Dconformer: A Denoising Convolutional Transformer with Joint Learning Strategy for Intelligent Diagnosis of Bearing Faults

Sheng Li [1] J.C. Ji[2] Yadong Xu[3*] Ke Feng[4*] Ke Zhang [1]
Jingchun Feng [1] Michael Beer[5,6,7] Qing Ni [2] Yuling Wang[3]

[1] Business School, Hohai University, Nanjing, Jiangsu 211100, China

[2]School of Mechanical and Mechatronic Engineering, University of Technology Sydney, NSW 2007, Australia

[3]School of Mechanical Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu 210094, China

[4]School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China

[5]Institute for Risk and Reliability, Leibniz University Hannover, Callinstr. 34, Hannover, Germany

[6]Institute for Risk and Uncertainty and School of Engineering, University of Liverpool, Peach Street, Liverpool L69 7ZF, UK

[7] Department of Civil Engineering, Tsinghua University, Beijing, 100190, China

Abstract

Rolling bearings are the core components of rotating machinery, and their normal operation is crucial to entire industrial applications. Most existing condition monitoring methods have been devoted to extracting discriminative features from vibration signals that reflect bearing health status. However, the complex working conditions of rolling bearings often make the fault-related information easily buried in noise and other interference. Therefore, it is challenging for existing approaches to extract sufficient critical features in these scenarios. To address this issue, this paper proposes a novel CNN-Transformer network, referred to as Dconformer, capable of extracting both local and global discriminative features from noisy vibration signals. The main contributions of this research include: 1) Developing a novel joint-learning strategy that simultaneously enhances the performance of signal denoising and fault diagnosis, leading to robust and accurate diagnostic results; 2) Constructing a novel CNN-transformer network with a multi-branch cross-cascaded architecture, which inherits the strengths of CNNs and transformers and demonstrates

Corresponding authors: Yadong Xu & Ke Feng
E-mail: ydxu@seu.edu.cn & ke.feng@outlook.com.au

superior anti-interference capability. Extensive experimental results reveal that the proposed Dconformer outperforms five state-of-the-art approaches, particularly in strong noisy scenarios.

Index Terms

Rolling bearing, fault diagnosis, vibration signal, Dconformer; complex working conditions, noisy scenarios.

I. Introduction

Rolling bearings play a vital role in supporting components in rotating machinery such as precision machine tools, high-speed trains, aircraft engines, and other major equipment. The faults and failure of rolling bearings can affect the normal operation of the equipment and even cause catastrophic accidents [1]. To ensure the safe operation of machinery and equipment, it is necessary to implement real-time condition monitoring and fault diagnosis for rolling bearings [2]–[4].

Conventional diagnostic approaches usually employ diverse signal analysis methods to extract fault frequencies from the vibration signals provided, which heavily rely on expert knowledge and experience. [5]. Even though these techniques have been widely employed, they still have the following limitations:

(1) The non-smooth and non-linear characteristics of mechanical signals make the traditional fault diagnosis procedure time-consuming and laborious, as it requires a large amount of analysis and verification [6] [7].

(2) Critical information reflecting the equipment status can be easily overwhelmed in the measured vibration signals with ambient noises and interference, making it challenging to extract valuable features using traditional methods [8].

In recent years, deep learning has become a promising tool for bearing fault diagnosis, overcoming the deficiencies of traditional data-driven algorithms [9]. In particular, due to their robust feature mapping capability, convolutional neural networks (CNNs) have been widely utilized in automatic feature learning of mechanical signals [10]. Wang et al. [11] proposed a novel CNN combined with symmetrized dot pattern representation, allowing the network to capture the fault-related features effectively. Huang et al. [12] developed an improved complete ensemble empirical mode decomposition method and utilized a 1-D CNN model to learn the high-frequency components of measured signals. Chen et al. [13] constructed a multi-scale CNN with the feature alignment module to improve the feature fusion performance of CNN. Ma et al. [14] presented a new CNN architecture with the probability confidence module embedded, to identify the unknown faults. Gao et al. [15] developed a new algorithm by combining the multi-strategy cuckoo search algorithm with a 1-D convolutional neural network for intelligent fault detection. Gao et al. [16] proposed a hierarchical training convolution network for fault diagnosis under imbalanced data conditions.

Although CNN-based diagnostic approaches yield promising results, several studies have indicated that while convolution-based networks are efficient in learning local features, they may fail to extract global long-distance dependencies [17]. These deficiencies hinder CNN-based approaches from constructing a complete and comprehensive representation of bearing health status, subsequently declining the performance of these architectures in fault diagnosis tasks. Recent approaches have attempted to integrate various attention mechanisms into CNNs to address the aforementioned deficiencies [18] [19] [20]. For instance, Zhao et al. [21] introduced a channel-wise attention mechanism to enhance feature representation in CNNs, thereby improving the diagnostic performance of these networks for the dual active bridge converter. Similarly, Wang et al. [22] developed a discrete wavelet attention mechanism, thereby enhancing the feature extraction capability of the convolution layer by mapping input time domain signals to wavelet space. Different from the above plug-and-play attention mechanisms, Liao et al. [23] developed a neuron-induced attention structure, termed qttention, which can efficiently facilitate the interpretable bearing fault diagnosis. These improvements are promising but somewhat limited, as the primary function of these attention mechanisms remains the enhancement or recalibration of local features in CNNs. They still lack the capability of modeling global long-distance dependencies, which is crucial for constructing a comprehensive fault representation.

In recent research on artificial intelligence and machine learning, the integration of the self-attention mechanism into CNNs, i.e., CNN-transformer structures, demonstrates significant potential in capturing both complex global feature dependencies and local information [19]. Based on this advancement, some approaches have developed various multi-branch architectures that combine CNNs with the transformer. For instance, Bai et al. [24] utilized a novel multi-branch vision transformer structure to facilitate the hyperspectral image classification. While, Liu et al. [25] developed a dual-branch network integrating a lightweight CNN branch and a compact vision transformer branch for high-resolution synthetic aperture radar image recognition tasks. Additionally, similar dual-branch CNN-transformer networks have also demonstrated superior feature extraction capabilities in other domains, including medical image segmentation [26]–[28], human face recognition [29], image fusion [30], and other areas [31]–[33] [34]. However, the aforementioned CNN-transformer-based architectures are typically developed for various image processing tasks, and their adaptation for vibration signals still requires in-depth research. Since the vibration signal is an information carrier completely different from the image, it is necessary to explore a specialized CNN-transformer architecture based on the unique characteristics of the signal in the fault diagnosis task.

Furthermore, mechanical systems often operate in harsh environments, bringing non-negligible interference to mechanical signals [35] [36]. Multiple studies demonstrated that the interference information can greatly affect the performance of deep learning-based models [37] . To enhance the noise resistance of the diagnostic model, many robust methods have been proposed to learn features from noisy signals. Xu et al. [6] developed a novel attention-based denoising network by stacking multiple multi-scale denoising modules. Yao et al. [38] proposed a novel acoustic-based diagnosis method to remove the non-stationary noise components embedded in vibration signals. Xiong et al. [39] developed an adaptive residual network to learn fault features using the long short-term memory module, and utilized a nonlinear transform layer to reduce noise. Han et al. [40] utilized multiple non-local fully convolutional blocks to formulate a novel robust denoising method. Zhi et al. [41] proposed a novel denoising algorithm with a joint wavelet regional correlation threshold to learn enough important features under noisy conditions. Zhao et al. [42] proposed a novel hybrid pre-training strategy to eliminate the interference of noise on CNN's diagnostic performance. Li et al. [43] constructed a graph wavelet denoising network to extract features from multi-respective, and obtained superior diagnostic performance. Zhao et al. [44] developed a deep rational attention-based network with a soft threshold unit to obtain outstanding diagnostic results. However, in these aforementioned studies, the denoising process and fault diagnosis process are implemented separately, which may cause useful information to be filtered out, leading to the under-exploration of the measured signals. [45]. Consequently, efficiently integrating the denoising process with the fault diagnosis task remains worthy of further exploration.

Based on the above discussions, there are two promising improvement directions in fault diagnosis:

(1) Specialized CNN-transformer architecture needs to be developed that can simultaneously extract both local features and global dependency information from vibration signals.

(2) Effective integration of the signal denoising process with the fault diagnosis task is a promising method to mutually enhance the performance of both.

To address these two challenges, we propose a novel CNN-transformer network, which will be referred to as Dconformer in this study. First, an Attention-Guided Multi-Scale Branch (AMB) and a Signal Transformer Branch (STB) are utilized to capture and fuse both local and global information. Subsequently, the proposed network architecture can fuse the extracted local and global information, thereby facilitating the construction of rich and comprehensive fault representations. Second, an encode-decode-based Signal Denoising Branch (SDB) is introduced to incorporate anti-interference capability into the overall architecture by filtering out noisy information intelligently. Finally, a Dynamic Weight Average (DWA) strategy is adopted to facilitate signal denoising and fault diagnosis simultaneously. In summary, the main contributions of this paper can be summarized as follows:

(1) We propose a novel CNN-transformer network featuring a multi-branch cross-cascaded architecture, termed Dconformer, which inherits the advantages of both CNNs and transformer structures and displays superior anti-interference capability.

(2) We have developed a joint-learning approach for simultaneous intelligent signal denoising and fault identification. Within joint-learning method, the adoption of a Dynamic Weight Average (DWA) strategy allows the dual tasks to mutually enhance each other, enabling the architecture to achieve favorable diagnostic results.

(3) We conduct two case studies using the ABLT-1A bearing dataset (constant-speed dataset) and Spectra Quest bearing dataset (variable-speed dataset) to validate the efficacy of the developed Dconformer. Extensive experimental results reveal that the Dconformer outperforms five state-of-the-art approaches, especially in noisy scenarios.

The structure of the paper is as follows. Section II provides a detailed explanation of the proposed method. In Section III, the performance of Dconformer is evaluated using the constant- and variable-speed bearing datasets. In Section IV, the superiority of the Dconformer is further verified and discussed. Finally, the conclusion and future research directions are discussed in Section V.

II. The Proposed Method

A. *Overview*

In this paper, we propose a novel end-to-end CNN-transformer network, termed Dconformer, for fault diagnosis of rotating machinery in various non-stationary scenarios. As illustrated in Fig. 1, this network captures and fuses both local and global information from vibration signals and adaptively filters out irrelevant noise components. Consequently, it achieves robust and satisfactory diagnostic results in harsh environments.

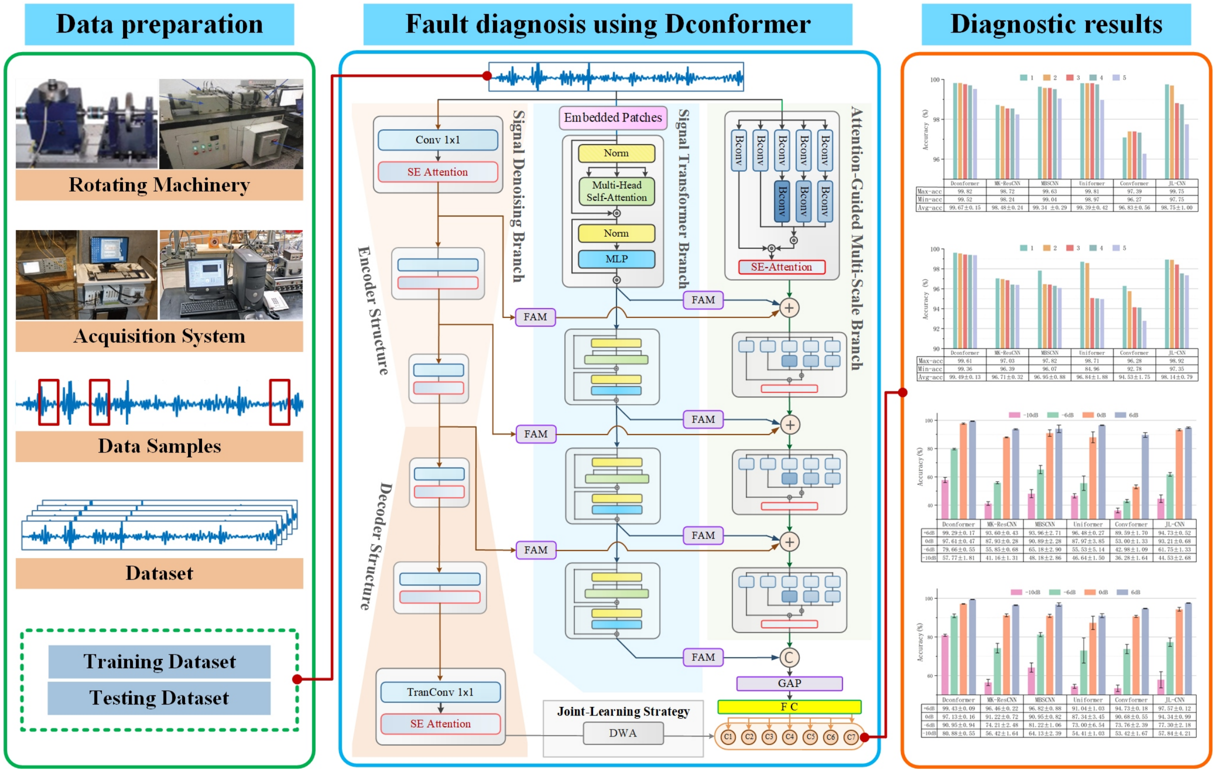


Fig. 1. Overall framework of the proposed Dconformer.

Inspired by previous work [6], our study introduces more refined improvements, primarily including:

(1) Contrasted with the plug-and-play denoising mechanism utilized in AM-DRCN [6], our work incorporates a specialized denoising branch structure into the overall architecture. This fashion enables a more efficient integration of the denoising process with the fault diagnosis task.

(2) Compared to the single-backbone structure of AM-DRCN, we have further developed a more refined multi-branch cross-cascaded architecture. This design allows for multiple branches to mutually communicate and integrate the strengths of each branch.

To be specific, the proposed Dconformer consists of three parallel branches, namely, an Attention-guided Multi-scale Branch (AMB), a Signal Denoising Branch (SDB), and a Signal Transformer Branch (STB). The AMB (light green), serving as the backbone of Dconformer, is applied to extract multi-scale local features from vibration signals. The STB (light blue) uses a self-attention mechanism to integrate global feature dependencies. The SDB (light red) applies to encode and decode operations to denoise the mechanical signals intelligently. In addition, a Feature Alignment Module (FAM), functioning as a bridge module, is developed to facilitate cross-cascaded connections among the AMB, STB, and SDB. This connection fashion effectively fuses global information from the STB and denoising features from the SDB into the backbone branch (AMB). Consequently, the integration of multi-scale local features with global dependencies co-constructs rich and comprehensive fault representations. Meanwhile, denoising features contribute to the removal of noise components within these representations. During the training phase, two loss functions are employed, supervising the denoising and diagnosis tasks separately. Furthermore, a Dynamic Weight Average (DWA) strategy is adopted to assign dynamic weights to each loss function, thereby controlling the balance between both tasks.

B. Attention-guided multi-scale branch

The Attention-guided Multi-Scale Branch (AMB) contains several Multi-Scale Attention (MSA) blocks, as shown in Fig.2. Inspired by the structure of the retinal fovea in the human visual system, the MSA block consists of two sub-modules: the multi-scale module and the squeeze-and-excitation (SE) attention module. The multi-scale module employs Bconv (a combination of a convolutional layer), a BN layer, and a Leaky ReLU activation function (Conv + BN + Leaky ReLU). This module has five branches, denoted as $b_i, (i = 1, 2, \ldots, 5)$. The first convolution layer of all branches is set to 1×1 to control the channel size to 16. In the second layer, Bconv is set with a kernel size of $1 \times (2i - 1)$. Then, the concatenation operator is applied to the four branches, and the identity shortcut branch is added to the concatenated feature vector using the elementwise summation operator. Finally, the whole module is fed into an SE attention module.

The SE module is leveraged to enhance and calibrate the extracted features from the multi-scale module. The structure of SE attention is given in Fig. 1. We use $x \in R^{C \times W}$ to denote the input for the SE module. To begin, the spatial information within each channel of the input is compressed into a scalar value by:

$$q_i = F_{sq}(x_i) = \frac{1}{W}[\sum_{k=1}^{W} x_i(c,k)]_{c=1}^{C}, \tag{1}$$

where $x_i(c,k)$ refers to the component of $x \in R^{C \times W}$, and $q_i \in R^{C \times 1}$ represents the squeezed feature map. Then, we construct a fully connected layer that comprises two linear layers and their subsequent activation functions to process the features aggregated by the previous squeeze operation, which can be expressed as:

$$z_i = F_{ex}(q_i) = \sigma_{Sig}(W_2 \cdot \sigma_{ReLU}(W_1 q_i)) \tag{2}$$

where $W_1 \in R^{\frac{C}{r} \times C}$, $W_2 \in R^{C \times \frac{C}{r}}$, $z_i \in R^{C \times 1}$, and r denotes the channel reduction ratio. $\sigma_{Sig}(\cdot)$ and $\sigma_{ReLU}(\cdot)$ are the sigmoid and ReLU activation functions, respectively.

Lastly, the output is produced through a scaling operation, which can be written as:

$$x_i^* = F_{scale}(x_i, z_i) = [z_i(c) \cdot x_i(c)]_{c=1}^{C}, \tag{3}$$

where $z_i(c) \in R^1$, $x_i(c) \in R^{1 \times W}$. The feature vector x_i^* denotes the output of the MSA block. By simply stacking Multiple MSAs, we enable the attention-guided multi-scale branch (AMB) to extract deep and advanced information from vibration signals.

C. Signal transformer branch

Inspired by the Vision Transformer (ViT) [46], we developed the Signal Transformer Branch, which includes a signal embedding layer and several stacked 1-D transformer encoders. These encoders have been specifically modified to process 1-D vibration signals.

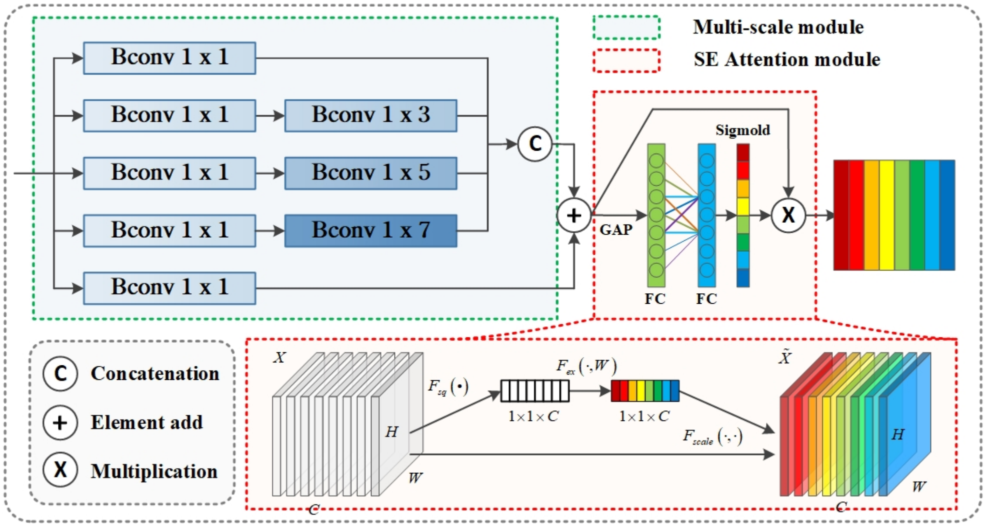


Fig. 2. The structure of the MSA block in AMB.

As illustrated in Fig. 3(b), the input vibration signals are initially divided into N patch embeddings (denoted as $x^i, i \in \{1, 2, ..., N\}$). They are then subjected to the linear operation (denoted as $Linear(\cdot)$) to produce corresponding projected embeddings and an additional learnable position embedding p^*, formulated as follows:

$$\{M^n\}_{n=1}^{N+1} = Linear(x^1, x^2, ..., x^N) + p^* \tag{4}$$

Subsequently, these embeddings are fed into the modified transformer encoder, primarily composed of a multi-head self-attention (MHSA) layer and a followed MLP layer. Layer Normalizations are applied before each layer, and residual connections are implemented in both the MHSA and MLP layers, which can be formulated as follows:

$$\hat{M^n} = Norm(F_{MHSA}(M^n)) + M^n, n = 1, 2, ..., N \tag{5}$$

where $Norm(\cdot)$ denotes the layer normalization operation. $F_{MHSA}(\cdot)$ represents the multi-head self-attention operation and can be described as follows:

$$F_{MHSA}(M^n) = F_{Softmax}(\frac{q(M^n)k(M^n)^T}{\sqrt{d_k}})v(M^n) \tag{6}$$

where $q(\cdot)$, $k(\cdot)$, and $v(\cdot)$ represent the linear projection operations that produce the corresponding query, key, and value matrices, respectively. d_k denotes the number of attention heads used for normalization. Subsequently, an MLP layer is utilized to further aggregate the extracted long-range dependencies, formulated as follows:

$$M^n_{mlp} = Norm(F_{MLP}(\hat{M^n})) + \hat{M^n} \tag{7}$$

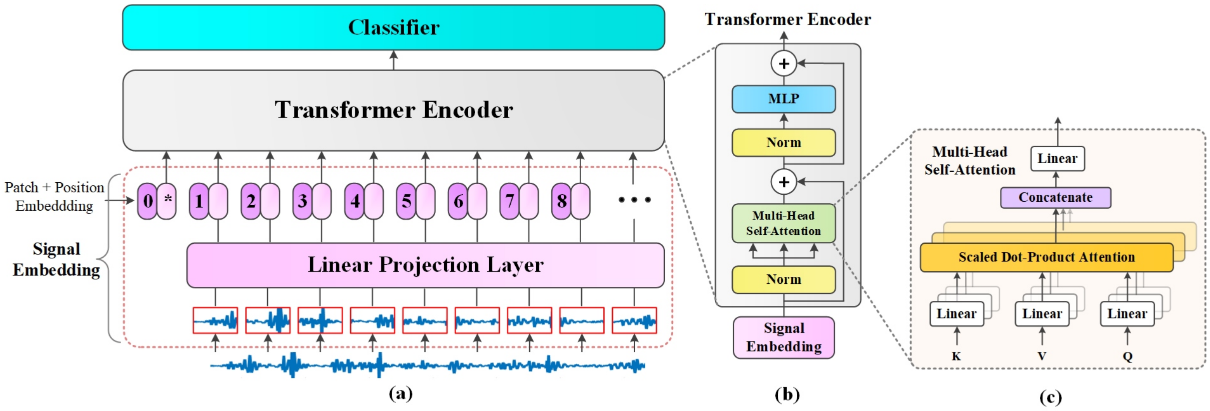


Fig. 3. (a) Structure of the Signal Transformer Branch. (b) Detailed structure of the 1-D transformer encoder. (c) Structure of the multi-head self-attention block.

where, $F_{MLP}(\cdot)$ is an MLP layer, can be formulated as:

$$F_{MLP}(\hat{M^n}) = \sigma_{ReLU}(Linear(\sigma_{ReLU}(Linear(\hat{M^n})))) \tag{8}$$

In our work, to tokenize the input, each input vibration signal is compressed into multiple 1×128 patch embeddings without overlap through a linear projection layer. Additionally, a class token is introduced to these patch embeddings for the subsequent fault diagnosis task.

D. Signal denoising branch

The Signal Denoising Branch (SDB) aims to reduce the interference noise components in the signal and reconstruct a noise-free vibration signal. Inspired by the deep learning-based denoising mechanism, we have designed a symmetrical encoder-decoder denoising network employing convolution and transposed convolution operations. Specifically, the encoder structure is composed of multiple 1-D convolution layers, which are denoted as $c_i, (i = 1, 2, \dots, n-1)$, and n is equal to the number of MFA blocks in the backbone branch. We set the kernel size of the convolution layer in the encoder to 1×1, the number of channels to $32 \times 2_{i-1}$, and the stride to 2.

In the decoder structure, we replace the convolution layers with the transposed convolution operation, which are denoted as $t_i, (i = 1, 2, \dots, n-1)$. The kernel size, the number of channels, and the stride are set to 1×1, $\frac{C}{2^i}(i > 1)$, and 2, respectively. In the last layer of the decoder, the dimension of the output feature map is set the same as the input signal. Each convolution and transposed convolution layer is followed by an SE attention module (as described in Section 2.2) and a Leaky ReLU activation function.

The encoder acts as a feature extractor, which encodes the main components of the signal while eliminating noise components. Then, the decoder decodes the encoded features to reconstruct the signal details. We use an SE attention module to enhance and calibrate useful signal details layer by layer, which helps the signal-denoising branch reconstruct the crucial and elaborate features of the vibration signal accurately.

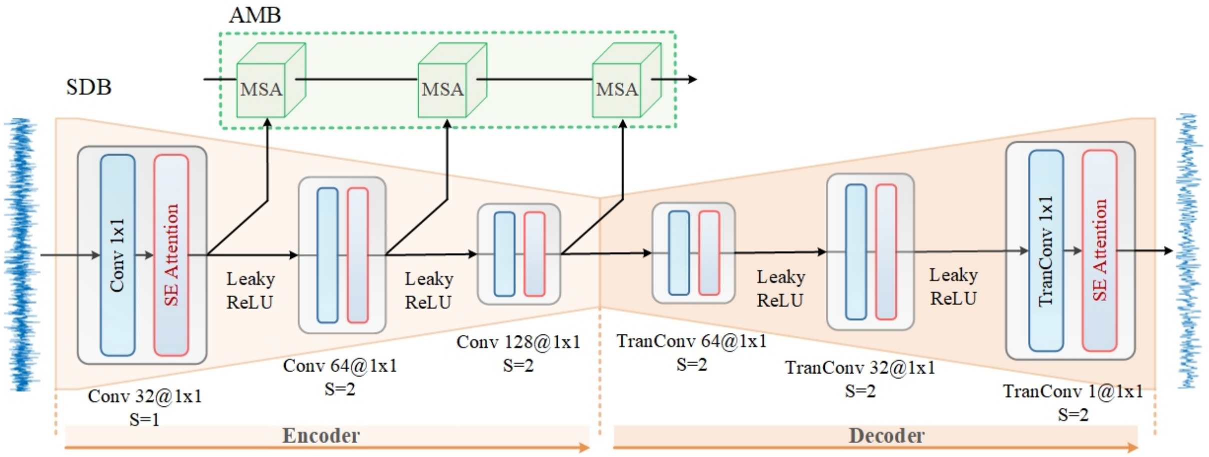


Fig. 4. Structure of the signal denoising branch.

E. Multi-branch cross-cascaded connection via the feature alignment module

A simple yet effective bridge module, termed the Feature Alignment Module (FAM), has been developed to facilitate cross-cascaded connections by eliminating feature misalignment among branches. The FAM consists of a convolution layer (1 × 1) and a Global Average Pooling (GAP) layer in series. For instance, when we connect STB with AMB, the feature maps from the transformer encoders first apply a 1×1 convolution to align the number of channels. Then, we employ a GAP operation to align the spatial dimensions. Finally, we utilize the element-wise summation operator to add the aligned features obtained in the previous step, as shown by the red circle in Fig.5(a).

The SDB and AMB are cross-cascaded and connected similarly as described above. The difference is that the final feature vector from SDB is not directly input into the backbone branch (AMB), instead, it is fed into a specialized loss function. This fashion significantly enhances the overall architecture's capability of identifying noisy characteristics during the denoising process.

Subsequently, a fused feature representation is employed for diagnostic classification. Specifically, let O_t^N and O_a^N represent the final feature vectors extracted from the transformer branch and CNN branch, respectively. Notably, these vectors respectively encompass the fault-related global long-distance feature dependencies and local information. We first perform a concatenation operation to generate a fused feature representation, which is formulated as follows:

$$O_c^N = Concat(F_{fam}^{t \to a}(O_t^N), O_a^N) \tag{9}$$

where $Concat(\cdot)$ denotes the concatenation operation. $F_{fam}^{t \to a}(\cdot)$ indicates the use of the feature alignment module to align the shapes of O_t^N with the backbone output O_a^N. Then, the output probability L_j for status category j can be calculated as:

$$L_j = \frac{exp(\theta^j | F_c(F_{gap}(O_c^N)))}{\sum_{j=1}^{J} exp(\theta^j | F_c(F_{gap}(O_c^N))))}, j = 1, 2, ..., J \tag{10}$$

where θ^j denotes the learned parameter of this fully connected classifier. $F_c(\cdot)$ represents the fully connected operation. $F_{gap}(\cdot)$ is the global average pooling operation.

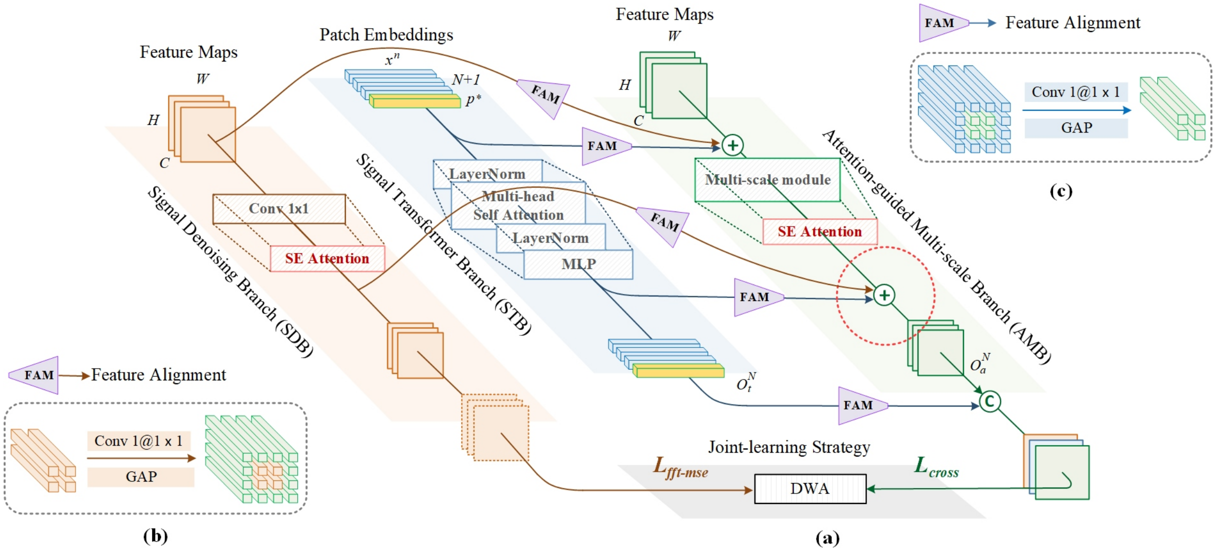


Fig. 5. (a) Schematic of the cross-cascaded connection among the Signal Denoising Branch (SDB), the Attention-guided Multi-scale Branch (AMB), and the Signal Transformer Branch (STB). (b) Feature alignment module that connects SDB with AMB. (c) Feature alignment module that connects STB with AMB.

F. Joint-learning strategy

In this section, a joint-learning strategy is developed to train the proposed Dconformer architecture. Within the joint-learning strategy, two loss functions, i.e., the Fast Fourier Transform-based Mean Squared Error (FFT-MSE) loss and the Cross-Entropy loss, are utilized to independently supervise the denoising and diagnosis tasks. Meanwhile, the joint-learning strategy also incorporates a Dynamic Weight Average (DWA) strategy to maintain a balance between both tasks. Specifically, let $x_i^n \in R^{N\times 1}$ denote the input noisy signal with its corresponding raw signal $g_i^n \in R^{N\times 1}$, and $x_i^s \in R^{N\times 1}$ be the reconstructed denoised signal from the SDB. Within the designed FFT-MSE loss, we initially perform a fast Fourier transform operation to the raw signal and denoised signal, respectively. Followed by the use of mean squared error to calculate their similarity in the frequency domain. In summary, given a mini-batch of N_t samples, the FFT-MSE loss can be formulated as:

$$L_{fft-mse} = \frac{1}{N_t}\sum_{n=1}^{N_t}(F_{fft}(x_i^s) - F_{fft}(g_i^n))^2 \tag{11}$$

In addition, we perform the cross entropy loss (denoted as L_{cross}) to supervise the shift between the predicted distribution y_i and the real distribution c_i, which can be formulated as:

$$L_{cross} = -\frac{1}{N_t}\sum_{n=1}^{N_t} c_i log(y_i) \tag{12}$$

Furthermore, we employ a simple yet effective Dynamic Weight Average (DWA) strategy to adaptively weigh the two loss functions during the training phase. The DWA strategy is defined as follows:

$$L = \omega_f L_{fft-mse} + \omega_c L_{cross} \tag{13}$$

TABLE I
FLOPS AND PARAMETERS OF COMPARISON NETWORKS (%).

Indicator	Dconformer	MK-ResCNN	MBSCNN	Uniformer	Convformer	JL-CNN
FLOPS	1.35×10^9	7.5×10^8	6.2×10^8	1.7×10^9	4.8×10^7	1.5×10^9

where $L_{fft-mse}$ and L_{cross} represent the loss functions supervising the denoising process and the diagnosis task, respectively. $\omega_i, i \in \{f, c\}$ are the weightings of the two loss functions. $\omega_i, i \in \{f, c\}$ can be calculated by:

$$\omega_i(t) = \frac{2 \times exp(\alpha_i(t-1)/\psi)}{exp(\alpha_f(t-1)/\psi) + exp(\alpha_c(t-1)/\psi)}, i \in \{f, c\} \tag{14}$$

$$\alpha_f(t-1) = \frac{L_{fft-mse}(t-1)}{L_{fft-mse}(t-2)} \tag{15}$$

$$\alpha_c(t-1) = \frac{L_{cross}(t-1)}{L_{cross}(t-2)} \tag{16}$$

where, $\alpha_i(\cdot), i \in \{f, c\}$ calculates the relative decrease rate of each loss, i.e., $L_{fft-mse}$ and L_{cross}. t denotes the iteration index. ψ is a parameter that regulates the softness of task weighting. Given a sufficiently large ψ, resulting in $\alpha_i(\cdot) \approx 1, i \in \{f, c\}$, which indicates the tasks are weighted equally. The developed joint learning strategy facilitates mutual enhancement between the diagnostic and denoising tasks, realizing superior performance for both tasks simultaneously, compared to a separate learning approach.

III. EXPERIMENTAL VALIDATION

A. Experimental settings

The code of Dconformer is implemented in Python3.9 and Pytorch 1.10 environment. All experiments are conducted on a workstation with Windows 11 OS, Intel i5-12400F CPU, and GTX3060Ti GPU. Dconformer is compared with five comparative models, namely a multi-kernel-based residual CNN model (MK-ResCNN) [47], a multibranch and multiscale CNN model (MBSCNN) [48], an attention-guided joint learning CNN (JL-CNN) [49], a lightweight CNN-transformer model (Convformer) [50], a CNN and self-attention-based model (Uniformer) [51]. The training strategies of these models are the same as Dconformer. The FLOPS, indicating the computational complexity, of the comparison methods are presented in Table 1. In the training phase, the Adam optimization algorithm with a learning rate of 0.0005 and a batch size of 128 is utilized.

B. Fault diagnosis of the constant-speed bearing dataset

1) Data description: The experimental data set was collected from the ABLT-1A bearing run-to-failure test bench. As shown in Fig. 6, the main components of the ABLT-1A test rig include a loading apparatus, a bearing test module, a driving system, and an electrical control system. The type HRB6205 of bearing is used as the rolling bearing. The sampling frequency is set to 12 kHz, and the acquisition time is one second per set. During the experiment, the motor speed is manually adjusted to 1500 rpm. A total of 7 health states are considered for the

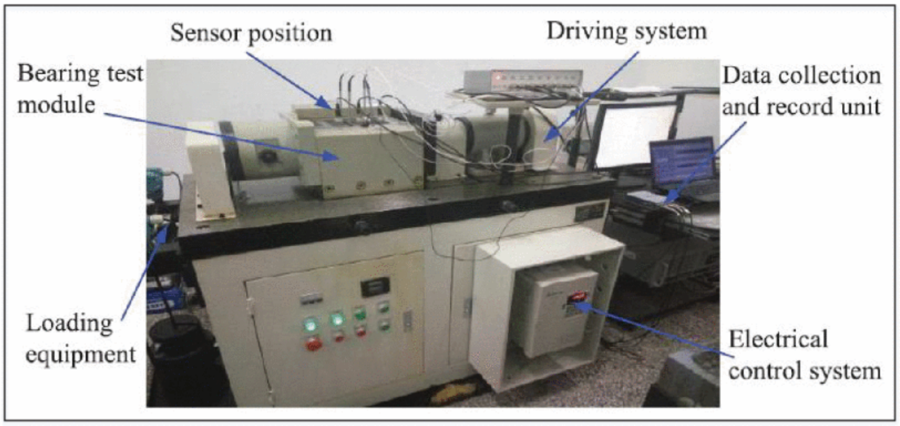


Fig. 6. Test bench of the ABLT-1A bearing.

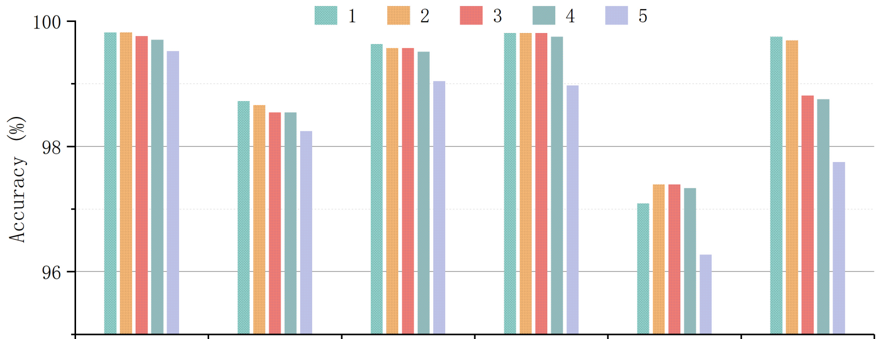
TABLE II
A COMPREHENSIVE EXPLANATION OF THE ABLT-1A BEARING DATASET

Label	Fault Type
H1	Normal condition
H2	Rolling element fault
H3	Inner ring fault
H4	Outer ring fault
H5	Compound fault of outer ring and rolling element
H6	Compound fault of inner and outer ring
H7	Compound fault of the inner-outer ring and rolling element

rolling bearings. There are 800 samples in each category, out of which 400 are allocated for training purposes, while the remaining 400 samples are kept for testing. Consequently, the dataset contains a total of 5600 samples, with 2800 samples reserved for training and another 2800 for testing. Each sample consists of 1024 data points. For additional information regarding the rolling bearing datasets, please refer to Table 2.

2) Comparison with state-of-the-art approaches: In this section, the overall performance of Dconformer is compared with five state-of-the-art methods. Each model is implemented five times to ensure the reliability of the experimental results, and the results are shown in Fig. 7. The average diagnostic accuracy of the proposed model reaches as high as 99.67%. In comparison with MK-ResCNN, MBSCNN, JL-CNN, Uniformer, and Convformer, Dconformer improves the diagnostic accuracy by 1.19%, 0.33%, 0.28%, 2.84%, and 0.92%, respectively. Overall, the proposed Dconformer demonstrates admirable diagnostic performance on the constant-speed bearing dataset.

This section further evaluates the performance of Dconformer in different noisy scenarios. We add different levels of Gaussian white noise to the raw vibration signal to simulate the noise signal under industrial noisy conditions. We set four SNR scenarios (-10, -6, 0, and 6 dB) to simulate extreme, strong, moderate, and weak noise, respectively. As shown in Fig. 8. Dconformer obtains the optimal diagnostic performance in each SNR scenario.



	Dconformer	MK-ResCNN	MBSCNN	Uniformer	Convformer	JL-CNN
Max-acc	99.82	98.72	99.63	99.81	97.39	99.75
Min-acc	99.52	98.24	99.04	98.97	96.27	97.75
Avg-acc	99.67±0.15	98.48±0.24	99.34 ±0.29	99.39±0.42	96.83±0.56	98.75±1.00

Fig. 7. Accuracy of the comparison methods for diagnosing faults in the ABLT-1A benchmark bearing dataset.

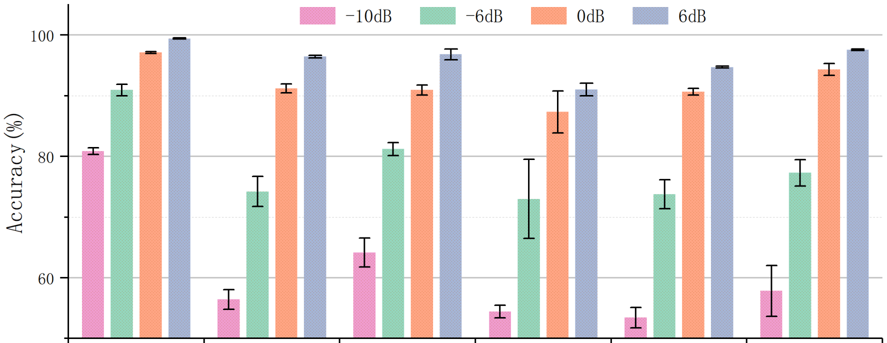
Even when SNR = -10 dB, Dconformer is still able to obtain 80.88% diagnostic accuracy. Compared with the other five methods, Dconformer improves diagnostic accuracy by 16.47%, 9.73%, 17.95%, 17.19%, and 13.65% in the strong noisy condition (SNR= -6 dB), respectively. The experiment result indicates that Dconformer has robust anti-noise ability, and as the noise level increases, the superiority of the proposed method is more significant. Notably, Dconformer outperforms Uniformer and Convformer significantly, indicating that by combining CNN and transformer structure in a multi-branch cross-cascaded architecture, the developed Dconformer can extract abundant local and global features. Dconformer outperforms JL-CNN in all noisy scenarios, indicating that the joint learning strategy promotes the developed framework to obtain better diagnostic results.

Nevertheless, the proposed Dconformer has certain limitations. As depicted in Table 1, its computational complexity, measured in FLOPS, is higher than that of MK-ResCNN, MBSCNN, and notably Convformer. This implies that Dconformer may require more advanced hardware, potentially restricting its deployment in mobile or embedded systems. Nevertheless, Dconformer still can be regarded as a promising network architecture due to its superior and robust diagnostic and denoising performance. For instance, in extremely noisy conditions (SNR = -10 dB), it demonstrates improvements of 24.46%, 16.75%, 26.47%, 27.46%, and 23.04% over the competing approaches, respectively.

We utilize the t-SNE algorithm to visualize the distribution of the extracted features in two-dimensional space under noisy conditions with SNR = -6 dB. As shown in Fig. 9, the features extracted by Dconformer have the best discriminability, showcasing that the network structure of Dconformer can effectively deal with the interference of strong noise and learn more effective fault information from the complex vibration signals.

3) Visualization of denoising results: In this section, the denoising performance of Dconformer is compared with two traditional denoising methods, namely resonance-based sparsity signal decomposition (RSSD) and variational mode decomposition (VMD). To demonstrate the performance of different methods, we perform a fast Fourier transform on the denoised signals and show their square envelope spectrums.

Experiment results indicate that Dconformer can achieve admirable denoising performance, which shows substan-



	Dconformer	MK-ResCNN	MBSCNN	Uniformer	Convformer	JL-CNN
+6dB	99. 43±0. 09	96. 46±0. 22	96. 82±0. 88	91. 04±1. 03	94. 73±0. 18	97. 57±0. 12
0dB	97. 13±0. 16	91. 22±0. 72	90. 95±0. 82	87. 34±3. 45	90. 68±0. 55	94. 34±0. 99
-6dB	90. 95±0. 94	74. 21±2. 48	81. 22±1. 06	73. 00±6. 54	73. 76±2. 39	77. 30±2. 18
-10dB	80. 88±0. 55	56. 42±1. 64	64. 13±2. 39	54. 41±1. 03	53. 42±1. 67	57. 84±4. 21

Fig. 8. Diagnostic results of the six models under four SNR scenarios.

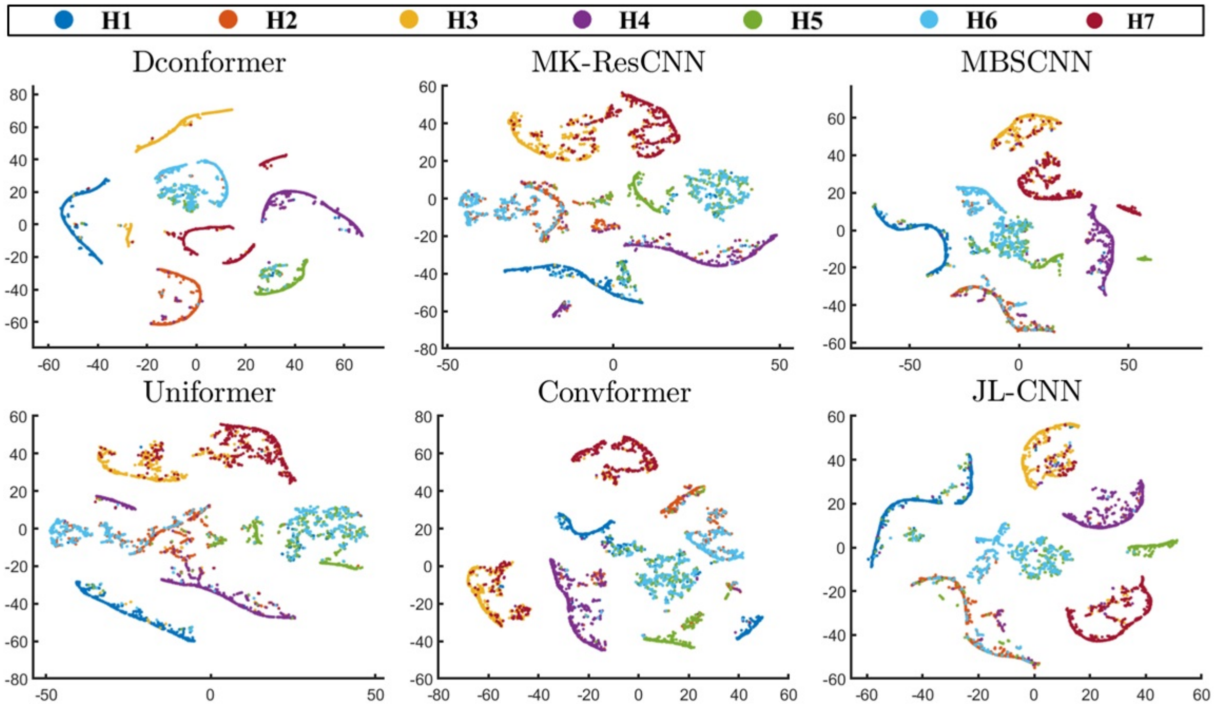


Fig. 9. Visualization of features extracted by the comparative networks under strong noise conditions (SNR = -6 dB).

tial superiority compared to the RSSD and VMD. As shown in Fig. 10, the signals denoised by Dconformer display more visible fault features in the frequency domain, and the waveforms are closer to the raw signals. In addition, in the extreme noise scenario (SNR = -10 dB), the proposed Dconformer can effectively retain the fault-related frequency, while RSSD and VMD excessively remove the fault-related characteristics in the denoising process, as shown in Fig.10 (d).

The superior denoising performance of Dconformer benefits from the developed joint-learning strategy. This

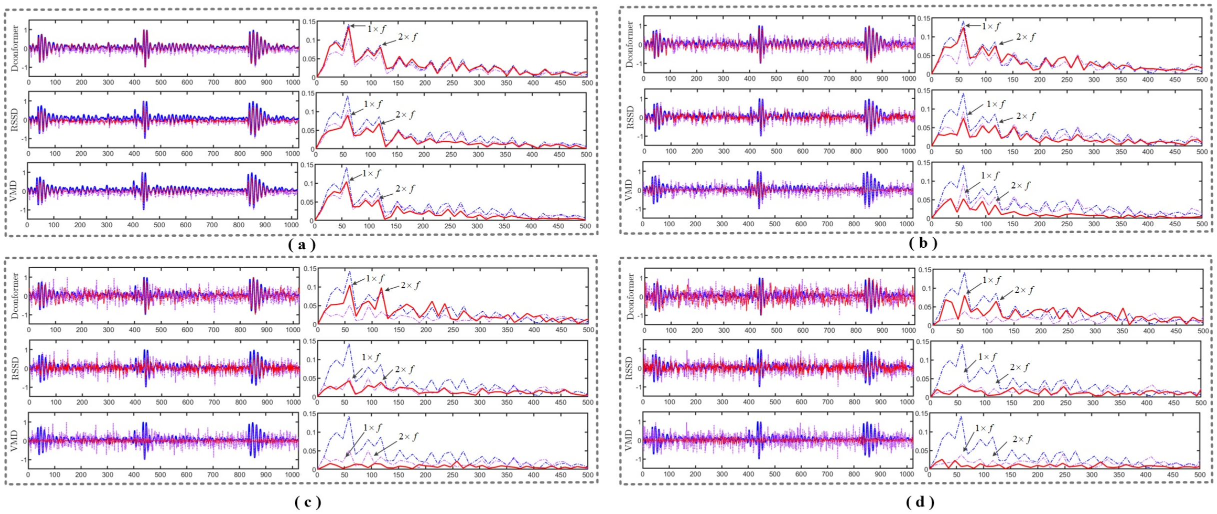


Fig. 10. Visualization of denoised signals by Dconformer, RSSD, and VMD on the ABLT-1A bearing dataset. Waveform and the square envelope spectrum of raw signals, noise-added signals, and denoised signals are displayed. (a), (b), (c), and (d) indicate the results under noise conditions SNR=+6 dB, SNR=0 dB, SNR=-6 dB, and SNR=-10 dB respectively. Blue denotes the raw signals without noise, purple denotes the signals with Gaussian noise added, and red denotes the denoised signals by selected methods.

learning strategy combines the signal-denoising process with the fault diagnosis task, allowing the signal-denoising process to reconstruct signal components that are critical for fault identification. Nevertheless, the denoising process based on RSSD or VMD primarily relies on manually set parameters and operates independently from the diagnostic task. Therefore, these approaches may remove some crucial fault-related features in diagnostic tasks.

C. Fault diagnosis of the variable-speed bearing dataset

1) Data description: The Spectra Quest Variable-Speed (SQV) dataset was obtained from the Spectra Quest composite mechanical fault simulation test bench, as illustrated in Fig. 11. The test bench includes a motor, a rotor system, and a load imposed by a tensioned belt. Vibration signals from the driven end of the motor were recorded using an acceleration sensor and a data acquisition device. The fault simulation experiment used NSK6023 rolling bearings. This dataset was collected during continuous variation of speed. Seven states of the bearing data were selected and described in detail in Table 2. Each state includes a complete process of gradually accelerating from a standstill to 3000rpm, then stabilizing for a period of time, and finally gradually decelerating until it stops. As shown in Fig. 12, the minimum speed for sample acquisition is set to 1050 rpm to ensure that each sample contains enough vibration points for one complete rotational period. The sampling frequency was set to 25.6 kHz in the experiment. Following the previously established sample acquisition criterion, 800 samples were randomly chosen from each state, with 400 samples designated for training and the remaining samples for testing. The dataset consists of a total of 5600 samples, with 2800 samples designated for training and 2800 for testing. Each sample includes 1024 measured vibration points.

2) Comparison with state-of-the-art approaches: In this section, the overall performance of Dconformer is compared with five methods on the variable-speed bearing dataset. Each model is implemented five times to ensure the reliability of the experimental results, and the results are shown in Fig. 13. The overall diagnostic accuracy of

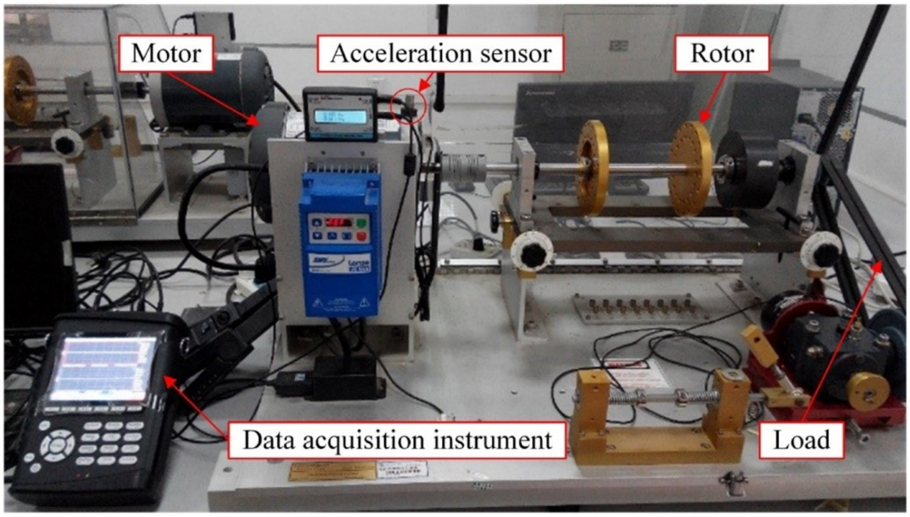


Fig. 11. Spectra Quest Variable-Speed (SQV) bearing test bench.

TABLE III
COMPREHENSIVE INFORMATION REGARDING THE SQV BEARING DATASET

Label	Fault Type
C1	Healthy status
C2	Mild inner race fault
C3	Moderate inner race fault
C4	Severe inner race fault
C5	Mild outer race fault
C6	Moderate outer race fault
C7	Severe outer race fault

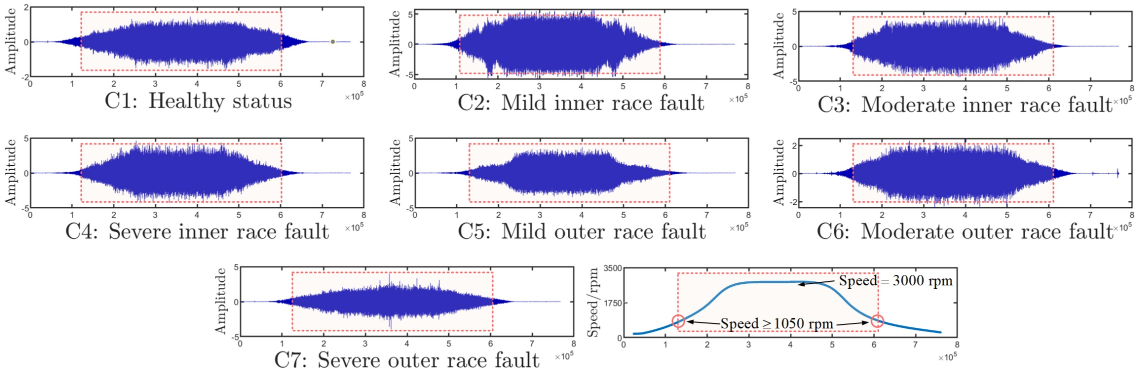
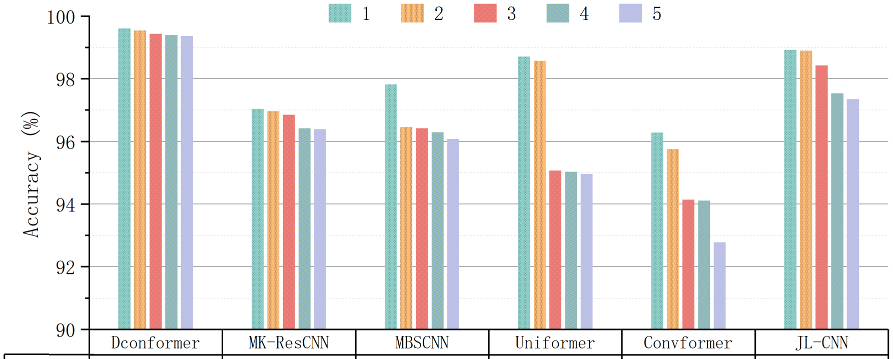


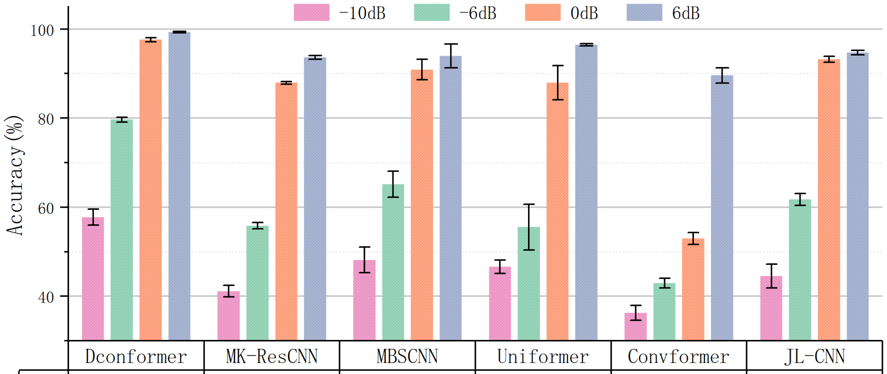
Fig. 12. Detailed description and visualization of the SQV bearing dataset. The red-colored dashed rectangles indicate the ranges of data sample acquisition.

the Dconformer model reaches as high as 99.49%. Compared with MK-ResCNN (96.71%), MBSCNN (96.95%), Uniformer (96.84%), Convformer (94.53%), and JL-CNN (98.14%), Dconformer improves the diagnostic accuracy



	Dconformer	MK-ResCNN	MBSCNN	Uniformer	Convformer	JL-CNN
Max-acc	99.61	97.03	97.82	98.71	96.28	98.92
Min-acc	99.36	96.39	96.07	84.96	92.78	97.35
Avg-acc	99.49±0.13	96.71±0.32	96.95±0.88	96.84±1.88	94.53±1.75	98.14±0.79

Fig. 13. Diagnostic accuracy of the comparison methods on the variable-speed bearing dataset.



	Dconformer	MK-ResCNN	MBSCNN	Uniformer	Convformer	JL-CNN
+6dB	99.29±0.17	93.60±0.43	93.96±2.71	96.48±0.27	89.59±1.70	94.73±0.52
0dB	97.61±0.47	87.93±0.28	90.89±2.28	87.97±3.85	53.00±1.33	93.21±0.68
-6dB	79.66±0.55	55.85±0.68	65.18±2.90	55.53±5.14	42.98±1.09	61.75±1.33
-10dB	57.77±1.81	41.16±1.31	48.18±2.86	46.64±1.50	36.28±1.64	44.53±2.68

Fig. 14. Diagnostic accuracy of the six comparison methods under the four SNR scenarios.

by 2.78%, 2.54%, 2.65%, 4.96%, and 1.35%, respectively. The standard deviation of Dconformer is 0.13, which is lower than the other five state-of-art methods. This showcases that Dconformer maintains robust performance under non-stationary speed conditions. In summary, the proposed Dconformer obtains the finest diagnostic performance on the variable-speed bearing dataset.

We further study the performance of Dconformer on the variable-speed dataset in different noise scenarios (-10, -6, 0, and 6 dB), and the result is shown in Fig. 14. Overall, the diagnostic accuracy of the proposed method on the variable-speed dataset is better than that on the constant-speed dataset. The proposed Dconformer obtains the finest performance among the six methods in each SNR scenario. Compared with MK-ResCNN (93.60%), MBSCNN (93.96%), Uniformer (96.48%), Convformer (89.59%), and JL-CNN (94.73%), The proposed method (99.29%) improves the diagnostic accuracy by 5.69%, 5.33%, 2.81%, 9.70%, and 4.56% in the weak noisy condition (SNR = 6 dB), and by 16.61%, 9.59%, 11.13%, 21.49%, and 13.24% in the extreme noisy condition (SNR = -10 dB).

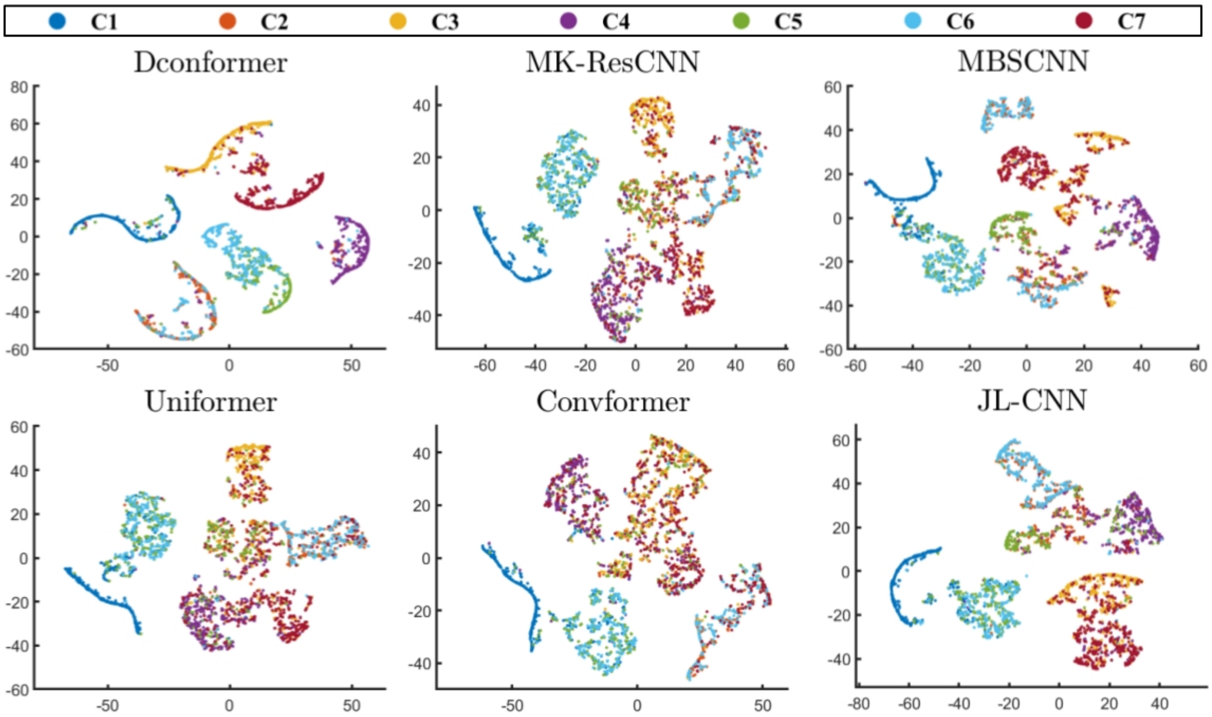


Fig. 15. Visualization of features learned by the six approaches under strong noise conditions with SNR = -6 dB.

We employ the t-SNE algorithm to visualize the feature distribution extracted by six approaches from the variable-speed dataset in a strong noise scenario (SNR=-6dB). As illustrated in Fig. 15, the features learned by Dconformer have the best discriminability, showcasing that Dconformer effectively mitigates the interference of strong noise under variable-speed conditions. It efficiently extracts both fault-related local features and global feature dependencies from the vibration signals, thereby co-constructing a discriminative and comprehensive representation.

3) Visualization of denoising results: In this section, the denoising performance of Dconformer is compared with two traditional denoising methods, i.e., RSSD and VMD. To demonstrate the denoising performance of different methods, we perform a fast Fourier transform on the denoised signals and show their square envelope spectrums.

The experiment results indicate that Dconformer also achieves the finest denoising performance under variable-speed conditions compared to RSSD and VMD. As shown in Fig. 10, the signals denoised by Dconformer display more visible fault features in the frequency domain, and the waveforms are closer to the raw signals. In addition, in the extreme noise scenario (SNR = -10 dB), the proposed Dconformer can effectively retain the fault-related frequency, while RSSD and VMD excessively remove the fault-related characteristics in the denoising process, as shown in Fig.10 (d).

IV. Ablation Study

A. Validation of the signal transformer branch

In this section, we investigate the effectiveness of the signal transformer branch (STB) on the overall performance improvements of the Dconformer. A new architecture termed DCNN is constructed for the experiment. The difference

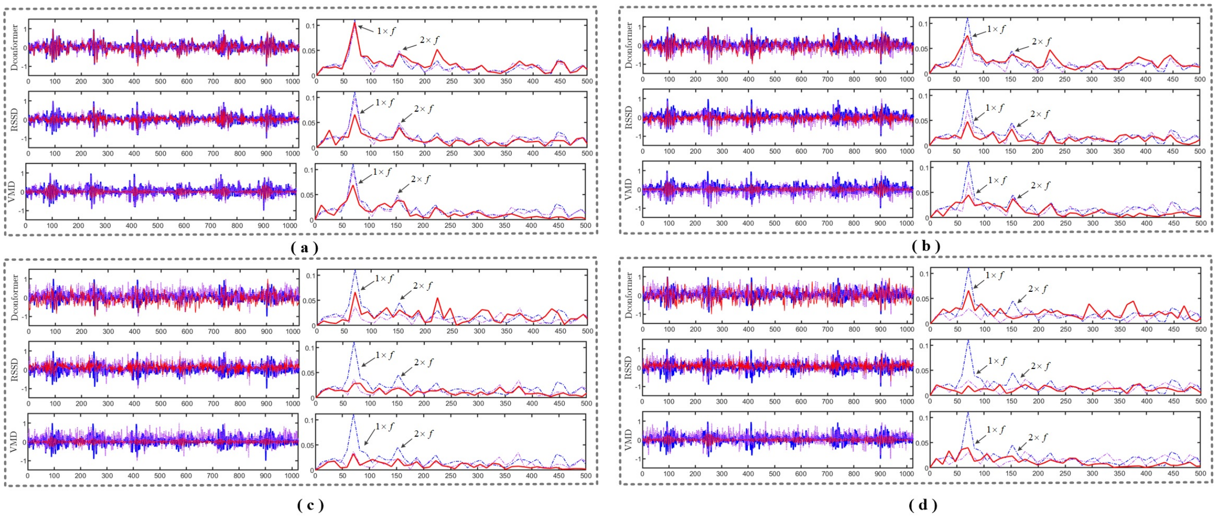


Fig. 16. Visualization of denoised signals by Dconformer, RSSD, and VMD on the SQV bearing dataset. Waveform and the square envelope spectrum of raw signals, noise-added signals, and denoised signals are displayed. (a), (b), (c), and (d) indicate the results under noise conditions SNR=+6 dB, SNR=0 dB, SNR=-6 dB, and SNR=-10 dB respectively. The blue line denotes the raw signals without noise, purple denotes the signals with noise added, and red denotes the denoised signals by selected methods.

TABLE IV
FAULT DIAGNOSTIC ACCURACY OF DCONFORMER AND DCNN IN VARIOUS SCENARIOS [%].

Dataset	Model	Raw signal	SNR = 6 dB	SNR = 0 dB	SNR = -6 dB	SNR = -10 dB
ABLT-1A	DCNN	96.78 ± 0.54%	96.12 ± 0.25%	95.21 ± 0.82%	87.92 ± 0.54%	72.22 ± 0.82%
	Dconformer	99.67 ± 0.15%	99.43 ± 0.09%	97.13 ± 0.16%	90.95 ± 0.94%	80.88 ± 0.55%
SQV	DCNN	97.11 ± 0.19%	96.57 ± 0.43%	96.01 ± 0.70%	78.24 ± 0.47%	55.80 ± 0.89%
	Dconformer	99.49 ± 0.13%	99.29 ± 0.17%	97.61 ± 0.47%	79.66 ± 0.55%	57.77 ± 1.81%

between DCNN and Dconformer is that DCNN removes the STB in the architecture. The ABLT-1A and SQV bearing datasets are utilized for this experiment, and the results are shown in Table 4. We can see that the Dconformer outperforms DCNN by 2.89% and 2.38% on the constant- and variable-speed bearing datasets. We further evaluate the performance of the two methods in different noisy scenarios. Specifically, on the constant-speed dataset, the Dconformer outperforms DCNN by 3.31%, 1.92%, 3.03%, and 3.66% under the noisy conditions with SNR = 6 dB, SNR = 0 dB, SNR = -6 dB, and SNR = -10 dB, respectively. Similarly, on the variable-speed dataset, the Dconformer outperforms DCNN by 2.38%, 2.72%, 1.60%, and 1.97% under the noisy conditions with SNR = 6 dB, SNR = 0 dB, SNR = -6 dB, and SNR = -10 dB, respectively. The results show that the specially designed Signal Transformer Branch (STB) improves the diagnostic accuracy of the Dconformer across all noisy scenarios. This suggests that the global feature dependencies extracted by the STB contribute to constructing a discriminative representation, thereby yielding satisfactory results in harsh environments.

TABLE V
FAULT DIAGNOSTIC ACCURACY OF DCONFORMER AND CONFORMER IN DIFFERENT SCENARIOS [%].

Dataset	Model	Raw signal	SNR = 6 dB	SNR = 0 dB	SNR = -6 dB	SNR = -10 dB
ABLT-1A	Conformer	99.12 ± 0.22%	98.52 ± 0.32%	95.62 ± 0.26%	85.89 ± 0.85%	75.39 ± 0.52%
	Dconformer	99.67 ± 0.15%	99.43 ± 0.09%	97.13 ± 0.16%	90.95 ± 0.94%	80.88 ± 0.55%
SQV	Conformer	99.35 ± 0.45%	99.14 ± 0.91%	94.30 ± 0.45%	73.29 ± 0.43%	52.01 ± 1.34%
	Dconformer	99.49 ± 0.13%	99.29 ± 0.17%	97.61 ± 0.47%	79.66 ± 0.55%	57.77 ± 1.81%

B. *Validation of the signal-denoising branch*

We further study the effectiveness of the signal denoising branch (SDB) on the overall performance improvements of Dconformer. A new architecture termed Conformer is constructed in this experiment. The difference between Conformer and Dconformer is that Conformer removes the SDB from the Dconformer architecture. We utilize the constant-speed and variable-speed bearing datasets in this experiment, and the results are shown in Table 5.

We can see that the Dconformer outperforms Conformer by 0.55% and 0.14% on the constant-speed and variable-speed datasets, respectively. We further test the performance of the Dconformer and Conformer in different noisy scenarios. Specifically, on the constant-speed dataset, the Dconformer outperforms Conformer by 0.91%, 1.51%, 5.06%, and 5.49% under the noisy conditions with SNR = 6 dB, SNR = 0 dB, SNR = -6 dB, and SNR = -10 dB, respectively. Similarly, on the variable-speed dataset, the Dconformer outperforms Conformer by 0.15%, 3.31%, 6.34%, and 5.76% under the noisy conditions with SNR = 6 dB, SNR = 0 dB, SNR = -6 dB, and SNR = -10 dB, respectively. It can be seen that the SDB improves the anti-noise ability of Dconformer. The effect of SDB is more pronounced as the noise level increases. The results indicate that SDB can reduce the irrelevant noise components embedded in the signal and retain the crucial status signal details.

C. *Validation of the joint-learning strategy*

During the training phase, the proposed Dconformer architecture is trained using a developed joint-learning strategy. This strategy integrates the signal-denoising task with the fault diagnosis task, simultaneously enhancing the performance of both tasks. The core of the joint-learning strategy is the application of the Dynamic Weight Average (DWA) strategy, which supervises the balance between dual tasks. To illustrate the effectiveness of the DWA strategy, an experiment was conducted comparing it to the fixed weight (FW) strategy commonly employed in related research [49]. Experimental results of the two strategies on the constant- and variable-speed bearing datasets are shown in Table 5. For the FW strategy, we manually set the weights for the two loss functions. Specifically, we test the performance of three types of FW strategy, termed Type1, Type2, and Type3, respectively. Type1 represents the case where the fixed weights of the dual-loss values are set to 0.2 and 0.8 respectively; Type2 is the case where the fixed weights of the dual-loss values are both set to 0.5; and Type3 corresponds to the case where the fixed weights of the dual-loss values are set to 0.8 and 0.2 respectively. As indicated in Table 5, the DWA strategy surpasses the Fixed Weight (FW) strategy in both constant- and variable-speed datasets. This suggests that the DWA strategy enhances the training of the Dconformer more effectively through a more refined weighting fashion.

TABLE VI
FAULT DIAGNOSTIC ACCURACY OF DWA AND DIFFERENT FW STRATEGIES IN SNR SCENARIOS [%].

Dataset	Model	Raw signal	SNR = 6 dB	SNR = 0 dB	SNR = -6 dB	SNR = -10 dB
ABLT-1A	DWA	99.67 ± 0.15%	99.43 ± 0.09%	97.13 ± 0.16%	90.95 ± 0.94%	80.88 ± 0.55%
	Type1	98.49 ± 0.31%	98.38 ± 0.07%	96.74 ± 0.35%	87.60 ± 0.78%	78.64 ± 0.72%
	Type2	99.43 ± 0.24%	98.32 ± 0.12%	96.91 ± 0.24%	89.01 ± 0.95%	78.85 ± 0.69%
	Type3	96.73 ± 0.30%	95.74 ± 0.22%	92.66 ± 0.38%	84.43 ± 0.35%	75.20 ± 0.53%
SQV	DWA	99.49 ± 0.13%	99.29 ± 0.17%	97.61 ± 0.47%	79.66 ± 0.55%	57.77 ± 1.81%
	Type1	99.20 ± 0.21%	98.19 ± 0.22%	95.88 ± 0.23%	76.91 ± 0.68%	55.46 ± 1.61%
	Type2	98.45 ± 0.41%	98.96 ± 0.07%	97.12 ± 0.21%	78.79 ± 0.84%	56.51 ± 0.89%
	Type3	97.32 ± 0.15%	96.22 ± 0.01%	93.69 ± 0.39%	74.92 ± 0.73%	55.37 ± 1.73%

V. CONCLUSION

This paper proposed a novel end-to-end CNN-transformer network, termed Dconformer, for intelligent fault diagnosis of rolling bearings. First, an attention-guided multi-scale branch was developed to extract deep and advanced local features from multiple scales. Meanwhile, a signal transformer branch (STB) was adopted to capture global long-distance feature dependencies from the input signals. Further, a signal denoising branch (SDB) was introduced to reduce the irrelevant noise components embedded in the signal and reconstruct the noise-free vibration signal. Finally, a joint-learning strategy was utilized to facilitate mutual improvement between the diagnostic and denoising tasks, leading to superior performance for both tasks simultaneously.

Two case studies, utilizing constant- and variable-speed bearing datasets, were conducted to validate the effectiveness of the developed Dconformer. Experimental results indicate that Dconformer achieves overall accuracies of 99.67% and 99.49% on these datasets, respectively. Furthermore, under noisy conditions with a signal-to-noise ratio of -6 dB, Dconformer still maintains diagnostic accuracies of 90.95% and 79.66%, respectively. Extensive experimental results showcase that Dconformer possesses exceptional diagnostic capabilities and noise robustness, surpassing state-of-the-art competing approaches, particularly in strong noise scenarios. Additionally, this study validates the effectiveness of the key components in the Dconformer, i.e., the signal transformer branch, the signal denoising branch, and the joint-learning strategy.

Nevertheless, although the Dconformer shows significant potential in intelligent bearing fault diagnosis, it still possesses certain limitations, such as high computational complexity, which need further improvement. Thereby, future research aims to incorporate more lightweight designs into the current architecture and to extend the application of this method to other industrial domains, including gears and robot reducers.

ACKNOWLEDGMENT

This work was partly supported by the Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant No. KYCX23_0647) and partly by the China Scholarship Council (CSC).

References

[1] K. Feng, J. Ji, Q. Ni, and M. Beer, "A review of vibration-based gear wear monitoring and prediction techniques," *Mechanical Systems and Signal Processing*, vol. 182, p. 109605, 2023.

[2] Y. Xu, X. Yan, B. Sun, and Z. Liu, "Global contextual residual convolutional neural networks for motor fault diagnosis under variable-speed conditions," *Reliability Engineering & System Safety*, DOI https://doi.org/10.1016/j.ress.2022.108618, p. 108618, 2022.

[3] B. Hou, D. Wang, T. Xia, Z. Peng, and K.-L. Tsui, "Difference mode decomposition for adaptive signal decomposition," *Mechanical Systems and Signal Processing*, vol. 191, p. 110203, 2023.

[4] X. Liu, Y. Lei, N. Li, X. Si, and X. Li, "Rul prediction of machinery using convolutional-vector fusion network through multi-feature dynamic weighting," *Mechanical Systems and Signal Processing*, vol. 185, p. 109788, 2023.

[5] B. Hou, X. Feng, J.-Z. Kong, Z. Peng, K.-L. Tsui, and D. Wang, "Optimized weights spectrum autocorrelation: A new and promising method for fault characteristic frequency identification for rotating machine fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 191, p. 110200, 2023.

[6] Y. Xu, X. Yan, K. Feng, X. Sheng, B. Sun, and Z. Liu, "Attention-based multiscale denoising residual convolutional neural networks for fault diagnosis of rotating machinery," *Reliability Engineering & System Safety*, DOI https://doi.org/10.1016/j.ress.2022.108714, p. 108714, 2022.

[7] S. Li, J. Ji, Y. Xu, X. Sun, K. Feng, B. Sun, Y. Wang, F. Gu, K. Zhang, and Q. Ni, "Ifd-mdcn: Multibranch denoising convolutional networks with improved flow direction strategy for intelligent fault diagnosis of rolling bearings under noisy conditions," *Reliability Engineering & System Safety*, vol. 237, p. 109387, 2023.

[8] Y. Xu, X. Yan, B. Sun, and Z. Liu, "Dually attentive multiscale networks for health state recognition of rotating machinery," *Reliability Engineering & System Safety*, DOI https://doi.org/10.1016/j.ress.2022.108626, p. 108626, 2022.

[9] K. Feng, J. Ji, and Q. Ni, "A novel adaptive bandwidth selection method for vold–kalman filtering and its application in wind turbine planetary gearbox diagnostics," *Structural Health Monitoring*, p. 14759217221099966, 2022.

[10] Y. Xu, X. Yan, B. Sun, and Z. Liu, "Hierarchical multiscale dense networks for intelligent fault diagnosis of electromechanical systems," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[11] H. Wang, J. Xu, R. Yan, and R. X. Gao, "A new intelligent bearing fault diagnosis method using sdp representation and se-cnn," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 5, pp. 2377–2389, 2019.

[12] D. Huang, S. Li, N. Qin, and Y. Zhang, "Fault diagnosis of high-speed train bogie based on the improved-ceemdan and 1-d cnn algorithms," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, 2021.

[13] J. Chen, R. Huang, K. Zhao, W. Wang, L. Liu, and W. Li, "Multiscale convolutional neural network with feature alignment for bearing fault diagnosis," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–10, 2021.

[14] B. Ma, W. Cai, Y. Han, and G. Yu, "A novel probability confidence cnn model and its application in mechanical fault diagnosis," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, 2021.

[15] S. Gao, S. Shi, and Y. Zhang, "Rolling bearing compound fault diagnosis based on parameter optimization mckd and convolutional neural network," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–8, 2022.

[16] Y. Gao, L. Gao, X. Li, and S. Cao, "A hierarchical training-convolutional neural network for imbalanced fault diagnosis in complex equipment," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 11, pp. 8138–8145, 2022.

[17] Y. Xu, X. Yan, B. Sun, and Z. Liu, "Deep coupled visual perceptual networks for motor fault diagnosis under nonstationary conditions," *IEEE/ASME Transactions on Mechatronics*, vol. 27, no. 6, pp. 4840–4850, 2022.

[18] Y. Xu, X. Yan, B. Sun, J. Zhai, and Z. Liu, "Multireceptive field denoising residual convolutional networks for fault diagnosis," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 11, pp. 11 686–11 696, 2021.

[19] G. Li, J. Wu, C. Deng, Z. Chen, and X. Shao, "Convolutional neural network-based bayesian gaussian mixture for intelligent fault diagnosis of rotating machinery," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–10, 2021.

[20] H. Wang, J. Xu, R. Yan, and R. X. Gao, "A new intelligent bearing fault diagnosis method using sdp representation and se-cnn," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 5, pp. 2377–2389, 2019.

[21] Y. Zhao, Y. He, Z. Xing, Y. Fu, J. Chen, B. Du, and L. Wang, "Multibranch 1-d cnn based on attention mechanism for the dab converter fault diagnosis," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, DOI 10.1109/TIM.2022.3203445, pp. 1–12, 2022.

[22] H. Wang, Z. Liu, D. Peng, and M. J. Zuo, "Interpretable convolutional neural network with multilayer wavelet for noise-robust machinery fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 195, p. 110314, 2023.

[23] J.-X. Liao, H.-C. Dong, Z.-Q. Sun, J. Sun, S. Zhang, and F.-L. Fan, "Attention-embedded quadratic network (qttention) for effective and interpretable bearing fault diagnosis," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–13, 2023.

[24] J. Bai, Z. Wen, Z. Xiao, F. Ye, Y. Zhu, M. Alazab, and L. Jiao, "Hyperspectral image classification based on multibranch attention transformer networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, DOI 10.1109/TGRS.2022.3196661, pp. 1–17, 2022.

[25] X. Liu, Y. Wu, W. Liang, Y. Cao, and M. Li, "High resolution sar image classification using global-local network structure based on vision transformer and cnn," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, DOI 10.1109/LGRS.2022.3151353, pp. 1–5, 2022.

[26] F. Yuan, Z. Zhang, and Z. Fang, "An effective cnn and transformer complementary network for medical image segmentation," *Pattern Recognition*, vol. 136, p. 109228, 2023.

[27] G. I. Okolo, S. Katsigiannis, and N. Ramzan, "Ievit: An enhanced vision transformer architecture for chest x-ray image classification," *Computer Methods and Programs in Biomedicine*, vol. 226, p. 107141, 2022.

[28] X. Jiang, Y. Zhu, G. Cai, B. Zheng, and D. Yang, "Mxt: A new variant of pyramid vision transformer for multi-label chest x-ray image classification," *Cognitive Computation*, vol. 14, no. 4, pp. 1362–1377, 2022.

[29] J. Shi, Y. Wang, Z. Yu, G. Li, X. Hong, F. Wang, and Y. Gong, "Exploiting multi-scale parallel self-attention and local variation via dual-branch transformer-cnn structure for face super-resolution," *IEEE Transactions on Multimedia*, 2023.

[30] J. Chen, X. Chen, S. Chen, Y. Liu, Y. Rao, Y. Yang, H. Wang, and D. Wu, "Shape-former: Bridging cnn and transformer via shapeconv for multimodal image matching," *Information Fusion*, vol. 91, pp. 445–457, 2023.

[31] P. Song, J. Li, Z. An, H. Fan, and L. Fan, "Ctmfnet: Cnn and transformer multiscale fusion network of remote sensing urban scene imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2022.

[32] G. Zhao, Q. Ye, L. Sun, Z. Wu, C. Pan, and B. Jeon, "Joint classification of hyperspectral and lidar data using a hierarchical cnn and transformer," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2022.

[33] B. Zhang, Y. Chen, Y. Rong, S. Xiong, and X. Lu, "Matnet: A combining multi-attention and transformer network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[34] S. Li, Q. Jiang, Y. Xu, K. Feng, Y. Wang, B. Sun, X. Yan, X. Sheng, K. Zhang, and Q. Ni, "Digital twin-driven focal modulation-based convolutional network for intelligent fault diagnosis," *Reliability Engineering & System Safety*, vol. 240, p. 109590, 2023.

[35] Y. Xu, X. Yan, K. Feng, Y. Zhang, X. Zhao, B. Sun, and Z. Liu, "Global contextual multiscale fusion networks for machine health state identification under noisy and imbalanced conditions," *Reliability Engineering & System Safety*, vol. 231, p. 108972, 2023.

[36] L. Wang, Z. Liu, H. Cao, and X. Zhang, "Subband averaging kurtogram with dual-tree complex wavelet packet transform for rotating machinery fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 142, p. 106755, 2020.

[37] S. Li, Y. Xu, K. Feng, Y. Wang, B. Sun, X. Yan, X. Sheng, K. Zhang, J. Zheng, and Q. Ni, "Joint threshold learning convolutional networks for intelligent fault diagnosis under nonstationary conditions," *IEEE Transactions on Instrumentation and Measurement*, 2023.

[38] Y. Yao, G. Gui, S. Yang, and S. Zhang, "A recursive denoising learning for gear fault diagnosis based on acoustic signal in real industrial noise condition," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–15, 2021.

[39] H. Xiong, Z. Wang, G. Wu, Y. Pan, Z. Yang, and Z. Long, "Steering actuator fault diagnosis for autonomous vehicle with an adaptive denoising residual network," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–13, 2022.

[40] H. Han, H. Wang, Z. Liu, and J. Wang, "Intelligent vibration signal denoising method based on non-local fully convolutional neural network for rolling bearings," *ISA transactions*, vol. 122, pp. 13–23, 2022.

[41] Z. Zhi, L. Liu, D. Liu, and C. Hu, "Fault detection of the harmonic reducer based on cnn-lstm with a novel denoising algorithm," *IEEE Sensors Journal*, vol. 22, no. 3, pp. 2572–2581, 2021.

[42] B. Zhao, C. Cheng, Z. Peng, Q. He, and G. Meng, "Hybrid pre-training strategy for deep denoising neural networks and its application in machine fault diagnosis," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, 2021.

[43] T. Li, C. Sun, S. Li, Z. Wang, X. Chen, and R. Yan, "Explainable graph wavelet denoising network for intelligent fault diagnosis," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[44] D. Zhao, H. Zhang, S. Liu, Y. Wei, and S. Xiao, "Deep rational attention network with threshold strategy embedded for mechanical fault diagnosis," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–15, 2021.

[45] S. Zhang, Z. Liu, Y. Chen, Y. Jin, and G. Bai, "Selective kernel convolution deep residual network based on channel-spatial attention mechanism and feature fusion for mechanical fault diagnosis," *ISA transactions*, 2022.

[46] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[47] R. Liu, F. Wang, B. Yang, and S. J. Qin, "Multiscale kernel based residual convolutional neural network for motor fault diagnosis under nonstationary conditions," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 6, pp. 3797–3806, 2019.

[48] D. Peng, H. Wang, Z. Liu, W. Zhang, M. J. Zuo, and J. Chen, "Multibranch and multiscale cnn for fault diagnosis of wheelset bearings under strong noise and variable load condition," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 7, pp. 4949–4960, 2020.

[49] K. Li, Y. Wang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, "Uniformer: Unified transformer for efficient spatiotemporal representation learning," *arXiv preprint arXiv:2201.04676*, 2022.

[50] S. Han, H. Shao, J. Cheng, X. Yang, and B. Cai, "Convformer-nse: A novel end-to-end gearbox fault diagnosis framework under heavy noise using joint global and local information," *IEEE/ASME Transactions on Mechatronics*, 2022.

[51] H. Wang, Z. Liu, D. Peng, and Z. Cheng, "Attention-guided joint learning cnn with noise robustness for bearing fault diagnosis and vibration signal denoising," *ISA transactions*, vol. 128, pp. 470–484, 2022.