Time-frequency denoising and optimization network for characterizing nonstationary rotating machinery signals

Depei Shao , Dezun Zhao , Tianyang Wang

PII: S0031-3203(26)00621-7
DOI: https://doi.org/10.1016/j.patcog.2026.113656
Reference: PR 113656

To appear in: *Pattern Recognition*

Received date: 11 December 2025
Revised date: 1 March 2026
Accepted date: 30 March 2026

Please cite this article as: Depei Shao , Dezun Zhao , Tianyang Wang , Time-frequency denoising and optimization network for characterizing nonstationary rotating machinery signals, *Pattern Recognition* (2026), doi: https://doi.org/10.1016/j.patcog.2026.113656

Highlights:

- A data-driven time-frequency denoising and optimized network is proposed to characterize TFR with high quality.
- Attention-guided sparse denoising sub-network is designed to eliminate noise aliasing interference.
- Stacked Transformer-based blocks are constructed to enhance time-frequency resolution.
- The superiority is validated in characterizing the signal and pattern recognition of the fault.

Time-frequency denoising and optimization network for characterizing nonstationary rotating machinery signals

Depei Shao[1], Dezun Zhao[1,2,*], Tianyang Wang[3]

[1]Beijing Key Laboratory of Advanced Manufacturing Technology, Beijing University of Technology, Beijing 100124, China

[2]Beijing Engineering Research Center of Precision Measurement Technology and Instruments, Beijing University of Technology, Beijing 100124, China

[3]Department of Mechanical Engineering, Tsinghua University, Beijing, 100084, China

Abstract:

Due to the strong non-stationarity of rotating machinery vibration signals under coupling effects of noise environments and time-varying conditions, traditional time-frequency analysis (TFA) methods and existing time-frequency networks struggle to dynamically characterize closely-spaced instantaneous frequencies (IFs) under noisy environments. Therefore, the time-frequency denoising and optimization network (TFDON) is proposed. In the TFDON, an attention-guided sparse denoising sub-network (SDSN) is first designed to eliminate noise aliasing interference and obtain the clean time-frequency representation (TFR). Then, the time-frequency optimization sub-network (TFOSN) with three-stage hybrid Transformer blocks (HTB) cascade is constructed. Within each HTB, an efficient grouped Swin-Transformer (EGST) is developed to compute the spatiotemporal characteristics, and guided by a dual-layer attention mechanism, the time-frequency concentration is iteratively enhanced. Additionally, a weight-controllable joint loss function tailored for the TFR denoising and optimization is designed to achieve the optimal balance in two tasks. The performance of the TFDON in characterizing and noise suppression is verified by a simulated signal with closely-spaced IFs. Meanwhile, two bearing and a planetary gearbox vibration signal added noise are further analyzed, and the TFDON achieves the lowest Rényi entropy of 6.055, 6.387, and 6.077 at −5 dB,

*Corresponding author.

respectively, reducing values by 6%–21% compared with existing TFA methods. Results show the TFDON improving the robustness and characterization concentration of closely-spaced IFs under noise aliasing.

Keywords: Data-driven time-frequency analysis; Transformer networks; Signal processing.

Nomenclature

Abbreviations		Variables	
TFA	Time-frequency analysis	$x(\tau)$	The input signal of the STFT
IF	Instantaneous frequency	$*$	Two-dimensional convolution operation
TFDON	Time-frequency denoising and optimization network	d_i	Dilation rates
TFR	Time-frequency representation	$\beta(\cdot)$	Batch normalization operation
HTB	Hybrid Transformer blocks	$\delta(\cdot)$	The gated Sigmoid function
SDSN	Sparse denoising sub-network	T	The Tanh activation function
TFOSN	Time-frequency optimization sub-network	λ	Weighting factor in the loss function
CT	Chirplet transform	PE	Patch embedding
GLCT	Generalized linear CT	PU	Patch unembedding
STFT	Short-time Fourier transform	1_H	All-one matrix of size $H\times H$
SST	Synchrosqueezing transform	H	Number of attention heads
MSST	Multisynchrosqueezing transform	A_l	Amplitude intensity of each signal
SSET	Second-order synchroextracting transform	$h_l(t)$	Randomly added Gaussian white noise
VSLCT	Velocity synchronous linear CT	A_n^p	The assignment in the penalty matrix

1. Introduction

As the power core of the modern industrial system, rotating mechanical equipment occupies an irreplaceable position in key fields such as wind turbines, aeroengine and so on [1]. However, prolonged operation under extreme conditions of high speed, strong noise, and variable load can lead to performance degradation and a decline in the structural integrity of critical components, such as bearings and gears within planetary gearboxes [2]. The degradation directly threatens the continuity of production and the intrinsic safety of the entire industrial system. Under this background, the development of fault recognition technology with robust characteristics has become a strategic requirement to ensure operational reliability and optimize the maintenance strategies of industrial assets [3].

E-mail address: dzzhao0903@bjut.edu.cn (D. Zhao), shaodepei@emails.bjut.edu.cn (D. Shao), wty19850925@tsinghua.edu.cn (T. Wang).

The TFA is a core mathematical tool for processing non-stationary signals, which can accurately reveal the energy distribution variation law in two dimensions of time and frequency. Hence, the TFA provides a powerful means for characterizing non-stationary signals generated in rotating machinery [4]. At present, mainstream TFA algorithms can be categorized into basis function transformation-based and deep learning-based TFA.

Within the category of basis function transformation-based methods, representative approaches such as the short-time Fourier transform (STFT) apply a sliding window to obtain TFRs [5]. Because the STFT is limited by the Heisenberg uncertainty principle, it cannot provide high time-frequency resolution at the same time. As an alternative approach, the Wigner–Ville distribution (WVD) was developed, whose principle is to apply the Fourier transform on the instantaneous auto or cross correlation function to obtain the TFR of the signal [6]. However, the WVD will produce cross terms for multi-component signals, which is also the main defect in the application.

To further improve the time-frequency concentration, more researchers have developed post-processing TFA techniques [7]. For instance, Oberlin T et al. [8] proposed the synchrosqueezing transform (SST), which compresses time-frequency coefficients of the STFT result into the IFs along the frequency direction, thereby improving the readability of the TFR. However, the SST can only provide a better TFR for weakly frequency-modulated and amplitude-modulated signals. To enhance the applicability of synchrosqueezing methods, Yu et al. [9,10] further presented the multisynchrosqueezing transform (MSST) and local maximum SST (LMSST). These two methods improve sensitivity to subtle time-varying features by iteratively estimating the IFs by multiple synchrosqueezing operations and a local search strategy, respectively, but they are difficult to deal with signals with closely spaced frequencies. Building on the advantages of the synchrosqueezing processing concept, another post-processing TFA method, termed the synchro-extracting transform (SET), was developed by Yu et al. [11]. The SET extracts the ridges of the STFT by a synchroextracting operator, but it only retains the time-frequency coefficients at fixed points of the

original IF estimation, resulting in unsatisfactory reconstruction results. To optimize the reconstruction effect and time-frequency energy of time-varying signals, Bao et al. [12] designed the second-order synchroextracting transform (SSET) by introducing the second-order phase derivative, achieving more precise energy concentration. Nevertheless, when dealing with complex signals in noisy environments, the reconstruction accuracy remains insufficient.

In contrast to the above non-parametric methods, parametric TFA constructs high-resolution TFRs based on assumed signal models and parameter estimation. Initially, the chirplet transform (CT) was proposed by Mann et al. [13], which introduced a chirp rate on the basis of the STFT and can match linear frequency-modulated signals, but it is difficult to characterize diverse nonlinear signals. To extend the applicability of the CT, Peng et al. [14] proposed the polynomial CT (PCT). In the PCT, the linear chirp kernel is replaced with a polynomial kernel to fit nonlinear frequency-modulated signals, but its performance is limited for multi-component signals. To address this issue, Yu et al. [15] developed the general linear CT (GLCT), which reconstructed the TFR by selecting the optimal value from multiple CTs. However, the GLCT introduces cross-terms when processing a multi-component signal, reducing representation accuracy. To continually optimize the characterization precision, Guan et al. [16] proposed the velocity synchronous linear CT (VSLCT), which dynamically adjusts the time-frequency resolution by selecting appropriate basis functions. However, the VSLCT is limited to a single fundamental frequency within each analysis window and is only suitable for signals with proportional frequency components. To overcome the limitation of VSLCT in representing non-proportional frequency components, Zhao et al. proposed the HDSCT by introducing an additional chirp rate and window length dimension, thereby enhancing the concentration of complex IF components. However, due to the influence of the original parametric methods, the overall noise robustness requires further improvement [18].

In recent years, data-driven methods have been introduced into the TFA domain, and a complete technical route has been established from post-processing optimization to end-to-end modeling [19]. In terms of post-processing networks, Wang et al. [20] constructed a convolutional encoder-decoder model based on deconvolution theory to balance time and frequency resolution of the STFT, and then obtained a better time-frequency kernel to enable high-resolution representation of simple, few-component signals. To characterize the more complex time-frequency ridges, Zhao et al. [21] proposed the CTNet, which integrated a deeper residual auto-encoder network and the convolutional block attention mechanism to eliminate cross-terms of the TFR of the GLCT. However, these methods are mainly built on simple encoder–decoder architectures, which lead to information loss during feature transformation and limit their representation capability. To fully integrate the advantages of traditional TFA and deep learning theories, Chen et al. [22] presented the QTFN by learning data-driven basis functions and incorporating attention and multi-scale learning mechanisms, thereby enhancing the processing capabilities for multi-component signals. Subsequently, Zhao et al. [23] constructed the time-frequency self-similarity enhancement network (TFSSEN) guided by physical knowledge. In the TFSSEN, the first layer introduces kernel function constraints, and the cascaded residual group architecture is constructed to enhance generalization in signal processing. However, the QTFN and TFSSEN both rely on prior time-frequency knowledge. To get rid of this dependence, Pan et al. [24] designed the TFA-Net to directly characterize the 2D TFR from vibration signals. The core of the TFA-Net is learning basis functions by a single-scale complex-valued convolutional layer and enhancing the time-frequency concentration through a residual network. Due to the limitations of single-scale basis function learning, the TFA-Net struggles to extract more comprehensive features from complex vibration signals. Based on this, Chen et al. [25] constructed an AMTFN, which learns more diverse basis functions from multi-scale layers and introduces an attention mechanism for adaptive selection among multiple basis functions, improving the

performance of the time-frequency network. However, the methods described above are not specifically designed for noise removal. Recent studies have further introduced multi-stage signal processing techniques into the denoising tasks [26]. For example, Biswas et al. [27] proposed the HRSpecNet, which enhances the signal-to-noise ratio (SNR) via an autoencoder and integrates convolutional STFT with a U-Net to generate high-resolution TFRs. However, the characterization effect of signals with closely-spaced frequencies still needs to be improved under the condition of low SNR.

