre-training stage can be described as

$$\hat{f}=\arg\min_{f\in F}\sum_{i=1}^{n}l\left(f\left(x_i\right),p_i\right) \tag{1}$$

where $\hat{f}$ is the obtained model in this stage, F is overall hypothesis set, $l(\cdot)$ represents loss function, and p_i is the proxy label for the sample x_i designed by human experts manually, which is derived from the sample x_i itself but does not depend on the class label y_i. By self-supervised pre-training, the model f can obtain reliable weights set, which will be reused in the fine-tuning stage [26]. In the fine-tuning stage, the training objective can be described as

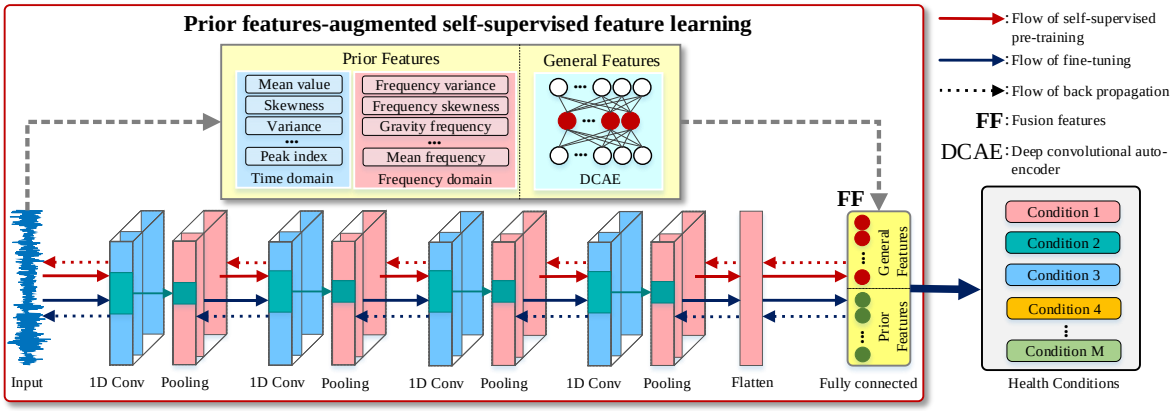

Fig. 2. Framework of the proposed model.

$$f^* = \arg\min_{\hat{f} \in F} \sum_{i=1}^{n} l\left(\hat{f}\left(x_i\right), y_i\right) \tag{2}$$

where f^* is the final model.

How to construct proxy labels for data is the key of self-supervised learning. In the field of computer vision, some scholars artificially rotated images at different angles and used the rotated angles as the proxy labels for supervised model training [27]. The experimental results in [27] demonstrate that the model can learn semantic features of images by predicting the rotation angle, thus providing an effective feature representation for downstream tasks. In addition, Noroozi et al. [28] constructed a position prediction task for image region blocks by segmenting images and using their positions as proxy labels to achieve self-supervised pre-training of the network. Recently, the idea of self-supervised representation learning has also been introduced into the field of intelligent fault diagnosis [23], [24], [29], [30]. For example, Zhang et al. [23] proposed a self-supervised feature learning-based federated learning strategy for fault diagnosis of machines, in which the authors construct fake data by swapping the positions of data fragments, and the model judges the data as real or fake for self-supervised training. Senanayaka et al. [24] used a self-supervised convolutional neural network for fault feature extraction, the temporary class labels given by a one-class support vector machine were served as the proxy labels for self-supervised training.

Existing works provide preliminary evidence that self-supervised learning has potential in feature mining, especially in the processing of unlabeled data. In engineering scenarios, unlabeled data is easier to obtain than labeled data, so the application of self-supervised learning in building intelligent diagnosis models will predictably increase. However, as a flexible learning strategy, how to make models learn prior diagnosis knowledge from raw monitoring data based on self-supervised learning has not been explored by scholars. For example, the self-supervised task is constructed by swapping signal fragments in [22], and the proxy labels in [23] are derived from a one-class SVM, which are not strongly correlated with prior diagnosis knowledge. In this paper, we try to construct a self-supervised training task for the model using fault features from existing diagnosis knowledge as the proxy labels, so that the model can achieve the purpose of learning prior diagnosis knowledge from raw monitoring data.

III. Proposed Method

A. Overview of the proposed model

In this section, we describe the proposed model in detail, as shown in Fig. 2. Given the training set $X = \{x_i\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^{n\times 1}$ is the i th sample with n data points. Further, the training set X can be divided into two subsets, i.e., the labeled subset $X_l = \{x_i, y_i\}_{i=1}^{K}$ and the unlabeled subset $X_u = \{x_i\}_{i=1}^{N-K}$, where $0 < K \le N$. The sample $x_i \in X_l$ is labeled with the health condition label $y_i \in \{1, 2, ..., M\}$, where M is the total number of health conditions.

First of all, as the source of the prior diagnosis knowledge in the proposed model, the feature indicators of monitoring signals are commonly used by domain experts to determine the health status of machines. Therefore, we select 24 feature indicators based on the recommendation of [3] and [21], and the validity of these indicators has been verified in [3] and [21]. As given in Table I, they contain 12 time-domain indicators and 12 frequency-domain indicators. The 24 indicators are served as prior features set $\{p_1^i, p_2^i, ..., p_{24}^i\}_{i=1}^{N}$ after data standardization, which can be calculated as

$$p^i = \frac{p^i - mean\left(p^i\right)}{std\left(p^i\right)}. \tag{3}$$

Then, we use a deep convolutional auto-encoder (DCAE), which has been proved to be effective in unsupervised feature extraction from vibration data [6], to extract the general features $\{g_1^i, g_2^i, ..., g_{24}^i\}_{i=1}^{N}$ of the signal. The general features learned by deep neural networks may contain fault information that the prior features do not have. To construct the proxy labels for training, we fused the prior features and the general features and standardized them using Eq. 3 to obtain the fusion features set $\{f_1^i, f_2^i, ..., f_{48}^i\}_{i=1}^{N}$. And then, a deep convolutional neural network (DCNN) based fault identifier, which has strong feature extraction ability and its effectiveness in fault identification has been proven in a wide range of work [31], is

