蘭州理工大學

科研成果汇总

学　　号： 221081101009

研 究 生： 强睿儒

导　　师： 霍海峰 教授　赵小强 教授

研究方向： 滚动轴承故障诊断

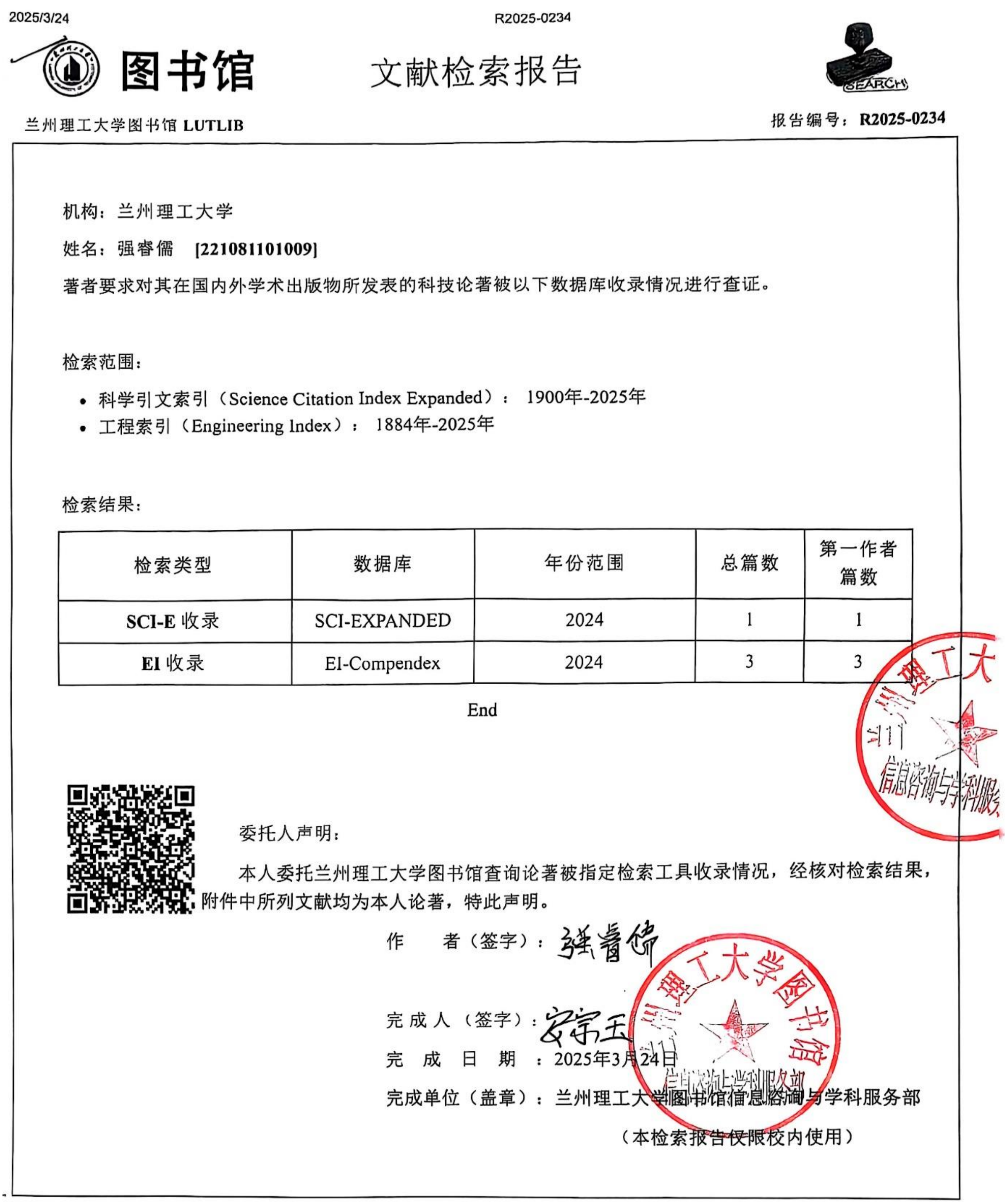
论文题目： 基于深度学习的不完备数据下滚动轴承故障诊断方法研究

学　　科： 控制理论与控制工程

学　　院： 电气工程与信息工程学院

入学时间： 2022 年 9 月

目录



2025/3/24 R2025-0234

图书馆

文献检索报告

兰州理工大学图书馆 LUTLIB

报告编号：**R2025-0234**

机构：兰州理工大学

姓名：强睿儒 **[221081101009]**

著者要求对其在国内外学术出版物所发表的科技论著被以下数据库收录情况进行查证。

检索范围：

- 科学引文索引（Science Citation Index Expanded）： 1900年-2025年
- 工程索引（Engineering Index）： 1884年-2025年

检索结果：

检索类型	数据库	年份范围	总篇数	第一作者篇数
SCI-E 收录	SCI-EXPANDED	2024	1	1
EI 收录	EI-Compendex	2024	3	3

End

委托人声明：

本人委托兰州理工大学图书馆查询论著被指定检索工具收录情况，经核对检索结果，附件中所列文献均为本人论著，特此声明。

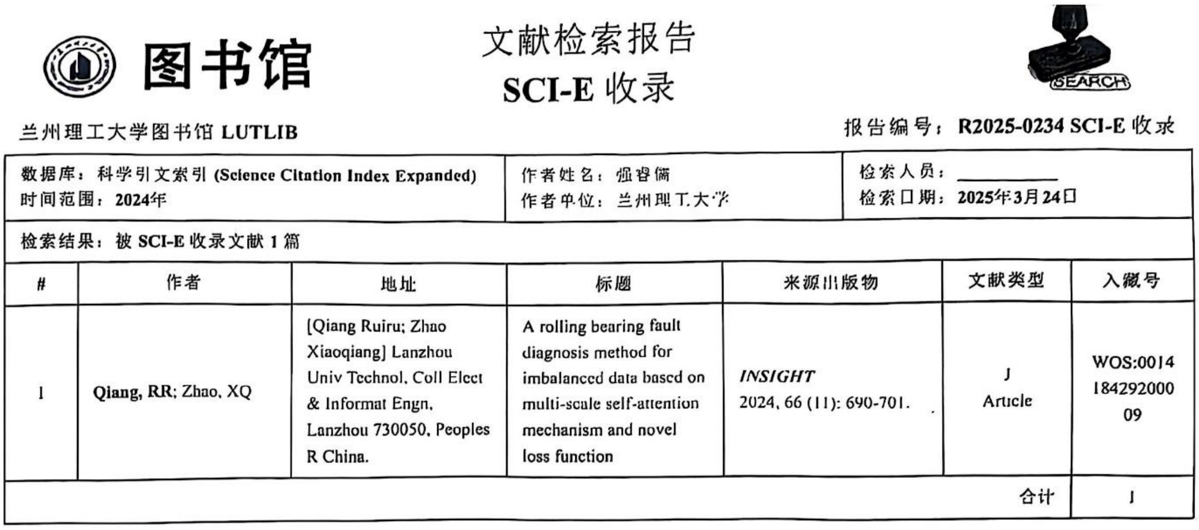
作　者（签字）：强睿儒

完成人（签字）：安宗玉

完　成　日　期　：2025年3月24日

完成单位（盖章）：兰州理工大学图书馆信息咨询与学科服务部

（本检索报告仅限校内使用）



图书馆

文献检索报告

SCI-E 收录

兰州理工大学图书馆 LUTLIB

报告编号：R2025-0234 SCI-E 收录

数据库：科学引文索引 (Science Citation Index Expanded) 时间范围：2024年	作者姓名：强睿儒 作者单位：兰州理工大学	检索人员：________ 检索日期：2025年3月24日

检索结果：被 SCI-E 收录文献 1 篇

#	作者	地址	标题	来源出版物	文献类型	入藏号
1	**Qiang, RR**; Zhao, XQ	[Qiang Ruiru; Zhao Xiaoqiang] Lanzhou Univ Technol, Coll Elect & Informat Engn, Lanzhou 730050, Peoples R China.	A rolling bearing fault diagnosis method for imbalanced data based on multi-scale self-attention mechanism and novel loss function	*INSIGHT* 2024, 66 (11): 690-701.	J Article	WOS:001418429200009
					合计	1

图书馆

文献检索报告
EI 收录

兰州理工大学图书馆 LUTLIB　　　　报告编号：**R2025-0234 EI 收录**

数据库：工程索引 **(Engineering Index)** 时间范围：**2024**年	作者姓名：强睿儒 作者单位：兰州理工大学	检索人员：__________ 检索日期：**2025**年**3**月**24**日

检索结果：被 **EI** 收录文献 **3** 篇

#	作者	地址	标题	来源出版物	文献类型	入藏号
1	**Ruiru, Qiang**; Xiaoqiang, Zhao	College of Electrical Engineering and Information Engineering, Lanzhou University of Technology, Gansu, Lanzhou	An Intelligent diagnosis method for rolling bearings based on Ghost module and adaptive weighting module	***Multimedia Tools and Applications*** 2024.	Article in Press	20242916727846
2	**Qiang, Ruiru**; Zhao, Xiaoqiang	College of Electrical and Information Engineering, Lanzhou University of Technology, Gansu, Lanzhou	A Small Sample Rolling Bearing Fault Diagnosis Method Based on Gramian Angular Difference Field and Generative Adversarial Network 基于格拉姆角差场和生成对抗网络的小样本滚动轴承故障诊断方法	***Huanan Ligong Daxue Xuebao/Journal of South China University of Technology (Natural Science)*** 2024, 52 (10): 64-75.	Journal article (JA)	20244517316686
3	**Ruiru, Qiang**; Xiaoqiang, Zhao	The College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou, Gansu	A rolling bearing fault diagnosis method for imbalanced data based on multi-scale self-attention mechanism and novel loss function	***Insight: Non-Destructive Testing and Condition Monitoring*** 2024, 66 (11): 690-701.	Journal article (JA)	20245017509757
					合计	3

DOI: 10.1784/insi.2024.66.11.690

A rolling bearing fault diagnosis method for imbalanced data based on multi-scale self-attention mechanism and novel loss function

Qiang Ruiru and Zhao Xiaoqiang

Deep learning methods are widely used in the field of rolling bearing fault diagnosis and produce good results when faced with datasets with roughly equal numbers of normal and faulty samples. However, real-world data often has a serious imbalance, with the number of fault samples being significantly less than the number of normal samples. This dataset imbalance challenges the performance of traditional deep learning methods. To address this problem, this paper proposes an efficient imbalanced data rolling bearing fault diagnosis method. The method consists of two parts: a deep learning network based on a multi-scale self-attention mechanism and a novel loss function. In terms of the deep learning network, firstly, the one-dimensional vibration signal is converted into a two-dimensional image through the Gramian angular field. This conversion maximises the inherent feature extraction capability of the network. Subsequently, the multi-scale learning capability of the network is enhanced by implementing different expansion rates for the head of the multi-scale self-attention mechanism. This nuanced approach allows the network to capture the underlying information more efficiently. Finally, the inclusion of Ghost bottlenecks and feature pyramid networks (FPNs) helps to optimise network efficiency and improve generalisation performance. A novel loss function is also proposed to make the method more suitable for imbalanced data. During the training process, the classification of samples in each class is analysed using the recall metric of imbalanced classification and the real-time recall is used as a weight to weaken the dominance of the majority class. This weighting ensures the adaptability of the method to imbalanced datasets. The proposed method is evaluated using rolling bearing datasets from Case Western Reserve University, USA, and Southeast University, China. Comparison results with other state-of-the-art deep learning methods show that the proposed method has a robust performance when dealing with imbalanced data.

Keywords: deep learning, rolling bearing fault diagnosis, imbalanced data, multi-scale self-attention mechanism, novel loss function.

1. Introduction

The evolution of modern industry has led to an escalating complexity in large industrial equipment[1]. Rotating machinery, a pivotal component in various pieces of large industrial equipment, is particularly crucial. The malfunction of rotating machinery can lead to production halts, escalated costs and potentially catastrophic accidents[2]. Thus, the efficiency of mechanical system maintenance can be greatly enhanced by investigating failures in rotating machinery. Timely fault detection and effective fault diagnosis not only prevent the exacerbation of faults but also reduce equipment downtime, thereby improving overall productivity[3]. Among the core components of rotating machinery, rolling bearings directly influence the proper functioning of the machinery. Consequently, research on fault diagnosis for rolling bearings holds immense significance in modern industry[4].

The fault diagnosis methods employed for rolling bearings can be broadly categorised into model-based and data-driven methods[5,6]. Model-based methods necessitate substantial expert knowledge for analysing failure mechanisms and constructing analytical models. However, the intricate operating conditions of rolling bearings make it challenging to establish specific models[7]. In contrast, data-driven methods for rolling bearing fault diagnosis, relying on data collected by sensors, have shown superior diagnostic outcomes without the need for *a priori* knowledge[8,9]. Data-driven diagnostic methods based on machine learning typically utilise techniques such as support vector machines[10] and extreme learning machines[11]. In recent years, the field of data-driven diagnostics has witnessed increased attention towards rolling bearing fault diagnosis methods based on deep learning techniques. The robust feature learning capabilities of deep learning methods enable effective extraction of intricate fault features, adapting well to large-scale datasets[12].

Various researchers have proposed innovative deep learning models for rolling bearing fault diagnosis. For instance, Jin *et al*[13] introduced a multi-layer adaptive convolutional neural network (CNN) with improved adaptive capabilities through multi-scale convolution and adaptive batch normalisation. Li *et al*[14] designed a recurrent neural network (RNN) with pooled bi-level attention, incorporating an attention mechanism for enhanced

● *Submitted 29.02.24 / Accepted 28.08.24*

Qiang Ruiru and Zhao Xiaoqiang are with the College of Electrical and Information Engineering, Lanzhou University of Technology, Gansu, Lanzhou 730050, China. Email: 853924752@qq.com / xqzhao@lut.edu.cn

fault classification. Mao *et al*[15] proposed a deep autoencoder that fused discriminative information of multiple fault types, introducing a relationship matrix of fault types for improved structural representation. The self-attention mechanism, acclaimed for its potent global learning ability, and Transformer have been introduced into fault diagnosis. Ding *et al*[16] proposed an end-to-end fault diagnosis framework based on time-frequency Transformer, addressing the limitations of traditional convolutional kernel recursive structures. Xu and Zhang[17] introduced a rolling bearing fault diagnosis method based on the Transformer encoder structure with one-dimensional vision, enabling end-to-end fault diagnosis by directly inputting raw one-dimensional data into the model.

Despite the successes of using deep learning for rolling bearing fault diagnosis, numerous challenges persist. Existing deep learning-based methods often require a substantial quantity of high-quality labelled data, with the number of fault state data samples nearly equalling that of normal state data samples[18]. However, in industrial production reality, rolling bearings spend the majority of their time in normal working conditions, with failure periods being relatively short. Consequently, the number of normal state samples far exceeds that of fault samples in realistically collected bearing operational data. While a large number of labelled normal samples can be accurately identified, the diagnostic outcomes for a smaller number of faulty data samples are often unsatisfactory[19,20]. As such, improving the performance of fault diagnosis methods for imbalanced data becomes a pressing challenge.

In the realm of rolling bearing fault diagnosis, addressing imbalance problems generally falls into two categories. The first category focuses on data preprocessing methods, aiming to alter the imbalance of the dataset at the data level[21]. This involves algorithmically generating artificial samples of minority class samples[22] or reducing the number of samples from the majority class to achieve equilibrium[23]. Zhou *et al*[24] and others utilised generative adversarial networks (GANs) to address data imbalance by generating additional fault samples. Han *et al*[25] effectively removed noisy samples by generating artificial samples using an improved synthetic minority over-sampling technique (SMOTE) algorithm. However, this category, though effective, consumes considerable time in sample generation and is prone to overfitting through the SMOTE algorithm[26,27]. The second category involves introducing a penalty factor to the algorithm to undermine the dominance of the majority class samples by adjusting the cost of different misclassifications[28,29]. However, cost-sensitive approaches often struggle to accurately define classification error costs and lack an efficient way of evaluating the performance of cost-sensitive classifiers. Consequently, effectively extracting features from minority class samples in an imbalanced dataset and developing fault diagnosis methods targeting imbalanced data pose significant challenges[30].

In summary, while progress has been made in addressing imbalanced sample issues in fault diagnosis, several challenges persist. To address the imbalanced data in rolling bearing fault diagnosis effectively, this paper proposes an imbalanced data rolling bearing fault diagnosis method based on a multi-scale self-attention mechanism and a novel loss function. The proposed method transforms one-dimensional vibration signals into two-dimensional images, which efficiently extracts shallow features with a multi-scale self-attention mechanism. A Ghost bottleneck is used to reduce the network computations and enhance network generalisation by connecting different layers of features through a feature pyramid network (FPN). The novel loss function aims to improve the performance of handling imbalanced data by diminishing the dominance of majority class samples and simultaneously boosting the confidence of minority class samples.

The main contributions of this paper include:

- Enhancement of the network's multi-scale learning ability through a multi-scale inflationary self-attention (MSIA) mechanism, addressing the deficiency of traditional self-attention mechanisms in multi-scale learning for shallow features and reducing redundant computations.
- Achievement of efficient network operation by introducing a Ghost bottleneck, which reduces model parameters and operations, thereby improving diagnostic speed. The use of a feature pyramid network connects low-level and high-level features, enhancing network generalisation and ensuring diagnostic accuracy.
- Design of a novel loss function to mitigate the dominance of majority class samples by dynamically adjusting the weights of the geometric mean confidence of each class. This ensures the network is unbiased in training and more adaptable to complex sample distributions.

The subsequent sections of this paper are structured as follows: Section 2 provides background knowledge, Section 3 details the proposed fault diagnosis methodology for imbalanced data, Section 4 validates the methodology using rolling bearing datasets and Section 5 concludes the paper.

2. Related work

2.1 Self-attention mechanism and vision transformer

In 2017, to address the limitations of traditional RNNs and CNNs in handling long sequential data within natural language processing (NLP), Vaswani *et al*[31] introduced a groundbreaking structure in their paper, titled: 'Attention is all you need': namely the self-attention mechanism. The self-attention mechanism, which is capable of comprehensively understanding each position within input data while simultaneously considering information from other positions in the sequence (*ie* emphasising global data information), proves exceptionally effective in grasping the nuances between distant words. As self-attention continues to demonstrate remarkable performance in various NLP tasks, some researchers have shifted their focus towards its applications in computer vision. Remarkably, self-attention has yielded outstanding results in image recognition, classification and image super-resolution reconstruction[32-34].

The structure of the conventional self-attention mechanism is shown in Figure 1. Self-attention mechanisms are good at solving the problem of feature interaction of input data at different spatial locations, leading to a better understanding of contextual information. Given a 2D feature map X of size $H \times W \times C$ (H: height, W: width, C: number of channels) as input, X is converted to a key $K = XW_k$, query $Q = XW_q$ and value $V = XW_v$ by embedding matrices W_k, W_q and W_v. Notably, the implementation of the embedding matrix is carried out using a 1 × 1 convolution in concrete operation. After that, the local relationship matrix $R_l \in^{-H \times W \times (n \times n \times C_h)}$ between the key K and the query Q is given by Equation (1):

$$R_l = K \otimes Q \qquad (1)$$

where C_h is the number of heads and ⊗ denotes the local matrix multiplication operation used to compute the pairwise relationship between each query Q and the corresponding key K in the local $n \times n$ grid in space, where n is the size of the grid in K. Thus, each feature R_i in the ith spatial location of the matrix R_i can be represented by a vector of size $n \times n \times C_h$ and all consist of maps of relationships (size: $n \times n$) between all the heads C_h, local queries Q and keys K. R_i further enriches the location information for each $n \times n$ grid:

$$\widehat{R} = R + P \otimes Q \text{ .... } (2)$$

where P denotes the 2D relative position embedding within each $n \times n$ grid and is shared among all heads. Next, as shown in Equation (3), the enhanced local relationship matrix $\widehat{R}$ is normalised by applying the Softmax function to the channel dimensions of each head to obtain the attention matrix A_m.

$$A_m = \text{Softmax}(\widehat{R}) \text{ .... } (3)$$

Finally, the feature vectors at each spatial location of A_m are reshaped into C_h local attention matrices and aggregated with all values within each $n \times n$ grid to obtain the final output feature map Y ($H \times W \times C$).

$$Y = V \otimes A_m \text{ .... } (4)$$

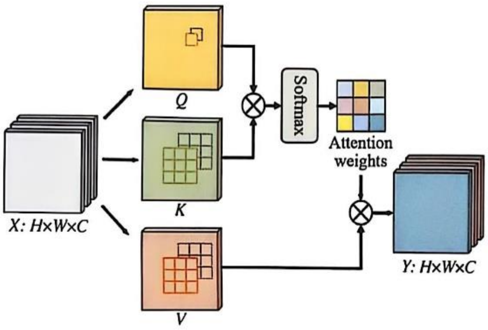


Figure 1. Self-attention mechanism

2.2 Ghost bottleneck

The Ghost module is a novel and efficient neural network proposed by Han *et al*[35] in 2020. It aims to solve the problems of feature redundancy with highly similar feature maps and an overly large number of parameters when using mainstream CNNs to extract target features. This is because redundant feature maps are unavoidable in convolutional operations and neural networks also need redundant feature maps to fully understand the input data. Rather than circumventing redundant feature maps, in the case of the Ghost module it is believed that it is better to use them to reduce the amount of computation. The ordinary convolution operation is shown in Figure 2(a).

Let the dimensions of the input data X be: $H \times W \times C$; then, the output obtained after an ordinary convolution operation is as follows:

$$Y = X * f + b \text{ .... } (5)$$

where * is the convolution operation, f is the convolution filter in the convolution layer, b is the bias in the convolution and $Y \in^{-H' \times W' \times C}$ is the output feature map with C channels, where H' and W' are the height and width of the output data, respectively. Let k be the size of the convolution kernel and n be the number of convolution kernels, then the number of floating-point operations per second (FLOPS) in the convolution process can be calculated as:

$$\text{FLOPS} = n \cdot H' \cdot W' \cdot C \cdot k \cdot k \text{ .... } (6)$$

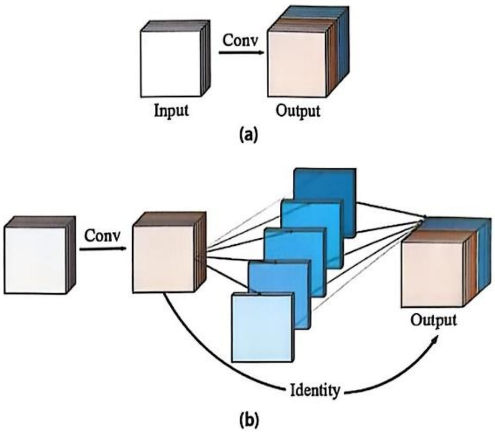


Figure 2. Conventional convolution and Ghost module: (a) conventional convolution; and (b) Ghost module

Unlike the direct generation of feature maps using ordinary convolution shown in Figure 2(a), the Ghost module first generates m intrinsic feature maps by conventional convolution. Then, by performing s cheap linear operations on the intrinsic feature maps, feature maps that are similar to each other are generated and called Ghost feature maps. Finally, the intrinsic feature maps are fused with the Ghost feature map as the new output. In this way, the Ghost module can provide an output feature map of the same size as the convolution with a lower number of parameters and reduced computation time. The number of FLOPS required for the Ghost module can be calculated as:

$$\text{FLOPS} = \frac{n}{s} \cdot H' \cdot W' \cdot C \cdot k \cdot k + (s-1) \cdot \frac{n}{s} \cdot H' \cdot W' \cdot d \cdot d \text{ ... } (7)$$

where d is the kernel size for linear operations, $n = m \times s$ and the speed-up ratio of the Ghost module with respect to ordinary convolution is:

$$\begin{aligned} ratio_s &= \frac{n \cdot H' \cdot W' \cdot C \cdot k \cdot k}{\frac{n}{s} \cdot H' \cdot W' \cdot C \cdot k \cdot k + (s-1) \cdot \frac{n}{s} \cdot H' \cdot W' \cdot d \cdot d} \\ &= \frac{n \cdot k \cdot k}{\frac{1}{s} \cdot C \cdot k \cdot k + \frac{s-1}{s} \cdot d \cdot d} \approx \frac{s \cdot C}{s + C - 1} \approx s \end{aligned} \text{ ... } (8)$$

The model parameter compression ratio is:

$$ratio_p = \frac{n \cdot C \cdot k \cdot k}{\frac{n}{s} \cdot C \cdot k \cdot k + \frac{s-1}{s} \cdot d \cdot d} \approx \frac{s \cdot C}{s + C - 1} \approx s \text{ ... } (9)$$

The Ghost bottleneck consists of the Ghost module, as shown in Figure 3.

2.3 Feature pyramid network

FPNs[36] are commonly used in the field of target detection. Compared to a traditional pyramid network, the FPN can fuse feature maps with strong low-resolution semantic information and high-resolution feature maps with weak semantic information but rich spatial information with less computational increase. This allows the features with high resolution and high semantic information to be obtained at the same time, improving the network's perceptual ability and detection accuracy. In addition, the FPN structure is simple and easy to implement and is able to match the backbone network fusion feature map with a small amount of computation to enhance the network feature representation capability. The FPN fuses feature maps of different resolutions by means of top-down paths and lateral connections to form a multi-scale feature pyramid, as shown in Figure 4.

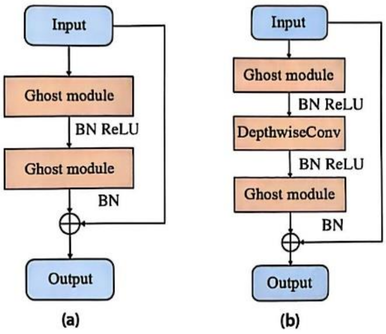


Figure 3. Ghost bottleneck for different stride lengths: (a) stride = 1; and (b) stride = 2

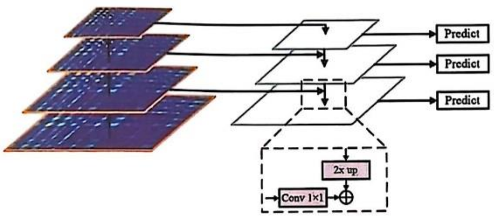


Figure 4. FPN structure

3. The proposed method

In this section, the proposed method is presented in detail in three parts. Firstly, the MSIA mechanism is introduced. Then, the novel loss function (Rloss) is introduced. Finally, the overall framework of the fault diagnosis model for imbalanced data is presented.

3.1 One-dimensional data transformation based on Gramian angular field

In recent years, some scholars have proposed a variety of methods for converting one-dimensional signals into two-dimensional images. These methods often rely on expert experience and expertise while preserving fault characteristics[37]. This dependency limits the universality of these methods. In order to address these issues, a Gramian angular difference field (GADF)-based transformation[38] is used to convert 1D signals into images as a way of visualising 1D time-series.

Let $T = \{t_1, t_2, \ldots, t_n\}$ be a time-series with n samples. The GADF transform of T is divided into three steps, as follows:

Step 1: Scale the time-series by normalising the input time-series data to the range $[-1, 1]$:

$$\tilde{t}_i = \frac{t_i - \min(t)}{\max(t) - \min(t)}, \forall i \in \{1, \ldots, n\} \text{......} (10)$$

Step 2: The normalised and scaled time-series signals are transformed from Cartesian coordinates to polar coordinates; this transformation preserves the temporal information in the input signals and is calculated as follows:

$$\varphi_i = \arccos\left(\tilde{t}_i\right), \forall i \in \{1, \ldots, n\} \text{..........} (11)$$

where φ_i is the polar coordinate of the angle.

Step 3: Identify temporal correlations in different time intervals by calculating the delta function difference between the polar coordinates of each time point and encode them into the geometric structure of the Gramian matrix:

$$\begin{aligned} GADF_{i,j} &= \left[\sin\left(\varphi_i - \varphi_j\right)\right] = \\ &\sqrt{I - \tilde{T}'^2} \cdot \tilde{T} - \tilde{T}' \cdot \sqrt{I - \tilde{T}^2}, \forall i, j \in \{1, \ldots, n\} \end{aligned} \quad \text{...} (12)$$

where I is a unit row vector and $\tilde{T}'$ and $\tilde{T}$ denote different row vectors. The main diagonal of the matrix contains the raw values of the time-domain signals and the angle information. Using the main diagonal, the GADF transform reconstructs the time-series into high-level features similar to those used in deep learning and further transforms the Gramian matrix into an image. The GADF transform is non-parametric, does not require prior assumptions about the data distribution or model and is applicable to a wide range of time-series data. In addition, the GADF transform captures the non-linear relationships and temporal patterns in the pairs of data with good tableau capability and robustness.

3.2 MSIA mechanism

Due to the complex working environment of rolling bearings and variable working loads, the vibration signals of bearings often have multi-scale complexity. The traditional self-attention mechanism lacks a certain multi-scale learning ability and has excessive redundant computation in the shallow attention matrix. Therefore, an MSIA mechanism was designed for fault diagnosis, as shown in Figure 5.

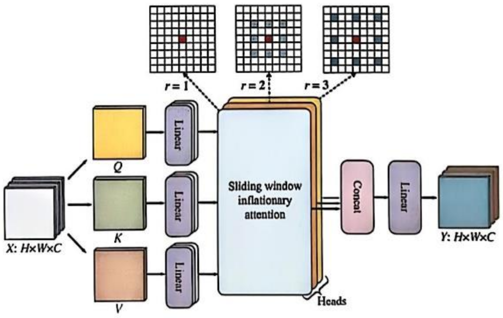


Figure 5. MSIA structure

Unlike a traditional self-attention mechanism, the MSIA mechanism is inspired by null convolution and enables the network to acquire multi-scale learning capabilities through the use of sliding window inflationary attention (SWIA). SWIA uses a sliding window of size $w \times w$ to sparsely select a key K and a value V centred on a position (i, j) in the original feature map and then performs self-attention. SWIA is described below:

$$Y = SWIA(Q, K, V, r) \text{..........................} (13)$$

where r is the inflation coefficient, which is used to control the sparsity, and Q is the query matrix. For the position (i, j), the component y_{ij} corresponding to the output Y of the SWIA operation is calculated as follows:

$$y_{ij} = Att\left(q_{ij}, K_r, V_r\right) = \text{Softmax}\left(\frac{q_{ij}K_r^T}{\sqrt{d_k}}\right)V_r,\ 1<i<W,\ 1<j<H \quad \text{.......... (14)}$$

where K_r and V_r denote the key K and the value V selected from the feature map and q_{ij} is a query positioned at coordinates (i, j). If the position (i, j) is given, the sliding window selects the key K and the value V at (i, j) for self-attention. SWIA simply makes use of the zero-padding strategy typically used in convolution to maintain the size of the feature map, thus querying the edges of the feature map. By sparsely selecting query-centric keys K and values V, SWIA can be made to satisfy locality and sparsity, thus effectively establishing remote dependencies.

In addition to this, unlike traditional self-attention mechanisms, for a given input X, the MSIA mechanism obtains the corresponding query Q, key K and value V by linear projection. The feature-mapped channels are then partitioned into Num_h different heads, and SWIA is executed in different heads using different inflation rates, thus taking full advantage of the sparsity of the self-attention mechanism at different scales. MSIA is described as follows:

$$h_i = SWIA\left(Q_i, K_i, V_i, r_i\right),\ 1 \le i \le Num_h \quad \text{....... (15)}$$

$$Y = Linear\left(Concat\left[h_1, \ldots, h_{num_h}\right]\right) \quad \text{.............. (16)}$$

where r_i is the inflation rate of the ith head and Q_i, K_i and V_i denote the feature map slices input to the ith head. The outputs of each head are concatenated together and then output to a linear layer for feature aggregation. By setting the inflation rate for different heads, the MSIA mechanism efficiently aggregates the input data to focus on the perceptual domains within different scales and effectively reduces redundant computations without adding additional computational cost.

3.3 Rloss

In traditional deep learning-based models for fault diagnosis, the standard cross-entropy (CE) loss function[39] is usually used for training. Higher accuracy can be achieved using CE when the amount of data for each type of fault is roughly equal to the amount of normal data. Let $\{d_n, l_n\} \in \{1, 2, \ldots, N\}$, where d_n are the training data and l_n are the labels corresponding to these data. The expression for CE is as follows:

$$loss_{CE} = -\sum_{n=1}^{N}\log\left(P_n^{l_n}\right) = -\sum_{c=1}^{C}\sum_{n:l_n=c}\log\left(P_n^{l_n}\right) = -\sum_{c=1}^{C}Num_c\log\left(P^c\right) \ldots \text{(17)}$$

where P_n denotes the prediction probability of the input data d_n on all categories and $P_n^{l_n}$ denotes the probability of the ith category. $P^c = \left(\prod_{n:l_n=c}P_n^{l_n}\right)^{\frac{1}{Num_c}}$ denotes the mean geometric confidence of category c and Num_c denotes the number of samples in category c.

In reality, however, the amount of fault data for rolling bearings is usually much smaller than the amount of normal data. When training data with imbalanced categories are present, using only the accuracy rate can no longer fully measure the diagnosis. Taking the binary classification problem as an example, the confusion matrix can be used to visualise the classification results, as shown in Table 1.

Table 1. Confusion matrix

Real situation	Positive	Negative
Positive	True positive (TP)	False negative (FN)
Negative	False positive (FP)	True negative (TN)

In Table 1, the TP and TN categories represent the samples with correct classification results and the FP and FN categories represent the samples with incorrect classification. On this basis, five classification metrics can be used, namely: accuracy (Acc), G-mean value (G), precision (P), recall (R) and F_1 value (F), which are calculated as follows:

$$Acc = \frac{TP+TN}{TP+FN+FP+TN} \quad \text{...................... (18)}$$

$$G = \sqrt{\frac{TP}{TP+FN}\cdot\frac{TN}{TN+FP}} \quad \text{..................... (19)}$$

$$P = \frac{TP}{TP+FP} \quad \text{................................ (20)}$$

$$R = \frac{TP}{TP+FN} \quad \text{................................ (21)}$$

$$F = \frac{2\times P\times R}{P\times R} \quad \text{.............................. (22)}$$

From Equation (18), it is easy to see that if the number of positive class samples is much larger than that of the negative class samples, the classification result of the negative class does not have a significant impact on Acc.

In fault diagnosis for imbalanced data, there is a large gap between the numbers of normal and faulty samples. Using Equation (17), CE uses the number of samples in each category as a weight, thus optimising the average geometric confidence for that category. This inevitably leads to a loss function that is biased in favour of a category when the Num_c of that category is large, which then leads to inaccurate model predictions. In such cases, the diagnostic results obtained from the continued use of CE would not be satisfactory. In order to deal with the above problem of multi-fault diagnosis with category imbalance, a CE-based variant of the function is proposed: the recall loss function (Rloss).

As can be seen from Equation (21), there is no FP in the denominator of R, thus providing an intuitive indication of whether a particular class of samples has been accurately classified. Badrinarayanan *et al*[40] demonstrated that the inverse of the frequency of occurrence of a certain type of sample can be used as a weight for CE. Inspired by them, the CE was weighted using R as a weight, as follows:

$$Rloss = -\sum_{c=1}^{C}\frac{FN_c}{FN_c+TP_c}Num_c\log\left(P^c\right) = -\sum_{c=1}^{C}\left(1-\frac{TP_c}{FN_c+TP_c}\right)Num_c\log\left(P^c\right) = -\sum_{c=1}^{C}\left(1-R_c\right)Num_c\log\left(P^c\right) \quad \text{... (23)}$$

where FN_c and TP_c are the FN and TP of the category c sample. In Rloss, the weights are defined as the R for that category. This is because in imbalanced classification problems, the majority class has a smaller FN and a larger R, so Rloss suppresses the gradient of the majority class samples. Conversely, the gradient of a minority class would be elevated. The weights take into account the classification performance of the network for each category in order to better handle category imbalance.

However, in network training, the weights of each class change following the update of network parameters. As a result, the instantaneous performance of the network also varies with the weights of each class. In order to react to the instantaneous performance of the network, a time factor is introduced as follows:

$$Rloss = -\sum_{c=1}^{C}\left(1-R_c\right)Num_c \log\left(P^c\right) = -\sum_{c=1}^{C}\sum_{n:l_i=c}\left(1-R_{c,t}\right)\log\left(p^{n,t}\right) \quad \ldots (24)$$

where $R_{c,t}$ is the recall of category c at moment t and $n:l_i = c$ is all samples labelled c. With Rloss, the weights provide a different loss contribution for each category, making the network more inclined to improve the poorly performing categories, thus improving the overall classification performance. The steps to implement Rloss are shown in Table 2.

It is worth mentioning that, in practice, a zero FN and TP for a sample type in a batch at the same time would result in a zero denominator for the weight calculation. To prevent this, $R_{c,t} = 0$ is made for this sample to increase the weight of the sample. The core idea of Rloss is to integrate the factors of true positives, false negatives and category imbalance to make the network better adapt to complex data distributions.

3.4 Efficient rolling bearing imbalanced data diagnosis framework based on the MSIA mechanism and Rloss

In summary, the proposed rolling bearing fault diagnosis framework for imbalanced data in this paper is shown in Figure 6. The steps are as follows:

Step 1: Acquisition of rolling bearing vibration signals.
Step 2: Conversion of the one-dimensional vibration signals into a two-dimensional image using the GADF transform and division of the data into a training set, testing set and validation set.
Step 3: Training of the network using the training strategy shown in Figure 6 and saving the network parameters that have the best performance on the test set.
Step 4: Validation of the trained network using the validation set data and outputting the diagnostic effect.

The network backbone integrates the MSIA mechanism and Ghost bottleneck. Initially, the input image is resized to 224 × 224 and the network achieves multi-scale learning by incorporating MSIA blocks in the lower layers. After effectively capturing low-level information, the features are downsampled once. Subsequently, a layer of 3 × 3 convolution, followed by two Ghost bottlenecks, is employed to reduce computation and extract advanced data information, generating high-level features. Feature fusion through an FPN enhances the generalisation and robustness of the network. Finally, a dropout operation mitigates overfitting after another layer of convolution and average pooling.

For network training, the parameters include a 30 epoch training cycle, an initial learning rate set to 0.001 and a batch size of 32. Firstly, the network parameters are initialised and then the data are used for training. The training process includes calculating the loss function, updating the weights by backpropagation using the Softmax classification function and optimising using the Adam optimiser. When the epoch reaches the specified number, the training ends and the network parameters are saved.

Table 2. Rloss implementation steps

Algorithm: Rloss	
Input:	Training data *D*, each corresponding to the true label *l*; the network outputs a probability distribution *p*
Output:	Rloss
Start	
Step 1:	Initialised number of categories
Step 2:	Determine the prediction category for each sample based on the network's output probability distribution *p*
Step 3:	Find the samples where the network prediction matches the true label *l* and save them in the set *TP*
Step 4:	Find unique category tags in the real label *l* and count the number of times they appear in the real label
Step 5:	Calculate the number of true positive samples for each category
Step 6:	Find the samples where the network prediction does not match the true label *l* and save them in the set *FN*
Step 7:	Find unique category labels in the samples that do not match the true labels and count the number of times they appear in the mismatched samples
Step 8:	Calculate the number of false negative samples for each category
Step 9:	Calculate $R_{c,t}$
Step 10:	Calculation of weights
Step 11:	Apply weights to the cross-entropy loss for each sample to obtain a weighted loss value
End	

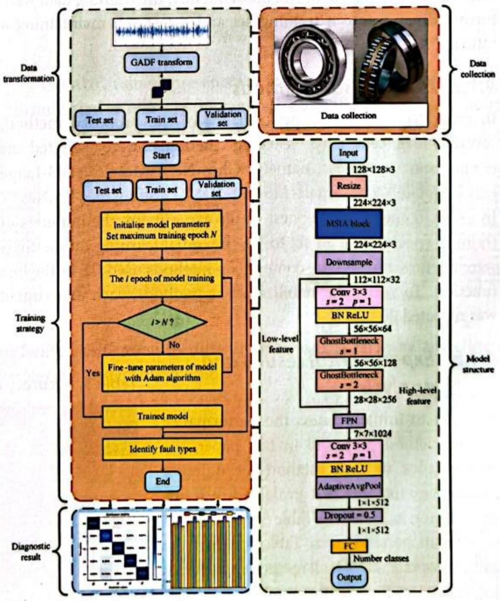


Figure 6. Schematic diagram of imbalanced rolling bearing fault diagnosis framework based on the multi-scale self-attention mechanism and novel loss function

4. Experiments and analyses

In order to verify the fault diagnosis performance of the method proposed in this paper for imbalanced data, fault diagnosis experiments were conducted using the Case Western Reserve University (CWRU), USA, bearing dataset[8] and the Southeast University, China, bearing dataset in[9]. The fault diagnosis capability of the method proposed in this paper for different imbalance rates was experimentally verified. To ensure the fairness of the experiments, all experiments were run on an AMD Ryzen7 5800H central processing unit (CPU) @ 3.2 GHz, Nvidia RTX 3060 (12 GB) with 16 GB random-access memory (RAM) and the framework used for the experiments was PyTorch1.12.

4.1 Case 1

4.1.1 Dataset description

The experimental dataset here originates from the rolling bearing test-bed at Case Western Reserve University, utilising an SKF 6205 bearing model. Bearing failures, induced artificially, are categorised into three types: ball failure (BF), inner ring failure (IF) and outer ring failure (OF). Additionally, fault points with diameters of 0.007", 0.014", 0.021" and 0.028" were machined using the electro-discharge method. These faults were further classified into 12 status labels based on diverse locations and sizes, as outlined in Table 3. Acceleration sensor data were collected at a sampling rate of 12 kHz under load conditions of 0 hp, 1 hp, 2 hp and 3 hp. For this study, the data under the 0 hp load condition were utilised. The number of sampling points was set to 128 and the normal state data were randomly selected and converted using a GADF to generate 1000 samples after sliding sampling. Fault data were selected based on varying imbalance rates, as detailed in Table 4. The dataset was then randomly divided into training data and a validation set in a ratio of 7:3. Subsequently, the training data were further divided into a training set and a test set, maintaining a ratio of 7:3.

4.1.2 Comparison methods

In order to validate the performance of the proposed method, several advanced deep learning methods were selected as comparison methods, namely GhostNet, MobileNetV3-Large (ML), MobileNetV3-Small (MS), ResNet-34 and ACmix-ResNet[41]. In order to ensure the fairness of the experiment, the numbers of training epochs were all set to 30, the initial learning rate settings were all the same and the comparison methods used CE as the loss function. To verify the stability of the methods, each experiment was repeated 20 times.

4.1.3 Experimental results and analysis

In order to initially assess the performance of the method proposed in this paper, the performance of each method for different imbalance rates was first evaluated in terms of accuracy, as shown in Table 5.

As can be seen from Table 5, when the imbalance rate was 1:1, all methods achieved a high accuracy because the number of samples of each type of faulty data was approximately equal to the number of samples of normal data. When the imbalance rate rose to 5:1, the accuracies of the comparison methods began to decline by varying degrees. This is due to the fact that the increase in the imbalance rate meant that the network did not learn enough about the minority type of samples and the network started to favour the majority type of samples. When the imbalance rate reaches 10:1 and 20:1, it can be seen that the classification accuracies of some of the comparison methods have fallen to the lowest level, with GhostNet even falling below 80%. At the same time, it can be seen that MS has the shortest diagnosis time of all models thanks to the excellent lightweighting. ResNet-34 and ACmix-ResNet have longer diagnosis times due to the deeper network and the complexity of the network structure due to the extensive use of the self-attention mechanism. The proposed approach did not have the shortest diagnosis time, due to the fact that it also uses a variant of the self-attention mechanism, but gives the best diagnosis results.

Table 3. Details of Case 1 data

Class label	Fault location	Fault size (in)
00	Normal	–
01	BF	0.007
02	IF	0.007
03	OF	0.007
04	BF	0.014
05	IF	0.014
06	OF	0.014
07	BF	0.021
08	IF	0.021
09	OF	0.021
10	BF	0.028
11	IF	0.028

Table 4. Imbalance rate division of data for Case 1

Number of normal data	Number of fault data	Imbalance rate
1000	1000	1:1
1000	500	2:1
1000	200	5:1
1000	100	10:1
1000	50	20:1
1000	20	50:1
1000	10	100:1

Table 5. Accuracy of each method at different imbalance rates

Imbalance rate	GhostNet	ML	MS	ResNet-34	ACmix-ResNet	Our method
1:1	97.50%	97.47%	97.02%	**98.26%**	97.65%	98.08%
2:1	96.88%	**98.15%**	96.41%	97.79%	98.00%	98.10%
5:1	88.75%	92.60%	89.17%	95.20%	95.31%	**97.92%**
10:1	78.42%	88.10%	82.38%	85.71%	88.41%	**98.25%**
20:1	77.42%	88.39%	87.53%	92.47%	93.55%	**98.27%**
50:1	87.43%	92.35%	91.80%	95.63%	92.62%	**99.18%**
100:1	92.49%	95.50%	94.29%	96.69%	96.99%	**99.09%**
Average time	384.79 s	400.39 s	**274.65 s**	1043.58 s	975.79 s	396.58 s

It is worth noting that as the imbalance rate rose still further, the data were already in a highly imbalanced state. At this point, the accuracies of the comparison algorithms started to gradually increase and even when the imbalance rate reached 100:1, some of the comparison methods, such as ResNet-34 and ACmix-ResNet, achieved roughly the same accuracy as when the imbalance rate was 1:1. However, this does not mean that the diagnostic results of both methods were excellent. To further assess the diagnostic effectiveness, the precision P, recall R, G-mean and F_1-value achieved by each method were calculated for comparison when the imbalance rate was higher than 5:1, as shown in Figure 7.

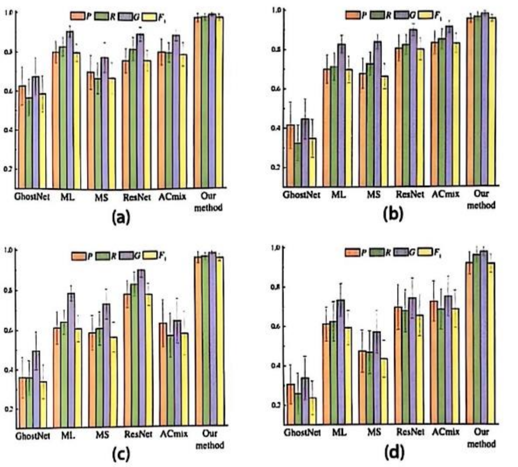


Figure 7. Failure classification metrics for different methods at different imbalance rates: (a) 10:1; (b) 20:1; (c) 50:1; and (d) 100:1

Combining Table 5 with Figure 7, it can be seen that the P, R, G and F_1 values obtained using the comparison methods were lower than 0.9 when the imbalance rate was 10 and the accuracies obtained using them were also lower than 90%. As the imbalance rate increased further, although the average accuracies of some of the comparison methods increased, the classification performance of the comparison methods gradually decreased, as can be seen from the remaining four metrics. This suggests that the comparison methods were unable to accommodate highly imbalanced data distributions and were less diagnostic. The proposed method showed a stable performance for different imbalance rates and even when the data were in a highly imbalanced state, the classification indices could reach more than 95%. This verifies that the proposed method not only has better feature learning ability but it also has better imbalanced classification ability after weighting the minority class samples using Rloss.

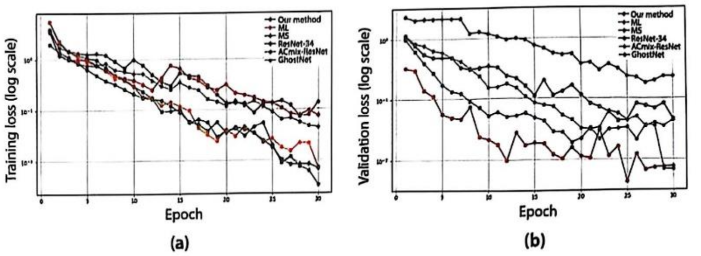


Figure 8. Comparison of loss functions for an imbalance ratio of 10:1: (a) training loss; and (b) validation loss

Meanwhile, taking the imbalance ratio of 10:1 as an example, the iterative curves of the loss function were plotted for different models, as shown in Figure 8. It can be seen that all the algorithms completed convergence within 30 epochs, among which the proposed method converged the fastest and had the lowest final loss value.

To further demonstrate the reliability of the proposed method, it was picked alongside GhostNet and ResNet-34, the receiver operating characteristic (ROC) curves[42] were plotted and the average area under the ROC curve (AUC) values was calculated to show the classification performance when the imbalance rate was higher than 5:1. When a certain class of samples are considered as a positive class of samples, the rest of the samples are all negative classes. ROC curves can only reflect a binary classification problem, so the false positive rate (FPR) was plotted as the horizontal coordinate of the ROC curve and the true positive rate (TPR) was plotted as the vertical coordinate (see Figures 9-12). These quantities were calculated as follows:

$$FPR = \frac{FP}{FP + TN} \quad \text{.............................. (25)}$$

$$TPR = \frac{TP}{TP + FN} \quad \text{.............................. (26)}$$

The AUC value represents the area under the ROC curve and is mainly used to measure the generalisation performance of a model.

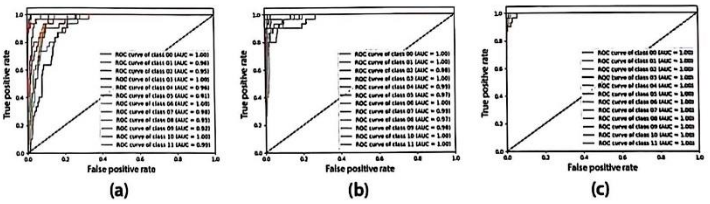


Figure 9. ROC curves for fault classification by different methods at an imbalance rate of 10:1: (a) GhostNet; (b) ResNet-34; and (c) our method

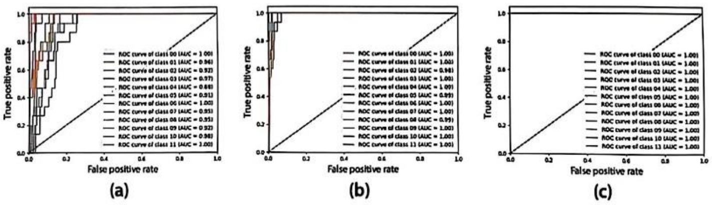


Figure 10. ROC curves for fault classification by different methods at an imbalance rate of 20:1: (a) GhostNet; (b) ResNet-34; and (c) our method

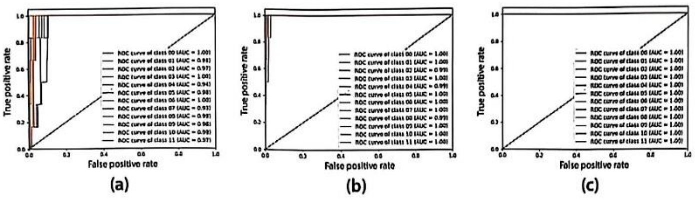


Figure 11. ROC curves for fault classification by different methods at an imbalance rate of 50:1: (a) GhostNet; (b) ResNet-34; and (c) our method

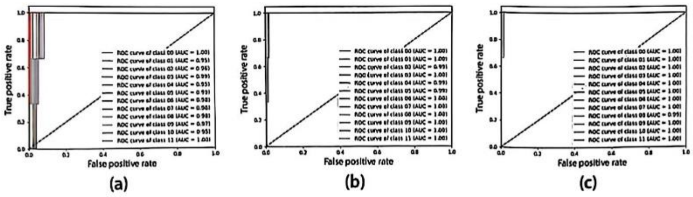


Figure 12. ROC curves for fault classification by different methods at an imbalance rate of 100:1: (a) GhostNet; (b) ResNet-34; and (c) our method

As shown in Figures 9-12, the AUC value takes the range of [0,1]. When 0.5 < AUC < 1, the classifier has some predictive value and the larger the AUC value, the higher the model reliability. When AUC = 0.5, it means that the classifier relies on random guesses for classification and the model has no predictive value. When AUC < 0.5, it means that the classifier is worse than a random classification and is not reliable. Since this case used ROC curves to judge the effectiveness of multi-class fault diagnosis, when a class was considered as a positive class, all the remaining classes were considered as negative. This operation produced 12 ROC curves for each method and the average AUC values for each method were calculated at different imbalance rates, as shown in Table 6.

Table 6. Mean AUC values for each method at different imbalance rates

Imbalance rate	GhostNet	ResNet-34	Our method
10:1	0.9689	0.9901	**0.9993**
20:1	0.9520	0.9956	**0.9998**
50:1	0.9762	0.9973	**0.9999**
100:1	0.9746	0.9967	**0.9995**

Combining Figures 9-12 with Table 6, this again proves that the proposed method can achieve better diagnostic results for fault diagnosis for imbalanced data. The method had the highest troubleshooting accuracy and the best stability for different imbalance rates. In summary, the proposed method has a significant advantage over advanced deep learning algorithms in terms of fault diagnosis performance in experiments with different imbalance rates.

4.2 Case 2

In order to verify the state-of-the-art of the proposed method, ResNet-34 was retained while also selecting two new rolling bearing fault diagnosis methods for imbalanced data as comparison methods. Among them is the generative adversarial network based on deep feature enhancement (DFEGAN)[43]; while TF-SAMB-NN[44] is an integrated multi-task rolling bearing diagnosis method based on representation learning for imbalanced sample conditions.

4.2.1 Data description

Experimental data were obtained from the driveline power simulator at the School of Mechanical Engineering, Southeast University, China[8]. The system operating condition is 20 Hz (1200 r/min)-0 V and there are five sample types, which are normal, rolling body failure (BF), compound failure (CF), inner ring failure (IF) and outer ring failure (OF), as shown in Table 7. As for Case 1, this case took 1000 samples of normal data and took different fault data according to imbalance rates of 10:1, 20:1 and 50:1. The training set, test set and validation set were divided in the same way as in Case 1.

4.2.2 Experimental results and analysis

The accuracy rates obtained for different imbalance rates are shown in Table 8. It can be seen that when the imbalance rate was 10:1, DFEGAN had the highest accuracy rate of 95%. When the imbalance ratio was 20:1, DFEGAN and the prosposed method achieved equal accuracies of 96.39%, which is significantly better than the other two compared methods. When the imbalance rate was 50:1, the prosposed method achieved the highest accuracy of 97.79%, which is higher than 97.22% for DFEGAN, 94.44% for ResNet-34 and 95.68% for TF-SAMB-NN.

Table 7. Details of Case 2 data

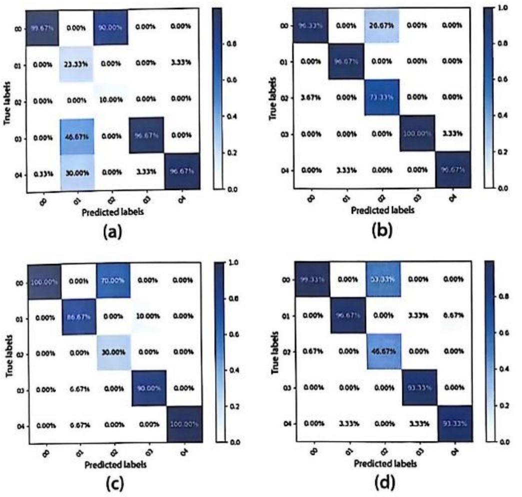
Class label	Fault location	Sample size (10:1)	Sample size (20:1)	Sample size (50:1)
00	Normal	1000	1000	1000
01	BF	100	50	20
02	CF	100	50	20
03	IF	100	50	20
04	OF	100	50	20

Table 8. Accuracy of each method at different imbalance rates

Imbalance ratio	ResNet-34	DFEGAN	TF-SAMB-NN	Our method
10:1	87.38%	**95.00%**	93.33%	94.52%
20:1	91.11%	**96.39%**	95.28%	**96.39%**
50:1	94.44%	97.22%	95.68%	**97.79%**
Average time	348.56 s	4256.78 s	213.74 s	**183.59 s**

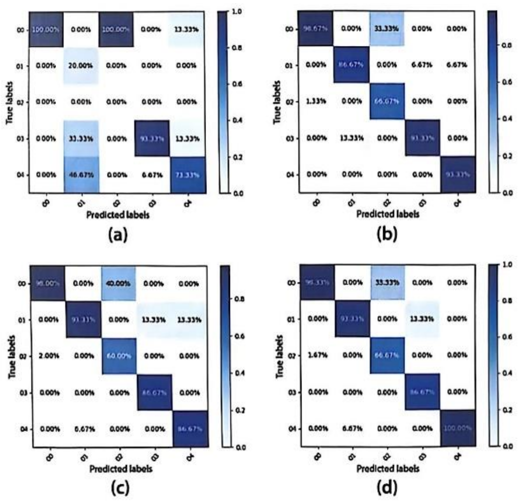
In order to analyse the classification results of all methods more intuitively, the fault diagnosis results were analysed by introducing confusion matrices, as shown in Figures 13-15, where the horizontal coordinates represent the predicted labels and the vertical coordinates represent the true labels. It can be seen that ResNet-34, trained on data with three different imbalance rates, had a large number of sample classification errors in the validation set, and the diagnostic results were unsatisfactory. Combining Figures 13-15 with Table 8, it can be seen that the remaining three methods performed better at the three imbalance rates. Among them, DFEGAN had the best diagnostic effect when the imbalance rate was 10:1. When the imbalance rate was 20:1, DFEGAN and the prosposed method were not only optimal in terms of accuracy, but there were three types of sample in the confusion matrix with classification accuracy higher than 90%. When the imbalance rate was 50:1, the prosposed method and DFEGAN had a classification accuracy higher than 95% for two types of sample and 80% for two types of sample. Compared to TF-SAMB-NN, a smaller classification error rate was achieved for the composite fault (CF) type. The excellent performance of DFEGAN was due to the large number of training samples generated by DFEGAN. However, training DFEGAN requires at least 100 epochs to be set up, which will take a lot of time and computer resources. In contrast, the prosposed method does not need to generate artificial samples and the time required for diagnosis is greatly reduced.

In summary, considering the time required for diagnosis and the diagnostic effect, the prosposed method has a certain degree of sophistication.



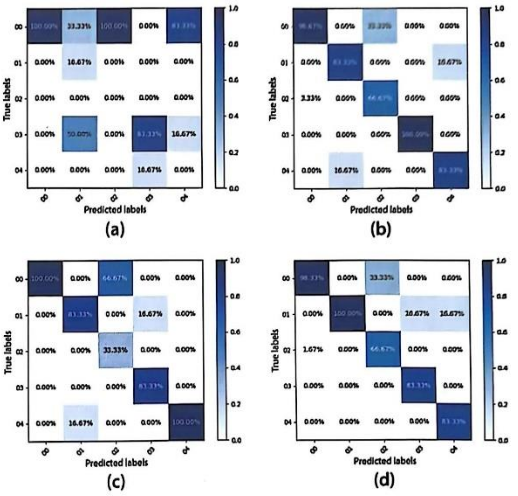
(a) (b) (c) (d)

Figure 13. Confusion matrices for fault classification by different methods at an imbalance ratio of 10:1: (a) ResNet-34; (b) DFEGAN; (c) TF-SAMB-NN; and (d) our method



(a) (b) (c) (d)

Figure 14. Confusion matrices for fault classification by different methods at an imbalance ratio of 20:1: (a) ResNet-34; (b) DFEGAN; (c) TF-SAMB-NN; and (d) our method



(a) (b) (c) (d)

Figure 15. Confusion matrices for fault classification by different methods at an imbalance ratio of 50:1: (a) ResNet-34; (b) DFEGAN; (c) TF-SAMB-NN; and (d) our method

4.3 Ablation experiments

In order to verify the impact of Rloss on the results in fault diagnosis, the experimental data from the first two cases was used, trained with CE and Rloss, respectively, so as to compare the impact of the two loss functions on the classification results. In order to ensure the stability of the experiment, each loss function was performed 20 times and averaged. The accuracy results are shown in Table 9 and the comparisons of precision rate *P* and other indices are shown in Tables 10-13.

Table 9. Comparison of accuracy achieved with different loss functions

Case_1 Data			
Imbalance rate	10:1	20:1	50:1
CE	94.13%	93.54%	95.35%
Rloss	97.93%	98.49%	98.91%
Case_2 Data			
Imbalance rate	10:1	20:1	50:1
CE	89.29%	90.83%	95.68%
Rloss	93.81%	95.56%	96.63%

It can be seen that when the network used CE as the loss function, the diagnostic effect was similar to that of ResNet-34 when combining the various classification metrics. This shows that the network is not inferior to good deep learning networks in terms of feature extraction capability. However, the use of CE makes the network ill-equipped to handle imbalanced data and, as the imbalance rate increases, the problem of degradation of classification performance arises. Better diagnostic results can be achieved when Rloss is used as the loss function. This again proves that the use of Rloss can indeed be effective in improving the classification performance of the network for imbalanced data, making the network better handle the data under realistic working conditions.

Table 10. Comparison of precision obtained with different loss functions

Case_1 Data			
Imbalance rate	10:1	20:1	50:1
CE	0.8972	0.8333	0.7639
Rloss	**0.9639**	**0.9611**	**0.9444**
Case_2 Data			
Imbalance rate	10:1	20:1	50:1
CE	0.8301	0.7962	0.5979
Rloss	**0.8567**	**0.926**	**0.8293**

Table 11. Comparison of recall achieved with different loss functions

Case_1 Data			
Imbalance rate	10:1	20:1	50:1
CE	0.9132	0.8512	0.8906
Rloss	**0.9657**	**0.9617**	**0.9554**
Case_2 Data			
Imbalance rate	10:1	20:1	50:1
CE	0.7891	0.8076	0.5786
Rloss	**0.8955**	**0.8709**	**0.7886**

Table 12. Comparison of G-mean values obtained with different loss functions

Case_1 Data			
Imbalance rate	10:1	20:1	50:1
CE	0.9501	0.9146	0.9367
Rloss	**0.9815**	**0.9797**	**0.9760**
Case_2 Data			
Imbalance rate	10:1	20:1	50:1
CE	0.8364	0.8713	0.6661
Rloss	**0.9359**	**0.9102**	**0.8539**

Table 13. Comparison of F_1 values obtained with different loss functions

Case_1 Data			
Imbalance rate	10:1	20:1	50:1
CE	0.8892	0.8304	0.7778
Rloss	**0.9639**	**0.9607**	**0.9450**
Case_2 Data			
Imbalance rate	10:1	20:1	50:1
CE	0.8093	0.7979	0.5798
Rloss	**0.8687**	**0.8878**	**0.7902**

5. Conclusion

The method proposed in this paper can be used for rolling bearing fault diagnosis for imbalanced data. One-dimensional vibration signals are converted into two-dimensional images using the GADF transform, thus taking full advantage of the feature extraction capability of convolutional neural networks. In the network design, a multi-scale inflationary attention mechanism is introduced to enhance the network's multi-scale learning ability for shallow features and reduce redundant computation. Then, a Ghost bottleneck module is added to ensure the network learning ability while reducing the number of network parameters and computing time. The generalisation of the network is then improved by connecting low-level features with high-level features through an FPN module. Finally, novel loss functions are proposed so that the network can be trained to handle imbalanced data more efficiently. The reliability of this method has been effectively verified using the CWRU dataset and the Southeast University dataset for experimental analyses in comparison with state-of-the-art deep learning methods. Future work will focus on the lightweighting of the model and the data imbalance between different working conditions.

Funding statement

This research work has been awarded by the National Natural Science Foundation of China (No 62263021) and the College Industrial Support Project of Gansu Province (2023CYZC-24).

Data availability

Data will be made available on request.

Acknowledgements

The authors would like to thank all those who have reviewed and contributed to this paper for their valuable assistance.

References

1. R Huang, J Xia, B Zhang *et al*, 'Compound fault diagnosis for rotating machinery: state-of-the-art, challenges and opportunities', Journal of Dynamics, Monitoring and Diagnostics, Vol 2, No 1, pp 13-29, 2023.
2. Y Qin, L Jin, A Zhang and B He, 'Rolling bearing fault diagnosis with adaptive harmonic kurtosis and improved bat algorithm', IEEE Transactions on Instrumentation and Measurement, Vol 70, pp 1-12, 3508112, 2021. DOI: 10.1109/TIM.2020.3046913
3. F Zhang, M Chen, Y Zhu *et al*, 'A review of fault diagnosis, status prediction and evaluation technology for wind turbines', Energies, Vol 16, No 3, 1125, 2023.
4. M Mohiuddin, M S Islam, S Islam *et al*, 'Intelligent fault diagnosis of rolling element bearings based on modified AlexNet', Sensors, Vol 23, No 18, 7764, 2023.
5. R Yuan, Y Lv, Z Lu *et al*, 'Robust fault diagnosis of rolling bearing via phase space reconstruction of intrinsic mode functions and neural network under various operating conditions', Structural Health Monitoring, Vol 22, No 2, pp 846-864, 2023.
6. P Liang, W Wang, X Yuan *et al*, 'Intelligent fault diagnosis of rolling bearing based on wavelet transform and improved ResNet under noisy labels and environment', Engineering Applications of Artificial Intelligence, Vol 115, 105269, 2022.
7. N Diao, Z Wang, H Ma *et al*, 'Fault diagnosis of rolling bearing under variable working conditions based on CWT and T-ResNet', Journal of Vibration Engineering and Technologies, Vol 11, No 8, pp 3747-3757, 2023.
8. W A Smith and R B Randall, 'Rolling element bearing diagnostics using the Case Western Reserve University data:

a benchmark study', Mechanical Systems and Signal Processing, Vol 64-65, pp 100-131, 2015.

9. S Shao *et al*, 'Highly accurate machine fault diagnosis using deep transfer learning', IEEE Transactions on Industrial Informatics, Vol 15, No 4, pp 2446-2455, 2018.
10. M Cui, Y Wang, X Lin *et al*, 'Fault diagnosis of rolling bearings based on an improved stack autoencoder and support vector machine', IEEE Sensors Journal, Vol 21, No 4, pp 4927-4937, 2020.
11. F Liu, H Wang, W Li *et al*, 'Fault diagnosis of rolling bearing combining improved AWSGMD-CP and ACO-ELM model', Measurement, Vol 209, 112531, 2023.
12. X Li, W Zhang, Q Ding *et al*, 'Multi-layer domain adaptation method for rolling bearing fault diagnosis', Signal Processing, Vol 157, pp 180-197, 2019.
13. T Jin, C Yan, C Chen *et al*, 'New domain adaptation method in shallow and deep layers of the CNN for bearing fault diagnosis under different working conditions', International Journal of Advanced Manufacturing Technology, Vol 124, No 11, pp 3701-3712, 2023.
14. J Li, Y Liu and Q Li, 'Intelligent fault diagnosis of rolling bearings under imbalanced data conditions using attention-based deep learning method', Measurement, Vol 189, 110500, 2022.
15. W Mao, W Feng, Y Liu *et al*, 'A new deep autoencoder method with fusing discriminant information for bearing fault diagnosis', Mechanical Systems and Signal Processing, Vol 150, 107233, 2021.
16. Y Ding, M Jia, Q Miao *et al*, 'A novel time-frequency Transformer based on self-attention mechanism and its application in fault diagnosis of rolling bearings', Mechanical Systems and Signal Processing, Vol 168, 108616, 2022.
17. P Xu and L Zhang, 'A fault diagnosis method for rolling bearing based on 1D-ViT model', IEEE Access, Vol 11, pp 39664-39674, 2023. DOI: 10.1109/ACCESS.2023.3268534
18. Q Hang, J Yang and L Xing, 'Diagnosis of rolling bearing based on classification for high-dimensional unbalanced data', IEEE Access, Vol 7, pp 79159-79172, 2019.
19. H Zhang, R Wang, R Pan *et al*, 'Imbalanced fault diagnosis of rolling bearing using enhanced generative adversarial networks', IEEE Access, Vol 8, pp 185950-185963, 2020.
20. Y Gao, L Gao, X Li *et al*, 'A hierarchical training-convolutional neural network for imbalanced fault diagnosis in complex equipment', IEEE Transactions on Industrial Informatics, Vol 18, No 11, pp 8138-8145, 2022.
21. X Liu, W Sun, H Li *et al*, 'Imbalanced sample fault diagnosis of rolling bearing using deep condition multi-domain generative adversarial network', IEEE Sensors Journal, Vol 23, No 2, pp 1271-1285, 2022.
22. Y Yu, L Guo, H Gao *et al*, 'PCWGAN-GP: a new method for imbalanced fault diagnosis of machines', IEEE Transactions on Instrumentation and Measurement, Vol 71, 3515711, 2022.
23. W C Lin, C F Tsai, Y H Hu *et al*, 'Clustering-based undersampling in class-imbalanced data', Information Sciences, Vol 409-410, pp 17-26, 2017.
24. F Zhou, S Yang, H Fujita *et al*, 'Deep learning fault diagnosis method based on global optimisation GAN for unbalanced data', Knowledge-Based Systems, Vol 187, 104837, 2020.
25. Y Han, B Li, Y Huang *et al*, 'Imbalanced fault classification of rolling bearing based on an improved oversampling method', Journal of the Brazilian Society of Mechanical Sciences and Engineering, Vol 45, No 4, 223, 2023.
26. D Dablain, B Krawczyk and N V Chawla, 'DeepSMOTE: fusing deep learning and SMOTE for imbalanced data', IEEE Transactions on Neural Networks and Learning Systems, Vol 34, No 9, pp 6390-6404, September 2023. DOI: 10.1109/TNNLS.2021.3136503
27. A Arafa, N El-Fishawy, M Badawy *et al*, 'RN-SMOTE: reduced noise SMOTE based on DBSCAN for enhancing imbalanced data classification', Journal of King Saud University – Computer and Information Sciences, Vol 34, No 8, pp 5059-5074, 2022.
28. Z Ren, Y Zhu, W Kang *et al*, 'Adaptive cost-sensitive learning: improving the convergence of intelligent diagnosis models under imbalanced data', Knowledge-Based Systems, Vol 241, 108296, 2022.
29. Z Wu, H Zhang, J Guo *et al*, 'Imbalanced bearing fault diagnosis under variant working conditions using cost-sensitive deep domain adaptation network', Expert Systems with Applications, Vol 193, 116459, 2022.
30. H Karamti, M M A Lashin, F M Alrowais *et al*, 'A new deep stacked architecture for multi-fault machinery identification with imbalanced samples', IEEE Access, Vol 9, pp 58838-58851, 2021.
31. A Vaswani, N Shazeer, N Parmar *et al*, 'Attention is all you need', Advances in Neural Information Processing Systems, 30, 2017.
32. X Liu, S Chen, L Song *et al*, 'Self-attention negative feedback network for real-time image super-resolution', Journal of King Saud University – Computer and Information Sciences, Vol 34, No 8, pp 6179-6186, 2022.
33. J Jiao, Y M Tang, K Y Lin *et al*, 'DilateFormer: multi-scale dilated transformer for visual recognition', IEEE Transactions on Multimedia, Vol 25, pp 8906-8919, 2023.
34. Y Xu, B Du and L Zhang, 'Self-attention context network: addressing the threat of adversarial attacks for hyperspectral image classification', IEEE Transactions on Image Processing, Vol 30, pp 8671-8685, 2021.
35. K Han, Y Wang, Q Tian *et al*, 'GhostNet: more features from cheap operations', Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, Washington, USA, pp 1580-1589, 13-19 June 2020.
36. T Y Lin, P Dollár, R Girshick *et al*, 'Feature pyramid networks for object detection', Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, USA, pp 2117-2125, 21-26 July 2017.
37. X Li, T Li and Y Wang, 'GW-DC: a deep clustering model leveraging two-dimensional image transformation and enhancement', Algorithms, Vol 14, No 12, 349, 2021.
38. X Li, Y Li, Y Cao *et al*, 'Fault diagnosis method for aircraft EHA based on FCNN and MSPSO hyperparameter optimisation', Applied Sciences, Vol 12, No 17, 8562, 2022.
39. S Han *et al*, 'End-to-end chiller fault diagnosis using fused attention mechanism and dynamic cross-entropy under imbalanced datasets', Building and Environment, Vol 212, 108821, 2022.
40. V Badrinarayanan, A Kendall and R Cipolla, 'SegNet: a deep convolutional encoder-decoder architecture for image segmentation', IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 39, No 12, pp 2481-2495, 2017.
41. X Pan, C Ge, R Lu *et al*, 'On the integration of self-attention and convolution', Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, Louisiana, USA, pp 815-825, 19-24 June 2022.
42. C Wang, C Deng and S Wang, 'Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost', Pattern Recognition Letters, Vol 136, 190-197, 2020.
43. Y Kang, G Chen, W Pan *et al*, 'A dual-experience pool deep reinforcement learning method and its application in fault diagnosis of rolling bearing with unbalanced data', Journal of Mechanical Science and Technology, Vol 37, No 6, pp 2715-2726, 2023.
44. J Zhang, K Zhang, Y An, H Luo and S Yin, 'An integrated multi-tasking intelligent bearing fault diagnosis scheme based on representation learning under imbalanced sample condition', IEEE Transactions on Neural Networks and Learning Systems, Vol 35, No 5, pp 6231-6242, May 2024. DOI: 10.1109/TNNLS.2022.3232147

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19838-w

An Intelligent diagnosis method for rolling bearings based on Ghost module and adaptive weighting module

Qiang Ruiru[1] · Zhao Xiaoqiang[1]

Received: 27 June 2023 / Revised: 7 May 2024 / Accepted: 6 July 2024

Abstract

The vibration signals of rolling bearings are inevitably affected by noise and working conditions. The use of one-dimensional raw signals converted into images for rolling bearing fault diagnosis has achieved good results, but ignores the large model and diagnostic speed, and thus it is not suitable for practical fault diagnosis. To address this problem, we propose a method based on Ghost module and adaptive weighting module. The method utilizes Ghost modules and coordinated attention to make the model lightweight while improving the network's ability to extract features of the input data. Additionally, in order to effectively utilize the similar feature maps generated by convolution, an adaptive weighting module is proposed to further simplify the learning process and reduce the network training time. The validation using the datasets from Case Western Reserve University and the Association for Mechanical Failure Prevention Technology demonstrates the effectiveness of the proposed method. Under the same conditions, our method achieves 98.62% accuracy with only one-tenth of the parameters of classical neural networks. In noise environment simulations, our method exhibits strong noise immunity with 98.64% accuracy, along with robust diagnostic and generalization performance under various loads. Compared to the advanced fault diagnosis algorithms, our method boasts a 4% higher average accuracy and has its superiority in rolling bearing fault diagnosis.

Keywords Fault diagnosis · Rolling bearing · Ghost module · Coordinate attention · Adaptive weighting module

✉ Qiang Ruiru
853924752@qq.com

Zhao Xiaoqiang
xqzhao@lut.edu.cn

[1] College of Electrical Engineering and Information Engineering, Lanzhou University of Technology, Lanzhou, Gansu 730050, China

Published online: 17 July 2024

1 Introduction

In modern industrial systems, mechanical rotating body has evolved towards high speed, mass and integration, and plays a vital role in aviation, industrial production and high-speed railways [1]. Rolling bearings, as important parts of mechanical equipment, play a role in supporting the rotating body, reducing friction and driving the transmission. If a bearing failure occurs, it would lead to an unexpected stoppage of the mechanical rotating body or even the entire mechanical equipment, thus causing serious casualties and economic losses [2, 3]. As rolling bearings are in a working environment with frequent load changes, short start-stop cycles and many sources of interference for long periods of time [4], therefore, the study of rolling bearing fault diagnosis has been a hot topic of research [5].

The current mainstream methods for rolling bearing fault diagnosis can be divided into two types: fault diagnosis methods based on fault mechanisms and data-driven fault diagnosis methods [6]. The fault mechanism-based diagnosis methods analyze the vibration characteristics of a damaged bearing by studying fault mechanism and building a kinetic model [7], which usually require extensive a priori knowledge and accurate system models, they are difficult to implement for complex systems. The data-driven methods do not rely on fault generating mechanisms and allow fault diagnosis to be carried out in the absence of a priori knowledge. As a result, some classical machine learning methods have been proposed, such as KNN [8], plain Bayesian [9], SVM [10, 11] and artificial neural networks [12]. These methods are also known as traditional data-driven methods. However, they still require a priori knowledge of the domain for fault extraction and the shallower network structure limits the fault diagnosis performance. In recent years, machinery and equipment inspection and fault diagnosis have entered the "era of big data analysis". Deep learning has received a lot of attention from researchers due to its powerful feature extraction capability. He et al. [13] used Fractional Fourier Transform (FRFT) and Deep Belief Network (DBN), and obtained good diagnostic results. Gao et al. [14] used multi-channel continuous wavelet transform (MCCWT) followed by joint long and short-term memory network (LSTM) to mine temporal features and local features. Kong et al. [15] combined the different types of features obtained by training several deep self-encoders (DAE) with different activation functions into a single feature pool, and then evaluated the features to construct a classifier. Zhang et al. [16] developed an enhanced Wasserstein GAN with a gradient penalty to generate high quality synthetic samples for faulty sample sets, thus solving the data imbalance problem.

Convolutional neural network (CNN) is one of the most important models in deep learning. With its powerful feature extraction capability, CNN has achieved significant results in fields, such as image classification, surface defect detection [17], speech recognition [18] and text translation. Eren et al. [19] introduced CNN to the field of fault diagnosis by using a one-dimensional convolutional neural network to train a deep learning model directly on the original signals. Zhang et al. [20] used a deep convolutional network with a wide first layer (WDCNN) to extract and suppress high-frequency noise to improve diagnostic accuracy. Zhang et al. [21] used a neural network with a residual structure, which greatly improved the flow of information throughout the network. Hoogen et al. [22] proposed a model built on the WDCNN framework to improve the performance of classification of fault types by using multivariate time series data. Although the above methods have achieved good results in dealing with rolling bearing fault diagnosis problems. However, they don’t take into account the complex characteristics of vibration signals on different time scales. Multiscale learning allows access to feature information at different time scales, thus improving the feature learning capability of the network [23]. Shi et al. [24] combined the concept of multiscale learning with attention mechanisms and residual learning to enable the network to extract richer fault features directly from the original vibration signals. Wang et al. [25] established a convolutional neural network with multi-scale feature fusion, thus solving the problem of noise interference and workload variation. Although the above works have yielded good results in terms of fault diagnosis. However, in the era of big data, problems such as data diversity lead to the fact that the 1D CNNs used in the above methods cannot fully utilize the feature extraction capability of the CNNs, thus affecting the diagnostic results of the network. Therefore, rolling bearing fault diagnosis researchers have begun to convert fault signals into image data, thus taking full advantage of the special diagnostic extraction capabilities of CNNs. Liang et al. [26] used wavelet variations to convert one-dimensional vibration data into two-dimensional time–frequency maps as input data for CNN, thus enabling composite fault diagnosis of rolling bearings. Zhang et al. [27] used the time–frequency map obtained by using Short Time Fourier Transform (STFT) as the input data for CNN and achieved good diagnostic results. Yao et al. [28] used an efficient neural network combined with the CBAM attention mechanism to achieve fault diagnosis of bearings in urban railways. Wang et al. [29] proposed a new method based on erosion operations (EOSTI) to convert time-domain vibration signals into RGB images, and verified the feasibility of the method using AlexNet-based CNN. The above methods make full use of the performance of CNNs by converting 1D signals into 2D images, but with little consideration of the model size and diagnostic efficiency. This results in such methods generating a large number of redundant features during computation, severely wasting the computational resources of the device and consuming a large amount of time, rendering the above methods inapplicable under practical conditions [30].

In order to solve the above problems, we first convert one-dimensional vibration signals into grey-scale images, so as to reduce the interference of noise in the original signals. The Ghost bottleneck is improved by using coordinate attention (CA) to enhance its performance. In addition, we propose an adaptive weighting module to avoid spending a lot of computational resources while making the model understand the input signals more comprehensively. Finally, the learned features are fed into the global average pooling and the Softmax function is used to achieve fault diagnosis of rolling bearings. The proposed method is validated by using the datasets of Case Western Reserve University (CWRU) and the Mechanical Failure Prevention Technology Society (MFPT). The experimental results show that the method proposed in this paper has high accuracy, good noise immunity, small model size and good diagnostic efficiency. The contributions of this paper are summarized as follows:

(1) One-dimensional vibration signals are converted into grey-scale images as inputs to reduce the interference of noise. We use CA to improve the Ghost bottleneck and introduce it into the model. Reducing the number of model parameters and diagnosis time while improving the network's ability to focus on 2D image coordinates improves the generalization of the network.

(2) We propose an adaptive weighting module based on dynamic convolution to efficiently process the redundant feature maps generated by the network. The representational

power of the network is improved by changing the existing equal processing of attention and by constructing adaptive weighting modules to dynamically learn signal features.

(3) The experiments on the CWRU and MFPT datasets show that the proposed method has good noise immunity and generalization capability, and has a small model size and a short diagnostic time. It is demonstrated that the method proposed has a high fault diagnosis performance and can accurately classify the types of faults under various loads.

The remainder of this paper is organized as follows. Section 2 introduces GhostNet, the coordinate attention mechanism, and dynamic convolution. Section 3 describes the proposed method in detail. Section 4 evaluates the proposed method through the experiments and analyses the results. The conclusions of this paper are given in Sect. 5.

2 Theoretical background

The main content of this section is a brief introduction to the Ghost module, dynamic convolution and the coordinate attention mechanism.

2.1 Ghost module

GhostNet[31] was a novel efficient neural network proposed by Han et al. in 2020. GhostNet solves the redundancy problem in mainstream CNNs by embracing rather than avoiding redundancy. Instead of discarding redundant feature mapping, GhostNet generates a few initial feature maps and then produces similar ones through inexpensive linear operations, termed Ghost feature maps. These Ghost feature maps are then fused with the initial ones to create the final output.

As shown in Fig. 1(a), for the input data $X \in \mathbb{R}^{c \times h \times w}$, where c is the number of channels of the input data, and h and w are the height and width of the input data, the output obtained after the ordinary convolution operation is as follows:

$$Y = X * f + b \tag{1}$$

where, $*$ is the convolution operation, $f \in \mathbb{R}^{c \times k \times k \times n}$ is the convolution filter in the layer, b is the bias term in the convolution, and $Y \in \mathbb{R}^{h\prime \times w\prime \times n}$ is the output feature map with n channels. In addition, k is the size of the convolution filter, n is the number of filters, $h\prime$ and $w\prime$ are the height and width of the output feature map, respectively. It follows that the FLOPs in this convolution process can be calculated as:

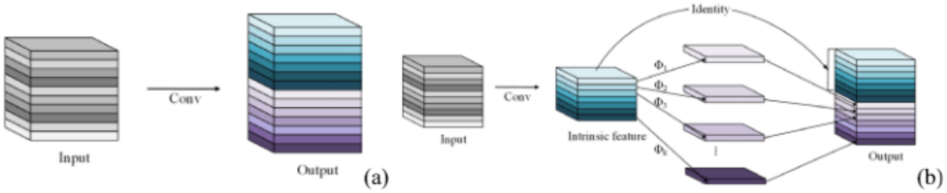


Fig. 1 Ordinary convolution and Ghost module: (**a**) Ordinary convolution, (**b**) Ghost module

$$\mathrm{FLOPs} = n \cdot h' \cdot w' \cdot c \cdot k \cdot k \tag{2}$$

Although redundant feature maps introduce a large number of parameters and computational effort into the convolution, this does not mean that redundant feature maps are useless. In fact, convolutional network requires redundant feature maps to provide a more comprehensive view of the data.

In the Ghost module, as shown in Fig. 1(b), the convolution filter in the convolution kernel is $F' \in \mathbb{R}^{c \times k \times k \times m}$, where $m \leq n$. To simplify the operations, the convolution in the Ghost module omits the bias term. The hyperparameters for filter size, stride and padding are the same as in Eq. 1. By the first convolution, m feature maps are generated and are referred to as intrinsic feature maps. Next, the Ghost module performs a series of cheap linear operations on each intrinsic feature map, so that each intrinsic feature map generates s ghost feature maps, thus expanding the feature maps to n, where $n = m \cdot s$. The average kernel size for each linear operation is d, and the number of linear operations is $m \cdot (s-1) = \frac{n}{s} \cdot (s-1)$. From this, the FLOPs required for the Ghost module can be calculated as:

$$\mathrm{FLOPs} = \frac{n}{s} \cdot h' \cdot w' \cdot c \cdot k \cdot k + (s-1) \cdot \frac{n}{s} \cdot h' \cdot w' \cdot d \cdot d \tag{3}$$

According to Eqs. 2 and 3, the acceleration ratio of the Ghost module to the ordinary convolution is:

$$\begin{aligned} ratio_s &= \frac{n \cdot h' \cdot w' \cdot c \cdot k \cdot k}{\frac{n}{s} \cdot h' \cdot w' \cdot c \cdot k \cdot k + (s-1) \cdot \frac{n}{s} \cdot h' \cdot w' \cdot d \cdot d} \\ &= \frac{n \cdot k \cdot k}{\frac{1}{s} \cdot c \cdot k \cdot k + \frac{s-1}{s} \cdot d \cdot d} \approx \frac{s \cdot c}{s+c-1} \approx s \end{aligned} \tag{4}$$

The compression ratio of the model parameters is:

$$ratio_c = \frac{n \cdot c \cdot k \cdot k}{\frac{n}{s} \cdot c \cdot k \cdot k + \frac{s-1}{s} \cdot d \cdot d} \approx \frac{s \cdot c}{s+c-1} \approx s \tag{5}$$

The Ghost bottleneck consists of the Ghost module, shown in Fig. 2.

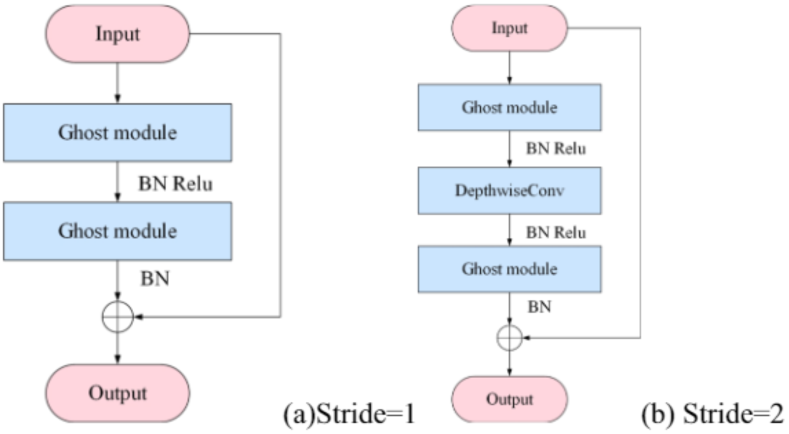


Fig. 2 Ghost bottleneck

2.2 Dynamic convolution

Dynamic convolution came about with the introduction of the concept of dynamic learning to deep learning for the first time by Jia et al. [32]. Compared to a single convolution kernel in normal convolution, dynamic convolution adapts the convolution parameters to the input and integrates multiple parallel convolution kernels via the Softmax function. Dynamic Convolution is fused in a non-linear form through an attention mechanism, which not only provides greater expressiveness but also allows for more efficient computation without the need to adjust the depth and width of the network [33]. The ordinary convolution formula is defined as: $y = g\left(W^T x + b\right)$ where W denotes the weight, b is the bias term, and $g(\cdot)$ denotes the activation function, the dynamic convolution is formulated as follows.

$$\begin{cases} y = g\left(\widetilde{W}^T x + \widetilde{b}\right) \\ \widetilde{W}(x) = \sum_{k=1}^{K} \pi_k(x)\widetilde{W}_k \\ \widetilde{b}(x) = \sum_{k=1}^{K} \pi_k(x)\left(\widetilde{b}\right)_k \\ \sum_{k=1}^{K} \pi_k(x) = 1{,}0 \le \pi_k(x) \le 1 \end{cases} \tag{6}$$

where $\widetilde{W}(x)$ and $\widetilde{b}(x)$ denote the weighted convolutional kernel and bias term, respectively, $\widetilde{W}_k$ and $\widetilde{b}_k$ denote the kth convolutional kernel and bias term, respectively, and π_k denotes the attentional weight of the k th convolutional kernel. Figure 3 shows the framework diagram of dynamic convolution.

2.3 Coordinate attention

The attention mechanism, inspired by human vision research, focuses processing resources on relevant visual information [34]. In CNNs, it's an additional neural network that assigns different weights to input information to highlight important parts and boost model performance. Common attention mechanisms include Squeeze-and-Excitation (SE) [35],

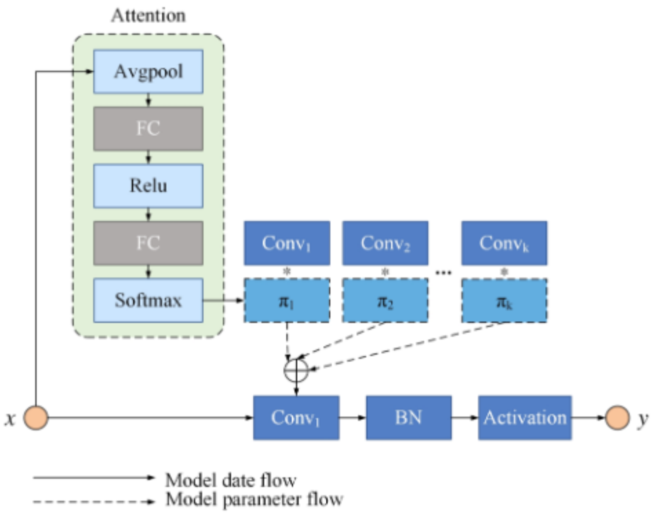


Fig. 3 dynamic convolution

Bottleneck-Attention-Module (BAM) [36], and Convolutional-Block-Attention-Module (CBAM) [37]. However, SE overlooks spatial location importance, while BAM and CBAM try to capture location information through global pooling. CBAM's weight considers local regions but lacks global context.

Coordinate attention [38] decomposes channel attention into two 1D feature encoding processes, which aggregate the features along two spatial directions, respectively. In this way, remote dependencies can be captured along one spatial direction, while retaining precise location information along the other. In simple terms, coordinate attention is achieved by averaging pooling horizontally and vertically, then encoding the spatial information, and finally fusing the spatial information in a channel-weighted manner. This approach is flexible and lightweight, and can easily be inserted into existing networks to enhance model performance. Figure 4 shows the structure of the coordinate attention.

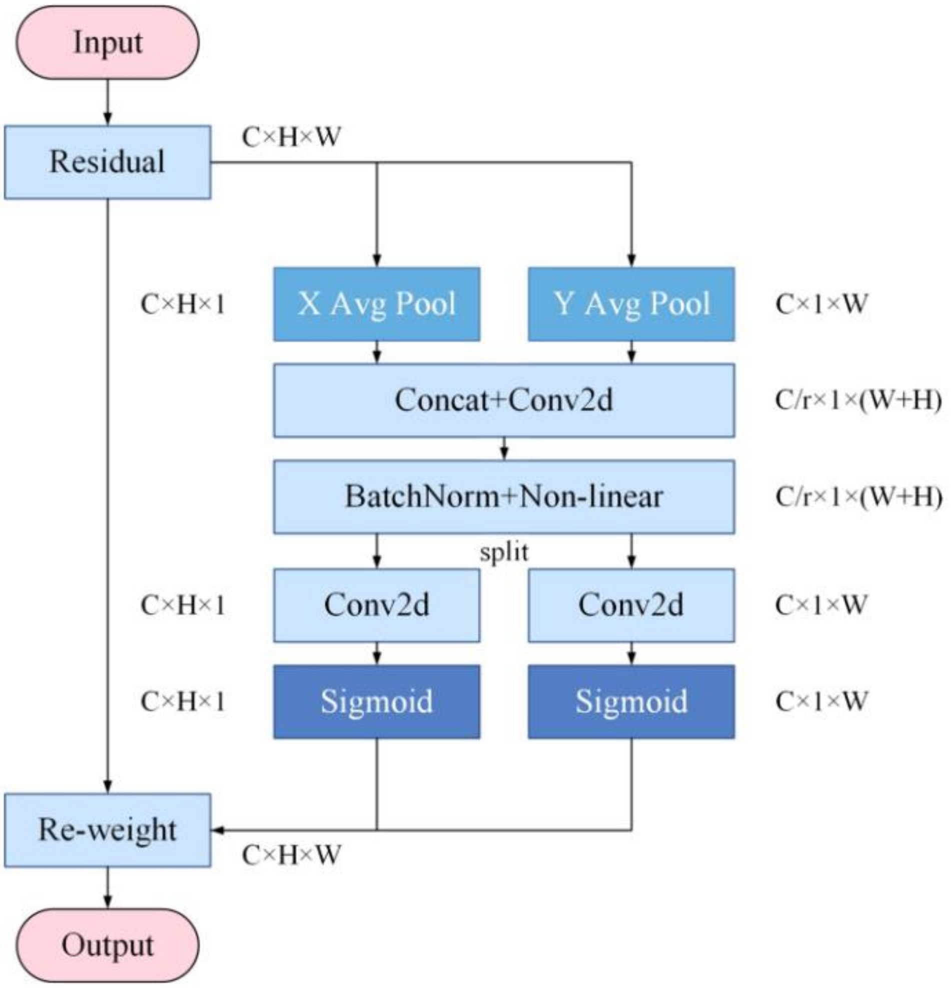


Fig. 4 Coordinate attention

3 The proposed method

This section presents a fault diagnosis method for rolling bearings based on Ghost module and attention mechanism. First, the process that one-dimensional vibration signals converted into a greyscale image is introduced, then the structure of the network model is shown, and finally the training strategy for the model is given.

3.1 Conversion of 1D vibration signals to image

In order to reduce the effect of noise in the 1D signals and to better exploit the advantages of CNN in image classification, the 1D signals are converted into a 2D image, which consists of 3 steps:

Step 1: Assumed that an image of size $N \times N$ is ultimately obtained, N columns of sub-signals of length N are selected in the one-dimensional vibrational signals by using sliding window fetching.
Step 2: Combine the randomly selected sub-signals from step 1 to obtain an $N \times N$ column of signals. The intensity of each signal is noted as $L(i)$,$i = 1{,}2, ...N^2$.
Step 3: The signals combined in step 2 are converted to a greyscale image in the following manner:

$$P(j,k) = round\left\{ \frac{L((j-1) \times N + k) - Min(L)}{Max(L) - Min(L)} \times 255 \right\} \tag{7}$$

where $round(\cdot)$ represents the rounding function which normalizes all pixels to the range of 0 to 255, which is the range of pixel values in a grayscale map.$P(j,k), j = 1......N, k = 1......N$ denotes the pixel value of each image after transformation. The transformation process is shown in Fig. 5. The result of data conversion is shown in Fig. 6.

Figure 5 shows a sample of the generated images, with image size N set to 32 in this paper.

3.2 Network model structure

The construction of the model consists mainly of the Ghost bottleneck sequence with the addition of coordinate attention and the adaptive weighting module proposed in this paper.

3.2.1 Improved Ghost bottleneck

Figure 7 represents the improved Ghost bottleneck. In the improved Ghost bottleneck, coordinate attention is added to model the relationship between channels, while remote dependencies are captured by using precise positional information as a means of improving the performance of image classification.

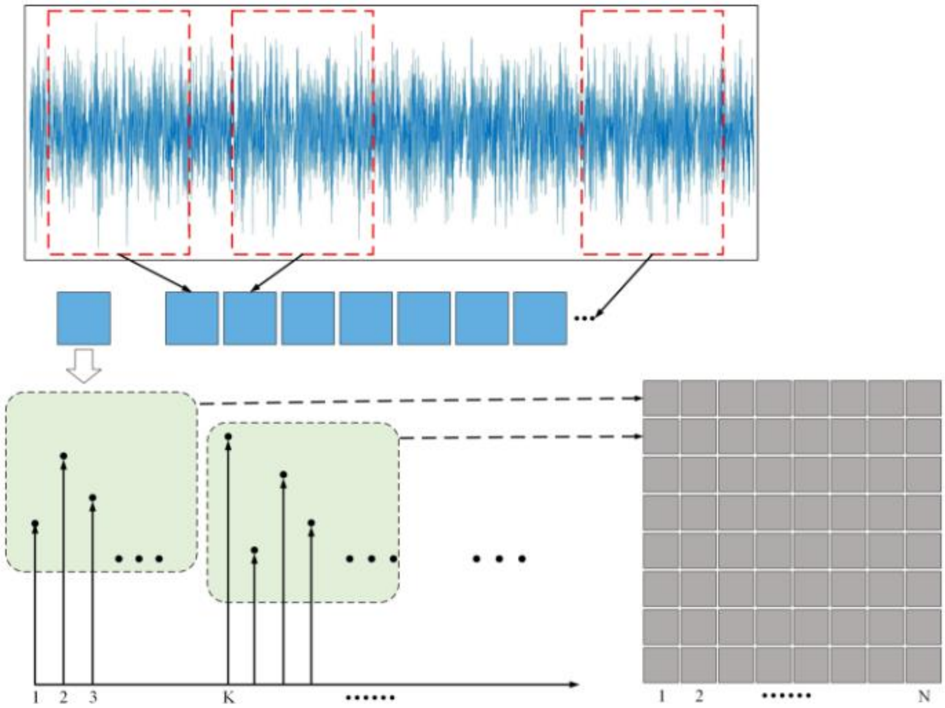


Fig. 5 Transformation process

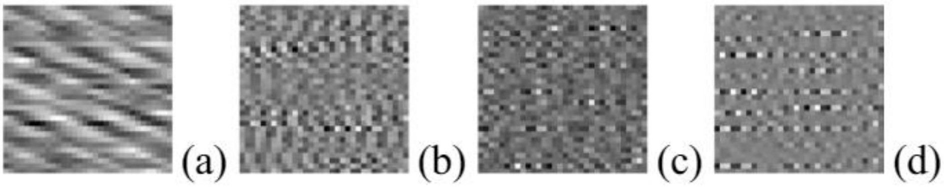


Fig. 6 Vibration image sample of (**a**) Normal sample, (**b**) Rolling body fault sample, (**c**) Sample of inner ring failures, (**d**) Sample of outer ring failures

3.2.2 Adaptive weighting module

The input data are passed through the improved Ghost bottleneck when the ghost feature map is generated by a series of cheap linear calculations. Studies have shown that the information in low-resolution images is rich in low-frequency and valuable high-frequency components. In order to avoid spending a lot of computational resources while making full use of the information contained in similar feature maps, an adaptive weighting module (AWM) is proposed in this paper. AWM contains three branches: the pixel attention branch, the channel attention branch and the adaptive weighting fusion branch, as shown in Fig. 8.

The input is X_{n-1}. The pixel attention branch contains a 1×1 convolution kernel pixel attention block (PA). In the channel attention branch, a 1×1 convolution kernel and a channel attention block (CA) are used, followed by feature recombination using a 1×1

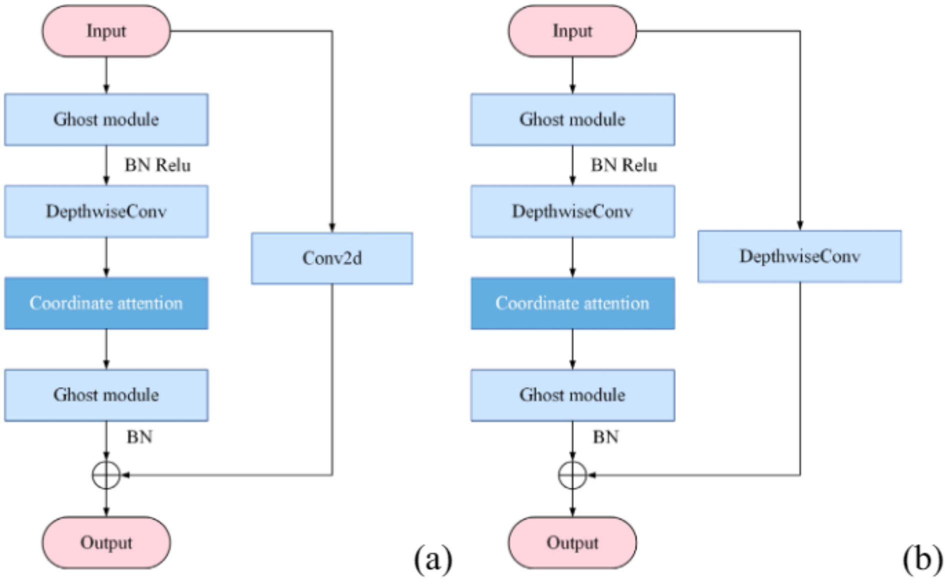


Fig. 7 Improved Ghost bottleneck of (**a**)Stride = 1, (**b**)Stride = 2

convolution for weight fusion with the adaptive weight fusion branch. At the same time, two cross structures are added into the two attentional branches (as shown in Fig. 9) to compensate for the features that are overlooked between the different attentions by means of feature reuse. Inspired by the literature [20] and similar to dynamic convolution, the adaptive weights branch is divided into a third branch of the module. This branch uses weighted summation to assign the weights to the pixel attention branch and the channel attention branch, automatically discard some unimportant attention features to achieve dynamic balance between the two branches. The output X_{PA} of the pixel attention branch and the output X_{CA} of the channel attention branch are feed into the 1×1 convolution layer to adjust the number of channels respectively, then are multiplied by different weights λ^{CA} and λ^{PA} for corresponding element summation, and finally feed into 1×1 convolution layer and output, and summed with the adaptive module residuals to obtain the final output X_n. AWM uses adaptive weighting to fuse the branches to dynamically adjust the weight share of two branches, improving the representational power of the network.

3.2.3 Network model structure

In summary, considering the limited data samples used, we use two improved Ghost bottleneck and two adaptive weight modules as the backbone of the network and name the proposed model in this paper as Coordinate-Ghost-Adaptive-Net (CGA-Net). The network structure is shown in Fig. 10, so that the size of the input image is 32×32, which is transformed and resized to 224×224. The transformed image enters the Ghost bottleneck after the first 3×3 convolution to generate the Ghost feature map, and then enters the adaptive weight layer. After two iterations, the model is made to fully integrate the features of the input signals. To avoid overfitting, the Dropout operation is used after a final layer of 3×3 convolution layers and average pooling, and the value of the Dropout parameter is set to 0.5.

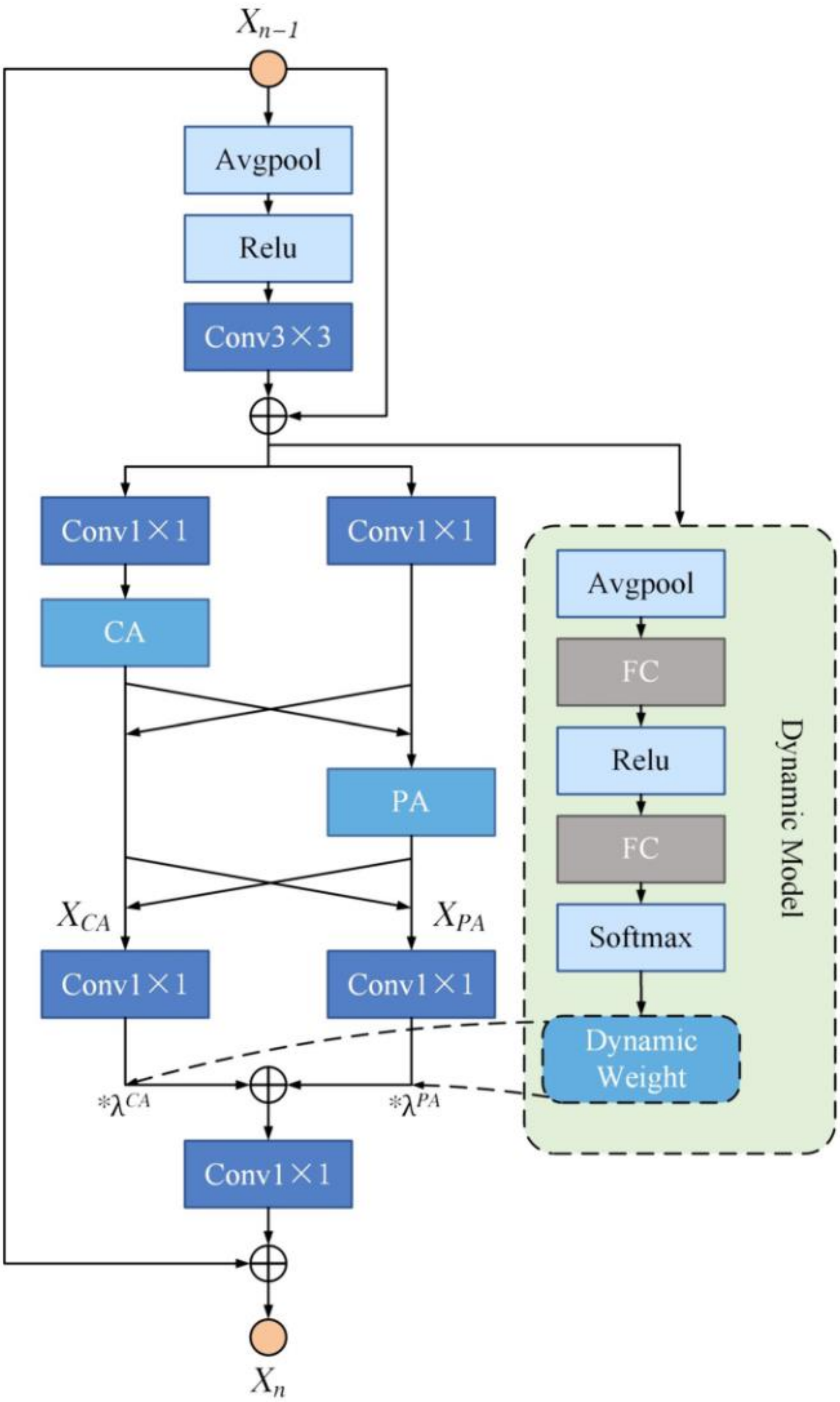


Fig. 8 Adaptive weighting module

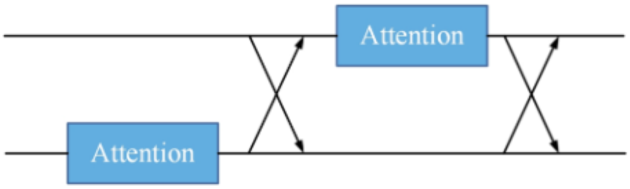


Fig. 9 Double cross structure

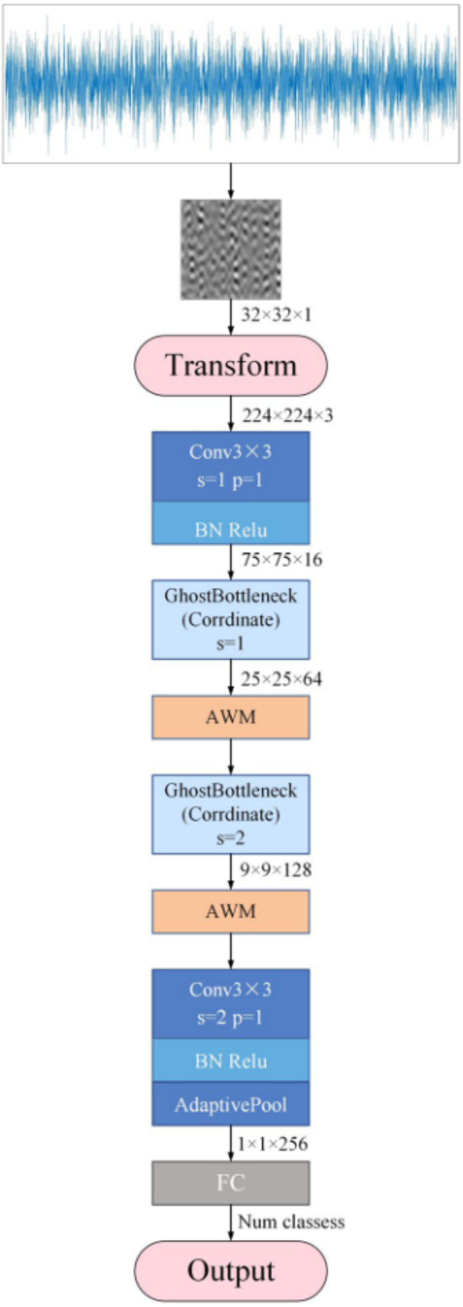


Fig. 10 Network model structure

3.3 Fault detection model framework and training strategy

The proposed method in this paper is shown in Fig. 11, which can be divided into 3 parts. The first part is data acquisition, where one-dimensional vibration signals are obtained by means of experimental equipment and sensors, which are then converted into 32×32 greyscale image, and the image dataset is divided into a training set and a test set according to scale. The second part is model training, the training epoch is set to 30, the initial learning rate is set to 0.001 and each batch is set to 64. With the model established in this paper, the model parameters are initialized and then CGA-Net is trained by using the data. The training process includes: calculating the loss function, updating the weights by back propagation by using the Softmax classification function, and using the Adam optimizer to optimize. Finally, when the epoch reaches the specified number, the training ends and

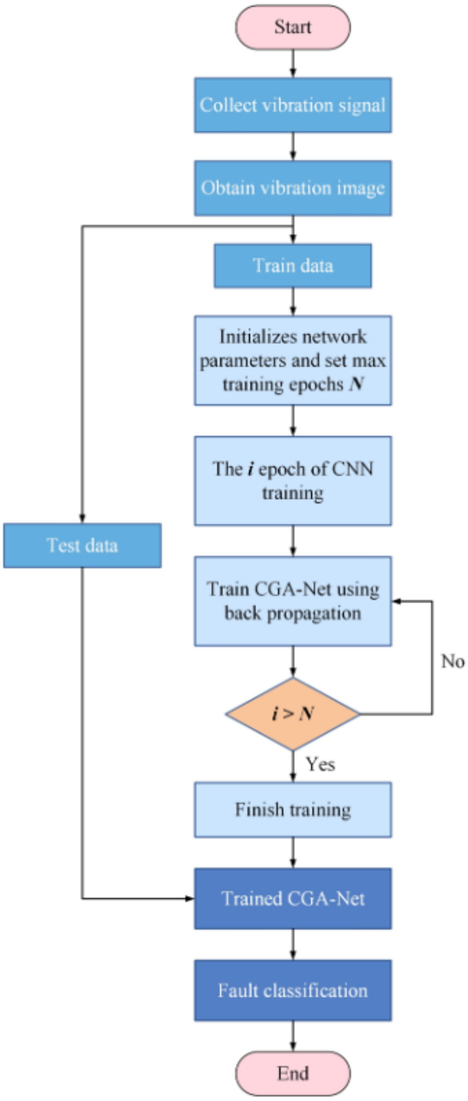


Fig. 11 The flow chart of CGA-Net for fault diagnosis

the model is saved. The third part is fault diagnosis, where the test dataset is fed into the trained model and finally the fault classification results are output.

4 Experiments and results

In this section, we evaluate the fault diagnosis performance of CGA-Net by using two rolling bearing datasets and experimentally validate the noise immunity, generalization and fault diagnosis capabilities of the method. The deep learning framework used in all simulations is Pytorch, with an AMD Ryzen 7 5800H CPU, an NVIDIA GeForce RTX 3060 GPU, and 16 GB RAM.

4.1 Case 1

4.1.1 Dataset description

The experimental data from operating SKF 6205 rolling bearings at Case Western Reserve University (CWRU), USA, which experimental platform is shown in Fig. 12, consist of four types: normal, ball failure (BF), inner ring failure (IF), and outer ring failure (OF). Each fault type is further categorized by size into 0.007, 0.014, 0.021, and 0.028 inches, resulting in 12 labels. The dataset encompasses the data collected under four loads labeled A, B, C, and D. Sampling is done at 12 kHz for 10 s, resulting in 406 sampling points per cycle. Each data sample is set to 2048 sampling points to ensure fault data reliability. With 150 samples per fault type under each load condition, totaling 1800 samples per load, the dataset comprises 7200 samples across the four loads, and is divided into a 7/3 training/test set ratio (Table 1 and 2).

4.1.2 Comparison methods

To verify the superiority of CGA-Net, we select the classic lightweight neural network model MobileNetV3 and various deep learning models for experimental comparison with CGA-Net, including GhostNet, AlexNet, ResNet34, and GoogLeNet. The initial hyper-parameters for all models were set as follows: learning rate = 0.001, Batch size = 32, Epochs = 30, and Optimizer as Adam optimizer.

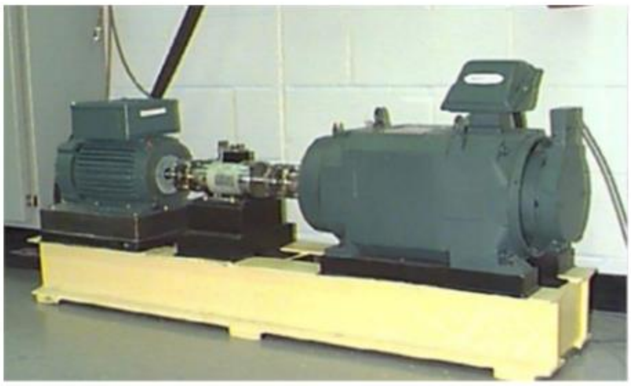


Fig. 12 Rolling bearing test bench

Table 1 Description of the dataset under the same load

Class label	Fault location	Fault size(in)	Number of samples
00	Normal	/	150
01	BF	0.007	150
02	IF	0.007	150
03	OF	0.007	150
04	BF	0.014	150
05	IF	0.014	150
06	OF	0.014	150
07	BF	0.021	150
08	IF	0.021	150
09	OF	0.021	150
10	BF	0.028	150
11	IF	0.028	150

Table 2 Description of the four types of load data

Dataset	Load(hp)	Sample size of the dataset	Training set/Test set
A	0hp	1800	1260/540
B	1hp	1800	1260/540
C	2hp	1800	1260/540
D	3hp	1800	1260/540

4.1.3 Performance testing of the original signals

In this experiment, as can be seen from Table 3 and Fig. 13,we evaluate six methods using original bearing signals for classification. CGA-Net consistently achieves accuracy rates of 97% or higher across all datasets, with an average accuracy of 98.62% and the shortest running time. Notably, CGA-Net achieves over 99% correct classification in Dataset C and Dataset D, which indicates that it has exceptional performance. While ResNet34 exhibits comparable accuracy (97.87%), it runs 7.4 times slower than CGA-Net. AlexNet's

Table 3 Accuracy of the six methods at the original signal

Algorithms	Dataset A	Dataset B	Dataset C	Dataset D	Average	Average time	Number of parameters
CGA-Net	**97.19%**	**98.94%**	**99.04%**	**99.3%**	**98.62%**	**124.19 s**	**0.84 MB**
GhostNet	90.71%	93.13%	95.31%	97.2%	94.09%	326.18 s	5.48 MB
AlexNet	76.21%	74.18%	83.66%	88.79%	80.71%	134.8 s	61 MB
ResNet34	95.48%	97.98%	98.74%	99.29%	97.87	918.53 s	47.51 MB
MobileNetV3	93.36%	96.36%	97.66%	98.91%	96.57	343.26 s	3.96 MB
GoogLeNet	81.21%	85.94%	86.91%	90.31%	86.09	319.64 s	7 MB

Bold text represents the best diagnostic with the smallest number of model parameters

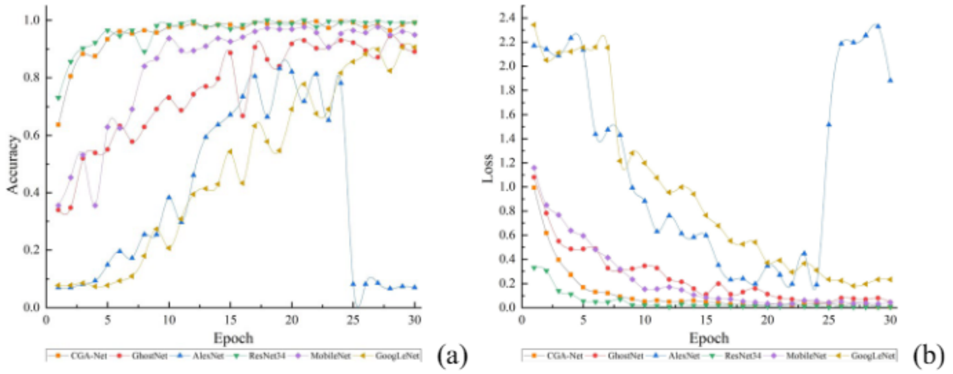


Fig. 13 (**a**) Test Accuracy, (**b**) Test losses

poor diagnostic results are attributed to gradient explosion from its large parameter count. CGA-Net's advantage lies in its small model size of 0.84 MB. Compared to GhostNet, CGA-Net outperforms due to improved feature extraction capabilities, particularly with the Ghost bottleneck and AWM. Overall, CGA-Net demonstrates satisfactory feature extraction capabilities.

4.1.4 Performance testing in noisy environments

In real-world rolling bearing environments, noise from vibration and friction is inevitable and can interfere with diagnostic accuracy when it is picked up by sensors along with original vibration signals. To simulate varying noise intensities, Gaussian white noise with different signal-to-noise ratios (SNR) is added to the original signals. The original signals serve as the training set, while the samples with added noise form the test set, allowing assessment of the proposed method's noise immunity across different noise levels. SNR is defined as:

$$\mathrm{SNR}_{\mathrm{dB}} = 10\log(\frac{P_{signal}}{P_{noise}}) \tag{8}$$

where P_{signal} and P_{noise} denote the powers of the original signals and the noise signals respectively. In this experiment, we use dataset B to train the model and add 3 dB, 6 dB, 9 dB and 12 dB of Gaussian white noise to the original signals for the test set. The results of the noise signal addition and image transformation for the 0.007" inner ring fault sample are shown in Fig. 14.

Table 4 and Fig. 15 compare CGA-Net's diagnostic performance with other methods under various noise levels. CGA-Net consistently achieves higher diagnostic accuracy across all noisy environments, with the shortest diagnostic time. When SNR = 9, CGA-Net exceeds 99% accuracy. ResNet34 follows with the second-highest accuracy, indicating its stability and feature extraction capability, however, its longer runtime limits practicality. Compared to GhostNet, CGA-Net improves accuracy by 4.64% while reducing diagnostic time to 2/5. This highlights CGA-Net's enhanced feature extraction with the Ghost bottleneck and AWM, showcasing its noise immunity, versatility, and robustness.

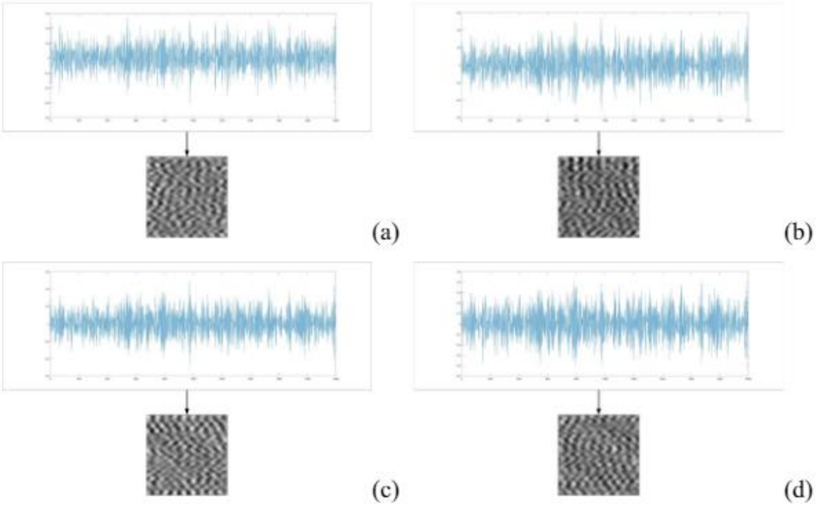
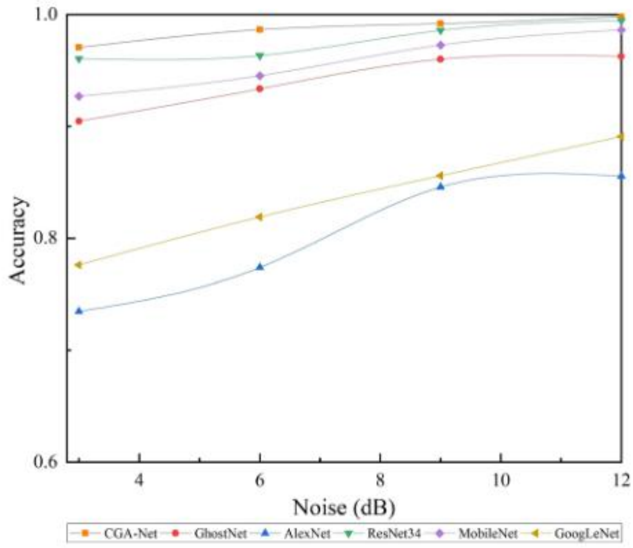


Fig. 14 Results of image conversion with different SNR. (**a**) SNR = 3, (**b**) SNR = 6, (**c**) SNR = 9, (**d**) SNR = 12

Table 4 Accuracy of the six methods with added noise

Algorithms	3dB	6dB	9dB	12dB	Average	Average time
CGA-Net	**97.06%**	**98.65%**	**99.18%**	**99.8%**	**98.67%**	**119.76 s**
GhostNet	90.47%	93.36%	96.01%	96.26%	94.03%	305.22 s
AlexNet	73.48%	77.38%	84.57%	85.55%	80.25%	135.5 s
ResNet34	96.05%	96.33%	98.59%	99.45%	97.61%	888.17 s
MobileNetV3	92.7%	94.53%	97.27%	98.63%	95.78%	324.09 s
GoogLeNet	77.61%	81.9%	85.59%	89.1%	83.55%	314.68 s

Fig. 15 Diagnostic results with different SNR



4.1.5 Performance testing under different working conditions

Rolling bearings operate in noisy environments under varying loads, presenting complex troubleshooting challenges. To simulate these conditions, we add Gaussian white noise under different loads to the CWRU drive-end bearing dataset, and divide the dataset as training set A and test sets B, C, and D to evaluate the adaptability of CGA-Net. The experimental results show that GhostNet, ResNet34, MobileNetV3, and GoogLeNet perform poorly when SNR = 3, with average accuracies ranging from 73.39% to 86.73%, as shown in Fig. 16a. Although the accuracy of CGA-Net is lower than 90% under the three variable conditions,

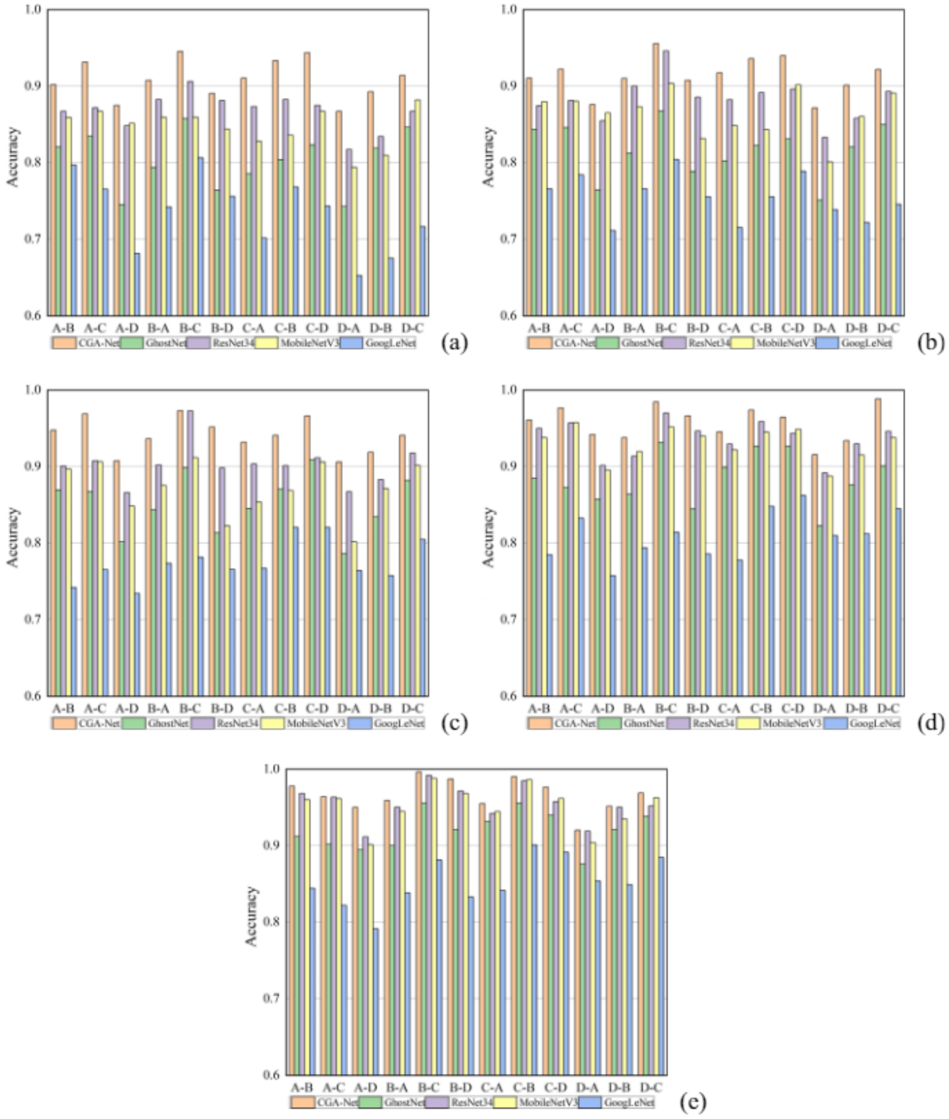


Fig. 16 Diagnostic results for variable working conditions with different SNR. (**a**) SNR = 3, (**b**) SNR = 6, (**c**) SNR = 9, (**d**) SNR = 12, (**e**) Without noise

4.1.5 Performance testing under different working conditions

Rolling bearings operate in noisy environments under varying loads, presenting complex troubleshooting challenges. To simulate these conditions, we add Gaussian white noise under different loads to the CWRU drive-end bearing dataset, and divide the dataset as training set A and test sets B, C, and D to evaluate the adaptability of CGA-Net. The experimental results show that GhostNet, ResNet34, MobileNetV3, and GoogLeNet perform poorly when SNR = 3, with average accuracies ranging from 73.39% to 86.73%, as shown in Fig. 16a. Although the accuracy of CGA-Net is lower than 90% under the three variable conditions,

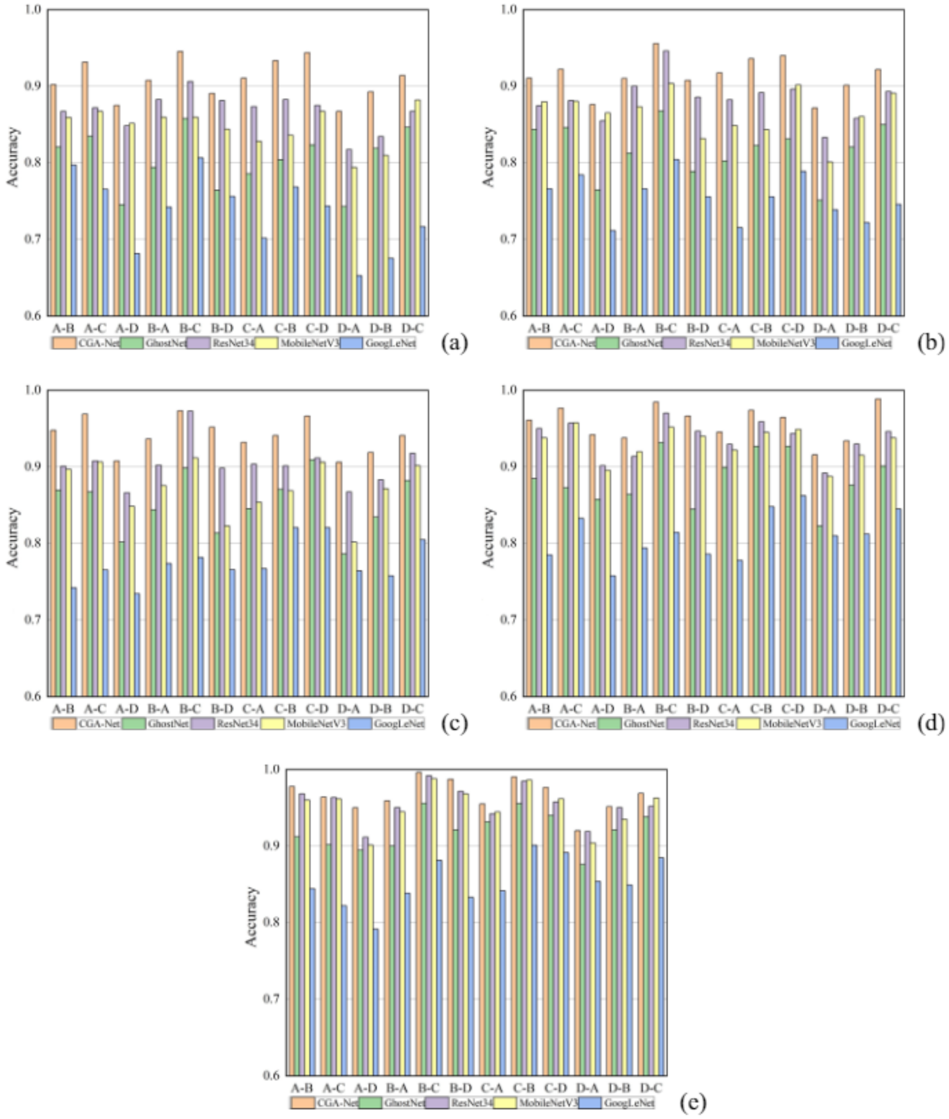


Fig. 16 Diagnostic results for variable working conditions with different SNR. (**a**) SNR = 3, (**b**) SNR = 6, (**c**) SNR = 9, (**d**) SNR = 12, (**e**) Without noise

its average accuracy is 90.94%, which is the highest among all the tested methods, indicating its strong adaptive ability. When the load difference between the training set and the test set is large, the overall diagnostic accuracy decreases due to the changes in signal characteristics. However, CGA-Net consistently outperforms the other methods, as shown Fig. 16d and Fig. 16e. The average accuracy is 95.74% at SNR = 12 and 96.62% at SNR = 0, indicating that there is almost no effect on the model performance when the noise is weak.

4.2 Case 2

4.2.1 Dataset description

To verify the generalization performance of CGA-Net, the experimental data are obtained from the American Society for Mechanical Failure Prevention Technology (MFPT). The bearing type used in this dataset is NICE. In the dataset, there are 3 normal data with a load of 270 lbs and 3 outer-raceway fault data with a load of 270 lbs. In addition, the MFPT dataset has 7 outer ring failure data and 7 inner-raceway failure data with an input shaft speed of 25 Hz and a sampling frequency of 48,828 sps for 3 s, corresponding to the loads of 25 lbs, 50 lbs, 100 lbs, 150 lbs, 200 lbs, 250 lbs and 300 lbs. We select seven inner-raceway failure data and seven outer-raceway failure data as experimental data, and the details of data are shown in Table 5. 1000 samples are collected for each type of fault data, with a training set/test set of 7/3.

4.2.2 Performance testing of the original signals

To further validate the performance of CGA-Net, we use a multi-scale lightweight-based rolling bearing fault diagnosis model(MLFD) [39], a lightweight ResNet rolling bearing fault diagnosis model(LResNet) [40], and a multi-scale residual fault diagnosis model as comparative algorithms(MRSCNN) [41], with each method run 20 times. We use the confusion matrix to represent the experimental results of CGA-Net in the test set, as shown in Fig. 17. Meanwhile, the results of the comparison algorithm experiments are shown in Fig. 18.

As shown in Fig. 17, for both inner and outer-raceway faults, CGA-Net can achieve the highest accuracy rates of 98.97% and 96.72% respectively. From Fig. 18, it can be seen that the performance of CGA-Net has a significant advantage over the comparison algorithms. This shows that CGA-Net has good generalization performance and achieves good classification accuracy for all types of bearing datasets.

Table 5 Description of the MFPT dataset

Fault	Inner-raceway						
Load(lbs)	0	50	100	150	200	250	300
Label	00	01	02	03	04	05	06
Train	700	700	700	700	700	700	700
Test	300	300	300	300	300	300	300
Fault	Outer-raceway						
Load(Lbs)	0	50	100	150	200	250	300
Label	00	01	02	03	04	05	06
Train	700	700	700	700	700	700	700
Test	300	300	300	300	300	300	300

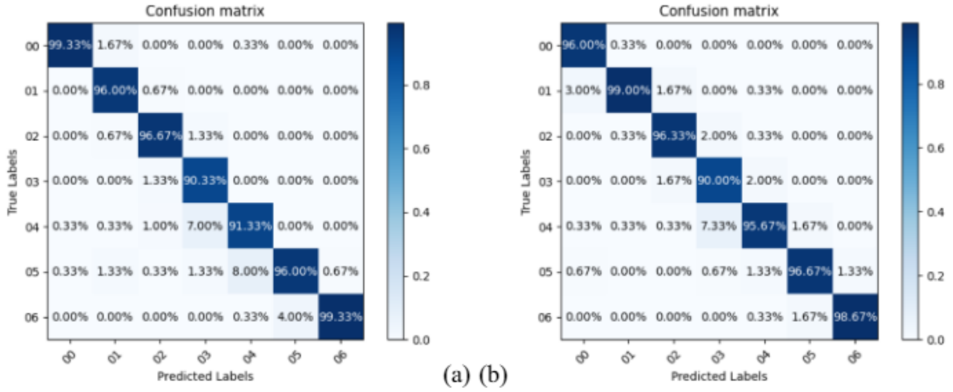
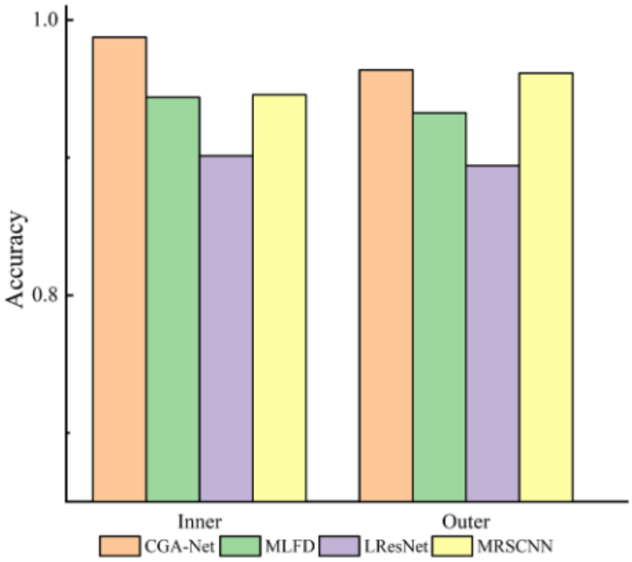


Fig. 17 Confusion matrix results of CGA-Net for fault classification in the test set (**a**) inner-raceway, (**b**) outer-raceway

Fig. 18 MFPT dataset diagnostic results



4.2.3 Performance testing in noisy environment

Similarly, to further analyze the noise immunity of CGA-Net versus the comparative methods. We still add Gaussian white noise with different signal-to-noise ratios to the seven inner and seven outer-raceway faults in the MPTF dataset as experimental data. As in Case 1, we use the dataset without added noise as the training set, and set the data with added noise as the test set, and the experimental results are shown in Table 6, it can be seen that compared with advanced rolling bearing fault diagnosis algorithms, CGA-Net still has more obvious advantages. Whether it is a strong noise environment with a signal-to-noise ratio of 3 dB or a weak noise environment with a signal-to-noise ratio of 12 dB, the accuracy of faults obtained by CGA-Net is above 90%. The accuracy of LResNet is lower than 85% in the case of strong noise with a signal-to-noise ratio of 3 dB, which shows that its noise immunity is poor. MLFD and MRSCNN have better noise immunity due to their

Table 6 MPTF dataset variable noise data tests

Algorithms	Inner-raceway diagnostic accuracy (%)			
	3dB	6dB	9dB	12dB
CGA-Net	**90.78%**	**94.45%**	**95.89%**	**97.14%**
MLFD	88.17%	91.45%	92.19%	96.24%
LResNet	82.15%	86.58%	88.73%	91.08%
MRSCNN	89.89%	92.46%	93.28%	96.09%
Algorithms	Outer-raceway diagnostic accuracy (%)			
	3dB	6dB	9dB	12dB
CGA-Net	**90.04%**	**92.12%**	**94.36%**	**95.69%**
MLFD	85.79%	89.53%	91.87%	94.13%
LResNet	83.26%	85.49%	86.61%	90.22%
MRSCNN	86.72%	88.91%	91.45%	93.98%

multi-scale network structure, but their diagnostic performance is not as good as that of CGA-Net. To summarize, CGA-Net still has a performance advantage over the advanced fault diagnostic algorithms.

4.3 Ablation experiments

In order to explore the impact of each part of CGA-Net, we perform ablation experiments on GA-Net, CG-Net, and G-Net using the CWRU dataset. Among them, the specific details of the branching structure of the comparison network are shown in Table 7, and to ensure the fairness of the experiment, all training parameters are kept consistent with CGA-Net.

4.3.1 Performance testing of the original signals

The experimental design is the same as 4.1.3 and the results are shown in Table 8, it can be seen that the performance of CGA-Net is better than the four methods. GA-Net uses the Ghost bottleneck without CA, and the model size does not change, but the performance shows a significant degradation, with the average accuracy dropping to 92.23%. This proves that CA does not enhance the size of the model, but can significantly improve the model performance. CG-Net uses CA's improved Ghost bottleneck, but drops AWM. It can be seen that the model size is reduced to one-half of

Table 7 Details of the branching structure of the ablation experiment

model	Coordinate Attention Mechanism	Adaptive weighting module
CGA-Net	Yes	Yes
GA-Net	No	Yes
CG-Net	Yes	No
G-Net	No	No

Table 8 Accuracy of four variants on original signals

Algorithms	Dataset A	Dataset B	Dataset C	Dataset D	Average	Average time	Number of parameters
CGA-Net	**97.19%**	**98.94%**	**99.04%**	**99.3%**	**98.62%**	124.19 s	0.84 MB
GA-Net	92.71%	92.37%	91.26%	92.98%	92.33%	113.35 s	0.84 MB
CG-Net	84.37%	84.93%	85.19%	84.29%	84.6%	97.06 s	**0.4 MB**
G-Net	80.59%	80.14%	79.38%	80.14%	80.06%	**93.73 s**	**0.4 MB**

CGA-Net, and the diagnostic time is significantly reduced. However, CG-Net shows a significant decrease in performance. This demonstrates that although AWM boosts the size of the model and extends the diagnostic time, it has a higher impact on the model performance than CA's improved Ghost bottleneck. G-Net does not use AWM and CA's improved Ghost bottleneck, and the performance is the worst, with an average accuracy of only 80.06%. In conjunction with 4.1.3, the experiments demonstrate that by using CA's improved Ghost bottleneck and AWM, the performance and timeliness of CGA-Net has a very significant advantage.

4.3.2 Performance testing in noisy environment

The experimental design is the same as 4.1.4 and the results are shown in Table 9. The average accuracy of G-Net is only 76.29%, this is due to the fact that it cannot understand the input data well without reasonable improvement in the case of insufficient network depth of the model. CGA-Net has good noise immunity and the average accuracy is 22% higher than G-Net. It can be seen that using CA's improved Ghost bottleneck and AWM allow the model to learn the input data better, resulting in a huge improvement in the model's noise immunity.

In summary, CGA-Net contains both CA and AWM modules and has the best performance.GA-Net contains only AWM without CA and has a higher performance than CG-Net.This shows that the gain of AWM on model performance is greater than the gain of CA on model performance. With the help of CA and AWM, CGA-Net is able to perform better feature extraction, better utilize spatial information as well as get better robustness.

Table 9 Noise immunity experiments with four branching variants

Algorithms	3dB	6dB	9dB	12dB	Average	Average time
CGA-Net	**97.06%**	**98.65%**	**99.18%**	**99.8%**	**98.67%**	119.76 s
GA-Net	82.19%	85.18%	86.27%	90.14%	85.95%	117.42 s
CG-Net	72.64%	74.21%	77.96%	82.14%	76.74%	98.58 s
G-Net	70.61%	72.56%	73.48%	78.52%	73.79%	**92.16 s**

5 Conclusion

The CGA-Net proposed in this paper can be used for fault diagnosis in noisy environments, load variation conditions. The vibration signal is first converted into an image and adequate feature extraction is performed using a 2D CNN. Then, in order to reduce the number of model parameters and enhance the model representation, coordinate attention is introduced into the Ghost bottleneck, which significantly reduces the model computation. Finally, AWM is designed to characterize the input information more efficiently by dynamically exploiting the features generated by the convolution. We validate the proposed CGA-Net by using the CWRU and MFPT datasets, respectively. In the CWRU dataset, the results demonstrate that CGA-Net can efficiently mitigate the degradation of diagnostic performance of diagnostic networks due to noise interference and variable operating conditions. In the MFPT dataset experiments, CGA-Net still has high diagnostic accuracy in the face of large data variability, and has a strong feature learning capability and satisfactory diagnostic results while greatly reducing the size of the model. In summary, CGA-Net has good fault identification and generalization performance under different operating conditions.

Acknowledgements This research work has been awarded by the National Natural Science Foundation of China (No.62263021), the College Industrial Support Project of Gansu Province (2023CYZC-24)

Data availability All data that support the findings of this study are included within the article (and any supplementary files).

Declarations

Conflicts of interests The authors declare that they have no conflicts of interest to this work. The people involved in the experiment have been informed and formally accepted.

References

1. Peng B, Bi Y, Xue B et al (2022) A survey on fault diagnosis of rolling bearings[J]. Algorithms 15(10):347
2. Liang H, Cao J, Zhao X (2022) Multi-scale dynamic adaptive residual network for fault diagnosis[J]. Measurement 188:110397
3. Wan L, Li Y, Chen K et al (2022) A novel deep convolution multi-adversarial domain adaptation model for rolling bearing fault diagnosis[J]. Measurement 191:110752
4. Hoang DT, Kang HJ (2019) A survey on deep learning based bearing fault diagnosis[J]. Neurocomputing 335:327–335
5. Yan G, Chen J, Bai Y et al (2022) A survey on fault diagnosis approaches for rolling bearings of railway vehicles[J]. Processes 10(4):724
6. AlShorman O, Irfan M, Saad N et al (2020) A review of artificial intelligence methods for condition monitoring and fault diagnosis of rolling element bearings for induction motor[J]. Shock Vib 2020:1–20
7. Mishra C, Samantaray AK, Chakraborty G (2017) Ball bearing defect models: A study of simulated and experimental fault signatures[J]. J Sound Vib 400:86–112
8. Pandya DH, Upadhyay SH, Harsha SP (2013) Fault diagnosis of rolling element bearing with intrinsic mode function of acoustic emission data using APF-KNN[J]. Expert Syst Appl 40(10):4137–4145
9. Muralidharan V, Sugumaran V (2012) A comparative study of Naïve Bayes classifier and Bayes net classifier for fault diagnosis of monoblock centrifugal pump using wavelet analysis[J]. Appl Soft Comput 12(8):2023–2029
10. Yan X, Jia M (2018) A novel optimized SVM classification algorithm with multi-domain feature and its application to fault diagnosis of rolling bearing[J]. Neurocomputing 313:47–64

11. Goyal D, Choudhary A, Pabla BS et al (2020) Support vector machines based non-contact fault diagnosis system for bearings[J]. J Intell Manuf 31:1275–1289
12. Al-Raheem KF, Roy A, Ramachandran KP, et al (2008) Application of the Laplace-wavelet combined with ANN for rolling bearing fault diagnosis[J]. https://doi.org/10.1115/1.2948399
13. He X, Ma J (2020) Weak fault diagnosis of rolling bearing based on FRFT and DBN[J]. Syst Sci Control Eng 8(1):57–66
14. Gao D, Zhu Y, Ren Z et al (2021) A novel weak fault diagnosis method for rolling bearings based on LSTM considering quasi-periodicity[J]. Knowl-Based Syst 231:107413
15. Kong X, Mao G, Wang Q et al (2020) A multi-ensemble method based on deep auto-encoders for fault diagnosis of rolling bearings[J]. Measurement 151:107132
16. Zhang H, Wang R, Pan R et al (2020) Imbalanced fault diagnosis of rolling bearing using enhanced generative adversarial networks[J]. IEEE Access 8:185950–185963
17. Pourkaramdel Z, Fekri-Ershad S, Nanni L (2022) Fabric defect detection based on completed local quartet patterns and majority decision algorithm[J]. Expert Syst Appl 198:116827
18. Yang CHH, Qi J, Chen SYC, Chen PY, Siniscalchi SM, Ma X, Lee CH (2021, June) Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition. In: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 6523–6527
19. Eren L (2017) Bearing fault detection by one-dimensional convolutional neural networks[J]. Math Probl Eng 2017:1–9
20. Wang P, Song L, Guo X et al (2021) A high-stability diagnosis model based on a multiscale feature fusion convolutional neural network[J]. IEEE Trans Instrum Meas 70:1–9
21. Zhang W, Li X, Ding Q (2019) Deep residual learning-based fault diagnosis method for rotating machinery[J]. ISA Trans 95:295–305
22. van den Hoogen JO, Bloemheuvel SD, Atzmueller M (2020) An improved wide-kernel cnn for classifying multivariate signals in fault diagnosis. In: 2020 International Conference on Data Mining Workshops (ICDMW). IEEE, pp 275–283
23. An Z, Li S, Wang J et al (2019) Generalization of deep neural network for bearing fault diagnosis under different working conditions using multiple kernel method[J]. Neurocomputing 352:42–53
24. Shi Y, Deng A, Deng M et al (2020) Enhanced lightweight multiscale convolutional neural network for rolling bearing fault diagnosis[J]. IEEE Access 8:217723–217734
25. Zhang W, Peng G, Li C et al (2017) A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals[J]. Sensors 17(2):425
26. Liang P, Deng C, Wu J et al (2019) Compound fault diagnosis of gearboxes via multi-label convolutional neural network and wavelet transform[J]. Comput Ind 113:103132
27. Zhang Y, Xing K, Bai R et al (2020) An enhanced convolutional neural network for bearing fault diagnosis based on time–frequency image[J]. Measurement 157:107667
28. Yao D, Liu H, Yang J et al (2021) Implementation of a novel algorithm of wheelset and axle box concurrent fault identification based on an efficient neural network with the attention mechanism[J]. J Intell Manuf 32:729–743
29. Wang Z, Zhao W, Du W et al (2021) Data-driven fault diagnosis method based on the conversion of erosion operation signals into images and convolutional neural network[J]. Process Saf Environ Prot 149:591–601
30. Yao D, Liu H, Yang J et al (2020) A lightweight neural network with strong robustness for bearing fault diagnosis[J]. Measurement 159:107756
31. Han K, Wang Y, Tian Q, Guo J, Xu C, Xu C (2020) Ghostnet: more features from cheap operations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1580–1589
32. Jia X, De Brabandere B, Tuytelaars T, Van Gool L (2016, January) Dynamic filter networks for predicting unobserved views. In: Proceedings ECCV 2016 workshops, pp 1–2
33. Zhang Y, Zhang J, Wang Q, Zhong Z (2020) Dynet: dynamic convolution for accelerating convolutional neural networks. arXiv preprint arXiv:2004.10694. https://doi.org/10.48550/arXiv.2004.10694
34. Niu Z, Zhong G, Yu H (2021) A review on the attention mechanism of deep learning[J]. Neurocomputing 452:48–62
35. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, pp 7132–7141
36. Park J, Woo S, Lee JY, Kweon IS (2018) Bam: bottleneck attention module. arXiv preprint arXiv:1807.06514. https://doi.org/10.48550/arXiv.1807.06514

37. Woo S, Park J, Lee J Y, et al (2018) Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). pp 3–19. https://doi.org/10.1007/978-3-030-01234-2_1
38. Hou Q, Zhou D, Feng J (2021) Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 13713–13722
39. Meng Z, Luo C, Li J et al (2023) Research on fault diagnosis of rolling bearing based on lightweight model with multiscale features[J]. IEEE Sens J 23(12):13236–13247
40. Chang M, Yao D, Yang J (2023) Intelligent fault dignosis of rolling bearings using efficient and lightweight resnet networks based on an attention mechanism (september 2022)[J]. IEEE Sens J 23(9):9136–9145
41. Zhao X, Zhang Y (2022) An intelligent diagnosis method of rolling bearing based on multi-scale residual shrinkage convolutional neural network[J]. Meas Sci Technol 33(8):085103

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

第52卷 第10期
2024年10月

华南理工大学学报（自然科学版）
Journal of South China University of Technology
（Natural Science Edition）

Vol. 52 No. 10
October 2024

doi：10.12141/j.issn.1000-565X.240021

基于格拉姆角差场和生成对抗网络的小样本滚动轴承故障诊断方法

强睿儒 赵小强

（兰州理工大学 电气工程与信息工程学院，甘肃 兰州 730050）

摘 要：针对基于深度学习的滚动轴承故障诊断算法需要从大量标注数据中学习，且面对样本数量受限时诊断效果不佳的问题，文中提出了一种基于格拉姆角差场（GADF）和生成对抗网络（GAN）的小样本滚动轴承故障诊断方法。首先，提出了基于GADF变换的数据增强方式，将少数1维振动信号通过GADF变换转换为2维GADF图像，并通过裁剪得到GADF子图，从而得到大量的图像样本；然后，将条件生成对抗网络（CGAN）与带有梯度惩罚的Wasserstein GAN（WGAN-GP）相结合，构建一种新的生成对抗网络，该网络通过条件辅助信息与梯度惩罚增强模型训练稳定性，并设计动态坐标注意力机制以增强模型的空间感知能力，从而生成高质量样本；最后，使用生成的样本对分类器进行训练，并在验证集上得到诊断结果。文中分别使用东南大学数据集和美国凯斯西储大学（CWRU）数据集进行了两组小样本环境下的轴承故障诊断实验。结果表明，与传统生成对抗网络以及先进的小样本故障诊断方法相比，文中所提方法的准确率和精确率等5项故障诊断指标均获得最好的结果，可以准确诊断出小样本条件下的轴承故障类型。

关键词：小样本轴承故障诊断；格拉姆角差场；生成对抗网络；注意力机制

中图分类号：TP185；TH133.3 **文章编号：**1000-565X（2024）10-0064-12

在现代工业中，旋转机械设备通过旋转动作来实现特定的功能，在各个工业领域都得到了广泛的应用[1]。滚动轴承作为旋转机械产品的关键部件之一，通常在高转速以及重负载等复杂的条件下工作，往往容易形成各种类型的缺陷和故障。在实际工业环境中，滚动轴承损坏可能导致重大事故的发生，造成巨大的经济损失和安全问题[2]。因此，开展滚动轴承故障诊断研究、预防这些严重后果具有重要的意义。近年来，智能传感器和人工智能技术的蓬勃发展，推动了基于深度学习的故障诊断方法的出现[3-4]。相较于传统的故障诊断方法，通过数据驱动的深度学习方法不需要大量的专家知识，仅需通过神经元对数据进行特征提取就可以取得良好的诊断效果[5]。

然而，基于数据驱动的故障诊断方法往往需要大量的标注数据对模型进行训练[6]，但在实际的工作环境中，旋转机械设备往往在大多数时间内处于健康工作状态，很少出现故障。这导致故障样本很难充分获取，用于模型训练的数据较少[7]。在这种情况下，有限的样本无法代表故障数据的全部特征，

收稿日期：2024-01-11

基金项目：国家自然科学基金资助项目（62263021，62163023）；甘肃省高校产业支撑计划项目（2023CYZC-24）

Foundation items：Supported by the National Natural Science Foundation of China（62263021，62163023）and the College Industrial Support Project of Gansu Province（2023CYZC-24）

作者简介：强睿儒（1996—），男，博士生，主要从事数据挖掘、故障诊断研究。E-mail：853924752@qq.com

†**通信作者：**赵小强（1967—），男，博士，教授，主要从事数据挖掘、故障诊断研究。E-mail：xqzhao@lut.edu.cn

如果直接使用这些数据进行训练，模型的性能无法较好地泛化至验证集中，导致诊断效果不佳[8]。

样本不足带来的小样本问题给滚动轴承故障诊断研究领域带来了极大的挑战，研究人员针对该问题进行了大量的研究。根据处理策略的不同，小样本故障诊断研究主要分为3类：基于统计学的方法、基于物理知识的方法以及基于深度学习的方法。尽管小样本数据不足以支持传统的统计分析方法，然而仍有一些统计方法可以在小样本的条件下进行故障诊断。如Indira等[9]提供的功率分析方法，有效地保证了统计稳定性，在确认最小样本故障诊断时取得良好的效果。Liu等[10-11]基于物理知识，提出了一种有限元模型仿真诊断网络，通过建立的有限元仿真模型，在数据量较少的情况下可以模拟获得大量的故障数据，从而解决数据不足的问题。然而，上述两种方法都需要大量的专家知识，建立复杂的数学模型，相较于基于深度学习的方法太过复杂。

同样地，基于深度学习的小样本故障诊断方法也分为3类：基于迁移学习的方法、基于元学习的方法和基于数据增强的方法。其中，基于迁移学习的方法是利用已经训练好的模型或特征提取器，在小样本的条件下进行微调并迁移到目标任务上，以此更好地利用有限的样本进行故障诊断。Hu等[12]使用人工损伤轴承产生的数据代替真实轴承产生的数据来训练模型，并通过实验验证了完整关系网络结构的性能改进能够较好地完成学习任务。然而，迁移学习的源域与目标域之间的差距过大时，不同设备、不同环境下的故障样本可能存在显著的分布差异。这种领域适应问题可能会导致迁移学习性能下降。基于元学习的方法是通过在大量不同任务上进行学习，让模型具备更好的泛化能力和适应新任务的能力。Chen等[13]将各种工况下的故障诊断问题转化为几次分类问题，并采用模型不可知的元学习模型来解决问题，该方法只需使用少量样本就可以快速适应新任务。Xia等[14]提出了一种基于增强的鉴别元学习方法，通过信号转换和多尺度学习的结合来提高特征学习的鲁棒性和特征嵌入的自适应性，并通过实验验证该方法的有效性。但元学习方法通常需要大量的元数据来学习任务间的共性和关系。对于小样本故障诊断问题，由于样本数量有限，可能难以提供足够多的元训练数据，导致算法受限。基于数据增强的方法是使用数据生成的方法对数据集进行扩充。Han等[15]使用改进的合成少数类过采样技术(SMOTE)对样本之间进行线性插值，从而获取大量数据样本。但使用SMOT进行插值时具有一定的盲目性，无法准确地生成足够的高质量样本。Yang等[16]将振动信号转换为单通道灰度图后，使用条件生成对抗网络(CGAN)对数据进行扩充，配合2维卷积神经网络对样本进行分类。Fan等[17]将振动信号转换为灰色纹理图像，并利用带有梯度惩罚的Wasserstein GAN[18](WGAN-GP)生成新的数据集，以解决样本不足的问题。但生成对抗网络在训练过程中容易出现模式崩溃、梯度消失和梯度爆炸等问题，导致训练过程不稳定，生成低质量样本，无法用于训练。

综上所述可知，尽管针对小样本故障诊断的研究取得了一定的成果，但仍然存在诸多问题。基于迁移学习的方法在任务数据差异性较大时，迁移学习诊断效果不佳。基于元学习的方法受限于元数据的制约，导致算法性能受限。基于数据增强的方法依赖原始样本质量和生成算法的稳定性。因此，为有效处理小样本条件下的故障诊断问题，文中提出了一种基于格拉姆角差场(GADF)和生成对抗网络的小样本故障诊断方法。该方法首先使用GADF变换将1维时序信号转换为2维GADF图，并通过剪裁实现初步的数据增强；然后，为了进一步扩充数据，文中提出了一种新的生成对抗网络模型，通过该模型可以稳定生成高质量样本，以实现数据增强的目的；最后，将生成的数据作为训练数据用于分类模型的训练，并在验证集中得到诊断结果，以此完成小样本条件下的故障诊断任务。

1　基础理论

1.1　基于格拉姆角差场的1维数据转换

近年来，学者们提出了多种将1维信号转化为2维图像的方法，这些方法在保留故障特征的同时，往往需要依赖专家经验和专业知识，这种依赖性限制了这些方法的普适性。为解决这些问题，Wang等[19]提出了GADF变换，以可视化解释1维时间序列。

设$X=\{x_1,x_2,\cdots,x_n\}$为具有n个样本的时间序列，对X进行GADF变换的步骤如下：

(1)将输入的时间序列数据归一化在$[-1,1]$的范围内，以缩放时间序列，即

$$\tilde{x}_i=\frac{x_i-\min x}{\max x-\min x},\quad \forall i\in\{1,2,\cdots,n\}\qquad(1)$$

(2)将归一化和缩放后的时间序列信号由笛卡尔坐标转换为极坐标，这种变换保留了输入信号的时间信息，即

$$\varphi_i = \arccos \tilde{x}_i, \quad \forall i \in \{1, 2, \cdots, n\} \tag{2}$$

式中，φ_i为角余弦的极坐标。

(3)通过计算每个时间点的极坐标的三角函数差来识别不同时间间隔内的时间相关性，并将其编码到格拉米矩阵的几何结构中，即

$$\boldsymbol{M}_{i,j} = \left[\sin\left(\varphi_i - \varphi_j\right)\right] = \sqrt{\boldsymbol{I} - \boldsymbol{X}_1^2}\,\boldsymbol{X}_2 - \boldsymbol{X}_1\sqrt{\boldsymbol{I} - \boldsymbol{X}_2^2}, \quad \forall i,j \in \{1, 2, \cdots, n\} \tag{3}$$

式中，$\boldsymbol{I} = [1, 1, \cdots, 1]$是单位行向量，$\boldsymbol{X}_1$和$\boldsymbol{X}_2$表示不同的行向量。矩阵的主对角线包含时域信号的原始值和角度信息，利用主对角线，GADF变换将时间序列重建为类似深度学习中的高层特征，进一步将格拉米矩阵转换为图像。GADF变换是非参数化的，不需要事先对数据分布或模型做出假设，适用于各种时间序列数据。此外，GADF变换还能捕捉数据中的非线性关系和时序模式，具有较好的表征能力和鲁棒性[20]。

1.2 生成对抗网络

1.2.1 条件生成对抗网络

生成对抗网络是由Goodfellow等[21]首先提出的一种无监督式的深度学习框架，该框架由生成器G和判别器D两部分组成。如图1所示，选取一组随机噪声$\boldsymbol{z}$作为生成器G的输入，生成器G负责将生成数据$G(\boldsymbol{z})$与原始数据的分布进行匹配，然后将生成数据与真实数据$\boldsymbol{x}$一并传递给判别器D。判别器D负责对生成数据和真实数据进行分辨，并将判别梯度传递给生成器G，以指导生成数据的训练方向。GAN的最终目标为通过生成器G与判别器D的相互博弈，使得判别器D无法分辨数据来源于生成数据还是真实数据，以此得到完美的结果。

由于原始GAN生成的图像的可控性不强，生成目标不明确，CGAN[22]将条件辅助信息$\boldsymbol{y}$引入生成器G和判别器D中，作为生成辅助信息来控制GAN生成的图像，以提高样本的生成质量。CGAN的训练过程与GAN相同，目标函数为

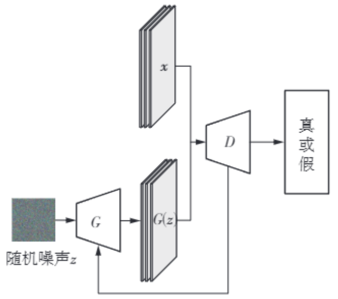


图1 生成对抗网络的结构

Fig. 1 Structure of the generative adversarial network

$$L_{\mathrm{CGAN}} = \min_G \max_D V(D, G) = E_{\boldsymbol{x} \sim p(\boldsymbol{x})}\left[\ln D\left(\boldsymbol{x} \middle| \boldsymbol{y}\right)\right] + E_{\boldsymbol{z} \sim p(\boldsymbol{z})}\left[\ln\left(1 - D\left(G\left(\boldsymbol{z} \middle| \boldsymbol{y}\right)\right)\right)\right] \tag{4}$$

式中，$p(\boldsymbol{x})$为真实数据$\boldsymbol{x}$的概率分布，$p(\boldsymbol{z})$为随机噪声$\boldsymbol{z}$的分布，$D(\boldsymbol{x})$为判别器判断数据是否属于$p(\boldsymbol{x})$的概率，$E_{\boldsymbol{x} \sim p(\boldsymbol{x})}$表示从真实数据分布$p(\boldsymbol{x})$中采样的$\boldsymbol{x}$的期望，$E_{\boldsymbol{z} \sim p(\boldsymbol{z})}$表示从随机噪声分布$p(\boldsymbol{z})$中采样的$\boldsymbol{z}$的期望，$G(\boldsymbol{z})$为生成器输出的假样本。

在GAN与CGAN的训练过程中，使用了Jensen-Shannon(JS)散度来度量生成样本与真实样本之间的差异，JS散度用于最小化生成样本与真实样本之间的分布差距，从而促进生成器G生成更高质量的样本。然而，由于JS散度在两个分布重叠部分的梯度会消失，而在分布之间存在明显差异时，梯度可能会变得非常大，导致在训练过程中可能会出现梯度消失或梯度爆炸等问题，使得训练不稳定。为解决这些问题，文中拟对GAN进行改进，使模型在梯度传播和训练过程中更加稳定。

1.2.2 带有梯度惩罚的Wasserstein GAN

为解决由于JS散度而带来的训练不稳定问题，文献[18]使用Wasserstein距离替代JS散度来衡量生成样本和真实样本之间的距离。相比于JS散度，Wasserstein距离对生成器的梯度信号更明确，使得训练更加稳定。WGAN的目标函数为

$$L_{\mathrm{WGAN}} = \min_G \max_{D \in \Delta} V(D, G) = E_{\boldsymbol{x} \sim p(\boldsymbol{x})}\left[D\left(\boldsymbol{x}\right)\right] - E_{\boldsymbol{z} \sim p(\boldsymbol{z})}\left[D\left(G\left(\boldsymbol{z}\right)\right)\right] \tag{5}$$

式中，Δ为1-Lipschitz函数的集合。由于该函数集中的函数梯度范围不超过1，因此通过限制判别器函数为1-Lipschitz函数集，WGAN可以对判别器的参数进行剪裁(或权重剪裁)，以此限制函数的梯度范围，确保函数的局部变化不会过大，从而有效地防止梯度消失或梯度爆炸。此外，WGAN可以更好地近似Wasserstein距离，从而提高模型的稳定性和生成样本的质量。

然而，使用1-Lipschitz函数集对判别器的权重进行剪裁，虽然可以将各层权重限制在一个很小的范围内，但这种方式往往无法精确地限制判别器函数的梯度范围，容易导致模型出现训练不稳定和梯度消失的问题，从而影响生成样本的质量。为此，Gulrajani等[23]提出了WGAN-GP，使用梯度惩罚

(GP)代替权重剪裁，以解决WGAN中的问题。WGAN-GP的目标函数为

$$L_{\mathrm{WGAN\text{-}GP}} = \min_{G} \max_{D} V(D,G) = -E_{z\sim p(z)}\left[D\left(G(z)\right)\right] + E_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[D(\boldsymbol{x})\right] + \lambda E_{\hat{\boldsymbol{x}}\sim p(\hat{\boldsymbol{x}})}\left[\left(\left\|\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}})\right\|_2 - 1\right)^2\right] \tag{6}$$

式中：λ为梯度惩罚系数；∇为梯度算子；$\hat{\boldsymbol{x}}$为真实数据与生成数据间的随机插值，计算式为

$$\hat{\boldsymbol{x}} = \zeta \boldsymbol{x} + (1-\zeta)G(z) \tag{7}$$

ζ服从[0,1]内的均匀分布。WGAN-GP解决了WGAN中可能存在的模式崩溃和训练不稳定等问题，提高了模型的稳定性、收敛速度和生成样本的质量。

1.2.3　坐标注意力

注意力机制是深度学习中广泛应用的技术，它通过模仿人类视觉的注意力机制，对输入信息的不同部分赋予不同的权重来提高模型对于输入信息的关注度和选择性，从而更加集中地处理重要的信息[24]。不同于传统移动网络中所使用的挤压–激发注意力(SEA)，坐标注意力机制(CAM)[25]更像是一个计算单元，其目的是增强网络对特征的学习能力。

CAM的结构如图2所示，对任意输入张量$\boldsymbol{T} = [\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_C] \in \mathbf{R}^{C\times H\times W}$，其中$C$为输入张量的通道数，$H$和$W$分别为输入数据的高和宽。CAM分别使用两个尺寸为$(H,1)$或$(1,W)$的池化核沿水平坐标和垂直坐标对每个通道进行编码，从而处理输入数据的不同空间方向的特征，使得注意力模块在空间上使用精确的位置信息捕捉远程交互。其计算过程为

$$\boldsymbol{s}_c = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{t}_c(i,j) \tag{8}$$

式中，$\boldsymbol{s}_c$是第c个通道的输出。

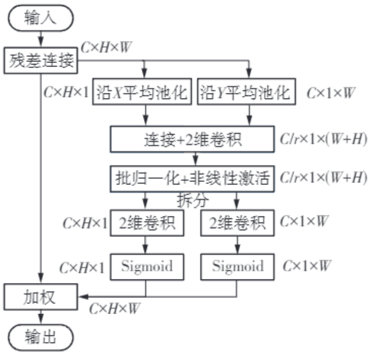


图2　坐标注意力机制的结构

Fig. 2　Structure of coordinate attention mechanism

上述变换分别沿两个空间方向聚合特征，产生一对方向感知特征图。为有效利用该过程产生的特征图，CAM设计了坐标注意力生成。连接生成的聚合特征图，通过一个共享的1×1卷积变换函数F_1，得到水平方向和垂直方向对空间信息进行编码的中间特征图，即

$$\boldsymbol{f} = \delta\left(F_1\left(\left[\boldsymbol{s}^h, \boldsymbol{s}^w\right]\right)\right) \tag{9}$$

式中，$\boldsymbol{f} \in \mathbf{R}^{C/r\times(H+W)}$为中间特征图，[·,·]为沿空间维度的连接操作，$\delta$为非线性激活函数，$r$为控制块大小的缩减比例。

然后，CAM沿空间维度将$\boldsymbol{f}$拆分为两个单独的张量$\boldsymbol{f}^h \in \mathbf{R}^{C/r\times H}$和$\boldsymbol{f}^w \in \mathbf{R}^{C/r\times W}$，利用另外两个1×1卷积变换分别将$\boldsymbol{f}^h$和$\boldsymbol{f}^w$转换为与输入$\boldsymbol{T}$通道数相同的张量$\boldsymbol{g}^h$和$\boldsymbol{g}^w$，随后将输出$\boldsymbol{g}^h$和$\boldsymbol{g}^w$分别拓展并用作注意力权重。CAM最终输出具有增广表示的变换张量$\boldsymbol{Y}$，$\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_C]$，$\boldsymbol{y}_c(i,j) = \boldsymbol{x}_c(i,j) \times \boldsymbol{g}_c^h(i) \times \boldsymbol{g}_c^w(j)$，$\boldsymbol{X}$与$\boldsymbol{Y}$同尺寸。

经实验验证，CAM可以方便地嵌入到现有网络中，从而提升卷积神经网络的性能。

2　小样本故障诊断模型

2.1　GADF样本增强

在基于深度学习的小样本故障检测中，由于样本数过小，深度学习模型无法学习到充分的特征信息，从而导致检测结果不准确。因此，文中引入GADF变换，对1维振动信号进行数据增强。首先，通过GADF变换将1维振动信号转换成大小为$h\times h$的2维原始图像。由式(3)可知，大小为$h\times h$的2维图像是经过h个时间序列点转化而成的。然后，对原始GADF的2维图像进行裁剪，使其生成多幅大小为$l\times l$的子图。为了保证完全裁剪原始图像，需要将h设为l的整数倍，即$h = al$，$a \in \mathbf{N}^{+}$。自此，原始GADF图像被裁减为a^2幅子图像。由于这些子图像包含着不同的特征信息且相互独立，可以作为输入数据，因此可在一定程度上解决样本不足的问题。

2.2　带有梯度惩罚的条件Wasserstein GAN

尽管WGAN-GP通过引入Wasserstein距离和梯度惩罚来解决传统GAN存在的一些训练不稳定性问题，但仍可能遇到训练困难的情况，尤其是在复杂数据集上。此外，WGAN-GP需要调整一些超参数，如惩罚系数等，以确保良好的性能，这些超参数的选择可能会对模型的训练和性能产生较大的影响。

为克服CGAN在训练过程中出现的梯度消失和不稳定问题，以及WGAN-GP存在的缺陷，文中将WGAN-GP与CGAN相结合，一方面通过WGAN-GP的梯度惩罚机制解决CGAN在训练过程中的梯度消失和不稳定问题；另一方面，通过CGAN中的条件辅助信息$\boldsymbol{y}$，使生成器可以根据特定条件生成不同类别的样本，从而增加生成样本的多样性。WGAN-GP与CGAN结合后的目标函数为

$$\begin{aligned} L = \min_G \max_D V(D,G) = E_{\boldsymbol{x} \sim p(\boldsymbol{x})}\left[D\left(\boldsymbol{x}\middle|\boldsymbol{y}\right)\right] - \\ E_{z \sim p(z)}\left[D\left(G\left(z\middle|\boldsymbol{y}\right)\right)\right] + \\ \lambda E_{\hat{\boldsymbol{x}} \sim p(\hat{\boldsymbol{x}})}\left[\left(\left\|\nabla_{\hat{\boldsymbol{x}}} D\left(\hat{\boldsymbol{x}}\middle|\boldsymbol{y}\right)\right\|_2 - 1\right)^2\right] \end{aligned} \tag{10}$$

这样，二者结合的新型生成对抗网络由CGAN提供了对生成过程的更精确控制，可以通过调整条件来指导生成器生成特定类型的样本，而WGAN-GP的梯度惩罚有助于提高生成样本的质量和真实度，并且WGAN-GP的Wasserstein距离和梯度惩罚机制可以帮助模型更好地学习数据分布，而CGAN的条件生成机制可以在学习数据分布的同时，学习条件与生成样本之间的关系。CGAN与WGAN-GP相结合的方法可以使模型具有更好的泛化能力，能够更好地处理不同类型的数据集和任务。

为更有效地提取输入信息的局部和全局特征，文中使用卷积神经网络作为生成器G和鉴别器D的主体结构。

此外，动态权重调整[26]在视觉超分辨率重建等任务中可以很好地调整注意力权重和非注意力权重，通过自动舍去一些不重要的注意力特征使两个分支达到动态平衡，从而充分利用特征图所包含的信息。因此，文中设计动态坐标注意力机制（DCAM），其结构如图3所示。坐标注意力提取到特征后，进一步将特征输入至动态权重模块中，并通过残差块进行连接。动态权重模块包含3个分支：非注意力分支、注意力分支和动态权重融合分支。非注意力分支包含一个1×1卷积进行通道调节，以及一个3×3卷积进行特征映射。注意力分支包含在经过一个1×1卷积后将特征数据输入至一个3×3卷积和像素注意力块（PAB）中，该模块可以为不同的通道分配不同的权重。再次经过一个3×3卷积后使用1×1卷积进行特征重组，以便与动态权重融合分支进行权重融合。动态融合分支可以利用加权求和的方式为注意力分支和非注意力分支分配不同的权重λ_1和λ_2，进行对应元素相加后，与动态权重模块的残差相加，得到最终的特征。

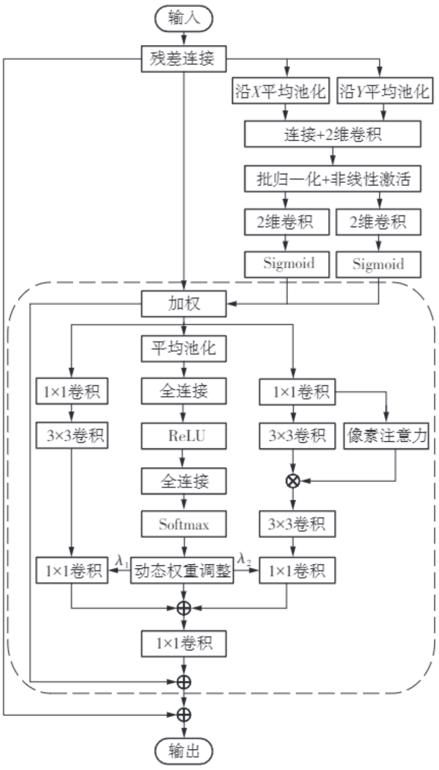


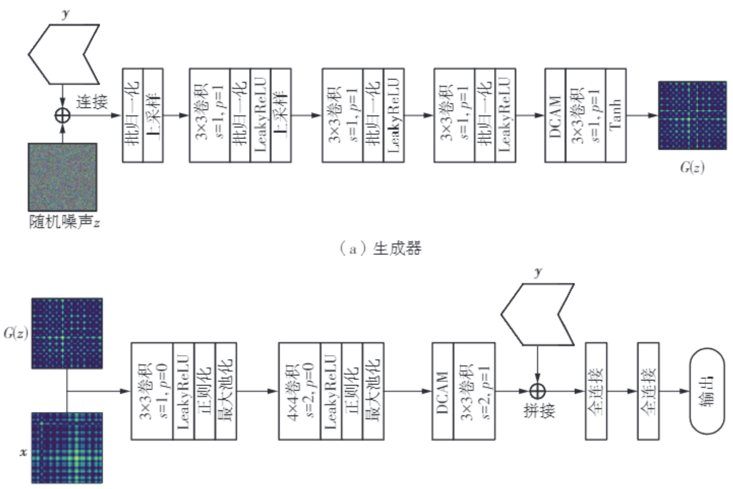
图3 动态坐标注意力机制的结构

Fig. 3 Structure of the dynamic coordinate attention mechanism

引入DCAM可以增强模型空间感知能力，提升整体泛化性能和鲁棒性。文中所提出的生成对抗网络的生成器和判别器结构如图4所示。

如图4(a)所示，生成器G的输入为随机噪声z和辅助信息$\boldsymbol{y}$的特征向量拼接在一起的新向量，通过一系列的卷积层、批归一化层和LeakyReLU激活函数，对输入特征进行处理和转换；在最后一个卷积层前引入DCAM，用于增强网络对空间信息的处理能力，最后通过Tanh激活函数对生成的图像进行归一化。其中，通过上采样操作逐渐将低分辨率特征映射转换为高分辨率的图像，从而创建细节更加丰富的图像。

如图4(b)所示，以生成器生成图像$G(z)$和真实图像$\boldsymbol{x}$为输入数据，经过一系列卷积层和激活函数进行特征提取与空间降维。同样地，在最后一个卷积层前加入DCAM。在鉴别器D中使用最大池化



（a）生成器

（b）判别器

图4　生成器和判别器的结构

Fig. 4　Structure of generator and discriminator

以减少特征图的空间尺寸，从而降低计算复杂度和参数数量。最后，将卷积结果与辅助信息$\boldsymbol{y}$进行拼接，拼接后的向量经过一系列线性层，最终输出一个标量值，表示输入图像为真实图像的概率。

2.3　小样本故障诊断流程

针对故障诊断过程中的样本不足问题，文中提出了一种基于GADF变换和生成对抗网络的小样本滚动轴承故障诊断模型。模型的总体框架与流程图分别如图5和图6所示，具体的诊断步骤如下：①利用GADF变换将1维信号转换为GADF图像，并将原始图像裁剪为子图像，初步达到数据增强的目的；②将裁剪后的数据按比例分为真实数据和验证集；③使用真实数据训练生成器G和判别器D，从而生成高质量数据，用于丰富样本多样性；④将生成数据与真实数据合并，按比例划分为训练集和测试集，

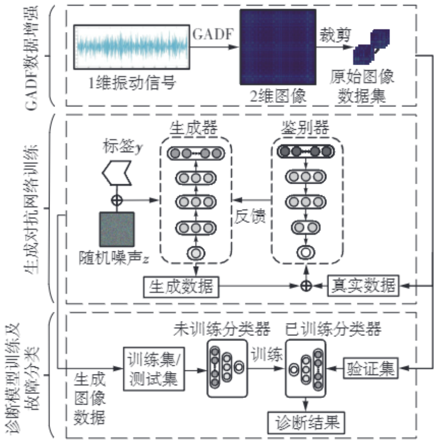


图5　文中诊断模型的总体框架

Fig. 5　General framework of the proposed diagnostic model

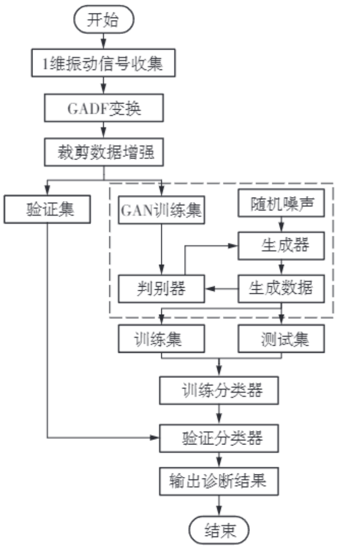


图6　文中诊断模型流程图

Fig. 6　Flowchart of the proposed diagnostic model

用以训练分类模型；⑤将验证集数据输入至训练好的分类模型中，输出诊断结果。

3 实验与结果分析

分别采用东南大学轴承数据集和美国凯斯西储大学(CWRU)轴承数据集作为实验数据。实验用的计算机配置为AMD锐龙R7-5800H CPU，NVIDIA GeForce RTX 3060(12GB)和16GB RAM，实验框架为PyTorch & Python。

3.1 实例1

实验数据由东南大学机械工程学院在传动系动力模拟器上采集，试验平台如图7所示。系统的工况为20Hz(1200r/min)-0V，该工况下有5类故障类型，分别为正常、滚动体故障、复合故障、内圈故障和外圈故障。因此，在实例1中，任务是对5种滚动轴承故障进行准确诊断。

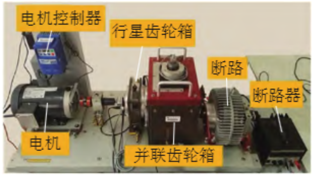


图7 东南大学数据采集平台

Fig. 7 Data collection platform of Southeast University

为应对小样本问题，设置采样点个数为1024，并进行滑动采样。对每一类故障数据进行随机截取，分别截取5个和10个样本。然后，使用GADF变换将这些样本转换为1024×1024大小的图像，分别生成了5幅和10幅图像。为了满足神经网络的输入需求，对这些图像进行裁剪，将其裁剪为128×128大小的子图像。由于1024可以被128整除，每幅1024×1024大小的图像可以被裁剪为64幅128×128大小的子图像。因此，在小样本故障诊断中，原始的5个和10个1维样本经过数据增强操作后，分别生成了320幅和640幅2维图像。接着，按照7:3的比例将这些图像分为原始数据样本和验证样本。实验数据详情如表1所示。

表1 实例1的实验数据详情

Table 1 Details of experimental data for case 1

数据标签	故障类型	10样本裁剪划分的样本数		5样本裁剪划分的样本数	
		原始数据	验证集	原始数据	验证集
00	正常	448	192	224	96
01	滚动体故障	448	192	224	96
02	复合故障	448	192	224	96
03	内圈故障	448	192	224	96
04	外圈故障	448	192	224	96

表2 参数设置

Table 2 Settings of parameters

方法	学习率	优化器	Batch Size	Epoch
CGAN	0.00020	Adam	32	1000
WGAN	0.00005	RMSprop	32	1000
WGAN-GP	0.00020	Adam	32	1000
文中方法	0.00020	Adam	32	1000

使用CGAN、WGAN和WGAN-GP作为对比算法进行实验，对比的生成对抗网络以及文中方法的具体参数设置如表2所示。此外，为了防止单分类器对诊断效果的影响，同时选择3种性能优秀的分类模型(MobileNetV3、ResNet34和GhostNet)进行比较，分类器优化器均设置为Adam优化函数，学习率为0.001，迭代次数为30。

4种生成对抗网络在10样本下的损失函数随着迭代次数的变化如图8所示。由图中可以看出：虽然CGAN在训练初期能够很快收敛在0附近，但其损失函数变化非常剧烈，在迭代次数不足10000时出现梯度爆炸，并在之后的训练中多次出现梯度爆炸现象，而采用Wasserstein距离的其余3个生成对抗网络，其训练过程要比CGAN稳定，表明使用Wasserstein距离可以提高训练稳定性；文中所提模型的收敛速度最快，在迭代2000次左右后便完成收敛，模型为样本生成做好了准备。鉴于不同生成对抗网络所使用的损失函数不同，具体的样本生成质量需要进一步观察分类器的训练结果。

如表3所示，仅通过原始数据集对3种分类模型进行训练，诊断效果较差。分析其原因，可能是：尽管使用GADF变换与裁剪进行数据增强，但原始数据集样本数量仍然不足，无法充分代表整个数据集的分布，模型很难学习到足够的信息进行准确分类；训练数据较少，更容易导致过拟合，模型性能无法较好地泛化在验证集中。而通过4种生成对抗网络模型对原始数据进行补充后，诊断结果得到了较大的提升，在10样本条件下，采用文中所提方法对原始数据进行增强后，MobileNetV3模型的诊断准确率为99.33%，ResNet34和GhostNet的诊断准确率分别为96.35%和96.25%，明显优于其余生成对抗网络模型的准确率。这是因为文中方法不仅利用CGAN的条件辅助信息控制样本的特定

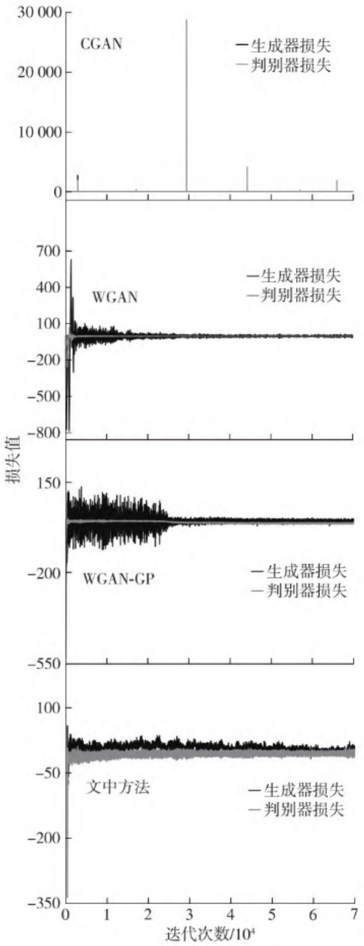


图8　不同生成对抗网络的损失函数随着迭代次数的变化曲线

Fig. 8　Changing curves of loss function with iteration number for different GANs

属性，提高样本的生成质量，同时通过WGAN-GP中的梯度惩罚机制使模型具有更稳定的训练过程。此外，文中方法的生成器和判别器中的卷积神经网络结构能够更有效地提取图像信息，使生成器和判别器更容易学习。因此，文中方法可以生成更逼真的数据，为分类器准确分类提供保障。从表中还可以发现，5样本条件下的诊断准确率整体低于10样本条件下，这是因为10样本条件下的数据包含的信息多于5样本条件，模型可以学习更多的信息。

为进一步分析基于不同生成对抗网络模型的分类结果，文中选取5样本条件并使用MobileNetV3作为分类器，通过引入混淆矩阵来显示故障诊断结果，如图9所示。由图9(a)可知，当仅使用原始样本做训练集时，大量样本被错误分类，诊断效果无法令人满意。由图9(b)至图9(e)可知，采用3种生成对抗网络(CGAN、WGAN、WGAN-GP)对数据进行强化后，虽然整体诊断准确率有所提高，但仍然存在某类故障诊断错误率接近20%的情况，而采用文中方法对数据进行强化后，在所有故障类别中均取得了90%以上的分类准确率。这表明，文中方法可提升小样本情况下的故障诊断性能。

3.2　实例2

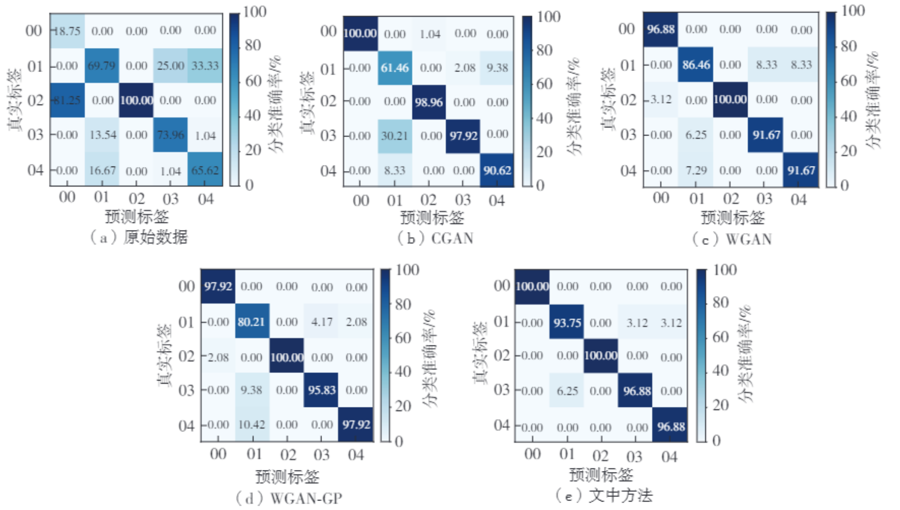
为验证文中方法的先进性，选取MobileNetV3作为分类器，并与4种新的小样本故障诊断方法进行对比。CGAN+2DCNN[16]是基于深度对抗网络和灰度图的方法，MAML[27]是基于模型不可知论的元学习故障诊断方法，TRN[28]是融合小样本学习机制与迁移学习的新型迁移关系网络，GWDConv[29]是基于离散图小波框架的图小波去噪卷积方法。

实验数据集来自美国凯斯西储大学，实验平台如图10所示。该数据集包含4种故障类型，分别为正常、滚珠故障、内圈故障和外圈故障。此外，根据故障尺寸，将每种故障细分为0.017 78、0.035 56、0.053 34和0.071 12 cm。文中将采集在工作负载为0 W下的不同位置以及不同尺寸的故障分为12个标签(包括1个正常标签和11个故障标签)。因此在实例2中，任务是对12种滚动轴承故障进行准确

表3　实例1的轴承故障诊断结果

Table 3　Bearing fault diagnosis results for case 1

方法	10样本下的诊断准确率/%			5样本下的诊断准确率/%		
	MobileNetV3	ResNet34	GhostNet	MobileNetV3	ResNet34	GhostNet
原始数据	71.67	68.23	63.54	63.75	62.91	62.29
CGAN	94.31	91.77	90.73	91.08	88.13	89.59
WGAN	94.21	92.44	93.76	93.79	90.13	89.58
WGAN-GP	95.98	94.83	94.10	94.01	92.75	93.03
文中方法	99.33	96.35	96.25	97.74	94.96	93.79



(a) 原始数据

(b) CGAN

(c) WGAN

(d) WGAN-GP

(e) 文中方法

图9　5样本下不同方法的混淆矩阵

Fig. 9　Confusion matrix of different methods with 5 samples

诊断。对数据的GADF变换处理方式与实例1相同，实验数据详情如表4所示。实验结果如表5所示，从表中可知，在5样本和10样本实验条件下，文中方法的诊断准确率均明显高于4种对比方法，表明了文中方法的优越性。

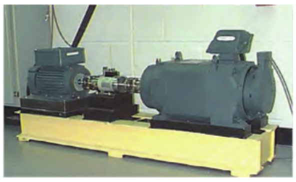


图10　CWRU数据采集平台

Fig. 10　CWRU data acquisition platform

为进一步说明文中方法的先进性，采用另外4个指标来判断诊断结果，分别为精确率P、召回率R、特异性S和F_1值[30]。精确率衡量模型预测为正例的准确性；召回率衡量模型对正例的查全率；特异性衡量模型对负例的识别能力；F_1值同时考虑精确率和召回率，衡量模型的综合性能。实验结果如图11所示。从图中可以看出，在两种实验条件下，文中方法的各项指标结果均优于对比方法，同时具有最小的数据误差，这进一步说明了文中方法的先进性。

为进一步分析GADF变换与图像裁剪相结合对小样本故障检测性能的影响，文中对比了其他几种

表4　实例2实验数据详情

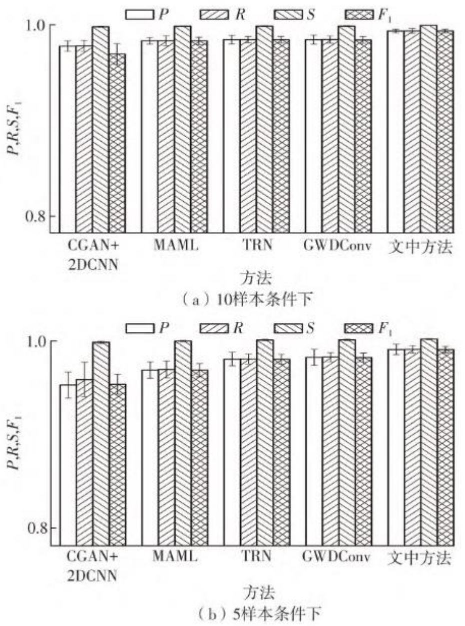
Table 4　Details of experimental data for case 2

数据标签	故障位置	故障尺寸/cm	5样本裁剪划分的样本数		10样本裁剪划分的样本数	
			原始数据	验证集	原始数据	验证集
00	正常		224	96	448	192
01	BF	0.01778	224	96	448	192
02	IF	0.01778	224	96	448	192
03	OF	0.01778	224	96	448	192
04	BF	0.03556	224	96	448	192
05	IF	0.03556	224	96	448	192
06	OF	0.03556	224	96	448	192
07	BF	0.05334	224	96	448	192
08	IF	0.05334	224	96	448	192
09	OF	0.05334	224	96	448	192
10	BF	0.07112	224	96	448	192
11	IF	0.07112	224	96	448	192

表5　实例2的轴承故障诊断结果

Table 5　Bearing fault diagnosis results for case 2

方法	准确率/%	
	10样本	5样本
CGAN+2DCNN	97.74	95.05
MAML	98.31	96.61
TRN	98.64	97.79
GWDConv	98.44	97.96
文中方法	99.35	98.78



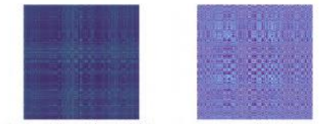
（a）10样本条件下

（b）5样本条件下

图11　几种方法的分类性能对比

Fig. 11　Comparison of classification performance of several methods

将1维信号转化为2维图像的方法。其中，1维信号生成分辨率为1 024×1 024灰度图要求原始数据长度过长，无法完成转化；基于连续小波变换的方法经过裁剪后出现较严重的特征碎片化现象，无法进行实验。因此，文中分别使用递归图、马尔可夫迁移场，将1维数据转换为2维图像后进行裁剪，再输入至文中所提出的生成对抗网络中进行训练。递归图和马尔可夫迁移场的数据转换结果如图12所示，轴承故障诊断结果如表6所示。



（a）马尔可夫迁移场　　（b）递归图

图12　不同转换方法所得特征图

Fig. 12　Characteristic pattern obtained by different conversion methods

从表6和前面的实验结果可以看出，使用经GADF变换并裁剪后得到的数据进行故障诊断，可以获得最高的准确率。这是因为GADF矩阵中每个位于非对角线上的值是不同的，因此经过裁剪后的GADF子图像互相独立，每一幅子图像都具有不同的特征，不会出现严重的特征碎片化现象，因此利用GADF变换之后再进行裁剪，对处理小样本下的滚动轴承故障诊断是有效的。

表6　不同数据转换方法的故障诊断结果

Table 6　Fault diagnosis results of different data conversion methods

方法	准确率/%	
	10样本	5样本
马尔可夫迁移场	91.53	87.96
递归图	93.21	88.38

4　结语

文中提出了一种基于GADF和生成对抗网络的轴承故障诊断方法，旨在解决小样本环境下滚动轴承故障诊断效果不佳的问题。首先，通过GADF变换将1维信号转换为GADF图像并进行裁剪，从而得到大量的图像样本，完成对样本数据的初步强化；接着，设计新的生成对抗网络模型，该模型融合了CGAN和WGAN-GP的优点，并使用卷积神经网络作为生成器和判别器的结构，使其可以更好地学习样本特征；然后，设计动态坐标注意力以进一步加强模型的空间感知能力；最后，使用东南大学轴承数据和CWRU轴承数据进行实验。结果表明：在2种小样本条件下，文中方法可以对不同的故障类型进行准确地分类；文中方法的精确率、召回率、特异性和F_1值均优于新型小样本故障诊断方法，进一步说明了文中方法的先进性。

尽管文中方法针对小样本故障诊断问题具有良好的性能，但生成对抗网络对图像的生成训练需要较长的训练时间和计算资源，这将导致诊断所需时间较长，对设备的要求较高。因此，未来将针对轻量化生成对抗网络和1维生成对抗网络做进一步研究。

参考文献：

[1] HUANG R，XIA J，ZHANG B，et al. Compound fault diagnosis for rotating machinery：state-of-the-art，challenges，and opportunities［J］. Journal of Dynamics，Monitoring and Diagnostics，2023，2(1)：13-29.

[2] WANG Q，XU F. A novel rolling bearing fault diagnosis method based on adaptive denoising convolutional neural network under noise background［J］. Measurement，2023，218：113209/1-13.

[3] 陈新度，扶治森，吴智恒，等. 基于多头卷积和差

分自注意力的小样本故障诊断方法［J］. 华南理工大学学报(自然科学版)，2023，51(7)：21-33.

CHEN Xindu，FU Zhisen，WU Zhiheng，et al. Small-sample fault diagnosis method based on multi-head convolution and differential self-attention ［J］. Journal of South China University of Technology (Natural Science Edition)，2023，51(7)：21-33.

［4］ NING S，WANG Y，CAI W，et al. Research on intelligent fault diagnosis of rolling bearing based on improved ShufflenetV2-LSTM ［J］. Journal of Sensors，2022，2022：8522206/1-13.

［5］ 陈仁祥，唐林林，胡小林，等. 不同转速下基于深度注意力迁移学习的滚动轴承故障诊断方法［J］. 振动与冲击，2022，41(12)：95-101，195.

CHEN Renxiang，TANG Linlin，HU Xiaolin，et al. A rolling bearing fault diagnosis method based on deep attention transfer learning at different rotations ［J］. Journal of Vibration and Shock，2022，41(12)：95-101，195.

［6］ ZHANG X，ZHAO B，LIN Y. Machine learning based bearing fault diagnosis using the Case Western Reserve University data：a review ［J］. IEEE Access，2021，9：155598-155608.

［7］ ZHANG J，ZHANG K，AN Y，et al. An integrated multitasking intelligent bearing fault diagnosis scheme based on representation learning under imbalanced sample condition ［J］. IEEE Transactions on Neural Networks and Learning Systems，2024，35(5)：6231-6242.

［8］ REN C，JIANG B，LU N. Task adaptation meta learning for few-shot fault diagnosis under multiple working conditions ［C］// Proceedings of 2023 the 6th International Symposium on Autonomous Systems. Nanjing：IEEE，2023：10164461/1-5.

［9］ INDIRA V，VASANTHAKUMARI R，SUGUMARAN V. Minimum sample size determination of vibration signals in machine learning approach to fault diagnosis using power analysis ［J］. Expert Systems with Applications，2010，37(12)：8650-8658.

［10］ LIU X，HUANG H，XIANG J. A personalized diagnosis method to detect faults in a bearing based on acceleration sensors and an FEM simulation driving support vector machine ［J］. Sensors，2020，20(2)：420/1-13.

［11］ LIU X，HUANG H，XIANG J. A personalized diagnosis method to detect faults in gears using numerical simulation and extreme learning machine ［J］. Knowledge-Based Systems，2020，195：105653/1-13.

［12］ HU Y，XIONG Q，ZHU Q，et al. Few-shot transfer learning with attention for intelligent fault diagnosis of bearing ［J］. Journal of Mechanical Science and Technology，2022，36(12)：6181-6192.

［13］ CHEN J，HU W，CAO D，et al. A meta-learning method for electric machine bearing fault diagnosis under varying working conditions with limited data ［J］. IEEE Transactions on Industrial Informatics，2022，19(3)：2552-2564.

［14］ XIA P C，HUANG Y X，WANG Y X，et al. Augmentation-based discriminative meta-learning for cross-machine few-shot fault diagnosis ［J］. Science China Technological Sciences，2023，66(6)：1698-1716.

［15］ HAN Y，LI B，HUANG Y，et al. Imbalanced fault classification of rolling bearing based on an improved oversampling method ［J］. Journal of the Brazilian Society of Mechanical Sciences and Engineering，2023，45(4)：223/1-11.

［16］ YANG J，LIU J，XIE J，et al. Conditional GAN and 2-D CNN for bearing fault diagnosis with small samples ［J］. IEEE Transactions on Instrumentation and Measurement，2021，70：3525712/1-12.

［17］ FAN H，MA J，ZHANG X，et al. Intelligent data expansion approach of vibration gray texture images of rolling bearing based on improved WGAN-GP ［J］. Advances in Mechanical Engineering，2022，14(3)：1-11.

［18］ ARJOVSKY M，CHINTALA S，BOTTOU L. Wasserstein generative adversarial networks ［C］// Proceedings of the 34th International Conference on Machine Learning. Sydney：MLResearchPress，2017：214-223.

［19］ WANG Z，OATES T. Imaging time-series to improve classification and imputation ［EB/OL］.(2015-06-01)［2023-11-27］. http://arxiv.org/abs/1506.00327.

［20］ THANARAJ K P，PARVATHAVARTHINI B，TANIK U J，et al. Implementation of deep neural networks to classify EEG signals using gramian angular summation field for epilepsy diagnosis ［EB/OL］.(2020-05-08)［2023-11-27］. https://arxiv.org/abs/2003.04534.

［21］ GOODFELLOW I，POUGET-ABADIE J，MIRZA M，et al. Generative adversarial nets ［C］// Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge：MIT Press，2014：2672-2680.

［22］ MIRZA M，OSINDERO S. Conditional generative adversarial nets ［EB/OL］.(2014-11-06)［2023-11-27］. https://arxiv.org/abs/1411.1784.

［23］ GULRAJANI I，AHMED F，ARJOVSKY M，et al. Improved training of Wasserstein GANs ［C］// Proceedings of the 31st International Conference on Neural

Information Processing Systems. Red Hook: Curran Associates Inc., 2017: 5769-5779.

[24] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module [C]// Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018: 3-19.

[25] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design [C]// Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Taipei: IEEE, 2021: 13713-13722.

[26] CHEN H, GU J, ZHANG Z. Attention in attention network for image super-resolution [EB/OL]. (2021-04-19) [2023-11-27]. https://arxiv.org/abs/2104.09497.

[27] ZHANG S, YE F, WANG B, et al. Few-shot bearing fault diagnosis based on model-agnostic meta-learning [J]. IEEE Transactions on Industry Applications, 2021, 57(5): 4754-4764.

[28] LU N, HU H, YIN T, et al. Transfer relation network for fault diagnosis of rotating machinery with small data [J]. IEEE Transactions on Cybernetics, 2021, 52(11): 11927-11941.

[29] LI T, SUN C, LI S, et al. Explainable graph wavelet denoising network for intelligent fault diagnosis [J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 35(5): 8535-8548.

[30] WANG L, ZHANG L, QI X, et al. Deep attention-based imbalanced image classification [J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(8): 3320-3330.

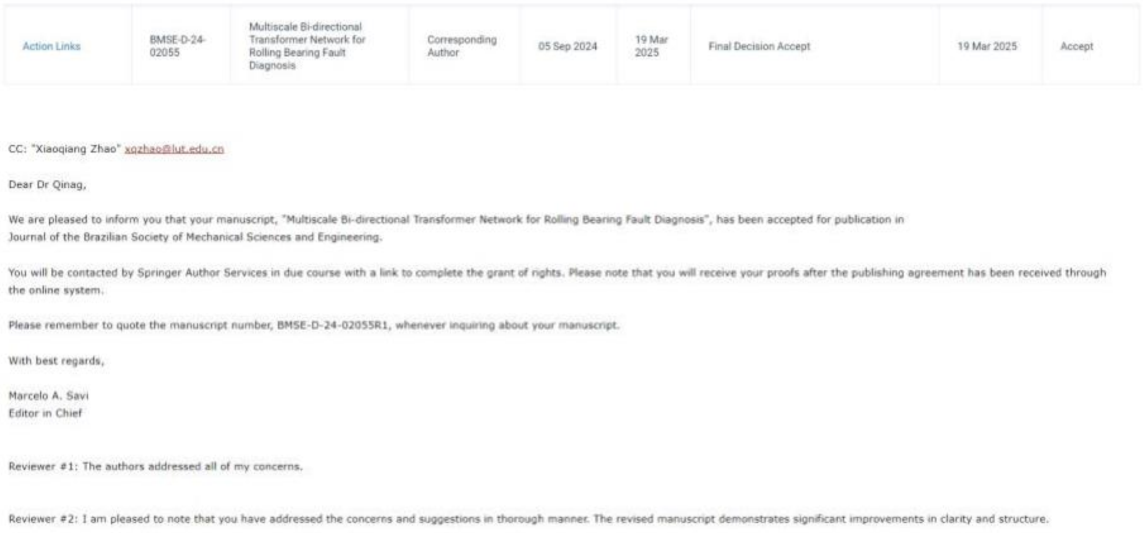
A Small Sample Rolling Bearing Fault Diagnosis Method Based on Gramian Angular Difference Field and Generative Adversarial Network

QIANG Ruiru　ZHAO Xiaoqiang

(College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, Gansu, China)

Abstract: Aiming at the problem that deep learning-based rolling bearing fault diagnosis algorithms need to learn from a large amount of labeled data and face poor diagnosis effect when the number of samples is limited, this paper proposed a small-sample rolling bearing fault diagnosis method based on the Gramian angular difference field (GADF) and generative adversarial networks (GAN). Firstly, a data enhancement method based on GADF transform was proposed, and it converts a few 1D vibration signals into 2D GADF images by GADF transform. GADF subgraphs are obtained by cropping to obtain a large number of image samples. Then, a conditional generative adversarial network (CGAN) was combined with Wasserstein GAN with gradient penalty (WGAN-GP) to construct a novel generative adversarial network, which enhances the model training stability by conditional auxiliary information with gradient penalty and designs dynamic coordinate attention mechanism to enhance the spatial perception of the model, so as to generate high-quality samples. Finally, the generative samples were used to train the classifier, and the diagnosis results were obtained on the validation set. Two sets of bearing fault diagnosis experiments in a small sample environment were conducted using the Southeast University dataset and the Case Western Reserve University dataset, respectively. The results show that, compared with traditional generative adversarial networks as well as advanced small-sample fault diagnosis methods, the proposed method can obtain the best results in five fault diagnosis metrics, including accuracy and precision, and can accurately diagnose the type of bearing faults under small-sample conditions.

Key words: small sample bearing fault diagnosis; Gramian angular difference field; generative adversarial network; attention mechanism



Action Links	BMSE-D-24-02055	Multiscale Bi-directional Transformer Network for Rolling Bearing Fault Diagnosis	Corresponding Author	05 Sep 2024	19 Mar 2025	Final Decision Accept	19 Mar 2025	Accept

CC: "Xiaoqiang Zhao" xqzhao@lut.edu.cn

Dear Dr Qinag,

We are pleased to inform you that your manuscript, "Multiscale Bi-directional Transformer Network for Rolling Bearing Fault Diagnosis", has been accepted for publication in Journal of the Brazilian Society of Mechanical Sciences and Engineering.

You will be contacted by Springer Author Services in due course with a link to complete the grant of rights. Please note that you will receive your proofs after the publishing agreement has been received through the online system.

Please remember to quote the manuscript number, BMSE-D-24-02055R1, whenever inquiring about your manuscript.

With best regards,

Marcelo A. Savi
Editor in Chief

Reviewer #1: The authors addressed all of my concerns.

Reviewer #2: I am pleased to note that you have addressed the concerns and suggestions in thorough manner. The revised manuscript demonstrates significant improvements in clarity and structure.

Multiscale Bi-directional Transformer Network for Rolling Bearing Fault Diagnosis

Qiang Ruiru[1]. Zhao Xiaoqiang[1].

(College of Electrical and Information Engineering, Lanzhou University of Technology, Gansu Lanzhou 730050, China)

Abstract: In industrial applications, the vibration signals of rolling bearings are often subjected to strong noise interference, variations in operating conditions, and fluctuating rotational speeds, resulting in high signal complexity and challenging fault diagnosis. Recent studies have leveraged the synergy between the Transformer's multi-head self-attention mechanism and convolutional networks to enhance feature extraction. However, these approaches often introduce excessive model complexity, leading to high computational costs and limiting their deployment in real-world industrial scenarios. To address these challenges, this paper proposes a lightweight Multi-scale Bi-directional Self-attentive Diagnosis Method (MBSADM). First, a multi-scale attention mechanism is designed to effectively capture discriminative features across different scales of vibration signals. Second, a multi-scale feature extraction module integrates multi-scale dilated convolution blocks with the multi-scale attention mechanism, enabling a multi-local receptive field with reduced computational overhead and fewer model parameters. Finally, to fully exploit temporal dependencies, we introduce a bi-directional Transformer that leverages a reverse mechanism to construct sequence representations containing spatially inverted information, thereby enhancing the temporal modeling capability of extracted features. Extensive experiments under strong noise, different load, and fluctuating speed conditions demonstrate the robustness and superior classification performance of the proposed MBSADM. Compared to five state-of-the-art fault diagnosis methods, MBSADM achieves higher diagnostic accuracy and demonstrates stronger industrial applicability, making it a promising solution for real-world bearing fault detection.

Keywords: Deep learning, rolling bearing fault diagnosis, multi-scale attention mechanism, Bi-directional Transformer

1. Introduction

Rotating machinery is widely used in industrial production, and rolling bearings, as key components, play a crucial role in ensuring equipment safety and stability [1, 2]. Due to prolonged exposure to complex operating environments, rolling bearings are susceptible to fatigue, pitting, and overload, which may ultimately lead to mechanical failures [3, 4]. Therefore, developing efficient and accurate fault diagnosis methods is essential for mitigating production risks and reducing economic losses [5, 6].

Traditional rolling bearing fault diagnosis methods primarily rely on signal processing techniques and expert knowledge. While these methods perform well under specific working conditions, they struggle to adapt to highly dynamic environments [7]. In recent years, deep learning has gained widespread adoption in fault diagnosis due to its powerful feature extraction and automatic learning capabilities. Among these methods, convolutional neural network (CNN)-based end-to-end approaches eliminate the need for manual feature engineering [8], directly extracting key features from raw signals and achieving high-precision fault classification [9]. Guo et al[10] proposed an end-to-end fault diagnosis approach that integrates attention-based CNNs with bidirectional long short-term memory networks (BiLSTM). Dong et al[11] designed a one-dimensional attention-enhanced neural network based on empirical wavelet transform to address the non-stationarity and non-linearity of rolling bearing vibration signals. Lin et al[12] tackled the limitations of conventional CNNs, which focus solely on single-scale features while neglecting multi-scale deep information, by proposing an improved multi-scale attention-based CNN for bearing fault diagnosis. Additionally, Jia et al[13] introduced a denoising strategy based on the periodic self-similarity of vibration signals and leveraged an end-to-end CNN model to suppress irrelevant noise in vibration signals.

Although CNNs effectively extract local features from input data without requiring prior knowledge, they primarily capture localized patterns. When fault signals are affected by noise interference or variations in rotational speed, relying solely on local information makes it difficult to extract global fault patterns[14]. Furthermore, CNNs employ a shared weight mechanism, which lacks the ability to model long-term dependencies in time-series data[15].

To enhance global feature extraction, Transformer models have recently been introduced into the fault diagnosis domain. By leveraging the self-attention mechanism, Transformers effectively model long-range dependencies and have achieved remarkable success in natural language processing (NLP)[16] and computer vision (CV)[17]. Some studies have attempted to apply Transformers to rolling bearing fault diagnosis. For example, Hou et al[18] proposed a fault diagnosis model combining fast fourier transform (FFT) with Transformer networks, while Tang et al[19] employed discrete wavelet transform to decompose vibration signals into sub-signals across different frequency bands, which were then fed into independent Transformer models for diagnosis. However, Transformers require large-scale data for training, whereas vibration signal datasets are typically limited in size[20], making it challenging for the model to learn complex fault patterns effectively. Additionally, Transformers struggle with capturing fine-grained local details, limiting their ability to extract multi-scale features from bearing vibration signals[21].

To address the respective shortcomings of CNNs and Transformers, researchers have proposed hybrid CNN-Transformer approaches, which leverage CNNs for local feature extraction and Transformers for global modeling. For instance, Gao et al[22] developed a fault diagnosis model that integrates CNNs with a dual-channel Transformer to achieve collaborative extraction of local and global features. Liu et al[23] proposed a lightweight diagnostic framework based on multi-scale convolution and broadcast self-attention, designed to handle varying rotational speeds.

In summary, although CNN-Transformer-based fault diagnosis methods have demonstrated promising results, several challenges remain: (1) CNNs require additional convolutional kernels to extract multi-scale features, significantly increasing computational complexity; (2) existing attention mechanisms often focus on single-scale features while overlooking the complex spatiotemporal dependencies in vibration signals; and (3) most methods adopt a unidirectional modeling approach that only considers historical information, failing to fully utilize future contextual information to enhance fault pattern representation.

To address these issues, this paper proposes a multi-scale bidirectional self-attention fault diagnosis method. The proposed method directly processes raw data without requiring any preprocessing, reducing dependence on domain-specific signal processing expertise. Furthermore, by constructing a multi-scale feature extraction module and a bidirectional Transformer module, our approach effectively captures features across different scales and learns richer contextual information. The main contributions of this paper are as follows:

(1) Construction of a multi-scale attention mechanism using squeeze-and-excitation networks and convolutional neural networks. Specifically, to enable the attention mechanism to aggregate information from different spatial scales, we first utilize the squeeze-and-excitation (SE) module and convolutional modules to extract multi-scale spatial information separately, followed by mutual weighting. Then, the spatial attention maps from both scales are summed to generate weights that are applied to the original features.

(2) Proposal of a multi-scale feature extraction module combining dilated convolutions and multi-scale attention mechanisms. By incorporating dilated convolutions, the network's feature learning capability is enhanced while effectively reducing the computational cost of convolution operations.

(3) Construction of a bidirectional transformer (Bi-Transformer) for temporal feature enhancement. By leveraging rolling bearing time-series data from both past and future time steps, the proposed Bi-Transformer enhances the temporal modeling capability of the network. This allows the model to capture both historical and future temporal contexts within vibration signals, leading to more comprehensive fault pattern representation.

The rest of the paper is organized as follows. Section 2 presents the theoretical background involved in the proposed method. Section 3 presents the multiscale bi-directional self-attentive fault diagnosis method. Section 4 presents an experimental evaluation of the proposed method. Section 5 concludes the paper.

2. Theoretical background

2.1. Convolution and Depth Separable Convolution

CNN have made very significant achievements in the field of image processing[24, 25]. In order to solve the rolling bearing fault diagnosis problem, the researchers converted 2D CNN to 1D CNN to analyze 1D timing signals. For conventional 1D CNNs, the width of the input features is assumed to be W, the size of the convolutional kernel is k, and C_{in} and C_{out} are the number of channels for the input and output data, respectively. The operation of one-dimensional convolution is described as follows:

$$k \times C_{in} \times C_{out} \times W \tag{1}$$

Depth separable convolution[26] consists(DSC) of depth convolution (DW) and pointwise convolution (PW), where PW convolution is the traditional convolution with convolution kernel 1. Each convolution channel in DW convolution is 1. Therefore, DW convolution requires convolution of each channel of the input data, and the number of channels of the output data is equal to the number of channels of the input data. Once the output features are obtained using DW convolution, the number of output data channels is customized using PW convolution. The operational cost of DS convolution is the sum of DW convolution and PW convolution, described as follows:

$$k \times C_{in} \times W + C_{in} \times C_{out} \times W \tag{2}$$

Compared with the traditional 1D convolution, the computational cost of DS convolution is $\frac{1}{C_{out}} + \frac{1}{k}$ times less than that of traditional convolution, which reduces the computational cost and realizes the lightweighting of the model.

2.2. Dilated Convolution

Receptive field is an important concept in CNNs, which represents the process in which a neuron receives a portion of the input image and performs feature extraction based on this information[27]. In layman's terms, the receptive field is the ratio of the individual pixels of the feature map to the pixels of the original image during the convolution process. The larger the receptive field, the richer the information in the original image contained in the feature map. The dilation convolution changes the receptive field by introducing a hyperparameter "dilation rate" to control the spacing of neighboring samples of the convolution kernel. As shown in Figure 1, the receptive field of a conventional one-dimensional convolution with a kernel size of 3×1 is 3. When the dilation rate is 2, the receptive field of the same 3×1 convolutional kernel dilates from 3 to 7. The same receptive field size requires a convolution kernel of 5 to do so, but the number of parameters in the dilation convolution is much smaller than that of the traditional convolution, which greatly reduces the computational cost.

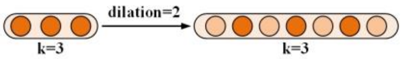


Figure 1. 1D dilation convolution

2.3. Self-attention mechanism (SA)

Attention mechanisms are widely used in CNNs for image processing tasks. The aim is to give CNN the ability to focus on and understand key regions of an image similar to the human eye. The core idea of the attention mechanism is to allow CNN to focus on important information in the input data while ignoring unimportant details. The self-attention mechanism[28] (SA) has achieved excellent results in tasks in the field of NLP since its proposal. Due to its excellent contextual understanding, SA began to be introduced by researchers into the field of computer vision.

The structure of SA is shown in Figure 2. Let X be the input data, the query Q, the key K and the value V are obtained by linear transformation. Subsequently, for each position i in the sequence, the attention score between it and all positions j in the sequence is computed, described as follows:

$$Score_{ij} = \frac{Q_i K_j^T}{\sqrt{d_k}} \tag{3}$$

where the scaling factor d_k is the dimension of the key vector used to stabilize the gradient of the softmax function. Then, the scores calculated above are converted to probability distribution by softmax function to get the attention weight matrix. The description is as follows:

$$a_i = \text{Softmax}\left(\frac{Q_i K_j^T}{\sqrt{d_k}}\right) \tag{4}$$

Finally, based on the computed attention weight matrix, the V of all positions are weighted and summed to obtain the contextual representation of the current position. The description is as follows:

$$C = \sum_{j=1}^{n} a_{ij} V_j \tag{5}$$

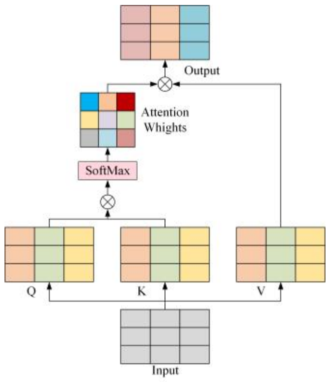


Figure 2. Self-attention mechanism

3. Proposed method

This section will be divided into two parts, the first part describes the main components of the method and the second part describes the overall method architecture and flow.

3.1. Module Composition

3.1.1. Multi-scale attention mechanisms

The traditional attention mechanism serves to strengthen the learning ability of CNNs on input data. Such as SE-Net focuses on different weights for the feature channels. The convolutional block attention module (CBAM) can focus on both channel and spatial attention. However, all of these attention mechanisms focus only on single-scale features and ignore feature representation in multi-scale spaces[29]. In order to compensate for the lack of attention to multi-scale space of existing attention mechanisms, we proposed a multi-scale squeeze-excitation attention module (MSSE).

MSSE consists of 3 branches: the SE branch, the convolution branch and the residual branch.

The SE branch has the same structure as SE-Net. The feature map is first compressed by global average pooling (GAP), thus preserving the global information of the features. Second, the compressed vectors are fully concatenated twice, the first layer is used to reduce the dimensionality and extract important features, and the second layer is activated using the ReLU activation function followed by normalization using the Sigmoid function to generate a dynamic weight vector. Finally, this weight is multiplied point-by-point with the original feature map to achieve re-weighting of the original features. Convolutional branch constructs another scale of spatial modeling by extracting contextual features of the input data through a 3×1 convolution. Since then, SE branch and convolutional branch have different scales of spatial representation.

Different from the information fusion methods such as summing and splicing in traditional attention mechanisms, we use a new information fusion method that allows MSSE to aggregate information in different scale spaces. The output of the SE branch is first subjected to a batch normalization operation (batchnorm), then it is normalized using the softmax function, and finally the normalized channel representations are subjected to a matmul operation with the output that has been convolved with 3×1. Unlike the point-by-point multiplication in SE-Net, matmul is equivalent to weighted summation of all channels at each pixel point to obtain a total channel feature. This operation is equivalent to tweaking the 3×1 convolution result using the output of the SE. Similarly, the same operation is performed in the convolution branch, using the weights generated by the 3×1 convolution to adjust the output of the SE. Finally, the original features are weighted by summing the spatial attention of the two scales and generating weights using Sigmoid again.

The structure of MSSE is shown in Figure 3.

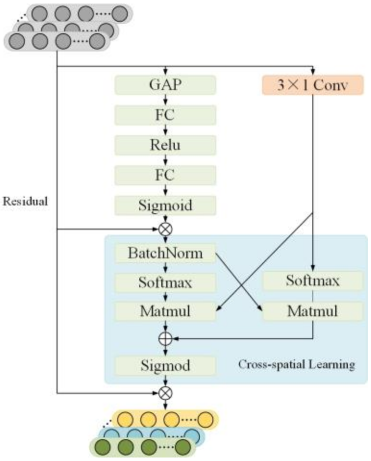


Figure 3. Structure of MSSE

3.1.2. Multi-scale feature extraction module

Since different dilation rates can bring different receptive fields to the convolution, different receptive fields can bring different feature scales. Therefore, we propose an dilated convolution based multi-scale feature extraction module (MSFE). The structure of MSFE is shown in Figure 4.

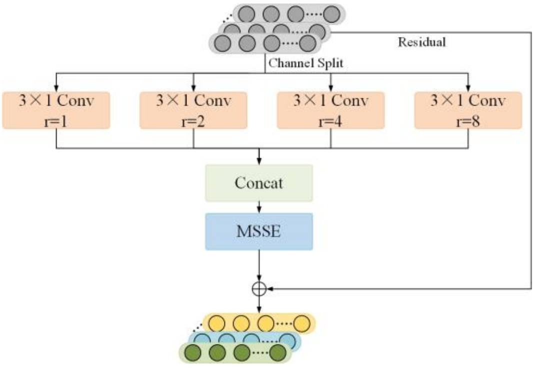


Figure 4.Structure of MFSE

MSFE consists of two main branches: the multiscale extraction branch and the residual branch. In this case, the multi-scale extraction branch divides the channel of input features into four branches, and features in each branch are extracted using dilated convolutions with different dilation rates. Through experiments, we set up four dilated convolution branches, and the dilation rate of each branch is set to 1, 2, 4, and 8, respectively. Input features are passed through the MSSE module after four branches, which are stitched together and passed through the MSSE module using concat. Finally, the output of the multiscale branch is summed with the residual branch. The MSFE module extracts a richer representation of the features using a smaller number of parameters.

3.1.3. Bi-directional Transformer for Time Signals

Rolling bearings, due to their complex operating environment, make bearing signals often present as complex time series. These complex time series are not only reflected in previous bearing operating cycles, but can also indicate impending fault. Therefore, capturing bi-directional temporal dependencies from the context in time-series signals of rolling bearings is important for accurate fault diagnosis.

Transformer was originally designed to handle tasks within the NLP field by using a multi-heads self-attention mechanism that essentially models remote dependencies in sequences. However, Transformer tends to focus only on the context prior to the current position while encoding, ignoring subsequent positions. To address this problem, inspired by Bi-LSTM and bi-directional gated recurrent unit (Bi-GRU), we extend the Transformer and introduce a Bi-Transformer. The Bi-Transformer can utilize time series from previous and subsequent rolling bearings, which can be used to augment the temporal relationships in the extracted features of the model. Since the input to the Bi-Transformer is feature information extracted from previous convolutional layers that encode high-level spatial information, they will serve as valuable contextual information for subsequent temporal relation learning. Bi-Transformer introduces an reverse mechanism that generates a sequence of reversed spatial information containing information about the input sequence as additional input. This allows Bi-Transformer to consider both past and future contexts when predicting a specific location. By combining outputs from both directions, our approach enables richer contextual understanding, effectively capturing complex relationships in sequences. The structure of Bi-Transformer is shown in Figure 5.

Applying the SA to the transformation of input feature maps. The input feature map is projected into the query Q, key K and value V for each header by linear mapping. The description is as follows:

$$\begin{cases} m_Q^h = W_Q^h \cdot m \\ m_K^h = W_K^h \cdot m \\ m_V^h = W_V^h \cdot m \end{cases} \tag{6}$$

where m is the input feature map, m_Q^h, m_K^h, and m_V^h are the Q, K, and V matrices on the h_{th} head. W_Q^h, W_K^h, and W_V^h are the projection matrices on the h_{th} head. The self-attention mechanism on each head is then utilized to obtain the attention weights. The description is as follows:

$$A^h = Attention\left(m_Q^h, m_K^h, m_V^h\right) = \\ \text{softmax}\left(\frac{m_Q^h \left(m_K^h\right)^T}{\sqrt{d_k}}\right) m_V^h \tag{7}$$

where d_k is the size of the m_K^h of the h_{th} head. The multi-head attention(MHA) mechanism obtains the final output by concatenating the results of all heads and applying another linear projection. The description is as follows:

$$MHA(m,m,m) = \text{Concat}\left(A^1, A^2 \dots A^H\right) \cdot W_O \tag{8}$$

where $MHA(m,m,m)$ denotes the MHA's output and W_O is the output projection matrix of the final multi-head attention output.

The key operation of the Bi-Transformer is the introduction of bi-directional attention. To accomplish this, we reverse the input feature map to account for past and future information. The description is as follows:

$$m_r[i,j] = x[i,d-j-1] \tag{9}$$

In general, the transformer encoder consists of L identical layers. Each of these layers has two sub-layers: the MHA and the fully connected feed-forward network. Thus, the above operation needs to be performed L times. Each layer in such a stack processes the input data m in turn, i.e., the operations from Eq. (6) to Eq. (10), thus capturing a more detailed representation of the context. In this way the dependencies in the input data are learned hierarchically. The output of each layer in the Bi-Transformer serves as an input to the subsequent layers, thus gradually extracting complex patterns. The final output of the Bi-Transformer is obtained by stepwise extraction of the input feature map. The description is as follows:

$$Bi-Transformer(m) = \\ Layer_L\left(Layer_{L-1}\left(\dots\left(Layer_1(m)\right)\dots\right)\right) \tag{10}$$

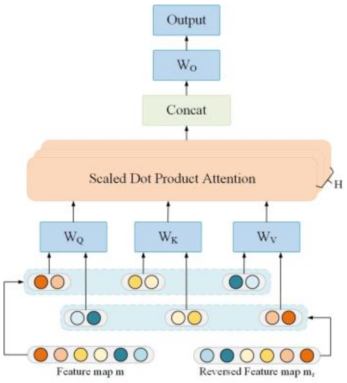


Figure 5.Structure of Bi-Transformer

3.2. Multi-scale bi-directional self-attentive diagnosis method framework

The framework of the multiscale bi-directional self-attentive diagnosis method (MBSADM) proposed in this paper is shown in Figure 6. The diagnostic steps are as follows:

Step1: Collect rolling bearing timing signals under different working conditions;

Step2: Crop the timing signal using a horizontal sliding window to construct the dataset;

Step3: Divide the dataset into training set, validation set and test set;

Step4: Train MBSADM using training set and test set to learn sample features;

Step5: Validate the model diagnosis results using the validation set.

In particular, the backbone of MBSADM consists of the MSFE module and the Bi-Transformer. First, DSW is used and one spatial downsample is performed using maxpool. Then, the MSFE module is stacked twice for multi-scale feature extraction. Subsequently, the data is reversed to capture the contextual relationships in the data using Bi-

Transformer. After using one convolution, the results are finally output through a fully connected layer.

4. Experimentation and analysis

In order to validate the fault diagnosis performance of MBSADM, we conduct experiments using three rolling bearing datasets so as to verify the noise immunity, hybrid fault diagnosis capability, and generalization of MBSADM. The device configuration used for all experiments was AMD Ryzen7 5800H CPU@3.2GHz, NVIDIA RTX3060 (12G) and 16G RAM, and the framework used for the experiments was pytorch1.12.

4.1. Comparison methods

In order to validate the performance of the proposed model, we chose advanced deep learning methods as comparison methods, including MobileNetV2, Transformer1D, CLFormer[30], Liconvformer[31] and Convformer_NSE[32]. Among them, MobileNetV2 is a mature deep learning model. transformer1D combines CNN and Transformer and is used as a model for fault diagnosis. CLFormer is a lightweight Transformer based on convolutional embedding and linear self-attention (LSA). Liconvformer is a fault diagnosis model based on separable multiscale modules with broadcast self-attention modules. Convformer_NSE uses sparse-corrected multi-self attention and constructs a novel senet (NSE) for channel adaptive learning.

To ensure the fairness of the experiments, the training epochs were all set to 50, and the initial learning rates were all set to 0.0003. To verify the stability of the method, each experiment was repeated 20 times.

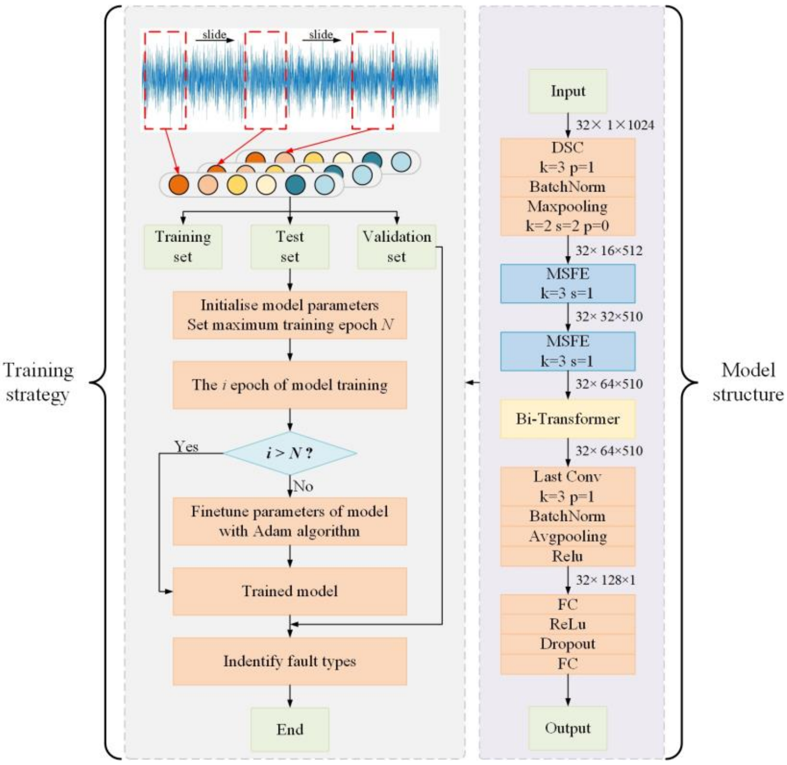


Figure 6. Fault diagnosis model framework

4.2. Case 1

4.2.1. description of the dataset

Experimental data were obtained from Case Western Reserve University (CWRU)[33]. As shown in Figure 7, bearing type SKF 6205 is used. Artificial failures were categorized into three types: ball faults (BF), inner-ring faults (IF), and outer-ring faults (OF), each of which in turn contained points of failure with diameters of 0.007 inches, 0.014 inches, and 0.021 inches. A total of 10 fault labels (including 1 health label and 9 fault labels) were obtained by combining the fault type and size permutations. Additionally, we use bearing data collected under three different loads (1 hp, 2 hp, and 3 hp, corresponding to forces of 416.7 N, 833.4 N, and 1250.1 N, respectively) to evaluate the model's generalization capability under complex operating conditions. The sampling frequency of the CWRU dataset is 12,000 Hz, and the bearing rotates at 1,797 r/min, generating approximately 400 samples per revolution. To ensure the reliability of the samples, we apply a moving window of length 1,024 to extract samples, ensuring that each sample contains sufficient information. Furthermore, we split the dataset into training, validation, and test sets in a 7:2:1 ratio. The detailed data distribution is presented in Tables 1 and 2.

Figure7. Rolling bearing test bench

Table 1.Fault label details

Class label	Fault location	Fault size(in)	Load(hp)	Dataset
00	Normal	/	1,2,3	A, B, C,
01	BF	0.007	1,2,3	A, B, C,
02	IF	0.007	1,2,3	A, B, C,
03	OF	0.007	1,2,3	A, B, C,
04	BF	0.014	1,2,3	A, B, C,
05	IF	0.014	1,2,3	A, B, C,
06	OF	0.014	1,2,3	A, B, C,
07	BF	0.021	1,2,3	A, B, C,
08	IF	0.021	1,2,3	A, B, C,
09	OF	0.021	1,2,3	A, B, C,

Table 2. Details of dataset division

Sample	Dataset A	Dataset B	Dataset C	Dataset D
Training	1631	1631	1631	1631
Validation	466	466	466	466
Test	233	233	233	233

4.2.2. Evaluation of MBSADM performance with original signals

In order to make a preliminary assessment of the feature extraction capability of MBSADM, we conducted diagnostic experiments on dataset A, B and C using MBSADM with the comparison method, and the results of the experiments are shown in Table 3. In the three datasets, MBSADM achieved an average accuracy of over 99%, with an average accuracy of 99.81%. This means that MBSADM categorizes almost every sample correctly. In dataset A, the advantages of MBSADM are most obvious. In dataset C, Liconvformer had the best diagnostic performance at 99.49%, but still differed from MBSADM by 0.48%. The above experimental results show that MBSADM has excellent feature extraction capability and can extract better features in the original signal.

Table 3.Accuracy of the six methods under the original signal(%)

methods	Dataset A	Dataset B	Dataset C	Average	Time
MobileNetV2	96.52	97.31	98.79	97.54	15.47s
Transformer1D	95.98	97.82	98.51	97.44	14.91s
CLFormer	97.36	98.97	99.21	98.51	12.74s
Liconvformer	98.45	99.31	99.49	99.08	**9.53s**
Convformer_NSE	97.84	98.98	99.05	98.62	13.32s
MBSADM	**99.56**	**99.91**	**99.97**	**99.81**	27.51s

4.2.3. Performance evaluation in noisy environments

In the actual working environment of rolling bearings, the interference of strong noise is often accompanied, which challenges the performance of the fault diagnosis method. In this experiment, we utilize the original signals for experiments and add Gaussian white noise with different signal-to-noise ratios to the data in the test set to simulate the ambient noise, so as to verify the noise immunity of the model. In this experiment, we utilize the original signals for training and add Gaussian white noise with different signal-to-noise ratios(SNR)[34] to the data in the test set to simulate the ambient noise, so as to verify the method's noise immunity. SNR is defined as follows:

$$\mathrm{SNR}_{\mathrm{dB}} = 10\log(\frac{P_{signal}}{P_{noise}}) \tag{11}$$

where P_{signal} and P_{noise} denote the power of the original and noise signals, respectively. From the above equation, it can be seen that when SNR<0, the noise power is greater than the original signal power, and when SNR>0, the noise signal power is less than the original signal power. In this experiment, we use the data in dataset A to do the training and add -6db, -3db, -2db, 3db and 6db Gaussian white noise signals of 5 SNRs to the test set to test the noise immunity performance of the proposed method. The experimental results are shown in Table 4.

Table 4.Accuracy of 6 methods in noisy environment(%)

methods	-6dB	-3dB	-2dB	3dB	6dB
MobileNetV2	70.80	75.34	77.67	94.32	96.88
Transformer1D	62.57	70.29	72.13	81.92	86.16
CLFormer	79.61	81.41	83.54	87.68	89.12
Liconvformer	77.32	82.91	85.92	89.46	93.52
Convformer_NSE	82.57	86.78	89.82	93.03	97.76
MBSADM	**85.48**	**89.69**	**93.74**	**97.64**	**99.17**

As can be seen from Table 4, the accuracy of MBSADM in the five noise environments is significantly higher than the other compared methods. In particular, MBSADM can still obtain more than 85% accuracy at SNR=-6dB, which is 3.89% higher than that of Convformer_NSE, which indicates that MBSADM still possesses obvious advantages in strong noise environments, and such noise-resistant performance makes MBSADM have a strong application value in practical application environments. It is worth noting that the Liconvformer, which performed sub-optimally in the original signal performance test, did not exhibit excellent noise immunity when encountering strong noise (SNR = -6dB, SNR = -3dB). This is because Liconvformer was designed with lightweight in mind and neglected to dig deeper into the abstract features of the input data. In contrast, MBSADM can understand the input signal more comprehensively as it is designed for multi-scale convolution while also utilizing the multi-scale attention mechanism to extract features from the input signal, and finally using Bi-Transformer to capture the bi-directional time dependence and extract features layer by layer. In addition, experiments with Transformer1D have shown that using only a simple combination of CNNs + Transformer does not provide noise immunity for the method. MobileNetV2 cannot achieve satisfactory diagnostic results when SNR<0, but it can achieve better diagnostic results when SNR>0. This indicates that MobileNetV2 cannot resist the interference of strong noise, but it can achieve better results under weak noise. In summary, it can be seen that MBSADM with both multi-scale convolution, multi-scale attention mechanism and Bi-Transformer can extract richer and more comprehensive features from complex vibration signals, which makes MBSADM have good noise immunity.

4.2.4. Performance Evaluation in Different Loading Conditions

Rolling bearings and rotating machinery often operate under varying load fluctuations. The rolling bearing data under different loads do not have the same feature distribution in the feature space. This requires that the fault diagnosis methods needs to overcome the difficulty of inconsistency not only in the feature space and class space, but also in the feature distribution. In this experiment, we choose three loaded data as training set respectively, while the other two datasets are used as test sets, so as to verify the load domain adaptation of the method. That is, when we choose dataset A as th9e training set, B,C are done as the test set respectively. The experimental results are shown in Figure 8. It can be observed that in the different loading experiments MBSADM also achieved the smallest average standard deviation while obtaining the highest average accuracy. MobileNetV2 had the most pronounced fluctuations, with an accuracy of 99.55% in the AB group of experiments and only 76.78% in the CA group of experiments. MobileNetV2 had the most pronounced fluctuation in performance, with 99.55% accuracy in the AB group of experiments and only 76.78% accuracy in the CA group of experiments. Meanwhile CLFormer achieved the worst results in the CA group of experiments, while MBSADM achieved an average accuracy of 96.26% in this group of experiments, which suggests that MBSADM can overcome the difficulty of inconsistent feature distributions to some extent. Furthermore, Liconformer achieved the second highest accuracy in this experiment, which demonstrates that the design of multi-scale modules allows the method to overcome feature inconsistencies. From this experiment, we can get the following two conclusions: (a) when the load gap between the two datasets is larger, the feature similarity of the data is smaller, and the diagnosis is more difficult; (b) when the load gap between the two datasets is smaller, the feature similarity is smaller, and better diagnosis can be achieved.

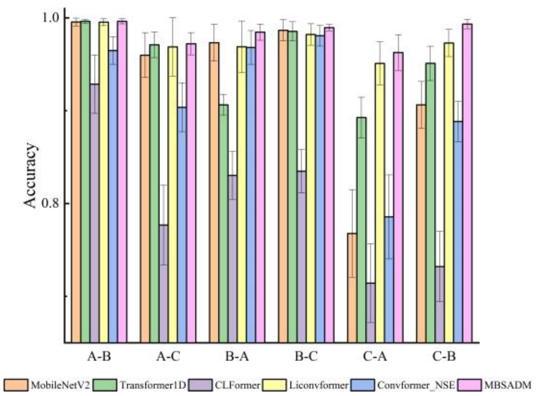


Figure 8. Accuracy of 6 methods in different loading environment

4.2.5. Performance Evaluation Under Limited Training Samples

Since rolling bearings operate in a healthy state for most of their lifespan, fault data is difficult to obtain. The limited number of fault samples fails to comprehensively represent the fault characteristics, and directly using such data for training hinders the generalization capability of fault diagnosis models to the validation set, leading to suboptimal diagnostic performance. Therefore, a key challenge in rolling bearing fault diagnosis is how to extract effective features from limited training data and achieve accurate classification. To evaluate the performance of MBSADM under limited training samples, we randomly reduced the CWRU dataset to one-fifth of its original size before partitioning it into training, validation, and test sets for experiments. The experimental results under limited samples are presented in Table 5.

Table5. Accuracy of the six methods under the Limited Training Samples

methods	Dataset A	Dataset B	Dataset C	Average
MobileNetV2	92.16	91.98	93.1	92.41
Transformer1D	86.21	84.87	85.64	85.57
CLFormer	91.55	92.03	91.44	91.67
Liconvformer	91.97	92.31	92.08	92.12
Convformer_NSE	92.25	93.59	93.15	92.99
MBSADM	**98.34**	**97.92**	**98.46**	**98.24**

A comparison between Table 3 and Table 5 reveals that all six methods experience a certain degree of performance degradation when faced with limited training data. This is because the insufficient training samples hinder the fault diagnosis models from effectively generalizing the learned features to the validation set. Among all methods, MBSADM exhibits the least performance decline, achieving an average accuracy of 98.24%, with only a 1.57% decrease. In contrast, the accuracy of MobileNetV2, Transformer1D, CLFormer, Liconvformer, and Convformer_NSE decreases by 5.13%, 11.87%, 6.84%, 6.96%, and 5.63%, respectively. These results demonstrate that MBSADM can still capture deep features from input data and generalize effectively to the validation set even when the training set is limited in size.

4.3. Case 3

4.3.1. Data set description

The homemade dataset was experimented and collected on an MFS test bed manufactured by Spectrum Quest Incorporated (SQI). The experimental equipment is shown in Figure 9. The test data uses data from the bearing drive end. Four fault types were simulated under normal conditions by laser etching: ball fault (BF), inner-ring fault (IF) and outer-ring fault (OF), and Composite fault (CF). The signals were collected in groups for four rotational speeds of 1130r/min, 1251r/min, 1378r/min and 1449r/min with a sampling frequency of 15.6Khz. The experimental equipment was loaded by applying a radial load of 50N through a 5.1kg rotor disk mounted between two bearings. The vibration signals were collected by connecting the signal collector and acceleration sensor using a 1-channel data cable and transferring the signals to a computer via the USB interface. We set the data into four sets A,B,C,D according to the four rotational speeds. The training set, validation set and test set are made as in Case 1. The detailed data are shown in Table 6 and Table7.

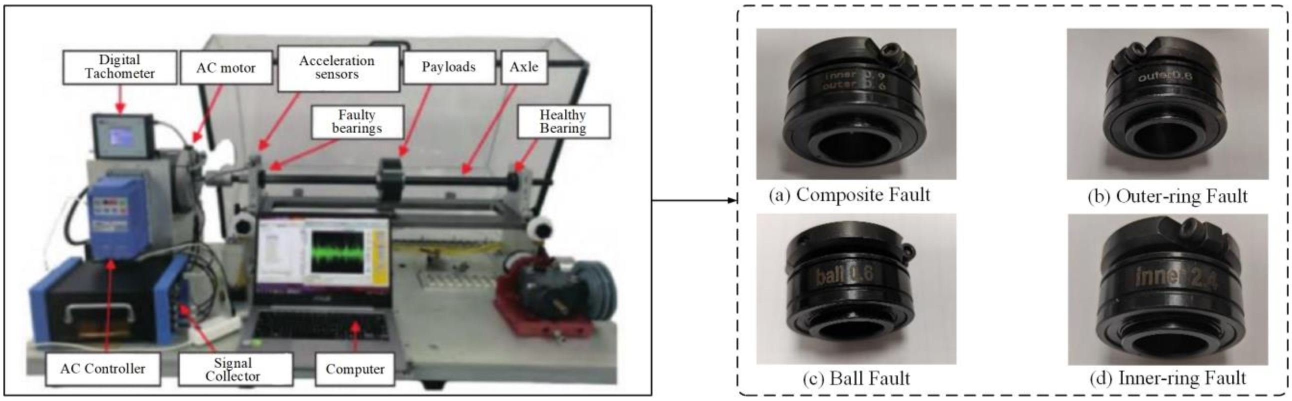


Figure 9. homemade dataset experimental equipment and fault types

Table6. Fault label details for homemade dataset

Class label	Fault location	Speed(r/min)	Dataset
00	BF	1130,1251,1378,1449	A, B, C, D
01	CF	1130,1251,1378,1449	A, B, C, D
02	IF	1130,1251,1378,1449	A, B, C, D
03	OF	1130,1251,1378,1449	A, B, C, D

Table7. Details of the division of the homemade dataset

Sample	Dataset A	Dataset B	Dataset C	Dataset D
Training	714	714	714	714
Validation	204	204	204	204
Test	102	102	102	102

4.3.2. Performance evaluation under original signal

To preliminarily evaluate the performance of MBSADM under this dataset, we use the confusion matrix to obtain the classification results at four rotational speeds, as shown in Figure 10(a)-(d). The vertical coordinate of the graph represents the real fault labels and the horizontal coordinate represents the predicted fault labels. The sample size of the test set for each fault is 102. The accuracy of each fault type can be observed from the main diagonal. From Figure 10, it can be seen that MBSADM achieves high accuracy for fault identification at four different rotational speeds. It can be seen that in the experiment with a rotational speed of 1378 r/min, the fault classification accuracy is 100% except for the 4 good label misclassification. In the other RPM experiments, all fault labels were not misclassified and the diagnostic accuracy was 100%. The above results demonstrate the good fault classification performance of MBSADM in homegrown datasets.

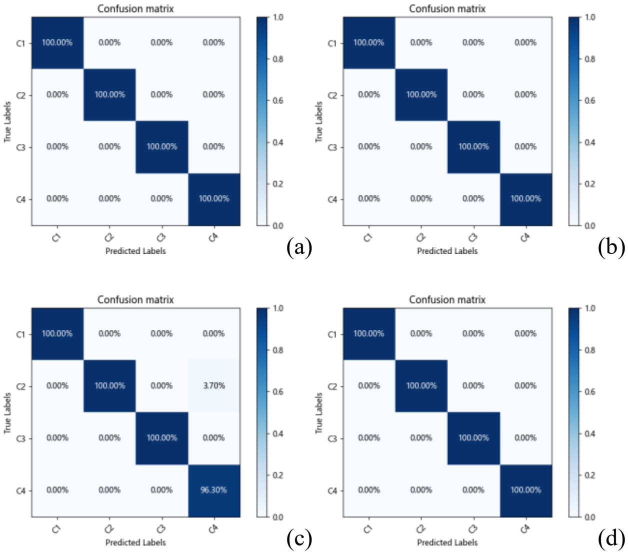


Figure10. Confusion matrix results for fault classification at (a) 1130r/min, (b) 1251r/min, (c) 1378r/min, (d) 1449r/min.

4.3.3. Evaluation of performance in a noisy environment

Same as Case 1, we use the training set of dataset A for training, and at the same time add the Gaussian white noise signals with five SNRs of -6db, -3db, -2db, 3db and 6db to the test set to test the noise immunity of the proposed method, respectively. The experimental results are shown in Figure 11.

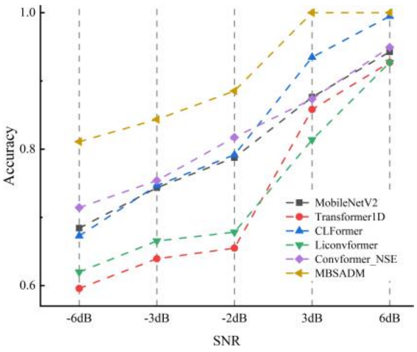


Figure 11. Accuracy of comparison methods in noisy

environment

As can be seen in Figure 11, even though the homemade dataset has only four fault labels, experiments in noisy environments still pose a challenge to the comparison methodology. When SNR < 0, the accuracy of the comparison methods are all less than 80%, which indicates that the comparison methods are less noise-resistant on the homemade dataset. Also the accuracy of MBSADM was significantly higher than the comparison methods. In particular, the accuracy achieved by MBSADM is 100% when SNR>0 and the diagnostic accuracy is also above 80% when facing strong noise with SNR<0. Combined with Case 1, we demonstrate that our design idea can effectively improve the noise immunity of the network and has strong robustness and generalization.

4.3.4. Performance evaluation under composite operating conditions

Rolling bearings operate not only in noisy environments, but also at different speeds. In order to realize the fault diagnosis of rolling bearings under compliant operating conditions, we select the dataset with different rotational speeds as the training set under two different strong noise environments (SNR=-6,-3), and use the remaining three datasets as the test set, that is, we use the data A as the training set, and the other three datasets as the test set, so as to validate the fault diagnosis performance of the MBSADM for the composite operating conditions of different rotational speeds under different noises. fault diagnosis performance under different noises. The experimental results are shown in Figure 12. From Figure 12(a), it can be seen that Transformer1D, CLFormer, and MobileNetV2 have poor domain adaptation when SNR = -3, with average accuracies of 75.09%, 80.73%, and 91.04% for the 12 cases. In contrast, the average accuracy of MBSADM was 97.48%, which is a significant advantage over the comparison methods. It is worth noting that most of the methods achieve good accuracy when the difference in rotational speed between the training set and the test set is small (e.g., when a dataset A with a rotational speed of 1130r/min is used as the training set and a dataset with a rotational speed of 1251r/min is used as the test set). However, the accuracy of all these methods decreases when the RPMs of the training and test sets differ significantly. This suggests that the gap between the data features in the two RPM domains that differ by a large amount is also large, and this gap poses a challenge to the adaptive nature of the fault diagnosis methodology. And MBSADM shows a very stable performance in the experiments and achieves an average accuracy of 88.54% even in the cross-domain experiments of A-D. This verifies that MBSADM possesses strong adaptivity. In Figure 12(b), it can be seen that despite the further enhancement of noise that causes a certain degree of decrease in the accuracy of the various methods, MBSADM still achieves the highest average fault accuracy of 97.48%, which verifies that MBSADM possesses better stability than the other methods. In summary, MBSADM can achieve the highest accuracy in cross-domain fault diagnosis in both strong noise scenarios, indicating that MBSADM has obvious advantages in fault diagnosis performance.

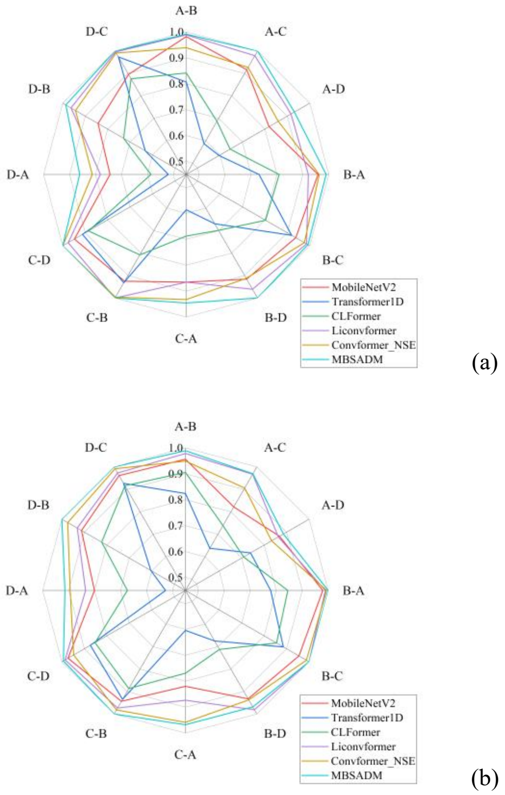


Figure 12. Accuracy of different speed experiment under (a) SNR=-3, (b) SNR=-6

4.4. Ablation experiments

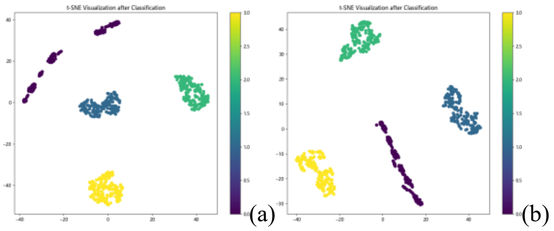
In order to validate the network module structure proposed in this paper in MBSADM, we perform ablation experiments on MBSADM using the homemade dataset in Case 2. The details of the contrasting network structures are shown in Table 8. where SADM uses ordinary convolution instead of MSFE module and traditional self-attention mechanism instead of Bi-Transformer.BSADM does not contain MSFE module but contains Bi-Transformer module.MSADM contains MSFE module but uses traditional self-attention mechanism instead of Bi-Transformer.

Table 8. ablation experimental model branching details

model	MSFE	Bi-Transformer
SADM	No	No
BSADM	No	Yes
MSADM	Yes	No
MBSADM	Yes	Yes

4.4.1. original signal set performance evaluation

We used a dataset with a rotational speed of 1251 r/min for the experimental data, and experiments were conducted using four variants of the method, and all four methods were able to achieve an average accuracy of 100%. In this case, in order to visualize and more intuitively show the features learned by the network, we use the t-SNE visualization technique to show the distribution of features as the data passes through the last layer of the network. This technique is commonly used to validate the effectiveness of fault diagnosis methods. The t-SNE visualization results are shown in figure 13. In figure 13, the coordinates of each point represent the location of the point in the 2D space, and different labels represent different fault types. It can be seen that although SADM, BSADM and MSADM can all separate fault points, the three methods are unable to cluster some fault points well together, and there is a clear intra-class separation. And through figure 13(d) it can be seen that MBSADM separates the four fault types completely and clusters them best. This illustrates the enhancement of modeling capability by MSFE and Bi-Transformer.



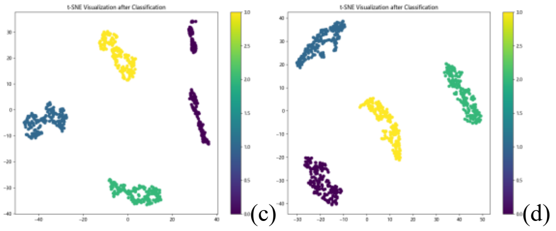


Figure 13. t-SNE visualization of the 4 methods on raw data

4.4.2 Evaluation of performance in a noisy environment

As in Case 2, we use the training set of dataset B for training, while adding -6db, -3db, -2db, 3db, and 6db of Gaussian white noise signals with five SNRs to the test set to test the method's noise immunity performance, respectively. The experimental results are shown in Table 9. Combined with Case 2, it can be seen that the accuracy of SADM is very similar to that of Transformer1D, due to the fact that the network structure of SADM is very similar to that of Transformer1D after using ordinary convolution in place of MSFE as well as using the conventional Transformer module in place of Bi-Transformer, and therefore similar diagnostic results. BSADM and MSADM, on the other hand, are not as good in terms of noise immunity due to the lack of MSFE and Bi-Transformer modules, respectively. This is because the lack of MSFE prevents BSADM from analyzing the input data at multiple scales, while the lack of Bi-Transformer in MSADM prevents a better understanding of the context.

Table 9.4 Accuracy of the methods in a noisy environment(%)

Methods	-6dB	-3dB	-2dB	3dB	6dB
SADM	57.42	62.79	64.95	83.24	91.96
BSADM	70.59	73.76	75.89	88.91	93.45
MSADM	72.43	74.39	76.34	87.36	94.92
MBSADM	**82.09**	**85.35**	**87.61**	**99.46**	**100**

We similarly induced the t-SNE technique to visualize the classification results, with the experimental context of a test set noise of -2 dB, as shown in Fig. 14. It can be seen that SADM, BSADM and MSADM all show significant class overlap, with SADM having the worst clustering effect. In figure 14(b)-(c) it can be seen that BSADM and MSADM only cluster the faults better for two categories, and the other two categories have a very severe overlap of fault points. This shows that SADM, BSADM and MSADM are not able to fulfill the fault diagnosis task well in the strong noise environment. At the same time, it can be seen that MBSADM has the best clustering effect, although there will be individual point overlapping phenomenon, but the number is not large. It shows that with the help of MSFE and Bi-Transformer module, MBSADM has better noise immunity and robustness.

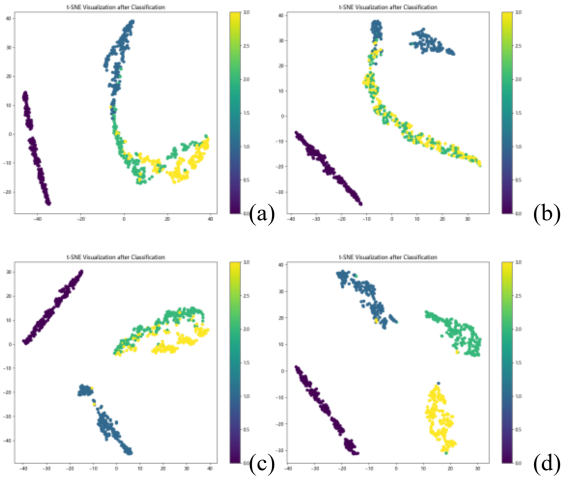


Fig. 14. t-SNE visualization of the 4 methods in a noisy environment

5. Conclusion

This paper proposes a novel fault diagnosis method, MBSADM, which demonstrates exceptional robustness under noisy environments and varying rotational speeds. The method directly takes raw one-dimensional data as input and achieves efficient feature learning through Multi-Scale Feature Extraction (MSFE) and a Bidirectional Transformer (Bi-Transformer). Specifically, MSFE employs Multi-Scale Subspace Encoding (MSSE) to capture features at different scales and integrates an attention mechanism to enhance the extraction of critical information. Meanwhile, Bi-Transformer incorporates a reversal mechanism to strengthen the modeling of temporal dependencies. Experimental results on the CWRU bearing fault dataset and a self-developed complex working condition dataset show that MBSADM maintains high accuracy even under strong noise and varying loads. Moreover, it exhibits robust fault recognition capabilities in extreme noise conditions. Ablation studies further validate the key roles of MSFE and Bi-Transformer in feature extraction and temporal modeling. Overall, MBSADM demonstrates superior fault diagnosis performance, noise resistance, and generalization ability across different working conditions, making it a reliable solution for intelligent maintenance and equipment health monitoring systems.

Data availability

All data that support the findings of this study are included within the article (and any supplementary files).

Acknowledgment

This research work has been awarded by the National Natural Science Foundation of China (No.62263021), the College Industrial Support Project of Gansu Province (2023CYZC-24)

Conflict of interest

The authors declare that there is no conflict of interest.

Reference

[1] Cui W, Jiao S, Gao R, et al. Continuous unsaturated second-order hybrid multi-stable stochastic energy resonance and its application in rolling bearing fault diagnosis[J]. Applied Acoustics, 2025, 228: 110298.

[2] Pandiyan M, Babu T N. Systematic review on fault diagnosis on rolling-element bearing[J]. Journal of Vibration Engineering & Technologies, 2024, 12(7): 8249-8283.

[3] Wang L, Zou T, Cai K, et al. Rolling bearing fault diagnosis method based on improved residual shrinkage network[J]. Journal of the Brazilian Society of Mechanical Sciences and Engineering, 2024, 46(3): 172.

[4] Davari S, Ommi F, Saboohi Z. Investigating the Effects of adding butene, homopolymer to gasoline on engine performance parameters and pollutant emissions: empirical study and process optimization[J]. Journal of The Institution of Engineers (India): Series C, 2022, 103(3): 421-434.

[5] Senthilnathan N, Babu T N, Varma K S D, et al. Recent advancements in fault diagnosis of spherical roller bearing: A short review[J]. Journal of Vibration Engineering & Technologies, 2024, 12(4): 6963-6977.

[6] Shojaeefard M H, Hosseini S E, Zare J. CFD simulation and Pareto-based multi-objective shape optimization of the centrifugal pump inducer applying GMDH neural network, modified NSGA-II, and TOPSIS[J]. Structural and Multidisciplinary Optimization, 2019, 60(4): 1509-1525.

[7] Wang B, Guo Y, Zhang Z, et al. Developing and

applying OEGOA-VMD algorithm for feature extraction for early fault detection in cryogenic rolling bearing[J]. Measurement, 2023, 216: 112908.

[8] Han T, Zhang L, Yin Z, et al. Rolling bearing fault diagnosis with combined convolutional neural networks and support vector machine[J]. Measurement, 2021, 177: 109022.

[9] Meng Z, Luo C, Li J, et al. Research on fault diagnosis of rolling bearing based on lightweight model with multiscale features[J]. IEEE Sensors Journal, 2023, 23(12): 13236-13247.

[10] Guo Y, Mao J, Zhao M. Rolling bearing fault diagnosis method based on attention CNN and BiLSTM network[J]. Neural processing letters, 2023, 55(3): 3377-3410.

[11] Dong Z, Zhao D, Cui L. An intelligent bearing fault diagnosis framework: one-dimensional improved self-attention-enhanced CNN and empirical wavelet transform[J]. Nonlinear Dynamics, 2024, 112(8): 6439-6459.

[12] Deng L, Zhang Y, Shi Z. An Improved Fault Diagnosis Method of Rolling Bearings Based on Multi-Scale Attention CNN[J]. Journal of Failure Analysis and Prevention, 2024, 24(4): 1814-1827.

[13] Jia L, Chow T W S, Yuan Y. GTFE-Net: A gramian time frequency enhancement CNN for bearing fault diagnosis[J]. Engineering Applications of Artificial Intelligence, 2023, 119: 105794.

[14] Xu Z, Tang X, Wang Z. A multi-information fusion ViT model and its application to the fault diagnosis of bearing with small data samples[J]. Machines, 2023, 11(2): 277.

[15] Xu P, Zhang L. A fault diagnosis method for rolling bearing based on 1D-ViT model[J]. IEEE Access, 2023, 11: 39664-39674.

[16] Patwardhan N, Marrone S, Sansone C. Transformers in the real world: A survey on nlp applications[J]. Information, 2023, 14(4): 242.

[17] Chen Z, Zhang Y, Gu J, et al. Dual aggregation transformer for image super-resolution[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 12312-12321.

[18] Hou Y, Wang J, Chen Z, et al. Diagnosisformer: An efficient rolling bearing fault diagnosis method based on improved transformer[J]. Engineering Applications of Artificial Intelligence, 2023, 124: 106507.

[19] Tang X, Xu Z, Wang Z. A novel fault diagnosis method of rolling bearing based on integrated vision transformer model[J]. Sensors, 2022, 22(10): 3878.

[20] Liang P, Yu Z, Wang B, et al. Fault transfer diagnosis of rolling bearings across multiple working conditions via subdomain adaptation and improved vision transformer network[J]. Advanced Engineering Informatics, 2023, 57: 102075.

[21] Hou S, Lian A, Chu Y. Bearing fault diagnosis method using the joint feature extraction of Transformer and ResNet[J]. Measurement Science and Technology, 2023, 34(7): 075108.

[22] Gao Z, Wang Y, Li X, et al. Twins transformer: rolling bearing fault diagnosis based on cross-attention fusion of time and frequency domain features[J]. Measurement Science and Technology, 2024, 35(9): 096113.

[23] Liu G, Zhang C, Xu S, et al. A new lightweight fault diagnosis framework towards variable speed rolling bearings[J]. IEEE Access, 2024.

[24] Kattenborn T, Leitloff J, Schiefer F, et al. Review on Convolutional Neural Networks (CNN) in vegetation remote sensing[J]. ISPRS journal of photogrammetry and remote sensing, 2021, 173: 24-49.

[25] Hosseini S E, Jafaripanah S, Saboohi Z. CFD simulation and aerodynamic optimization of two-stage axial high-pressure turbine blades[J]. Journal of the Brazilian Society of Mechanical Sciences and Engineering, 2024, 46(11): 666.

[26] Chollet F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1251-1258.

[27] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions[J]. arXiv preprint arXiv:1511.07122, 2015.

[28] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[29] Dai Z, Liu H, Le Q V, et al. Coatnet: Marrying convolution and attention for all data sizes[J]. Advances in neural information processing systems, 2021, 34: 3965-

3977.

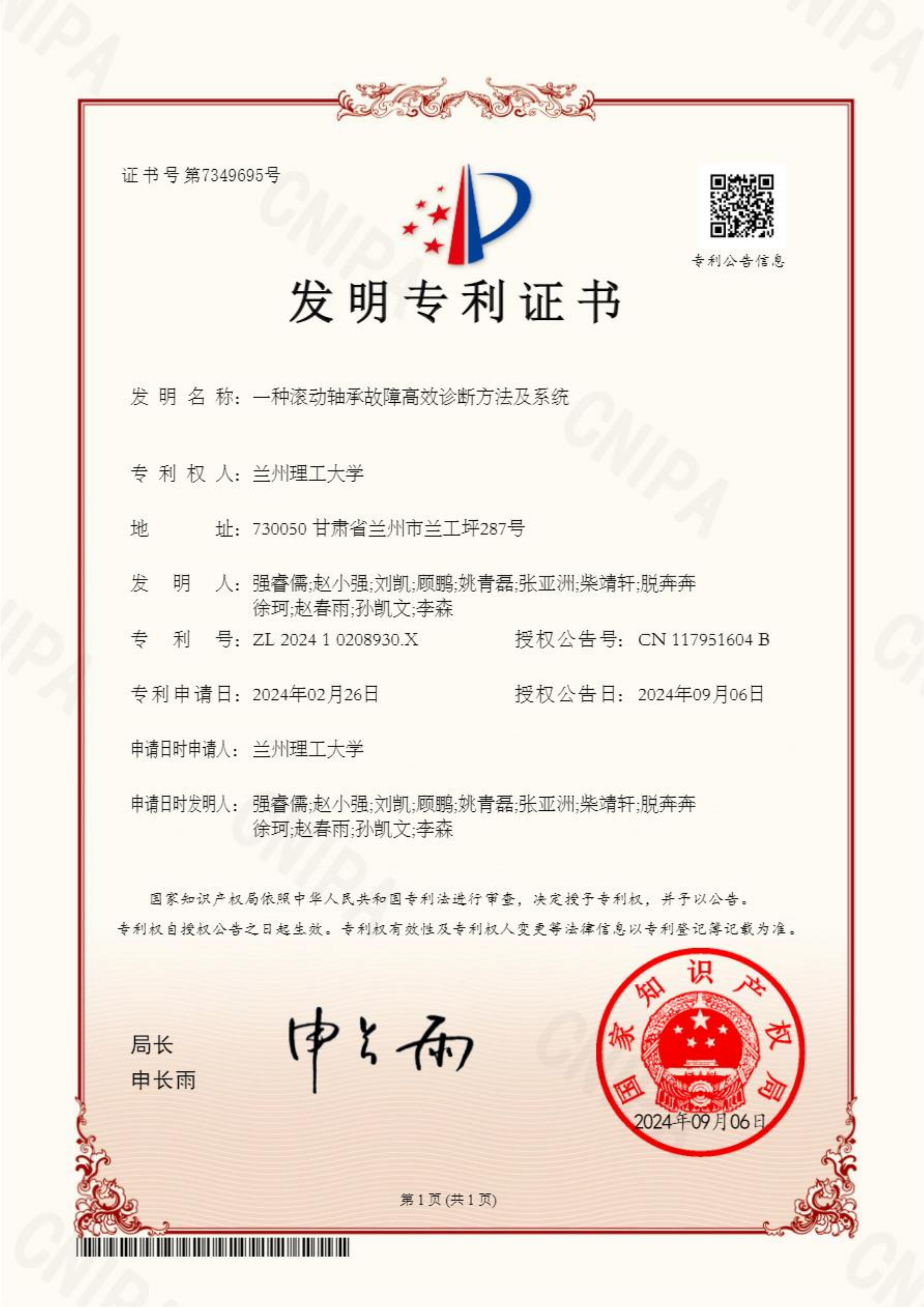
[30] Fang H, Deng J, Bai Y, et al. CLFormer: A lightweight transformer based on convolutional embedding and linear self-attention with strong robustness for bearing fault diagnosis under limited sample conditions[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 71: 1-8.

[31] Yan S, Shao H, Wang J, et al. LiConvFormer: A lightweight fault diagnosis framework using separable multiscale convolution and broadcast self-attention[J]. Expert Systems with Applications, 2024, 237: 121338.

[32] Han S, Shao H, Cheng J, et al. Convformer-NSE: A novel end-to-end gearbox fault diagnosis framework under heavy noise using joint global and local information[J]. IEEE/ASME Transactions on Mechatronics, 2022, 28(1): 340-349.

[33] Smith W A, Randall R B. Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study[J]. Mechanical systems and signal processing, 2015, 64: 100-131.

[34] Liang H, Cao J, Zhao X. Multi-scale dynamic adaptive residual network for fault diagnosis[J]. Measurement, 2022, 188: 110397.



证书号第7349695号

专利公告信息

发明专利证书

发 明 名 称：一种滚动轴承故障高效诊断方法及系统

专 利 权 人：兰州理工大学

地　　　址：730050 甘肃省兰州市兰工坪287号

发　明　人：强睿儒;赵小强;刘凯;顾鹏;姚青磊;张亚洲;柴靖轩;脱奔奔
徐珂;赵春雨;孙凯文;李森

专　利　号：ZL 2024 1 0208930.X　　　授权公告号：CN 117951604 B

专利申请日：2024年02月26日　　　授权公告日：2024年09月06日

申请日时申请人：兰州理工大学

申请日时发明人：强睿儒;赵小强;刘凯;顾鹏;姚青磊;张亚洲;柴靖轩;脱奔奔
徐珂;赵春雨;孙凯文;李森

国家知识产权局依照中华人民共和国专利法进行审查，决定授予专利权，并予以公告。

专利权自授权公告之日起生效。专利权有效性及专利权人变更等法律信息以专利登记簿记载为准。

局长
申长雨

申长雨

国家知识产权局
2024年09月06日