Deep Residual Joint Transfer Strategy for Cross-Condition Fault Diagnosis of Rolling Bearings

Songjun Han and Zhipeng Feng

School of Mechanical Engineering, University of Science and Technology Beijing, Beijing 100083, China

Abstract: Rolling bearings are key components of the drivetrain in wind turbines, and their health is critical to wind turbine operation. In practical diagnosis tasks, the vibration signal is usually interspersed with many disturbing components, and the variation of operating conditions leads to unbalanced data distribution among different conditions. Although intelligent diagnosis methods based on deep learning have been intensively studied, it is still challenging to diagnose rolling bearing faults with small amounts of samples. To address the above issue, we introduce the deep residual joint transfer strategy method for the cross-condition fault diagnosis of rolling bearings. One-dimensional vibration signals are pre-processed by overlapping feature extraction techniques to fully extract fault characteristics. The deep residual network is trained in training tasks with sufficient samples, for fault pattern classification. Subsequently, three transfer strategies are used to explore the generalizability and adaptability of the pre-trained models to the data distribution in target tasks. Among them, the feature transferability between different tasks is explored by model transfer, and it is validated that minimizing data differences of tasks through a dual-stream adaptation structure helps to enhance generalization of the models to the target tasks. In the experiments of rolling bearing faults with unbalanced data conditions, localized faults of motor bearings and planet bearings are successfully identified, and good fault classification results are achieved, which provide guidance for the cross-condition fault diagnosis of rolling bearings with small amounts of training data.

Keywords: fault diagnosis; feature transferability; rolling bearing; transfer strategy; wind turbine

I. INTRODUCTION

Rolling bearings play a vital role in the transmission system of wind turbines. Rolling bearing faults are likely to result in unexpected downtime and economic loss [1–3]. Therefore, rolling bearing fault diagnosis is important to the steady operation of wind turbines.

As vibration signals contain diverse components, such as rotational frequencies and their harmonics, noise, or even different signal coupling, extracting fault information from vibration signals has become the typical approach for fault diagnosis of rotating machinery. Some researchers have effectively explored the fault mechanism of rotating machinery using signal analysis techniques. For example, Feng et al. [4,5] calculated the fault feature frequencies of damaged components in planetary gearboxes, thus identifying the gearbox faults. Jain et al. [6] built a dynamics model of planet gear system by the planet bearing fault mechanism for analyzing the frequency domain characteristics of vibration response. However, traditional methods largely rely on the prior knowledge about research objects to extract obvious fault features, such as eigenvalue and fault feature frequency of signals. Therefore, extracting the hidden weak fault features from lots of signals is highly challenging. With the application of intelligent diagnosis technology, such as machine learning, the fault information of monitoring data can be automatically mined and processed, thereby avoiding the deficiency of traditional methods [7–10]. Driven by the network structures and training algorithms, deep learning methods [11–14] based on

sufficient labeled data have been extensively investigated in the intelligent diagnosis of rotating machinery. Jiao et al [14] established multivariate encoder information dataset by blending multiple encoder data to effectively diagnose different fault categories of planetary gearboxes. Xu and Souza et al [15,16] improved the deep network by shallow algorithm structures for health identification of bearings and thus classified different fault types by the constructed classification module. Verstraete et al [17] performed a featureless learning method for digging out fault features of rolling bearing datasets, thus addressing the uncertainty and bias during the feature learning. Although deep learning methods have achieved excellent results in identifying faults of mechanical equipment, they are still highly controversial. The above-mentioned methods usually presume that feature distribution of test sample sets is the equivalent to that of training sample sets, and train deep networks on monitoring data with abundant health label information. However, the available fault data in engineering applications are much less than normal data, that is, the lack of faulty training samples, which makes the generalization of trained models insufficient in target tasks.

Therefore, transfer learning methods [18] were introduced to the cross-condition fault diagnosis of rotating machines. Transfer learning is intuitively divided into multiple categories depending on the learning style [19], with feature transfer learning receiving the most extensive research. Some researchers [20,21] have presented to incorporate transfer learning into the training of deep networks or to transfer models with strong generalization among different diagnostic tasks, thereby resolving the problems of few labeled samples and vibration data diversity in bearing diagnosis tasks. The diagnostic performance of these

___

Corresponding author: Zhipeng Feng (e-mail: fengzp@ustb.edu.cn).

methods may exhibit instability due to insufficient training samples and large feature differences among different tasks. The tight internal structure of bearings increases the difficulty of feature extraction, which leads to the inability to adequately extract fault information for bridging the data differences among different conditions. Thus, some domain adaptation methods based on metric distance are applied to the cross-condition fault diagnosis of rotating equipment. He et al [22,23] utilized maximum mean discrepancy (MMD) metric to calculate the data difference among training and target tasks for fault diagnosis of machinery. Sun et al [24] adapted deep correlation alignment (D-CORAL) to calculate the data discrepancy (CORAL loss) instead of MMD loss. Other researchers utilized some distance metrics and divergences, such as Wasserstein distance [25] and Kullback–Leibler divergence, to gauge the data differences between different tasks. The speed or load of rotating equipment in engineering applications changes frequently, resulting in widely varying data distributions in different conditions. Therefore, when the data discrepancy is too large or the metric distance is inappropriate, negative transfer effects can occur in the cross-condition transfer diagnosis. According to this, we explored the feature transferability between different diagnosis tasks by freezing and fine-tuning the parameters and compared the effects of different metric distances on transfer tasks.

In engineering applications, the collected normal signals are usually much more than the fault signals, resulting in the lack of fault data and data imbalance during the network training. Therefore, this paper proposes a method combining deep residual networks and transfer diagnosis strategy to accomplish the cross-condition diagnosis for rolling bearing faults. Sufficient training samples are utilized in the source task to improve the generalization ability of network models, and the fault information is hierarchically compiled into weight parameters of network layers. The trained models are utilized in the new diagnosis task of bearings by means of feature transfer. Moreover, a transfer strategy is proposed to evaluate the data distribution between different tasks using MMD distance. Compared with traditional deep learning method, this method is able to effectively mine the fine-grained feature information for fault diagnosis of target tasks, rather than taking advantage of numerous labeled samples with same data distribution to train the network. The major contributions of this paper are categorized into two areas:

(1) A transfer framework is designed on the basis of pretrained models in the source domain, and the weights of models are recompiled to adapt to target tasks, thus exploring the feature transferability among different conditions using different transfer strategies.

(2) A dual-stream feature adaptation strategy is proposed on the basis of fine-tuning the weights. This method gauges the distribution distance among different tasks to evaluate the data differences of tasks, thus bridging the gap of data distribution among different diagnostic tasks, which can enhance the generalization of models to target tasks.

The rest of this paper is organized as follows. Section II describes a brief research background. We propose a deep transfer learning framework for bearing fault diagnosis in Section III. Section IV introduces experimental setting and displays the experimental results. Lastly, conclusions are derived in Section V.

II. RESEARCH BACKGROUND

A. DIAGNOSTIC TASK DESCRIPTION

Vibration signal from different conditions exhibits diverse data characteristic distribution. Conventional supervised deep learning demands adequate labeled fault samples to boost the diagnostic performance of network models, and separate models have to be constructed for different diagnostic tasks. Bearing fault diagnosis in engineering applications can hardly meet the above requirements. Moreover, vibration signals of bearings are interspersed with high-frequency noise and other disturbing components, and the frequent changes in operating conditions lead to increased data discrepancies among signals. Therefore, we introduce transfer diagnosis methods for the cross-condition fault diagnosis of bearings. Domain is used to delineate data with different feature distributions, that is, source domain $\mathcal{D}_S$ and target domain $\mathcal{D}_T$, where the probability distribution $\mathcal{P}(\mathbf{x})$ is the indicator to distinguish the two domains. In the transfer cases, the probability distributions within both domains are dissimilar ($\mathcal{P}^S(\mathbf{x}) \neq \mathcal{P}^T(\mathbf{x})$). The training aim of the source domain is to uncover a mapping condition in similar tasks that enables automatic learning of feature information in the target task. The core of transfer learning is to mine feature similarities between different diagnostic tasks and leverage the characteristics in labeled tasks $\mathcal{D}_S = \{x_i^S, y_i^S\}^n$ to complete the diagnosis of the unlabeled tasks $\mathcal{D}_T = \{x_j^T\}^m$, among which the data probability distribution between diagnostic tasks is different.

B. FEATURE LEARNING

When in excess of a certain depth of layers in the network, the training effectiveness of deep convolutional neural networks (CNNs) will degrade in the absence of training samples. Because the gradient accumulates during the backpropagation, the gradient vanishing/exploding issue may occur. Therefore, the residual network (ResNet) structures with multiple residual blocks [26] are proposed. The residual block structure diagram is shown in Fig. 1. From the structure diagram, the underlying mapping $H(x)$ is the sum of the mapping $F(x,w)$ and input x. The learning goal is no longer a stand-alone outcome, but rather a differential value between the mapping $H(x)$ and the value x. The residual block changes the training goal of networks, and the mapping $H(x)$ does not vanish when $F(x,w)$ is close to 0. The output $G(x)$ is the activation value of the mapping $H(x)$ through the Rectified Linear Unit.
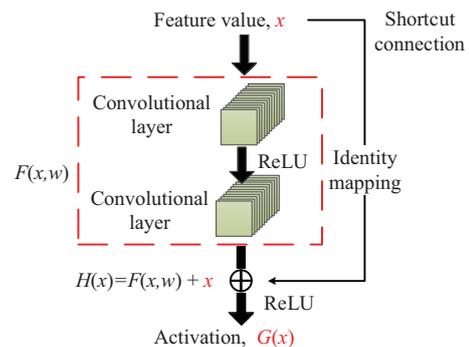

Fig. 1. The residual block structure diagram.

The convolutional layers complete the convolution operation by sliding feature extraction filters on signal matrices.

Convolutional values are further processed by the activation unit to obtain a series of feature maps. The operational formulation of convolution process is given by:

$$X_j^l = f\left(\sum_i X_i^{l-1} * \omega_{ij}^l + b_j^l\right) \qquad (1)$$

where X denotes the value of each neuron in network layers; ω means the weight parameters of convolutional kernels; b is the bias; and $f(\cdot)$ is the nonlinear activation function:

$$f(x) = \max(0, \lg(1 + e^x)) \qquad (2)$$

Activation features obtained after convolution operation are still high-dimensional matrices. To remove redundant features and decrease the computation amount, the dimensionality of activation features is reduced via the down-sampling layers, but feature scale invariance is maintained to a certain extent. Therefore, the max-pooling method is followed to diminish dimension of feature matrices. The expression of the maximum pooling process is described as:

$$x_j^{l+1}(i) = \max_{(i-1)W+1 \le t \le iW} x_j^l(t) \qquad (3)$$

where $x(\cdot)$ indicates the feature value; W is a local region for pooling operations.

After several convolution operations for feature extraction, the final result in the convolution layer is still the multidimensional features. Therefore, few fully connected layers and a classification layer are stitched together after the convolution layers to flatten and classify the multidimensional features. In the final classification layer, the output is the probability value associated with fault classes, namely the activated value of the weighted sum among feature values. The *softmax* activation function is represented as:

$$p = \text{softmax}(w^k x^{k-1} + b^k) \qquad (4)$$

where p is the probability value associated with fault classes.

III. DEEP TRANSFER DIAGNOSTIC FRAMEWORK DESIGN

The CNN constitutes a hierarchical feature representation system through the weights and biases of a large number of neurons, thereby obtaining the powerful feature extraction function. By training the network, fault features and category information are effectively compiled into model parameters. The CNN structure can usually be regarded as two structures: the feature extraction structure (convolution block) and the classification structure (classifier). The convolutional layers and pooling layers complete feature learning and feature representation in source domains and compile the learned features into weight parameters (red dashed region in Fig. 2). Additionally, the fully connection layers and classification layer classify the extracted features, and the operation process is treated as a classifier (blue dashed region in Fig. 2). In the transfer framework, the trained parametric model in source tasks has certain generalization for the target task, thus transferring the trained models as the initial model in target task. Whereafter, the parameters within the model are recompiled using few samples similar to the target task, thus completing the feature learning in the target task. The fault diagnosis framework of bearings based on deep feature transferability is shown in Fig. 2.

After a series of feature extraction in the forward propagation process, the feature information learned from many samples is transformed into the probability distribution across the relevant label. To optimize the weight
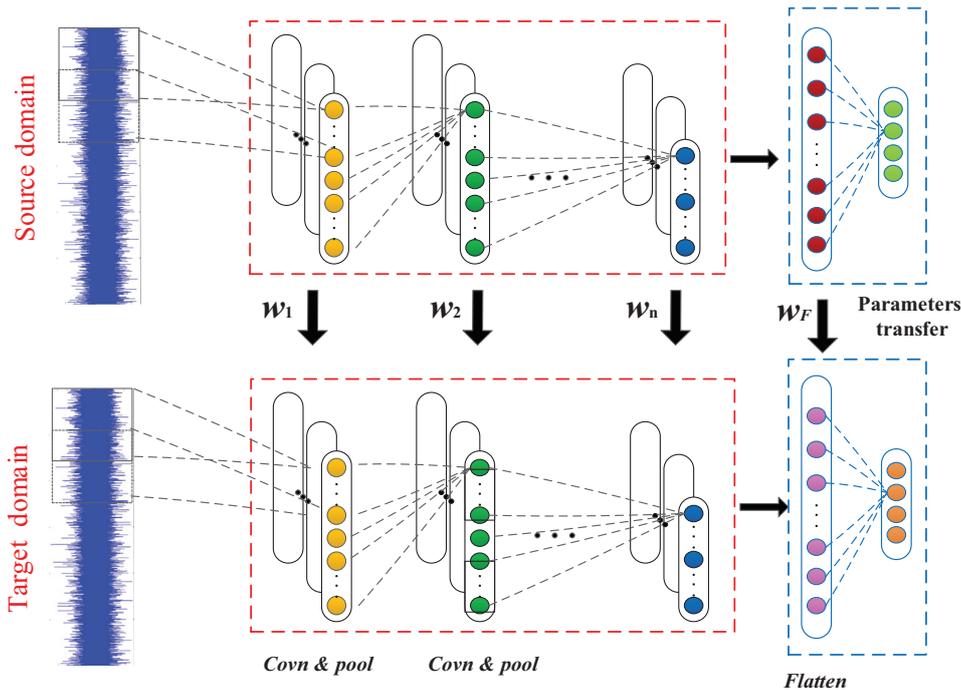

Fig. 2. Schematic diagram of deep transfer framework.

parameters, the error between the true and predicted labels is measured through the cross-entropy loss function during the backpropagation:

$$L = -\sum_{i=1}^{N} y^{(i)} \log \widehat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \widehat{y}^{(i)}) \quad (5)$$

where L denotes the cross-entropy loss; y and $\widehat{y}$ indicate the true and predicted label of category; and N is the total sample category.

In each category of the source domain, there are enough samples to train the basic network. The inherent characteristics of each fault are compiled into the weight parameters of the network. For a specified network structure, to obtain the optimal network model from large-scale source samples $\mathcal{D}_S = \{x_i^S, y_i^S\}^n$, the optimization problem in the training process is expressed as:

$$arg \min_{w} \sum_{i} L(y_i^s, f(x_i^s, w)) \quad (6)$$

$$w = \{w_l\}_{l=1}^{n} \quad (7)$$

where y_i^s and $f(x_i^s, w)$ denote the true and predicted labels of category in source domain, respectively; w denotes the parameter collection.

The task of Eq. (6) is to minimize the loss between the predicted and true values of the fault category. Adaptive moment estimation (Adam) algorithm dynamically adjusts the adaptive learning rate of each parameter with moment estimation of the gradient, thereby optimizing the model parameters. The optimization process is summarized as:

$$g_t = \nabla_w G_t(w) \quad (8)$$

$$\alpha_t = \frac{\alpha \cdot \sqrt{1 - \beta_2^t}}{1 - \beta_1^t} \quad (9)$$

$$w_t \leftarrow w_{t-1} - \alpha_t \cdot m_t / (\sqrt{v_t} + \widehat{\varepsilon}) \quad (10)$$

where t denotes the number of update steps; $G_t(w)$ denotes the gradient with respect to w; α denotes the learning rate; β denotes the moment attenuation coefficients; m_t and v_t denote the moments of the gradient g_t, respectively; and $\widehat{\varepsilon}$ is a hyper-parameter.

For the cross-condition diagnosis task of bearings, the pre-trained model of source tasks is used to initialize the network parameters for target tasks. To make the pre-trained model suitable for new target tasks, some samples $\{x_j^T, y_j^T\}^l$ similar to the target task are chosen to further recompile the partial weight parameters of networks, and the optimization process can be expressed as:

$$arg \min_{w'} \sum_{i} L(y_i^t, f(x_i^t, w')) \quad (11)$$

where w' represents the parameter collection of the target domain.

Due to the lack of fault samples, the trained model of source tasks is loaded directly into the unknown tasks, which increases the risk of over-fitting on the target task. Assuming that fine-tuning samples of the target task are far less than samples to be tested, over-fitting may also appear during the fine-tuning process. To avoid this issue, it is vital to improve the generalization of pre-trained models for target diagnosis tasks. The feature transferability of bearings in different diagnostic tasks is investigated through three strategies with different transfer ways. The transfer processes are shown in Fig. 3.

We propose three transfer strategies to further discuss the advantages and disadvantages of this method, which are described in detail as follows:

Strategy 1: The source model is directly transferred to target tasks. The parameters of convolution blocks are frozen, and the parameters of the classifier are
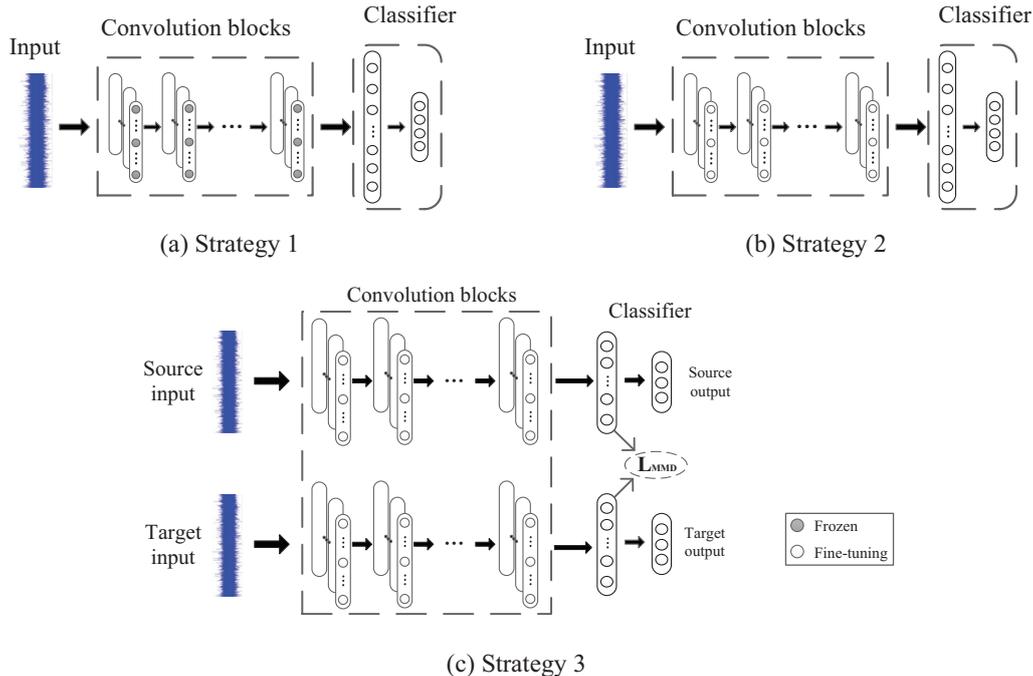

(a) Strategy 1

(b) Strategy 2

(c) Strategy 3

Fig. 3. The transfer process of three strategies.

recompiled according to the target task, thus reducing the number of training parameters.

Strategy 2: Unlike strategy 1, the trained model of ResNet is transferred to the target task as initial parameters of the network, and few samples similar to the target task are utilized to recompile the whole parameters to suit the diagnostic task.

Strategy 3: This strategy utilizes the dual-stream structure to train an effective model for the target task by measuring the distance between feature distributions in different diagnostic tasks. We utilize the standard metric [27,28] (MMD) to minimize the feature error among samples in different tasks. The loss function is defined as:

$$L = L_C(X_L, y) + \lambda MMD^2(X_S, X_T) \qquad (12)$$

where L_C is the loss of the source task; λ is a hyperparameter; and MMD denotes the distribution distance among different samples. The expression of MMD is as follows:

$$MMD(X_S, X_T) = \|E_{x_s \in X_S}[\phi(x_s)] - E_{x_t \in X_T}[\phi(x_t)]\| \qquad (13)$$

where $\phi(\cdot)$ represents the data point operation.

The main steps of the transfer process are shown in Algorithm 1.

IV. EXPERIMENTAL VALIDATION

A. EXPERIMENT SETUP

Planetary gearboxes are the most common variable speed mechanism used in wind turbines. Therefore, as shown in Fig. 4, we construct a test rig of one-staged planetary gearbox to simulate the scenario in wind turbines, where the planetary gearbox is used as a speed increaser. In Fig. 4, an induction motor (driving motor) drives the planetary gearbox through a coupling, and the sun gear shaft of the gearbox is linked to an AC generator, thus dissipating the generated energy through a resistor bank within the generator. Also, an encoder and a torque sensor are successively connected between the gearbox and the generator. The detailed parameters of the planetary gearbox are given in Tables I and II. Three sets of planet bearing faults, that is, inner race fault (IRF), outer race fault (ORF), and rolling element fault, are set up in the experiment. These fault types are all localized damage and are displayed in Fig. 5.

B. DATASETS

The first diagnostic dataset comprises of bearing fault data from Case Western Reserve University (CWRU) [29]. This bearing dataset is recognized as a world-wide standard dataset. The bearing signal acquisition platform is shown in Fig. 6. This platform is made up of a drive motor, a torque transducer, a dynamometer, and the corresponding control electrical device. We choose four health statuses in the

---

Algorithm 1: Transfer process

Input: A source task with labels $\mathcal{D}_S = \{x_i^S, y_i^S\}^n$, and a target domain without labels $\mathcal{D}_T = \{x_j^T\}^m$.

Output: The network model after fine-tuning.

  begin

 1: Initialize the model parameters of network.

 2: Set up the adaptation layers of network.

 3: Utilize the source data to obtain pre-trained models with different generalization abilities.

 If strategy 1 or 2:

  4: Fine-tune the parameters of the pre-trained model with a few samples to fit the target task.

 If strategy 3:

  4: Measure the distance between source and target samples by MMD on the adaptation layers.

 5: Calculate the loss between the predicted and the true labels.

 6: Use the calculated loss to optimize the network parameters.

 7: **Until** the loss converges in the target task.

Fig. 4. Test rig of one-staged planetary gearbox.

Table I.   Number of gear teeth

Sun	Planet	Ring
36	35 (3)	108

Note: The number of planet gears in parenthesis.

Table II.   Configuration parameters of planet bearings

Diameter of rollers (mm)	Diameter of pitch circle (mm)	Number of rollers	Contact angle (°)
3.5	19.5	10	0

bearing fault dataset for experimental analysis, including healthy operation conditions (normal condition (NC), IRF, ORF, and ball fault). Drive end bearing faults are used as targets for analysis, where the sampling rate is 12000 Hz. And we also select different fault diameters for transfer diagnosis, that is, 0.007 inches and 0.014 inches. The dataset settings are shown in Table III. Specifically, data from the load zone centered at 6:00 direction are analyzed in the ORF. In addition, the four motor loads (0HP, 1HP, 2HP, 3HP) correspond to motor speeds of 1797 rpm (M_1), 1772 rpm (M_2), 1750 rpm (M_3), and 1730 rpm (M_4) at the fault diameter of 0.007 inches. Similarly, the motor speeds corresponding to the motor loads are noted as N_1 (1797 rpm), N_2 (1772 rpm), N_3 (1750 rpm), and N_4 (1730 rpm) in the condition of 0.014 inches fault diameter. According to the speeds and fault diameters, the transfer tasks of this dataset are shown in Table IV.

The second diagnostic dataset consists of multiple planet bearing vibration signals. The experiments are carried out under different bearing fault conditions, including NC when all planet bearings are healthy, IRF where only one planet bearing has faulty inner race, ORF where only one planet bearing has faulty outer race, and rolling element fault (REF) where only one planet bearing has one faulty rolling element. In the constant working condition, each health condition is respectively experimented at different motor speeds: 360 rpm (B_1), 480 rpm (B_2), and 600 rpm (B_3). Depending on the internal structure of the planetary gearbox, we collect vibration signals from the top surface of the gearbox shell using accelerometers, thus minimizing energy loss, as seen at acquisition point in Fig. 4. To obtain enough fault information, we set the sampling rate to 20480 Hz during the signal acquisition. The vibration signals are segmented using overlapping slices to obtain

the corresponding sample datasets. In particular, to obtain the high-quality signal dataset with uniform data distribution, the signal distributed in the 15s~45 s period (the overall time period of 60 s) is divided into sample dataset. Each sub-sample contains 10240 discrete points, which reflects the characteristic information within signals over 0.5-second time span. Accordingly, the transfer diagnosis tasks of planet bearing faults are listed as shown in Table V.

C. EXPERIMENTAL ANALYSIS SYSTEM

All experiments adopted end-to-end diagnosis procedure on deep diagnostic framework to enhance the diagnostic adaptability of this method. Increasing the network depth by simply stacking convolutional layers can drastically degrade the diagnostic performance of the network, while causing vanishing gradient during the training process. Compared with other deep networks, the ResNets are able to utilize multiple residual blocks to maintain good feature extraction performance. To avoid the risk of over-fitting and gradient disappearance or explosion, ResNet18 is used to learn feature information within signals. The batch normalization is able to dramatically increase the training speed of networks, making the training loss more quickly converge. Especially for strategies 1 and 2, we selected 15% of target data as fine-tuned samples to recompile the pre-trained model weights from the source task. In strategy 3, we extract feature information in both the source and target tasks and measure the distribution differences between the features using MMD. In addition, we adopt the correlation alignment (CORAL) algorithm and the adversarial algorithm (DANN) as comparison methods for comparing MMD in strategy 3
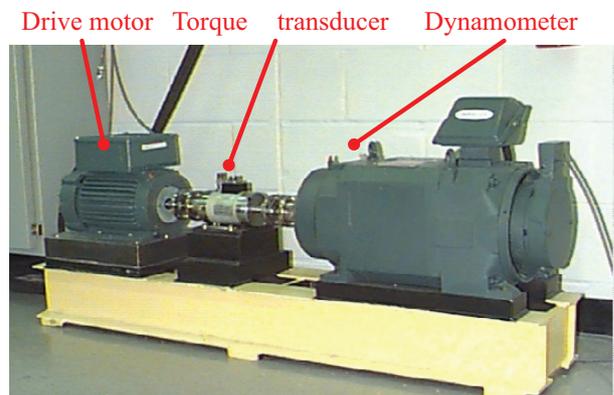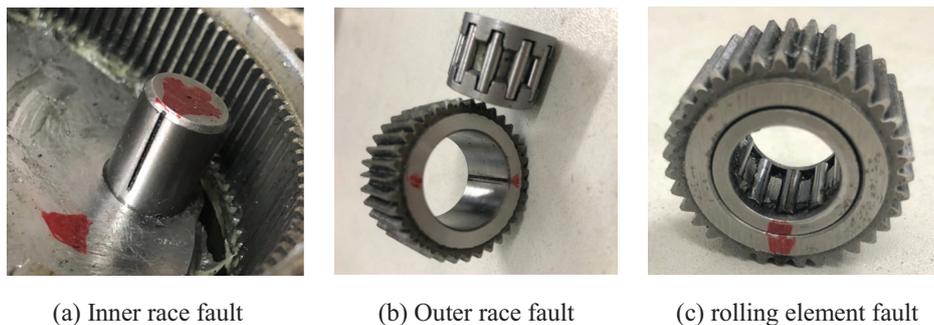

Fig. 6.   Signal acquisition platform.

(a) Inner race fault   (b) Outer race fault   (c) rolling element fault

Fig. 5.   The damaged planet bearings.

Table III.   Dataset settings

Fault diameter	Motor load (HP)	Approx. motor speed (rpm)	Inner race	Ball	Outer race (centered @6:00)
0.007"	0	1797	IR007_0	B007_0	OR007@6_0
	1	1772	IR007_1	B007_1	OR007@6_1
	2	1750	IR007_2	B007_2	OR007@6_2
	3	1730	IR007_3	B007_3	OR007@6_3
0.014"	0	1797	IR014_0	B014_0	OR014@6_0
	1	1772	IR014_1	B014_1	OR014@6_1
	2	1750	IR014_2	B014_2	OR014@6_2
	3	1730	IR014_3	B014_3	OR014@6_3

Table IV.   Transfer tasks of motor bearings

	Transfer task	Motor speed (rpm)		Health status
		Source task	Target task	
Speed transfer	M_1–M_2	1797	1772	
	M_2–M_3	1772	1750	Normal condition (NC)
	M_3–M_4	1750	1730	Inner race fault (IRF)
Fault transfer	M_1–N_1	1797	1797	Outer race fault (ORF)
	M_2–N_2	1772	1772	Ball fault (BF)
	M_3–N_3	1750	1750	

Table V.   Transfer tasks of planet bearings

Transfer task	Motor speed (rpm)		Health status
	Source task	Target task	
B_1–B_2	360	480	
B_1–B_3	360	600	Normal condition (NC)
B_2–B_1	480	360	Inner race fault (IRF)
B_2–B_3	480	600	Outer race fault (ORF)
B_3–B_1	600	360	Rolling element fault (REF)
B_3–B_2	600	480	

framework, respectively. All experiments are implemented in the Pytorch1.3 framework. We performed five experimental analyses with random initial parameters for each task and calculated the average of results to assess the diagnostic performance, thus ensuring the reliability of the experiment.

D. RESULTS

The generalization ability of pre-trained models in the small sample set is essential for different diagnostic tasks. In general, improving the generalizability of networks to target tasks with small sample set is determined by two aspects: fault category and sample size. Therefore, the diagnostic cases for motor bearings and planet bearings are separately constructed in the proposed transfer framework.

1) RESULTS OF CWRU BEARING DATASETS.   Table VI lists the detailed diagnostic results for six transfer cases. It can be seen that strategy 1 has the worst classification result, as fine-tuning the classifier parameters by similar data cannot bridge the feature distribution differences among different tasks very well. Compared with the results (i.e., 73.2%, 77.8%) of the traditional method, the pre-trained model of source dataset (i.e., 72.3%, 74.9%) cannot effectively learn features of target tasks, which leads to negative transfer. The main method of strategy 2 is to recompile all parameters of the trained model through similar samples, and the classification results have been greatly improved compared with strategy 1. However, when the pre-trained model is fine-tuned between different conditions, the classification results (94.6% and 87.3%) showed large differences. Compared with strategy 1 and strategy 2, strategy 3 (with average result of 98.4%) can utilize the similarity of data to effectively minimize the data differences between the source and target domains, resulting in better classification results. As can be found in

Table VI.   Diagnosis results of different strategies in motor bearing dataset

	ResNet18	Strategy 1	Strategy 2	Strategy 3	CORAL	DANN
M_1–M_2	72.2	76.7	87.3	98.6	98.9	98.3
M_2–M_3	73.2	72.3	88.3	99.9	99.3	99.3
M_3–M_4	77.8	74.9	91.8	98.4	98.6	97.5
M_1–N_1	39.1	35.8	92.7	97.3	92.8	93.3
M_2–N_2	37.7	31.0	94.6	99.2	91.3	89.6
M_3–N_3	40.1	38.2	90.9	99.7	93.4	90.4
Average	**56.7**	**54.8**	**90.9**	**98.9**	**95.7**	**94.7**

Table VI, the diagnostic performance of strategy 3 is also better than other feature assessment methods (CORAL (95.7%), DANN (94.7%)). Meanwhile, comparing these results reveals that evaluating the data distribution between different tasks during the training process can effectively compensate for the feature differences in the tasks.

We compare the predicted labels of target samples with the true labels to present the classification results more clearly. The transfer task M_2-M_3 is selected to calculate the confusion matrix corresponding to the optimal accuracy of each strategy. The results are shown in Fig. 7; it can be found that different transfer strategies exhibit obvious discrepancy under the same condition. There are a large number of sample misclassifications in ResNet18, strategy 1, and strategy 2, whereas strategy 3 can accurately classify samples of different health states. Fig. 8 shows the training process of different strategies. The loss of strategy 3 reaches convergence quickly and is lower than the other strategies. The diagnostic accuracy is higher and more stable than other methods. These results illustrate that strategy 3 has better diagnostic performance and can effectively augment the generalization of models by minimizing the feature differences among different tasks.

2) RESULTS OF PLANET BEARING FAULT DIAGNOSIS.
The diagnosis results of planet bearing faults are shown in Table VII. It can be found that the classification result of strategy 2 is higher than that of traditional method and strategy 1, but the diagnostic performance of strategy 2 (83.7%, 65.9%, 84.6%) is unsatisfactory in several transfer tasks. Compared with strategy 2, the transfer ability of strategy 3 has been significantly improved in various

transfer tasks. For example, the classification result of strategy 3 (98.7%) is superior to other methods (84.0%, 73.3%, 94.9%) in the transfer task B_2-B_1. Fig. 9 displays the training process of different methods in the task B_3-B_1. It can be clearly seen that the test losses of other methods clearly exhibit instability in comparison to strategy 3. From Fig. 9(b), the diagnostic performance of the traditional method and strategy 1 appears to fluctuate significantly in the cross-condition diagnostic task, indicating that the feature information learned from source tasks cannot effectively complete the diagnosis of the target task, while strategy 3 exhibits the stable training process by evaluating the feature differences between different diagnostic tasks.

As can be seen in Table VII, strategy 3 is able to effectively bridge the discrepancy of data distribution

Table VII. Diagnosis results of different strategies in planet bearing dataset

	ResNet18	Strategy 1	Strategy 2	Strategy 3
B_1−B_2	57.3	70.4	83.7	98.4
B_1−B_3	48.8	53.6	65.9	97.0
B_2−B_1	84.0	73.3	94.9	98.7
B_2−B_3	87.5	58.7	97.9	98.2
B_3−B_1	66.6	59.7	84.6	99.7
B_3−B_2	74.9	65.7	94.9	98.1
Average	**69.9**	**63.6**	**86.9**	**98.4**

(a) ResNet18   (b) strategy 1   (c) strategy 2   (d) strategy 3
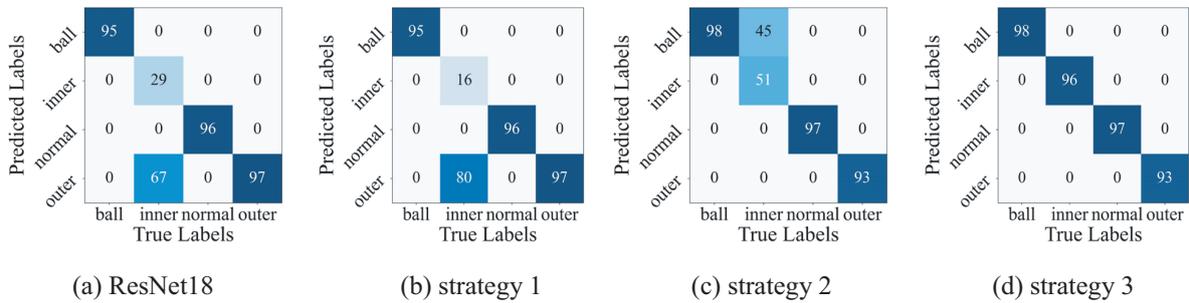
Fig. 7. Confusion matrixes of different methods in the task M_2–M_3.
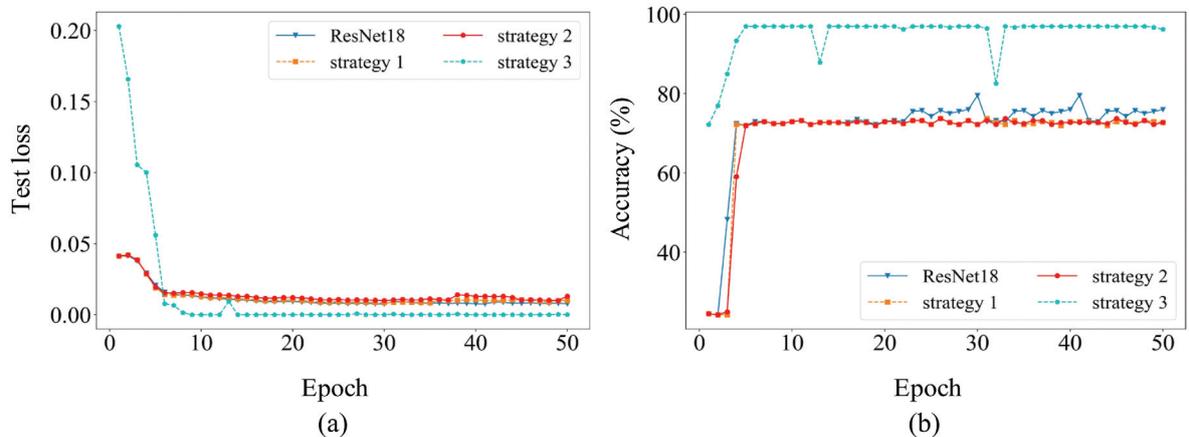

(a)   (b)

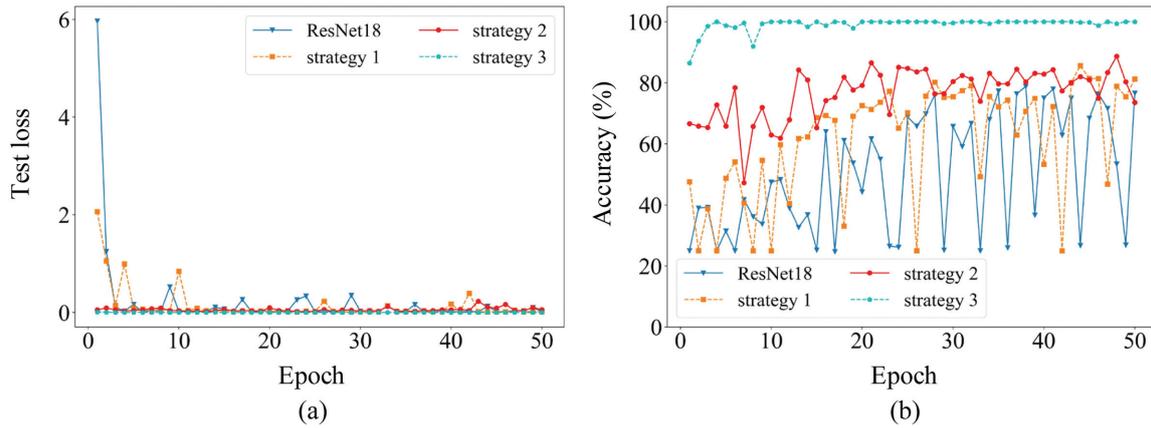**Fig. 8.** The training process of different methods in the task M_2–M_3.

Fig. 9. The training process of different methods in the task B_3–B_1.

Table VIII.   Effect of learning rate on diagnostic results

		B_1–B_2	B_2–B_3	B_3–B_1	Average
$\beta = 0.1$	Strategy 3	53.2	63.5	38.2	**51.6**
	CORAL	50.6	56.2	45.6	**50.8**
	DANN	46.5	52.3	35.2	**44.7**
$\beta = 0.01$	Strategy 3	86.3	76.9	65.4	**76.2**
	CORAL	78.4	75.1	59.8	**71.1**
	DANN	70.6	65.3	65.7	**67.2**
$\beta = 0.001$	Strategy 3	98.4	98.2	99.7	**98.8**
	CORAL	81.7	96.7	49.9	**76.1**
	DANN	86.5	96.3	34.3	**72.4**

between different tasks in the cross-condition diagnosis task compared to strategies 1 and 2. Specifically, the learning rate of the optimizer in the dual-stream adaptation structure has a significant impact on the update of weight parameters, so we complete ablation experiments to explore the impact of the learning rate. As shown in Table VIII, three learning rates are set for comparison in the three transfer tasks. It is clearly seen that the learning rate of 0.001 optimizes the network weight parameters better. However, it can be seen from the diagnostic results of different transfer tasks that strategy 3 using MMD can effectively measure the feature differences between the source and target tasks, making the diagnostic performance of network stable. For example, the accuracy of DANN is 34.3% in the task B_3-B_1 and 96.3% in the task B_2-B_3, which illustrates the poor generalization of adversarial training to different data distributions. These results indicate that the dual-stream adaptation structure can effectively evaluate the data distribution between the source and target domains in the cross-condition diagnosis task of rolling bearings, where evaluating data differences using MMD can be widely applied in different transfer scenarios.

V. CONCLUSIONS

Rolling bearings are a common component in wind turbines, and bearing faults may cause low transmission efficiency of wind turbines. In real-world engineering applications, the lack of fault data, the feature differences among different conditions, and the relatively weak fault information caused by resonance modulation cause difficulty in extracting fault information from bearing signals. To address the above issues, this paper proposes a deep residual joint transfer strategy method to implement cross-condition fault diagnosis of bearings. This method explores the feature transferability between different tasks by means of model transfer and recompiles the model weights to fit the data distribution of target tasks by different transfer strategies. Moreover, a dual-stream feature adaptation transfer strategy is constructed through the feature adaptation layer, thus improving the generalization of models to target tasks by gauging the data differences between different tasks. Compared with traditional methods, this method is able to mine more fine-grained feature information for network training, rather than using lots of labeled samples to train the network from scratch. Experimental results demonstrate the applicability of this method, thereby effectively bridging the gap between the source and target diagnostic tasks. The deep residual joint transfer strategy method successfully diagnoses bearing health conditions in different bearing fault datasets and provides an effective method for bearing fault diagnosis.

CONFLICTS OF INTEREST

The author(s) declared no potential conflicts of interest concerning the research, authorship, and/or publication of this article.

CRediT AUTHORSHIP CONTRIBUTION STATEMENT

Zhipeng Feng and Songjun Han designed and defined the experimental protocol. Songjun Han performed the experiments and analyzed the data. Zhipeng Feng provided guidance and advice for research, and Songjun Han participated in the content and writing of the manuscript. All authors have read and approved the final manuscript.

DATA AVAILABILITY

The raw/processed data required to reproduce these findings cannot be shared at this time as the data also form part of an ongoing study.

CODE AVAILABILITY

The algorithm involved in this paper is still being studied by the research group, so it is not publicly disclosed.

References

[1] C. Zhao and Z. P. Feng, "Application of multi-domain sparse features for fault identification of planetary gearbox," *Measurement*, vol. 104, pp. 169–179, 2017.

[2] Y. G. Lei, F. Jia, D. T. Kong, J. Lin, and S. B. Xing, "Opportunities and challenges of machinery intelligent fault diagnosis in big data era," *J. Mech. Eng.*, vol. 54, no. 5, pp. 94–104, 2018.

[3] Z. P. Feng and M. Liang, "Fault diagnosis of wind turbine planetary gearbox under nonstationary conditions via adaptive optimal kernel time–frequency analysis," *Renew. Energy*, vol. 66, pp. 468–477, 2014.

[4] Z. P. Feng and M. J. Zuo, "Vibration signal models for fault diagnosis of planetary gearboxes," *J. Sound Vib.*, vol. 331, no. 22, pp. 4919–4939, 2012.

[5] Z. Feng and M. J. Zuo, "Fault diagnosis of planetary gearboxes via torsional vibration signal analysis," *Mech. Syst. Signal Process.*, vol. 36, no. 2, pp. 401–421, 2013.

[6] S. Jain and H. Hunt, "Vibration response of a wind-turbine planetary gear set in the presence of a localized planet bearing defect," *Proc. ASME Int. Mech. Eng. Cong. Exposit.*, vol. 7, pp. 943–952, 2011.

[7] P. Henriquez, J. B. Alonso, M. A. Ferrer, and C. M. Travieso, "Review of automatic fault diagnosis systems using audio and vibration signals," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 44, no. 5, pp. 642–652, 2014.

[8] A. Widodo, B. S. Yang, D. S. Gu, and B. K. Choi, "Intelligent fault diagnosis system of induction motor based on transient current signal," *Mechatronics*, vol. 19, pp. 680–689, 2009.

[9] T. Han, C. Liu, W. Yang, W. G. Yang, and D. X. Jiang, "Learning transferable features in deep convolutional neural networks for diagnosing unseen machine conditions," *ISA Trans.*, vol. 93, pp. 341–353, 2019.

[10] Y. G. Lei, *Intelligent Fault Diagnosis and Remaining Useful Life Prediction of Rotation Machinery*. Oxford: Elsevier Butterworth-Heinemann, 2016.

[11] R. N. Liu, G. T. Meng, B. Y. Yang, C. Sun, and X. F. Chen, "Dislocated time series convolutional neural architecture: an intelligent fault diagnosis approach for electric machine," *IEEE Trans. Ind. Inf.*, vol. 13, no. 3, pp. 1310–1320, 2017.

[12] T. Han, C. Liu, W. G. Yang, and D. X. Jiang, "A novel adversarial learning framework in deep convolutional neural network for intelligent diagnosis of mechanical faults," *Knowl-Based Syst.*, vol. 165, pp. 474–487, 2019.

[13] Y. H. Chang, J. L. Chen, C. Qu, and T. Y. Pan, "Intelligent fault diagnosis of wind turbines via a deep learning network using parallel convolution layers with multi-scale kernels," *Renew. Energy*, vol. 153, pp. 205–213, 2020.

[14] J. Y. Jiao, M. Zhao, J. Lin, and J. Zhao, "A multivariate encoder information based convolutional neural network for intelligent fault diagnosis of planetary gearboxes," *Knowl.-Based Syst.*, vol. 163, pp. 237–250, 2018.

[15] Y. Xu, Z. X. Li, S. Q. Wang, W. H. Li, T. Sarkodie-Gyan, and S. Z. Feng, "A hybrid deep-learning model for fault diagnosis of rolling bearings," *Measurement*, vol. 169, p. 108502, 2021.

[16] R. M. Souza, E. G. S. Nascimento, U. A. Miranda, W. J. D. Silva, and H. A. Lepikson, "Deep learning for diagnosis and classification of faults in industrial rotating machinery," *Comput. Ind. Eng.*, vol. 153, p. 107060, 2020.

[17] D. Verstraete, A. Ferrada, E. Lopez Droguett, V. Meruane, and M. Modarres, "Deep learning enabled fault diagnosis using time-frequency image analysis of rolling element bearings," *Shock Vib.*, vol. 6, pp. 1–17, 2017.

[18] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2: pp. 199–210, 2011.

[19] K. Weiss, T. M. Khoshgoftaar, and D. D. Wang, "A survey of transfer learning," *J. Big Data*, vol. 3, p. 9, 2016.

[20] L. Wen, X. Li, and L. Gao, "A transfer convolutional neural network for fault diagnosis based on ResNet-50," *Neural Comput. Appl.*, vol. 32, no. 10, pp. 6111–6124, 2020.

[21] X. Wang, C. Q. Shen, M. Xia, D. Wang, and Z. K. Zhu, "Multi-scale deep intra-class transfer learning for bearing fault diagnosis," *Reliab. Eng. Syst. Saf.*, vol. 202, p. 107050, 2020.

[22] Z. Y. He, H. D. Shao, L. Jing, J. S. Cheng, and Y. Yang, "Transfer fault diagnosis of bearing installed in different machines using enhanced deep auto-encoder," *Measurement*, vol. 152, p. 107393, 2020.

[23] M. Azamfar, J. Singh, X. Li, and J. Lee, "Cross-domain gearbox diagnostics under variable working conditions with deep convolutional transfer learning," *J. Vib. Control*, vol. 27, no. 7–8, pp. 854–864, 2021.

[24] B. C. Sun and K. Saenko, "Deep CORAL: correlation alignment for deep domain adaptation," *Comput. Vision – Eccv 2016 Workshops*, vol. 9915, pp. 443–450, 2016.

[25] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN. Machine learning," arXiv:1701.07875.

[26] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conf. Comput. Vision Pattern Recogn.*, 2016, pp. 770–778.

[27] K. M. Borgwardt et al., "Integrating structured biological data by Kernel Maximum Mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. 49–57, 2006.

[28] E. Tzeng et al., "Deep domain confusion: maximizing for domain invariance," *Comput. Sci.*, 2014. https://doi.org/10.48550/arXiv.1412.3474

[29] Case Western Reserve University Bearing Data Center Website, https://engineering.case.edu/bearingdatacenter/download-data-file.