TABLE I
24 PRIOR FEATURE INDICATORS

Time-domain		Frequency-domain	
Name	Equation	Name	Equation
Mean value	$p_1=\frac{1}{N}\sum_{n=1}^{N}x(n)$	Frequency mean value	$p_{13}=\frac{1}{N}\sum_{n=1}^{N}s(n)$
Standard deviation	$p_2=\sqrt{\frac{1}{N-1}\sum_{n=1}^{N}[x(n)-p_1]^2}$	Frequency variance	$p_{14}=\frac{1}{N-1}\sum_{n=1}^{N}[s(n)-p_{13}]^2$
Square root amplitude	$p_3=(\frac{1}{N}\sum_{n=1}^{N}\sqrt{\lvert x(n)\rvert})^2$	Frequency skewness	$p_{15}=\frac{1}{Np_{14}^{3/2}}\sum_{n=1}^{N}[s(n)-p_{13}]^3$
Absolute mean value	$p_4=\frac{1}{N}\sum_{n=1}^{N}\lvert x(n)\rvert$	Frequency steepness	$p_{16}=\frac{1}{Np_{14}^{2}}\sum_{n=1}^{N}[s(n)-p_{13}]^4$
Skewness	$p_5=\frac{1}{N}\sum_{n=1}^{N}(x(n))^3$	Gravity frequency	$p_{17}=\sum_{n=1}^{N}f_i s(n)\Big/\sum_{n=1}^{N}s(n)$
Kurtosis	$p_6=\frac{1}{N}\sum_{n=1}^{N}(x(n))^4$	Frequency standard deviation	$p_{18}=\sqrt{\sum_{n=1}^{N}(f_i-p_{17})^2 s(n)\Big/N\sum_{n=1}^{N}s(n)}$
Variance	$p_7=\frac{1}{N}\sum_{n=1}^{N}(x(n))^2$	Frequency root mean square	$p_{19}=\sqrt{\sum_{n=1}^{N}f_i^2 s(n)\Big/\sum_{n=1}^{N}s(n)}$
Kurtosis index	$p_8=p_6/(\sqrt{p_7})^2$	Average frequency	$p_{20}=\sqrt{\sum_{n=1}^{N}f_i^4 s(n)\Big/\sum_{n=1}^{N}f_i^2 s(n)}$
Peak index	$p_9=\max\lvert x(n)\rvert/p_2$	Regularity degree	$p_{21}=\sum_{n=1}^{N}f_i^2 s(n)\Big/\sqrt{\sum_{n=1}^{N}s(n)\Big/\sum_{n=1}^{N}f_i^4 s(n)}$
Waveform index	$p_{10}=p_2/p_4$	Variation parameter	$p_{22}=p_{18}/p_{17}$
Pulse index	$p_{11}=\max\lvert x(n)\rvert/p_4$	Eighth-order moment	$p_{23}=\sum_{n=1}^{N}(f_i-p_{17})^3 s(n)\Big/Np_{18}^3$
Skewness index	$p_{12}=p_5/(\sqrt{p_7})^3$	Sixteenth-order moment	$p_{24}=\sum_{n=1}^{N}(f_i-p_{17})^4 s(n)\Big/Np_{18}^4$

where $x(n)$ is a signal sample containing N data points, $s(n)$ is a spectrum, and f_i is the frequency value of the i th spectrum line.

trained in a self-supervised way, where the sample $x_i \in X$ is fed into DCNN, and the proxy-labels $\left\{f_1^i, f_2^i, ..., f_{48}^i\right\}$ are output.

Finally, to obtain higher fault identification accuracy, a few labeled samples are used to fine-tune the parameters of DCNN, where the sample $x_i \in X_l$ is input into DCNN, and the health condition label y_i is output.

B. Network structure

The proposed model involves two network modules, i.e., a DCAE and a DCNN. DCAE contains a convolutional encoder, which has four convolutional layers and four pooling layers, and a convolutional decoder, which has four convolutional layers and four up-sampling layers. The structure of the DCNN-based fault identifier is given in Table II. In the proposed model, the fault identifier will extract features from raw signals by the calculation of the convolutional layers and the pooling layers.

Given the kernel in the i th convolutional layer is k_i, the bias is b_i, and the output c_i can be calculated as

$$c_i^j = lrelu\left(\sum_{k\in M_i} p_{i-1}^j \times k_i^{j,k} + b_i^j\right) \tag{4}$$

where M_i is the feature vector of the i th convolutional layer, $lrelu$ is the LeakyReLU activation function

$$lrelu = \begin{cases} x, x>0 \\ 0.2*x, x<0 \end{cases}, \tag{5}$$

and p_{i-1} is the output of the $i-1$th pooling layer.

$$p_{i-1}^j = \max_{k\in\rho} c_{i-1}^{k+j*s} \tag{6}$$

TABLE II
THE STRUCTURE OF THE FAULT IDENTIFIER

Layer	Channels @ Kernel size× Stride / Pool size×Stride	Output shape	Padding	Activation function
Input	/	1×1024	/	/
1D Convolution	32@4×1	32×1024	Same	LeakyReLU
Max-pooling	2×2	32×512	Valid	/
1D Convolution	32@4×1	32×512	Same	LeakyReLU
Max-pooling	2×2	32×256	Valid	/
1D Convolution	64@4×1	64×256	Same	LeakyReLU
Max-pooling	2×2	64×128	Valid	/
1D Convolution	64@4×1	64×128	Same	LeakyReLU
Max-pooling	2×2	64×64	Valid	/
Flatten	/	4096	/	/
Fully connected	/	48	/	LeakyReLU
Classification	/	7	/	Softmax

where ρ is the pooling window size and s is the stride. As given in Table II, the output of the fourth pooling layer p_4 is downscaled to 48 dimensions using a fully connected layer, which can be described as

$$g_q = lrelu\left(p_4 w_q + b_q\right) \tag{7}$$

where g_q is the output of the fully connected layer, w_q and b_q are the weight matrix and bias of this layer.

To achieve fault classification, the extracted features g_q are fed into a softmax classifier. The output of the softmax classifier is $h(x_i)$.

$$h(x_i) = \begin{bmatrix} p(y_i=1\mid x_i;\theta) \\ p(y_i=2\mid x_i;\theta) \\ \cdots \\ p(y_i=M\mid x_i;\theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^{M} e^{\theta_j^T x_i}} \begin{bmatrix} e^{\theta_1^T x_i} \\ e^{\theta_2^T x_i} \\ \cdots \\ e^{\theta_M^T x_i} \end{bmatrix} \tag{8}$$

where θ is the parameter of the softmax classifier.

C. Training strategy

The training of the proposed model consists of the following three steps.

1) Proxy labels construction

The 24 prior features set $\left\{p_1^i, p_2^i, ..., p_{24}^i\right\}_{i=1}^N$ of the sample $x_i \in X$ are calculated and standardized. The general features $\left\{g_1^i, g_2^i, ..., g_{24}^i\right\}_{i=1}^N$ of the sample $x_i \in X$ are extracted by DCAE. The prior features and the general features are fused and standardized to form the fusion features set $\left\{f_1^i, f_2^i, ..., f_{48}^i\right\}_{i=1}^N$, which is served as the proxy labels.

In this step, DCAE achieves the extraction of the general features by optimizing the data reconstruction loss L_{dr}.

$$L_{dr} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{n}\left(x_i^j - \tilde{x}_i^j\right)^2}{nN} \tag{9}$$

where $\tilde{x}$ is the reconstructed data sample.