Although the aforementioned deep learning-based TFA methods have demonstrated outstanding performance in signal processing, they are difficult to handle signals with closely-spaced IFs and heavily disturbed by noise, which limits their practical application scope. Based on this, to enhance fault pattern recognition of rotating machinery under noisy conditions, a novel deep learning model based on CNN networks and the Transformer model is proposed, termed the time-frequency denoising and optimization network (TFDON). The main contributions are as follows:

1) An attention-guided sparse denoising sub-network (SDSN) is developed to mitigate noise aliasing in nonstationary time-frequency representations. By leveraging multi-scale receptive fields and adaptive feature refinement, SDSN enhances noise suppression while preserving closely-spaced IF structures.

2) The time-frequency optimization sub-network (TFOSN) based on cascaded hybrid Transformer blocks is proposed to enhance time-frequency concentration. Within this framework, the efficient grouped Swin-Transformer (EGST) is designed for effective global and local dependency modeling, allowing progressive refinement of the time-frequency distribution and improved robustness in complex noise environments.

3) A controllable multi-task joint training strategy is designed to coordinate denoising and optimization objectives. Through weighted loss fusion, the framework stabilizes training and achieves a balanced trade-off between noise removal and time-frequency concentration enhancement.

The rest of the paper is organized as follows. Section 2 introduces the mathematical formulation of TFDON. Sections 3 and 4 present numerical and experimental validations, respectively. Section 5 concludes the paper.

2. The TFDON

A novel time-frequency network, termed TFDON, is developed, and its core structure comprises the SDSN and TFOSN. In this section, the mathematical theory and training process of the algorithm are described in detail, and the TFDON-based dynamic characterization framework for rotating machinery nonstationary signals is constructed.

2.1 Theory of the TFDON

The core idea of the TFDON lies in the construction of two sub-networks, namely SDSN and TFOSN, which are respectively employed for denoising the noisy TFR and optimizing its energy concentration. Specifically, the SDSN is first designed, utilizing dynamically configured and alternately arranged multi-scale dilated convolutions of varying sizes to introduce three different receptive fields. Subsequently, the TFOSN is constructed, employing a hierarchical design comprising multiple groups of alternately arranged EGSTs and dual-layer attention mechanisms.

2.1.1 The developed SDSN

The proposed SDSN consists of a feature extraction module (FEM), multiple deep sparse dynamic blocks (DSDB), a pyramid fusion module (PFM), and a reconstruction layer (REM), with the specific architecture illustrated in Fig. 1 [29]. Among these components, the FEM is employed for preliminary extraction of noise information. The main structure consists of eight alternating layers of DSDB for learning the distribution of noise. The joint design of the PFM and the REM leverages parallel convolutional kernels to integrate features with varying receptive fields, capturing both local and global structures, which aids in more comprehensive signal reconstruction post-denoising. Finally, a residual connection is utilized to reconstruct a clean TFR. The detailed theoretical

derivations are presented as follows.

The input TFR is obtained by the STFT, and its calculation expression is:

$$TFR_{in} = STFT(t, f) = \int_{-\infty}^{+\infty} x(\tau) g(\tau - t) e^{-j2\pi f \tau} d\tau \tag{1}$$

where $x(\tau)$ is the input signal, f represents the continuous frequency, $e^{-j2\pi f\tau}$ denotes the STFT basis function, and $g(\tau\text{-}t)$ represents the window function time-shifted by t.

Then, the $TFR_{in} \in \mathbb{R}^{B\times C\times H\times W}$ is input data, where B, C, H, and W are the batch size, channel number, height, and width of the feature, respectively. The TFR_{in} is input into the FEM:

$$TFR_F = \sigma(C * TFR_{in} + b) \tag{2}$$

where $*$ is a two-dimensional convolution, C is a kernel size 3×3 convolution weight matrix, b represents a bias term, $\sigma(\cdot)$ is the ReLU activation function, TFR_F is the output feature, and its channel dimension is expanded to C_{out}.

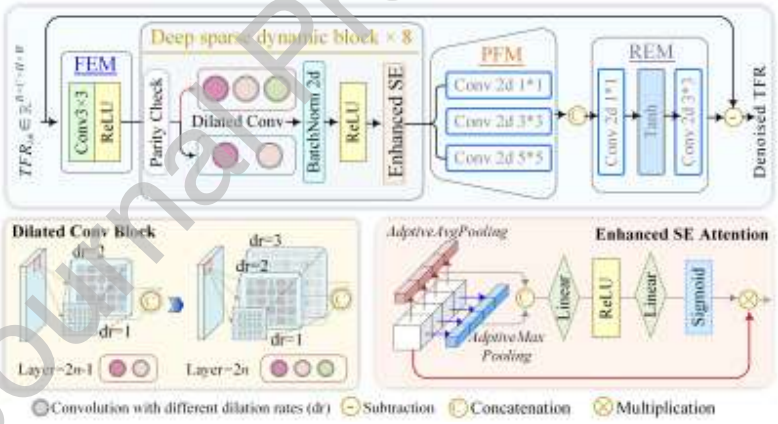


Fig. 1. The structure of the SDSN.

The time-frequency features are then further input into the DSDB, where the multi-scale dilated convolution block with the set of dilation rates $\{d_1, d_2\}$ or $\{d_1, d_2, d_3\}$ can be expressed as:

$$F_k = \sigma(\beta(\sum_{i=1}^{C} W_k^{(i)} *_{d_k} TFR_F^{(i)})), \quad \forall k \in \{1, \ldots, K\} \tag{3}$$

where, $*_{d_k}$ denotes the extended convolution operation with expansion ratio d_k, $\beta(\cdot)$ is a batch normalization operation. In addition, it should be noted that the size of K is alternately designed as 2 and 3, that is, when the

number of layers is odd 2*n*-1, *K*=2, and when the number of layers is even 2*n*, *K*=3. This design introduces different receptive fields to effectively deal with noise in different time and frequency intervals. Further, the features in each path are fused:

$$TFR_k = W_f * [F_1 || \cdots || F_K] \tag{4}$$

where '||' denotes channel splicing, W_f is a convolution kernel of 1×1. The obtained TFR_k is input into EnhanceSE for feature extraction, and its mathematical expression is:

$$\begin{cases} z_{avg} = \dfrac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} TFR_k(:,:,i,j) \\ z_{max} = max_{i,j} TFR_k(:,:,i,j) \end{cases} \tag{5}$$

where z_{avg} captures global context by calculating the mean across spatial dimensions, while z_{max} focuses on the most significant local response using the maximum value. Then, the concatenation $[z_{avg} || z_{max}] \in \mathbb{R}^{2C}$ is further operated to realize collaborative fusion of two statistical features and generate attention weights based on two-way features:

$$s = \delta(W_2 \cdot \sigma(W_1 \cdot [z_{avg} || z_{max}])) \tag{6}$$

This operation consists of dimensionality reduction, nonlinear activation, and dimensionality restoration. The reduction layer $W_1 \in \mathbb{R}^{(C/r) \times 2C}$ compresses the features to reduce parameters, while the expansion layer $W_2 \in \mathbb{R}^{C \times (C/r)}$ restores the channel dimension and generates attention weights. The gated Sigmoid function $\delta(\cdot)$ restricts the weights to (0,1), implementing soft attention. Then, the features are recalibrated, and the adjusted features can be calculated as:

$$TFR_k = TFR_k \odot s \tag{7}$$

where '$\odot$' denotes a channel-by-channel multiplication operation to spatially scale the feature map. The EnhancedSE attention improves the feature selection ability through the processing of feature extraction and dynamic weighting.

Then input $\widetilde{TFR_k}$ to the PFM, where the process of multi-kernel convolution operation can be expressed as:

$$\begin{cases} F_1 = W_1 * TFR_k \\ F_3 = W_3 *_{p=1} TFR_k \\ F_5 = W_5 *_{p=2} TFR_k \end{cases} \tag{8}$$

where $*_p$ denotes a convolution with fill size p. Then, the reconstruction of features is implemented in the REM:

$$F_R = W_3 * (\mathrm{T}(W_1 * [F_1 || F_3 || F_5])) \tag{9}$$

where T represents the Tanh activation function. Finally, residual operation is performed on the characteristic map obtained by the main path and the input TFR to calculate:

$$TFR_{deno} = TFR_{in} - F_R \tag{10}$$

where the TFR_{deno} represents the clean TFR after denoising.

2.1.2 The proposed TFOSN

The TFSON, composed of input convolution layers, the three-stage cascaded HTB, and a residual connection, is designed to optimize the time-frequency concentration. The network performs feature mapping through convolutional layers, leverages the spatiotemporal feature transformation capabilities of the EGST, and employs a dual attention mechanism to iteratively optimize the TFR. The above sections are calculated as:

$$F_{TFS}(TFR_{deno}) = C_{in}(TFR_{deno}) + \sum_{k=1}^{3} HTB_k(C_{in}(TFR_{deno})) \tag{11}$$

where C_{in} is the input convolutional layer, the kernel size is 3×3, HTB_k is the k-th HTB, where a single HTB consists of two EGST, an MDCA, an ESA, and a convolutional layer designed at intervals, and the calculation process can be described by the mathematical expression as:

$$HTB(X) = C_{3\times3} \cdot ESA \cdot EGST_2 \cdot MDCA \cdot EGST_1(X) \tag{12}$$

where '·' denotes the compound operation of the function. $EGST_i$ denotes the EGST of the i-th layer.

A. The developed EGST

The Swin-Transformer block demonstrates strong performance in computer vision by exploiting self-attention

to model similarities among image patches, but it incurs considerable GPU memory overhead [30,31]. To solve this problem, the EGST model with high efficiency is constructed. The EGST consists of two continuous blocks. Different from Swin-Transformer, window-based multi-point self-attention (W-MSA) and locally shifted window-based multi-point self-attention (SW-MSA) are improved to efficient W-MSA (EW-MSA) and efficient SW-MSA (ESW-MSA), respectively. Then, the LayerNorm (LN) layer is applied before the MSA, and the multilayer perceptron (MLP)module, and a residual connection is carried out in the two parts. Its structure is shown in Fig. 2.

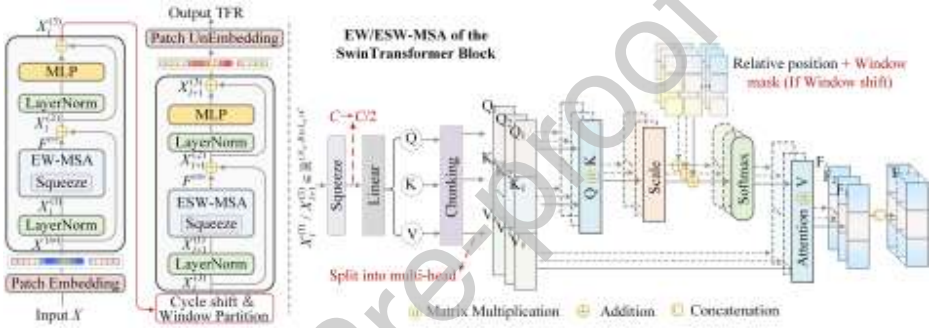


Fig. 2. Architecture of the developed EGST.

The input TFR of the module is denoted as *X*, and its processing is divided into four steps. Step 1: The patch segmentation module divides the input TFR into equal-sized, non-overlapping patches, and each patch is regarded as a “token”, as shown in the left panel of Fig. 3. Each patch corresponds to a local time-frequency region, preserving localized energy distribution characteristics. Step 2: The patches are grouped into self-attention windows. In EW-MSA, the patches are divided into four equal-sized non-overlapping windows, while the window division mode of ESW-MSA is shown in the right panel of Fig. 3. To enhance cross-window interaction while maintaining computational efficiency, a shifted window strategy is adopted, as illustrated in Fig. 4. Specifically, a left shift with right filling is first performed, followed by an upward shift with lower filling. This operation enables information exchange across adjacent time-frequency regions. Step 3: The window-based self-attention mechanism is applied

within each window, allowing adaptive weighting of time-frequency components inside a local region. Step 4: A reverse shift operation is performed to restore the original spatial arrangement, thereby maintaining structural consistency of the time-frequency representation. The calculation of the above process can be described as follows:

$$ST(X) = PU \cdot (\prod_{l=1}^{2} ESW_l) \cdot PE(X) \tag{13}$$

where PE is patch embedding, PU is patch unembedding, and ESW_l represents a single ESW module in a contiguous block. The sequential computation process using an alternating EW-MSA and ESW-MSA structure is as follows, where the first operation is:

$$X_l^{(1)} = \mathrm{LN}(X^{(in)}) \tag{14}$$

where LN represents the layer normalization operation, and $X_1^{(1)} \subset \mathbb{R}^{(N_w \cdot B) \times L_w \times C}$, where N_w and L_w denote the number and length of windows, respectively. Then, the input dimension is compressed to half by the linear transformation, i.e., the channel dimension C is changed to $C/2$:

$$X_l^{reduce} = X_l^{(1)} W_{reduce} \tag{15}$$

$$Q, K, V = \mathrm{LinearSplit}(X_l^{reduce}) \tag{16}$$

Eq. (16) shows that the feature map is projected into three elements: query (Q), key (K), and value (V) by using a linear layer. To reduce the calculation cost, Q, K, and V are divided into i sequences of equal size along the sequence dimension. Attention is calculated for each sub-block Q_i, K_j, and V_j:

$$F_i = \mathrm{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}} + B[R]\right) V_i \tag{17}$$

where $B[R]$ represents the offset value obtained by looking up the index matrix R. It should be noted that when the shift window operation is performed, the pre-stored mask is used for fusion, and the attention score is corrected as:

$$F^{ew} = F_i + Mask \otimes 1_H \tag{18}$$

where '$\otimes$' represents copying the mask to each attention head by window, 1_H represents the all-one matrix, and H is

the number of heads. Then, the residual sum is added with the original input:

$$X_l^{(2)} = F + X^{(in)} \tag{19}$$

Finally, after the LM and MLP operation processing, the residual connection is performed with $X^{(2)}$:

$$X_l^{(3)} = X_l^{(2)} + \mathrm{MLP}(\mathrm{LN}(X_l^{(2)})) \tag{20}$$

Features are then input into the l+1 layer of the EGST. Wherein, the ESW-MSA is similar to the EW-MSA, and the former performs window-based multi-point self-attention with a shifted window partition operation, followed by window restoration.

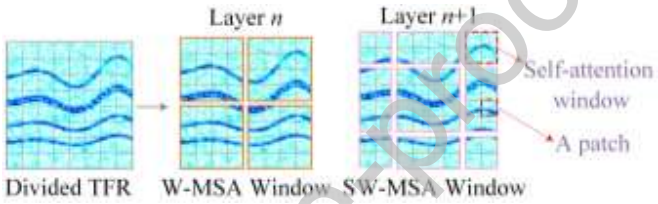


Fig. 3. The division of the self-attention window.

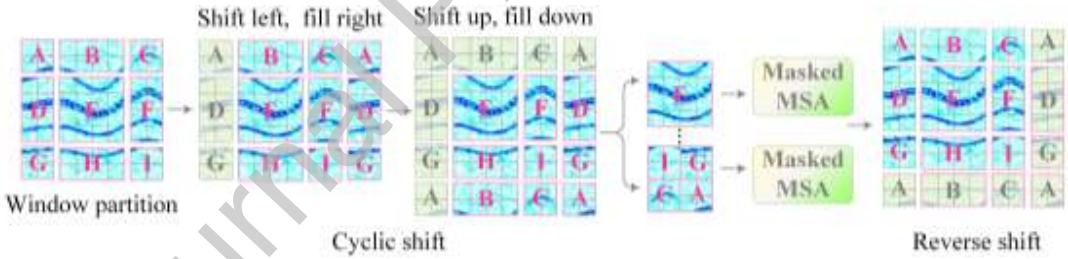


Fig. 4. The window partition and cycle process of the EW-MSA.

B. The MDCA and ESA

The proposed MDCA achieves efficient time-frequency feature optimization through lightweight architecture design and a dynamic feature enhancement mechanism, and its structure adopts the design of deep separable convolution and multi-scale feature fusion. The structure details are shown in Fig. 5.

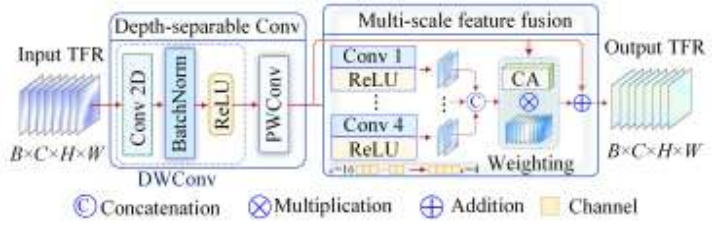


Fig. 5. The MDCA module.

In the model, depth-separable convolutions, including DwConv and PwConv, are first designed, which are composed of deep and dot convolutions to extract local features. In the time-frequency optimization task, deep convolution can capture local patterns in the TFRs, while dot convolution integrates channel information:

$$Y_{dw} = \sigma(\mathrm{DWConv}_{3\times3}(X)) \tag{21}$$

$$Y_{pw} = \mathrm{Conv}_{1\times1}(Y_{dw}) \tag{22}$$

Then, the four parallel 1×1 convolutions are used for multi-scale feature fusion, reducing the number of channels to 1/4, and then concatenating. This process helps capture features at different time or frequency ranges, such as high-frequency details and low-frequency contours. The process can be expressed as:

$$\{M_s\} = \{\mathrm{Conv}^{s}_{1\times1}(Y_{pw})\}_{s=1}^{4} \tag{23}$$

where s denotes parallel convolution branches. A channel attention (CA) mechanism is embedded, which generates channel weights through global average pooling and two fully connected layers for enhancing the response of important channels and suppressing irrelevant information and noise channels. The calculation process is as follows:

$$\alpha = \delta(\mathrm{Conv}(\mathrm{Avgpool}(\oplus M_s))) \tag{24}$$

Finally, residual connections are used to prevent information loss and facilitate gradient flow, while preventing the model from losing original information when enhancing key features:

$$\mathrm{MDCA}(X) = X \odot \alpha + X \tag{25}$$

In addition, an enhanced spatial attention (ESA) mechanism [32] is introduced into the main branch of the HTB

to capture significant energy regions through maximum pooling. The calculation process is as follows:

$$X_1^{ESA} = \mathrm{Conv}_{1\times1}(X) \tag{26}$$

In the above equation, channel compression is performed by a 1×1 convolution. Spatial dimensionality is then reduced by a step convolution layer and a maximum pooling layer, and then recovered by an upsampling layer:

$$X_2^{ESA} = H_i(\mathrm{Conv}_{3\times3}(\sigma(\mathrm{MaxPool}(\mathrm{Conv}_{3\times3}(X_1^{ESA}))))) \tag{27}$$

where H_i is bilinear interpolation to restore the resolution. Finally, the attention-weighted output result is obtained:

$$X_{out}^{ESA} = X \odot \delta(\mathrm{Conv}_{1\times1}(X_2^{ESA} + X_1^{ESA})) \tag{28}$$

Hence, the HTB uses global Transformer and local attention alternately, taking into account long-range dependence and detail enhancement. In the HTB, the EGST is used to extract time-frequency global features, the first MDCA performs channel-dimensional dynamic calibration of the TFR, the second EGST is used to refine local features, and finally ESA is introduced to enhance spatial saliency.

2.2 The training details

The NVIDIA GeForce RTX 3080 parallel computing architecture is used to accelerate the training of the TFDON [33]. The initial learning rate is set to 1e-3, and the adaptive adjustment of parameter update is realized based on the phased reduction strategy. The batch size is set to 8. Fig. 6 shows the complete training framework, which includes two steps: first, build training and test datasets; second, the joint loss function of the denoising task and the time-frequency enhancement task is designed.

2.2.1 Dataset construction

The constructed datasets include noise-containing and noise-free signals calculated from the STFT, ideal and penalty TFR data. Each type of sample contains 8.8×10^3 sets of signals. The samples are divided into training and test sets according to a sample ratio of 10:1. The training set and the test set are independent of each other [21].

The construction of the dataset is determined by a sinusoidal FM signal and a cubic signal with several randomly assigned variables, mathematically expressed as:

$$s(t)=\sum_{l=5}^{L}\left(A_l \sin(j2\pi(\int_0^t f_1(\tau))+A_l \sin(j2\pi(\int_0^t f_2(\tau)\right)+h_l(t) \tag{29}$$

where, L takes the value of 9, indicating that the number of signal components is selected between [5, 9]; A_l indicates the amplitude intensity of each component signal, and the follows distribution rule A_l=0.01+5|u_n|; $h_l(t)$ is the randomly added Gaussian white noise with noise intensity [-10,10] dB; where the expressions for IF curves $f_1(\tau)$ and $f_2(\tau)$ are:

$$\begin{cases} f_1(\tau)=2\pi/a_l*(b_l\tau^2+c_l\tau)*sin(2\pi\tau/a_l)-(2b_l\tau+c_l)*\cos(2\pi\tau/a_l)+d_l \\ f_2(\tau)=a_l\tau^3-b_l\tau^2+c_l\tau+d_l \end{cases} \tag{30}$$

The sampling frequency of the signal is f_s=2048Hz, and the sampling time is 2.0 s; the values of each parameter in Eq. (30) belong to the range shown in Table 1, where the U indicates that the variable is uniformly distributed.

Table 1. Parameters ranges

Types	a_l	b_l	c_l	d_l
$f_1(\tau)$	U(0.20, 2)	U(0.15, 1.8)	U(0.25, 1.5)	U(25, 460)
$f_2(\tau)$	U(25, 70)	U(20, 150)	U(5, 75)	U(50, 450)

According to Eq. (29) and (30), the STFT TFR dataset D_1 without noise and the dataset D_2 containing noise are first generated, respectively. The two datasets are used as training input signals of the TFDON and learning targets of the time-frequency denoising task, respectively. Then, the ideal TFR dataset D_3 and the penalty matrix dataset D_4 are generated according to Eq. (29) and (30). The generation of the ideal TFR requires preparing an all-zero matrix first, and filling random variables A_n conforming to the following rules into the matrix, and the physical meaning of the filling positions in the matrix is the IF trend of the signal, which is determined by the following equation:

$$\mathrm{D}_3^n(x,y)=\overline{A}_n \quad \text{if}(\,\mathrm{mod}(\mathrm{IF}_n(y),\,f_s)<(x+1)\Delta f\,)\cap(\,\mathrm{mod}(\mathrm{IF}_n(y),\,f_s)\geq x\Delta f\,) \tag{31}$$

where x and y represent frequency and time indices, respectively; the matrix size is $2N_f\times T$, and the value in this paper

is 256*256; mod represents modulo operation; $\Delta f = f_s/(2N_f)$ denotes frequency interval.

The penalty matrix dataset D_4 differs from D_3 in that it first generates an all-ones matrix with a size of $2N_f \times T$, and then reassigns the value A_n^p in position of the IF curves to enhance the degree of attention to the region to be characterized, which is calculated as follows:

$$\mathrm{D}_4^n(x,y) = A_n^p \quad \text{if } \mathrm{mod}(\mathrm{IF}_n(y), f_s) < (x+1)(f_s/(2N_f)) \cap \mathrm{mod}(\mathrm{IF}_n(y), f_s) \geq x(f_s/(2N_f)) \tag{32}$$

where A_n^p is the assignment in the penalty matrix, calculated as:

$$A_n^p = (log(\max_i(A_n)))^2 \big/ 20\log 10(A_n + 1) \tag{33}$$

The penalty weighting matrix can be used as supervision information to guide the network in characterizing IF ridges more accurately during training.

2.2.2 Design of loss function

The TFDON training involves two parts: one is the denoising task of the input TFR, and the other is the optimized task of time-frequency energy. Hence, the loss function of the joint training framework is designed, which also consists of two parts: time-frequency denoising and TFR optimization.

For the training task of the first stage of the SDSN, the MAE is used to measure the denoising effect:

$$L_{denoise} = \frac{1}{N}\sum_{i=1}^{N}\left|\widehat{D}_2^i - D_2^i\right| \tag{34}$$

where $\widehat{D}_2^i$ is the denoising matrix of the model output, D_2^i is the true noise-free signal, and N is the number of samples. For the time-frequency optimization task of the second stage TFOSN model, the MSE is used to optimize the TFR:

$$L_{opt} = \frac{1}{N}\sum_{i=1}^{N}\left((\widehat{D}_3^i - D_3^i) \odot D_4^i\right) \tag{35}$$

where $\widehat{D}_3^i$ is the TFR predicted by the TFOSN, D_3^i represents the ideal TFR label, and D_4^i is the penalty matrix label. The errors are regionally weighted by element-by-element multiplication, especially by giving higher weights to energy concentration regions and reducing the weights of noise-dominant regions. Force the model to fit key regions first, so that the model can pay differential attention to different regions of the time-frequency plane when

training.

Finally, the two parts of the loss of the joint training task are fused by weighting:

$$L_{total} = (\lambda \cdot L_{denoise} + (1-\lambda)L_{opt})s \tag{36}$$

where λ is an adjustable weighting factor that can be adjusted according to the different emphasis of the task; λ=0.5 is set in this paper; in addition, since the loss values MAE and MSE are of magnitude 10^{-3}, a scaling factor s is used to scale the final total loss value to obtain L_{total}. After scaling, the gradient is more significant to avoid gradient disappearance. The design of the joint loss function takes into account the cooperative joint training of multiple tasks and realizes the priority trade-off between denoising and optimization tasks by controllable weight allocation.

Regarding training, the TFDON adopts a segmented strategy rather than an end-to-end framework, motivated by the distinct objectives of denoising and optimization [34]. Specifically, first, the segmented training strategy separates the denoising and optimization objectives to avoid mutual interference. Since denoising targets noise suppression and frequency preservation, and the optimization stage emphasizes energy enhancement, sequential training allows each stage to converge under its own dominant objective. Second, the segmented strategy improves training stability under noisy and nonstationary conditions. By stabilizing the denoising module before feature optimization, it prevents unstable gradient propagation caused by noise-sensitive instantaneous frequency structures.

2.3 The TFDON-based dynamic characterization framework

Based on the TFA method of the TFDON, the novel dynamic characterization technology for rotating mechanical equipment is constructed. As depicted in Fig. 6, the methodology encompasses three stages:

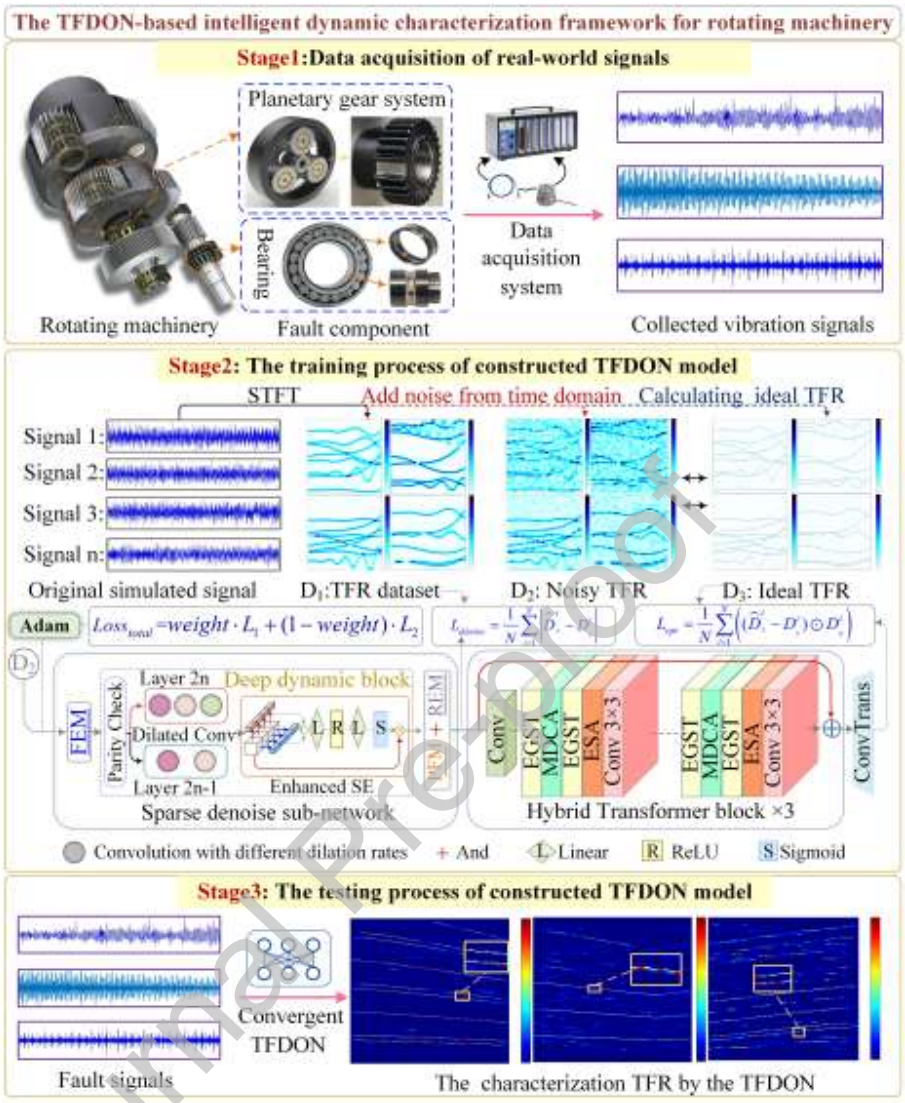


Fig. 6. The TFDON-based intelligent dynamic characterization framework for rotating machinery.

Stage 1, data acquisition of real signals: experiment vibration signals are collected from bearings and planetary gearboxes of mechanical equipment operating at variable rotational speeds using a signal acquisition system, such as wind turbines, steam turbines, aerospace equipment, etc..

Stage 2, training process of constructed TFDON model: first, the time-frequency network TFDON is designed, and the training parameters are determined; second, the various datasets for training and testing are constructed; finally, the training datasets are input into the developed TFDON for periodic repeated training.

Stage 3, the testing process of constructed TFDON model: the model is first validated using simulated noisy signals, and then applied to experimental data to generate TFRs, where theoretical fault characteristic frequency (FCF) curves are matched with time–frequency ridges for nonstationary condition monitoring.

3. Numerical experiment

In this section, a five-component numerical signal with closely-spaced IFs is defined to evaluate the effectiveness of the TFDON under strong noise environments. Then, the TFDON is compared with traditional and advanced TFA methods. In addition, each TFA technique is compared under various noise intensity using two quantization indices.

The simulation signal is defined as follows:

$$s_{multi}(t) = \sum_{n=1}^{5} \sin(2\pi \int_0^t f_n(\tau) d\tau) + k(t) \tag{37}$$

where $k(t)$ is the Gaussian white noise. To simulate the noise interference, the noise with an SNR of 0 dB is added to the signal. The IF curves $f_1(\tau)$ and $f_5(\tau)$ are proportional to the fundamental frequency $f(\tau)$, with multiplication factors of 1.5, 2.3, 3.33, 3.5, and 4.7, respectively. The expression of fundamental frequency $f(\tau)$ is:

$$f(\tau) = (20\tau + 100)cos(2\pi\tau / 1.25) - 2\pi / 1.25(10\tau^2 + 100\tau)sin(2\pi\tau / 1.25) - 80 \tag{38}$$

The sampling frequency of the signal is set to 1024 Hz. The waveform and IF ridges are shown in Fig. 7(a) and (b), respectively. The TFDON result is shown in Fig. 8(a), and it can remove most of the background noise. In addition, the five IF curves are completely characterized, and the two frequency curves f_3 and f_4 with closely-spaced distribution can also distinguish the accurate position, almost free from noise interference.

Fig. 8(b)-(f) shows the results of the HDSCT, SSET, SST, TFA-Net, and VSLCT, respectively. In the result of the HDSCT, the frequency curve f_2 is blurred, and the closely-spaced curves f_3 and f_4 are not distinguished. Meanwhile, the time-frequency characterization results of the SSET and SST face the same problem, that is, the five

IF ridges are more seriously overlapped with the surrounding noise signals. The TFA-Net struggles to smoothly characterize closely-spaced frequencies, and the background noise is prominent. The VSLCT result can characterize the IF distribution of each curve, but it shows low energy concentration and is easy to identify random background noise as the real IF curve.

In addition, the Rényi entropy index is introduced to quantitatively evaluate the performance of each method under low SNR. The Rényi entropy can be defined as [4]:

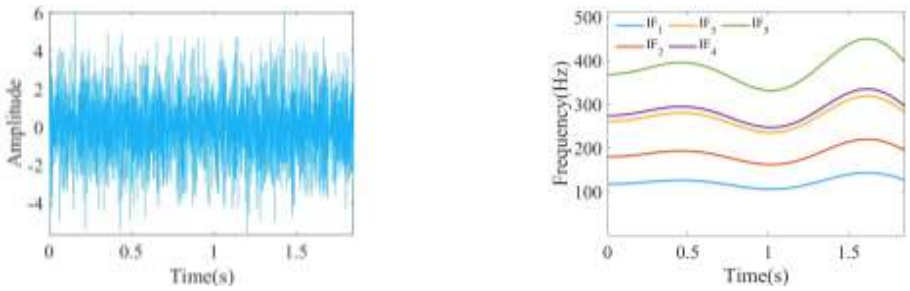
$$R^{\alpha} = \frac{1}{1-\alpha}\log_2 \iint \left(\frac{TFR(t,\eta)}{\iint TFR^3(t,\eta)\,d\eta dt}\right)^{\alpha} d\eta dt \tag{39}$$

where α in Eq. (39) is 3, and R is inversely proportional to time-frequency concentration, i.e., the smaller the R, the better the concentration effect. To ensure statistical reliability, each method is evaluated with 10 independent trials at every 2 dB SNR step from -6 dB to 14 dB, and the mean Rényi entropy along with tenfold standard deviations are presented in Fig. 9. It is evident from the figure that the TFDON achieves the minimum Rényi entropy across the entire SNR range, accompanied by relatively low standard deviations, indicating both superior concentration and stability.

To further quantitatively evaluate different methods, the Earth mover's distance (EMD) is used as a metric to measure the discrepancy between the normalized TFR and ideal TFR, and it can be defined as [21]:

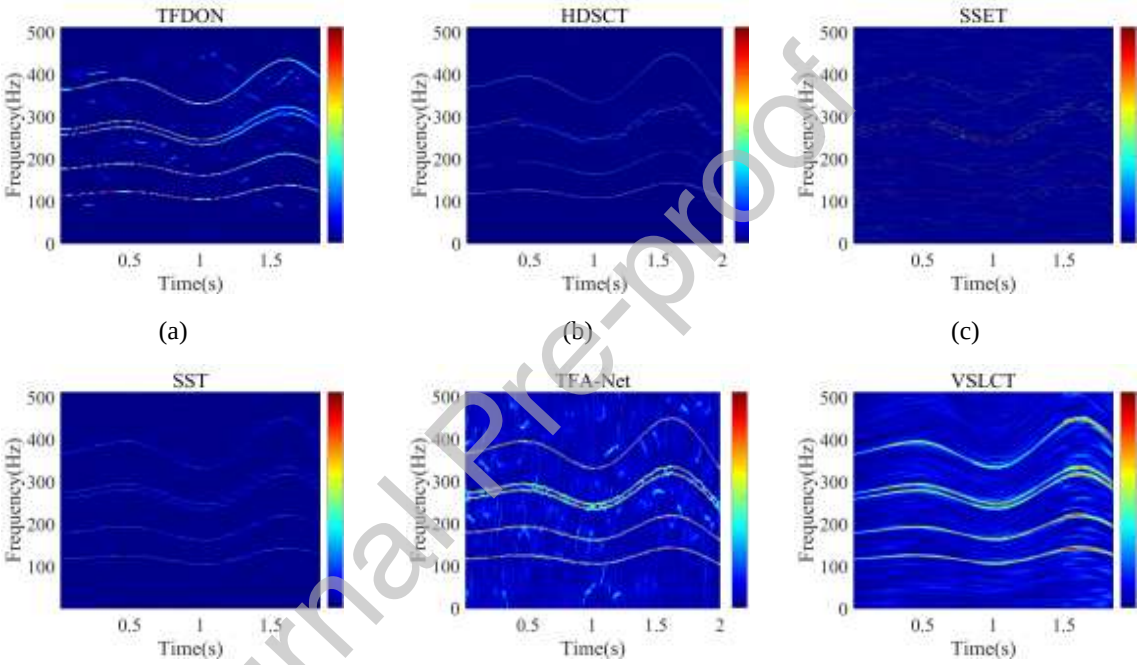
$$E = \iint_{s} |\,\mathrm{TFR}_i - \mathrm{ITFR}_i\,|\,ds \tag{40}$$

where a smaller E value indicates a higher similarity to the ideal TFR. Each method is tested 10 times under four noise levels, and the mean EMD, with its standard deviations magnified tenfold for clarity, is presented in Fig. 9. From the figure, the TFDON achieves the minimum EMD with low standard deviations under all noise conditions, indicating both stable computational accuracy and noise robustness.



(a) (b)

Fig. 7. Simulated signal: (a) waveform; and (b) IF ridges.



(a) (b) (c)

(d) (e) (f)

Fig. 8. TFRs of numerical signal from (a) TFDON; (b) HDSCT; (c) SSET; (d) SST; (e) TFA-Net; and (f) VSLCT.

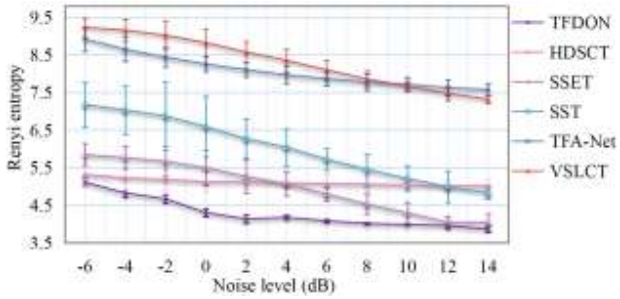


Fig. 9. Results of Rényi entropy by six TFA methods with different SNR values.

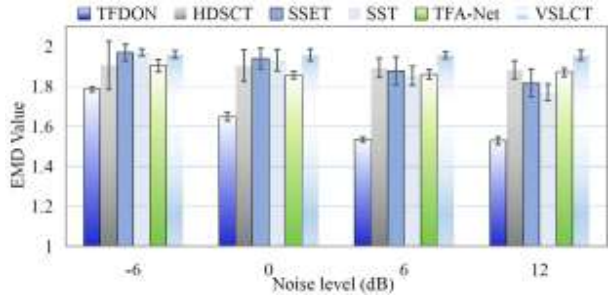


Fig. 10. EMD values by six TFA methods with different SNR values.

4. Real signals experiment

In this section, the comprehensive performance of the TFDON is further analyzed using two case studies of bearing and one set of planetary gearbox signals under strong noise environments. Meanwhile, the advantages of the TFDON are empirically revealed by comparison with various TFA techniques, combined with the intuitive visual performance of the TFR and the calculation results of the Rényi entropy index.

4.1 Bearing signal with single fault

Bearing is the core component used to support rotating parts in mechanical systems, and it is widely applied in wind turbines, steam turbines, aerospace equipment, etc. Hence, it is of great engineering value to carry out bearing fault diagnosis research for such advanced industrial equipment [35,36].

In this subsection, the Ottawa bearing signal with a single inner ring fault is used to evaluate the performance of the TFDON [37]. The test experimental device is shown in Fig. 11, which shows the specific installation positions of the AC driver, motor, encoder, accelerometer, and two bearings. Among them, the AC driver is used to control the shaft speed; the incremental encoder and accelerometer are mounted to measure the speed and vibration data of the shaft, respectively; two ball bearings support the shaft, and the right and left bearings are faulty and healthy, respectively.

Using the bearing vibration dataset I-D-2 with inner ring fault, the sampling frequency and time are 2×10^5 Hz

and 2.5 s, respectively, and its operating rotational frequency (RF) decreases from 25.3 Hz to 14.8 Hz and then increases to 19.4 Hz. Fig. 12(a) and (b) show the bearing's measured vibration signal waveform and RF ridge, respectively.

The selected ball bearing model is ER16K. The bearing diameter is 38.52mm, the ball diameter is 7.94mm, and the number of balls is 9. According to the bearing parameters, the fault characteristic coefficient of the inner ring of the bearing is calculated as 5.43, so the calculation method of the FCF of the inner ring is $f_{\text{inner}} = 5.43\ fr$.

The collected signal is first subjected to the initial down-sampling process and normalized the signal amplitude to the [-1,1] interval. Then, the 5 dB SNR noise is introduced in the fault signal. The TFR obtained by the TFDON is shown in Fig. 13(a). In the TFR, four IF ridges are detected, including the inner ring FCF and its 2^{nd}, $3^{rd,}$ and 5^{th} harmonics. It can be seen from the zoom that the frequency ridges are relatively complete and of high concentration. Therefore, the proposed TFDON can accurately determine the fault in the bearing's inner ring.

The TFRs calculated from the STFT, SSET, SST, GLCT, and VSLCT are shown in Fig. 13(b)-(f). In the STFT result, frequency ridges show the energy redundancy phenomenon. From the TFRs of the SSET and SST, the energy of the FCF curve is weak, and it can be observed in the zoom that it is seriously overlapped with noise. A large number of cross terms are generated in the TFR of the GLCT, which is not readable well. In addition, the result of the VSLCT is easy to identify background noise as weak FCFs, and some unrelated frequency curves are produced.

The Rényi entropy is used to further evaluate the performance of the TFDON, and the results are calculated every 2 dB under noise intensities from -5 dB to 15 dB. Meanwhile, the TFDON is compared with the above methods. The test results are shown in Fig. 14. From the results, the TFDON obtains minimum values at different SNRs, indicating that the TFDON has better time-frequency concentration performance than the other five comparison methods.

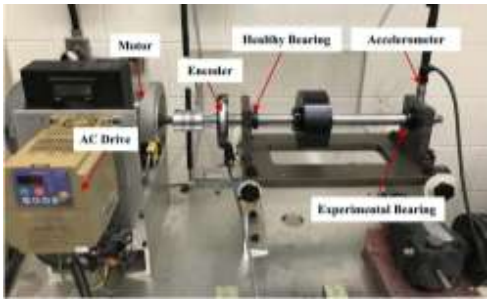
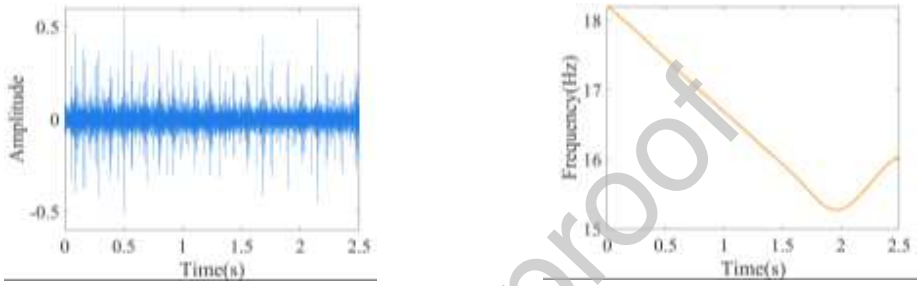
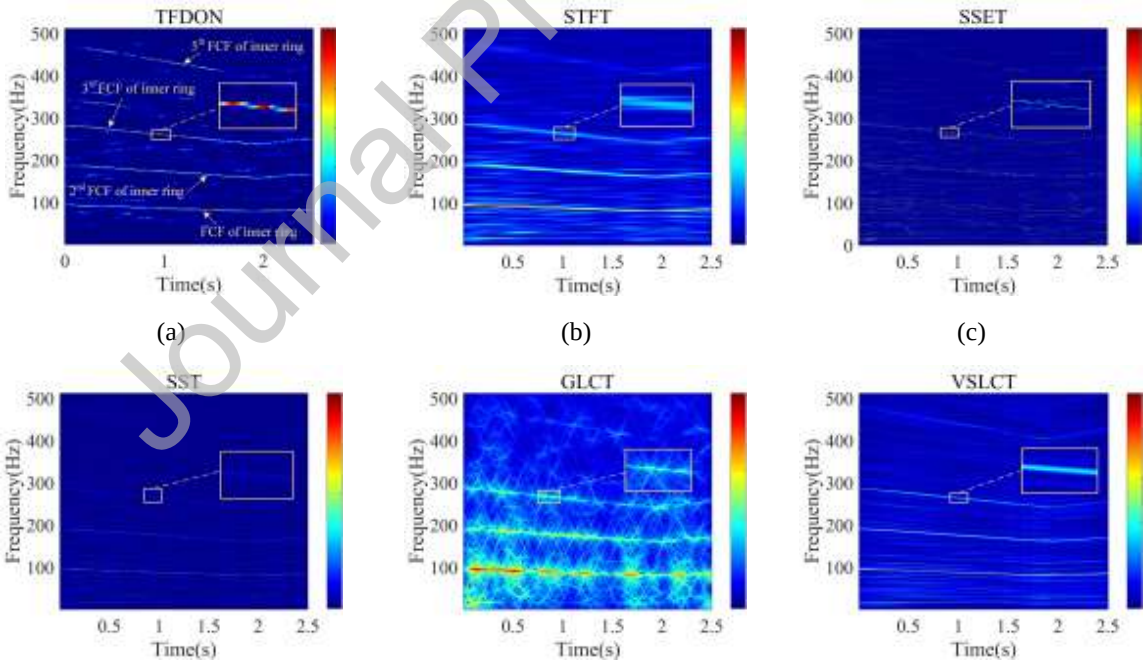


Fig. 11. Ottawa bearing experimental bench.



(a) (b)

Fig. 12. Ottawa signal: (a) waveform; and (b) RF.



(a) (b) (c)

(d) (e) (f)

Fig. 13. TFRs generated from (a) TFDON; (b) STFT; (c) SSET; (d) SST; (e) GLCT; and (f) VSLCT.

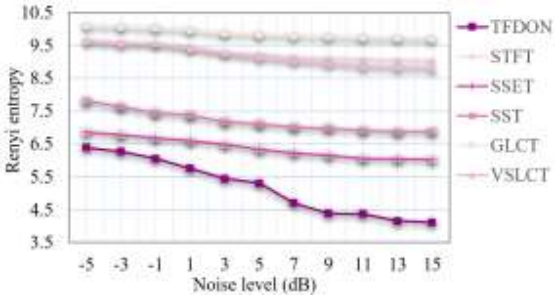


Fig. 14. Rényi entropy on bearing signal with single fault *vs*. SNR levels.

4.2 Bearing signal with multi-fault

The subsection further verifies the effectiveness of the TFDON in identifying multi-component faults. The multi-fault bearing signal is collected from the test bench as shown in Fig. 15 (a). The figure shows the main components, such as the motor, fault bearing, sensor, speed controller, and so on. The selected bearing model is 6000, and Fig. 15(b)-(c) present the cracks of outer and inner ring faults, and their width and depth are 0.2 and 0.4 mm, respectively.

As shown in Fig. 16 (a)-(b), bearing vibration signal and its frequency curves are collected by the acceleration sensor and encoder, respectively. The sampling rate is 2.4×10^4 Hz. The mathematical expression of the RF curve is obtained by quadratic function fitting: $f_r(t)$ = -1.02t^2-5.34t+52.84. According to the bearing model, the fault characteristic coefficients of the inner ring, outer ring, and ball can be calculated as 4.4, 2.5, and 1.7, respectively. Then, the FCF curve expression of the inner ring and outer ring of the bearing can be expressed as follows:

$$\begin{cases} f_{inner}(t) = 4.4 *_r (f_r(t)) = -4.48 * t^2 - 23.49 * t + 232.49 \\ f_{outer}(t) = 2.5 *_r (f_r(t)) = -2.55 * t^2 - 13.35 * t + 132.09 \end{cases} \tag{41}$$

The bearing signal is processed by the spectral kurtosis-based bandpass filter algorithm and Hilbert transform, and 5 dB Gaussian white noise is added after normalization. The TFR shown in Fig. 17(a) is obtained by the TFDON. From the figure that 9 frequencies are characterized, including three groups of closely-spaced IFs, and the added Gaussian white noise can be removed. The theoretical FCF curve calculated by Eq. (41) is matched with each curve

in the TFR. The main FCF curves, such as the second RF harmonic, FCF of outer ring, FCF of inner ring, and its second harmonic, are recognized, so that the fault characteristic type of bearing can be accurately diagnosed.

The TFRs calculated by the comparison methods STFT, SSET, SST, GLCT, and VSLCT are presented in Fig. 17(b)-(f). Wherein, the TFRs obtained from the STFT, GLCT, and VSLCT have low time-frequency concentration, and strong background noise interference, especially the curves at high frequency are seriously confused with noise. In addition, the overall readability of the time-frequency results of the SSET and SST is poor, and the energy of each frequency curve is weak, which makes it difficult to accurately identify important information.

Fig. 18 shows the Rényi entropy results of the TFDON and other methods under SNR from -5 dB to 15 dB. The TFDON always obtains the minimum value, specifically, under -5 dB extreme noise, the Rényi entropy of the TFDON is reduced by about 0.73 compared with the SSET. This shows that the proposed method achieves the optimal time-frequency concentration effect.

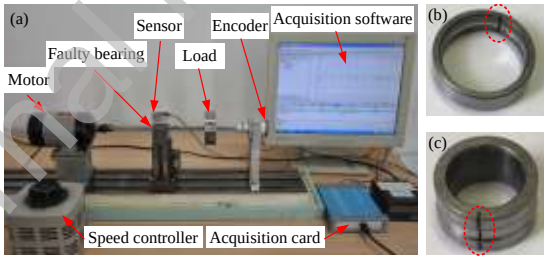
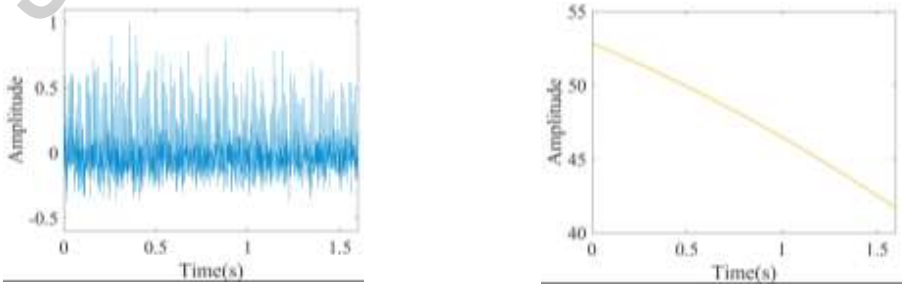


Fig. 15. (a) Bearing fault test rig layout; (b) outer fault ring; and (c) inner fault ring.



(a) (b)

Fig. 16. Bearing signal with multi-fault: (a) waveform; and (b) RF.

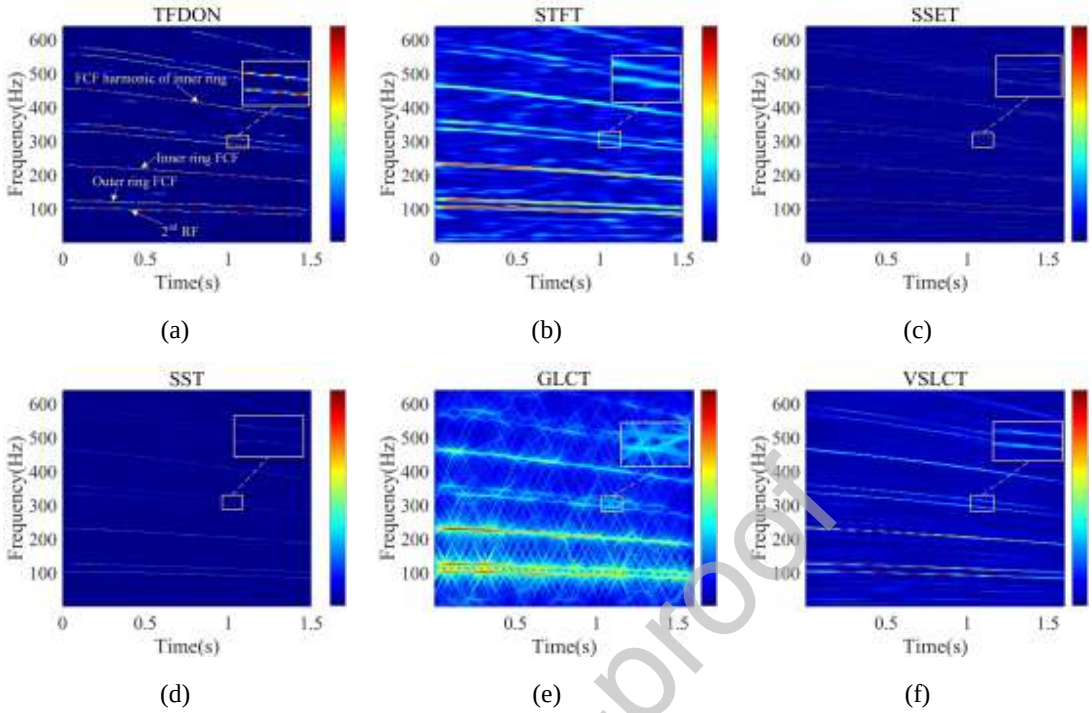


Fig. 17. TFRs generated from (a) TFDON; (b) STFT; (c) SSET; (d) SST; (e) GLCT; and (f) VSLCT.

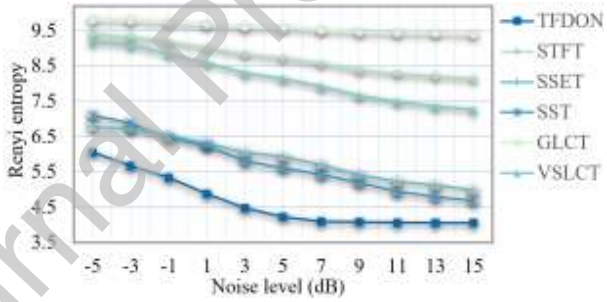


Fig. 18. Rényi entropy on bearing signal with multi-fault *vs*. SNR levels.

4.3 Planetary gearbox fault signal

As a core component of modern transmission systems, the planetary gearbox plays a vital role in rotating machinery, yet its structural and modulation complexity make fault detection challenging. Therefore, effective TFA technology can identify the fault pattern and improve the equipment reliability [38,39].

In this subsection, the effectiveness of the TFDON is verified by the vibration signal measured from a planetary gearbox. Fig. 19(a) and (b)-(e) show the partial details of the test rig, respectively. The planetary gearbox consists of

a sun gear, planet gear, and ring gear; the number of teeth of the three kinds of gear is 28, 36, 100, respectively; the gear transmission ratio is 8, and the number of planet gears is 4.

The sampling rate is 2.4×10^4 Hz, and the sampling time is set as 1.5 s. The sensor and signal acquisition card are used to record the vibration signal waveform and frequency curve, and draw them in Fig. 20(a)-(b), respectively. According to measured speed signal, fit frequency curve: $f_d(t)$= 1.2275t+42.499. The RF of the sun gear is consistent with the RF of the driving motor, i.e., the FCF and meshing frequency of the planetary gearbox can be calculated according to the RF information and the above-mentioned gearbox parameters, wherein the sun gear FCF is determined:

$$f_s(t)=\frac{f_m(t)}{Z_s}N_p=\frac{Z_rN_p}{Z_s+Z_r}f_s^r(t) \tag{42}$$

wherein, $f_m(t)$ is the meshing frequency, $f_s^r(t)$ is the sun gear RF, and Z_s, Z_r, and N_p represent the number of teeth of the sun gear, the number of teeth of the planet gear, and the number of planet gears, respectively. Then the characteristic meshing frequencies and sun gear FCFs in the gearbox can be determined as:

$$\begin{cases} f_m(t)=f_d(t)/8*175 \\ f_s(t)=f_d(t)/8*25 \end{cases} \tag{43}$$

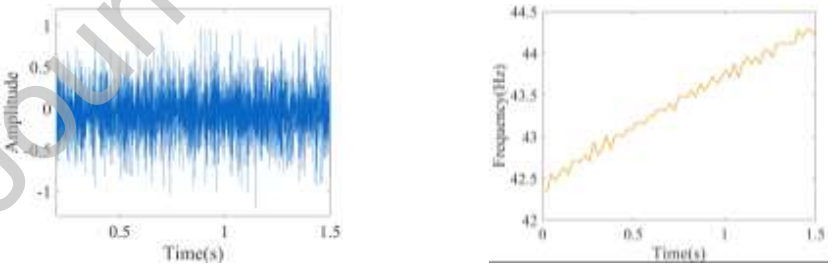
The vibration signal with the sampling frequency of 2440 Hz is obtained after the original signal processing by the down-sampling algorithm, and the signal is normalized and added with 5 dB Gaussian white noise. Fig. 21 presents the [900, 1220] Hz segment of the processing results of each method. The TFR obtained by the TFDON processing is shown in Fig. 21(a). In the TFR, 8 frequency trajectories with physical significance can be identified, among which 6 are combined ridges related to the sun ring fault, and the closely-spaced characteristic components are characterized, such as [f_m] and [f_m-0.25f_s+f_d]. Moreover, it can be observed from the zoom that TFR has high energy concentration. The other frequencies are identified as: [f_m+f_d], [f_m+f_s-1.5f_d], [f_m+f_s-f_d], [f_m+2f_s-2f_d], [f_m+2f_s-1.5f_d], and [f_m+2f_s-f_d]. Hence, the TFDON is able to successfully detect the sun gear fault in the planetary gearbox signal.

The TFRs obtained by the STFT, SSET, SST, GLCT, and VSLCT are shown in Fig. 21(b)-(f). The results show

that the calculation results of the STFT, SSET and GLCT are difficult to identify closely-spaced FCFs, and most time-frequency ridges are seriously submerged by background noise, so it is hard to diagnose bearing faults. In addition, although the SST and VSLCT can identify closely-spaced curves, they show poor energy concentration.

Finally, the Rényi entropy is used to analyse the performance of the TFDON under different SNR intensities from -5 to 15 dB. The test results are shown in Fig. 22. The TFDON has obvious performance improvement compared with other methods, and the minimum value is obtained in all SNR ranges. Specifically, the Rényi entropy obtained at the lowest SNR of -5 dB is still greater than the value of the sub-optimal method SSET in the whole SNR range. Therefore, the results further confirm that the TFDON has better energy concentration performance.

Fig. 19. Planetary gearbox test rig and its partial details.



(a) (b)

Fig. 20. Planetary gearbox signal: (a) waveform; and (b) motor RF.

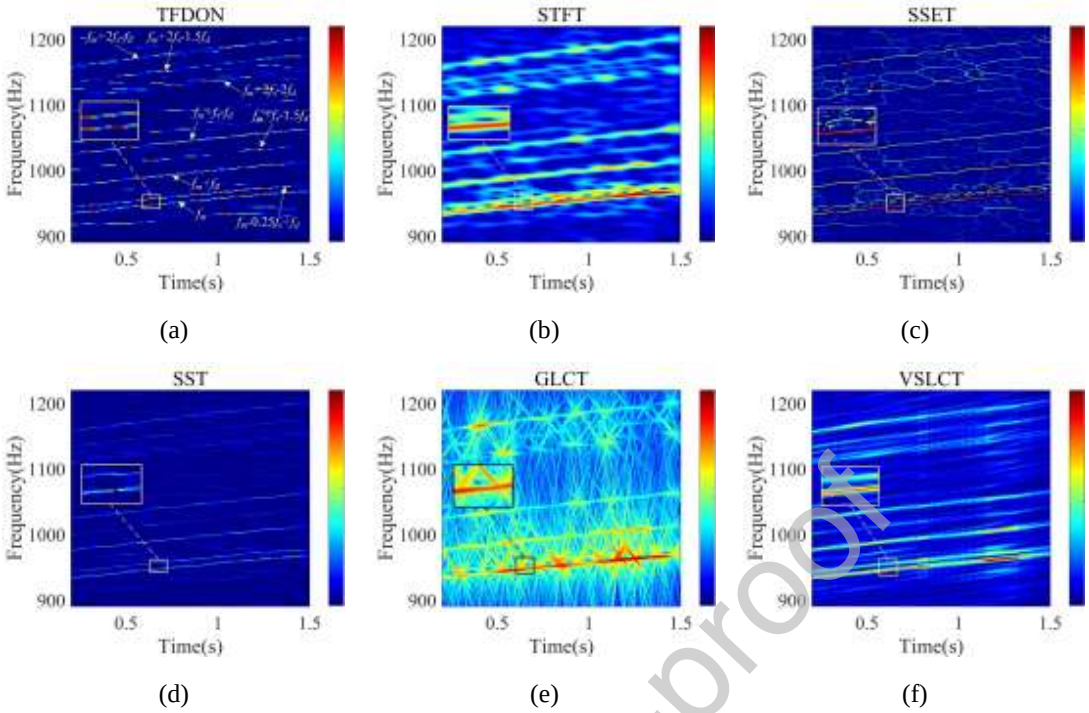


Fig. 21. TFRs generated from (a) TFDON; (b) STFT; (c) SSET; (d) SST; (e) GLCT; and (f) VSLCT.

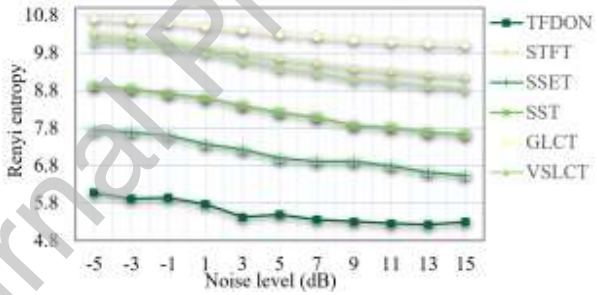


Fig. 22. Rényi entropy on planetary gearbox fault signal *vs.* SNR levels.

5. Conclusion

In this paper, the TFDON is proposed to robustly characterize the nonstationary signals of rotating machinery under strong noise environment. In the TFDON, the SDSN is first constructed, where the multi-scale expansion convolution layer and EnhancedSE attention are designed to extract noise information, and then the TFR with clean background is reconstructed by the residual connection. Second, the TFOSN, composed of cascaded HTBs, is developed to refine the denoised TFRs. Each HTB integrates alternating EGST and dual-attention mechanisms to

model long-range dependencies across time–frequency features, and optimizes time-frequency energy progressively.

The performance of the TFDON is analyzed by a closely-spaced simulation signal with 0 dB Gaussian white noise, and the results show that the TFDON can obtain a concentrated TFR and eliminate noise interference accurately. Furthermore, the experimental results show that the TFDON can characterize the bearings and planetary gearbox signals with closely-spaced FCFs under strong noise aliasing, and further detect the fault pattern. Compared with the STFT, SSET, SST, GLCT, and VSLCT, the TFDON achieves the minimum Rényi entropy at -5 to 15 dB SNRs, and exhibits better noise robustness and concentration characteristics, which verifies the theoretical feasibility of the TFDON to provide a more reliable scientific basis for rotating machinery engineering decisions in high noise scenarios.

The proposed TFDON still faces challenges in computational efficiency and physical interpretability in complex scenarios. To this end, future research will explore lightweight model design, such as structured pruning, to perform real-time computations more efficiently. In addition, physical information of mechanical parameters can be incorporated during dataset construction to improve model interpretability and generalization.

CRediT authorship contribution statement

Depei Shao: Writing - original draft, Writing - review & editing, Software, Formal analysis, Validation, Visualization. **Dezun Zhao:** Writing - original draft, Writing - review & editing, Conceptualization, Methodology, Data curation, Funding acquisition, Resources, Supervision. **Tianyang Wang:** Data curation, Resources, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

Generative AI was used solely for language polishing and grammatical improvement. All research design, analysis, results, and conclusions were conducted by the authors, who take full responsibility for the manuscript.

Data Availability Statement

The authors do not have permission to share data.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (51905292).

Reference

[1] Jianfei Z, Dongnan C, Changhua H, et al. TBiGAN-based parallel networks for remaining useful life prediction of multi-stage degraded bearings[J]. Pattern Recognition, 2025: 112349.

[2] Shi G, Qin C, Zhang Z, et al. Sparsity-assisted variational nonlinear component decomposition[J]. IEEE Transactions on Industrial Informatics, 2023, 20(3): 4173-4186.

[3] Sharma A K, Verma N K. A novel vision transformer with selective residual in multihead self-attention for pattern recognition[J]. Pattern Recognition, 2025: 112497.

[4] Zhao D, Huang X, Wang T, et al. Generalized reassigning transform: Algorithm and applications[J]. Reliability Engineering & System Safety, 2025, 255: 110677.

[5] Ohamouddou S, El Afia A, Ohamouddou M, et al. Introducing the short-time fourier Kolmogorov Arnold network: A dynamic graph CNN approach for tree species classification in 3D point clouds[J]. Pattern Recognition, 2025: 112584.

[6] M. Al-Sa'd, B. Boashash, M. Gabbouj, Design of an optimal piece-wise spline Wigner-ville distribution for tfd performance evaluation and comparison, IEEE Trans. Signal Process. 2021, 69: 3963–3976.

[7] Liu T, Li X, Sun J, et al. A post-processing method called fourier transform based on local maxima of autocorrelation function for extracting fault feature of bearings[J]. Advanced Engineering Informatics, 2024, 62: 102766.

[8] Oberlin T, Meignen S, Perrier V. The Fourier-based synchrosqueezing transform[C]//2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2014: 315-319.

[9] Yu G, Wang Z, Zhao P. Multisynchrosqueezing transform[J]. IEEE Transactions on Industrial Electronics, 2018, 66(7): 5441-5455.

[10] Yu G, Wang Z, Zhao P, et al. Local maximum synchrosqueezing transform: An energy-concentrated time-frequency analysis tool[J]. Mechanical Systems and Signal Processing, 2019, 117: 537-552.

[11] Yu G, Yu M, Xu C. Synchroextracting transform[J]. IEEE Transactions on Industrial Electronics, 2017, 64(10):

8042-8054.

[12] Bao W, Li F, Tu X, et al. Second-order synchroextracting transform with application to fault diagnosis[J]. IEEE Trans Instrum Meas, 2020, 70: 1–9.

[13] Mann S, Haykin S. The chirplet transform: Physical considerations[J]. IEEE Transactions on Signal Processing, 1995, 43(11): 2745-2761.

[14] Peng Z K, Meng G, Chu F L, et al. Polynomial chirplet transform with application to instantaneous frequency estimation[J]. IEEE Transactions on Instrumentation and Measurement, 2011, 60(9): 3222-3229.

[15] Yu G, Zhou Y. General linear chirplet transform[J]. Mechanical Systems and Signal Processing, 2016, 70: 958-973.

[16] Guan Y, Liang M, Necsulescu D S. Velocity synchronous linear chirplet transform[J]. IEEE Transactions on Industrial Electronics, 2018, 66(8): 6270-6280.

[17] Zhao D, Wang H, Cui L. Frequency-chirprate synchrosqueezing-based scaling chirplet transform for wind turbine nonstationary fault feature time–frequency representation[J]. Mechanical Systems and Signal Processing, 2024, 209: 111112.

[18] Shi G, Qin C, Xia P, et al. Generalized envelope nonlinear Gini index-gram guided two-stage chirp mode decomposition for shield machine main bearing fault diagnosis[J]. Advanced Engineering Informatics, 2026, 71: 104354.

[19] Pan P, Zhang Y, Deng Z, et al. Deep learning-based 2-D frequency estimation of multiple sinusoidals[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(10): 5429-5440.

[20] Wang Z, Chen L, Xiao P, et al. Enhancing time-frequency resolution via deep-learning framework[J]. IET Signal Processing, 2023, 17(4): e12210.

[21] Zhao D, Shao D, Cui L. CTNet: a data-driven time-frequency technique for wind turbines fault diagnosis under time-varying speeds[J]. ISA Transactions, 2024, 154: 335-351.

[22] Chen T, Jiao Y, Xie L, et al. QTFN: A General End-to-End Time-Frequency Network to Reveal the Time-Varying Signatures of the Time Series[J]. Big Data Mining and Analytics, 2024, 7(3): 905-919.

[23] Zhao D, Shao D, Wang T, et al. Time-frequency self-similarity enhancement network and its application in wind turbines fault analysis[J]. Advanced Engineering Informatics, 2025, 65: 103322.

[24] Pan P, Zhang Y, Deng Z, et al. TFA-Net: A deep learning-based time-frequency analysis tool[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 34(11): 9274-9286.

[25] Chen T, Chen Q, Zheng Q, et al. Adaptive multi-scale TF-net for high-resolution time–frequency representations[J]. Signal Processing, 2024, 214: 109247.

[26] Shi G, Qin C, Zhang Z, et al. A novel shield machine main bearing health evaluation approach based on two-stage signal decomposition[J]. IEEE Transactions on Industrial Informatics, 2026.

[27] Biswas S, Alam A M, Gurbuz A C. HRSpecNET: A deep learning-based high-resolution radar micro-Doppler signature reconstruction for improved HAR classification[J]. IEEE Transactions on Radar Systems, 2024, 2: 484-497.

[28] Zhao D, Shao D, Wang T. Dynamic cross-scale time-frequency network for characterizing non-stationary fault characteristic frequency of offshore wind turbine[J]. Ocean Engineering, 2025, 332: 121367.

[29] Tian C, Xu Y, Li Z, et al. Attention-guided CNN for image denoising[J]. Neural Networks, 2020, 124: 117-129.

[30] Zhang J, Zhang F, Wei H. PSSS-EEG: A Probabilistic-masking Self-Supervised Swin-transformer model for EEG-based drowsiness recognition[J]. Pattern Recognition, 2025, 158: 111005.

[31] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.

[32] Fang J, Lin H, Chen X, et al. A hybrid network of cnn and transformer for lightweight image super-resolution[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 1103-1112.

[33] Zhang Y, Wang X, Shakeel M S, et al. Learning upper patch attention using dual-branch training strategy for masked face recognition[J]. Pattern Recognition, 2022, 126: 108522.

[34] Zeng R, Song Y, Zhong Y. An interpretable unsupervised capsule network via comprehensive contrastive learning and two-stage training[J]. Pattern Recognition, 2025, 158: 111059.

[35] Zhao D, Cai W, Cui L. Multi-perception Graph Convolutional Tree-embedded Network for Aero-engine Bearing Health Monitoring with Unbalanced Data[J]. Reliability Engineering & System Safety, 2025: 110888.

[36] Kun Zhang, Yanlei Liu, Long Zhang, Chaoyong Ma, Yonggang Xu, Frequency slice graph spectrum model and its application in bearing fault feature extraction, Mechanical Systems and Signal Processing, 2025, 226:

112383.

[37] Huang H, Baddour N. Bearing vibration data collected under time-varying rotational speed conditions[J]. Data Brief, 2018, 21: 1745–9.

[38] Zhou P, Tong Q, Chen S, et al. Eace: Explain anomaly via counterfactual explanations[J]. Pattern Recognition, 2025, 164: 111532.

[39] Cai W, Zhao D, Wang T. Multi-scale dynamic graph mutual information network for planet bearing health monitoring under imbalanced data[J]. Advanced Engineering Informatics, 2025, 64: 103096.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: