

# A Multiscale Feature Residual Network With Staged Optimization for Unsupervised Cross-Domain Fault Diagnosis

Changbo He<sup>1</sup>, Qi Meng<sup>1</sup>, Peng Chen<sup>1</sup>, *Member, IEEE*, Yaqiang Jin<sup>2</sup>, Wei Fan<sup>3</sup>, *Member, IEEE*, and Zhibo Yang<sup>4</sup>, *Member, IEEE*

**Abstract**—In the fault diagnosis of industrial equipment, the traditional supervised learning method is difficult to effectively deal with the unlabeled data in the target domain because of the significant differences of working modes. To address this issue, a multiscale feature residual network (MFSResnet) with staged optimization is proposed in this article. First, an enhanced dilated convolution module and two novel residual blocks are designed to improve feature representation ability. Building on these components, a MFSResnet is then proposed, integrating the enhanced dilated convolution module, the residual blocks, and an SE attention module for effective feature extraction. Additionally, a staged optimization method is employed throughout the training process to optimize multiple loss functions. Finally, experiments analysis conducted on two bearing datasets, CWRU and PU, demonstrate the efficacy of the proposed approach in addressing the challenge of unlabeled data in the target domain, and its superiority over existing methods.

**Index Terms**—Fault diagnosis, group normalization (GN), residual network (ResNet), rotating machinery, transfer learning.

## I. INTRODUCTION

ROTATING machinery [1] constitutes a fundamental component in numerous industrial operations. Thus, the failure of rotating machinery may cause production interruption, economic losses, and even serious safety accidents,

Received 27 March 2025; revised 20 May 2025; accepted 29 May 2025. Date of publication 18 June 2025; date of current version 26 June 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 52305085 and Grant 52105111, in part by China Postdoctoral Science Foundation under Grant 2023M740021, and in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2025A1515012256. The Associate Editor coordinating the review process was Dr. Jing Yuan. (*Corresponding authors: Wei Fan; Qi Meng.*)

Changbo He and Qi Meng are with the College of Electrical Engineering and Automation, Anhui University, Hefei 230601, China (e-mail: changbh@ahu.edu.cn; 609379887@qq.com).

Peng Chen is with the College of Engineering and the Key Laboratory of Intelligent Manufacturing Technology, Ministry of Education of China, Shantou University, Shantou, Guangdong 515063, China (e-mail: dr.pengchen@foxmail.com).

Yaqiang Jin is with the School of Qilu Transportation, Shandong University, Jinan 250061, China (e-mail: yaqiang.jin@outlook.com).

Wei Fan is with the School of Mechanical Engineering, Jiangsu University, Zhenjiang 212013, China, and also with the School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: weifan@uj.sjtu.edu.cn).

Zhibo Yang is with the State Key Laboratory for Manufacturing and Systems Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: phdapple@mail.xjtu.edu.cn).

Digital Object Identifier 10.1109/TIM.2025.3580878

endangering life and property safety. The advantage of fault diagnosis is that it can detect potential faults, reduce the number of production interruptions, and avoid catastrophic consequences. Consequently, fault diagnosis of rotating machinery is a key technology to ensure the safe operation of industrial equipment and reduce maintenance costs.

Traditional fault diagnosis methods [2] require manual feature extraction, which introduces subjectivity and uncertainty. The diagnostic performance is heavily influenced by prior expert knowledge, making it difficult to establish corresponding diagnostic models for varying working conditions. Meanwhile, due to the high latitude and nonlinearity of modern data, traditional methods are difficult to extract complex features, resulting in insufficient reliability of diagnosis results.

In recent years, scholars have begun to use methods based on deep learning [3] to break through the limitations of traditional techniques. Deep learning models can automatically extract features, reduce reliance on manual labor, and effectively capture weak features in high-dimensional data, improving the accuracy of fault diagnosis. Abdeljaber et al. [4] utilized a 1-D adaptive CNN that integrates feature extraction and classification modules for fault diagnosis. Ma et al. [5] introduced a deep residual learning approach for diagnosing nonstationary operating states of planetary gearboxes using demodulated time–frequency features. Zhang et al. [6] presented a method leveraging a hybrid attention-enhanced Resnet [7]. This approach introduces a novel Resnet with attention module to enhance fault diagnosis accuracy.

Traditional supervised transfer learning [8] refers to training a model using source domain data samples with known labels, and adjusting model parameters using target domain data samples with some known labels, for pursuing better performance. In the process of fault diagnosis, due to the high cost of data annotation and the significant differences in working modes, it usually leads to data lacking labels in the target domain. However, supervised transfer learning requires the dataset to have precise fault information as labels to enable the model to identify data differences during the learning process. For example, different fault types are needed for the fault diagnosis. Faced with this challenge, traditional supervised learning methods are clearly powerless.

Therefore, scholars began to pay attention to unsupervised transfer learning (UTL) [9], hoping to tackle unlabeled data

in the cross-domain diagnosis. UTL refers to learning features from source domain data and using transfer learning methods to adapt the model to a target domain with unknown labels, thereby addressing the issue of missing labels in the target domain data. Transfer learning methods can be categorized into four types: instance-based, map-based, network-based, and adversarial-based methods. Among them, the map-based unsupervised fault diagnosis method represents a highly effective approach. This kind of method involves feature extraction through a feature extractor, followed by mapping the features from both the source and target domains to a higher dimensional space for discrimination. Various metrics have been proposed to measure dissimilarity between the domains, including Euclidean distance, Minkowski distance, correlation alignment (CORAL [10]), maximum mean discrepancy (MMD), multi-kernel MMD (MK-MMD [11]), and joint MMD (JMMD [12]), among others. Li et al. [13] presented a method for diagnosing faults in rolling bearings through an adaptive domain approach, utilizing deep learning techniques. By minimizing the MK-MMD across various levels, they made certain that the representations learned from the source domain could be effectively transferred to the target domain. Long et al. [11] developed a new architecture for deep adaptation networks (DANs), which embeds hidden representations from each task-specific layer into the kernel Hilbert space with MMD. Wang et al. [14] propose a dynamic collaborative adversarial domain adaptation network (DCADAN) that significantly improves cross domain diagnostic performance by dynamically adjusting network architecture and multisource domain collaboration. Yang et al. [15] suggested a polynomial kernel-induced MMD (PK-MMD) distance metric, providing an alternative method for assessing domain dissimilarity. Long et al. [12] developed the joint adaptation network (JAN), a technique that utilizes the JMMD criterion to align the joint distributions of domain-specific layers, thus facilitating the learning of a transfer network. Kang et al. [16] introduced a contrastive adaptation network (CAN), in which they optimized a novel metric to explicitly simulate intraclass domain differences as well as interclass domain differences. Zheng et al. [17] proposed a structure optimized convolutional neural network for UTL. This method utilizes MMD and fast batch nuclear norm maximization to further improve the discrimination of target domain data. Jia et al. [18] combined data and physical information to design a method based on physical information to improve the reliability and interpretability of UTL.

The above methods have achieved certain results, but there is also a problem that ignore the importance of the loss function during the training phase. In fact, multiple loss functions have their own functions during the training process. However, it is not suitable to simply view them as a holistic optimization objective, and each type of loss function should be used in stages based on their respective characteristics.

To address this problem, this article proposes a multiscale feature residual network (MFSResnet) with staged optimization that can improve the coordination of loss items and the interpretability of the training process. The key contributions of this article can be summarized in the following manner.

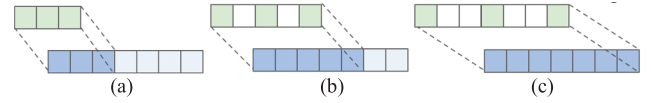


Fig. 1. Dilated convolution kernels with different dilated rate. (a) Dilated rate = 1. (b) Dilated rate = 2. (c) Dilated rate = 3.

- 1) Utilizing novel residual building blocks, improved dilated convolution and a SE attention module, a multiscale feature Resnet is proposed to improve feature extraction ability at different scales.
- 2) Integrating multiple loss items, a staged optimization method is designed to make different loss items cooperate with each other to improve the accuracy and interpretability of the model.
- 3) Introduce an innovative approach for diagnosing bearing faults across diverse operating conditions. Through the comparative analysis of two experiments, it is proved that the proposed method has an excellent performance for fault diagnosis.

The structure of this article is organized in the following manner. Section II delves into the theoretical background. Section III offers a detailed discussion on the architecture of the proposed transfer learning model. Section IV presents and examines the experimental results. Finally, Section V concludes this article by summarizing the results and their implications.

## II. BASIC THEORY

### A. Dilated Convolution

Traditional convolution layers typically have fixed kernel sizes, limiting their ability to extract features at multiple scales and their focus on global features. Differently, the dilated convolution layer can address this problem by selecting parameters such as convolution kernels and dilated rates [19].

Let  $X$  denotes the input,  $Y$  represents the output characteristics, and  $K_i$  represents the dilated convolution kernels. Thus, dilated convolution can be expressed as the following equation:

$$Y = K_{i,d_i} * X. \quad (1)$$

The asterisk (\*) symbol denotes the convolution operation. The dilated rate of the convolution kernel, denoted as  $d_i$ , determines the relative positions of the elements inside the convolution kernel. The dilated rate of one indicates the standard convolution operation. Conversely, a dilated rate greater than one signifies the spatial gap between neighboring elements within the convolution kernel, as shown in Fig. 1.

### B. Group Normalization

Group normalization (GN) [20] divides feature into multiple groups and normalizes each group's features. Compared to batch normalization (BN) [21], GN does not require calculating statistical information for each channel, reducing computational consumption under small batches.

In the case of 1-D data, it is assumed that input feature  $X$  has a shape of  $N \times C \times H$ , where  $N$  represents the batch size,  $C$  represents the number of channels, and  $H$  represents the

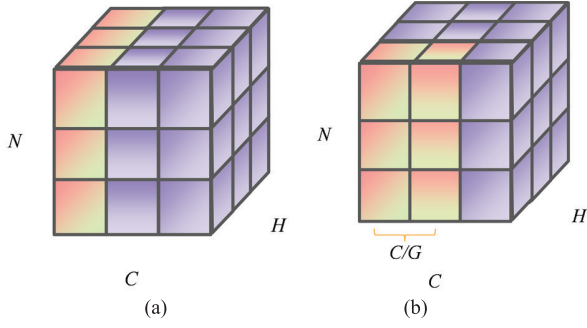


Fig. 2. (a) BN and (b) GN.

feature length. GN divides  $C$  feature channels into  $G$  groups, so that  $X$  can be divided into  $G$  subsets  $\{X_1, X_2, \dots, X_G\}$ , representing the features of the  $i$ th group. Finally, the statistics (mean and standard deviation) are obtained by normalization.

The set in BN can be expressed as (2), as shown in Fig. 2(a)

$$S_i = \{k \mid k_C = i_C\} \quad (2)$$

where  $k_c$  and  $i_c$  are indexes along channel  $C$ .

The set of GN is defined by (3) as shown in Fig. 2(b)

$$S_i = \left\{ k \mid k_N = i_N, \left\lfloor \frac{k_C}{C/G} \right\rfloor = \left\lfloor \frac{i_C}{C/G} \right\rfloor \right\} \quad (3)$$

where  $C/G$  represents the number of channels in each group.

### C. SE Attention Module

SE attention [22] is a technique comprising two primary components: squeeze and excitation. This attention first uses the squeeze part to perform the global averaging pooling operation. Then, the resulting output is taken as the input for the excitation part. Finally, an attention matrix is obtained to focus on the importance of different features.

Suppose the input feature is  $X$ , and the output expression  $Y$  of squeeze operation is the following equation:

$$Y = \frac{1}{w} \sum_{i=1}^w X_i. \quad (4)$$

$Y$  is also the input of excitation, and the expressions are shown in the following equations:

$$Z = \xi(W_2 \sigma(W_1 Y)) \quad (5)$$

$$u = Z \cdot X \quad (6)$$

where  $W_1, W_2$  denotes the weight matrix,  $\sigma$  represents activation function, and  $\xi$  represents the Sigmoid function.

### D. Loss Function

Loss function [23] plays an important role in deep learning, quantifying the disparity between a model's predicted outcomes and the actual results, thereby serving as a metric for evaluating the model's effectiveness.

1) *Cross-Entropy Loss*: Cross-entropy loss [24] is an effective performance measure that provides clear gradient information during training to help the optimization algorithm better adjust the model parameters and it is frequently employed to evaluate the discrepancy between actual labels and model predictions. This function is expressed as the following equation:

$$L_c = - \sum_{i=1}^K y_i \log(p_i) \quad (7)$$

where  $K$  represents the number of classes,  $y_i$  represents the value of the true label of the  $i$ -class, and  $p_i$  is the probability of the  $i$ -class predicted by the model.

2) *Joint MMD Loss*: Joint MMD aims to address the variance in data features across different distributions by maximizing the mean difference between data from two domains.

Assuming that the source domain data is  $X_s = \{x_1^{s1}, \dots, x_{n_s}^{sL}\}$  and the target domain data is  $X_t = \{x_1^{t1}, \dots, x_{n_t}^{tL}\}$ , the features extracted by the feature extraction network are  $Z_s = \{z_1^{s1}, \dots, z_{n_s}^{sL}\}$  and  $Z_t = \{z_1^{t1}, \dots, z_{n_t}^{tL}\}$ , where  $n_s$  and  $n_t$  are the number of sample. The expression for the JMMD loss term can be obtained as shown in the following equations:

$$K^l(z_i^{sl}, z_j^{sl}) = \exp\left(-\frac{\|z_i^{sl} - z_j^{sl}\|^2}{2\sigma^2}\right) \quad (8)$$

$$\begin{aligned} L_{\text{JMMD}}(P, Q) &= \frac{1}{n_s^2} \sum_{i=1}^{n_s} \sum_{j=1}^{n_s} \prod_{l=L} K^l(Z_i^{sl}, Z_j^{sl}) \\ &+ \frac{1}{n_t^2} \sum_{i=1}^{n_t} \sum_{j=1}^{n_t} \prod_{l=L} K^l(Z_i^{tl}, Z_j^{tl}) \\ &- \frac{1}{n_s n_t} \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} \prod_{l=L} K^l(Z_i^{sl}, Z_j^{tl}). \end{aligned} \quad (9)$$

3) *NT-Xent Loss*: NT-Xent loss [25] is employed to compute the likelihood of samples being similar or dissimilar. Similar samples are anticipated to have shorter distances, while dissimilar ones are expected to have longer distances. Given  $N$  is the number of samples. This process can be computed as the following equations:

$$L_{\text{NT-Xent}} = \frac{1}{2N} \sum_{i,j=1, i \neq j}^N [l_{i,j} + l_{j,i}] \quad (10)$$

$$l_{i,j} = -\log\left(\frac{\exp\left(\frac{s(f_i^t, f_j^t)}{\tau}\right)}{\sum_{k=1}^{2N} [k \neq i] \exp\left(\frac{s(z_i, z_k)}{\tau}\right)}\right) \quad (11)$$

$$s(f_i^t, f_j^t) = \frac{(f_i^t)^T f_j^t}{\|f_i^t\| \|f_j^t\|} \quad (12)$$

where  $f_i^t, f_j^t$  represent the target domain features via the feature extraction module. The value of  $[k \neq i]$  is equal to 1 if  $k \neq i$ .  $\tau$  is meaningful to affect the similarity distribution between samples.

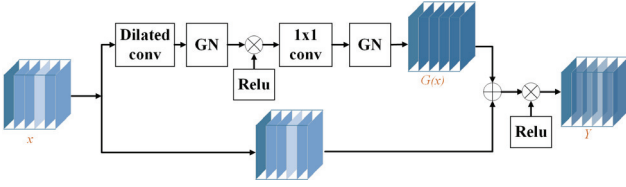


Fig. 3. RBB1.

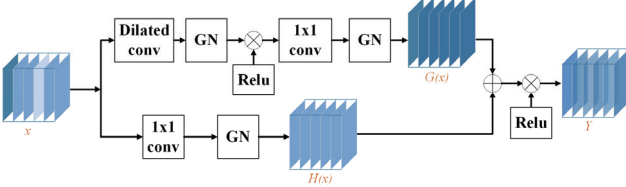


Fig. 4. RBB2.

4) *Self-Doubting Loss*: For samples in the target domain, where correct classification labels are lacking, the self-doubting loss term [26] is chosen to reduce the likelihood of the maximum probability label and enhance the classification accuracy of the samples. This is expressed in the following equations:

$$L_D = \frac{1}{n^t} \sum_{i=1}^{n^t} (y_i^t)^T \ln o_i^t \quad (13)$$

$$y_i^t = [y_{i,[1]}^t, \dots, y_{i,[L]}^t, \dots, y_{i,[C]}^t] \quad (14)$$

$$y_{i,[L]}^t = \begin{cases} 1, & j = \arg \max (o_i^t) \\ 0, & \text{else} \end{cases} \quad (15)$$

where  $o_i^t$  represents the output probability matrix of the  $i$ -target domain sample.  $n^t$  is the number of samples in the target domain.

### III. PROPOSED METHOD

#### A. Feature Extraction Module

1) *Improved Residual Basic Blocks*: The feature extraction network of MFSResnet primarily relies on two enhanced residual blocks, called residual building block one (RBB1) (Fig. 3) and residual building block two (RBB2) (Fig. 4).

RBB1 employs convolution layers in conjunction with GN to create a residual block. Its main purpose is reducing the dependence on batches and improving the feature extraction ability as much as possible in the case of small batches.

Assume the input to the residual block is  $X$ , and the output  $Y$  after passing through RBB1 is (14)

$$Y = \text{Relu}(G(x) + x). \quad (16)$$

RBB2 adds a convolution layer and GN to the residual connection. The main purpose is to make the input size and output size of residual connection consistent when changing the feature dimension, while weakening the dependence on the batch size.

Assume the input to the residual block is  $X$ , and the output  $Y$  after passing through RBB2 is the following equations:

$$Y = \text{Relu}(G(x) + H(x)) \quad (17)$$

$$H(x) = \text{GN}(\text{conv}(x)). \quad (18)$$

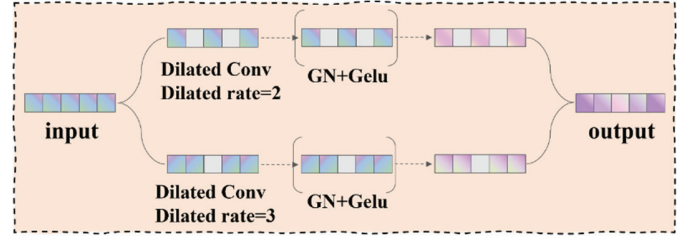


Fig. 5. Multiscale dilated convolution module.

2) *Improved Multiscale Dilated Convolution Module*: Inspired by Xiao et al. [27], this article constructs an improved multiscale dilated convolution module (Fig. 5), mainly to increase the model's receptive field and pay more attention to global features, so that the network can better extract fault feature information.

The whole module is divided into two parallel branches, each of which is composed of different dilated convolution layers, GN and GELU activation function [28], and the output of the final module is equal to the sum of the two branches output. The first branch mainly uses a dilated convolution layer with a dilated rate of 2, and obtains partial output through GN and GELU. The second branch uses a dilated convolution layer with a dilated rate of 3, also with GN and GELU for partial output. The reason for choosing dilated rates of 2 and 3 is that excessive dilated rates may result in the loss of important feature information. The sum of the last two outputs acts as the input for the next stage.

Assumed the input to multiscale dilated convolution module is  $x$ , and output of branch  $i$  is  $y_i$ . Therefore, the detailed operation at each branch can be expressed as

$$y_i = \text{GeLU} \left( \text{GN} \left( \sum_{k=1}^K w_k \cdot x_{i+r-k} \right) \right) \quad (19)$$

where  $w_k$  represents learnable kernel weights,  $r$  is dilated rate, and  $K$  is kernel size.

Furthermore, the complete multiscale operation is formally defined as

$$Y = \sum_{s=1}^S f_s(X) \quad (20)$$

where  $S$  represents the number of branches and  $f_s$  represents the feature transformation of branch  $s$ .

#### B. Staged Optimization

According to the mentioned different loss function terms in Chapter 2, this article adopts a multistaged optimization method. The training process is divided into three stages to provide corresponding theoretical support for different stages of training and enhance the model's performance, as shown in Fig. 6.

The first stage involves incorporating the cross-entropy loss term  $L_c$  and the domain adversarial loss term function  $L_{ad}$ . These components are utilized to enable the model classify the fault state of the source domain initially. Its equation is shown as

$$\text{Loss} = L_c + L_{ad}. \quad (21)$$

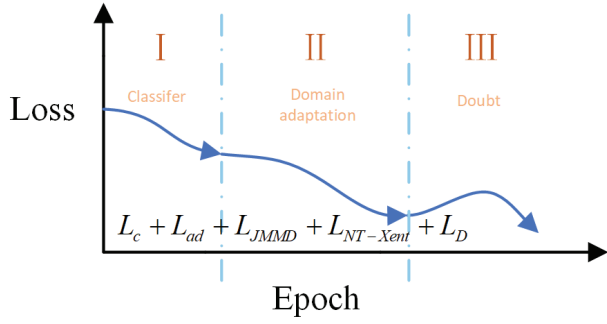


Fig. 6. Staged optimization.

In the second stage, JMMD loss term  $L_{JMMD}$  and NT-Xent loss term  $L_{NT-Xent}$  are incorporated into the loss function. The goal of these enhancements is to approximate the correspondence between the target and source domains, minimize boundary distribution samples, and align the features effectively. This loss function is shown as the following equation:

$$\text{Loss} = L_c + L_{ad} + \alpha(L_{JMMD} + \theta L_{NT-Xent}). \quad (22)$$

The third stage adds the loss term named self-doubting loss  $L_D$ , which adjusts the samples from the target domain to further improve classification accuracy.

The final optimization object is shown in the following equation:

$$\text{Loss} = L_c + L_{ad} + \alpha(L_{JMMD} + \theta L_{NT-Xent}) + (1 - \alpha)\beta L_D. \quad (23)$$

In this formula,  $\alpha$  is an adaptive parameter that decreases from 1 to 0. The specific value of  $\alpha$  in the training process is shown in the following equation:

$$\alpha = \frac{1}{1 + e^{\frac{\text{spo}-100}{10}}}. \quad (24)$$

Its purpose is to dynamically adapt to the training process, emphasize relatively important optimization objectives in the hierarchical training process, and balance the influence of multiple loss terms on the overall process.  $\beta$  and  $\theta$  are weight coefficients, serving as artificial hyperparameters.

### C. Overall Framework

The overall framework of the proposed methodology in this study comprises three primary stages, and each is designed to address specific challenges in cross-domain fault diagnosis for bearing vibration signals. The first stage involves the construction of data loaders for both source and target domains. In this stage, bearing vibration data is systematically extracted from the two domains, with a strategic partitioning ratio of 8:2 between source and target domain data, adhering to the conventional practice of allocating 80% data for training purposes. This partitioning scheme ensures adequate representation of both domains while maintaining sufficient data for effective model training.

The second stage focuses on advanced feature extraction through a sophisticated neural network architecture. A residual network (ResNet) is developed and incorporated two distinct residual building blocks (RBB1 and RBB2) to facilitate deep

feature learning. To further enhance the network's capability to capture multiscale characteristics of vibration signals, a MFSResnet is implemented by integrating a multiscale dilated convolution module. This architectural enhancement enables the model to simultaneously process local and global features, thereby improving its ability to discern subtle fault patterns across different operational conditions.

The final stage implements a staged optimization approach for model training, which is particularly crucial for handling unlabeled target domain data. This optimization strategy involves sequential training phases that progressively refine the model's diagnostic capabilities. The staged approach allows for systematic adaptation of the model from the source domain to the target domain, ensuring robust performance in fault diagnosis tasks. Through this comprehensive three-stage process, the proposed model demonstrates significant improvements in diagnostic accuracy and cross-domain adaptability, as evidenced by the experimental results presented in subsequent sections.

This hierarchical structure enables the model to effectively integrate and balance global and local features, which is essential for obtaining the inherent multiscale features of complex vibration signals. By emphasizing the collaborative combination of these features, the model improves its ability to identify subtle fault modes. In addition, the model optimizes the whole training process, improves the convergence speed and generalization performance, enhances its interpretability, and provides clearer insights for the decision process. This interpretability is particularly valuable, as understanding the basic mechanism of fault diagnosis aids root cause analysis and system improvement. Additionally, it contributes to establishing a more transparent and credible fault diagnosis framework. The final model is illustrated in Fig. 7.

## IV. EXPERIMENTAL AND RESULTS

### A. Case Western Reserve University Data

To confirm the efficacy of the suggested approach, bearing datasets from CWRU and PU were selected as the validation sets for this method. The training process is conducted using Python 3.10, with PyTorch 2.0 as the deep learning framework. The algorithm is executed on an Intel Core RTX 4090 GPU.

1) *Data Description*: The vibration data in this dataset are obtained from the SKF6205 bearing under test [29], with a sampling frequency of 12 kHz, situated at the drive end of the test stand. The test setup for the CWRU dataset is depicted in Fig. 8. The dataset comprises one healthy state ( $N$ ) and three fault states: roller element fault (RF), inner ring fault (IF), and outer ring fault (OF). Each fault state corresponds to three fault diameters: 0.007, 0.014, and 0.021 in. Consequently, the CWRU dataset can be categorized into nine fault states with varying severity levels and one healthy state. The experimental platform operates under four motor load conditions of 0, 1, 2, and 3HP. For simplicity, the datasets collected under these load conditions are labeled 0, 1, 2, and 3, respectively. Therefore, there are a total of 12 transition diagnosis tasks among the four datasets, denoted by  $T_{01}, T_{02}, T_{03}, T_{10}, T_{12}, T_{13}, T_{20}, T_{21}, T_{23}, T_{30}, T_{31},$  and  $T_{32}$ .

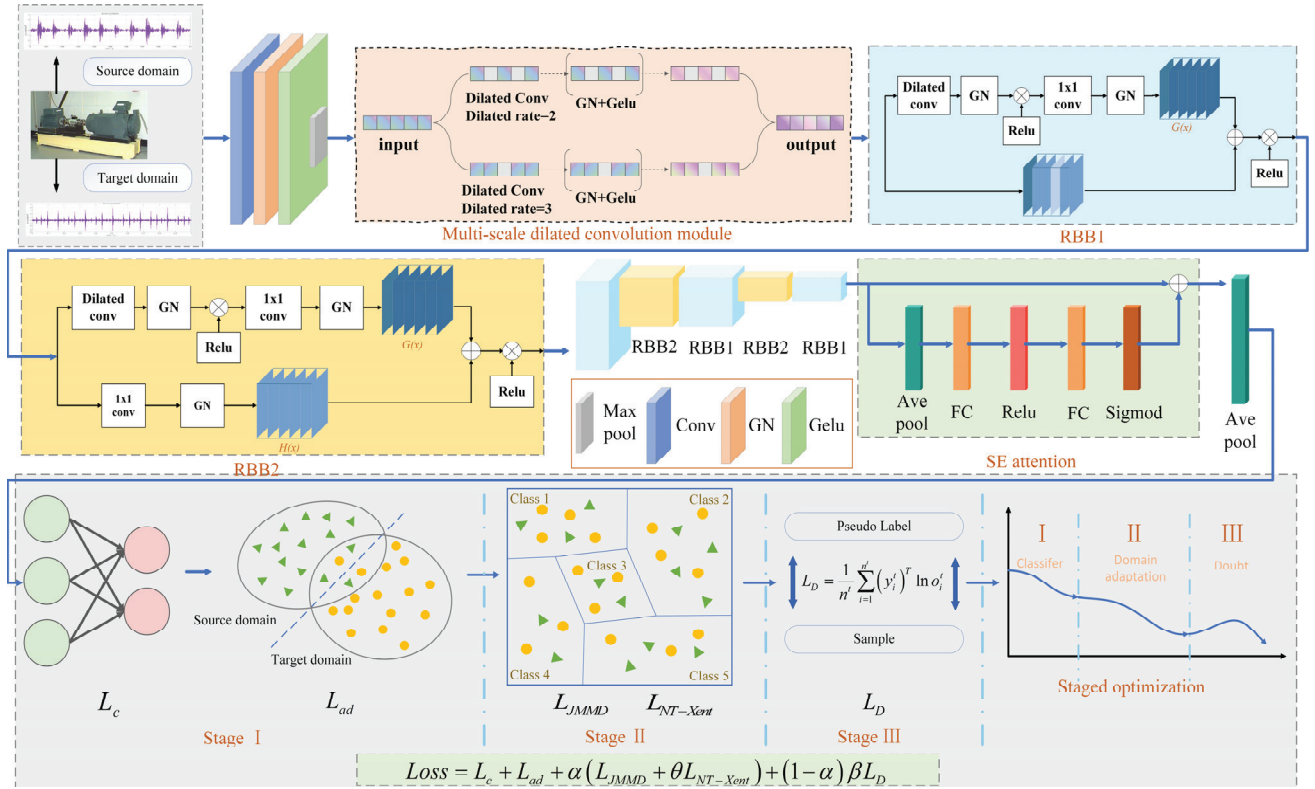


Fig. 7. Detailed structure of MFSResNet.

TABLE I  
RESULTS OF THE CWRU DATASET

Task	JAN	MKMMMD	CORAL	DANN	Resnet	LRSADTLM	WMGRNMM	Ours
$T_{01}$	92.09	93.15	92.43	85.67	98.31	100	97.78	<b>100</b>
$T_{02}$	91.31	89.56	88.43	83.59	99.68	99.72	100	<b>100</b>
$T_{03}$	87.15	85.67	81.83	77.34	99.42	99.72	99.72	<b>100</b>
$T_{10}$	94.96	97.25	96.25	92.24	98.37	<b>99.72</b>	97.50	99.50
$T_{12}$	95.86	98.95	96.34	91.35	98.59	100	100	<b>100</b>
$T_{13}$	92.14	96.54	95.65	90.45	99.26	99.72	97.50	<b>100</b>
$T_{20}$	92.44	98.33	96.27	89.24	98.36	98.89	97.22	<b>99.86</b>
$T_{21}$	93.69	97.45	95.76	91.36	98.92	98.89	97.50	<b>100</b>
$T_{23}$	95.86	97.93	97.95	98.85	99.21	100	100	<b>100</b>
$T_{30}$	92.05	92.18	91.34	92.55	96.08	99.17	98.61	<b>99.75</b>
$T_{31}$	92.95	93.43	92.35	92.27	97.13	99.44	97.22	<b>100</b>
$T_{32}$	96.84	96.74	95.87	93.56	99.76	100	100	<b>100</b>
Ave	93.12	94.76	93.36	89.87	98.59	99.61	98.59	<b>99.93</b>

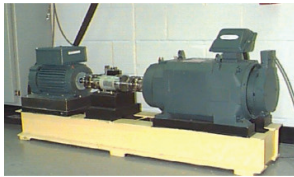


Fig. 8. CWRU dataset test bench.

The sample length used in this article is 1024. There are 1800 data samples, with 180 samples for each state.

2) *Experimental Results:* To verify the effectiveness of MFSResnet for UTL, 80% of the source domain data is designated as the test set, while 20% of the target domain data is allocated as the verification set. Specifically, 1440 samples

from the source domain data constitute the training set, and 360 samples from the target domain serve as the transfer learning verification set, with 36 samples in each category. Additionally, JAN [12], MKMMMD [30], CORAL [31], DANN [32], LRSADTLM [33], and WMGRNMM [34] are selected as comparison models to evaluate the experimental results.

The optimization method for JAN, MKMMMD, CORAL, and DANN is an overall optimization method with  $L_c$  and  $L_{JMMD}$ . And for LRSADTLM and WMGRNMM, the corresponding methods are introduced in [33] and [34]. According to the staged optimization in Chapter 3, the training process is divided into three stages and each stage comprises 50 training epochs. The final experimental results are presented in Table I.

Fig. 9 shows the accuracy of different models in all transfer tasks. From this figure, it is evident that the MFSResnet

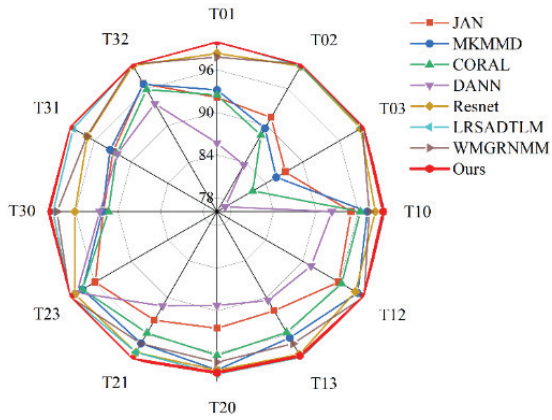


Fig. 9. Accuracy radar map of different models about CWRU dataset.

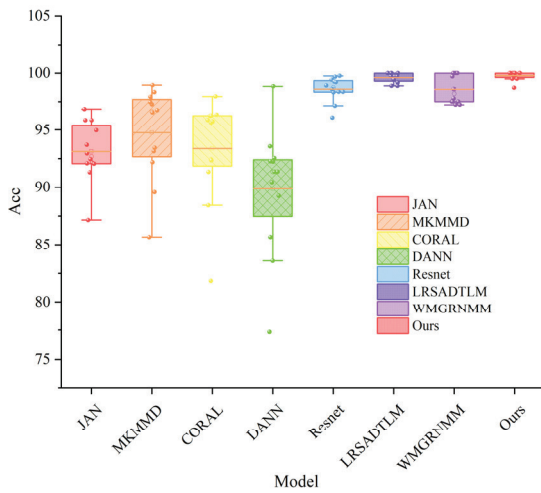


Fig. 10. Distribution of CWRU experimental results.

proposed in this article demonstrates excellent performance in the CWRU datasets. Specifically, MFSResnet achieves 100% accuracy in  $T_{01}$ ,  $T_{02}$ ,  $T_{03}$ ,  $T_{12}$ ,  $T_{13}$ ,  $T_{21}$ ,  $T_{23}$ ,  $T_{31}$ , and  $T_{32}$ . Additionally, MFSResnet achieves the lead on the majority of tasks, with an average accuracy of 99.93% for all tasks. It leads other models by at least 0.32% and up to 10.06%.

Fig. 10 shows the distribution and average accuracy of different models in all tasks. The MFSResnet's accuracy distribution of each task is relatively concentrated, with little difference, and the most other models' distributions are significantly more scattered. Although LRSADTLM's stability is also outstanding in this figure, MFSResnet's average accuracy and most task results are higher, which demonstrates the superiority of the improved model.

For instance, the t-SNE diagram in  $T_{03}$  (Fig. 11) demonstrates the model's effective classification. The data distribution is uniform across each class, and the differences between classes are evident, which further validates the powerful feature extraction and generalization capabilities of MFSResnet.

Fig. 12 shows the changes in loss value of the target domain data during the training process. From Fig. 12, it can be seen that under the proposed staged optimization method, the loss value of the target domain decreases according to the

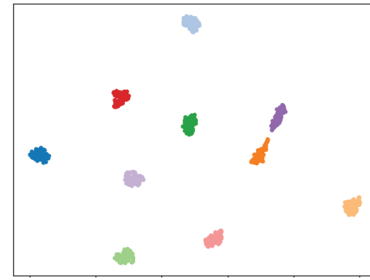


Fig. 11. T-SNE diagram.

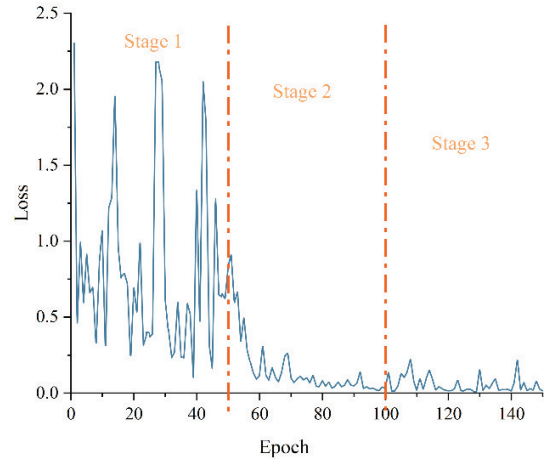


Fig. 12. Target domain data loss value during training process.

expected trend. In Stage 1, the data exhibit clear fluctuations with an overall downward trend. This is because in this stage the model focuses on classification and preliminary separation of samples in both domains. And the model training in this stage can facilitate unsupervised transfer in the subsequent domain adaptation stage. In Stage 2, due to the effects of the JMMD and NT-Xent loss terms, there is a significant decrease in the loss value. Compared to Stage 1, the loss value decreases significantly, which represents the effectiveness of the domain adaptation stage. This stage helps the proposed model align the data distribution of the source and target domains. In Stage 3, the self-doubting loss term doubts the correctness of target sample's classification result, enhancing the credibility of unsupervised transfer results.

In order to demonstrate the advantages of the proposed staged optimization method, the following paragraph takes  $T_{20}$  as an example to conduct ablation experiments on the NT-Xent loss term and self-doubting loss term.

The comparative experiment analysis for the optimization object with and without self-doubting loss is conducted, and the result is shown in Fig. 13. It can be seen that the overall trend of target domain loss value with self-doubting loss is consistent with no doubt. In stage 1 and stage 2, there is no significant difference between two experiments, because the two optimization objects contain the same loss items. In stage 3, the self-doubting loss is added in the optimization object, which makes the model begin to doubt the authenticity of target sample's classification result. Although the self-doubting loss item makes the target loss value begin to

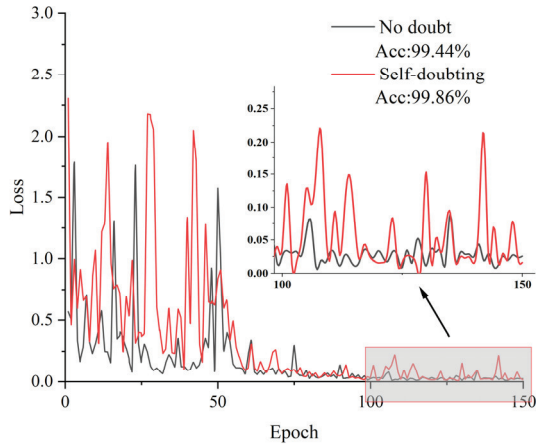


Fig. 13. Target domain data loss value in the  $L_D$  ablation experiment.

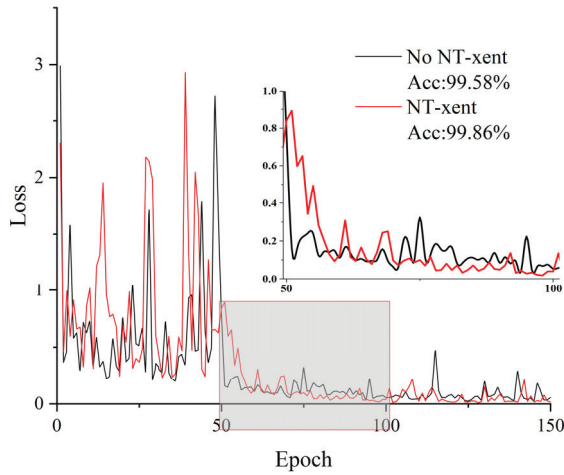


Fig. 14. Loss value in the  $L_{NT-Xent}$  ablation experiment.

fluctuate greatly, the final accuracy remains at the higher level. This phenomenon helps the model avoid over fitting or falling into the local optimal solution in the training process, and enhances the interpretability of the training process.

Similarly, an ablation experiment is conducted on the NT-Xent loss term to verify its effectiveness in the method proposed in this article. Fig. 14 shows the target domain data loss value about this ablation experiment. According to this figure, during the second stage of the training process,  $L_{NT-Xent}$  causes the loss of the target domain to gradually decrease. On the contrary, the loss value without  $L_{NT-Xent}$  suddenly decreases and remains at a relatively stable level. Then, the loss value in this article begins to be lower than the comparison scheme at around 70 rounds. This phenomenon indicates that  $L_{NT-Xent}$  enables the model to gradually learn the overall differences between source and target domain samples, which helps the model better distinguish samples between source and target domains during the domain adaptation stage.

### B. Paderborn University Bearing Data

1) *Data Description*: The vibration signals in this dataset are collected from a 6203-type rolling bearing, and they are captured using a piezoelectric accelerometer with a sampling frequency of 64 kHz. The PU dataset [35] predominantly

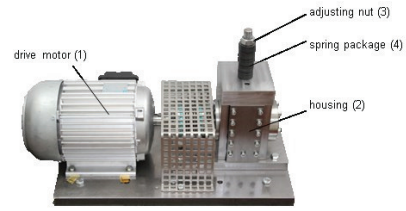


Fig. 15. PU dataset accelerated life test bench.

TABLE II  
PU DATASET PARAMETERS

Task	Rotational speed (rpm)	Load torque (Nm)	Radial force (N)	Sample size
0	1500	0.7	1000	3203
1	1500	0.1	1000	3219
2	1500	0.7	400	3251

TABLE III  
FAULT TYPE DETAILS

Bearing element	Damage detail	Extent of damage
OR	S + FP	$\leq 0.2\text{mm}$
OR	S + PI	$\leq 0.2\text{mm}$
OR	R + FP	$> 0.2\text{mm}$ and $\leq 4.5\text{mm}$
OR	S + FP	$\leq 0.2\text{mm}$
OR	R + PI	$\leq 0.2\text{mm}$
OR + IR	M + FP	$> 0.2\text{mm}$ and $\leq 4.5\text{mm}$
OR + IR	M + FP	$> 4.5\text{mm}$ and $\leq 13.5\text{mm}$
OR + IR	M + PI	$\leq 0.2\text{mm}$
IR	M + FP	$\leq 0.2\text{mm}$
IR	S + FP	$> 4.5\text{mm}$ and $\leq 13.5\text{mm}$
IR	R + FP	$\leq 0.2\text{mm}$
IR	S + FP	$> 0.2\text{mm}$ and $\leq 4.5\text{mm}$
IR	S + FP	$\leq 0.2\text{mm}$

comprises two types of data: artificial damage faults and real damage faults resulting from accelerated life experiments. To investigate the transfer diagnosis task involving real faults, this article selected real damage faults generated from accelerated life tests. The accelerated life experiment platform is depicted in Fig. 15.

The real damage faults dataset includes three variables: the rotational velocity of the driving mechanism, the lateral pressure applied to the experimental bearing, and the torque load on the driving system. This dataset covers three distinct experimental scenarios, labeled simply as 0, 1, and 2. The basic parameters are outlined in Table II. The final transfer diagnosis task is denoted as  $T_{01}$ ,  $T_{02}$ ,  $T_{10}$ ,  $T_{12}$ ,  $T_{20}$ , and  $T_{21}$ . The dataset encompasses 13 fault types. For ease of numerical representation of each fault type, detailed fault information is provided in Table III.

2) *Experimental Results*: Choose 80% of the data from the source domain to form the training set, the remaining 20% forms the test set. And allocate 20% from the target domain as the validation set. The training procedure remains segmented into three stages and each stage comprises 50 epochs. To evaluate the model's capability to generalize in UTL, the basic learning parameters of the model remain the same as the CWRU experiment. Similarly, to verify the effectiveness of

TABLE IV  
RESULTS OF THE PU DATASET

Task	JAN	MKMMMD	CORAL	DANN	LRSADTLM	Li	Ours
$T_{01}$	95.61	94.26	95.68	94.59	98.91	98.26	<b>100</b>
$T_{02}$	78.85	76.60	55.73	83.53	96.89	96.80	<b>99.91</b>
$T_{10}$	92.86	91.45	81.98	93.44	99.55	99.01	<b>100</b>
$T_{12}$	77.76	78.07	63.62	81.53	97.05	91.77	<b>99.33</b>
$T_{20}$	72.86	60.76	51.52	74.28	89.96	91.92	<b>98.83</b>
$T_{21}$	76.03	69.26	59.91	80.13	<b>89.32</b>	89.21	87.38
Average	82.33	78.40	66.41	84.58	95.28	94.49	<b>97.58</b>

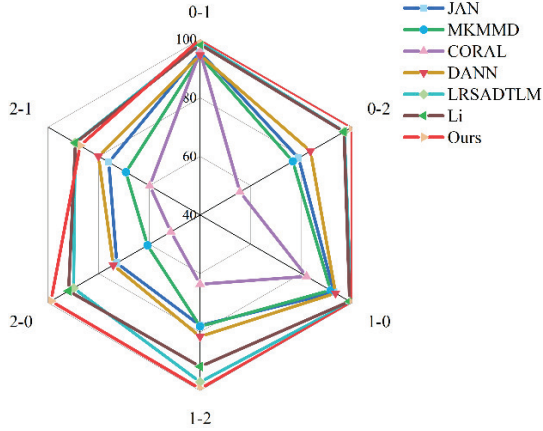


Fig. 16. Accuracy radar map of different models about PU dataset.

MFSResnet, JAN, MKMMMD, CORAL, DANN, LRSADTLM, and Li's model in [36] continue to be chosen for comparative experiments, with the results displayed in Table IV.

Fig. 16 shows the accuracy of different models in all transfer tasks. Compared to other models, MFSResnet achieves a maximum increase of 31.17% in average accuracy and a minimum increase of 13%. It also reduces the average standard deviation by up to 1.68. In addition, MFSResnet achieves 100% accuracy in  $T_{01}$  and  $T_{10}$ , surpassing the worst-performing model by 18.02%. As the load torque and radial force vary simultaneously, it is more difficult to extract fault feature in some tasks, such as  $T_{21}$ . In  $T_{21}$ , although MFSResnet's accuracy is lower than LRSADTLM and Li, it also obtains a competitive accuracy and just decreases by about 2%. From the average accuracy, MFSResnet is more powerful than LRSADTLM and Li.

As illustrated in the box Fig. 17, the red mean line distinctly indicates that the model proposed in this article achieves the highest average accuracy across all tasks. The concentrated data distribution of MFSResnet shows its superior performance, which contrasts sharply with the more dispersed distributions of other models. Such a concentrated distribution not only underscores the model's stability but also demonstrates its robustness in handling diverse fault diagnosis tasks. In comparison, other models exhibit wider data spreads, indicating less consistent performance.

Figs. 18 and 19 show the ablation experiments results on  $L_{NT-Xent}$  and  $L_D$ , which verifies the effectiveness of the stage

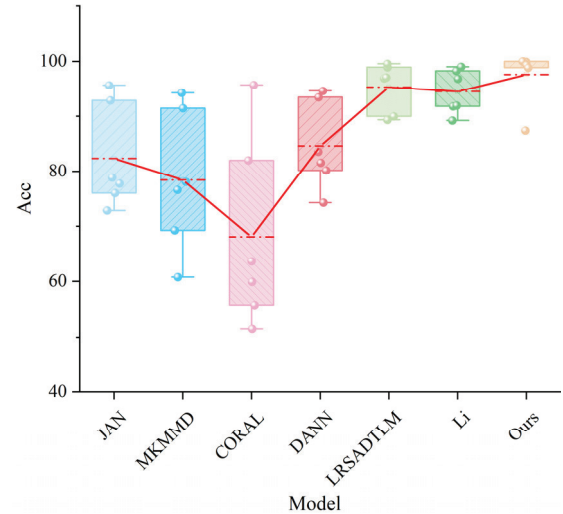


Fig. 17. Distribution of PU experimental results.

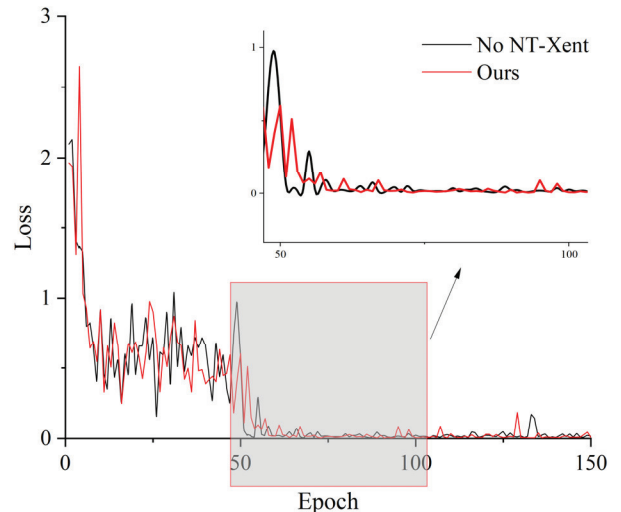


Fig. 18. Target domain data loss value in the  $L_{NT-Xent}$  ablation experiment.

optimization method proposed in this article. It can also be seen that the NT-Xent loss term effectively slows down the decrease of loss value in stage 2, which can effectively reduce gradient explosion. Thus, the domain adaptation stage avoids falling into local optima and enhances the effectiveness of unsupervised processes. As shown in Fig. 19, it is easy to see that the loss value increases quickly in the final training

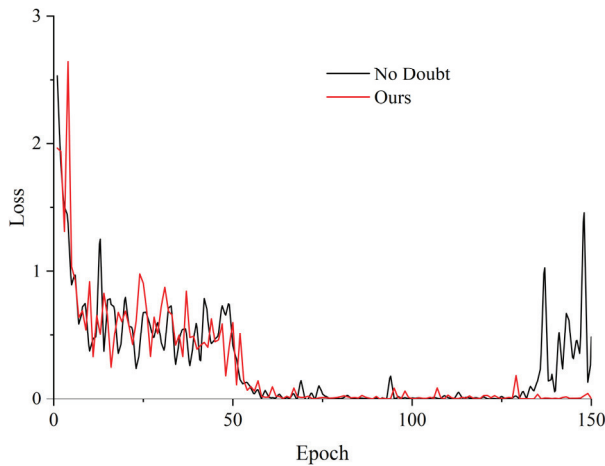


Fig. 19. Target domain data loss value in the  $L_D$  ablation experiment.

process. This represents a significant misclassification of the samples in the target domain, which further reflects the unreasonable optimization objectives or other issues such as overfitting. But, the self-doubting loss term helps the training process avoids this phenomenon. Therefore, the proposed optimization objective can be found greatly enhances the performance of the training process.

## V. CONCLUSION

This article introduces a MFSResnet for unsupervised cross-domain fault diagnosis. This model comprises a feature extraction network and a domain adaption module. The feature extraction network is based on an improved Resnet with the improved multiscale dilated convolution and SE attention module. Meanwhile, the domain adaption module combines multiple loss function terms to achieve staged training, containing cross-entry loss, adversarial loss, JMMD loss, NT-Xent loss, and self-doubting loss. Experimental results from two sets of cross-working conditions data demonstrate that the proposed model can accurately identify each fault category in the target domain, which showcases its effectiveness for cross-domain fault diagnosis. In the actual industrial scene, unknown fault types may occur in the target domain. Therefore, it is possible to further explore the diagnostic capabilities of the model in detecting unknown categories, and enhance its practical value in the future.

## REFERENCES

- [1] A. Heng, S. Zhang, A. C. C. Tan, and J. Mathew, "Rotating machinery prognostics: State of the art, challenges and opportunities," *Mech. Syst. Signal Process.*, vol. 23, no. 3, pp. 724–739, 2009.
- [2] B. Mahesh, "Machine learning algorithms—A review," *Int. J. Sci. Res.*, vol. 9, no. 1, pp. 381–386, Jan. 2020.
- [3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [4] O. Abdeljaber, O. Avci, S. Kiranyaz, M. Gabbouj, and D. J. Inman, "Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks," *J. Sound Vib.*, vol. 388, pp. 154–170, Feb. 2017.
- [5] S. Ma, F. Chu, and Q. Han, "Deep residual learning with demodulated time-frequency features for fault diagnosis of planetary gearbox under nonstationary running conditions," *Mech. Syst. Signal Process.*, vol. 127, pp. 190–201, Jul. 2019.
- [6] K. Zhang, B. Tang, L. Deng, and X. Liu, "A hybrid attention improved ResNet based fault diagnosis method of wind turbines gearbox," *Measurement*, vol. 179, Jul. 2021, Art. no. 109491.
- [7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [8] S. J. Pan, "Transfer learning," *Learning*, vol. 21, pp. 1–2, Jan. 2020.
- [9] Z. Zhao et al., "Applications of unsupervised deep transfer learning to intelligent fault diagnosis: A survey and comparative study," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–28, 2021.
- [10] B. Sun and K. Saenko, "Deep CORAL: Correlation alignment for deep domain adaptation," in *Proc. Comput. Vis.-ECCV Workshops*, Amsterdam, The Netherlands, 2016, pp. 443–450.
- [11] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37, Jul. 2015, pp. 97–105.
- [12] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2208–2217.
- [13] X. Li, W. Zhang, Q. Ding, and J.-Q. Sun, "Multi-layer domain adaptation method for rolling bearing fault diagnosis," *Signal Process.*, vol. 157, pp. 180–197, Apr. 2019.
- [14] X. Wang, H. Jiang, M. Mu, and Y. Dong, "A dynamic collaborative adversarial domain adaptation network for unsupervised rotating machinery fault diagnosis," *Rel. Eng. Syst. Saf.*, vol. 255, Mar. 2025, Art. no. 110662.
- [15] B. Yang, Y. Lei, F. Jia, N. Li, and Z. Du, "A polynomial kernel induced distance metric to improve deep transfer learning for fault diagnosis of machines," *IEEE Trans. Ind. Electron.*, vol. 67, no. 11, pp. 9747–9757, Nov. 2020.
- [16] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 4893–4902.
- [17] B. Zheng, J. Huang, X. Ma, X. Zhang, and Q. Zhang, "An unsupervised transfer learning method based on SOCNN and FBNN and its application on bearing fault diagnosis," *Mech. Syst. Signal Process.*, vol. 208, Feb. 2024, Art. no. 111047.
- [18] N. Jia, W. Huang, C. Ding, J. Wang, and Z. Zhu, "Physics-informed unsupervised domain adaptation framework for cross-machine bearing fault diagnosis," *Adv. Eng. Informat.*, vol. 62, Oct. 2024, Art. no. 102774.
- [19] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.
- [20] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
- [21] N. Bjorck, C. P. Gomes, B. Selman, and K. Q. Weinberger, "Understanding batch normalization," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2018, pp. 4470–4478.
- [22] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2018, pp. 7132–7141.
- [23] Z. Zhao et al., "Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study," *ISA Trans.*, vol. 107, pp. 224–255, Dec. 2020.
- [24] A. Mao, M. Mohri, and Y. Zhong, "Cross-entropy loss functions: Theoretical analysis and applications," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2023, pp. 23803–23828.
- [25] Y. An, K. Zhang, Y. Chai, Q. Liu, and X. Huang, "Domain adaptation network base on contrastive learning for bearings fault diagnosis under variable working conditions," *Expert Syst. Appl.*, vol. 212, Feb. 2023, Art. no. 118802.
- [26] Z. An, X. Jiang, J. Cao, R. Yang, and X. Li, "Self-learning transferable neural network for intelligent fault diagnosis of rotating machinery with unlabeled and imbalanced data," *Knowl.-Based Syst.*, vol. 230, Oct. 2021, Art. no. 107374.
- [27] Y. Xiao, H. Shao, Z. Min, H. Cao, X. Chen, and J. Lin, "Multiscale dilated convolutional subdomain adaptation network with attention for unsupervised fault diagnosis of rotating machinery cross operating conditions," *Measurement*, vol. 204, Nov. 2022, Art. no. 112146.
- [28] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.
- [29] W. A. Smith and R. B. Randall, "Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study," *Mech. Syst. Signal Process.*, vols. 64–65, pp. 100–131, Dec. 2015.
- [30] M. Long, Y. Cao, Z. Cao, J. Wang, and M. I. Jordan, "Transferable representation learning with deep adaptation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 3071–3085, Dec. 2019.

- [31] X. He, "Quantum correlation alignment for unsupervised domain adaptation," *Phys. Rev. A, Gen. Phys.*, vol. 102, no. 3, pp. 153–171, Sep. 2020.
- [32] Y. Ganin et al., "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1–35, 2016.
- [33] X. Yu, Y. Wang, Z. Liang, H. Shao, K. Yu, and W. Yu, "An adaptive domain adaptation method for rolling bearings' fault diagnosis fusing deep convolution and self-attention networks," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–14, 2023.
- [34] X. Yu et al., "A wavelet packet transform-based deep feature transfer learning method for bearing fault diagnosis under different working conditions," *Measurement*, vol. 201, Sep. 2022, Art. no. 111597.
- [35] C. Lessmeier, J. K. Kimotho, D. Zimmer, and W. Sextro, "Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification," in *Proc. PHM Soc. Eur. Conf.*, Jul. 2016, vol. 3, no. 1, pp. 1–8.
- [36] Y. Li, Y. Song, L. Jia, S. Gao, Q. Li, and M. Qiu, "Intelligent fault diagnosis by fusing domain adversarial training and maximum mean discrepancy via ensemble learning," *IEEE Trans. Ind. Informat.*, vol. 17, no. 4, pp. 2833–2841, Apr. 2021.



**Changbo He** received the B.S. degree from Northeastern University, Shenyang, China, in 2013, and the Ph.D. degree from Dalian University of Technology, Dalian, China, in 2019.

He is currently an Associate Professor with the School of Electrical Engineering and Automation, Anhui University, Hefei, China. His research interests include weak fault feature extraction and condition monitoring.



**Qi Meng** received the B.S. degree in electrical engineering and its automation from Nantong University, Nantong, China, in 2022. He is currently pursuing the master's degree in control engineering with Anhui University, Hefei, China.

His main research interests include signal processing and intelligent fault diagnosis.



**Peng Chen** (Member, IEEE) received the M.S. degree in mechatronic engineering and the Ph.D. degree in mechanical engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2016 and 2020, respectively, under the supervision of Prof. Ming J. Zuo.

From 2019 to 2020, he was an International Scholar with the Department of Mechanical Engineering, Katholieke Universiteit Leuven, Leuven, Belgium. He is currently an Assistant Professor with the College of Engineering, Shantou University, Shantou, Guangdong, China. His research interests include comprise signal and acoustics processing, deep learning, intelligent interactions, and fault diagnosis and prognostics for advanced industrial equipment.



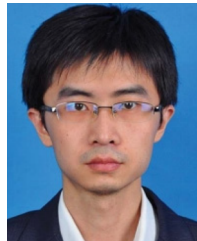
**Yaqiang Jin** received the M.S. degree in mechanical engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2018, and the Ph.D. degree in mechanical engineering from the University of Lyon, Lyon, France, in 2022.

He is currently a Post-Doctoral Fellow with Shandong University, Jinan, China. His research interests include condition monitoring and signal processing.



**Wei Fan** (Member, IEEE) received the B.S. and M.S. degrees from Soochow University, Suzhou, China, in 2012 and 2015, respectively, and the Ph.D. degree from the City University of Hong Kong, Hong Kong, in 2018.

She is currently a Professor with the School of Mechanical Engineering, Jiangsu University, Zhenjiang, China. Her research interests include signal processing and machinery fault diagnosis.



**Zhibo Yang** (Member, IEEE) received the Ph.D. degree in mechanical engineering from Xi'an Jiaotong University, Xi'an, China, in 2013.

He is currently a Professor of mechanical engineering with Xi'an Jiaotong University, Xi'an, China. His current research interests include the wavelet finite-element method and signal processing in machine condition monitoring.