2) Self-supervised pre-training

The sample $x_i \in X$ is input into DCNN, and the output of the fully connected layer in DCNN is taken to fit the proxy label $\left\{f_1^i, f_2^i, ..., f_{48}^i\right\}$ by optimizing the self-supervised loss L_{sp}.

$$L_{sp} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{48}\left(f_j^i - \tilde{f}_j^i\right)^2}{48*N} \tag{10}$$

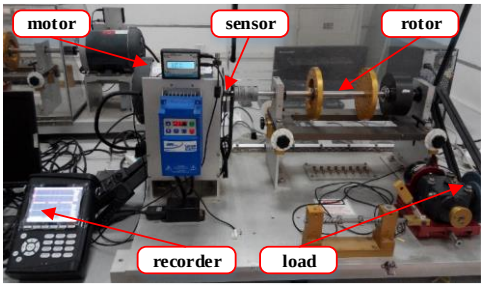

Fig. 3. SQ test bench.

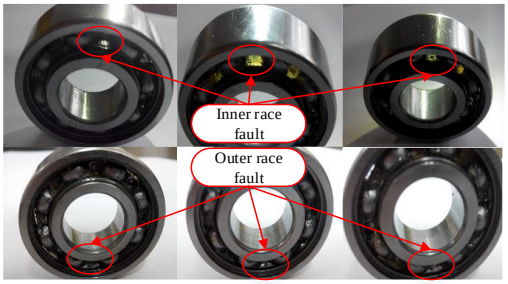

Fig. 4. Faulty bearings in the SQ experiment.

where $\tilde{f}$ is the output of the fully connected layer in DCNN.

3) Parameters fine-tuning

The sample $x_i \in X_l$ is input into DCNN, and the predicted label $h(x_i)$ is output directly. The parameters are fine-tuned by optimizing the cross-entropy loss function L_c.

$$L_c = -\frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{M} 1\{y_i = j\}\log\left(h(x_i)_j\right). \tag{11}$$

where $1\{\cdot\}$ is a function returning 1 if $y_i = j$, and 0 otherwise.

IV. Experimental Verification and Result Discussion

In this section, we will introduce two mechanical fault simulation experiments and verify the effectiveness of the proposed model based on the collected experimental data. The two experiments are carried out in Spectra Quest (SQ) mechanical fault simulation test bench and Tian-Xian (TX) mechanical fault simulation test bench respectively.

A. Experiments and data description

1) SQ mechanical fault simulation experiment

As shown in Fig. 3, the SQ test bench mainly consists of a drive motor, a rotor, a load, a data recorder, and vibration sensors. To simulate different health conditions of rolling bearings, we processed six kinds of single-point faults, which are labeled as minor inner race fault (IF-1), medium inner race fault (IF-2), severe inner race fault (IF-3), minor outer race fault (OF-1), medium outer race fault (OF-2), and severe outer race fault (OF-3), as given in Fig. 4. In addition, there is a healthy bearing in normal condition (NC-0) for comparison. In the vibration data collection experiments, the rotating speed is 40Hz, and the sampling frequency of the data recorder is 25.6kHz. We recorded bearing vibration data in seven kinds of health conditions in turn. To verify the proposed model, we construct a bearing vibration data samples set, which contains 1125 samples in each health condition, and each sample contains 1024 data points.

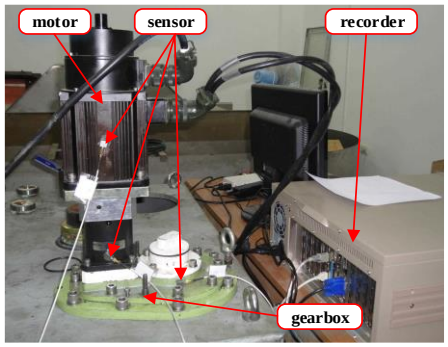

Fig. 5. TX test bench.

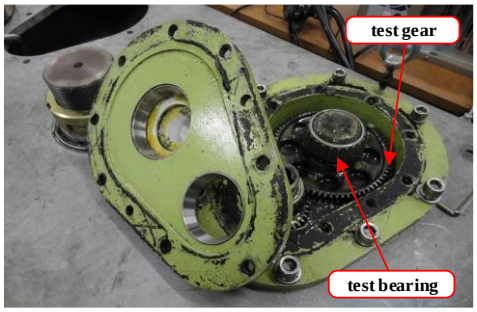

Fig. 6. Gearbox including tested bearings and gears.

2) TX mechanical fault simulation experiment

As given in Fig. 5 and 6, the TX test bench consists of a motor, a gearbox, a data recorder, and vibration sensors. Different from the SQ bearing fault simulation experiment, which only has one tested object, the TX fault simulation experiment tested both bearings and gears. In the experiment, we processed six kinds of bearing single point faults and one gear fault. Besides, a normal condition is also tested for comparison. The eight health conditions are labeled as normal condition (NC-0), bearing cage fault (CF-0), gear fault (GF-0), bearing roller fault (RF-0), bearing outer race spot welding (OF-0), bearing minor outer race fault (OF-1), bearing medium outer race fault (OF-2), and bearing severe outer race fault (OF-3). During the data collection process, the rotating speed of the motor is 800rpm, and the sampling frequency of the data recorder is 5kHz. Similarly, we construct a sample set containing 774 samples in each health condition, and each sample contains 1024 data points.

B. Fault diagnosis using small samples

First, we need to define the range of small samples. In the study of few-shot intelligent diagnosis, there are usually two options for defining the range of small samples. The first one is to use the ratio of the training samples number to the total samples number, for example, when the training samples do not exceed 10% of the total samples, it is considered as the small sample case [32]. The second one is to use the absolute number of training samples, for example, when the number of training samples does not exceed 30, it is considered as the small sample case [7]. Here, we are more in agreement with the latter, because even 10% of the data volume is substantial when the total number of samples to be analyzed is large enough. Therefore, in this paper, we define the small sample condition

