CoWS: Self-Supervised Representation Pre-Training for Cross-Machine Fault Diagnosis

Xiaohan Chen, Rui Yang, *Senior Member, IEEE,*, Yihao Xue, Zidong Wang, *Fellow, IEEE*

***Abstract*—Deep learning-based techniques have emerged as powerful tools for fault diagnosis. However, conventional deep learning methods require designing and training fault diagnosis models from scratch for each machine, necessitating a large-scale and high-quality labeled dataset. This limitation hinders the application of deep learning in practical settings, where labeled data is scarce and data distribution varies across different machines. To address these challenges, a novel self-supervised learning framework specifically designed for time-series data is proposed, aimed at enhancing cross-machine downstream fault diagnosis tasks with limited data. The proposed framework exploits inherent consistency between waveform and spectrogram representations, learning robust and transferable features from unlabeled data. Experimental results across three cross-machine fault diagnosis scenarios demonstrate that the proposed method outperforms existing state-of-the-art self-supervised methods, significantly reducing the reliance on labeled data and improving diagnostic performance. The pre-trained model and source code are available at: https://github.com/Xiaohan-Chen/CoWS.**

***Index Terms*—Self-supervised learning, Pre-training model, Knowledge transfer, Fault diagnosis**

I. Introduction

ROTATING machinery is a critical component of modern industrial systems, powering a wide range of applications in manufacturing, power generation, transportation, and other sectors. These machines are subject to continuous operation under varying loads and conditions, making them susceptible to faults that can lead to significant downtime, safety hazards, and economic losses. Effective fault diagnosis and maintenance are therefore crucial for ensuring the reliability and efficiency of rotating machinery [1]–[7].

Deep learning-based rotating machinery fault diagnosis methods have been given increased attention in recent years. Leveraging the powerful non-linear fitting capabilities of artificial neural networks, these methods have achieved remarkable success in fault detection and classification, as demonstrated by various studies [8]–[13]. However, a significant limitation of these approaches is their reliance on deep learning-based methods that require large volumes of labeled data for training. Acquiring such labeled datasets is often expensive and time-consuming. Usually, the number of fault samples is typically much smaller than normal samples, as machinery usually operates in a normal state, and once a fault is discovered it must be stopped immediately to avoid personal and property damage. Consequently, collecting sufficient labeled data for training neural networks becomes a considerable challenge [11], [14]–[17].

In practical applications, deep learning-based methods face substantial challenges in cross-machine fault diagnosis problems due to the inherent diversity in machinery design, operating conditions, and failure modes. Machines often differ substantially in their mechanical structures, component types, and control systems, leading to variations in the underlying data distributions. These differences make it difficult to apply diagnostic models trained on one machine to another directly. As mentioned before, normal operating data is abundant, fault data is scarce because faults occur infrequently. Furthermore, collecting labeled data for every machine type requires expert knowledge, making it costly and impractical. Moreover, conventional deep learning-based methods often face difficulties in generalizing to unseen fault types, which significantly limits their real-world applicability [18]–[21].

Pre-training knowledge transfer (PTKT) is a powerful technique that can be applied to resolve the above-mentioned limitations of deep learning methods. This technique involves adapting a pre-trained model from a data-rich source domain to a target domain with limited data. By fine-tuning the pre-trained model's learned features and representations to the target domain, PTKT enables efficient fault classification even with limited data [19]. PTKT offers several advantages compared with conventional deep learning methods. First, PTKT tackles the challenge of data scarcity by leveraging knowledge from a similar domain with abundant data. Second, pre-trained models can extract generalizable knowledge from massive data. Third, PTKT minimizes training time and resource costs by avoiding training from scratch. PTKT is a widely used technique within the computer vision (CV) natural language processing (NLP) fields. PTKT has revolutionized CV and NLP and achieved state-of-the-art performance by enabling the development of highly accurate and versatile models [22], [23]. However, PTKT remains relatively unexplored in fault diagnosis, and there are no publicly available pre-trained models for researchers to leverage.

PTKT falls into two main categories: supervised and self-supervised [24]. Supervised PTKT involves training a model on a massive, labeled dataset for a specific classification task;

This research is partially supported by: Jiangsu Provincial Scientific Research Center of Applied Mathematics (BK20233002), Jiangsu Provincial Qinglan Project (2021), XJTLU Research Enhancement Fund (REF-23-01-008) and Suzhou Science and Technology Programme (SYG202106).

X. Chen and Y. Xue are with the School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, China, 215123, and also with the Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, United Kingdom, L69 3BX.

R. Yang is with the School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, China, 215123.

Zidong Wang is with the Department of Computer Science, Brunel University London, Uxbridge, Middlesex UB8 3PH, United Kingdom.

Corresponding author: Rui Yang

while self-supervised PTKT learns general domain knowledge from extensive unlabeled datasets through pre-defined pretext tasks [25], and enables the model to discover meaningful patterns within the data without explicitly providing it with the correct annotations [23]. Compared to supervised PTKT, self-supervised PTKT leverages easily accessible unlabeled data for model pre-training, which eliminates the need for extensive manual labeling. This is particularly advantageous in the field of fault diagnosis, where labeling fault data requires domain experts with significant knowledge and experience in machinery faults, making the process costly, time-consuming, and often impractical. Self-supervised learning (SSL) alleviates these challenges by enabling the use of abundant unlabeled monitoring data, such as vibration signals, for pre-training. Furthermore, studies demonstrate that models pre-trained using SSL not only reduce dependence on labeled fault data but also exhibit superior transferability and generalization to diverse downstream tasks [24], [26], [27].

Existing SSL approaches in fault diagnosis often directly apply pretext tasks designed for CV and NLP to pre-train models, neglecting the unique characteristics of temporal fault signals [28]–[31]. Fundamental differences exist between temporal fault signals and images/text: CV pretext tasks prioritize spatial relationships and patterns within images, NLP tasks focus on understanding the semantics and relationships between words or phrases in text [22], while rotating machinery fault diagnosis emphasizes analyzing time, frequency, and time-frequency domain features to extract meaningful fault information and patterns from time-series signals [32]. Thus, the direct application of CV and NLP pretext tasks to temporal fault diagnosis can result in inefficient learning and hinder the model's ability to learn fault-discriminative features. Moreover, many existing SSL methods in fault diagnosis, such as simple Siamese framework (SSF) [33] and automatic fault feature extractor (AFFE) [34], suffer from additional challenges in cross-machine scenarios. Contrastive methods like self-supervised pre-training via contrast learning (SSPCL) [28] rely on large numbers of negative pairs, which may lead to false negatives when similar patterns from different domains are incorrectly treated as dissimilar. Other models like AFFE are limited to either temporal or spectral features and lack mechanisms to effectively align multiple modalities. These weaknesses become particularly problematic when fault data originate from heterogeneous sources with different sampling rates, machine structures, and sensor layouts. Consequently, existing SSL models often struggle to capture robust and transferable representations that generalize across different machines. Therefore, tailoring pretext tasks to the specific characteristics of temporal fault diagnosis signals, while addressing domain variation and feature alignment, is crucial for effective PTKT in real-world applications.

To address the aforementioned limitations, this paper proposes CoWS (Consistent Wave and Spectrogram), a novel SSL pretext task for rotating machinery fault diagnosis. Wave signals capture the continuous amplitude changes over time, while spectrograms represent the frequency content over the same period. Despite originating from different domains, waves and spectrograms convey the same underlying fault information. In the latent space, their representations should remain consistent, as both encode identical fault patterns. CoWS utilizes a dual-encoder and dual-projection head architecture, where both views are projected into a unified feature space via self-attention mechanisms. This design not only avoids dependence on large batches or negative pairs but also enhances cross-domain robustness by fusing complementary temporal and spectral features. In parallel, this paper investigates the importance of designing appropriate data augmentation hype-parameters to improve the model's generalization and robustness. Notably, the proposed CoWS pretext task does not require negative sample pairs [23], [26], [27], specialized model architectures [35], [36] and loss functions [26], yet it avoids generating collapsed solutions, such as producing the same representations for all wave and spectrogram pairs. The contributions of this paper are summarized as follows:

1) A novel SSL pretext task is proposed for cross-machine knowledge transfer, utilizing unlabeled fault monitoring data from rotating machinery;
2) A wave and spectrogram consistent representation framework with dual self-attention projection heads is proposed to automatically learn general and robust fault features;
3) The proposed CoWS pretext task is evaluated on various cross-machine downstream fault diagnosis tasks to verify its robustness and generalization;
4) Pre-trained models and source code are provided for both researchers and practitioners as open-access resources to promote development and innovation within the fault diagnosis community.

The rest of this paper is organized as follows. Section II reviews related works. Section III introduces the proposed CoWS framework in detail. Section IV presents the case studies and results analysis of the CoWS. Section V concludes the paper and discusses future work.

