Highlights

A Wavelet Transform and self-supervised learning-based framework for bearing fault diagnosis with limited labeled data

Yuhong Jin, Lei Hou, Ming Du, Yushu Chen

- A novel fault diagnosis framework for bearings with limited labeled data is developed.
- Satisfactory diagnosis accuracy is achieved on two bearing fault datasets consisting of only 1% labeled data.
- The effect of hyperparameters on fault diagnosis performance and computational complexity of the proposed framework is discussed in detail.
- Without any supervised sample, fault-specific features with inter-class separability are observed via visualization technique.
A Wavelet Transform and self-supervised learning-based framework for bearing fault diagnosis with limited labeled data

Yuhong Jin, Lei Hou, Ming Du and Yushu Chen

School of Astronautics, Harbin Institute of Technology, Harbin, 150001, P. R. China

ARTICLE INFO

Keywords:
Fault diagnosis
Self-supervised learning
Transformer
Limited labeled samples
Deep learning

ABSTRACT

Traditional supervised bearing fault diagnosis methods rely on massive labeled data, yet annotations may be very time-consuming or infeasible. The fault diagnosis approach that utilizes limited labeled data is becoming increasingly popular. In this paper, a Wavelet Transform (WT) and self-supervised learning-based bearing fault diagnosis framework is proposed to address the lack of supervised samples issue. Adopting the WT and cubic spline interpolation technique, original measured vibration signals are converted to the time-frequency maps (TFMs) with a fixed scale as inputs. The Vision Transformer (ViT) is employed as the encoder for feature extraction, and the self-distillation with no labels (DINO) algorithm is introduced in the proposed framework for self-supervised learning with limited labeled data and sufficient unlabeled data. Two rolling bearing fault datasets are used for validations. In the case of both datasets only containing 1% labeled samples, utilizing the feature vectors extracted by the trained encoder without fine-tuning, over 90% average diagnosis accuracy can be obtained based on the simple K-Nearest Neighbor (KNN) classifier. Furthermore, the superiority of the proposed method is demonstrated in comparison with other self-supervised fault diagnosis methods.

1. Introduction

Rotating machinery plays a vital role in modern industries, so its condition monitoring and health management are of great importance. According to statistics, about 45%-55% of rotating machinery and equipment failure is caused by damage to the bearing part [1]. Therefore, timely and accurate bearing fault diagnosis has always been highly demanded to enhance machine reliability [2, 3]. Traditional bearing fault diagnosis methods are based on the physical model, and the specific fault components are analyzed by various signal processing techniques [4–6]. However, these physical models are not universal in complex, high-dimensional systems. With the rapid development of industrial technology, the structure of rotating machinery becomes more and more complicated, and the limitations of conventional fault diagnosis methods are gradually highlighted.

In recent years, benefited from the advances in industrial computer and sensor technology, data-driven intelligent fault diagnosis method has attracted more and more attention from researchers due to the great merits of high accuracy and low requirement for prior knowledge [7]. Data-driven fault diagnosis methods can be divided in two categories: Machine Learning (ML) based approach and Deep Learning (DL) based approach. Currently, many ML models such as Support Vector Machine (SVM) [8–10], Self-Organized Map (SOM) [11, 12], Auto-Encoder (AE) [13, 14], and Radial Basis Function (RBF) neural network [15, 16] have been extensively employed in the field of fault diagnosis. Nevertheless, most of these approaches still require manual design features, which means limitations in adaptive feature extraction.

DL based techniques can automatically learn the feature representation from mass data according to the given task and have a strong feature extraction ability. These approaches have been widely developed, and numerous promising results have been acquired [17]. Assorted neural network architectures and enhanced techniques, such as Deep Belief Network (DBN) [18], Convolutional Neural Network (CNN) [19–21], Recurrent Neural Network (RNN) [22, 23], Attention mechanism [24], Transformer [25, 26] and their variants [27, 28] have also been generally exploited for the fault diagnosis. Jie et al. [29] established a novel Gaussian-Bernoulli deep belief network (GDBN) model for intelligent
fault diagnosis, where the graph regularization and sparse features learning are embedded. Combining the advantages of attention mechanism, Squeeze-and-Excitation Network (SENet) [30], and soft threshold, Zhao et al. [31] developed the Deep Residual Shrinkage Networks (DRSN). This new deep learning model can achieve a high fault diagnosis accuracy under strong noise interference. Aiming at the problem that the connection of the local fragments (namely quasi-periodicity) in the measured vibration signals is easy to be neglected, Gao et al. [32] performed a novel weak fault diagnosis method for the rolling bearings based on the Long Short Term Memory (LSTM) network and multichannel Continuous Wavelet Transform (CWT). Compared to the traditional CWT, MCCWT converts the original sample space into a multichannel representation, which improves the feature extraction capability of the LSTM. In recent work, Ding et al. [33] applied the Transformer architecture to fault diagnosis of rolling bearings. Based on the view that the time-frequency map obtained by signal processing is formed by splicing the instantaneous spectrum of the signal over a period of time, the time-frequency Transformer (TFT) model is presented. Through experiments, the effectiveness and superiority of the proposed TFT are validated. In brief, the advances in DL based fault diagnosis methods fully prove the potency of the emerging data-driven algorithm for processing complex mechanical systems [34].

Satisfactory fault diagnosis results reported in the above literature are premised on the sufficient labeled data. These DL based approaches are based on the paradigm of supervised learning for end-to-end training, thus acquiring the features with generalization ability [35]. However, in real engineering environment, labeling large amounts of measured data may be time-consuming and costly. Besides, considering the complexity and uncertainty of the machinery system, in many cases, the fault mode corresponding to the acquired vibration signal is virtually unidentified, which leads that supervised learning is not effective enough. Therefore, the problem of fault diagnosis with limited labeled data has aroused extensive attention from researchers. Currently, two main approaches to this problem are dataset extension and exploring unlabeled data. Dataset extension aims to generate the additional "labeled" data based on the limited labeled samples through a technique similar to interpolation, including various data augmentation methods [36, 37], Generative Adversarial Network (GAN) [38–40], etc. Li et al. [41] designed a data augmentation method that combined different signal processing techniques such as masking noise, signal translation, stretching, etc. The experimental results show that the diagnosis performance of the DL model can get a promotion from more generated samples. By employing the Sparse Auto-Encoder (SAE) to reduce the dimension of the original data, Ma et al. [42] improved the traditional GAN and proposed the Sparsity-Constrained Generative Adversarial Network (SCGAN), which better converges to Nash equilibrium, and higher diagnosis accuracy can be achieved with limited labeled data. Different from the above research, Wang et al. came up with a novel dataset extension method based on the Sub-Pixel Convolutional Neural Network (ESPCN) [43]. Through this method, generated data with high-resolution can be acquired. Experimental results show that the proposed method has strong feasibility to carry out data augmentation for fault diagnosis of rotating machines under the speed fluctuation condition. Overall, sample size can be enlarged using dataset extension, and the lack of labeled data is alleviated.

However, fake samples generated based on the dataset extension technology will inevitably be similar to the real samples, resulting in lacking data diversity, which will increase the risk of overfitting. In addition, mass unlabeled data without precise machine health conditions, which can be easily collected in general, does not get effective utilization in these methods. To solve these problems, fault diagnosis methods under the limited labeled samples case by exploring unlabeled data are presented, known as unsupervised learning [44, 45], self-supervised learning [46–48], and semi-supervised learning [49–51]. Zhang et al. [52] offered an unsupervised framework with Reconstruction Sparse Filtering (RSF) for rolling bearing diagnosis. The basis vectors are constrained explicitly by a Soft-Reconstruction Penalty (SRP), enabling RSF to learn a group of independent basis vectors to extract dissimilar features without applying any labeled sample. Li et al. [53] leveraged deep InfoMax (DIM) to improve the generalization ability of the features extracted by a CNN encoder, which can alleviate the overfitting problem when the labeled fault samples are limited. Li et al. [54] designed a three-stage semi-supervised fault diagnosis conjoin unsupervised clustering and supervised fine-tuning. Even only one labeled sample for each class, a high testing accuracy is obtained by this method. The above research results demonstrate the feasibility of exploring unlabeled data to address the challenging diagnostic tasks with limited labeled samples.

This paper focuses on the problem of bearing fault diagnosis with availabilities of limited labeled data and sufficient unlabeled data. Under this proposition, a Wavelet Transform and self-supervised learning-based framework is proposed. Herein, the Vision Transformer is adopted as an encoder to extract the feature vectors, which possesses the preponderance of great parallelism and strong scalability. Accordingly, the projector head is introduced on the top of the encoder to establish the model network for self-supervised learning, whose output is pseudo labels. Based on this, a pretext task called "local to global correspondence" is constructed by initializing the teacher and student networks with
A Wavelet Transform and self-supervised learning-based framework for bearing fault diagnosis with limited labeled data

the same architecture. Moreover, the centering and sharpening operations included in the self-distillation with no labels algorithm are integrated into the proposed framework to train the model networks, which enable us effectively avoid the mode collapse during the training procedure. Furthermore, the Wavelet Transform technique is used to convert the original time-domain vibration signals into time-frequency maps. This preprocessing method helps the encoder learn a better feature representation. Experiments on two bearing fault datasets where only 1% labeled data is contained are implemented to validate the proposed method, and over 90% average testing accuracy is achieved.