TABLE III
FAULT DIAGNOSIS RESULT 1 BASED ON TWO SET OF EXPERIMENTAL DATA

Data	Model	Number of training samples (All labeled)							
		10	15	20	25	30	40	50	100
SQ	PF-S	**0.7230±0.0322**	0.7500±0.0118	0.7763±0.0211	0.7693±0.0262	0.7776±0.0206	0.8010±0.0640	0.8378±0.0010	0.9037±0.0008
	GF-S	0.5297±0.0330	0.6948±0.0691	0.7830±0.0464	0.8083±0.0524	0.8405±0.0142	0.9027±0.0014	0.9335±0.0112	0.9526±0.0182
	FF-S	0.7148±0.0356	0.7966±0.0237	0.8032±0.0246	0.8119±0.0223	0.8681±0.0156	0.9098±0.0052	0.9530±0.0052	0.9620±0.0025
	DCNN	0.5514±0.0552	0.7316±0.0743	0.8355±0.0352	0.8884±0.0116	0.9375±0.0055	0.9535±0.0404	0.9634±0.0314	0.9757±0.0257
	SSAE	0.6479±0.0191	0.7620±0.0186	0.8723±0.0453	0.9219±0.0041	0.9556±0.0021	0.9668±0.0022	0.9683±0.0016	0.9886±0.0014
	R-Net	0.5808±0.0388	0.7484±0.0375	0.8637±0.0255	0.9176±0.0260	0.9440±0.0056	0.9564±0.0140	0.9617±0.0054	0.9868±0.0018
	S-Net	0.6189±0.0177	0.7309±0.0243	0.8339±0.0160	0.8827±0.0117	0.9193±0.0178	0.9417±0.0254	0.9479±0.0090	0.9608±0.0098
	PM	0.6692±0.0279	**0.8213±0.0497**	**0.9175±0.0203**	**0.9511±0.0235**	**0.9622±0.0106**	**0.9704±0.0277**	**0.9835±0.0157**	**0.9921±0.0201**
TX	PF-S	0.8128±0.0155	0.8058±0.0125	0.7947±0.0583	0.7855±0.0626	0.8050±0.0573	0.8066±0.0201	0.8252±0.0095	0.8352±0.0095
	GF-S	0.6713±0.0311	0.7160±0.0271	0.7136±0.0616	0.7561±0.0214	0.7702±0.0272	0.8292±0.0119	0.8817±0.0263	0.9248±0.0195
	FF-S	**0.8423±0.0262**	0.8263±0.0273	0.8586±0.0263	0.8863±0.0233	0.8772±0.0307	0.9002±0.0104	0.9240±0.0111	0.9567±0.0055
	DCNN	0.7851±0.0512	0.8071±0.0397	0.8507±0.0334	0.8854±0.0359	0.9136±0.0374	0.9288±0.0166	0.9337±0.0193	0.9588±0.0116
	SSAE	0.6943±0.0327	0.7073±0.0077	0.8055±0.0170	0.8309±0.0090	0.8486±0.0115	0.8822±0.0188	0.8971±0.0229	0.9323±0.0144
	R-Net	0.7370±0.0561	0.7817±0.0521	0.8553±0.0378	0.8792±0.0283	0.9060±0.0273	0.9233±0.0274	0.9610±0.0055	0.9756±0.0088
	S-Net	0.7531±0.0118	0.8256±0.0140	0.8639±0.0165	0.9023±0.0214	0.9261±0.0310	0.9338±0.0211	0.9530±0.0129	0.9592±0.0283
	PM	0.8094±0.0209	**0.8478±0.0252**	**0.8904±0.0229**	**0.9237±0.0162**	**0.9350±0.0206**	**0.9576±0.0186**	**0.9737±0.0152**	**0.9873±0.0243**

TABLE IV
FAULT DIAGNOSIS RESULT 2 BASED ON TWO SET OF EXPERIMENTAL DATA

Number of training samples		SQ experimental data			TX experimental data		
Total	Labeled	PF-DCNN	GF-DCNN	PM	PF-DCNN	GF-DCNN	PM
10	1	0.3238±0.0956	0.2878±0.1252	**0.5983±0.0541**	0.6045±0.0650	0.5744±0.1068	**0.7624±0.0215**
	5	0.5158±0.0538	0.5092±0.1588	**0.6347±0.0249**	0.6602±0.0320	0.7016±0.0353	**0.7683±0.0263**
	10	0.6400±0.0429	0.5597±0.1432	**0.6692±0.0279**	0.7519±0.0179	0.7639±0.0350	**0.8094±0.0209**
20	1	0.3998±0.0547	0.4326±0.1195	**0.8163±0.0404**	0.5549±0.1002	0.7096±0.0341	**0.7861±0.0401**
	5	0.7785±0.0995	0.6121±0.1038	**0.8186±0.0544**	0.8043±0.0385	0.7530±0.0248	**0.8637±0.0320**
	10	0.8212±0.0418	0.7537±0.0378	**0.8459±0.0379**	0.8544±0.0451	0.8327±0.0379	**0.8802±0.0242**
	15	0.8649±0.0207	0.7767±0.0146	**0.9150±0.0313**	0.8856±0.0241	0.8623±0.0159	**0.8946±0.0210**
	20	0.8859±0.0098	0.8199±0.0159	**0.9175±0.0203**	0.8759±0.0185	0.8708±0.0124	**0.8904±0.0229**
30	1	0.4848±0.0680	0.5212±0.1253	**0.8761±0.0492**	0.6061±0.0434	0.6881±0.0745	**0.8145±0.0597**
	5	0.8581±0.0644	0.8021±0.0734	**0.8942±0.0377**	0.8090±0.0468	0.7739±0.0356	**0.8923±0.0141**
	10	0.9034±0.0626	0.8795±0.0189	**0.9273±0.0160**	0.8677±0.0493	0.8443±0.0211	**0.8907±0.0198**
	15	0.9128±0.0355	0.8891±0.0095	**0.9571±0.0216**	0.9064±0.0316	0.8567±0.0088	**0.9184±0.0212**
	20	0.9436±0.0048	0.8973±0.0168	**0.9702±0.0190**	0.9122±0.0251	0.8642±0.0176	**0.9246±0.0181**
	25	0.9591±0.0107	0.9263±0.0220	**0.9646±0.0195**	0.9133±0.0358	0.9058±0.0025	**0.9243±0.0199**
	30	**0.9659±0.0026**	0.9485±0.0209	0.9622±0.0106	**0.9379±0.0115**	0.9316±0.0081	0.9350±0.0206
40	1	0.4524±0.0811	0.5938±0.1166	**0.9026±0.0392**	0.5744±0.0978	0.7220±0.1034	**0.8467±0.0389**
	5	0.8323±0.0536	0.9088±0.0267	**0.9606±0.0107**	0.7804±0.0496	0.8674±0.0325	**0.8958±0.0247**
	10	**0.9696±0.0215**	0.9209±0.0284	0.9647±0.0158	0.8858±0.0462	**0.9253±0.0120**	0.9200±0.0153
50	1	0.4041±0.0778	0.6383±0.0484	**0.9336±0.0307**	0.6090±0.0689	0.6847±0.0356	**0.8504±0.0459**
	5	0.7858±0.0585	0.9395±0.0480	**0.9762±0.0138**	0.7905±0.0375	0.9038±0.0307	**0.9247±0.0189**
	10	0.9387±0.0428	0.9574±0.0331	**0.9819±0.0097**	0.8931±0.0297	0.9520±0.0083	**0.9581±0.0123**
100	1	0.4888±0.0757	0.6234±0.0075	**0.9677±0.0032**	0.4702±0.1073	0.6972±0.0954	**0.8281±0.0545**
	5	0.8951±0.0564	0.9288±0.0401	**0.9683±0.0061**	0.8389±0.0692	0.9260±0.0238	**0.9382±0.0201**
	10	0.9527±0.0214	0.9623±0.0703	**0.9889±0.0061**	0.0692±0.0103	0.9508±0.0219	**0.9552±0.0071**

when the training samples number in each class does not exceed 50. Compared to the total number of samples in each class (1125 in the SQ data and 774 in the TX data), we consider 50 to be a small value. It should be noted that in this part, we also took 100 training samples to carry out experiments, which was to verify the validity of the model over a larger range.

Then, we selected nine methods for comparison.

1) PF-S: We extract 24 feature indicators from signal samples, and feed them into the Softmax classifier for fault classification directly. The 24 signal feature indicators are given in Table I.
2) GF-S: We use the DCAE in the proposed model to extract the general features of the signals, and input them into the Softmax classifier for classification directly.
3) FF-S: The extracted features in PF-S and GF-S are fused to form the fusion features. The fusion features are input into the Softmax classifier for classification directly.
4) DCNN: We use a DCNN to process the signal samples directly, and the health condition labels are output by DCNN automatically. The structure and parameters setting of DCNN are given in Table II.
5) PF-DCNN: A DCNN is pre-trained by self-supervised learning and then fine-tuned, in which the prior features in PF-S are used as the proxy labels. The fully connected layer in this DCNN has 24 nodes, and the rest of the parameters are set as in Table II.
6) GF-DCNN: We take the convolutional encoder from the trained DCAE in GF-S and add a Softmax classifier to the last layer to form a DCNN, and then fine-tune this DCNN using labeled data.
7) SSAE: Saufi et al. [33] presented a stacked sparse auto-encoder (SSAE) for machine fault identification with limited fault data, in which the L2 weight regularization and sparsity regularization were applied in SSAE.
8) R-Net: Yang et al. [34] proposed a residual wide-kernel deep net (R-Net) for bearing fault diagnosis using limited samples, in which residual learning blocks were used to prevent overfitting.
9) S-Net: A few-shot learning approach based on Siamese neural network (S-Net) was proposed in [11] for bearing fault classification, which obtained high classification accuracy with small training data.

Besides, the proposed model is abbreviated as PM in the analysis of experimental results.

For the setting of hyper-parameters, we conducted some experiments to determine the appropriate values of the learning rates and training epochs in the model. In the training process, the DCAE in the proposed model is trained 500 epochs with a learning rate of 0.01. The fault identifier is pre-trained 500 epochs with a learning rate of 0.01 and then fine-tuned 50 epochs with a learning rate of 0.001. Each set of the experiment is repeated 10 times, and the average results are recorded for analysis, as given in Tables III and IV.

As shown in Table III, the proposed model achieves the highest accuracy in the six experiments from 15 samples to 50 samples. At the sample size of 10, PF-S and FF-S, which directly utilize the prior features, perform better than the proposed model, which indicates that the prior features are effective for fault identification under small samples. Although the proposed model also learns the prior features in the self-supervised pre-training stage, the learned features are bound to have errors because the features learned by the model are obtained by the calculation of multilayer neural networks rather than by explicitly defined formulas. Especially when the amount of training data is small, this error will cause the proposed model to fail to achieve better results than PF-S and FF-S. Compared with PF-S, GF-S, and FF-S, DCNN is an end-to-end model, which can process data automatically. However, DCNN is prone to problems such as overfitting under conditions of small samples. The proposed model is also end-to-end and incorporates prior diagnosis knowledge and general features of data in the training process, so the fault features learned by the proposed model are richer than those of an ordinary DCNN. Moreover, compared with the three state-of-the-art methods, the results based on both two sets of fault experimental data show that the proposed model has higher diagnosis performance under small sample conditions.

In addition, in Table IV, the proposed model obtained by self-supervised pre-training with both prior and general features performs better than that obtained by pre-training with a single type of feature in most cases. It is worth noting that the smaller the number of training samples, the more obvious the advantage of the proposed model over other relevant models. Finally, under the condition of 100 training samples, the proposed model is still able to achieve higher fault identification accuracy compared to other methods.

The features utilized in the proposed model, including prior features (PF), general features (GF), and fusion features (FF), have a significant impact on the performance of the model. Therefore, it is necessary to verify the effectiveness of these features for fault classification. We selected five classification models and trained them directly using these features for comparison. The five models are described as follows.

1) F-CNN: A convolutional neural network with two convolutional layers and two pooling layers is selected, the parameters of the convolutional layers are 8@4×1 (channels @ kernel size ×stride) and the parameters of the pooling layers are 2×2 (pool size ×stride).
2) F-kNN: A k-Nearest Neighbors classifier with the number of neighbors of 5.
3) F-MLP: A two-layer Multi-layer Perceptron with 48 and 24 nodes when FF is used as input, and 24 and 12 nodes when PF or GF is used as input.
4) F-SVM: A Support Vector Machine with the kernel function of Gaussian radial basis function.
5) F-Softmax: A Softmax classifier. The experimental results of this model are given in Table III.

In the experiments, we use 20 samples as training samples and the rest as test samples, and the experimental results are shown in Fig. 7. In Fig. 7, the horizontal coordinate represents the input features. For example, SQ-PF represents the PF in the SQ experimental data. The vertical coordinate represents the average accuracy. With only 20 training samples, all three kinds of features achieved accuracy of more than 0.7 in all five models. It is noteworthy that the accuracy of using FF as input exceeds 0.8 in all experiments, which shows that the features used in the proposed model are effective for fault classification.

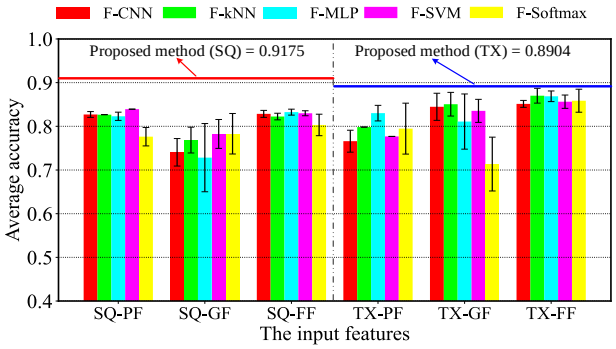

Fig. 7. Fault classification based on traditional classifiers using PF, GF, and FF under 20 training samples. The red line and the blue line represent the accuracy of the proposed model with SQ data and TX data respectively.

C. Fault diagnosis under data imbalance

For many industrial systems, there is an imbalance between the amount of normal data and faulty data. Therefore, the proposed model is validated using imbalanced data to show its applicability in more realistic industrial scenarios.

