PeriodNet: Noise-Robust Fault Diagnosis Method Under Varying Speed Conditions

Ruixian Li, Jianguo Wu, *Member, IEEE*, Yongxiang Li, *Member, IEEE*, and Yao Cheng, *Member, IEEE*

***Abstract*— Rolling bearings are critical components in modern mechanical systems and have been extensively equipped in various rotating machinery. However, their operating conditions are becoming increasingly complex due to diverse working requirements, dramatically increasing their failure risks. Worse still, the interference of strong background noises and the modulation of varying speed conditions make intelligent fault diagnosis very challenging for conventional methods with limited feature extraction capability. To this end, this study proposes a periodic convolutional neural network (PeriodNet), which is an intelligent end-to-end framework for bearing fault diagnosis. The proposed PeriodNet is constructed by inserting a periodic convolutional module (PeriodConv) before a backbone network. PeriodConv is developed based on the generalized short-time noise resist correlation (GeSTNRC) method, which can effectively capture features from noisy vibration signals collected under varying speed conditions. In PeriodConv, GeSTNRC is extended to the weighted version through deep learning (DL) techniques, whose parameters can be optimized during training. Two open-source datasets collected under constant and varying speed conditions are adopted to assess the proposed method. Case studies demonstrate that PeriodNet has excellent generalizability and is effective under varying speed conditions. Experiments adding noise interference further reveal that PeriodNet is highly robust in noisy environments.**

***Index Terms*— Bearing fault diagnosis, complex operating conditions, deep learning (DL), noise resist correlation, periodic convolutional module (PeriodConv).**

I. Introduction

ROTATING machines are the cornerstone of modern society, playing an irreplaceable role in manufacturing, transportation, etc. As productivity develops rapidly, they often operate under complex working conditions, especially at varying speeds [1]. Although highly reliable, bearings in mechanical equipment are prone to failure during long-time operations [2], whose faults may cause up to 41% of all rotating machine failures [3]. In this regard, bearing failures may lead to fatal accidents and substantial economic losses, especially in critical sectors such as manufacturing, energy, and transportation [4]. Hence, bearing fault diagnoses are crucial to ensure the safe and stable operation of rotating machines, and have attracted increasing interest from academia and industry [5].

Signal processing (SP) approaches are the most favorable among many fault diagnosis methods [6]. As rollers strike the damaged part of the faulty bearing, a series of characteristic impulses will occur in vibration signals, which could be captured easily by sensors thanks to the development of data collection techniques [7]. Thus, the bearing fault could be characterized by analyzing periodic patterns in vibration signals. However, varying speed conditions may dramatically change the characteristic frequency of signals and thus make the bearing fault diagnosis very challenging [8].

Many SP methods have been successfully applied to extract interpretable features for fault diagnoses, such as the order tracking analysis [9], the statistics and probability analysis [10], and the spectral analysis [11]. However, these methods are usually developed based on strict physical assumptions, and they commonly require labor-intensive participation by experts to make decisions [12]. Furthermore, selecting suitable parameters under varying speed conditions is considerably challenging [13].

Machine learning (ML)-based approaches have been introduced to automatically analyze the features extracted by SP methods to overcome the aforementioned difficulties [14]. Many traditional classifiers, such as the fuzzy neural network [15], Bayesian network [16], artificial neural network [17], support vector machine [18], and extreme learning machine [19], have been proposed for the bearing fault diagnosis. Although traditional ML algorithms have progressed considerably in the past two decades, their limited nonlinear mapping capacities hinder their application efficacy [20]. Hence, they cannot perform satisfactorily in solving complex fault diagnosis problems [21].

Deep learning (DL) has recently achieved unprecedented breakthroughs in various fields [22]. With advanced computing hardware, DL-based approaches can overcome the limitations of traditional ML-based methods. Thanks to their powerful feature learning capability, DL models, especially the convolutional neural networks (CNNs), have been successfully

Manuscript received 10 June 2022; revised 27 November 2022, 29 January 2023, and 11 April 2023; accepted 29 April 2023. The work of Jianguo Wu was supported in part by the National Natural Science Foundation of China under Grant 72171003. The work of Yongxiang Li was supported in part by the National Natural Science Foundation of China under Grant 72101147 and in part by the Shanghai Pujiang Program under Grant 21PJ1405500. *(Corresponding author: Yongxiang Li.)*

Ruixian Li and Yao Cheng are with the Department of Industrial and Manufacturing Systems Engineering, The University of Hong Kong, Hong Kong, China (e-mail: rxli@connect.hku.hk; yaocheng@hku.hk).

Jianguo Wu is with the Department of Industrial Engineering and Management, Peking University, Beijing 100871, China (e-mail: j.wu@pku.edu.cn).

Yongxiang Li is with the Department of Industrial Engineering and Management, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: yongxiangli@sjtu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2023.3274290.

Digital Object Identifier 10.1109/TNNLS.2023.3274290

used to perform bearing fault diagnosis tasks [23]. To fully utilize the strength of CNNs in processing images, vibration signals are first transformed into feature maps by SP methods in most cases. However, SP and DL are two separate steps in such diagnosis frameworks. Thus, the parameters of SP methods must be meticulously and manually adjusted by researchers with expertise according to bearing working conditions [24]. Moreover, it is complicated to manually specify optimal parameters to make feature maps match well with DL networks, especially under varying speed conditions [20].

End-to-end fault diagnosis frameworks, which map raw vibration signals onto bearing health states directly and automatically through one model [25], are widely applied in the literature to eliminate the reliance on the manual feature extraction process and expert involvement in fault identification [26]. These methods provide a better trade-off between efficiency and effectiveness [27]. For example, Lei et al. [28] adopted recurrent neural networks (RNNs) for the bearing fault diagnosis, and Wu et al. [29] modified conventional CNNs particularly to accommodate the 1-D signal input. Although these models perform satisfactorily in most conventional scenarios, complex operating conditions often severely affect their performance, where fault patterns are modulated by varying speeds and masked by intense background noises. This robustness issue is mainly because most existing networks are proposed for images and natural languages, whose noise mechanism differs from periodic vibration signals, which commonly are contaminated with strong noises under complex operating conditions. Processing such low signal-to-noise ratio (SNR) vibration signals requires specially designed SP approaches. Thus, the structure of existing networks may not be good at capturing informative features of these vibration signals.

Many researchers are keen on designing dedicated network structures for the bearing fault diagnosis to overcome these issues [30]. For example, conventional SP strategies, which guide the feature learning procedure of networks, are embedded into the network structure to address the interpretable problems and improve the fault diagnosis performance [31]. For example, Zhao et al. [32] proposed the adaptive intraclass and interclass CNN (AIICNN) for the intelligent fault diagnosis of gearboxes. However, AIICNN relies on frequency domain transformation to suppress noises, and its critical parameters have to be optimized manually. Zhao et al. [33] developed the deep residual shrinkage network (DRSN) to improve fault diagnosis accuracy in noisy environments. They adopted soft thresholding as a nonlinear mapping mechanism to eliminate unimportant features. Nevertheless, DRSN cannot permanently eliminate the masking effect of intense noises, which challenges the effectiveness of thresholding-based approaches.

This study focuses on the CNN-based methods because they can achieve better performance on the bearing fault diagnosis among DL methods [34]. In general, the first convolutional layer of CNN is critical to the fault diagnosis performance of the entire network [35]. For instance, Abid et al. [36] proposed SincNet by replacing the convolutional filter with the cardinal-sine (Sinc) function; Liu et al. [37] proposed TScatNet using Morlet wavelet as the predefined convolutional kernel; Li et al. [38] proposed WaveletKernelNet based on the wavelet kernel convolution. Although these networks have demonstrated considerable performance, the wavelet-basis functions used in the networks may be unable to sufficiently overcome the masking effect of noises and the modulation of varying speeds.

Besides the wavelet-based kernels, the variational kernels derived from the successive variational mode decomposition were integrated into 1-D CNNs [12]. However, these networks based on enhanced convolutional kernels all replace the first layer of the backbone network with a designed layer, which may break the structural integrity of the pretrained CNN and reduce their fault diagnosis performance. In contrast, Zhao et al. [13] proposed a denoising layer based on the reproducing kernel Hilbert space (RKHS) algorithm. They inserted it before the backbone network, which could obtain better results theoretically. However, RKHS can only deal with low SNR signals. Therefore, domain-oriented network modules that simultaneously handle strong noises and varying speeds still lack in-depth research.

Recently, the noise-resistant correlation (NRC) method was developed [39], demonstrating excellent performance in bearing fault diagnosis in noisy environments. Although NRC was designed under constant speed scenarios, its variants [40], [41] were developed to process signals contaminated with intense background noises and varying speeds. Due to its excellent performance in bearing fault diagnosis, this article proposes a generalized short-time NRC (GeSTNRC) to process varying speed signals contaminated with strong noises.

Specifically, a network module embedding the weighted GeSTNRC, called periodic convolutional module (PeriodConv), is first proposed in this study. PeriodConv is an organic fusion of the SP theories and DL techniques because the weighting process of GeSTNRC is realized via the self-attention mechanism. Then, periodic convolutional neural network (PeriodNet) is constructed by inserting PeriodConv before a backbone network. Actually, PeriodNet is an umbrella term for a series of networks whose structure depends on the selected backbone network. The proposed PeriodNet has excellent fault diagnosis performance under varying speed conditions in noisy environments. The main contributions and novelties of this study are summarized below.

1) GeSTNRC, an extension of NRC, is proposed. It can achieve an excellent balance between dealing with intense noises and varying speeds as other NRC variants. Further, GeSTNRC is a 2-D extension because it adopts the cross-autocorrelation, making it more suitable to fully utilize the strength of CNNs.
2) Based on weighed GeSTNRC, PeriodConv adapts well to the various working conditions because its parameters are optimized automatically during training. Moreover, PeriodConv is generalizable because it can be inserted before any existing convolution-based backbone network.
3) By inserting the proposed PeriodConv before a backbone network, PeriodNet is constructed as an end-to-end fault diagnosis method. The proposed PeriodNet can fully retain the integrity of existing CNN models and entirely

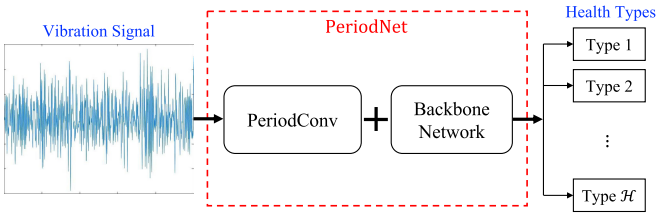

Fig. 1. Illustration of the end-to-end bearing fault diagnosis method based on PeriodNet.

use their pretrained weights. Thus, it is very effective under varying speed conditions and highly robust in noisy environments.

The rest parts are organized as follows. Section II describes the construction of PeriodNet in detail, Section III explains the preparation work for experiments, and Section IV verifies its effectiveness and robustness through three cases. Finally, the conclusions are summarized in Section V.

II. PeriodNet for Fault Diagnosis

A schematic of the proposed diagnosis method is illustrated in Fig. 1, where $\mathcal{H}$ represents the number of health types. With excellent generalizability, PeriodConv can be combined with arbitrary 2-D CNN, but this study recommends using pretrained CNNs as backbone networks.

As a robust SP algorithm, NRC could suppress the random effects of intense noises without prior knowledge. Thus, it can accurately extract the hidden periodic components in the low SNR vibration signals. Although one of the theoretical bases of vibration analysis is to diagnose bearing faults according to periodic components, it considers human thinking habits, which may not be suitable for ML. Therefore, this study removes the period detection paradigm and drastically expands NRC to a generalized version to retain more information. Furthermore, inspired by the window-based theory, especially the short-time Fourier transform (STFT) method [42], the proposed GeSTNRC method windows the input signal and processes the obtained sub-signals sequentially.

The GeSTNRC method combines convolution operations, residual connections, attention mechanisms, and other DL techniques to obtain PeriodConv. PeriodConv inputs the raw signals and outputs the feature maps, which could be optimized during the network training process. Compared with the backbone network, PeriodNet is more robust to the interference of strong background noises. At the same time, PeriodNet is more conducive to taking advantage of pretrained backbone networks, which could significantly improve their performance on bearing fault diagnosis.

A. Brief of NRC

The signals $\mathbf{y} = [y(0), \ldots, y(L-1)]^T$ are assumed to be the sum of the unknown pure signals $\mathbf{x} = [x(0), \ldots, x(L-1)]^T$ and the independently and identically normally distributed noises $\boldsymbol{\epsilon} = [\epsilon(0), \ldots, \epsilon(L-1)]^T$, where $\epsilon(t) \sim \mathcal{N}(0, \sigma^2)$ for any t. That is,

$$\mathbf{y} = \mathbf{x} + \boldsymbol{\epsilon}. \tag{1}$$

The NRC method suppresses the noises via two key steps called segmenting and averaging [39]. In the segmenting step, $\mathbf{y}$ is divided into $m = \lfloor L/l \rfloor$ segments. Each segment has a length of $l \in [1, l_{\max}]$, where $l_{\max} \leq L$ is the allowed maximum length for each segment. In this study, $\lfloor \cdot \rfloor$ represents the rounding down operation, and the last $L - ml$ signal points are not considered in the segmenting step for simplification. The obtained ith segment is denoted as

$$\mathbf{y}_i^{(l)} = \left[y((i-1)l), \ldots, y(il-1)\right]^T \tag{2}$$

for $i = 1, 2, \ldots, m$. In the averaging step, for any $l \in [1, l_{\max}]$, the average of the m segments is used to construct a mean vector of $\mathbf{y}$, denoted as

$$\boldsymbol{\mu}_{\mathbf{y}}^{(l)} = \left[\mu_{\mathbf{y}}^{(l)}(0), \ldots, \mu_{\mathbf{y}}^{(l)}(L-1)\right]^T \tag{3}$$

where its tth element is

$$\mu_{\mathbf{y}}^{(l)}(t) = \begin{cases} \dfrac{1}{m}\displaystyle\sum_{k=0}^{m-1} y(t+kl), & 0 \leq t < l \\ \mu_{\mathbf{y}}^{(l)}\left(t - \left\lfloor \dfrac{t}{l} \right\rfloor l\right), & l \leq t < L. \end{cases} \tag{4}$$

Note that the mean vector $\boldsymbol{\mu}_{\mathbf{y}}^{(l)}$ is periodic, and each period of $\boldsymbol{\mu}_{\mathbf{y}}^{(l)}$ is the average of the m segment signals $\{\mathbf{y}_i^{(l)}\}$.

Inspired by the autocorrelation function, the correlation vector between $\boldsymbol{\mu}_{\mathbf{y}}^{(l)}$ and $\mathbf{y}$ is proposed in NRC, which is denoted as $\mathbf{c}_{\mathbf{y}} = [c_{\mathbf{y}}(1), \ldots, c_{\mathbf{y}}(l_{\max})]^T$. The lth element of $\mathbf{c}_{\mathbf{y}}$ is calculated by

$$c_{\mathbf{y}}(l) = \frac{1}{L}\mathbf{y}^T \boldsymbol{\mu}_{\mathbf{y}}^{(l)}. \tag{5}$$

To cancel a drift term that increases with l, a robust autocorrelation vector $\mathbf{q}_{\mathbf{y}} = [q_{\mathbf{y}}(1), \ldots, q_{\mathbf{y}}(l_{\max})]^T$ is proposed in NRC, where

$$q_{\mathbf{y}}(l) = c_{\mathbf{y}}(l) - \frac{v_{\mathbf{y}}(l)}{m-1} \tag{6}$$

in which

$$v_{\mathbf{y}}(l) = \frac{1}{L}\left(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}}^{(l)}\right)^T \left(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}}^{(l)}\right). \tag{7}$$

It has been shown that if $\mathbf{x}$ is periodic with the period P, $q_{\mathbf{y}}(l)$ will display prominent peaks at integer multiples of P [39]. Thus, the bearing fault can be detected by the waveform of $\mathbf{q}_{\mathbf{y}}$ because the vibration signals of a faulty bearing are commonly periodic. However, these conventional fault detection methods typically require labor participation to make decisions.

B. Generalized Short-Time NRC

1) Construction of STNRC: During the operation, the speed variation of rotating machines is usually continuous. Thus, when the time interval is short enough, the sub-signal in each interval could be approximated to be stationary without period fluctuations. Therefore, inspired by window-based algorithms represented by STFT [42], this study intends to propose the short-time NRC (STNRC) method. Firstly, the window function resamples the original signals in the time

domain, obtaining a series of approximate stationary subsignals. Then the original NRC method is applied to the subsignals sequentially, realizing the processing of varying speed signals and getting time-time domain results.

For time-frequency domain analysis methods such as STFT, the window function needs to be chosen carefully, and its parameter setting is highly dependent on the experience of experts. Fortunately, as a time-time domain analysis method, STNRC does not need to transform the signals between domains, so the spectral response of the window function is not necessarily considered. In this way, STNRC could work without prior knowledge, such as filter design theories.

Denote the jth window function as $\boldsymbol{\delta}_j = [\delta_j(0), \ldots, \delta_j(L-1)]^T$ for $j \in [1, n]$, which has ℓ nonzero elements, where $n = \lfloor L/\ell \rfloor$ denotes the number of resampled subsets from $\mathbf{y}$. Specifically, $\delta_j(t) \neq 0$ if $(j-1)\ell \leq t < j\ell$, and $\delta_j(t) = 0$ otherwise. The sequence of ℓ nonzero elements forms the window of $\boldsymbol{\delta}_j$. This study assumes that all window functions share the same sequence of nonzero elements. This sequence is denoted as $\mathbf{g}$, which is a $\ell \times 1$ vector. The jth signal subset is resampled from $\mathbf{y}$ by extracting the nonzero elements from $\boldsymbol{\delta}_j \circ \mathbf{y}$, where $\circ$ denotes the element-wise product (Hadamard product). Thus, the jth subset is $\mathbf{g} \circ \mathbf{y}_j^{(\ell)}$, where $\mathbf{y}_j^{(\ell)}$ is obtained similarly to $\mathbf{y}_i^{(l)}$ in (2). It can be seen that the window of $\boldsymbol{\delta}_j$ does not overlap that of $\boldsymbol{\delta}_{j'}$ when $j \neq j'$. The nonoverlap fashion of the windows aims to avoid mixing new periodic components during the resampling process.

As the most straightforward window function, the rectangular window resamples the signals in length without changing their amplitude, which is suitable for STNRC. At the same time, the Extended NRC divides the signals into subsets and applies the NRC method to each subset [40]. Therefore, it could be regarded as a particular case of STNRC with a rectangular window function, whose overlapping interval is equal to the segment length of the subset signals. In this regard, the rectangular window is adopted in this article, leading to $\mathbf{g} = [1, \ldots, 1]_{1\times\ell}^T$.

Based on the jth resampled subset $\mathbf{g} \circ \mathbf{y}_j^{(\ell)}$, a robust autocorrelation vector of the STNRC method is proposed, denoted as

$$\mathbf{q} = \left[\mathbf{q}_{\mathbf{g}\circ\mathbf{y}_1^{(\ell)}}; \ldots; \mathbf{q}_{\mathbf{g}\circ\mathbf{y}_n^{(\ell)}}\right]. \tag{8}$$

By substituting $\mathbf{y}$ in (6) with $\mathbf{g} \circ \mathbf{y}_j^{(\ell)}$, the lth element of the autocorrelation vector $\mathbf{q}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}} = [q_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}(1), \ldots, q_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}(r)]^T$ is calculated by

$$q_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}(l) = c_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}(l) - \frac{v_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}(l)}{m-1} \tag{9}$$

where $m = \lfloor \ell/l \rfloor$ is the number of signal segments after segmenting $\mathbf{g} \circ \mathbf{y}_j^{(\ell)}$ with length l. Moreover, $c_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}(l)$ and $v_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}(l)$ are obtained by substituting $\mathbf{y}$ with $\mathbf{g} \circ \mathbf{y}_j^{(\ell)}$ in (5) and (7), respectively. In other words, $c_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}(l)$ and $v_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}(l)$ are the short-time versions of $c_{\mathbf{y}}(l)$ and $v_{\mathbf{y}}(l)$.

2) Construction of GeSTNRC: For period detection and signal reconstruction purposes, NRC adopts the conservative strategy of information utilization. However, this strategy would undoubtedly hinder the fault diagnosis performance of the NRC method. Recently, Wang et al. [1] observed a close correlation between health states and operation conditions in the vibration signals, then proposed the multitask attention CNN to learn the complementary information contained in multiple related tasks. Inspired by this observation, our study is devoted to mining richer information from vibration signals via cross-autocorrelation.

According to (1) and (5), $c_{\mathbf{y}}(l)$ can also be written as

$$c_{\mathbf{y}}(l) = \frac{1}{L}\left(\boldsymbol{\mu}_{\mathbf{y}}^{(l)}\right)^T \boldsymbol{\mu}_{\mathbf{y}}^{(l)} \tag{10}$$

which will improve the computational efficiency of NRC. Moreover, (10) shows that $c_{\mathbf{y}}(l)$ only considers the autocorrelation within $\boldsymbol{\mu}_{\mathbf{y}}^{(l)}$, and the cross-autocorrelation between $\boldsymbol{\mu}_{\mathbf{y}}^{(l)}$ and $\boldsymbol{\mu}_{\mathbf{y}}^{(l')}$ is ignored when $l \neq l'$. Instead, this study considers the cross-autocorrelation between $\boldsymbol{\mu}_{\mathbf{y}}^{(l)}$ and $\boldsymbol{\mu}_{\mathbf{y}}^{(l')}$ to obtain a generalized NRC method, where the integers $l, l' \in [1, r]$ are the length of the signal segment.

Based on (10), the autocorrelation matrix $\mathbf{C}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}$ for the resampled subset signal $\mathbf{g} \circ \mathbf{y}_j^{(\ell)}$ is

$$\mathbf{C}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}} = \begin{bmatrix} c_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}(1,1) & \cdots & c_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}(1,r) \\ \vdots & \ddots & \vdots \\ c_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}(r,1) & \cdots & c_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}(r,r) \end{bmatrix} \tag{11}$$

in which the lth row and l'th column is

$$c_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}(l,l') = \frac{1}{\ell}\left(\boldsymbol{\mu}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}^{(l)}\right)^T \boldsymbol{\mu}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}^{(l')} \tag{12}$$

where $\boldsymbol{\mu}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}^{(l)}$ denotes the mean vector of $\mathbf{g} \circ \mathbf{y}_j^{(\ell)}$ and is defined similar to $\boldsymbol{\mu}_{\mathbf{y}}^{(l)}$ in (3). According to (7), the generalized short-time version of $v_{\mathbf{y}}(l)$ corresponding to the subset signal $\mathbf{g} \circ \mathbf{y}_j^{(\ell)}$ could be calculated as

$$v_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}(l,l') = \frac{\left(\mathbf{g} \circ \mathbf{y}_j^{(\ell)} - \boldsymbol{\mu}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}^{(l)}\right)^T \left(\mathbf{g} \circ \mathbf{y}_j^{(\ell)} - \boldsymbol{\mu}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}^{(l')}\right)}{\ell}. \tag{13}$$

Denote the robust autocorrelation matrix of the GeSTNRC method as $\mathbf{Q} = [\mathbf{Q}_{\mathbf{g}\circ\mathbf{y}_1^{(\ell)}}, \ldots, \mathbf{Q}_{\mathbf{g}\circ\mathbf{y}_n^{(\ell)}}]$, where $\mathbf{Q}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}$ is the robust autocorrelation matrix of the jth subset $\mathbf{g} \circ \mathbf{y}_j^{(\ell)}$, and is given as the following:

$$\mathbf{Q}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}} = \begin{bmatrix} q_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}(1,1) & \cdots & q_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}(1,r) \\ \vdots & \ddots & \vdots \\ q_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}(r,1) & \cdots & q_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}(r,r) \end{bmatrix} \tag{14}$$

where its lth row and l'th column $q_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}(l,l')$ is the generalized short-time version of $q_{\mathbf{y}}(l)$ corresponding to $\mathbf{g}_j \circ \mathbf{y}$ and could be expressed as

$$q_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}(l,l') = c_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}(l,l') - \frac{v_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}(l,l')}{m-1}. \tag{15}$$

Equation (15) shows that the diagonal of $\mathbf{Q}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}$ equals $\mathbf{q}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}$, i.e., $\text{diag}(\mathbf{Q}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}) = \mathbf{q}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}$.

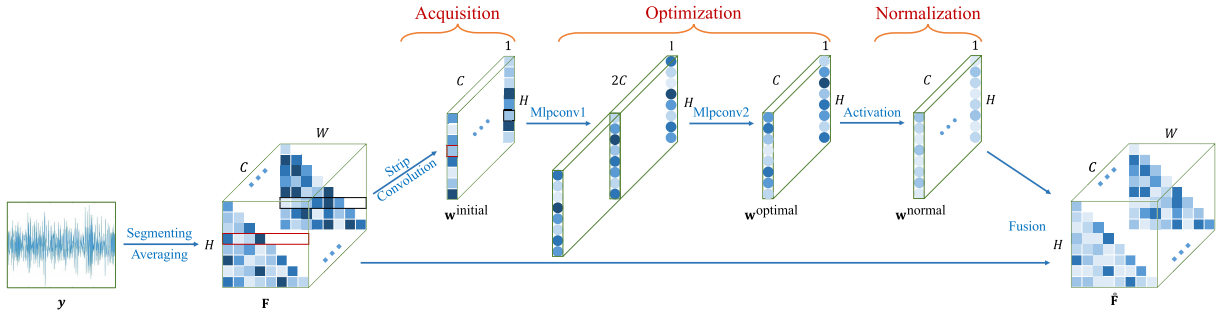

Fig. 2. Illustration of the weighting unit, where H, W, and C denote the dimension size (height, width, and channel) of feature maps.

C. Design of PeriodConv

1) Weighted GeSTNRC: Compared with the 1-D results obtained by the NRC method, the 2-D results obtained by the GeSTNRC method contain more valuable information. Thus, they can better empower the learning capability of CNN, leading to better fault diagnosis performance. However, the information in $\mathbf{Q}$ does not contribute equally to the final fault diagnosis results. Thus, the effective screening of the information becomes particularly important [43]. As a powerful tool for the neural network to achieve information screening, the attention mechanism shines in many fields, whose performance is also remarkable in bearing fault diagnosis [44]. Therefore, this study introduces the attention mechanism to GeSTNRC, and the obtained method is called the weighted GeSTNRC.

Unlike general images taken by cameras, the feature maps obtained by the GeSTNRC approach have a clear pattern during the construction, resulting in the applicability of the row-wise attention mechanism. As mentioned above, the core steps of the GeSTNRC method are segmenting and averaging, as shown in (2)–(4). For the jth subset signal $\mathbf{g} \circ \mathbf{y}_j^{(\ell)}$, the matrix consisting of all reconstructed $\boldsymbol{\mu}^{(l)}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}$ is denoted as

$$\mathbf{F}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}} = \left[\boldsymbol{\mu}^{(1)}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}, \ldots, \boldsymbol{\mu}^{(r)}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}\right]. \tag{16}$$

It can be seen from (10) and (12) that methods NRC and GeSTNRC consider the autocorrelations within and between columns of $\mathbf{F}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}$, respectively. Following these ideas, this study introduces the weight for the signal subset $\mathbf{g} \circ \mathbf{y}_j^{(\ell)}$ into (16), which is denoted as $\mathbf{w}_j = [w_j^{(1)}, \ldots, w_j^{(r)}]^T$. Furthermore, the weighted version of $\mathbf{F}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}$ is characterized as $\mathbf{F}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}$, which could be calculated as

$$\mathbf{F}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}} = \left[w_j^{(1)}\boldsymbol{\mu}^{(1)}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}, \ldots, w_j^{(r)}\boldsymbol{\mu}^{(r)}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}\right] \tag{17}$$

where $w_j^{(l)}$ is the weight of $\boldsymbol{\mu}^{(l)}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}$.

Due to the lack of prior knowledge, $\mathbf{w}_j$ is unknown. Hence, this study implements the weighted GeSTNRC method as a module in neural networks called PeriodConv. Thus, $\mathbf{w}_j$ would be optimized automatically with other network parameters during the training process, aiming to improve the feature extraction ability and alleviate the use threshold for PeriodConv.

2) Design of the Weighting Unit: As mentioned above, the feature maps obtained by the GeSTNRC method are 3-D, and these three dimensions can be divided into spatial and depth dimensions. The depth dimension is also called the channel dimension, which is generally used to store information that is relatively independent of each other. For example, the three channels of the RGB image store the grayscale information colors. Because the processes of each $\mathbf{Q}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}$ in $\mathbf{Q}$ are separate, their calculation process will not intersect with each other. Thus, this study intends to collapse the windowing results into the channel dimension. In this way, all signal subsets can be processed simultaneously via existing DL techniques, which can significantly improve the computational efficiency of the weighted GeSTNRC method.

Moreover, (4) indicates that the data of $\boldsymbol{\mu}_{\mathbf{y}}^{(l)}$ in the interval $t \in [l, L-1]$ just repeat those in the interval $t \in [0, l-1]$, which does not provide extra information. Thus, when the signal reconstruction is not considered, the duplicate elements in the interval $t \in [l, L-1]$ of $\boldsymbol{\mu}_{\mathbf{y}}^{(l)}$ can be omitted to improve computational efficiency. In this way, the element $\mathbf{F}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}$ of $\mathbf{F}$ used in PeriodConv is simplified to a lower triangular matrix. As the lower triangular matrix provides the same adequate information as the original matrix, no symbolic distinction is made in this study.

Many attention-based CNNs have been designed for fault diagnosis via the self-attention mechanism [33], which calculates weights from the input. Traditionally, the weights among channels are concerned because their feature maps are obtained mainly by different kernels and lack physical meaning. However, the $\mathbf{Q}$ consists of channels corresponding to different signal subsets. As shown in (17), this study considers the weights within the channel, which makes the standard weighting process infeasible. Therefore, as shown in Fig. 2, a novel weighting unit inspired by SENet [45] is proposed in this study.

As all channels of $\mathbf{F}$ will follow the same weighting process, this study only takes the jth channel as an example for illustration. As shown in Fig. 2, the body of the weighting unit comprises three steps as follows.

a) Acquisition of the initial weight: According to (17), a weight vector $\mathbf{w}_j$ with length $r \times 1$ is required for each $\mathbf{F}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}$ in $\mathbf{F}$. In general, the initial weight $\mathbf{w}_j^{\text{initial}}$ is acquired via average pooling [45]. However, the average operation is already adopted during the construction of

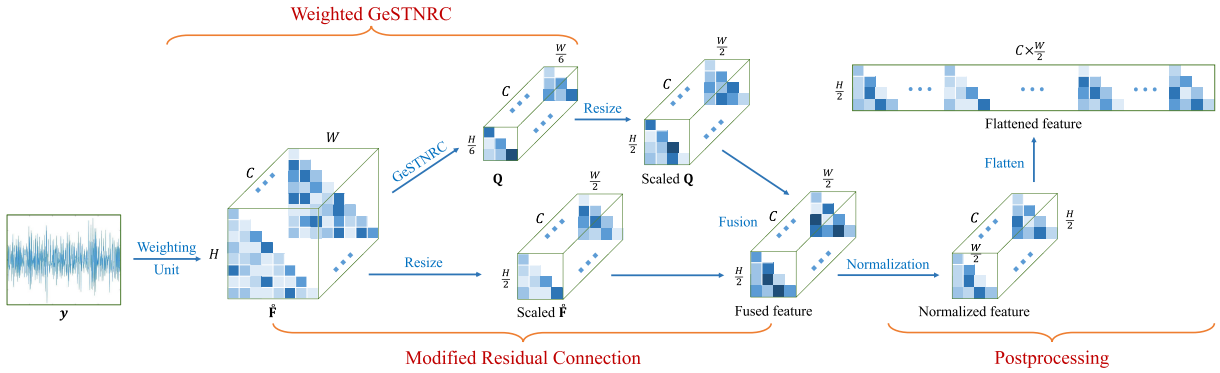

Fig. 3. Illustration of PeriodConv.

$\boldsymbol{\mu}^{(l)}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}$, making the average pooling can only capture minimal information.

Considering that the dimension of the matrix $\mathbf{F}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}$ is $r \times r$, a feature map of $r \times 1$ will be obtained using the strip convolution with kernel size $1 \times r$ [23], which exactly meets the dimension requirement of the $\mathbf{w}_j^{\text{initial}}$. Therefore, this study will use the channel-wise strip convolution to calculate the $\mathbf{w}_j^{\text{initial}}$. Denote the horizontal strip kernel for channel j as $\mathbf{k}_j^T$, where $\mathbf{k}_j = [k_j^{(1)}, \ldots, k_j^{(r)}]^T$. When the stride for the strip convolution equals the kernel size, the lth element of initial weights can be acquired as

$$\left(w_j^{\text{initial}}\right)^{(l)} = \mathbf{k}_j^T \boldsymbol{\mu}^{(l)}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}. \tag{18}$$

In addition, the strip convolution on matrix $\mathbf{F}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}$ that contains periodic signals is the core operation of the proposed module, and thus it is called PeriodConv.

b) Optimization of the initial weight: To better capture the dependencies among columns of $\mathbf{F}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}$, the $\mathbf{w}_j^{\text{initial}}$ needs to be further optimized, which is typically implemented by several fully connected (FC) layers. However, as global operations, FC layers cannot meet the channel-wise needs of PeriodConv. Although the channel-wise FC layer has been proposed [46], it may face computational issues because its code lacks optimization by developers. Fortunately, the Mlpconv layer could do a similar job as the FC layer [47], which is implemented by existing common layers. Therefore, this study uses the channel-wise Mlpconv layer to obtain the optimal weights $\mathbf{w}_j^{\text{optimal}}$.

When optimizing $\mathbf{w}_j^{\text{initial}}$ with a sub-network composed of FC layers, the activation layer is often adopted to improve its nonlinear learning ability, which may help learn more valuable features. However, unlike general feature maps obtained from convolutional kernels, the relationships among columns of the matrix $\mathbf{F}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}$ are interpretable, which significantly reduces the dependence on the sub-network. At the same time, an overly complex sub-network for $\mathbf{w}_j^{\text{optimal}}$ may increase the overfitting risk, resulting in poor generalization. Therefore, the sub-network for the optimization process consists of two channel-wise convolutional layers, where the first layer has two kernels for each channel. Mathematically, due to the kernel size of Mlpconv layers being 1×1, kernels of the first Mlpconv layer for channel j are denoted as $k_j^{(1)}$ and $k_j^{(2)}$, respectively, and the kernel of the second Mlpconv layer is denoted as $k_j^{(3)}$. Thus, $\mathbf{w}_j^{\text{optimal}}$ could be calculated as

$$\mathbf{w}_j^{\text{optimal}} = \left(k_j^{(1)} + k_j^{(2)}\right) k_j^{(3)} \mathbf{w}_j^{\text{initial}}. \tag{19}$$

c) Normalization of the optimal weight: After the optimization step, $\mathbf{w}_j^{\text{optimal}}$ will be normalized using the Sigmoid function. Thus, the lth element of final weight $\mathbf{w}_j^{\text{normal}}$ could be obtained

$$\left(w_j^{\text{normal}}\right)^{(l)} = \frac{1}{1 + e^{-\left(w_j^{\text{optimal}}\right)^{(l)}}}. \tag{20}$$

By repeating (18)–(20), $\mathbf{w}_j$ in PeriodConv can be optimized, which results in $\mathbf{w}_j^{\text{normal}}$. Then, $\mathbf{w}_j^{\text{normal}}$ will be fused with the input $\mathbf{F}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}$ according to (17), and the obtained $\mathbf{F}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}$ will be used to calculate $\mathbf{Q}_{\mathbf{g}\circ\mathbf{y}_j^{(\ell)}}$. Moreover, the optimization process of $\mathbf{w}_j$ is wholly based on the existing network structure, so it has a unique advantage in computational efficiency.

Furthermore, in the processes of the weighted GeSTNRC, the constructed matrix $\mathbf{F} = [\mathbf{F}_{\mathbf{g}\circ\mathbf{y}_1^{(\ell)}}, \ldots, \mathbf{F}_{\mathbf{g}\circ\mathbf{y}_n^{(\ell)}}]$ will be replaced by its weighted version $\tilde{\mathbf{F}} = [\mathbf{F}_{\mathbf{g}\circ\mathbf{y}_1^{(\ell)}}, \ldots, \mathbf{F}_{\mathbf{g}\circ\mathbf{y}_n^{(\ell)}}]$. Besides, other steps will not be repeated because they remain the same as GeSTNRC.

3) Implementation of PeriodConv: After too many middle layers of back-propagation, the gradients of the loss function will degrade more or less [48], which makes the parameters training in the beginning layers a challenging job. As a transgenerational network structure, the residual connection could address this difficulty through the identity shortcut [48]. With the great privilege of residual connections, the residual networks (ResNets) could be very deep and have been applied for the fault diagnosis successfully [49].

Considering the improvement of residual connections to fault diagnosis performance, this study designs a modified residual structure in PeriodConv. As shown in Fig. 3, PeriodConv is roughly divided into three operations, and $\mathbf{F}$ is processed by two parallel streams.

a) Execution of the weighted GeSTNRC: The weighted GeSTNRC consists of the weighting unit and the first part of the upper stream. As shown in Fig. 3, the spatial dimension of $\mathbf{Q}$ is different from that of $\mathbf{F}$. In practice, not all signal

TABLE I
HYPER-PARAMETERS OF PERIODCONV

Layer	Type	Name	Filters	Size / Stride	Input	Output
Weighting Unit	Input				$1 \times L \times 1$	$1 \times L \times 1$
	Designed	Segmenting & Averaging			$1 \times L \times 1$	$r \times r \times n$
	Group Convolution	Strip Convolution	n	$1 \times r$	$r \times r \times n$	$r \times 1 \times n$
	Group Convolution	Mlpconv1	$2n$	1×1	$r \times 1 \times n$	$r \times 2 \times n$
	Group Convolution	Mlpconv2	n	1×1	$r \times 2 \times n$	$r \times 1 \times n$
	Sigmoid	Activation			$r \times 1 \times n$	$r \times 1 \times n$
	Multiplication	Fusion			$r \times 1 \times n$, $r \times r \times n$	$r \times r \times n$
Upper Stream	Designed	GeSTNRC			$r \times r \times n$	$r/6 \times r/6 \times n$
	Resize	Resize			$r/6 \times r/6 \times n$	$r/2 \times r/2 \times n$
Lower Stream	Resize	Resize			$r \times r \times n$	$r/2 \times r/2 \times n$
Both	Addition	Fusion			$r/2 \times r/2 \times n$, $r/2 \times r/2 \times n$	$r/2 \times r/2 \times n$
	Layer Normalization	Normalization			$r/2 \times r/2 \times n$	$r/2 \times r/2 \times n$
	Depth To Space	Flatten			$r/2 \times r/2 \times n$	$r/2 \times r \times n/2$

points of $\mathbf{q_y}$ obtained by NRC are used for fault diagnosis. This is because the denoising effect will gradually degrade as l increases [39], where the noise term contained in the correlation vector $\mathbf{c_y}$ could be expressed as σ^2/m. Thus, the influence of noises increases as m decreases. As $m = \lfloor L/l \rfloor$, the end part of $\mathbf{q_y}$ is more affected by noises than the beginning part because the former has a larger index value l. For example, (4) indicates that only one signal point is involved in the averaging step when $l \in (L/2, L]$, only two signal points are involved in the averaging step when $l \in (L/6, L/2]$, and so on. Obviously, a smaller l means more signal points involved in the averaging step, resulting in a better denoising effect.

Considering that PeriodConv is not designed for period detection and CNNs usually have better pattern recognition capabilities than human experts, this study intends to retain more points of denoised signals than the NRC method by setting $l_{\max} = L/6$. Therefore, for the $\mathbf{F}$ whose dimension size of the height, width, and channel is $r \times r \times n$, the output of the GeSTNRC method in PeriodConv is cropped to $r/6 \times r/6 \times n$, and it is still denoted as $\mathbf{Q}$.

b) Construction of the modified residual connection: The two streams of $\mathbf{F}$ compose the modified residual connections, where the modification lies in resizing operations. Traditionally, the feature map fusion in the residual structure is performed via the additional layer. Thus, feature maps for fusion must have consistent dimensions. However, the size of $\mathbf{Q}$ is smaller than $\mathbf{F}$ as mentioned above, so the resize layers are adopted, whose output dimension is $r/2 \times r/2 \times n$. Therefore, the output size after the residual structure is half of the input in height and width.

c) Postprocessing of the feature maps: To improve the generalization capacity of PeriodConv, the fused feature map will pass through a normalization layer, which normalizes a mini-batch of data across all channels for each observation independently.

The final step of PeriodConv is to flatten the normalized feature map, which will restore the information of $\mathbf{F}$ that collapsed to the channel dimension back to the spatial dimension. The flattening operation is necessary because PeriodConv will be inserted before the first layer of the backbone network, which generally assumes the input feature map has interdependent channels. This assumption holds for standard RGB images but contradicts the GeSTNRC approach. Thus, the normalized feature map is flattened to prevent the backbone network from forcibly extracting dependencies from unrelated channels. Consequently, the utility of features extracted by the GeSTNRC method is maximized.

Figs. 2 and 3 together illustrate the design of PeriodConv. To introduce the structure of PeriodConv more clearly, its hyper-parameters are summarized in Table I.

D. Construction of PeriodNet

Based on PeriodConv, this study proposes a novel CNN model for end-to-end bearing fault diagnosis called PeriodNet. As PeriodConv is generalizable to be combined with any 2-D CNN and acts as its first layer, PeriodNet is an umbrella term for a series of models whose specific structure is determined by the backbone network.

As $\mathbf{w}_1, \ldots, \mathbf{w}_n$ in PeriodConv are trainable, the backbone network can affect the feature extraction capability of PeriodConv through the back-propagation process. In this way, the backbone network can affect the final fault diagnosis performance directly and indirectly. However, as the number of parameters in PeriodConv is usually much fewer than that in the backbone network, the choice of the backbone network will not play a decisive role in the feature extraction capability of PeriodConv.

At the same time, the contribution of PeriodConv to the final fault diagnosis task is mainly achieved indirectly by enhancing the fault feature learning ability of the backbone network. Since one of the primary purpose of this study is to explore the automatic feature extraction ability of the proposed PeriodConv from nonstationary vibration signals, this study selects the ResNet-18 [48] as the backbone network, which is one of the most popular pretrained networks.

The ResNet-18 was initially proposed for image recognition problems [48], which introduced the residual block to address the gradient degradation issue caused by the increased layer number in the network. There are many other popular variants of ResNets, such as the ResNet-50 and ResNet-101, whose suffixes are the number of network layers and their structure is almost the iterations of residual blocks. As this section aims to verify the effectiveness of PeriodConv rather than obtain the best fault diagnosis performance, the ResNet-18 is considered to save computation resources.

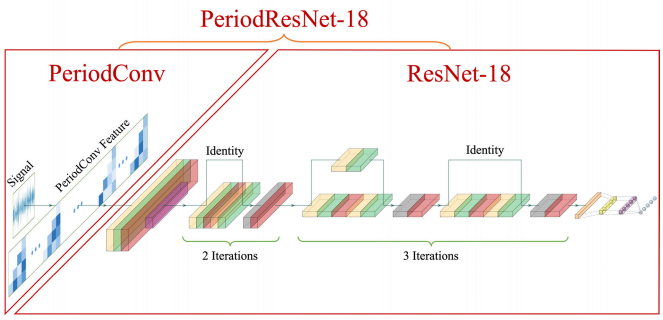

Fig. 4. Illustration of PeriodResNet-18.

For convenience, this study denotes PeriodNet based on the ResNet-18 as PeriodResNet-18, whose structure is shown in Fig. 4. Because this study only modifies the input and output layers of the original ResNet-18 to meet the dimension needs of the diagnosis problem, its hyper-parameters will not be described in detail. In addition, other PeriodNets are similar in structure to PeriodResNet-18, so their structures will not be described in the following.

The "Iteration" in Fig. 4 indicates that the group of repeated layers and the number in front represents the repetitions. It is worth noting that the layers that belong to the same "Iteration" in Fig. 4 have similar structures but different hyper-parameters. As the corresponding details have nothing to do with the research content of this study, the model structure shown in Fig. 4 is only for illustration.

E. Performance Assessment Indicator

The Macro-$F1$ score is utilized to validate the performance of the developed method of bearing fault diagnosis, whose calculation is described in

$$\begin{aligned} \text{Precision}_h &= \frac{\text{TP}_h}{\text{TP}_h + \text{FP}_h}, \quad \text{Recall}_h = \frac{\text{TP}_h}{\text{TP}_h + \text{FN}_h} \\ F1_h &= \frac{2 \times \text{Precision}_h \times \text{Recall}_h}{\text{Precision}_h + \text{Recall}_h} \\ \text{Macro-}F1 &= \frac{1}{\mathcal{H}} \sum_{h=1}^{\mathcal{H}} F1_h \end{aligned} \tag{21}$$

where TP_h denotes the true positive in recognizing the type h of bearing health state, FP_h signifies the false recognition of this state, and FN_h represents the number of false negatives. As shown in (21), the Precision_h and Recall_h of the health type h are calculated first, which will be used to calculate its $F1_h$. Then, the Macro-$F1$ will average all $F1_h$s for $h = 1, \ldots, \mathcal{H}$.

III. Experiments Preparation

This section introduces the necessary components for comparative experiments to explore the overall performance of PeriodNets on bearing fault diagnosis. This article utilizes an open-source dataset from Ottawa University (OU), which collects vibration signals under varying speed conditions. To verify the generalizability of PeriodConv on backbone networks, this article considers six more state-of-the-art pre-trained CNNs simultaneously besides ResNet-18. Since this article conducts experiments from three perspectives, including speed conditions, noise levels, and noise types, their details are presented in the following.

TABLE II
Description of Selected Sub-Datasets in the CWRU Dataset

Health Types	File Index	Fault Width (mm)	Location of Bearing Faults	Locations of Bearings
1	3001	0.72	IR	DE
2	118	0.18	BA	DE
3	197	0.36	OR centered	DE
4	290	0.36	BA	FE
5	298	0.18	OR orthogonal	FE
6	302	0.18	OR opposite	FE
7	97	N.A.	N.A.	FE

A. Data Description

1) OU Dataset: There are two versions of the dataset provided by the OU. This study adopts the second version [50], which considers a challenging scenario in which the gearbox affects the bearing fault diagnosis task.

Five bearing health states and four varying speed conditions result in twenty different setting cases. The health states include: 1) normal; 2) inner race fault (IR); 3) outer race fault (OR); 4) ball fault (BA); and 5) combined fault in three locations. Besides, the speed conditions include: 1) acceleration; 2) deceleration; 3) acceleration then deceleration; and 4) deceleration then acceleration, whose frequency range of the shaft rotation is 10–40 Hz. Each signal is sampled at 200 kHz and has a time duration of 10 s. Furthermore, three trials are conducted for each setting case to ensure data authenticity, which leads to 60 datasets.

However, the fixed pattern for the speed variation of bearings does not exist in practice, which means the operating conditions simulated by the four-speed conditions are essentially the same. Therefore, this study does not distinguish between various conditions in the case studies. As all five bearing health states are considered, and there is only one bearing per health state, the health type number is directly aligned with the health state. Therefore, as all files provided by the OU are adopted, information such as the file indices is omitted.

2) CWRU Dataset: Although PeriodNet is designed for varying speed conditions, it should perform well under constant speed conditions. Thus, the CWRU dataset [51], which is applied widely in the research of bearing fault diagnosis, is employed to validate the proposed algorithms.

There are more bearings in the CWRU dataset, which are deployed separately at the drive end (DE) and fan end (FE). Because there are many available data files in the CWRU dataset, the most challenging files recommended by a benchmark study [52] are utilized in this study. Specifically, the selected sub-dataset is sampled at 12 kHz via the accelerometer deployed at DE. There are seven bearing health states in total, whose details are presented in Table II.

B. Data Preprocessing

1) Sample Segmentation: The collected time-series data needs to be truncated into individual samples before the

training process, and the total number of samples in both datasets is set to three hundred. Since the sampling frequency of the OU dataset is 200 kHz, which is high enough to cause storage and transmission issues in practice. Thus, its signals are downsampled to 10 kHz, similar to the CWRU dataset. At the same time, since the OU provides twelve equivalent datasets in each health state, only twenty-five samples are needed to be obtained per dataset. Considering all varying speed conditions are analyzed together, $1/8$ of points at the beginning and end of the dataset are dropped in this study, which may help reduce the startup impact and further enhance the uniformity of the datasets under different varying speed conditions. Therefore, when no overlap is allowed between adjacent samples, each sample contains 3000 signal points after truncation. Then, all data samples in the OU dataset are further divided for training, validation, and test, respectively, with a ratio of 6:2:2, where the validation dataset is used for validation during training.

On the contrary, the challenging sub-datasets provided by the CWRU contain fewer signal points in each file, leading to 1000 signal points in each sample. Therefore, the collected time-series data of the CWRU dataset is first truncated into three parts for training, validation, and test, respectively, with the same ratio as the OU dataset. Then, inspired by Zhao et al. [53], the sliding window with a length of 1000 signal points is utilized to intercept the signal parts, which will enlarge the total number of samples to three hundred.

2) Noise Addition: To verify the robustness of PeriodNets under noisy scenarios, especially errors caused by sensors [14], white Gaussian noises with SNR of $[10, 3, 0, -3, -7, -10]$ dB are added to the OU dataset. In this way, six more noised samples could be synthesized from each original sample, while the inherent noises are ignored in this study. Therefore, seven noise levels are considered for the same PeriodNet.

At the same time, four types of colored noises: pink, brown, blue, and purple, are also considered to explore the generalizability of PeriodNets. Considering noise that follows a Gaussian distribution is one of the most common noises in bearing fault diagnosis [14], [16], [33], colored Gaussian noises with 0 SNR are employed in the OU dataset. Therefore, five noise types are considered for the same PeriodNet.

C. Backbone Networks Description

To explore the generalizability of the proposed PeriodConv, six more state-of-the-art pretrained CNNs are utilized as backbone networks to establish PeriodNets, which are listed in Table III. The DL Toolbox of MATLAB provides many prevalent models, which are initialized with weights pretrained on the ImageNet dataset [54]. Only models that balance the depth, size, and volume of parameters could be chosen as backbone networks. At the same time, only the shallowest one in the same series of models is selected. Among the involved networks, the suffixes consist of one letter and one number (including "V3," "V2," and "B0") representing the version of corresponding models. Furthermore, the suffix consists of one number (including "19") denotes the number of layers in the network. At the same time, ResNet-18, mentioned in Section II-D, is marked as M0 in the following.

TABLE III
SUMMARY OF BACKBONE NETWORKS IN COMPARATIVE EXPERIMENTS

Backbone Network	Depth	Size (MB)	Parameters (Millions)	Notation
Inception-V3	48	89	23.9	M1
MobileNet-V2	53	13	3.5	M2
Xception	71	85	22.9	M3
ShuffleNet	50	5.4	1.4	M4
DarkNet-19	19	78	20.8	M5
EfficientNet-B0	82	20	5.3	M6

The backbone networks listed in Table III get their corresponding PeriodNets similar to M0 (as shown in Fig. 4), which are named by adding the prefix "Period" to the notations. At the same time, the backbone networks M1–M6 and the corresponding PeriodM1–PeriodM6 are tested the same way as M0 and PeriodM0.

D. Experiment Settings

The implementation of the proposed PeriodNet relies on MATLAB 2021b, and the computing platform is an NVIDIA GeForce RTX 3080 with 12G video memory.

The loss function is set as cross-entropy, which is the most extensively used loss function in multiclassification problems. To accelerate the convergence speed of PeriodNet and save video memory, the mini-batch training mode is adopted, whose size is 32 for all experiments unless the memory is insufficient. Besides, the stochastic gradient descent with momentum is adopted as the optimization algorithm, and the momentum is set to 0.9. To prevent overfitting and loss oscillation, the learning rate is decayed by half every five epochs, whose initial value is 0.001. The training process will be terminated after two hundred epochs.

Meanwhile, the early stopping mechanism will be triggered if the performance of PeriodNet on the validation dataset is not improved after five validations, where the performance is validated per two epochs. In addition, a random seed is set to 0 in all experiments to ensure the same initialization condition, and ten trails are adopted to further suppress the randomness of experiments. At the same time, without loss of generality, the hyper-parameter n in Table I is set to 10.

IV. CASE STUDIES

This section conducts three comparative experiments to explore the overall performance of PeriodNets on bearing fault diagnosis, where the results of backbone networks are also included for comparison. It is worth noting that some layers of the original backbone networks must be modified to meet the dimensional constraints of the 1-D vibration signals. Although such modifications will inevitably waste some pretrained weights, it is the traditional way to use the pretrained networks for end-to-end bearing fault diagnosis. Besides this modification, the experiment settings and rest hyper-parameters are kept the same for both models.

Although PeriodNets are constructed based on GeSTNRC, which is an extension of NRC, the fault diagnosis performance

TABLE IV
COMPARATIVE RESULTS UNDER DIFFERENT SPEED CONDITIONS

Methods	OU Dataset (Varying speed condition)						CWRU Dataset (Constant speed condition)					
	Backbone Network		PeriodNet Version				Backbone Network		PeriodNet Version			
	Mean	Std	Mean	Std	RMean	RStd	Mean	Std	Mean	Std	RMean	RStd
M0	0.9191	0.0457	0.9950	0.0042	108.26%	9.19%	**0.9936**	**0.0059**	**0.9990**	**0.0012**	100.54%	20.34%
M1	**0.9660**	**0.0242**	0.9963	0.0048	103.14%	19.83%	0.7178	0.2218	0.9817	0.0140	136.77%	6.31%
M2	0.7836	0.1082	0.9957	0.0039	127.07%	3.60%	0.9132	0.0802	0.9931	0.0026	108.75%	3.24%
M3	0.4687	0.0828	0.9920	0.0042	211.65%	5.07%	0.3421	0.1023	0.9974	0.0024	291.55%	2.35%
M4	0.7495	0.0741	0.9903	0.0058	132.13%	7.83%	0.9692	0.0118	0.9826	0.0070	101.38%	59.32%
M5	0.9154	0.0515	0.9977	**0.0022**	108.99%	4.27%	0.7085	0.1939	0.9979	0.0024	140.85%	1.24%
M6	0.8216	0.0767	**0.9987**	0.0023	121.56%	3.00%	0.8367	0.0888	0.9902	0.0048	118.35%	5.41%
Average	0.8034	0.0662	0.9951	0.0039	130.40%	7.54%	0.7830	0.1007	0.9917	0.0049	142.60%	14.03%

of raw data rather than NRC features are adopted for comparison. This is because the combination of NRC results and backbone networks is a feature-based diagnosis method rather than an end-to-end method.

A. Studies on Speed Condition

The proposed PeriodNets are applied separately to test samples in the OU and CWRU datasets, whose performance on bearing fault diagnosis is summarized in Table IV. Table IV measures the performance through the mean and standard deviation (Std) of Macro-$F1$ scores, where the best values in the results obtained in each column are marked in bold. Moreover, Table IV calculates the relative mean (RMean) and relative standard deviation (RStd) of PeriodNets compared to backbone networks, aiming to present the performance of PeriodNets visually. As shown in Table IV, PeriodNets achieve higher means and lower Stds than backbone networks on both datasets, which verifies the effectiveness of the proposed PeriodConv under different speed conditions.

Table IV also shows that the performance gains brought by PeriodConv to backbone networks vary wildly among models, which roughly depends on the performance of backbone networks. For instance, the relative advantage of PeriodResNet-18 on the CWRU dataset is the worst of all experiments, whose mean value is only slightly improved. This poor improvement does not negate the generalizability of PeriodConv, because ResNet-18 already performs excellently on the CWRU dataset [23], leading to little room for improvement. Besides, the Std of PeriodResNet-18 is significantly reduced, proving the effectiveness of PeriodConv.

It is worth noting that although structures of models M1 (Inception-V3) and M3 (Xception) are inextricably linked, Table IV indicates that their performance is totally different. In particular, the relative performance of M1 and M3 on bearing datasets is diametrically opposite to their pretraining results on the ImageNet dataset. Therefore, this phenomenon provides evidence to reveal the difference between the bearing fault diagnosis and the image recognition, which also justifies the necessity of specially designing CNNs for bearing fault diagnosis.

The last row of Table IV calculates the average scores of different methods, where the scores are higher on the OU dataset than on the CWRU dataset. This score inversion contradicts the consensus that bearing fault diagnosis under varying speed conditions is more challenging. However, such a comparative conclusion heavily depends on the selected datasets, which is far from a general conclusion. Specifically, the OU dataset is very different from the CWRU dataset in many respects, and it is unfair to compare the model performance without control variables directly. Although the interference of the gearbox is deliberately introduced in the OU dataset, the CWRU dataset has a more complicated experimental environment. Worse still, this study selects the most challenging sub-dataset for diagnosis, making it normal for models to perform worse on the CWRU dataset.

Although Macro-$F1$ scores of PeriodNets on the CWRU dataset are not outstanding compared to the OU dataset, the RMean improvement brought by PeriodConv is more apparent. With the strong generalizability of PeriodConv, it is reasonable to believe that PeriodNet can achieve higher Macro-$F1$ scores when selecting more advanced pretrained CNN models as backbone networks, which will meet the requirements of intelligent diagnosis of bearing faults in practice adequately.

B. Studies on Noise Level

Table V summarizes the comparative results of all models on the OU dataset at different white Gaussian noise levels. Although mean values of Macro-$F1$ scores decrease inevitably as the noise level increases, PeriodNets exhibit more robustness than their backbone networks. For instance, comparing the results of M6 and PeriodM6 at noise level VI with those on the original data, it can be found that the performance degradation for both models on the OU dataset is 47.37% and 28.71%, respectively.

Table V shows that M1 and PeriodM5 perform slightly better than other models. Specifically, when noises do not exceed level III, the means of model M1 significantly overwhelm other models, and its Stds are also in the first echelon. Besides, the means of PeriodM5 achieved the highest score five times in seven experiments. However, its Stds do not outperform similar models, which only achieve the best value three times. Even so, PeriodM5 still has apparent advantages over PeriodM1 and PeriodM3. According to Table III, with the shallow depth and simple structure, DarkNet-19 performs better than others with similar model size and parameter volume when combined with PeriodConv, which provides valuable advice for the domain-oriented neural network design in bearing fault diagnosis.

At the same time, the Std of PeriodM6 at noise level V is greater than that of M6, which means the insertion of

TABLE V
COMPARATIVE RESULTS AT DIFFERENT NOISE LEVELS

		Backbone Network							PeriodNet Version						
	Noise Level	Original	I	II	III	IV	V	VI	Original	I	II	III	IV	V	VI
Mean	M0	0.9191	0.9112	0.8254	0.7775	0.7580	0.6060	0.5499	0.9950	0.9903	0.9747	0.9380	0.8810	0.7763	0.7055
	M1	**0.9660**	**0.9598**	**0.9458**	**0.9152**	0.8256	0.6568	0.6175	0.9963	0.9967	0.9833	0.9637	0.9020	0.7539	0.6710
	M2	0.7836	0.7627	0.7611	0.6745	0.7676	**0.6869**	0.6037	0.9957	0.9937	0.9746	0.9498	0.8717	0.7567	**0.7286**
	M3	0.4687	0.4725	0.4663	0.4542	0.4160	0.3759	0.3837	0.9920	0.9900	0.9800	0.9415	0.8587	0.7003	0.6699
	M4	0.7495	0.7063	0.6566	0.6738	0.7187	0.6273	**0.6349**	0.9903	0.9903	0.9709	0.9264	0.8801	0.7345	0.6858
	M5	0.9154	0.9268	0.8897	0.8125	**0.8523**	0.6128	0.5366	0.9977	**0.9980**	**0.9943**	**0.9810**	**0.9523**	**0.8185**	0.6921
	M6	0.8216	0.7435	0.6717	0.5310	0.4610	0.4270	0.4324	**0.9987**	0.9977	0.9896	0.9666	0.9262	0.7816	0.7120
Std	M0	0.0457	0.0721	0.0901	0.0744	0.0805	0.0464	0.0705	0.0042	0.0048	0.0092	0.0280	0.0160	**0.0153**	0.0167
	M1	**0.0242**	0.0339	**0.0315**	**0.0497**	0.0842	0.0414	0.0483	0.0048	0.0031	0.0094	0.0136	0.0278	0.0384	0.0266
	M2	0.1082	0.1246	0.0827	0.1078	0.0579	0.0503	0.0387	0.0039	0.0061	0.0061	0.0144	0.0313	0.0175	0.0196
	M3	0.0828	0.0553	0.0775	0.0559	0.0729	0.0649	0.0361	0.0042	0.0042	0.0074	0.0235	0.0293	0.0343	0.0197
	M4	0.0741	0.0973	0.0783	0.1012	**0.0483**	0.0570	**0.0283**	0.0058	0.0058	0.0116	0.0183	**0.0106**	**0.0153**	**0.0149**
	M5	0.0515	**0.0326**	0.0893	0.0874	0.0641	0.1233	0.0556	**0.0022**	0.0036	**0.0032**	**0.0065**	0.0127	0.0207	0.0228
	M6	0.0767	0.1058	0.0726	0.0691	0.0518	**0.0269**	0.0415	0.0023	**0.0027**	0.0060	0.0070	0.0198	0.0370	0.0233

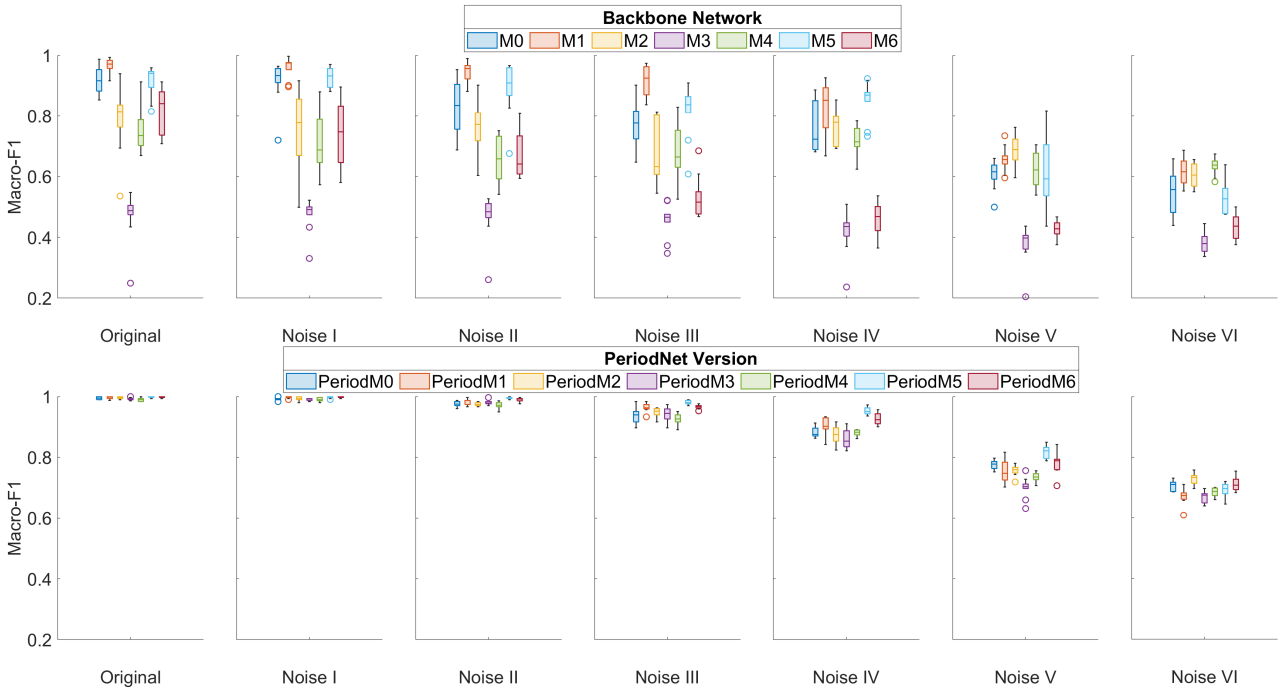

Fig. 5. Comparative results at different noise levels.

PeriodConv increases the instability of M6. However, the price is worth it as the mean is almost doubled. Moreover, the coefficient of variation of M6 is 0.0629, while the score of PeriodM6 is only 0.0473. Thus, the equivalent instability of PeriodM6 is only 3/4 of M6 after excluding the mean increase.

Moreover, since boxplots could display values of Macro-$F1$ scores and their variance simultaneously, they are helpful to compare the performance of all models more intuitively. Fig. 5 composes upper and lower parts corresponding to backbone networks and PeriodNets. Furthermore, each part consists of seven regions arranged horizontally, corresponding to the results at different noise levels. Thus, Fig. 5 has no "horizontal axis" in the strict sense, and the seven regions belonging to the same part share the same vertical axis and legend. For comparison purposes, the graphs belonging to a backbone network and its corresponding PeriodNet are drawn in the same color, and the ordinate interval of the figure is limited to [0.2, 1].

As shown in Fig. 5, the graph usually consists of four parts: the box area, outliers, the horizontal line, and vertical lines. The box area covers the upper (0.75) and lower (0.25) quartiles, and the distance between its vertical edges is the interquartile range (IQR). The outlier may not exist in the graph, which values more or less than 1.5 times the IQR of the vertical box edges. At the same time, the horizontal line inside the box corresponds to the median, while the vertical line connects the upper or lower quartile to the nonoutlier maximum or minimum.

Comparing both parts of Fig. 5, it is evident that PeriodConv has improved its performance significantly and made remarkable achievements in reducing its instability. For example, graphs of models M2 and M4 at all noise levels contain

TABLE VI
COMPARATIVE RESULTS UNDER DIFFERENT NOISE TYPES

		Backbone Network					PeriodNet Version				
	Noise Type	pink	white	brown	blue	purple	pink	white	brown	blue	purple
Mean	M0	0.9454	0.7775	**0.9900**	0.7046	0.7289	0.9492	0.9380	0.9543	0.9556	0.9409
	M1	**0.9727**	**0.9152**	0.9814	**0.8778**	0.8947	0.9681	0.9637	0.9837	0.9817	0.9810
	M2	0.9421	0.6745	0.8596	0.7730	0.8167	0.9626	0.9498	0.9769	0.9613	0.9680
	M3	0.4690	0.4542	0.4405	0.4308	0.4389	0.9643	0.9415	0.9793	0.9657	0.9789
	M4	0.8370	0.6738	0.9006	0.6284	0.6207	0.9230	0.9264	0.9443	0.9157	0.9186
	M5	0.8374	0.8125	0.9175	0.7804	**0.9022**	**0.9814**	**0.9810**	0.9821	**0.9920**	**0.9917**
	M6	0.8523	0.5310	0.8540	0.6669	0.7039	0.9777	0.9666	**0.9857**	0.9790	0.9856
Std	M0	**0.0305**	0.0744	**0.0087**	0.1152	0.1392	0.0124	0.0280	0.0115	0.0212	0.0190
	M1	0.0317	**0.0497**	0.0094	0.0550	0.0829	**0.0047**	0.0136	0.0089	0.0106	0.0083
	M2	0.0306	0.1078	0.1054	0.0951	0.0969	0.0111	0.0144	0.0096	0.0137	0.0097
	M3	0.0706	0.0559	0.1539	0.0974	**0.0540**	0.0189	0.0235	0.0068	0.0097	0.0088
	M4	0.0419	0.1012	0.0450	0.0989	0.0867	0.0101	0.0183	0.0156	0.0201	0.0112
	M5	0.2169	0.0874	0.1374	0.1112	0.0758	0.0080	**0.0065**	0.0065	**0.0042**	**0.0053**
	M6	0.0381	0.0691	0.0310	**0.0424**	0.0601	0.0068	0.0070	**0.0054**	0.0091	0.0065

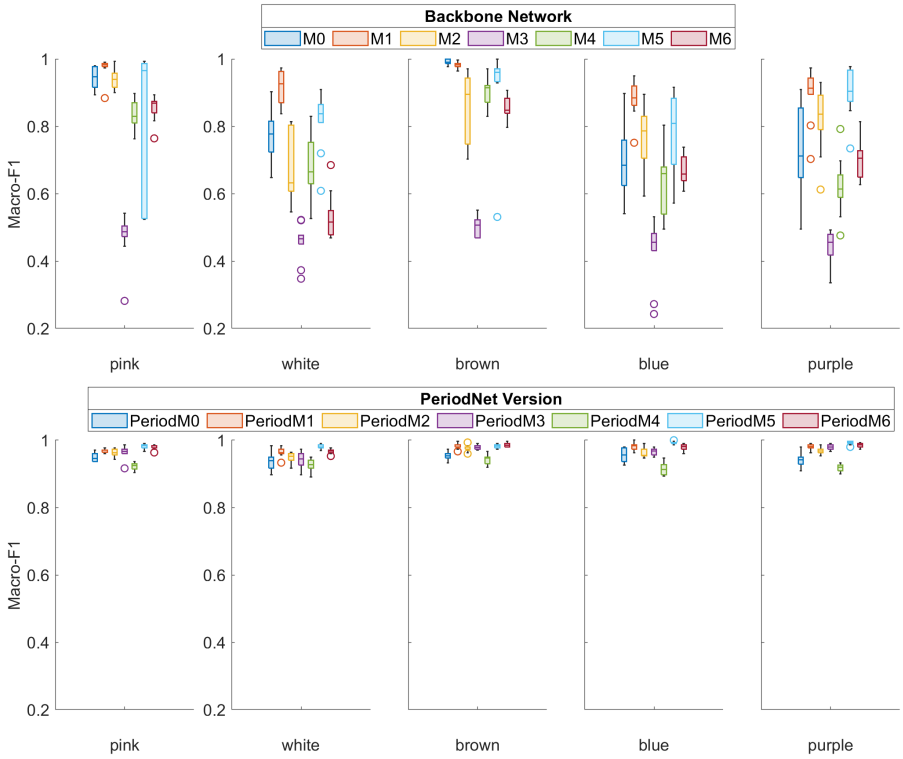

Fig. 6. Comparative results under different noise types.

long lines, indicating the wild results fluctuations. Worse still, an undervalued outlier represented by the yellow circle appears in the result of model M2 on the original data. On the contrary, graphs of corresponding PeriodNets only have short lines and even shrunk to points when the noise strength is lower than level III. Moreover, the variances of backbone networks fluctuate wildly with the change in noise level, and there is no apparent correlation between them. However, Fig. 5 confirms that variances of PeriodNets increase with noise levels, which aligns with the common knowledge that stronger noises lead to greater uncertainty.

Furthermore, the model performance decreases as the noise level increases obviously, as shown in the lower part of Fig. 5. Although this trend also exists theoretically in its upper part, it is unobservable when the noise strength is below level V due to the masking effect of sharp fluctuations. Moreover, graphs of M3 contain many purple circles at different noise levels, while the appearance of PeriodConv significantly alleviates these poor performance and outlier issues. In other words, PeriodConv makes the performance of PeriodM3 as good as others.

C. Studies on Noise Type

Table VI summarizes the comparative results of all models on the OU dataset under different noise types. As shown in Table VI, even with the same noise level, noise types have a decisive influence on the interference in the bearing fault

diagnosis task. For instance, the model M6 performs 60% better in pink noise than in white noise, demonstrating the robustness of backbone networks to noise types.

However, although PeriodConv only analyses the white noise during the design, it helps backbone networks deal well with the interference of all noise types. Actually, PeriodNets perform better than white noise in other noise types. One possible reason for this phenomenon is the Gaussian property shared by different types of noise, making the white noise representative of all noises.

According to Table VI, models M1 and PeriodM5 have the highest mean values, which is consistent with Table V. Moreover, the Std values of model PeriodM5 reveal its pronounced advantage on the uncertainty. Even with fewer layers and a more straightforward structure, M5 seems a better backbone network to combine with PeriodConv, which could obtain better performance under the interference of noises. Therefore, blindly increasing neural network complexity may not improve bearing fault diagnosis performance. In contrast, implementing SP algorithms in the network design is an effective solution.

Fig. 6 follows the drawing pattern of Fig. 5, but it has five horizontal regions corresponding to the results under different noise types. The upper part of Fig. 6 illustrates the poor performance of M3 under all noise types, and it is even inferior to M4, which is a lightweight model with only 6% of its size and parameter volume. However, with the help of PeriodConv, PeriodM3 achieves excellent performance like other models.

As shown in Fig. 6, PeriodConv significantly improves the mean and reduces the variance of backbone networks in most cases, except for M0 under the brown noise. However, this is not enough to challenge the robustness of the proposed method, because the results of M0 under other noises show that its result under brown noise is an outlier. Actually, Fig. 6 reveals some correlations in graphs of the same backbone network under different noise types. For example, M4 has a long line under white and blue noises, and M5 has a large box area under pink and blue noises. Differently, M0 performs much better under brown noise than under others, making its superiority isolated and unforeseen, which also occurs on the CWRU dataset under the white noise, as shown in Table IV. On the contrary, PeriodNets achieve excellent performance under all noise types, highlighting the need for domain-oriented design for bearing fault diagnosis.

V. Conclusion

A robust end-to-end method, called PeriodNet, is proposed in this study for intelligent bearing fault diagnosis under varying speed conditions in noisy environments. Experiments on two-speed conditions, seven noise levels, and five noise types verify the effectiveness of PeriodNets. By inserting PeriodConv before backbone networks, PeriodNets significantly improve the performance of backbone networks in terms of the mean and standard deviation of Macro-$F1$ scores. At the same time, the performance of PeriodNets is less affected by noises than backbone networks under the interference of various noise levels and noise types, which verifies the robustness of PeriodNets. Moreover, case studies demonstrate that PeriodConv can significantly reduce the uncertainty of the backbone network selection, which indicates that PeriodNet has excellent generalizability.

However, PeriodNet inevitably has some shortcomings. This study mainly considers the interference of a single noise, while several types of noises may contaminate the collected vibration signals simultaneously in particular scenarios, such as noise attacks. At the same time, PeriodConv resamples the signals through the rectangular window and folds the feature maps to the channel dimension, which may limit the potential of the module design, especially the window length. For example, the existing CNNs can only handle feature maps of the same size among channels, leading to the constant window length for all window functions. Therefore, the proposed method could be further improved in the following two aspects.

1) Considering the coupling effect of multiple noises. Because the numerous types of noises may mix, causing different interfering or even modulating impacts on the fault signal components, PeriodNet may need to be updated according to the theoretical properties of multiple noises.
2) Improving the potential of the module design. More advanced window functions may be appropriately applied in PeriodConv to further extract valuable information from the signals. Recent advances in DL could also be introduced to automatically adapt the window lengths.

References

[1] H. Wang, Z. Liu, D. Peng, M. Yang, and Y. Qin, "Feature-level attention-guided multitask CNN for fault diagnosis and working conditions identification of rolling bearing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4757–4769, Sep. 2022.

[2] B. Cai et al., "Artificial intelligence enhanced two-stage hybrid fault prognosis methodology of PMSM," *IEEE Trans. Ind. Informat.*, vol. 18, no. 10, pp. 7262–7273, Oct. 2022.

[3] K. N. Gyftakis and A. J. Marques-Cardoso, "Reliable detection of very low severity level stator inter-turn faults in induction motors," in *Proc. 45th Annu. Conf. IEEE Ind. Electron. Soc.*, vol. 1, Oct. 2019, pp. 1290–1295.

[4] E. T. Chelmiah and D. F. Kavanagh, "Hilbert marginal spectrum for failure mode diagnosis of rotating machines," in *Proc. 47th Annu. Conf. IEEE Ind. Electron. Soc.*, Toronto, ON, Canada, Oct. 2021, pp. 1–6.

[5] X. Wang, H. Gu, T. Wang, W. Zhang, A. Li, and F. Chu, "Deep convolutional tree-inspired network: A decision-tree-structured neural network for hierarchical fault diagnosis of bearings," *Frontiers Mech. Eng.*, vol. 16, no. 4, pp. 814–828, Oct. 2021.

[6] W. Caesarendra and T. Tjahjowidodo, "A review of feature extraction methods in vibration-based condition monitoring and its application for degradation trend estimation of low-speed slew bearing," *Machines*, vol. 5, no. 4, p. 21, Sep. 2017.

[7] X. Kong et al., "Optimal sensor placement methodology of hydraulic control system for fault diagnosis," *Mech. Syst. Signal Process.*, vol. 174, Jul. 2022, Art. no. 109069.

[8] H. Li, Q. Zhang, X. Qin, and S. Yuantao, "Raw vibration signal pattern recognition with automatic hyper-parameter-optimized convolutional neural network for bearing fault diagnosis," *Proc. Inst. Mech. Eng., C, J. Mech. Eng. Sci.*, vol. 234, no. 1, pp. 343–360, Jan. 2020.

[9] M. Zhao, J. Lin, X. Wang, Y. Lei, and J. Cao, "A tacho-less order tracking technique for large speed variations," *Mech. Syst. Signal Process.*, vol. 40, no. 1, pp. 76–90, Oct. 2013.

[10] G. Xin, N. Hamzaoui, and J. Antoni, "Semi-automated diagnosis of bearing faults based on a hidden Markov model of the vibration signals," *Measurement*, vol. 127, pp. 141–166, Oct. 2018.

[11] G. Manhertz and A. Bereczky, "STFT spectrogram based hybrid evaluation method for rotating machine transient vibration analysis," *Mech. Syst. Signal Process.*, vol. 154, Jun. 2021, Art. no. 107583.

[12] Z. Mo, Z. Zhang, and K.-L. Tsui, "The variational kernel-based 1-D convolutional neural network for machinery fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–10, 2021.

[13] B. Zhao, C. Cheng, G. Tu, Z. Peng, Q. He, and G. Meng, "An interpretable denoising layer for neural networks based on reproducing kernel Hilbert space and its application in machine fault diagnosis," *Chin. J. Mech. Eng.*, vol. 34, no. 1, p. 44, May 2021.

[14] B. Cai et al., "Fault detection and diagnostic method of diesel engine by combining rule-based algorithm and BNs/BPNNs," *J. Manuf. Syst.*, vol. 57, pp. 148–157, Oct. 2020.

[15] M. Seera and C. P. Lim, "Online motor fault detection and diagnosis using a hybrid FMM-CART model," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 4, pp. 806–812, Apr. 2014.

[16] B. Cai et al., "Data-driven early fault diagnostic methodology of permanent magnet synchronous motor," *Expert Syst. Appl.*, vol. 177, Sep. 2021, Art. no. 115000.

[17] R. Ahmed, M. El Sayed, S. A. Gadsden, J. Tjong, and S. Habibi, "Automotive internal-combustion-engine fault detection and classification using artificial neural network techniques," *IEEE Trans. Veh. Technol.*, vol. 64, no. 1, pp. 21–33, Jan. 2015.

[18] T. Han, D. Jiang, Q. Zhao, L. Wang, and K. Yin, "Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery," *Trans. Inst. Meas. Control*, vol. 40, no. 8, pp. 2681–2693, May 2018.

[19] T. Berghout, L.-H. Mouss, O. Kadri, L. Saïdi, and M. Benbouzid, "Aircraft engines remaining useful life prediction with an adaptive denoising online sequential extreme learning machine," *Eng. Appl. Artif. Intell.*, vol. 96, Nov. 2020, Art. no. 103936.

[20] 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 Trans.*, vol. 128, pp. 470–484, Sep. 2022.

[21] Y. Lei, F. Jia, J. Lin, S. Xing, and S. X. Ding, "An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data," *IEEE Trans. Ind. Electron.*, vol. 63, no. 5, pp. 3137–3147, May 2016.

[22] A. Esteva et al., "A guide to deep learning in healthcare," *Nature Med.*, vol. 25, no. 1, pp. 24–29, Jan. 2019.

[23] R. Li, L. Zhuang, Y. Li, and C. Shen, "Intelligent bearing fault diagnosis based on scaled Ramanujan filter banks in noisy environments," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[24] C. Liu, G. Cheng, B. Liu, and X. Chen, "Bearing fault diagnosis method with unknown variable speed based on multi-curve extraction and selection," *Measurement*, vol. 153, Mar. 2020, Art. no. 107437.

[25] Z. Shang, Z. Zhao, Z. Zhou, C. Sun, Y. Sun, and R. Yan, "Denoising fused wavelets net for aeroengine bevel gear fault diagnosis," in *Proc. Int. Conf. Sens., Meas. Data Anal. Artif. Intell. (ICSMD)*, Oct. 2021, pp. 1–6.

[26] X. Su, H. Liu, L. Tao, C. Lu, and M. Suo, "An end-to-end framework for remaining useful life prediction of rolling bearing based on feature pre-extraction mechanism and deep adaptive transformer model," *Comput. Ind. Eng.*, vol. 161, Nov. 2021, Art. no. 107531.

[27] J. O. D. van den Hoogen, S. D. Bloemheuvel, and M. Atzmueller, "An improved wide-kernel CNN for classifying multivariate signals in fault diagnosis," in *Proc. Int. Conf. Data Mining Workshops (ICDMW)*, Sorrento, Italy, Nov. 2020, pp. 275–283.

[28] J. Lei, C. Liu, and D. Jiang, "Fault diagnosis of wind turbine based on long short-term memory networks," *Renew. Energy*, vol. 133, pp. 422–432, Apr. 2019.

[29] C. Wu, P. Jiang, C. Ding, F. Feng, and T. Chen, "Intelligent fault diagnosis of rotating machinery based on one-dimensional convolutional neural network," *Comput. Ind.*, vol. 108, pp. 53–61, Jun. 2019.

[30] C. Tian, Y. Xu, and W. Zuo, "Image denoising using deep CNN with batch renormalization," *Neural Netw.*, vol. 121, pp. 461–473, Jan. 2020.

[31] F.-L. Fan, J. Xiong, M. Li, and G. Wang, "On interpretability of artificial neural networks: A survey," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 5, no. 6, pp. 741–760, Nov. 2021.

[32] X. Zhao et al., "Intelligent fault diagnosis of gearbox under variable working conditions with adaptive intraclass and interclass convolutional neural network," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jan. 5, 2022, doi: 10.1109/TNNLS.2021.3135877.

[33] M. Zhao, S. Zhong, X. Fu, B. Tang, and M. Pecht, "Deep residual shrinkage networks for fault diagnosis," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4681–4690, Jul. 2020.

[34] Z. Zhao et al., "Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study," *ISA Trans.*, vol. 107, pp. 224–255, Dec. 2020.

[35] J. Yuan, S. Cao, G. Ren, H. Jiang, and Q. Zhao, "SGWnet: An interpretable convolutional neural network for mechanical fault intelligent diagnosis," in *Neural Computing for Advanced Applications*, H. Zhang, Z. Yang, Z. Zhang, Z. Wu, and T. Hao, Eds. Singapore: Springer, 2021, pp. 360–374.

[36] F. B. Abid, M. Sallem, and A. Braham, "Robust interpretable deep learning for intelligent fault diagnosis of induction motors," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 6, pp. 3506–3515, Jun. 2020.

[37] C. Liu, C. Qin, X. Shi, Z. Wang, G. Zhang, and Y. Han, "TScatNet: An interpretable cross-domain intelligent diagnosis model with antinoise and few-shot learning capability," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–10, 2021.

[38] T. Li et al., "WaveletKernelNet: An interpretable deep neural network for industrial intelligent diagnosis," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 52, no. 4, pp. 2302–2312, Apr. 2022.

[39] W. Fan, Y. Li, K. L. Tsui, and Q. Zhou, "A noise resistant correlation method for period detection of noisy signals," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2700–2710, May 2018.

[40] Y. Li, H. Zhao, W. Fan, and C. Shen, "Extended noise resistant correlation method for period estimation of pseudoperiodic signals," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[41] Y. Li, H. Zhao, W. Fan, and C. Shen, "Generalized autocorrelation method for fault detection under varying-speed working conditions," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[42] L. Durak and O. Arikan, "Short-time Fourier transform: Two fundamental properties and an optimal implementation," *IEEE Trans. Signal Process.*, vol. 51, no. 5, pp. 1231–1242, May 2003.

[43] K. Zhong, M. Han, T. Qiu, and B. Han, "Fault diagnosis of complex processes using sparse kernel local Fisher discriminant analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1581–1591, May 2020.

[44] H. Wang, J. Xu, R. Yan, C. Sun, and X. Chen, "Intelligent bearing fault diagnosis using multi-head attention-based CNN," *Proc. Manuf.*, vol. 49, pp. 112–118, Jan. 2020.

[45] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[46] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2536–2544.

[47] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[49] W. Zhang, X. Li, and Q. Ding, "Deep residual learning-based fault diagnosis method for rotating machinery," *ISA Trans.*, vol. 95, pp. 295–305, Dec. 2019.

[50] H. Huang and N. Baddour, "Bearing vibration data collected under time-varying rotational speed conditions," *Data Brief*, vol. 21, pp. 1745–1749, Dec. 2018.

[51] *Case Western Reserve University Bearing Data Center Website*. Accessed: Dec. 2021. [Online]. Available: https://engineering.case.edu/bearingdatacenter

[52] W. A. Smith and R. B. Randall, "Rolling element bearing diagnostics using the case western reserve university data: A benchmark study," *Mech. Syst. Signal Process.*, vols. 64–65, pp. 100–131, Dec. 2015.

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

[54] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

Ruixian Li received the B.S. and M.S. degrees from the Department of Industrial Engineering and Management, Shanghai Jiao Tong University, Shanghai, China, in 2019 and 2022, respectively. He is currently pursuing the Ph.D. degree with the Department of Industrial and Manufacturing Systems Engineering, The University of Hong Kong, Hong Kong, China.

His research interests include signal processing, deep learning, and reliability modeling.

Jianguo Wu (Member, IEEE) received the B.S. degree in mechanical engineering from Tsinghua University, Beijing, China, in 2009, the M.S. degree in mechanical engineering from Purdue University, West Lafayette, IN, USA, in 2011, and the M.S. degree in statistics and the Ph.D. degree in industrial and systems engineering from the University of Wisconsin–Madison, Madison, WI, USA, in 2014 and 2015, respectively.

He was an Assistant Professor with the Department of Industrial, Manufacturing and Systems Engineering (IMSE), The University of Texas at El Paso (UTEP), El Paso, TX, USA, from 2015 to 2017. He is currently an Assistant Professor with the Department of Industrial Engineering and Management, Peking University, Beijing. His research interests are focused on data-driven modeling, monitoring, and analysis of advanced manufacturing processes and complex systems for quality control and reliability improvement.

Yongxiang Li (Member, IEEE) received the Ph.D. degree in data science from the City University of Hong Kong, Hong Kong, in 2019.

He is currently an Associate Professor with the Department of Industrial Engineering and Management and the Chinese Institute for Quality Research, Shanghai Jiao Tong University, Shanghai, China. His research interests include the theoretical and applied aspects of data science integrated with domain knowledge for quality and reliability engineering using methodologies from statistics, machine learning, and signal processing. He has been working on the research areas, such as computer experiments, statistical quality control, statistical anomaly detection, and intelligent fault diagnostics.

Yao Cheng (Member, IEEE) received the Ph.D. degree in industrial and systems engineering from Rutgers University, New Brunswick, NJ, USA, in 2018.

She is currently an Assistant Professor with the Department of Industrial and Manufacturing Systems Engineering, The University of Hong Kong, Hong Kong. Her research interests lie in reliability and resilience modeling.