2. Proposed method

The proposed method mainly comprises three major parts: (1) preprocessing pipeline of the raw vibration signals; (2) model network for feature representation; (3) training algorithm for the model network. In this section, we will illustrate our proposed methodology in detail.

2.1. Wavelet Transform and preprocessing pipeline

In the fault diagnosis of rotating machinery, background noise is pervasive due to sensors’ noise input and environmental factors, which will interfere with the time-domain signals. Additionally, the acquired vibration signals are usually non-stationary due to speed fluctuation and fault. In this case, neither time-domain analysis nor frequency-domain analysis is suitable. Therefore, it is necessary to convert the raw vibration signals into time-frequency representation (TFR). In this paper, we utilize the Wavelet Transform (WT) as the time-frequency domain analysis method. The WT of a given vibration signal $x(t)$ can be defined as

$$WT(a, \tau) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t) \psi\left(\frac{t - \tau}{a}\right) dt = \int_{-\infty}^{+\infty} x(t) \psi_{a, \tau}(t) dt \quad (1)$$

where $a$ is scaling parameter, $\tau$ is time translation parameter. $\psi_{a, \tau}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t - \tau}{a}\right)$ is wavelet basis function of the WT.

Fig.1 shows the TFR of the raw vibration signals based on the WT. We can find that the frequency-domain signals of each time step are well extracted. Nevertheless, it should be noted that the size of TFR is related to the number of sampling points, which means that the shape of TFR may be uncertain. Besides, numerous sampling points can result in a huge TFR, making subsequent calculations difficult. To fix the shape of TFR, as shown in Fig.1, we introduce a resize method. The specific approach is as follows: first, the amplitude range of TFR is scaled to [0, 1] by normalization. Then, the amplitude values of TFR are mapped to color values. Finally, the size of TFR is fixed at $224 \times 224$ by cubic spline interpolation. The TFR after the above treatments is denoted as time-frequency map (TFM), whose shape is $224 \times 224 \times 3$. TFM can be directly used as input to the model network.

![Figure 1: Preprocessing pipeline of the original vibration signal](image)

2.2. Model network

2.2.1. Encoder

We need two networks in the following training process: teacher network and student network. In this paper, both of them are collectively called the model network. As shown in Fig.2, the model network consists of two stages:
A Wavelet Transform and self-supervised learning-based framework for bearing fault diagnosis with limited labeled data

an encoder and a projector head. The encoder provides a feature extraction and representation from the input TFM. A variety of DL models, such as Residual Neural Network (ResNet) and Vision Transformer (ViT) [55], can be used as encoder’s backbone, and in this paper, we employ the ViT as our backbone. The impact of different backbones on the fault diagnosis performance will be discussed later. For the sake of presentation, the input TFM is denoted as \( x \in \mathbb{R}^{H \times W \times C} \) ignoring the batch size, where \((H, W)\) is the shape of the TFM, \(C\) is the number of channels. As described previously, \( H = W = 224 \), and \( C \) is 3. To conform the input form of the standard Transformer, we reshape the TFM into a sequence of flattened 2D patches, denoted as \( x_p \in \mathbb{R}^{N \times (P^2 \times C)} \) where \((P, P)\) is the resolution of each TFM patch, \(N\) is the number of patches, and \(N = HW / P^2\). This process is visually illustrated in Fig.2, which divides the input TFM evenly into patches of a specified size. Then, we reorder these patches into a sequence, as shown in Eq.(2)

\[
x \in \mathbb{R}^{H \times W \times C} \xrightarrow{\text{reshape}} x_p = [x_1^p, x_2^p, \ldots, x_N^p] \in \mathbb{R}^{N \times (P^2 \times C)}
\]  

(2)

where \(x_i^p\) is the \(i\)th patch, and \(i = 1, 2, 3, \ldots, N\). Then, like the procedure of word embedding, we flatten each patch and map them to the high-dimensional embedding space through a learnable linear projection, as given in Eq.(3)

\[
z_0 = [x_1^p, x_2^p, \ldots, x_N^p] \cdot W_{\text{emd}} \in \mathbb{R}^{N \times d}
\]  

(3)

where \(W_{\text{emd}} \in \mathbb{R}^{(P^2 \times C) \times d}\) is a trainable embedding matrix, \(d\) is the embedding dimension. It should be noted that all patches share the same embedding matrix, so Eq.(3) is mathematically equivalent to a 2D convolution operation. Then, since the Transformer does not contain the position information, position encoding is added to retain the absolute and relative position information of the patches. And beyond that, similar to BERT’s class token, we present a learnable vector (denoted as \(x_{\text{class}}\)) to serve as the feature representation, which can be obtained by Eq.(4)

\[
z_0 \leftarrow [x_{\text{class}}; z_0] + E_{\text{pos}}
\]  

(4)

where \(x_{\text{class}} \in \mathbb{R}^d\), \(E_{\text{pos}} \in \mathbb{R}^{(N+1) \times d}\). Both of them are learnable parameters. Eq.(2) to Eq.(4) are collectively called patch embedding, as shown in Fig.2. \(z_0\) is referred to as “embedding sequence.”

Figure 2: The architecture of the model network.

The following part of the encoder can be regarded as a feature extraction structure, which takes the embedding sequence \(z_0\) as input. In the ViT architecture, the core part of feature extraction is the Transformer layer, which comprises multiple Transformer basic blocks stacked on top of each other. Fig.2 introduces the structure of the
Transformer base blocks, including Multi-head Self-attention mechanism (MSA) and Multilayer perceptron blocks (MLP). MSA is the most critical definition in Transformer basic blocks, which is based on the attention mechanism. The attention mechanism can be explained in terms of soft addressing. Consider that an element in the memory is composed of a key (K) and value (V). Currently, there is a query (Q), and we need to pick out V based on the similarity between Q and K. Note that for the same Q value, instead of hard addressing, we may take all the Vs in the memory and sum them weighted based on their importance. The importance of V is measured by the comparability between Q and its corresponding K, which is denoted as the Attention Distribution (AD). The calculation method of AD is called the score function, such as Scaled Dot-Product Attention, Bahdanau Attention, and Content-based Attention. In this paper, we adopt the Scaled Dot-Product Attention in our backbone, which is described as

$$\text{attn}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_K}})V$$

(5)

where \(d_K\) is the embedding dimension of K. \(\frac{1}{\sqrt{d_K}}\) is a scaling factor for stabilizing the gradient. However, the attention mechanism in (5) cannot extract the information in the embedding sequence under the different subspace. Then, to avoid this deficiency, similar to the group convolution operation, researchers have introduced the Multi-head attention (MHA) mechanism, which can be given as

$$MHA(Q, K, V) = [head_1, head_2, ..., head_{\text{head}}] \cdot W_O$$

where \(head_i = \text{attn}(QW_i^Q, KW_i^K, VW_i^V)\)

(6)

where \(W_0 \in \mathbb{R}^{hd_x \times d}, W_i^Q \in \mathbb{R}^{d \times d_k}, W_i^K \in \mathbb{R}^{d \times d_k}, W_i^V \in \mathbb{R}^{d \times d_v}\). \(h\) is the number of head and \(d_v\) is the dimension of values. Furthermore, when \(Q = K = V\), Eq.(6) is also called Multi-head Self-attention mechanism, namely MSA. The output of MSA in \(l\)th Transformer basic block (denoted as \(z_i^{MSA}\)) with residual connection and Layer normalization (LN) can be described as

$$z_i^{MSA} = MSA(LN(z_{l-1})) + z_{l-1}$$

$$= MHA(LN(z_{l-1}), LN(z_{l-1}), LN(z_{l-1})) + z_{l-1}$$

(7)

where \(z_i\), \(l = 1, 2, 3, ..., depth\) is the output of the \(l\)th Transformer basic block in Transformer layer, and \(depth\) is the number of Transformer blocks. \(z_0\) is the embedding sequence.

Moreover, MLP blocks are applied after MSA in every Transformer basic block to achieve more complex nonlinear mapping. The MLP block in the \(l\)th Transformer basic block includes a nonlinear projection layer with an activation function and a linear projection layer, which is given in Eq.(8)

$$MLP(z_i^{MSA}) = GeLU(z_i^{MSA}W_1^l + b_1^l)W_2^l + b_2^l$$

(8)

where \(W_1^l \in \mathbb{R}^{d \times d_{MLP}}, b_1^l \in \mathbb{R}^{d_{MLP}}, W_2^l \in \mathbb{R}^{d \times d_{MLP}}, b_2^l \in \mathbb{R}^d\). \(d_{MLP}\) represents the embedding dimension of the nonlinear projection layer, and \(GeLU(x)\) indicates the Gaussian error Linear Unit (GeLU), which can be shown as Eq.(9)

$$GeLU(x) = x\phi(x) = x[1 + erf(x/\sqrt{2})]/2 \approx 0.5x(1 + tanh(\sqrt{2/\pi}(x + 0.045x^3)))$$

(9)

where \(\phi(x)\) is the standard Gaussian distribution function. GeLU is smoother over the entire input range than Rectified Linear Unit (ReLU) and Leaky ReLU, with no discontinuous gradient at 0. Combining the residual connection and LN, the output of MLP in the \(l\)th Transformer basic block (namely \(z_i^{MLP}\)) can be obtained as

$$z_i = MLP(LN(z_i^{MSA})) + z_i^{MSA}$$

(10)

Based on Eq.(7) and Eq.(10), the final output of the Transformer layer and the feature representation (denoted as \(y\)) in the ViT are shown in

$$z_i^{MSA} = MSA(LN(z_{l-1})) + z_{l-1}, \quad l = 1, 2, 3, ..., depth$$

$$z_i = MLP(LN(z_i^{MSA})) + z_i^{MSA}, \quad l = 1, 2, 3, ..., depth$$

(11)

$$y = LN(z_{depth}[0])$$
where $z_{\text{depth}}$ is the final output of the Transformer layer, and $z^0_{\text{depth}}$ means the class token in $z_{\text{depth}}$. To facilitate the expression in the following, we denote the set of learnable parameters in the model network as $\theta$ and denote the calculation process of the encoder part as $f_\theta$. Then the feature representation can be expressed as $y \triangleq f_\theta(x)$, where $x$ is the input TFM.

### 2.2.2. Projector head

The architecture of the projection head is straightforward, consisting of an $n$-layer MLP, an L2-normalization layer, and a weight normalized linear layer. The calculation of the $n$-layer MLP and linear layer is almost precisely the same as Eq.(8), not tired in words here. The L2-Normalization layer stabilizes the training process. Overall, the projector head plays almost the same role as the last layers of the various supervised learning backbones, which transforms the feature representation of the encoder into the probability distribution over $K$ dimension. The difference is that this probability distribution is meaningless, which means that the projector head outputs the pseudo labels. The hidden and bottleneck dimension of the MLP in the projector head is expressed as $d_{\text{head}}^{\text{MLP}}$. Finally, similar to the expressive method in the encoder part, the calculation process of the pseudo labels is denoted as $g_\theta(y) = g_\theta(f_\theta(x)) = q_\theta(x)$, where $q = g \circ f$, and $y$ is the feature representation.

### 2.3. Self-distillation with no labels

When the sample data has no labels or only a few labels, traditional supervised learning cannot effectively train the network. We adopt the self-distillation with no labels (DINO) algorithm [56] to train the model network for the fault diagnosis in the case of few labels. DINO is a self-supervised learning framework that can be interpreted as a form of knowledge distillation. DINO’s goal is to make the network learn a feature representation that can be used for downstream tasks by exploring the information of the unlabeled data. As described previously, DINO uses two model networks to learn: the teacher and student networks, as shown in Fig.3. The teacher and student networks share the same architecture (shown in Fig.2), but parameterized by $\theta_s$ and $\theta_l$ respectively. Then, we design a self-supervised learning task called "local to global correspondence" to train the teacher and student networks. As shown in Fig.3, given a TFM, we can construct different views, or called crops, of the original image through the data augmentation (including random crop, Gaussian blur, color jittering, and solarization). Then, it should be noted that two different random crop scale parameters are utilized to obtain the global and local views separately. Each local crop contains only a small area (scale range [0.05, 1]), while the global crops cover a large area (scale range [1, 1]) of the original TFM. $s$ is the scale parameter in the random crop process. In a random crop process, we can obtain a set of different crops, namely $X'$, which contains 2 global views (denoted as $X^g = \{x^g_1, x^g_2\}$) and $N$ local views (denoted as $X^l = \{x^l_1, x^l_2, \ldots, x^l_N\}$), $X' = X^l \cup X^g$. The teacher network only accepts global crops ($X^g$) as input, while the student network passes all crops ($X'$).

Next, we will discuss the other computational details in DINO’s algorithm. We want the student network’s output to match the given teacher network. From this perspective, DINO is very similar to knowledge distillation [57]. The difference lies in the teacher network in knowledge distillation is an extensive pre-trained network. In DINO, the teacher and student networks share the same structure and have not undergone any pre-training. Then, the probability of pseudo labels $P$ can be obtained by normalizing the output of the model network through a softmax function with temperature

$$P(x)^{(l)} = \frac{\exp(q_\theta(x)^{(l)}/\tau)}{\sum_{k=1}^{K} \exp(q_\theta(x)^{(k)}/\tau)}$$  \hspace{1cm} (12)$$

where $P$ is the probability distribution with a given input $x$, $\tau$ is a temperature parameter which can control the sharpness of $P$. A larger $\tau$ smoothes $P$, while a smaller $\tau$ encourages the sharper output distribution. For example, if we set $\tau = 0$, then Eq.(12) is equivalent to the One-hot encoding. Here, we adopt different temperature parameters in the teacher and student networks, denoted as $\tau_l$ and $\tau_s$, respectively. In addition, we require that $\tau_l$ must be lower than $\tau_s$, and this design is called “sharpening.” Sharpening is to avoid the mode collapse in the training process, which will be discussed in detail later. Moreover, as shown in Fig.3, compared with the student network, another design, called “centering”, has been added to the teacher network. The specific approach is to add a bias item $c \in \mathbb{R}^K$ to the output of the teacher network and update it among different batches based on the Exponential Moving Average (EMA) approach.
A Wavelet Transform and self-supervised learning-based framework for bearing fault diagnosis with limited labeled data

Figure 3: DINO’s overall process.

in the training process, which can be described as Eq.(13)

\[
q_{\theta_t}(x) \leftarrow q_{\theta_t}(x) + c
\]

\[
c \leftarrow m_c c + (1 - m_c) \frac{1}{B} \sum_{i=1}^{B} q_{\theta_t}(x[i])
\]  

(13)

where \(m_c \in [0, 1]\) is the momentum parameter in the centering, \(B\) is batch size. In the same way as sharpening, centering can also avoid the mode collapse, whose impact will be presented later. Based on the above statement, the goal of DINO is to minimize the loss function shown in Eq.(14)

\[
\mathcal{L}_{\theta_s, \theta_t} = \sum_{x_t \in X^t} \sum_{x_s \in X^s} \mathcal{L}_{ce}(P_s(x_s), P_t(x_t))
\]  

(14)

where \(\mathcal{L}_{ce}\) is the cross-entropy between the \(P_s\) and \(P_t\), and \(\mathcal{L}_{ce}(a, b) = -a \ln b\). We call the \(\mathcal{L}_{\theta_s, \theta_t}\) as “target entropy”.

Finally, we explain the parameter updating process of the teacher and student networks. As shown in Fig.3, \(\theta_s\) is iteratively updated by the backpropagation. Furthermore, when constructing the computation graph, we apply the stop-gradient in the teacher network, that is, \(\theta_t\) is not updated based on the loss function’s gradient. In each training epoch, the iteration form of \(\theta_t\) is similar to Eq.(13), EMA approach can be described as

\[
\theta_t \leftarrow m \theta_t + (1 - m) \theta_s
\]  

(15)

where \(m \in [0, 1]\) is the momentum parameter. The process of DINO is shown in Algorithm 1. As the training process goes on, \(\theta_t\) and \(\theta_s\) keep iterating and updating. When the training is over, all the encoder parameters in the teacher network and student networks are frozen, and the feature representation \(q_\theta(x)\) of the given TFM can be obtained.

2.4. Overview

We propose a new intelligent fault diagnosis method for the rolling bearings based on the DINO algorithm and ViT model to solve the fault diagnosis problem under the limited labeled data condition. The process framework is illustrated in Fig.4, and its specific stages can be described as follows

1) Converting the original vibration signals into TFMs through the WT and resize operation.
Algorithm 1: The process of DINO algorithm

**Input:** training TFMs dataset (without labels), Dataloader with batch size \( B \)

**Initialization:** teacher network \( q_\theta_t \), student network \( q_\theta_s \), bias term in centering \( c \)

**Define:** data augmentation strategy for global views \( \text{Aug}_g(\cdot) \), data augmentation strategy for local views \( \text{Aug}_l(\cdot) \)

1: for epoch in range(max_epoch) do
2:  
3:     for \( x \) in Dataloader do
4:         Obtain the global crops: \( X^g = \text{Aug}_g(x) \)
5:         Obtain the local crops: \( X^l = \text{Aug}_l(x) \), and let \( X = X^g \cup X^l \)
6:         Centering operation: \( q_t(x) \leftarrow q_t(x) + c \)
7:         Calculate the target entropy: \( H_s, t = \sum_{x_t \in X^g} \sum_{x_s \in X} L_{ce}(P_s(x_s), P_t(x_t)) \)
8:         Apply the stop-gradient in the teacher network: \( \text{stop-gradient}(q_t) \)
9:         Backpropagation for the student network: \( \theta_s \leftarrow \text{optimizer}(\theta_s, \nabla L_{\theta_s, \theta_t}) \)
10:        Update the teacher network by EMA: \( \theta_t \leftarrow m \theta_t + (1 - m) \theta_s \)
11:        Update the bias term by EMA: \( c \leftarrow m_c c + (1 - m_c) \frac{1}{B} \sum_{i=1}^{B} q_t(x[i]) \)
12:    end for
13: end for

2) Constructing the model network based on the ViT.
3) Combining the limited labeled and unlabeled data and conducting the self-supervised learning via the ViT model and DINO algorithm.
4) Freezing and saving the model networks’ parameters. Adopting the encoder part of the teacher network to extract the feature representation from the input TFMs.
5) Outputting the fault diagnosis results based on the limited labeled data and K-nearest neighbor (KNN) classifier.

3. Experimental setup

3.1. Datasets description

In this paper, we adopt the CWRU dataset [Case western reserve university], which is collected by the Bearing Data Center of Case Western Reserve University (CWRU), and the XJTU dataset [59], which is provided by the Xi’an Jiaotong University (XJTU), to train the model network and evaluate the effectiveness of the proposed method. Both are publicly available rolling bearing fault datasets widely used by many researchers. The following is a brief Introduction to the datasets.

1) The CWRU dataset is a rolling bearing prefabricated fault dataset. The dataset is composed of multivariate vibration signals generated by a bearing test-rig, as presented in Fig.5(a). The vibration signals are measured by the acceleration sensor with 12kHz sampling rates. We adopt the bearing data from the drive end of the motor (bearing type 6205-2RS JEM SKF) in this study, and three fault types are processed with the Electro-Discharge Machining (EDM), including inner race fault (IR), outer race fault (OR), and rolling ball fault (RB). Besides, various fault diameters ranging from 7 inches to 21 mils are introduced in each fault type, separately. In summary, the CWRU dataset contains ten failure modes (including the normal condition, denoted as NC). Then, based on the resampling method, the specific composition of the CWRU dataset in this paper is shown in Tab.1, where only 1% limited labeled data is involved.

2) The XJTU dataset comprises complete run-to-failure data of 15 rolling element bearings (same type, LDK UER204) acquired by conducting many accelerated degradation experiments, whose experimental device is shown in Fig.5(b). Vibration signals of the tested bearings are obtained by two acceleration sensors in horizontal and vertical directions with 25.6kHz sampling rates. In this paper, the recorded data from Bearing 3_1, Bearing 3_2, and Bearing 2_3 are chosen as the analysis data, which contains four fault modes: Outer race (OR), Inner race, ball, cage and outer race (IBCO), Inner race (IR), and Cage. Then, similar to the CWRU dataset, the XJTU dataset after resampling is shown in Tab.1, which also contains only 1% labeled data.

3.2. Parameters setup and other details

During the training process, the structural parameters of the model network adopted in this study are shown in Tab.2. The backbone of the encoder follows a lightweight ViT architecture, where the patch size of the input TFM \( P \)
**Figure 4:** The overall framework of the proposed method.

**Table 1**
Composition of the datasets

| Dataset | Fault modes | Sample size | The number of labeled data | The number of unlabeled data |
|---------|-------------|-------------|---------------------------|-----------------------------|
| CWRU    | NC, IR007 to IR021 OR007 to OR021, RB007 to RB021 | 1000 × 10 | 10 × 10 | 990 × 10 |
| XJTU    | OR, IBCO, IR, Cage | 2000 × 4 | 20 × 4 | 1980 × 4 |

...
A Wavelet Transform and self-supervised learning-based framework for bearing fault diagnosis with limited labeled data

Figure 5: Test rig of the datasets. (a) CWRU dataset. (b) XJTU dataset.

| Name     | Input size | $P$ | $d$ | $h$ | $d_K = d_V$ | $d_{MLP}$ | $depth$ | $K$ | $n$ | $d_{MLP}^{head}$ |
|----------|------------|-----|-----|-----|-------------|------------|---------|-----|-----|------------------|
| Value    | $224 \times 224 \times 3$ | 16  | 192 | 3   | 64          | $4 \times 192$ | 12      | 1024 | 3   | (2048, 2048, 256) |

Table 2
Structural parameters of the model network.

| Name     | $m$ | $m_c$ | $\tau_t$ | $\tau_i$ | $[0.05, s], [s, 1], s$ | $N$ |
|----------|-----|-------|----------|----------|-------------------------|-----|
| Value    | 0.996 $\rightarrow$ 1 | 0.9   | 0.04     | 0.1      | 0.4                     | 8   |

Table 3
Parameters of the DINO algorithm.

views is $[s, 1]$. The number of local crops $N$ is 8. The above parameters are adjustable hyperparameters, and we will discuss their impact on fault diagnosis performance in detail later.

Finally, we adopt the Adam optimizer to minimize the target entropy in Eq.(14), and set the weight decay equal to 0.04 to introduce the regularization. The batch size $B$ is 64, and the max training epoch is set to 100. Then, we use the warm-up strategy combined with cosine schedule to adjust the learning rate. At the first ten training epochs, the learning rate increases linearly from $1 \times 10^{-6}$ to $1.25 \times 10^{-4}$, and the subsequent learning rate follows the cosine schedule. The hardware environment is Ryzen 3995WX, NVIDIA RTX 3090, and we adopt Python 3.8, Pytorch 1.8.1, and CUDA 10.2 for the deep learning framework.

4. Experimental results and discussions

4.1. Mode collapse problem

In self-supervised learning, mode collapse is a crucial problem. Mode collapse refers to the situation in which the network gradually converges to trivial solutions due to the abnormal training process. For example, the trained adversarial generation network’s generator can only generate one kind of image. As shown in Fig.2 and Fig.3, DINO’s goal is to make the $K$-dimensional pseudo labels of the student network match the teacher network. Based on this, there are two forms of the collapse in the proposed method: 1) the outputs of the teacher and student networks are evenly distributed in each dimension (namely over-alignment), that is, the probability value of each pseudo label is $\frac{1}{K}$. 2) the output of the teacher and student networks is 1 in one dimension and 0 in all others (namely over-uniformity), such as $[0, 1, 0, 0, ..., 0]$. Then, as described previously, we introduce centering and sharpening to avoid collapse, and we will study their respective roles in this section. Firstly, the cross entropy in Eq.(14) can be decomposed into Kullback-Leibler


\[
\sum_{x_i \in X^s} \sum_{x_j \in X'} L_{ce}(P_s(x_i), P_t(x_j)) = \sum_{x_i \in X^s} \sum_{x_j \in X'} D_{KL}(P_s(x_i)\|P_t(x_j)) + \sum_{x_i \in X^s} h(P_t(x_i))
\]

(16)

where \( D_{KL} \) is the KL divergence between \( P_t \) and \( P_s \), and \( h \) is the entropy of the teacher network’s output \( P_t \). \( D_{KL} \) can calculate the match between \( P_t \) and \( P_s \). If \( P_t \) and \( P_s \) are identical, then \( D_{KL} \) equals zero, which means the mode collapse. Besides, \( h \) can measure the uncertainty of \( P_t \), which can be given as

\[
h(P_t) = \sum_{i=1}^{K} -P_t^{(i)} \ln P_t^{(i)}
\]

(17)

where \( P_t^{(i)} \) is \( i \)-th element in \( P_t \). As mentioned above, the value of \( D_{KL} \) can be used to judge whether the mode collapse occurs. Furthermore, we can determine the form of the mode collapse based on the value of \( h \). It can be seen from Eq.(17) that \( h \) will converge to \( \ln K \) for over-uniformity case, and will converge to zero for over-alignment case.

Subsequently, we present the evolution of \( L_{\theta_s,\theta_t}, D_{KL} \) and \( h \) under different datasets and various designs in Fig.6. To analyze the respective roles of the centering and sharpening, we set up four different designs: 1) The output of the teacher network goes through both centering and sharpening operations (namely Both), as shown in Fig.3. 2) There is only centering in the DINO framework. 3) There is only sharpening in the DINO framework. 4) The output of the teacher network is identical to the student network without any additional modifications (namely Neither). As shown in Fig.6(a) and (e), in the case of only centering, the target entropy will stay at an enormous value during the entire training process. Conversely, in the case of only sharpening, the target entropy converges rapidly to zero. Both cases indicate that the model networks have not been trained normally. Further, it can be seen from Fig.6(b) and (f) that the KL divergence between the teacher and student outputs degrades to zero promptly in all cases except Both, which means \( P_t = P_s \), indicating the mode collapse. However, the entropy converges to different values under the diverse designs. It can be seen from the data in Fig.6(c) and (g) that if there is only centering operation, the entropy of the teacher network output \( h \) will converge to \( \ln K \), indicating the over-uniformity. And \( h \) will converge to 0 if there is only sharpening in the DINO framework, which means the over-alignment case. In the absence of centering and sharpening, the result is the same as that of only centering, and over-uniformity occurs. Finally, from the blue lines in Fig.6 we can see that \( L_{\theta_s,\theta_t} \) and \( h \) show a tendency to decline gradually during training, and \( D_{KL} \) do not converge to 0, indicating that there is no mode collapse. These results suggest that the teacher network and the student network can carry out normal training only in the case of Both.

In summary, for the informants in this section, it can be observed that the centering and sharpening operations are complementary in avoiding the mode collapse. Missing either operation will cause the mode collapse in the proposed method. Centering can inhibit over-alignment, but encourage over-uniformity. Sharpening has the opposite effect, which can avoid over-uniformity but promote over-alignment. In addition, mode collapse will also occur if there is no centering or sharpening. Only by applying both operations simultaneously can the model network be appropriately trained.

4.2. Fault diagnosis results based on the proposed method

Based on the DINO algorithm, the model networks are trained on the CWRU and XJTU datasets without any labels. The parameters in the encoder part of the teacher network are frozen, and we can use it to extract the feature representation of the input TFM. The extracted feature representation is called feature vectors. Then, to diagnosis the input TFM, we calculate its feature vectors and compare it against the labeled data. Adopting the above process, we can make fault diagnose for many unlabeled data with the KNN classifier, even if the labeled data is minimal. Tab.4 provides the fault diagnosis results obtained by our proposed method. We select multiple the number of nearest neighbors (denoted as \( N_k^t \)), ranging from 10 to 100, to optimize the classifier parameters. In addition, to enhance the robustness of the classifier to \( N_k \), we introduce a temperature parameter \( \tau_k \) as a divisor when calculating the similarity, similar to Eq.(12). The effect of \( \tau_k \) is also set out in Tab.4. We find that \( N_k \) has a significant impact on the fault diagnosis accuracy of the KNN classifier if without \( \tau_k \). On the CWRU dataset, the accuracy of the classifier is 92.65% with \( N_k = 10 \), but when \( N_k \) equals 40, the accuracy drops to 79.19%. Analogously, the fault diagnosis accuracy of KNN is 83.08% with \( N_k = 10 \) and 74.07% with \( N_k = 40 \) on the XJTU dataset. These results show that the fault diagnosis performance of the classifier will be sensitive to \( N_k \) if there is no temperature parameter. Furthermore,
Table 4
Fault diagnosis results with only 1% labeled data (unit: %).

| Dataset \ $N_k$ | 10  | 20  | 30  | 40  | 50  | 60  | 70  | 80  | 90  | 100 |
|----------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| CWRU with $r_k$ | **95.27** | 93.69 | 93.44 | 93.46 | 93.47 | 93.47 | 93.47 | 93.47 | 93.47 | 93.47 |
| CWRU without $r_k$ | 92.65 | 81.42 | 79.22 | 79.19 | 79.49 | 80.05 | 80.58 | 80.73 | 81.22 | 81.22 |
| XJTU with $r_k$ | **89.33** | 88.32 | 88.40 | 88.37 | 88.37 | 88.37 | 88.37 | 88.37 | None | None |
| XJTU without $r_k$ | 83.08 | 75.74 | 74.44 | 74.07 | 75.25 | 75.34 | 75.52 | 75.77 | None | None |

the first and third rows in Tab.4 illustrate that $r_k$ can obviously improve the robustness of the classifier to $N_k$, and increase the fault diagnosis accuracy. Taking the CWRU dataset as an example, for the different values of $N_k$, the top and minimum accuracy of the classifier are 95.27% and 93.44%, and their fluctuation range is significantly reduced compared with the data without $r_s$. Then, by applying $r_s$, the accuracy of KNN on the CWRU and XJTU datasets achieve 95.27% and 89.33% with $N_k = 10$, respectively. This is a brilliant fault diagnosis result under the limited labeled data condition. Moreover, it should be pointed out that the feature vectors extracted by the trained encoder do not undergo any additional fine-tuning, and great fault diagnosis accuracy shows that feature vectors possess good generalization ability.

To analyze the detailed fault diagnosis results, Fig.7 provides the confusion matrices of the proposed method on the CWRU and XJTU datasets, where rows represent the actual fault mode and columns is the predicted fault mode. Specifically, on the CWRU dataset, promising fault diagnosis accuracy can be obtained in most fault modes, and the misidentification results mainly focus on the class RB014. This result may be explained by the fact that the features of rolling ball fault condition is general weak. However, we can still obtain higher than 77% identification accuracy with...
only 1% labeled data. In contrast, due to the earlier fault stage of the rolling bearing during the run-to-failure process, the proposed method achieves relatively lower diagnosis accuracy on the XJTU dataset. About 89.33% testing accuracy with the unlabeled data can be acquired, and 28.74% OR samples are misclassified as IR. This result indicates that the outer race fault in early stage may be a challenging fault mode. But, over 70% accuracy for OR can be still achieved. That shows that the proposed method has satisfactory fault diagnosis performance in the case of limited labeled data.

Overall, through the results of fault diagnosis accuracy and confusion matrices under the different datasets, the effectiveness of the proposed method is validated in this section. Feature vectors with strong generalization ability can be acquired utilizing the model networks and DINO algorithm, and high diagnosis accuracy is obtained with only 1% labeled data.

4.3. Influence of the hyperparameters

In this section, we will investigate the influence of the hyperparameters in the proposed method, including the encoder form, pseudo labels dimension $K$, temperature parameters in the teacher and student networks ($t$ and $s$), scale parameter $s$, momentum parameter in centering $m_c$, and the number of local crops $N$. The impact of the hyperparameters is evaluated from two aspects: accuracy and peak memory when running. The accuracy directly
reflected the fault diagnosis performance, while the peak memory can represent the model scale and computational complexity. These results are summarized in Tab.5. Firstly, the result in group A shows that we can obtain a higher accuracy based on the feature vectors extracted by the teacher network. A possible explanation for this might be that the outputs of the teacher network are sharper ($\tau_t < \tau_s$), which helps the model learn features with more generalization ability. Next, as one of the critical hyperparameters in the proposed method, the effects of the pseudo labels dimension $K$ are discussed in group B in Tab.5. One unanticipated finding is that a larger $K$ may not imply better performance. For example, the fault diagnosis accuracy of $K = 64$ is higher than that of $K = 512$. This result suggests that the performance of the proposed method may be further improved by fine-tuning $K$. Within the range that we set in this study, the accuracy is highest when $K$ equals 1024. Meanwhile, the data in group B also show that the size of $K$ has no significant impact on the peak memory. This is because $K$ is associated only with the last layer of the projector head, which is a negligible value compared to the entire model network.

Besides, the relationship between the temperature parameter of teacher network and the testing accuracy is presented in group C in Tab.5. We observe that the model is susceptible to $\tau_t$, and inappropriate $\tau_t$ will significantly reduce the model performance, even leading to a failed training process. Too small $\tau_t$ will make the teacher output too sharp, which causes the network easily fall into the local minimum, resulting in the decline of diagnosis accuracy. Nevertheless, as mentioned in Section 4.1, we need sharpening to avoid the mode collapse, so a too large $\tau_t$ is also unfitting. Actually, through the experiment, we find that when the teacher temperature is higher than 0.06 (such as 0.8), the entropy of teacher output consistently converges to $\ln K$, indicating over-uniformity. Therefore, the fault diagnosis performance is inferior with an overlarge $\tau_t$.

Next, group D in Tab.5 compares the testing accuracy and peak memory of the proposed method under diverse student temperature $\tau_s$. As described previously, the design of sharpening demands that $\tau_t$ must be lower than $\tau_s$ to avoid the mode collapse, further confirmed by the result in group D. If $\tau_t$ is higher than or close to $\tau_s$, such as 0.03 or 0.05, sharpening will fail, and the poor diagnosis performance will be obtained. In our experiment, the student temperature below 0.05 is not recommended. In addition, similar to $\tau_t$, oversize $\tau_s$ is able to reduce the fault diagnosis accuracy. Hence, a moderate $\tau_s$ is suggested in practical implementations.

Scale range $s$ is also an essential parameter that affects the fault diagnosis performance of the proposed method, and the testing results with different $s$ are presented in group E. A general pattern summarized from group E is that higher diagnosis accuracy can be acquired with a medium size of $s$. A lower $s$ means smaller areas of the original TFM adopted to generate the global and local crops, which encourages the model networks to concentrate on the local features of the input. On the contrary, a larger $s$ helps our proposed method to learn the global features of the TFM. Both local and global features are necessary for the fault diagnosis problem, so a moderate $s$ should be selected to balance their effects. In our study, the highest accuracy can be achieved with $s = 0.4$. Moreover, since the local and global crops are resized to a fixed size, $s$ does not influence the computational complexity.

Per our earlier discussion, there are two additional designs in the teacher output to prevent the mode collapse: sharpening and centering. $\tau_t$ and $\tau_c$ are related to sharpening, and their effects have been investigated in group C and D. $m_c$ is the momentum parameter in centering, and its impact is provided in group F. It is somewhat surprising that, compared with the temperature parameters, the convergence of the proposed method is robust to a wide range of $m_c$. We do not notice the over-uniformity or over-alignment with $m_c \in [0, 0.99]$ in our trials. The model network only occurs the model collapse when the update of bias term in centering is too slow, such as $m_c = 0.999$. Simultaneously, the maximum testing accuracy can be procured with $m_c$ equals 0.9 in our datasets.

Finally, the influence of the number of local crops $N$ is shown in group G in Tab.5. A universal display pattern is observed that a larger $N$ can improve the diagnosis performance. Nonetheless, the computational complexity of the proposed method also increases with the expansion of $N$, so we should weigh it according to the specific hardware environment in the actual application.

### 4.4. Attention maps visualization

Great interpretability is a significant advantage of the attention mechanism. This section employs attention maps (denoted as AM) visualization to explore the feature representation process qualitatively. Here, three dissimilar forms of AM are introduced: 1) Class token AM (CAM), which is calculated by the class token and embedding sequence on the heads of the last Transformer basic block in the model network. CAM can represent the importance of each patch in the input TFM to the feature representation. 2) Threshold class token AM (TAM). We visualize the marked patches obtained by thresholding the CAM to keep 90% of the attention, illustrating the areas that primarily affect the feature vectors. 3) Embedding sequence AM (EAM), computing from the embedding sequence in the last Transformer.
Table 5
Influence of the hyperparameters

| Group | Encoder form | Hyperparameters | N | Accuracy | Peak memory |
|-------|--------------|-----------------|---|----------|-------------|
| Baseline | Teacher | $K$ | $\tau_s$ | $s$ | $m_c$ | | | | |
| A | Student | 1024 | 0.04 | 0.1 | 0.4 | 0.9 | 8 | 95.27% | 6.34G |
| | | 94.34% | 6.34G |
| B | | 64 | 94.24% | 6.33G |
| | | 128 | 93.72% | 6.33G |
| | | 256 | 92.10% | 6.33G |
| | | 512 | 90.86% | 6.33G |
| C | 0.02 | 94.85% | 6.34G |
| | 0.06 | 88.79% | 6.34G |
| | 0.08 | 37.82% | 6.34G |
| D | 0.03 | 51.06% | 6.34G |
| | 0.05 | 52.00% | 6.34G |
| | 0.2 | 94.24% | 6.34G |
| | 0.4 | 86.72% | 6.34G |
| E | 0.08 | 87.91% | 6.34G |
| | 0.16 | 92.97% | 6.34G |
| | 0.24 | 93.82% | 6.34G |
| | 0.32 | 94.08% | 6.34G |
| | 0.48 | 92.09% | 6.34G |
| F | 0 | 91.32% | 6.34G |
| | 0.99 | 92.27% | 6.34G |
| | 0.999 | 63.60% | 6.34G |
| G | 0 | 86.67% | 4.27G |
| | 2 | 90.70% | 4.69G |
| | 4 | 91.58% | 5.14G |
| | 6 | 93.71% | 5.78G |

basic block. Note that there are three head in the MSA, so the EAM also can be described in three different subspaces, denoted as EAM$_1$, EAM$_2$, and EAM$_3$ respectively.

The results of the attention maps visualization obtained from the untrained and trained teacher network are compared in Fig.8. As shown in Fig.8(a), for the network adopting the random weight without any training, no salient attention value is found in CAM, and marked patches in TAM almost cover the whole input region. Additionally, the attention values in EAM also present a consistent distribution form, indicating a poor feature extraction ability. In contrast, the teacher network trained by self-supervised learning demonstrates an entirely different pattern. As can be seen from Fig.8(b), the prominent parts in TFM have more significant attention values in CAM, and the marked patches in TAM are also concentrated on the smaller areas. That is, the trained teacher network pays more attention to the patches where the amplitude in the TFM is more pronounced, which adheres to our intuition. Then, most of the attention values in EMA are centralized in a few patches, suggesting that the teacher network grasps the crucial information for fault diagnosis from the input TFM.

Taken together, through the attention maps visualization, the feature representation process in our proposed method is analyzed. Based on self-supervised learning, the teacher network spontaneously learns the fault-specific features without any labels, and it pays more attention to the patches where the amplitude in the TFM is more obvious.
4.5. Feature vectors visualization

In addition to the testing accuracy, the distribution form of feature vectors is also a meaningful indicator to evaluate the proposed fault diagnosis method. Therefore, we will focus on discussing the visualization results of feature vectors in this section.

In this study, the dimension of the feature vectors extracted by the model network is 192, which is a higher-dimensional space that cannot be visualized directly. Here, we adopt the t-distributed Stochastic Neighbor Embedding (t-SNE) technique as a dimension reduction method. As a comparison, visualization results of the input TFM are also necessary. However, it should be noted that the scale of the TFM is $224 \times 224 \times 3$, so processing it immediately with t-SNE will be very time-consuming and not effective. To address this issue, an approach combined the Principal Component Analysis (PCA) and t-SNE method is utilized. Firstly, PCA is used to decompose the first 192 principal...
components of the TFM and integrate them into a feature vector. Then, t-SNE is applied to reduce its dimension further and realize visualization. Based on the above analysis, Fig. 9 shows the feature vectors visualization results on the CWRU and XJTU datasets. A large amount of overlapping parts are observed for different fault modes in Fig. 9(a) and (c), demonstrating that it is not feasible to use the input TFM for fault diagnosis directly. Then, turning now to the visualization results of the feature vectors obtained by our proposed method. As shown in Fig. 9(b) and (d), most instances of the same fault mode are projected into the same region, and different fault modes are separated. That shows that the proposed method can extract the feature vectors with inter-class separability by exploring information in unlabeled data. However, we should point out that confusion still occurs between a few different samples due to the lack of supervised learning with labeled data, such as RB007 and RB021 in Fig. 9(b), and OR and IR in Fig. 9(d). These results are compatible with the above confusion matrix analysis results in Fig. 7, revealing some more challenging fault diagnosis tasks.

![Figure 9: Feature vectors visualization results. (a) and (b) CWRU dataset. (c) and (d) XJTU dataset. (a) and (c) The input TFMs. (b) and (d) Feature vectors extracted by our proposed method.](image)

### 4.6. Comparison with other methods

To further verify the superiority of the proposed method, comparisons with other fault diagnosis methods in practical application are necessary. Our method consists of two main parts: the training paradigm (DINO) and the backbone of the encoder (ViT). Therefore, the contrast experiments also include different training paradigms and backbones. In terms of the training paradigm, conventional supervised learning and other self-supervised learning methods are involved. The supervised learning technique only considers the cross-entropy loss of the limited labeled data, and the self-supervised learning approaches adopted for comparisons incorporate the SimCLR [60] and BYOL [61]. Then, in the aspect of the encoder backbone, ResNet18 [62] and DenseNet121 [63] are considered.

These training paradigms and encoder backbones are used for the CWRU and XJTU datasets, respectively, and their results are summarized in Tab. 6. Note that these fault diagnosis methods share the same data pre-processing pipeline shown in Fig. 1. Among them, due to the lack of labeled data, the self-supervised learning paradigm can not
Table 6  
Testing accuracy of the comparison methods on the CWRU and XJTU datasets

| Training paradigm | Encoder backbone | CWRU dataset | XJTU dataset | Average |
|-------------------|-----------------|--------------|--------------|---------|
| Supervised        | ResNet18        | 77.41%       | 63.86%       | 70.64%  |
|                   | DenseNet121     | 72.34%       | 54.63%       | 63.49%  |
|                   | ViT             | 46.86%       | 49.84%       | 48.35%  |
| SimCLR            | ResNet18        | 88.58%       | 82.75%       | 85.67%  |
|                   | DenseNet121     | 83.64%       | 87.92%       | 85.78%  |
|                   | ViT             | 87.29%       | 67.16%       | 77.23%  |
| BYOL              | ResNet18        | 91.05%       | 83.09%       | 87.07%  |
|                   | DenseNet121     | 89.91%       | 83.94%       | 86.93%  |
|                   | ViT             | 73.53%       | 63.84%       | 68.69%  |
| DINO              | ResNet18        | 93.56%       | 90.85%       | 92.21%  |
|                   | DenseNet121     | 92.90%       | 85.83%       | 89.37%  |
|                   | ViT             | 95.27%       | 89.33%       | 92.30%  |

effectively train the networks. It is difficult to obtain the feature vectors with preferable generalization ability through self-supervised learning, so its testing accuracy is relatively low. By contrast, better fault diagnosis performance can be acquired by training the model network with the self-supervised learning method. Regardless of the encoder backbone, the testing accuracy obtained by DINO are generally higher than those of other self-supervised learning method, suggesting that DINO is a more powerful training paradigm. Besides, surprisingly, adopting the ResNet18 rather than ViT as the encoder backbone can get a higher testing accuracy on the XJTU dataset. But overall, among all these solutions, ViT trained by DINO achieves the highest average diagnosis accuracy on the given datasets, which further proves the effectiveness and superiority of the proposed method for the rolling bearing fault diagnosis with limited labeled data.

5. Conclusions

Faced with the contradiction between the conventional supervised fault diagnosis methods’ dependency on massive labeled data and engineering practice, a Wavelet Transform (WT) and self-supervised learning-based fault diagnosis framework for bearing fault diagnosis with limited labeled data has been presented. In this method, the original vibration signals are pre-processed by WT and cubic spline interpolation to obtain the time-frequency maps (TFMs) with a specific scale. The teacher and student networks are established based on the Vision Transformer (ViT) encoder and projector head, and a pretext task called "local to global correspondence" is introduced for self-supervised learning. Adopting the Self-distillation with no labels (DINO) algorithm to train the teacher and student networks, fault-specific feature representation can be obtained. The main conclusions are summarized as follows.

1) Through the DINO algorithm, effective feature vectors from the complex TFMs can be extracted by the trained ViT encoder, and involving the centering and sharpening operations in the teacher network can avoid the problem of mode collapse efficaciously during the self-supervised learning procedure. The complementary effect of the centering and sharpening is observed, where centering encourages over-uniformity but inhibits over-alignment, while sharpening has the opposite function. Only by applying both operations simultaneously can the mode collapse be avoided.

2) In the situation that only 1% labeled samples are included in the CWRU and XJTU datasets, adopting the feature vectors extracted by the trained encoder without any fine-tuning, 95.27% and 89.33% testing accuracy can be obtained based on the simple K-Nearest Neighbor (KNN) classifier.

3) The influence of hyperparameters is discussed in detail. Teacher temperature and student temperature can significantly affect the fault diagnosis performance. Inappropriate values of them may directly cause the mode collapse, so a careful adjustment is recommended. Then, the scale range and momentum parameter in centering are also crucial, and there is an optimal value for both hyperparameters. The computational complexity and diagnosis accuracy of the proposed framework increase with the expansion of the number of local crops, so it should be weighed according to the specific hardware environment in the actual application. Finally, the explicit law of the influence of the pseudo labels...
A Wavelet Transform and self-supervised learning-based framework for bearing fault diagnosis with limited labeled data dimension is not observed, but on the whole, it has a relatively unapparent impact on the fault diagnosis accuracy of the proposed framework.

4) Different training paradigms, such as supervised learning, SimCLR, BYOL, DINO, and various encoder backbones, including ResNet18, DenseNet121, and ViT, are considered in twelve comparative fault diagnosis approaches. Among them, the proposed method has the highest average accuracy on the CWRU and XJTU datasets, demonstrating its effectiveness and superiority.

Since sufficient unlabeled data is needed in this study, the main limitation lies in that the proposed method is not suitable for the small sample. In future work, further research should focus on the improved diagnosis approach for small-sample learning.

Acknowledgements

It is very grateful for the financial supports from the National Major Science and Technology Projects of China (No. 2017-IV-0008-0045), the National Natural Science Foundation of China (Nos. 11972129, 11732005) and the Fundamental Research Funds for the Central Universities.

References

[1] S. Nandi, H. Toliyat, X. Li, Condition monitoring and fault diagnosis of electrical motors: A review, IEEE Transactions on Energy Conversion 20 (2005) 719–729.
[2] X. Zhao, M. Jia, J. Bin, T. Wang, Z. Liu, Multiple-order graphical deep extreme learning machine for unsupervised fault diagnosis of rolling bearing, IEEE Transactions on Instrumentation and Measurement 70 (2021) 1–12.
[3] Y. Liu, Y. Zhao, J. Li, H. Ma, Q. Yang, X. Yan, Application of weighted rating contribution of nonlinear output frequency response functions to rotor rub-impact, Mechanical Systems and Signal Processing 136 (2020) 106518.
[4] X. Jiang, C. Shen, J. Shi, Z. Zhu, Initial center frequency-guided vmd for fault diagnosis of rotating machines, Journal of Sound and Vibration 435 (2018) 36–55.
[5] A. Glowacz, R. Tadeusiewicz, S. Legutko, W. Caesarendra, M. Irfan, H. Liu, F. Brumercik, M. Gutter, M. Sulowicz, J. A. Antonino Daviu, T. Sarkodie-Gyan, P. Fracz, A. Kumar, J. Xiang, Fault diagnosis of angle grinders and electric impact drills using acoustic signals, Applied Acoustics 179 (2021) 108070.
[6] Z. Wang, N. Yang, N. Li, W. Du, J. Wang, A new fault diagnosis method based on adaptive spectrum mode extraction, Structural Health Monitoring 20 (2021) 3354–3370.
[7] Z. Wang, W. Zhao, W. Du, N. Li, J. Wang, Data-driven fault diagnosis method based on the conversion of erosion operation signals into images and convolutional neural network, Process Safety and Environmental Protection 149 (2021) 591–601.
[8] X. Yan, M. Jia, A novel optimized svm classification algorithm with multi-domain feature and its application to fault diagnosis of rolling bearing, Neurocomputing 313 (2018) 47–64.
[9] T. Han, L. Zhang, Z. Yin, A. C. Tan, Rolling bearing fault diagnosis with combined convolutional neural networks and support vector machine, Measurement 177 (2021) 109022.
[10] H. Yuan, N. Wu, X. Chen, Y. Wang, Fault diagnosis of rolling bearing based on shift invariant sparse feature and optimized support vector machine, Machines 9 (2021).
[11] D. Xiao, J. Ding, X. Li, L. Huang, Gear fault diagnosis based on kurtosis criterion vmd and som neural network, Applied Sciences 9 (2019) 5424–5449.
[12] H. Fan, Y. Yan, X. Zhang, X. Cao, J. Ma, Composite fault diagnosis of rolling bearing based on optimized wavelet packet ar spectrum energy entropy combined with adaptive no velocity term pso-som-bpnn, JOURNAL OF SENSORS 2021 (2021) 4138652.
[13] W. Mao, W. Feng, Y. Liu, D. Zhang, X. Liang, A new deep auto-encoder method with fusing discriminant information for bearing fault diagnosis, Mechanical Systems and Signal Processing 150 (2021) 107233.
[14] D. Yang, H. R. Karimi, K. Sun, Residual wide-kernel deep convolutional auto-encoder for intelligent rotating machinery fault diagnosis with limited samples, Neural Networks 141 (2021) 133–144.
[15] A. Afia, C. Rahmoune, B. Djamel, B. Merainani, S. Fedala, New intelligent gear fault diagnosis method based on autocorrelation and radial basis function neural network, Advances in mechanical engineering 12 (2020) 1687814020916593.
[16] L. Yang, H. Chen, Fault diagnosis of gearbox based on rbf-pf and particle swarm optimization wavelet neural network, Neural Computing & Applications 31 (2019) 4463–4478.
[17] B. Ding, J. Wu, S. Chuang, S. Wang, X. Chen, Y. Li, Sparsity-assisted intelligent condition monitoring method for aero-engine main shaft bearing, Transactions of Nanjing University of Aeronautics & Astronautics 37 (2020) 508–516.
[18] J. Yang, W. Bao, Y. Liu, X. Li, J. Wang, Y. Niu, J. Li, Joint pairwise graph embedded sparse deep belief network for fault diagnosis, Engineering Applications of Artificial Intelligence 99 (2021) 104149.
[19] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, D. J. Inman, 1d convolutional neural networks and applications: A survey, Mechanical Systems and Signal Processing 151 (2021) 107398.
[20] A. Kumar, G. Vashishtha, C. P. Gandhi, Y. Zhou, A. Glowacz, J. Xiang, Novel convolutional neural network (ncnn) for the diagnosis of bearing defects in rotary machinery, IEEE Transactions on Instrumentation and Measurement 70 (2021) 1–10.
[21] X. Wang, H. Gu, T. Wang, W. Zhang, A. Li, F. Chu, Deep convolutional tree-inspired network: a decision-tree-structured neural network for hierarchical fault diagnosis of bearings, Frontiers of Mechanical Engineering 16 (2021) 814–828.
A Wavelet Transform and self-supervised learning-based framework for bearing fault diagnosis with limited labeled data

[22] S. Liu, C. Shen, Z. Chen, W. Huang, Z. Zhu, A sudden fault detection network based on time-sensitive gated recurrent units for bearings, Measurement 186 (2021) 110214.

[23] M. Qiao, S. Yan, X. Tang, C. Xu, Deep convolutional and lstm recurrent neural networks for rolling bearing fault diagnosis under strong noises and variable loads, IEEE Access 8 (2020) 66257–66269.

[24] J. Li, Y. Liu, Q. Li, Intelligent fault diagnosis of rolling bearings under imbalanced data conditions using attention-based deep learning method, Measurement 189 (2022) 110500.

[25] X. Pei, X. Zheng, J. Wu, Rotating machinery fault diagnosis through a transformer convolution network subjected to transfer learning, IEEE Transactions on Instrumentation and Measurement 70 (2021) 1–11.

[26] X. Du, L. Jia, I. Ul Haq, Fault diagnosis based on spbo-sdae and transformer neural network for rotating machinery, Measurement 188 (2022) 110545.

[27] W. Wang, Y. Le, T. Yan, N. Li, A. Nandi, Residual convolution long short-term memory network for machines remaining useful life prediction and uncertainty quantification, Journal of Dynamics, Monitoring and Diagnostics 1 (2021) 2–8.

[28] Z. Zhao, T. Li, B. An, S. Wang, B. Ding, R. Yan, X. Chen, Model-driven deep unrolling: Towards interpretable deep learning against noise attacks for intelligent fault diagnosis, ISA Transactions (2022).

[29] J. Yang, W. Bao, X. Li, Y. Liu, Improved graph-regularized deep belief network with sparse features learning for fault diagnosis, Neural Computing & Applications (2022).

[30] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 42 (2020) 2011–2023.

[31] M. Zhao, S. Song, X. Fu, B. Tang, M. Pecht, Deep residual shrinkage networks for fault diagnosis, IEEE Transactions on Industrial Informatics 16 (2020) 4681–4690.

[32] G. Gao, Y. Zhi, Z. Ren, K. Yan, W. Kang, A novel weak fault diagnosis method for rolling bearings based on lstm considering quasi-periodicity, Knowledge-Based Systems 231 (2021) 107143.

[33] X. Ding, M. Jia, Q. Mao, Y. Cao, A novel time-frequency transformer based on self-attention mechanism and its application in fault diagnosis of rolling bearings, Mechanical Systems and Signal Processing 168 (2022) 108616.

[34] D. Hoang, J. Kang, A survey on deep learning-based bearing fault diagnosis, Neurocomputing 335 (2019) 327–335.

[35] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, R. X. Gao, Deep learning and its applications to machine health monitoring, Mechanical Systems and Signal Processing 115 (2019) 213–237.

[36] R. Bai, Q. Xu, Z. Meng, L. Cao, K. Xing, F. Fan, Rolling bearing fault diagnosis based on multi-channel convolution neural network and multi-scale clipping fusion data augmentation, Measurement 184 (2021) 109885.

[37] Y. Zhang, X. Li, L. Gao, L. Wang, L. Wen, Imbalanced data fault diagnosis of rotating machinery using synthetic oversampling and feature learning, Journal of Manufacturing Systems 48 (2018) 34–50. Special Issue on Smart Manufacturing.

[38] T. Zheng, L. Song, J. Wang, W. Teng, X. Xu, C. Ma, Data synthesis using dual discriminator conditional generative adversarial networks for imbalanced fault diagnosis of rolling bearings, Measurement 158 (2020) 107741.

[39] W. Wan, S. He, J. Chen, A. Li, Y. Feng, Qscgan: An un-supervised quick self-attention convolutional gan for lre bearing fault diagnosis under limited label-lacked data, IEEE Transactions on Instrumentation and Measurement 70 (2021) 1–16.

[40] H. Wang, Z. Liu, T. Ai, Long-range dependencies learning based on non-local 1d-convolutional neural network for rolling bearing fault diagnosis, Journal of Dynamics, Monitoring and Diagnostics (2022).

[41] X. Li, W. Zhang, Q. Ding, J. Sun, Intelligent rotating machinery fault diagnosis based on deep learning using data augmentation, Journal of Intelligent Manufacturing 31 (2020) 433–452.

[42] L. Ma, Y. Ding, Z. Wang, C. Wang, J. Ma, C. Lu, An interpretable data augmentation scheme for machine fault diagnosis based on a sparsity-constrained generative adversarial network, Expert Systems with Applications 182 (2021) 115234.

[43] M. A. Talab, S. Awang, S. A.-d. M. Najim, Super-low resolution face recognition using integrated efficient sub-pixel convolutional neural network(espcn) and convolutional neural network(cnn), in: 2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS), 2019, pp. 331–335. doi: 10.1109/I2CACIS.2019.8825083.

[44] X. Zhao, M. Jia, J. Jin, B. Tian, W. Li, Z. Liu, Multiple-order graphical deep extreme learning machine for unsupervised fault diagnosis of rolling bearing, IEEE Transactions on Instrumentation and Measurement 70 (2021) 1–12.

[45] S. Liu, J. Chen, S. He, E. Xu, H. Lv, Z. Zhou, Intelligent fault diagnosis under small sample size conditions via bidirectional infomax gan with unsupervised representation learning, Knowledge-Based Systems 232 (2021) 107488.

[46] Y. Ding, J. Zhuang, P. Ding, M. Jia, Self-supervised pretraining via contrast learning for intelligent incipient fault detection of bearings, Reliability Engineering & System Safety 218 (2022) 108126.

[47] M. Wei, Y. Liu, T. Zhang, Z. Wang, J. Zhu, Fault diagnosis of rolling machinery via contrast based self-supervised learning method and very few labeled samples, Sensors 22 (2022) 192.

[48] T. Wang, M. Qiao, M. Zhang, Y. Yang, H. Snoussi, Data-driven prognostic method based on self-supervised learning approaches for fault detection, Journal of Intelligent Manufacturing 31 (2020) 1611–1619.

[49] S. Zhang, F. Ye, B. Wang, T. G. Habetler, Semi-supervised bearing fault diagnosis and classification using variational autoencoder-based deep generative models, IEEE Sensors Journal 21 (2021) 6476–6486.

[50] Y. Wu, R. Zhao, W. Jin, T. He, S. Ma, M. Shi, Intelligent fault diagnosis of rolling bearings using a semi-supervised convolutional neural network, Applied Intelligence 51 (2021) 2144–2160.

[51] C. Jian, K. Yang, Y. Ao, Industrial fault diagnosis based on active learning and semi-supervised learning using small training set, Engineering Applications of Artificial Intelligence 104 (2021) 104365.

[52] W. Zhang, X. Li, H. Ma, Z. Luo, X. Li, Federated learning for machinery fault diagnosis with dynamic validation and self-supervision, Knowledge-Based Systems 213 (2021) 106679.
A Wavelet Transform and self-supervised learning-based framework for bearing fault diagnosis with limited labeled data

[53] G. Li, J. Wu, C. Deng, M. Wei, X. Xu, Self-supervised learning for intelligent fault diagnosis of rotating machinery with limited labeled data, Applied Acoustics 191 (2022) 108663.

[54] X. Li, X. Li, H. Ma, Deep representation clustering-based fault diagnosis method with unsupervised data applied to rotating machinery, Mechanical Systems and Signal Processing 143 (2020) 106825.

[55] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, 2021. arXiv:2010.11929.

[56] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9630–9640. doi:10.1109/ICCV48922.2021.00951.

[57] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, 2015. URL: https://arxiv.org/abs/1503.02531. doi:10.48550/ARXIV.1503.02531.

[Case western reserve university] Case western reserve university, Case western reserve university bearing data center, [Online]. http://csegroups.case.edu/bearingdatacenter/home Accessed: Aug. 2021.

[59] B. Wang, Y. Lei, N. Li, N. Li, A hybrid prognostics approach for estimating remaining useful life of rolling element bearings, IEEE Transactions on Reliability 69 (2020) 401–412.

[60] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, PMLR, 2020, pp. 1597–1607. URL: https://proceedings.mlr.press/v119/chen20j.html.

[61] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot, k. kavukcuoglu, R. Munos, M. Valko, Bootstrap your own latent - a new approach to self-supervised learning, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 21271–21284. URL: https://proceedings.neurips.cc/paper/2020/file/f3ada80d5c4ee70142b17b8192b2968e-Paper.pdf.

[62] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[63] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261–2269. doi:10.1109/CVPR.2017.243.