We conducted the following experiments based on SQ data. In the experiments, 20 samples of each fault type are taken as faulty training samples and the rest are used as test samples. The number of normal samples is determined according to the data imbalance ratio, and if the data imbalance rate is 50, then the number of normal training samples is 1000. In the case of data imbalance, the recall R for each type of data is an important indicator to evaluate the performance of the model.

$$R(class = i) = \frac{T_i}{P_i} \tag{12}$$

where P_i is the total number of samples contained in class i and T_i is the number of samples correctly identified in class i. We repeated the experiments 10 times and took the average results for analysis, as given in Table V.

From Table V, it can be seen that the recall for faulty samples fluctuates and tends to decrease as the imbalance ratio increases. And, the recall for normal samples is always 1.0, which is because the proportion of normal samples in the training set is larger than faulty samples, and therefore the trained model has a stronger identification of normal samples. In terms of the average accuracy, when the training set is balanced, the classification accuracy is 0.9175, and when the data imbalance ratio is as high as 50, the classification accuracy can still reach 0.8236. Thus, it can be seen that the proposed model can still achieve relatively high fault identification accuracy even if there is a certain degree of imbalance in the data.

TABLE V
RECALL R UNDER DATA IMBALANCE USING SQ DATA

	Imbalance ratio				
	10	20	30	40	50
IF-1	1.0000±0	0.9998±0.0005	0.9988±0.0018	0.9996±0.0009	0.9996±0.0009
IF-2	0.8162±0.1483	0.8732±0.0902	0.8038±0.1587	0.7464±0.1618	0.7550±0.2113
IF-3	0.7848±0.1778	0.5608±0.3782	0.6422±0.3244	0.6038±0.3386	0.6004±0.2219
OF-1	0.8386±0.1188	0.9768±0.0199	0.8846±0.0983	0.8922±0.2182	0.8272±0.1804
OF-2	0.9366±0.0878	0.8048±0.2140	0.8258±0.1640	0.9268±0.0470	0.8818±0.0565
OF-3	0.8436±0.2169	0.9222±0.0569	0.9326±0.0792	0.7274±0.1382	0.7016±0.2529
NC-0	1.0000±0	1.0000±0	1.0000±0	1.0000±0	1.0000±0
Average accuracy	0.8886±0.0384	0.8767±0.0441	0.8696±0.0801	0.8423±0.0522	0.8236±0.0427

D. Analysis of the learned features

Evaluation of the learned features is an important measure of the superiority of the model. We select seven models, i.e., PF-S, GF-S, FF-S, DCNN, PF-DCNN, GF-DCNN, and the proposed model, and then qualitatively observed the learned data features using t-SNE technology [35]. Based on the classification accuracies in Tables III and IV, we chose to visualize the learned features under 20 training samples to highlight the superiority of the proposed model, the results are shown in Fig.8 and 9.

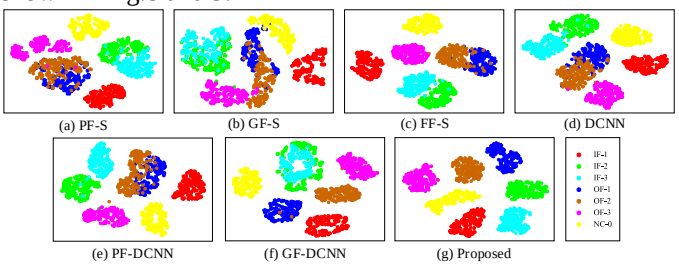

Fig. 8. Feature visualization of SQ experimental data.

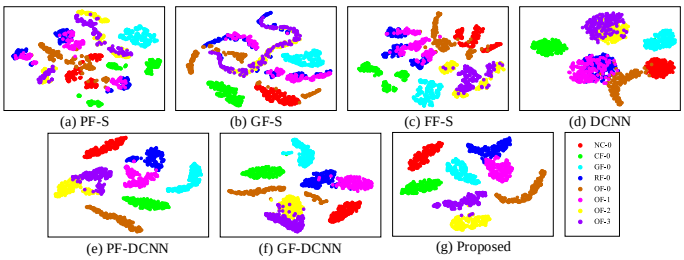

Fig. 9. Feature visualization of TX experimental data.

It can be found that the features used in PF-S and GF-S are relatively limited in the ability to characterize the different types of data, and there are multiple types of data that are misclassified. The distinguishability of features in FF-S has been improved due to the combination of prior and general features. Under the condition of small samples, the features extracted by DCNN are also prone to unclear classification boundaries. In contrast, the proposed model can obtain clearer classification boundaries than DCNN, therefore, the number of misclassified samples is reduced greatly. Finally, the proposed model based on fusion features learns more distinguishable fault features than the models (PF-DCNN and GF-DCNN) based on single feature sources.

Further, we use the Pearson linear correlation coefficient (Pcc) to quantify the correlation of the learned features before and after fine-tuning [7]. A higher correlation between the features before and after fine-tuning indicates that self-supervised pre-training is more useful for fault identification since the features learned by pre-training have high similarity to the features eventually used by the model. As shown in Fig. 10, the feature learned by the model through pre-training is a 48-dimensional vector (fusion features), of which the first 24 dimensions are the prior features and the last 24 dimensions are the general features. We will analyze the correlation of the features from three perspectives, the first one is the correlation between the overall features or fusion features, the second one is the correlation between the prior features, and the third one is the correlation between the general features. Taking the SQ data as an example, we focus on the correlation of the features before and after fine-tuning from 10 to 50 training samples. Each set of experiments was repeated 10 times and the average correlation was taken for analysis. We present the obtained results separately according to the health condition labels, as given in Fig. 11.

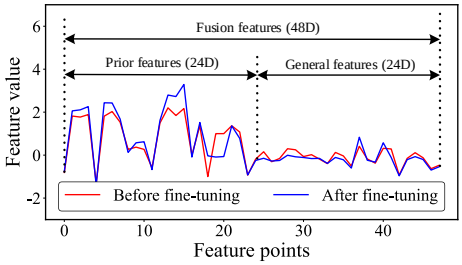

Fig. 10. Learned features of SQ data before and after fine-tuning (IF-1 condition and 20 training samples).

As can be seen from Fig. 11, the data features learned before and after fine-tuning have a certain degree of correlation. Moreover, for most kinds of health conditions, the feature correlation tends to increase slowly as the number of training samples increases, which indicates that the data features learned by self-supervised pre-training are indeed useful for the final fault identification. The specific analysis results are as follows.

1) For the inner race faulty data, when the number of training samples is small, the correlation of the prior features is higher than that of the general features, which indicates that the prior features play a greater role in the inner race fault identification than the general features at this time. And as the number of training samples increases, the correlation of the general features gradually exceeds that of the prior features.
2) For the data in the normal condition, feature correlations do not vary much with the number of training samples, and the correlations of the general features are slightly higher than the prior features in most cases.
3) For the outer race faulty data, we note that the correlations of the general features are higher than the prior features in almost all cases, which suggests that the general features may be more beneficial than the prior features in the SQ data for the identification of bearing outer race faults.

Overall, the data features artificially constructed enhance the model training through self-supervised learning. More importantly, unlike traditional end-to-end black-box models, the fault features learned by the proposed model are adapted from the artificially constructed ones, so the proposed model can be considered to have a certain degree of interpretability.

E. Discussion

Finally, the advantages of the proposed model are listed as follows.

1) The proposed model requires only a small number of fault signal samples for training. Through self-supervised pre-training, the proposed model can extract rich and discriminative features from a small amount of data, thus improving the accuracy of fault identification.
2) The diagnosis performance of the proposed model is enhanced using prior diagnosis knowledge, which is generally targeted and reliable. The interpretability of

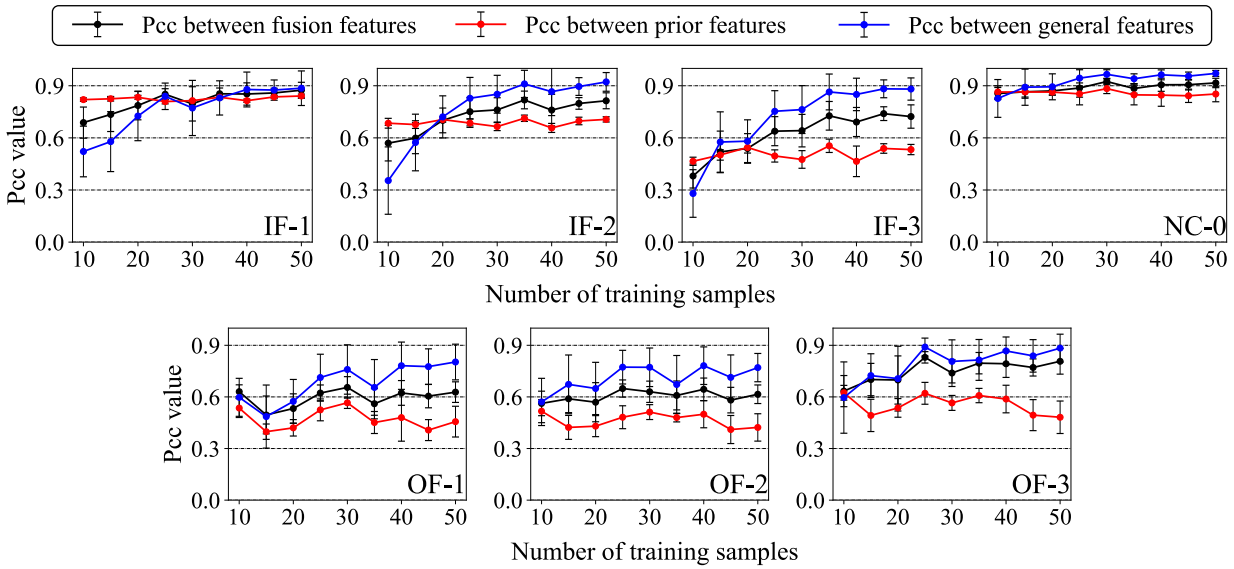

Fig. 11. Correlation analysis of the learned features (SQ experimental data).

the model is improved.

3) The network modules like DCAE and DCNN and the prior features set used in the presented model are substitutable and engineers can easily replace them according to the characteristics of the data to be analyzed.

Certainly, there are some shortcomings for further consideration. First, when the number of training samples is very small, the performance of the model still needs to be further improved. Since the prior features have been experimentally shown to be effective under small samples, how to make the model further fully learn and utilize the prior features may be an effective means to improve the model performance. Second, engineers can identify faults in new classes based on prior diagnosis knowledge, while the proposed model currently can only identify the faults contained in its training set and cannot predict new types of faults. Therefore, how to identify unseen faults is still a problem worth considering for the proposed model. In the future development, we may add a prior knowledge embedded zero-shot learning module [36] to the proposed model to achieve the identification of unseen fault types.

V. Conclusion

In this paper, we propose a prior knowledge augmented self-supervised feature learning scheme for intelligent fault diagnosis of machines under the condition of small samples. We build prior features set based on existing diagnosis knowledge and extract general features of data using a deep convolutional auto-encoder. A self-supervised learning framework is designed using the prior and general features as proxy labels for the training of the fault identifier. The trained fault identifier can mine rich features from a small number of signal samples and achieve relatively high fault identification accuracy. The validity of the proposed model is verified based on two sets of mechanical fault simulation data. The results show that, with the enhancement of prior diagnosis knowledge, the proposed model can achieve better performance than related methods in the case of small samples. And for future development, the proposed model should consider how to learn and utilize the prior knowledge more fully and how to use the prior knowledge to identify unseen fault types of machines.

References

[1] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, and A. K. Nandi, "Applications of machine learning to machine fault diagnosis: A review and roadmap," *Mech. Syst. Signal Process.*, vol. 138, p. 106587, 2020.

[2] J. Pan, Y. Zi, J. Chen, Z. Zhou, and B. Wang, "LiftingNet: A Novel Deep Learning Network with Layerwise Feature Learning from Noisy Mechanical Data for Fault Classification," *IEEE Trans. Ind. Electron.*, vol. 65, no. 6, pp. 4973–4982, 2018.

[3] T. Pan, J. Chen, S. He, *et al*, "A Novel Deep Learning Network via Multiscale Inner Product With Locally Connected Feature Extraction for Intelligent Fault Detection," *IEEE Trans. Ind. Informatics*, 2019.

[4] Y. Chang, J. Chen, C. Qu, and T. Pan, "Intelligent fault diagnosis of Wind Turbines via a Deep Learning Network Using Parallel Convolution Layers with Multi-Scale Kernels," *Renew. Energy*, 2020.

[5] T. Zhang *et al.*, "Intelligent fault diagnosis of machines with small & imbalanced data: A state-of-the-art review and possible extensions," *ISA Trans.*, 2021.

[6] T. Zhang, J. Chen, J. Xie, and T. Pan, "SASLN : Signals Augmented Self-taught Learning Networks for Mechanical Fault Diagnosis under Small Sample Condition," *IEEE Trans. Instrum. Meas.*, 2021.

[7] T. Zhang, J. Chen, F. Li, and T. Pan, "A Small Sample Focused Intelligent Fault Diagnosis Scheme of Machines via Multi-modules Learning with Gradient Penalized Generative Adversarial Networks," *IEEE Trans. Ind. Electron.*, 2021.

[8] S. Dixit and N. K. Verma, "Intelligent Condition Based Monitoring of Rotary Machines with Few Samples," *IEEE Sens. J.*, 2020.

[9] Z. Ren *et al.*, "A novel model with the ability of few-shot learning and quick updating for intelligent fault diagnosis," *Mech. Syst. Signal Process.*, vol. 138, 2020.

[10] F. Jia, S. Li, H. Zuo, and J. Shen, "Deep Neural Network Ensemble for the Intelligent Fault Diagnosis of Machines under Imbalanced Data," *IEEE Access*, vol. 8, pp. 120974–120982, 2020.

[11] A. Zhang, S. Li, Y. Cui, W. Yang, R. Dong, and J. Hu, "Limited Data Rolling Bearing Fault Diagnosis With Few-Shot Learning," *IEEE Access*, vol. 7, pp. 110895–110904, 2019.

[12] F. Li, J. Chen, J. Pan, and T. Pan, "Cross-domain learning in rotating machinery fault diagnosis under various operating conditions based on parameter transfer," *Meas. Sci. Technol.*, 2020.

[13] H. Kim and B. D. Youn, "A New Parameter Repurposing Method for Parameter Transfer with Small Dataset and Its Application in Fault Diagnosis of Rolling Element Bearings," *IEEE Access*, 2019.

[14] Z. He, H. Shao, P. Wang, J. Lin, J. Cheng, and Y. Yang, "Deep transfer multi-wavelet auto-encoder for intelligent fault diagnosis of gearbox with

few target training samples," *Knowledge-Based Syst.*, vol. 191, p. 105313, 2020.

[15] Z. Ren and E. B. Sudderth, "Clouds of Oriented Gradients for 3D Detection of Objects, Surfaces, and Indoor Scene Layouts," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

[16] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum PointNets for 3D Object Detection from RGB-D Data," 2018. doi: 10.1109/CVPR.2018.00102.

[17] S. Guo, B. Zhang, T. Yang, D. Lyu, and W. Gao, "Multitask Convolutional Neural Network with Information Fusion for Bearing Fault Diagnosis and Localization," *IEEE Trans. Ind. Electron.*, 2020.

[18] H. Zheng, Y. Yang, J. Yin, Y. Li, R. Wang, and M. Xu, "Deep Domain Generalization Combining A Priori Diagnosis Knowledge Toward Cross-Domain Fault Diagnosis of Rolling Bearing," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021.

[19] J. Yu and G. Liu, "Knowledge extraction and insertion to deep belief network for gearbox fault diagnosis," *Knowledge-Based Syst.*, 2020.

[20] J. Xie, Z. Li, Z. Zhou, and S. Liu, "A Novel Bearing Fault Classification Method Based on XGBoost: The Fusion of Deep Learning-Based Features and Empirical Features," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021.

[21] J. Chen, C. Wang, B. Wang, and Z. Zhou, "A visualized classification method via t-distributed stochastic neighbor embedding and various diagnostic parameters for planetary gearbox fault identification from raw mechanical data," *Sensors Actuators, A Phys.*, vol. 284, pp. 52–65, 2018.

[22] X. Liu *et al.*, "Self-supervised Learning: Generative or Contrastive," *IEEE Trans. Knowl. Data Eng.*, 2021, doi: 10.1109/tkde.2021.3090866.

[23] W. Zhang, X. Li, H. Ma, Z. Luo, and X. Li, "Federated learning for machinery fault diagnosis with dynamic validation and self-supervision," *Knowledge-Based Syst.*, vol. 213, 2021.

[24] J. S. L. Senanayaka, H. Van Khang, and K. G. Robbersmyr, "Toward Self-Supervised Feature Learning for Online Diagnosis of Multiple Faults in Electric Powertrains," *IEEE Trans. Ind. Informatics* , 2021.

[25] D. Zhang, Y. Chen, F. Guo, H. R. Karimi, H. Dong, and Q. Xuan, "A New Interpretable Learning Method for Fault Diagnosis of Rolling Bearings," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021.

[26] Q. Zhao and J. Dong, "Self-supervised representation learning by predicting visual permutations," *Knowledge-Based Syst.*, vol. 210, 2020.

[27] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *6th International Conference on Learning Representations*, 2018.

[28] M. Noroozi and P. Favaro, "Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles Mehdi Noroozi and Paolo Favaro," *Proc. Eur. Conf. Comput. Vison*, 2016.

[29] T. Wang, M. Qiao, M. Zhang, Y. Yang, and H. Snoussi, "Data-driven prognostic method based on self-supervised learning approaches for fault detection," *J. Intell. Manuf.*, vol. 31, no. 7, 2020.

[30] W. Zhang, D. Chen, and Y. Kong, "Self-supervised joint learning fault diagnosis method based on three-channel vibration images," *Sensors*, vol. 21, no. 14, 2021.

[31] J. Jiao, M. Zhao, J. Lin, and K. Liang, "A comprehensive review on convolutional neural network in machine fault diagnosis," *Neurocomputing*, 2020.

[32] Y. Ding, L. Ma, J. Ma, C. Wang, and C. Lu, "A generative adversarial network-based intelligent fault diagnosis method for rotating machinery under small sample size conditions," *IEEE Access*, 2019.

[33] S. R. Saufi, Z. A. Bin Ahmad, M. S. Leong, and M. H. Lim, "Gearbox Fault Diagnosis Using a Deep Learning Model with Limited Data Sample," *IEEE Trans. Ind. Informatics*, vol. 16, no. 10, pp. 6263–6271, 2020.

[34] D. Yang, H. R. Karimi, and K. Sun, "Residual wide-kernel deep convolutional auto-encoder for intelligent rotating machinery fault diagnosis with limited samples," *Neural Networks*, 2021.

[35] L. Van Der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, 2008.

[36] L. Feng *et al.*, "Fault Description Based Attribute Transfer for Zero-Sample Industrial Fault Diagnosis," *IEEE Trans. Ind. Informatics*, 2020.

Tianci Zhang received the B.S. degree in mechanical engineering from Xi'an Jiaotong University, Xi'an, China, in 2018. He is currently working towards the Ph.D. degree in mechanical engineering at Xi'an Jiaotong University.

His research interests include mechanical signal processing, intelligent fault diagnosis and machinery condition monitoring.

Jinglong Chen (M'16) received the Ph.D. degrees in mechanical engineering from Xi'an Jiaotong University, Xi'an, China, in 2014. He is currently an Associate Professor of mechanical engineering at Xi'an Jiaotong University. From 2012 to 2013, he was a Visiting PhD Student with the University of Alberta, Edmonton, AB, Canada.

His research interests focus on machinery condition monitoring and fault diagnosis, mechanical signal processing, and mechanical system reliability.

Shuilong He received the Ph.D. degrees in mechanical engineering from Xi'an Jiaotong University, Xi'an, China, in 2014. He is currently an Associate Professor of mechanical engineering at Guilin University of Electronic Technology.

His research interests focus on mechanical vibration data processing, and machinery condition monitoring and fault diagnosis.

Zitong Zhou received the Ph.D. degree in mechanical engineering from Xi'an Jiaotong University, Xi'an, China, in 2020.

He is currently a Digital Engineer with Shaanxi Fast Gear Company Ltd., Xi'an. His research interests include condition monitoring of mechanical equipment and mechanical signal processing.