II. Related Works

A. Deep Learning in Industrial Applications

In recent years, the integration of advanced deep learning techniques into industrial systems has significantly enhanced the performance of intelligent fault diagnosis and predictive maintenance. For instance, Huang et al. [37] proposed a deep adversarial capsule network for compound fault diagnosis under domain shifts, while Liang et al. [38] combined capsule networks with adversarial data augmentation for gearbox fault detection. Xu et al. [39] addressed nonstationary conditions using a cascadic multireceptive learning network with adaptive filtering and domain-constrained label alignment. In the context of prognostics, Piao et al. [40] proposed a conformal prediction-based framework to provide reliable remaining useful life (RUL) estimation, offering not only accurate predictions but also desired coverage for real-world decision-making. Conventional deep learning-based RUL methods may fail under real-world conditions where sensor data may be incomplete or variably available. To address this limitation, Zhang et al. [41] proposed a multiscale cross-channel attention network to handle missing sensor channels. Several comprehensive survey articles have reviewed these trends in depth, highlighting the increasing integration of deep learning with industrial systems [15], [19], [42], [43]. While effective, these methods often rely on task-specific supervision or domain-specific tuning, limiting their scalability across machines and conditions. The proposed PTKT approach is specifically tailored for time-series fault signals, thereby contributing to the broader landscape of data-efficient and transfer-capable industrial solutions.

B. Knowledge Transfer in Fault Diagnosis

The key idea of most existing knowledge transfer approaches in fault diagnosis is to learn domain-invariant knowledge from a well-labeled similar domain and then adapt the knowledge to a different task [19]. For example, Shao et al. [44] reused the pre-trained convolutional neural network to extract low-level features from time-frequency images and then fine-tuned the last two convolution blocks to extract hierarchical features for target domains. Guo et al. [45] reduced the domain discrepancy between the source and target data to learn common features and then transfer the shared knowledge to the target domain. In some practical application scenarios, collecting sufficient target domain data can be challenging. Chen et al. [46] proposed a novel momentum prototypical neural network that progressively updates the support encoder to learn transferable knowledge from the source domain, allowing the pre-trained model to be effective in the target domain with only one or five-shot query samples. Although existing methods mitigate the problem of limited target domain data, they still require substantial labeled source domain data for prior knowledge acquisition.

C. Self-Supervised Learning in Fault Diagnosis

SSL has found its way into the realm of fault diagnosis, offering a solution to the challenge of limited data labeling. Li et al. [47] proposed an SSL model to learn fault knowledge by maximizing the mutual information between the original unlabeled vibration signals and its low-dimensional feature maps. Peng et al. [34] proposed a based on "Bootstrap Your Own Latent" time-domain fault data SSL framework to learn latent fault features. The augmented vibration signals were augmented to predict its off-line target with the stop-gradient strategy. Ding et al. [28] proposed a self-supervised momentum contrast learning model to learn local instance-level representations from unlabeled data by distinguishing positive samples from massive negative samples. Zhang et al. [48] applied time and frequency domain data to construct the time-frequency consistency encoder for feature extraction via SSL. The encoder was pre-trained on the electroencephalogram data and then the encoder was transferred to the bearing fault dataset. While many existing SSL methods for fault diagnosis directly apply pretext tasks designed for CV and NLP to pre-train models, overlooking the unique characteristics of temporal fault signals. In order to bridge the knowledge gap, this paper proposes CoWS, a novel SSL pretext task specifically designed for time-series signals.

III. Methodology

The framework for the proposed SSL pretext task, CoWS, is illustrated in Fig. 1. The procedure for implementing CoWS in cross-machine fault diagnosis can be delineated into the following three steps:

1) Source domain data collection and pre-processing: This step involves collecting adequate unlabeled fault data from faulty bearings. The collected data are then normalized to a range between 0 and 1. Subsequently, the long sequence signals are segmented into parts, each containing 1024 sample points.
2) Self-supervised pre-training: The next step is to pre-train the wave and spectrogram encoders by modeling consistent representations of the wave and spectrogram within a shared latent space.
3) Downstream task knowledge transfer: In the final step, fault data from downstream fault diagnosis tasks are collected and pre-processed. The pre-trained encoder is

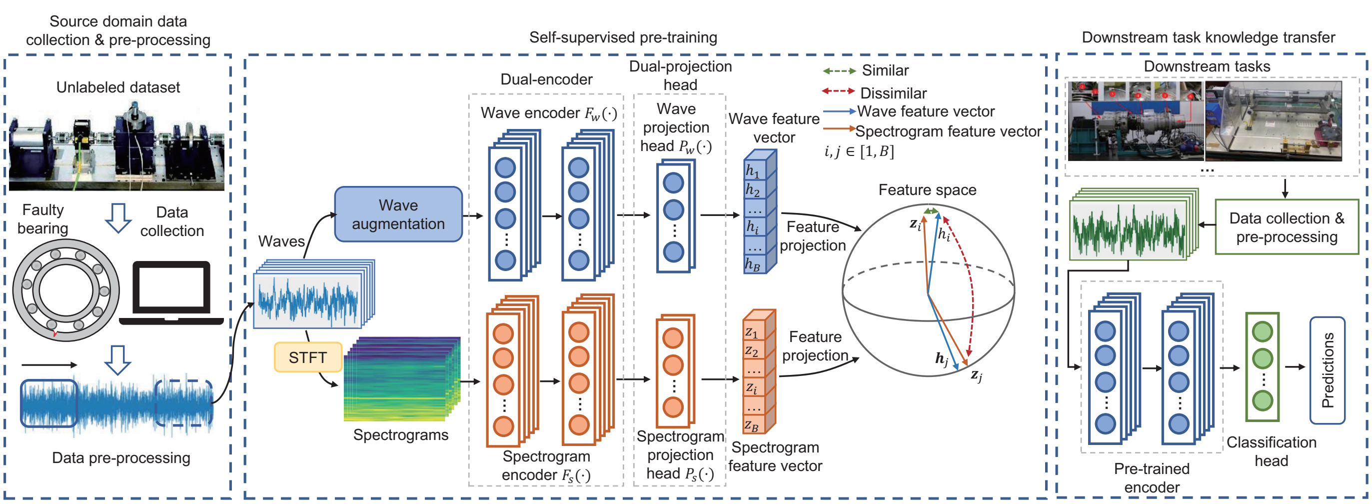


Fig. 1. An illustration of the proposed CoWS framework.

connected to a classifier head, and the model is fine-tuned using the limited available data.

In real-world applications, signal sampling rates may vary significantly across datasets due to differences in sensor configurations or machine types. To ensure consistency in input format, fixed-length input segments are extracted from each raw signal. While this standardizes the number of input points, it inherently leads to varying physical time durations per sample. To address this challenge, the proposed CoWS framework employs a dual-view architecture that processes both the waveform and its spectrogram representation. The waveform encoder captures temporal dependencies directly from the raw sampled signal, while the spectrogram encoder models frequency-domain patterns, which tend to be more invariant to sampling rate differences. In addition, the self-attention-based projection heads adaptively reweight and refine the extracted features, allowing the model to focus on semantically important patterns irrespective of their time scale. This combination enables CoWS to maintain robust performance across datasets with diverse sampling characteristics.

A. Wave Augmentation

Before inputting signals into encoders, data augmentation is applied to facilitate the model to extract generalized features. Unlike conventional SSL pre-text tasks that rely on various data augmentation methods, only Gaussian noise at a pre-determined signal-to-noise (SNR) level was selected due to its highest contribution from a series of wave and spectrogram augmentation methods, such as flip, amplitude scaling, time mask, shift, spectrogram mask, etc. [28], [33], [49]. As highlighted in [27], data augmentation plays a crucial role in influencing the performance of SSL pre-training, including factors such as the probabilities of applying augmentations and the intensity of the augmentations. Consequently, the effect of data augmentation hyper-parameter settings on generalization performance was analyzed in detail to determine the appropriate range of hyper-parameter configurations. Given a batch of pre-training wave signals $\{\mathbf{x}_i | i = 1, ..., B\}$, where each sample $\mathbf{x}_i \in \mathbb{R}^D$ consists of D features and B is the batch size. The wave signals are injected with a random SNR range to simulate real-world background noise:

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i + \mathrm{Noise}, \mathrm{Noise} \sim \mathcal{N}(0, \frac{\|\mathbf{x_i}\|^2 / N}{10^{\mathrm{SNR_{dB}}/10}}) \tag{1}$$

where N is the length of sample $\mathbf{x}_i$, and $\mathrm{SNR_{dB}}$ is the SNR that randomly chosen from a given range $[\mathrm{SNR_{min}}, \mathrm{SNR_{max}}]$.

B. Self-Supervised Pre-Training

CoWS leverages a dual-encoder architecture to extract both temporal and spectral features from raw wave signals and their corresponding spectrograms. The wave encoder, denoted as $F_w(\cdot)$, extracts temporal features $\mathbf{u}_i = F_w(\tilde{\mathbf{x}}_i)$ from the augmented wave signals. To capture spectral patterns, a standard short-time Fourier transform (STFT) with a fast Fourier transform (FFT) window length of 150 samples and a hop size of 75 samples is applied to convert wave signals into spectrograms, denoted as $\mathbf{x}'_i = \mathrm{STFT}(\mathbf{x}_i)$. The spectrogram encoder, denoted as $F_s(\cdot)$, extracts spectral features $\mathbf{v}_i = F_s(\mathbf{x}'_i)$ from the spectrograms. It is worth noting that the CoWS embraces versatility, accommodating diverse encoder architectures to suit specific needs. In this work, 1D and 2D ResNet-50 backbones are employed to extract domain-invariant features from wave and spectrogram signals, respectively. Both backbones follow a bottleneck architecture with four stages of residual blocks. Details of the architectural adaptation are summarized in Table I.

TABLE I
ARCHITECTURAL ADAPTATIONS OF THE 1D AND 2D RESNET-50 IN COWS.

Layer name	1D ResNet-50	2D ResNet-50
Conv1	7×1, stride 2	7×7, stride 2
MaxPool	3×1, stride 2	3×3, stride 2
Stage 1	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 1, 64 \\ 1 \times 1, 64 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 64 \end{bmatrix} \times 3$
Stage 2	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 1, 128 \\ 1 \times 1, 128 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 128 \end{bmatrix} \times 4$
Stage 3	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 1, 256 \\ 1 \times 1, 256 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 256 \end{bmatrix} \times 6$
Stage 4	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 1, 512 \\ 1 \times 1, 512 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 512 \end{bmatrix} \times 3$
AvgPool	AdaptiveAvgPool1D(1)	AdaptiveAvgPool2D(1×1)

In the proposed CoWS framework, a dual-projection head architecture is introduced to align and refine the temporal and spectral features extracted by the waveform and spectrogram encoders. Specifically, the dual-projection head consists of two dedicated modules: $P_w(\cdot)$ for processing waveform features and $P_s(\cdot)$ for processing spectrogram features. Conventional projection heads commonly use multi-layer perceptrons (MLP) [26], [36], these are often limited in their ability to emphasize modality, specific informative elements, especially when handling heterogeneous feature types such as time-domain waveforms and frequency-domain spectrograms. In contrast, self-attention (SA) mechanisms are particularly well-suited to model relationships across feature dimensions, allowing the model to focus on semantically important components and enhance cross-modal alignment [50].

Therefore, both $P_w(\cdot)$ and $P_s(\cdot)$ are implemented using two stacked self-attention blocks followed by a feed-forward network (FFN). These attention modules operate over the feature dimensions of each input vector. Given a feature vector $\mathbf{u}_i$ extracted from the waveform encoder, the self-attention mechanism computes query, key, and value representations as:

$$\mathbf{Q} = \mathbf{W}_Q \mathbf{u}_i, \quad \mathbf{K} = \mathbf{W}_K \mathbf{u}_i, \quad \mathbf{V} = \mathbf{W}_V \mathbf{u}_i \tag{2}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d \times d}$ are learnable weight matrices, and d is the feature dimension. The self-attention output is computed as:

$$\mathcal{A}(\mathbf{u}_i) = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V} \tag{3}$$

A residual connection and layer normalization are then applied:

$$\mathbf{u}_i' = \text{LayerNorm}(\mathbf{u}_i + \mathcal{A}(\mathbf{u}_i)) \tag{4}$$

The output $\mathbf{u}_i'$ is then processed by a feed-forward network:

$$\mathbf{h}_i = \text{FFN}(\mathbf{u}_i') \tag{5}$$

The same operation is applied independently to the spectrogram features $\mathbf{v}_i$ using $P_s(\cdot)$. The resulting outputs, $\mathbf{h}_i = P_w(\mathbf{u}_i)$ and $\mathbf{z}_i = P_s(\mathbf{v}_i)$, are used to obtain the contrastive loss in self-supervised learning. This attention-based design enables better feature discrimination and robust alignment between modalities.

C. Loss Function

CoWS facilitates the extraction of comprehensive fault features from both waveforms and spectrograms through a dual-stream SSL strategy. While waveforms and spectrograms provide distinct perspectives, they encode the same fault information in a shared latent space. By modeling consistency between the two, the model learns unified representations across both domains, capturing the shared underlying structure. Therefore, this approach maximizes the similarity between correlated temporal and spectral representation pairs in a shared latent space, denoted as $\langle \mathbf{h}_i, \mathbf{z}_i \rangle$, while simultaneously minimizing the similarity between uncorrelated pairs, $\langle \mathbf{h}_i, \mathbf{z}_j \rangle$, where $i \neq j$ and $i, j \in [1, B]$. Given a batch of input signals with a batch size B, the data inherently contains B similar representation pairs and $B^2 - B$ dissimilar pairs. To quantify the similarity relationships between temporal and spectral representations, a $B \times B$ similarity matrix $\mathbf{S}$ is constructed, where each element $\mathbf{S}_{\langle \mathbf{h}_i, \mathbf{z}_j \rangle}$ represents the similarity between the i-th temporal representation $\mathbf{h}_i$, and the j-th spectral representation $\mathbf{z}_j$. Cosine similarity is employed to measure the correlation between temporal and spectral representations, ensuring a normalized range between 0 and 1:

$$\mathbf{S}_{\langle \mathbf{h}_i, \mathbf{z}_j \rangle} = \frac{\mathbf{h}_i}{\|\mathbf{h}_i\|_2} \cdot \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|_2} \tag{6}$$

where $\|\cdot\|_2$ represents l_2-norm. The similarity pseudo labels between temporal and spectrogram representations can be defined as follows:

$$\mathbf{C}_{i,j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{7}$$

The cosine similarity values are scaled by a learnable temperature parameter, τ, which controls the softness of the distribution. Subsequently, a Softmax function is utilized along the horizontal dimension of the similarity matrix, normalizing the values into a probability distribution:

$$p_{\langle \mathbf{h}_i, \mathbf{z}_j \rangle} = \frac{\exp(\mathbf{S}_{\langle \mathbf{h}_i, \mathbf{z}_j \rangle} \cdot \exp(\tau))}{\sum_{k=1}^{B} \exp(\mathbf{S}_{\langle \mathbf{h}_i, \mathbf{z}_k \rangle} \cdot \exp(\tau))} \tag{8}$$

where $p_{\langle \mathbf{h}_i, \mathbf{z}_j \rangle}$ represents the scaled similarity score between the i-th temporal representation and the j-th spectral representation. Owing to CoWS's inherent design that eliminates the need for constructing numerous negative instances, the framework does not require a special design of the loss function [26], [27]. Instead, CoWS applies the widely used cross-entropy loss function to guide its SSL process. The loss function $\mathcal{L}_{\langle \mathbf{h}, \mathbf{z} \rangle}$ is defined as:

$$\mathcal{L}_{\langle \mathbf{h}, \mathbf{z} \rangle} = -\textstyle\sum_{i=1}^{B} \sum_{j=1}^{B} \mathbf{C}_{i,j} \cdot \log p_{\langle \mathbf{h}_i, \mathbf{z}_j \rangle} \tag{9}$$

Following [27], CoWS employs a symmetric loss function $\mathcal{L}$ defined as:

$$\mathcal{L} = \frac{1}{2}(\mathcal{L}_{\langle \mathbf{h}, \mathbf{z} \rangle} + \mathcal{L}_{\langle \mathbf{z}, \mathbf{h} \rangle}) \tag{10}$$

where $\mathcal{L}_{\langle \mathbf{z}, \mathbf{h} \rangle}$ mirrors the form of $\mathcal{L}_{\langle \mathbf{h}, \mathbf{z} \rangle}$ but calculates the loss from the perspective of the spectral representations:

$$\mathcal{L}_{\langle \mathbf{z}, \mathbf{h} \rangle} = -\textstyle\sum_{j=1}^{B} \sum_{i=1}^{B} \mathbf{C}_{i,j} \cdot \log {p_{\langle \mathbf{h}_i, \mathbf{z}_j \rangle}}^T \tag{11}$$

where ${p_{\langle \mathbf{h}_i, \mathbf{z}_j \rangle}}^T$ represents the transpose of the scaled similarity score matrix $p_{\langle \mathbf{h}_i, \mathbf{z}_j \rangle}$. The pseudocode for the symmetric loss function $\mathcal{L}$ is presented in Algorithm 1.

Algorithm 1 Pseudocode for loss function.

Input: Wave representations $\mathbf{h}_i, i \in [1, B]$, spectrogram representations $\mathbf{z}_j, j \in [1, B]$
Output: Loss $\mathcal{L}$

Initialize wave and spectrogram indexes: $i \leftarrow 1, j \leftarrow 1$
while $i \leq B$ **do**
  while $j \leq B$ **do**
    Obtain cosine similarities using (6):
    $$\mathbf{S}_{\langle \mathbf{h}_i, \mathbf{z}_j \rangle} \leftarrow \frac{\mathbf{h}_i}{\|\mathbf{h}_i\|_2} \cdot \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|_2}$$
    Obtain scaled similarity scores using (8):
    $$p_{\langle \mathbf{h}_i, \mathbf{z}_j \rangle} \leftarrow \frac{\exp(S_{\langle \mathbf{h}_i, \mathbf{z}_j \rangle} \cdot \exp(\tau))}{\sum_{k=1}^{B} \exp(S_{\langle \mathbf{h}_i, \mathbf{z}_k \rangle} \cdot \exp(\tau))}$$
    Initialize labels using (7):
    $$\mathbf{C}_{i,j} \leftarrow \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$
    $j \leftarrow j + 1$
  end while
  $i \leftarrow i + 1$
end while
Compute sub-loses using (9) and (11):
$$\mathcal{L}_{\langle \mathbf{h}, \mathbf{z} \rangle} \leftarrow -\textstyle\sum_{i=1}^{B} \sum_{j=1}^{B} \mathbf{C}_{i,j} \cdot \log p_{\langle \mathbf{h}_i, \mathbf{z}_j \rangle}$$
$$\mathcal{L}_{\langle \mathbf{z}, \mathbf{h} \rangle} \leftarrow -\textstyle\sum_{j=1}^{B} \sum_{i=1}^{B} \mathbf{C}_{i,j} \cdot \log {p_{\langle \mathbf{h}_i, \mathbf{z}_j \rangle}}^T$$
Compute symmetric loss using (10):
$$\mathcal{L} \leftarrow \tfrac{1}{2}(\mathcal{L}_{\langle \mathbf{h}, \mathbf{z} \rangle} + \mathcal{L}_{\langle \mathbf{z}, \mathbf{h} \rangle})$$

IV. Experiments

A. Datasets

To evaluate the performance of the proposed CoWS on SSL feature extraction and downstream knowledge transfer, four bearing fault datasets are used in this manuscript:

1) Paderborn University (PU) electromechanical drive systems benchmark [51]: The PU dataset contains run-to-failure bearing fault data with various fault categories such as inner race fault (IRF), outer race fault (ORF), and combined fault, collected at a sampling rate of 64 kHz.

TABLE II
DATASET DESCRIPTIONS.

Dataset	Year	No. of samples			Bearing type	Sampling frequency	Fault generation	Fault locations
		Training	Validation	Test				
PU	2016	4459	637	1274	6203 deep groove ball bearing	64 kHz	Run-to-failure	Inner race, outer race, combined
XJTU-SY	2018	20	400	400	LDK UER204 bearing	25.6 kHz	Run-to-failure	Inner race, outer race, cage, combined
HIT	2023	20	400	400	Inter-shaft bearing	2.5 kHz	Seeded	Inner race, outer race
HUST	2024	50	500	500	ER-16K ball bearing	25.6 kHz	Seeded	Inner race, outer race, ball, combined

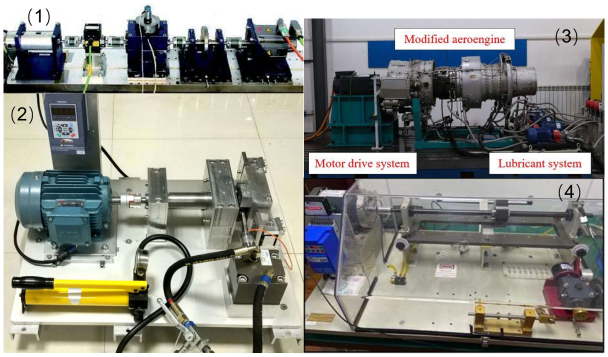


Fig. 2. The testbenches used in the experiments: (1) PU dataset, (2) XJTU-SY dataset, (3) HIT dataset, (4) HUST dataset.

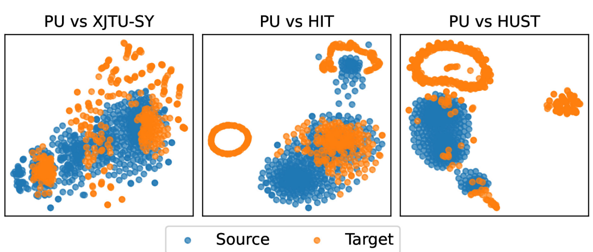


Fig. 3. Visualization of source (PU) and target datasets (XJTU-SY, HIT, HUST).

2) Xi'an Jiaotong University and the Changxing Sumyoung Technology Co., Ltd. (XJTU-SY) rolling element bearing accelerated life test [52]: The XJTU-SY dataset is a run-to-failure bearing fault dataset contains IRF, ORF, combined fault, and cage fault, collected at a sampling rate of 25.6 kHz. The test rig was run with three different working conditions.
3) Harbin Institute of Technology (HIT) aero-engine system benchmark [53]: the HIT dataset is an artificial fault dataset with faults seeded on the inner and outer race by wire cutting. The dataset consists of four working conditions, with data collected at a sampling rate of 2.5 kHz.
4) Huazhong University of Science and Technology (HUST) bearing fault benchmark [54]: The HUST dataset is based on Spectra-Quest machinery fault simulation platform. Inner race, outer race, ball, and combined faults were seeded on bearings. The dataset also consists of four working conditions, with data collected at a sampling rate of 25.6 kHz.

The experimental evaluation utilized four test benches, as shown in Fig. 2. The XJTU-SY dataset comprises three working conditions, denoted as WC1 to WC3. Conversely, the HIT and HUST datasets encompass four operating conditions, coded WC1 to WC4. Table II provides a comprehensive overview of the four datasets, including details on data splitting, bearing types, and other relevant information. Given the extensive collection of fault samples in the PU dataset, it was employed for pre-training the encoder using SSL approach. Subsequently, the pre-trained encoder and its associated parameters were transferred to the XJTU-SY, HIT, and HUST datasets for downstream fault diagnosis.

To illustrate the distributional differences across datasets, the source PU dataset and three representative targets (XJTU-SY, HIT, and HUST) are visualized using t-distributed stochastic neighbor embedding (t-SNE). As shown in Fig. 3, PU and XJTU-SY exhibit relatively greater overlap in the embedding space, suggesting smaller domain gaps. By contrast, PU and HUST show little overlap, reflecting a larger discrepancy due to structural differences. These observations highlight the core challenges of cross-machine fault diagnosis: differences in machine types, fault types, and sampling rates can lead to substantial domain gaps, which motivates the design of robust transfer methods.

B. Implementation

To evaluate the performance of the proposed CoWS SSL framework, one random initialization and six state-of-the-art SSL methods are compared: random initialization (random init), BYOL [35], SimSiam [36], SSPCL [28], SSF [33], AFFE [34], TFC [48], FSSSL [55], TFAI [56], and SMCLN [57]. To ensure equitable comparisons, all techniques utilize the same backbone ResNet-50 to extract fault representations from the unlabeled PU bearing fault dataset. All techniques utilize stochastic gradient descent with an empirically set momentum hyper-parameter of 0.9. The training process is conducted over 400 epochs, utilizing a mini-batch size of 256. The initial learning rate is established at 0.1, and a cosine annealing schedule is employed. Except for the above hyper-parameters, the other hyper-parameters are consistent with the original papers. Each method is executed 10 times, and the average accuracy is reported.

C. Case Study 1: Linear Evaluation Protocol

Consistent with previous works [35], [36], the linear evaluation protocol was employed to assess the self-supervised learned representations. Linear evaluation protocol freezes the pre-trained encoder and adds a trainable linear classifier, enabling fault prediction with limited labeled data. By freezing the encoder, the extracted features solely rely on the knowledge learned from unlabeled data during SSL. Therefore, linear evaluation protocol efficiently evaluates the effectiveness of the feature extraction ability via the predefined pretext tasks. In this study, the model is pre-trained using the unlabeled PU dataset, followed by fine-tuning the classifier with various proportions of labeled data. The classifier is fine-tuned for 40 epochs with a learning rate of 0.1.

The linear evaluation performance of the proposed method and seven comparative methods is illustrated in Fig. 4. Analyzing the experimental outcomes reveals that all SSL methods successfully learned meaningful fault features from unlabeled data through the specified pretext task. The pre-trained models showcase satisfactory performance even in scenarios with limited labeled data resources, highlighting the efficacy of SSL. As the proportion of linear evaluation samples increases, there is a general trend of increasing accuracy for most models. Notably, the proposed CoWS model achieves over 90% accuracy using only 1% of the linear evaluation samples. Whilst the SMCLN achieves second-best results, these are still less than the proposed CoWS by about 18%. TFC and SSPCL achieve an accuracy of approximately 60%, while random initialization achieves only 20%. When the proportion of linear evaluation samples increases to 15%, the accuracy of TFC and SimSiam becomes comparable to that of CoWS. The accuracy of SSPCL, AFFE, and SMCLN all exceeds 80%, and the performance of random initialization also surpasses 50%. Importantly, CoWS with only 25% labeled samples already achieves 99.41% accuracy, nearly matching the fully supervised counterpart (99.91%). This demonstrates that CoWS significantly reduces dependency on labeled data, achieving competitive performance with self-supervised pre-training on unlabeled data. Such efficiency highlights the practical effectiveness of CoWS in real-world industrial applications where labeled data are often scarce.

D. Case Study 2: Cross-Machine Knowledge Transfer

Cross-machine fault diagnosis presents a substantial challenge in practical applications due to the differences in bearing types, machine structures, sampling rates, etc. [19]. To investigate the generalizability of the CoWS to unseen machines, the model was pre-trained on the unlabeled PU dataset and the learned knowledge was transferred to the XJTU-SY, HIT, and HUST fault diagnosis tasks across various working conditions. The pre-trained model was fine-tuned 40 epochs with separate learning rate 0.001 (encoder) and 0.01 (classifier).

As illustrated in Table III, the proposed CoWS method achieves high average accuracy with low standard deviation across various machines and working conditions, indicating strong robustness and adaptability to diverse downstream fault diagnosis tasks compared with the other state-of-the-art SSL methods. Notably, on the XJTU-SY and HUST datasets, CoWS achieves an average accuracy improvement of about 5% compared to the second- and third-best methods (e.g., AFFE, SSPCL, and TFC). While AFFE performs competitively on the XJTU-SY and HIT datasets, it lags behind methods such as TFC and FFSSL on HUST. Only the proposed CoWS achieves consistently high performance across all eleven cross-machine transfer scenarios, clearly demonstrating its generalization ability and robustness to distribution shifts. It is noteworthy that, despite SimSiam demonstrating competitive outcomes in linear evaluation, it exhibits suboptimal performance in downstream transfer. This observation underscores the challenge of adapting representations learned without tailored structural design. Overall, the experimental results clearly demonstrate the positive impact of knowledge transfer from a related source, even without specific fault labels, in improving both fault diagnosis performance and overall robustness.

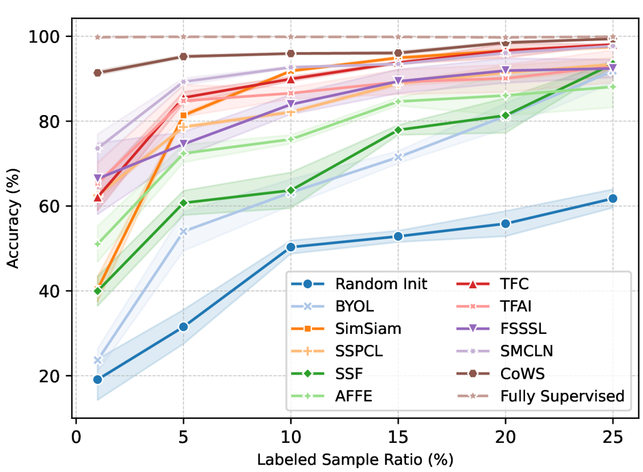


Fig. 4. Linear evaluation performance with different number of labeled data.

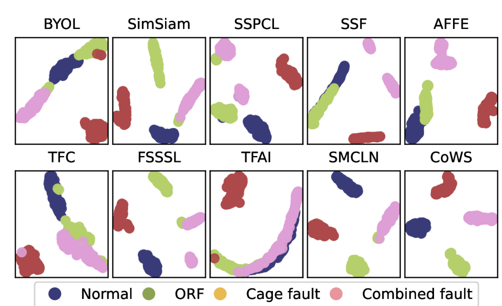


Fig. 5. Data distribution visualization on the XJTU-SY dataset.

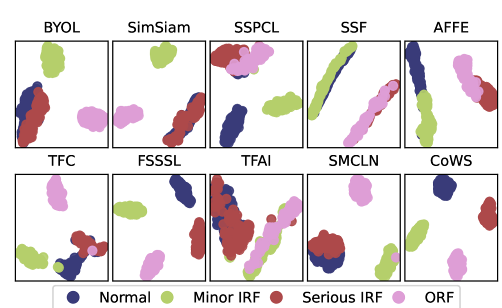


Fig. 6. Data distribution visualization on the HIT dataset.

To visualize the representations extracted by each SSL algorithm, t-SNE was employed for feature visualization. Fig.

TABLE III

CROSS-MACHINE FAULT DIAGNOSIS ACCURACY (%) AND STANDARD DEVIATION. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method	XJTU-SY			HIT				HUST			
	WC1	WC2	WC3	WC1	WC2	WC3	WC4	WC1	WC2	WC3	WC4
Random	42.20±13.08	43.91±6.65	42.61±7.33	36.45±4.04	38.03±5.45	55.16±11.15	35.77±1.79	37.50±8.95	34.22±6.85	24.83±5.76	24.71±5.91
BYOL	70.10±2.04	52.00±5.41	55.43±9.66	71.58±9.43	72.04±6.77	70.79±3.21	73.25±0.94	81.53±3.05	86.09±5.57	73.54±4.56	71.55±7.95
SimSiam	62.88±12.73	56.90±11.93	95.48±3.70	81.84±6.04	82.67±13.39	81.47±6.37	86.13±5.19	86.47±4.32	76.50±6.40	59.44±2.84	76.53±5.87
SSPCL	76.83±1.43	82.13±1.84	97.35±0.43	94.92±0.67	92.12±1.48	98.10±0.75	92.64±1.35	98.05±0.50	92.67±0.38	95.04±2.16	93.93±0.40
SSF	84.11±1.62	97.38±1.06	98.57±0.55	88.60±1.74	81.50±3.22	98.23±1.20	89.88±3.97	76.89±6.05	80.48±6.58	83.50±4.01	74.75±5.55
AFFE	98.26±1.43	88.03±3.11	99.16±0.28	97.35±3.04	98.78±1.68	**99.98±0.06**	93.24±4.18	96.09±2.13	92.59±0.49	87.26±6.16	90.16±3.50
TFC	72.96±4.27	97.51±1.68	97.56±2.35	95.20±0.87	91.83±1.56	96.27±2.30	92.94±2.13	98.87±0.64	94.12±3.53	94.96±4.25	88.98±3.75
FSSSL	89.12±4.43	85.79±6.60	96.39±5.73	97.32±1.05	98.20±2.52	99.31±1.05	93.50±4.01	96.83±1.49	92.57±3.68	91.53±4.88	87.83±3.49
TFAI	82.63±3.53	87.76±2.32	84.09±1.76	86.64±3.13	84.53±2.16	86.05±3.39	79.09±3.49	96.20±0.92	79.53±7.73	80.45±7.82	74.55±4.98
SMCLN	94.32±4.30	91.80±5.09	94.88±3.01	97.82±2.25	89.68±3.01	93.65±2.40	88.85±2.75	95.83±4.31	89.93±3.69	92.03±3.22	82.86±5.27
CoWS	**99.52±1.18**	**99.83±0.23**	**99.75±0.16**	**98.80±0.54**	**99.51±0.23**	99.86±0.08	**97.53±0.97**	**99.16±0.22**	**96.64±0.35**	**98.58±0.21**	**99.61±0.13**

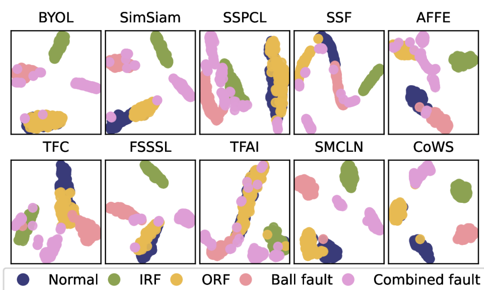


Fig. 7. Data distribution visualization on the HUST dataset.

5 illustrates the data distribution of the XJTU-SY dataset under working condition WC1. The proposed CoWS method effectively clusters each category, with clear classification boundaries. The second-best method, AFFE, also achieves good classification results. However, the phenomenon of normal samples and ORF samples overlapping is also observed in the SSF, TFC and TFAI methods. Notably, even though the source domain data from PU does not include cage fault data, all methods correctly classify the cage faults. Fig. 6 visualizes the data distribution of the HIT data under working condition WC2. The CoWS and FSSSL methods successfully distinguish all fault categories, with each class exhibiting well-defined and separable decision boundaries. In contrast, SSF shows noticeable overlap between normal and minor IRF samples, as well as between serious IRF and ORF samples. For TFC, although the boundaries between minor IRF and ORF are relatively clear, several samples are misclassified as normal or serious IRF faults. Fig. 7 visualizes the data distribution of the HUST data under working condition WC4. The CoWS method correctly clusters IRF, ball faults, and combined fault, but several normal samples and ORF samples are misclassified. SSPCL, the second-best method for the HUST dataset, correctly classifies IRF and ball faults, but there is overlapping between normal and ORF data. For random initialization, all data are mixed together, explaining why the random initialization method achieves only 24.71% accuracy.

As visualized in Fig. 3, datasets differ significantly. The proposed CoWS is specifically designed to reduce these gaps via self-attention and waveform–spectrogram alignment. The robustness of CoWS in such challenging scenarios can be explained by its architecture: the self-attention mechanism emphasizes fault-relevant transients and suppresses machine-specific noise, while the waveform–spectrogram alignment module enforces cross-view consistency, making the learned features less sensitive to the impact of variance in machine structures and mismatched sampling frequencies. Together, these design choices enable CoWS to extract frequency-invariant and structure-robust representations, which account for its superior performance in cross-machine transfer tasks.

To investigate the performance of the eight methods for each class on the HUST dataset under working condition WC4, the confusion matrix is provided in Fig. 8. TFC and FSSL achieve good performance on normal, IRF, ball fault, and combined fault prediction but still shows some confusion. SSPCL and SSF achieve 100% fault diagnosis accuracy on ORF fault, where most approaches fail to predict it accurately. AFFE and SSPCL show good boundaries for single faults, yet combined faults remain problematic. The CoWS method achieves almost perfect classification across most categories, with notably high precision and recall. Only a small number of combined faults are misclassified as ball faults. Importantly, CoWS eliminates most of the overlap observed in other methods, particularly for ORF and combined faults. Overall, the confusion matrix analysis confirms that CoWS not only improves overall accuracy but also enhances fault boundary clarity across all categories, especially for challenging cases such as ORF and combined faults.

Regarding the minor misclassifications of combined faults as ball faults, this can be attributed to the physical nature of compound faults in rolling bearings. Compound faults involve defects across multiple components, producing vibration signals that approximate the superposition of single-fault signatures. For example, the ball spin frequency and its harmonics, which are prominent in isolated ball defects, also appear in compound faults containing ball damage [58]. As a result, overlapping frequency components in the spectral domain make compound fault signatures partially resemble those of pure ball faults, particularly under noise or varying operating conditions. Such overlaps can introduce modulated harmonics that mimic single-fault patterns [59].

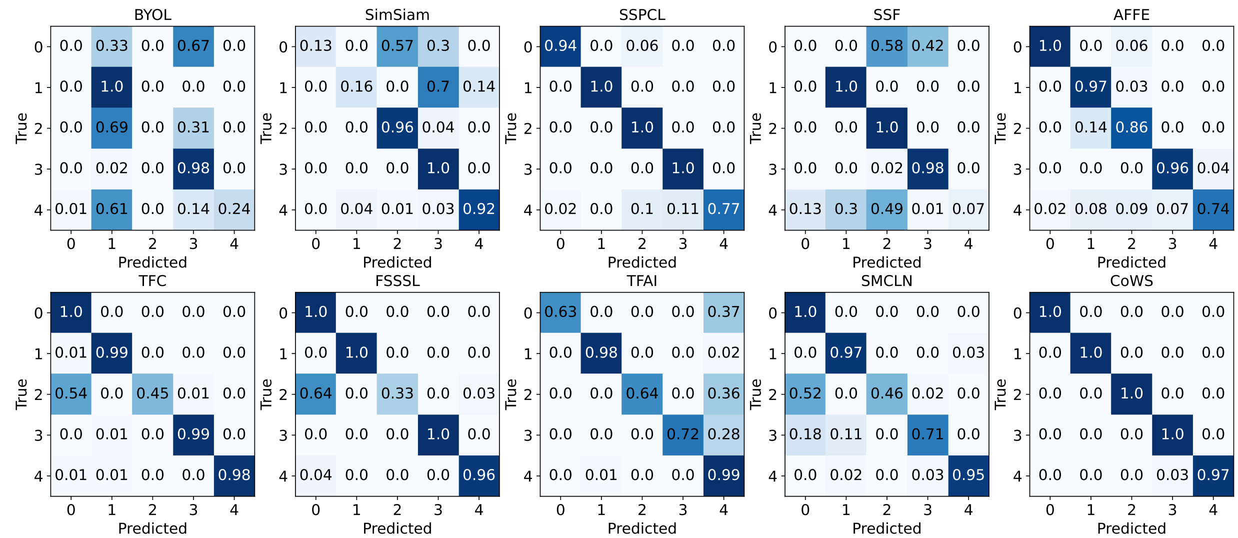


Fig. 8. The confusion matrix on the HUST dataset. label 0: normal, label 1: IRF, label 2: ORF, label 3: ball fault, and label 4: combined fault.

E. Ablation Study

1) Data Augmentation Hyper-Parameter Selection: To evaluate the impact of Gaussian noise at a predetermined SNR level augmentation during self-supervised pre-training, a grid search over multiple SNR ranges and application probabilities under the linear evaluation protocol is conducted, as illustrated in Fig. 9. The results reveal a distinct performance peak when the SNR is sampled from the range [0,10] and the augmentation is applied with 80% probability. This configuration achieves the highest accuracy of 95.25%, indicating that moderate noise levels introduce beneficial perturbations that enhance the generalization ability of the learned representations.

In contrast, overly strong augmentation, such as low Gaussian noise applied with high probability, can excessively distort the fault-related signal patterns, thereby degrading the semantic structure of the input and making it difficult for the model to extract meaningful representations from unlabeled data. On the other hand, overly weak augmentation, such as high noise applied with low probability, provides insufficient variation during pre-training, limiting the model's capacity to learn robust and transferable features. These results suggest that an appropriate level of augmentation is essential for balancing invariance and information preservation in self-supervised representation learning.

This observation aligns with prior findings in self-supervised learning. For instance, Chen et al. [27] emphasize that strong augmentations are crucial for effective representation learning, while Liu et al. [60] caution that excessively strong noise may distort data structures and reduce sensitivity to critical features. These results support the conclusion that augmentation strength must be carefully balanced to preserve discriminative information while encouraging robust feature learning from unlabeled data. The ablation study confirms that the chosen setting, Gaussian noise with SNR sampled from [0,10] and an application probability of 0.8, strikes an effective trade-off between regularization and information retention.

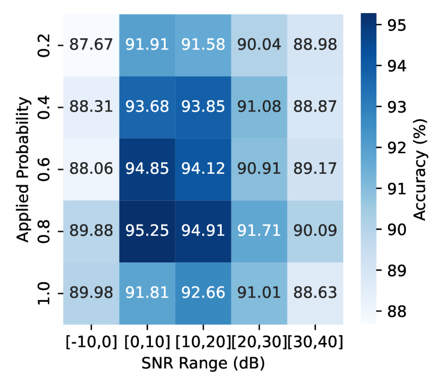


Fig. 9. Accuracy Across SNR Ranges and Augmentation Probabilities.

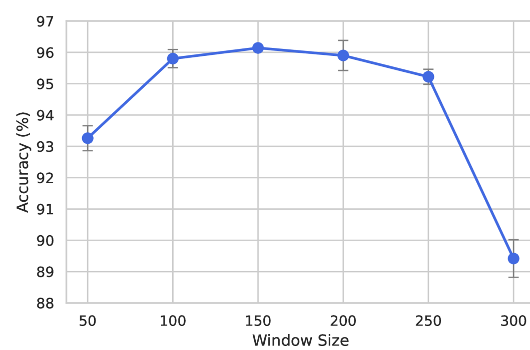


Fig. 10. Effect of Window Size on Performance.

2) STFT Window Length: The impact of different STFT window lengths on model performance is also evaluated. The window length controls the number of samples used in each local Fourier transform, thereby directly influencing the trade-off between time and frequency resolution in the resulting spectrograms. Smaller windows provide better temporal localization but may blur spectral details, while larger windows offer finer frequency resolution at the cost of time precision.

To identify an appropriate configuration, a grid search over six different window lengths is conducted: {50, 100, 150, 200, 250, 300} samples.

Fig. 10 illustrates the performance of CoWS across these settings on linear evaluation. The best accuracy is achieved when the STFT window length is set to 150 samples. When the window is too small, the frequency resolution is insufficient to capture discriminative spectral features, leading to degraded accuracy. Conversely, overly large windows may oversmooth temporal variations and introduce spectral leakage, resulting in a substantial performance drop. These results confirm that a moderate window size of 150 samples provides a balanced trade-off between time and frequency resolution, and enables effective representation learning from spectrograms in dual-view pretraining framework.

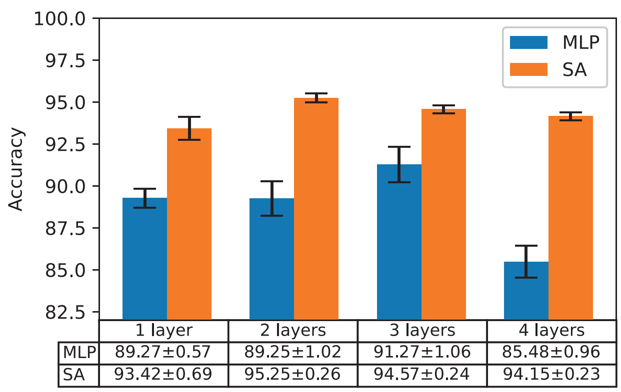


Fig. 11. Ablation study of different layers MLP and SA projection head.

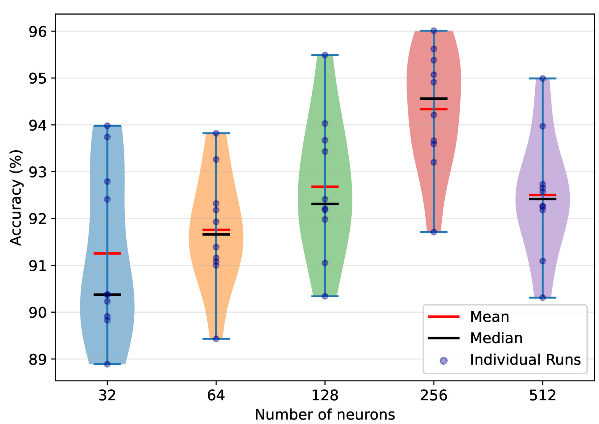


Fig. 12. Ablation study of SA projection head hidden dimensions.

3) Projection Head: The SA projection head is designed to model relationships between distant views of the same input data, thereby uncovering complex correlations that may be overlooked by an MLP projection head. This hypothesis is evaluated through an ablation study within the CoWS framework. Both wave and spectrogram projection heads are replaced with MLP and SA variants under different layer and parameter configurations, and the resulting representations are assessed using the linear evaluation protocol. As shown in Fig. 11, the SA head consistently outperforms the MLP head. For example, while a three-layer MLP projection head achieved 91.27% accuracy, a two-layer SA projection head reached 95.25%, with lower standard deviation (0.26 vs. 1.06). This validates that the SA head effectively guides the encoder in learning robust fault representations even without labels. Based on these results, a two-layer SA projection head is adopted, which provides a good balance between accuracy and efficiency.

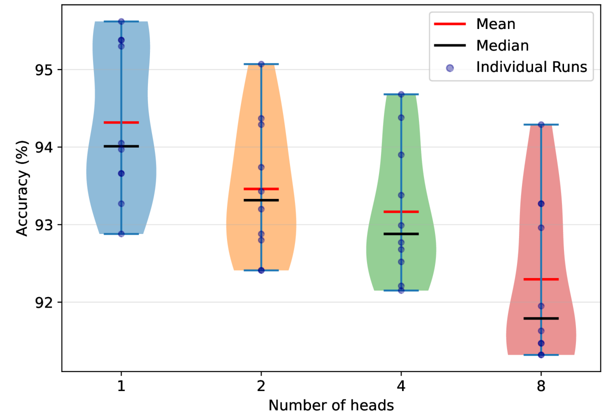


Fig. 13. Ablation study of the number of attention heads in the SA.

The sensitivity of the SA head to the hidden layer dimension is also analyzed. Fig. 12 shows the performance of SA with hidden sizes of 32, 64, 128, 256, and 512. The mean accuracy improves steadily from 91.25% at 32 units to 94.34% at 256 units, with 256 providing the best trade-off between accuracy and stability. In contrast, increasing the hidden size to 512 units decreases mean accuracy to 92.50% and does not provide additional benefits. Based on these findings, a two-layer SA projection head with 256 hidden units is adopted for both waveform and spectrogram inputs. Similarly, Fig. 13 reports results when varying the number of attention heads among 1, 2, 4, and 8. Accuracy drops from 94.32% with 1 head to 92.29% with 8 heads. Based on these results, a single attention head is chosen to maximize performance.

F. Computational Efficiency

To evaluate the practical applicability of the proposed framework, the self-supervised pre-training time for all compared methods was measured under identical hardware (NVIDIA Tesla V100) and training settings. As shown in Table IV, CoWS achieves the shortest total pre-training time of 1005 seconds, outperforming strong baselines such as AFFE (1547 s), SSPCL (1622 s), and SMCLN (1244 s). This is achieved despite CoWS employing dual encoders to process both waveform and spectrogram inputs, demonstrating that the additional modality processing does not introduce excessive computational cost. These results suggest that CoWS can be efficiently trained for real-world deployment without sacrificing learning effectiveness.

G. Limitations

Although CoWS shows strong performance across various cross-machine transfer scenarios, there remain several

TABLE IV
SELF-SUPERVISED PRE-TRAINING TIME OF DIFFERENT SSL METHODS UNDER IDENTICAL HARDWARE AND SETTINGS.

Method	BYOL	SimSiam	SSPCL	SSF	AFFE
Time (s)	1284.90	1046.14	1622.10	1217.00	1547.31
Method	TFC	FSSSL	TFAI	SMCLN	CoWS
Time (s)	2712.27	1548.92	2928.53	1244.36	**1005.27**

limitations worth discussing. First, in the current contrastive learning setup, positive pairs are formed between waveform and spectrogram representations of the same input signal, while all other combinations within a batch are treated as negatives. This may lead to false negatives when two different input segments exhibit near-identical signal patterns, e.g., from steady-state or repeated machine cycles. Although this issue is partially mitigated by modality-specific encoders, structured augmentations, and batch-level diversity, such negative pair mislabeling may still occur and could limit representation quality in edge cases. Second, while the current evaluation focuses on bearing fault datasets, the model's generalizability to other mechanical components (e.g., gearboxes, motors) and other sensing modalities (e.g., acoustics, pressure, thermal data) has not yet been explored. While CoWS demonstrates strong generalization within the bearing fault diagnosis domain, extending its robustness to broader mechanical contexts and further improving resistance to false negatives are important directions for future research.

V. CONCLUSION

This study introduces CoWS, a novel self-supervised pre-training framework specifically designed for rotating machinery fault diagnosis. CoWS employs a dual-encoder and dual-projection head architecture to model the intrinsic relationship between raw waveform signals and their spectrogram representations. To enhance alignment and feature fusion between the two modalities, the dual-projection head is replaced with a self-attention mechanism, enabling the model to refine cross-view features and focus on fault-relevant patterns. Extensive experiments on three benchmark bearing datasets demonstrate that CoWS achieves competitive performance across a variety of cross-machine transfer tasks.

To mitigate the issue of false negatives in contrastive learning, advanced strategies such as hard negative mining or semantic-aware contrastive learning can be introduced to separate different faults from visually similar but distinct samples. In addition, extending CoWS to cover a wide range of machine components and sensing modalities will be critical to assess its scalability and robustness in real-world industrial environments. Such efforts will help position CoWS as a general solution for intelligent fault diagnosis across heterogeneous machinery systems.

REFERENCES

[1] G. Cui, L.-B. Wu, and M. Wu, "Adaptive event-triggered fault-tolerant control for leader-following consensus of multi-agent systems," *International Journal of Systems Science*, pp. 1–13, 2024.

[2] Y. Liang, L. Tian, X. Zhang, X. Zhang, and L. Bai, "Multi-dimensional adaptive learning rate gradient descent optimization algorithm for network training in magneto-optical defect detection," *International Journal of Network Dynamics and Intelligence*, pp. 100 016–100 016, 2024.

[3] Y. Wang, C. Shen, J. Huang, and H. Chen, "Model-free adaptive control for unmanned surface vessels: a literature review," *Systems Science & Control Engineering*, vol. 12, no. 1, pp. 1–9, 2024.

[4] V. Kwao and I. Raptis, "Sensor fault diagnosis in nonlinear stochastic systems: a hybrid system formulation," *International Journal of Systems Science*, vol. 56, no. 8, pp. 1755–1773, 2025.

[5] M. Li, K. Zhang, Y. Ma, and B. Jiang, "Prescribed-time fault-tolerant control for the formation of quadrotors based on fully-actuated system approaches," *International Journal of Systems Science*, pp. 1–15, 2024.

[6] H. Hassani, R. Razavi-Far, M. Saif, J. Zarei, and F. Blaabjerg, "Intelligent decision support and fusion models for fault detection and location in power grids," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, no. 3, pp. 530–543, 2021.

[7] Y. Wang, Z. Wang, W. Liu, N. Zeng, S. Lauria, C. Prieto, F. Sikström, H. Yu, and X. Liu, "A novel depth-connected region-based convolutional neural network for small defect detection in additive manufacturing," *Cognitive Computation*, vol. 17, no. 1, p. 36, 2025.

[8] Z. Chen, J. Xu, T. Peng, and C. Yang, "Graph convolutional network-based method for fault diagnosis using a hybrid of measurement and prior knowledge," *IEEE Transactions on Cybernetics*, vol. 52, no. 9, pp. 9157–9169, 2021.

[9] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, 2017.

[10] X. Li, Z. Zhang, L. Gao, and L. Wen, "A new semi-supervised fault diagnosis method via deep CORAL and transfer component analysis," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, no. 3, pp. 690–699, 2021.

[11] F. Deng, Y. Ming, and B. Lyu, "CCE-Net: Causal convolution embedding network for streaming automatic speech recognition," *International Journal of Network Dynamics and Intelligence*, pp. 100 019–100 019, 2024.

[12] X. Chen, Y. Xue, M. Huang, and R. Yang, "Multi-modal self-supervised learning for cross-domain one-shot bearing fault diagnosis," *IFAC-PapersOnLine*, vol. 58, no. 4, pp. 746–751, 2024.

[13] Y. Xue, R. Yang, X. Chen, B. Song, and Z. Wang, "Separable convolutional network-based fault diagnosis for high-speed train: A gossip strategy-based optimization approach," *IEEE Transactions on Industrial Informatics*, 2024.

[14] Y. Wang, C. Wen, and X. Wu, "Fault detection and isolation of floating wind turbine pitch system based on kalman filter and multi-attention 1DCNN," *Systems Science & Control Engineering*, vol. 12, no. 1, p. 2362169, 2024.

[15] 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," *Mechanical Systems and Signal Processing*, vol. 138, p. 106587, 2020.

[16] T. Ouyang, Y. He, and H. Huang, "Monitoring wind turbines' unhealthy status: a data-driven approach," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 3, no. 2, pp. 163–172, 2018.

[17] Y. Xue, R. Yang, X. Chen, W. Liu, Z. Wang, and X. Liu, "A review on transferability estimation in deep transfer learning," *IEEE Transactions on Artificial Intelligence*, 2024.

[18] R. Yan, F. Shen, C. Sun, and X. Chen, "Knowledge transfer for rotary machine fault diagnosis," *IEEE Sensors Journal*, vol. 20, no. 15, pp. 8374–8393, 2019.

[19] X. Chen, R. Yang, Y. Xue, M. Huang, R. Ferrero, and Z. Wang, "Deep transfer learning for bearing fault diagnosis: A systematic review since 2016," *IEEE Transactions on Instrumentation and Measurement*, 2023.

[20] L. Zhao and B. Li, "Adaptive fixed-time control for multiple switched coupled neural networks," *International Journal of Network Dynamics and Intelligence*, pp. 100 018–100 018, 2024.

[21] C. Wang, Z. Wang, H. Liu, H. Dong, and G. Lu, "An optimal unsupervised domain adaptation approach with applications to pipeline fault diagnosis: Balancing invariance and variance," *IEEE Transactions on Industrial Informatics*, 2024.

[22] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[23] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3733–3742.

[24] T. Chen, J. Frankle, S. Chang, S. Liu, Y. Zhang, M. Carbin, and Z. Wang, "The lottery tickets hypothesis for supervised and self-supervised pre-training in computer vision models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 306–16 316.

[25] L. Ma, J. He, K. Lu, D. Wang, L. Yin, and Z. Li, "A contrastive learning and knowledge distillation-based framework for efficient federated intrusion detection in iot," *Systems Science & Control Engineering*, vol. 13, no. 1, p. 2518963, 2025.

[26] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[27] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1597–1607.

[28] Y. Ding, J. Zhuang, P. Ding, and M. Jia, "Self-supervised pretraining via contrast learning for intelligent incipient fault detection of bearings," *Reliability Engineering & System Safety*, vol. 218, p. 108126, 2022.

[29] C. Yang, J. Liu, K. Zhou, and X. Jiang, "Semisupervised machine fault diagnosis fusing unsupervised graph contrastive learning," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 8, pp. 8644–8653, 2023.

[30] Q. Li and Y. Wang, "Self-supervised pretraining based on noise-free motion reconstruction and semantic-aware contrastive learning for human motion prediction," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 8, no. 1, pp. 738–751, 2023.

[31] Z. Xiao, H. Xing, B. Zhao, R. Qu, S. Luo, P. Dai, K. Li, and Z. Zhu, "Deep contrastive representation learning with self-distillation," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2023.

[32] G. Yu, "A concentrated time–frequency analysis tool for bearing fault diagnosis," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 2, pp. 371–381, 2019.

[33] W. Wan, J. Chen, Z. Zhou, and Z. Shi, "Self-supervised simple siamese framework for fault diagnosis of rotating machinery with unlabeled samples," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[34] T. Peng, C. Shen, S. Sun, and D. Wang, "Fault feature extractor based on bootstrap your own latent and data augmentation algorithm for unlabeled vibration signals," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 9, pp. 9547–9555, 2022.

[35] J. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 271–21 284, 2020.

[36] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 750–15 758.

[37] R. Huang, J. Li, Y. Liao, J. Chen, Z. Wang, and W. Li, "Deep adversarial capsule network for compound fault diagnosis of machinery toward multidomain generalization task," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, 2020.

[38] P. Liang, C. Deng, X. Yuan, and L. Zhang, "A deep capsule neural network with data augmentation generative adversarial networks for single and simultaneous fault diagnosis of wind turbine gearbox," *ISA Transactions*, vol. 135, pp. 462–475, 2023.

[39] Y. Xu, S. Li, K. Feng, R. Huang, B. Sun, X. Yang, Z. Zhao, and G. Q. Huang, "Domain constrained cascadic multireceptive learning networks for machine health monitoring in complex manufacturing systems," *Journal of Manufacturing Systems*, vol. 80, pp. 563–577, 2025.

[40] S. Piao, R. Huang, and F. Tsung, "CRULP: reliable rul estimation inspired by conformal prediction," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–11, 2025.

[41] T. Zhang, L. Jiang, R. Huang, and X. Zhang, "A multi-scale cross-channel attention network for remaining useful life prediction with variable sensors," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–10, 2025.

[42] W. Li, R. Huang, J. Li, Y. Liao, Z. Chen, G. He, R. Yan, and K. Gryllias, "A perspective survey on deep transfer learning for fault diagnosis in industrial scenarios: Theories, applications and challenges," *Mechanical Systems and Signal Processing*, vol. 167, p. 108487, 2022.

[43] C. Wang, Z. Wang, Q. Liu, H. Dong, W. Liu, and X. Liu, "A comprehensive survey on domain adaptation for intelligent fault diagnosis," *Knowledge-Based Systems*, p. 114109, 2025.

[44] S. Shao, S. McAleer, R. Yan, and P. Baldi, "Highly accurate machine fault diagnosis using deep transfer learning," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2446–2455, 2018.

[45] L. Guo, Y. Lei, S. Xing, T. Yan, and N. Li, "Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 9, pp. 7316–7325, 2018.

[46] X. Chen, R. Yang, Y. Xue, C. Yang, B. Song, and M. Zhong, "A novel momentum prototypical neural network to cross-domain fault diagnosis for rotating machinery subject to cold-start," *Neurocomputing*, vol. 555, p. 126656, 2023.

[47] G. Li, J. Wu, C. Deng, M. Wei, and X. Xu, "Self-supervised learning for intelligent fault diagnosis of rotating machinery with limited labeled data," *Applied Acoustics*, vol. 191, p. 108663, 2022.

[48] X. Zhang, Z. Zhao, T. Tsiligkaridis, and M. Zitnik, "Self-supervised contrastive pre-training for time series via time-frequency consistency," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3988–4003, 2022.

[49] S. Padi, S. O. Sadjadi, R. D. Sriram, and D. Manocha, "Improved speech emotion recognition using transfer learning and spectrogram augmentation," in *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 645–652.

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[51] C. Lessmeier, J. K. Kimotho, D. Zimmer, and W. Sextro, "Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification," in *PHM Society European Conference*, vol. 3, no. 1, 2016.

[52] B. Wang, Y. Lei, N. Li, and N. Li, "A hybrid prognostics approach for estimating remaining useful life of rolling element bearings," *IEEE Transactions on Reliability*, vol. 69, no. 1, pp. 401–412, 2018.

[53] L. Hou, H. Yi, Y. Jin, M. Gui, L. Sui, J. Zhang, and Y. shu Chen, "An inter-shaft bearing fault diagnosis dataset from an aero-engine system," 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260604174

[54] C. Zhao, E. Zio, and W. Shen, "Domain generalization for cross-domain fault diagnosis: An application-oriented perspective and a benchmark study," *Reliability Engineering & System Safety*, p. 109964, 2024.

[55] H. Wang, X. Wang, Y. Yang, K. Gryllias, and Z. Liu, "A few-shot machinery fault diagnosis framework based on self-supervised signal representation learning," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–14, 2024.

[56] D. Fu, J. Liu, H. Zhong, X. Zhang, and F. Zhang, "A novel self-supervised representation learning framework based on time-frequency alignment and interaction for mechanical fault diagnosis," *Knowledge-Based Systems*, vol. 295, p. 111846, 2024.

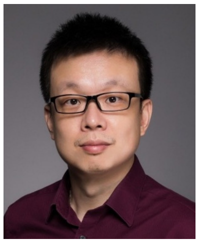
[57] Y. Xu, X. Lu, T. Gao, and R. Meng, "A self-supervised multi-view contrastive learning network for the fault diagnosis of rotating machinery under limited annotation information," *IEEE Transactions on Instrumentation and Measurement*, 2025.

[58] M. Afshar, M. Heydarzadeh, and B. Akin, "A comprehensive investigation of fault signatures and spectrum analysis of vibration signals in distributed bearing faults," *IEEE Transactions on Industry Applications*, vol. 61, no. 1, pp. 515–526, 2025.

[59] Y. Wang, H. Yang, S. Zhao, Y. Fan, and R. Dong, "Vibration characterization of rolling bearings with compound fault features under multiple interference factors," *PLoS ONE*, vol. 19, no. 2, p. e0297935, 2024.

[60] Y. Liu, C. Wang, Q. Yang, L. Li, and B. Wang, "Self-supervised learning video anomaly detection based on time interval prediction and noise classification," *Pattern Recognition*, p. 112198, 2025.

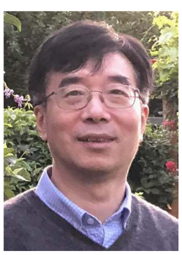
Xiaohan Chen received the Ph.D. degree in Electrical Engineering and Electronics from the University of Liverpool, Liverpool, U.K. His research interests include transfer learning, few-shot learning, and fault diagnosis, for complex systems. He is currently a Postdoctoral Researcher with the KIOS Research and Innovation Center of Excellence, University of Cyprus, Nicosia, Cyprus.



Rui Yang received the B.Eng. degree in Computer Engineering and the Ph.D. degree in Electrical and Computer Engineering from National University of Singapore in 2008 and 2013 respectively.

He is currently an Associate Professor in the School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, China, and an Honorary Lecturer in the Department of Computer Science, University of Liverpool, Liverpool, United Kingdom. His research interests include machine learning based data analysis and applications. He is the author or co-author of several technical papers and also a very active reviewer for many international journals and conferences. Dr. Yang is currently serving as an Associate Editor for Neurocomputing, Cognitive Computation and IEEE Transactions on Instrumentation and Measurement.

Yihao Xue received the B.Eng. degree in Robotics Engineering from Guangzhou University, Guangzhou, China, in 2021, and the Ph.D. degree in Electrical Engineering and Electronics from the University of Liverpool, Liverpool, U.K., in 2025. His research interests encompass deep learning, transfer learning, and multitask learning, with an emphasis on their applications to fault diagnosis of rotating machinery. He serves as an active reviewer for numerous international journals and conferences.



Zidong Wang (Fellow, IEEE) received the B.Sc. degree in mathematics from Suzhou University, Suzhou, China, in 1986, and the M.Sc. degree in applied mathematics and the Ph.D. degree in electrical engineering from the Nanjing University of Science and Technology, Nanjing, China, in 1990 and 1994, respectively.

From 1990 to 2002, he held teaching and research appointments in universities in China, Germany, and the U.K. He is currently a Professor of dynamical systems and computing with the Department of Computer Science, Brunel University London, Uxbridge, U.K. He has published a number of papers in international journals. His research interests include dynamical systems, signal processing, bioinformatics, control theory, and applications.

Prof. Wang is a member of the Academia Europaea and the European Academy of Sciences and Arts, an Academician of the International Academy for Systems and Cybernetic Sciences, a fellow of the Royal Statistical Society, and a member of the Program Committee of many international conferences. He holds the Alexander von Humboldt Research Fellowship of Germany, the JSPS Research Fellowship of Japan, and the William Mong Visiting Research Fellowship of Hong Kong. He serves (or has served) as the Editor-in-Chief for International Journal of Systems Science, Neurocomputing, and Systems Science and Control Engineering; and an Associate Editor for 12 international journals, including IEEE Transactions on Automatic Control, IEEE Transactions on Control Systems Technology, IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Signal Processing, and IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews.