Hindawi
Journal of Sensors
Volume 2021, Article ID 5714240, 13 pages
https://doi.org/10.1155/2021/5714240

Research Article

Bearing Fault Diagnosis under Variable Working Conditions Based on Deep Residual Shrinkage Networks and Transfer Learning

Xinyu Yang, Fulin Chi, Siyu Shao, and Qiang Zhang

Air and Missile Defense College, Air Force Engineering University, 710000 Xi'an, China

Correspondence should be addressed to Siyu Shao; cathygx.sy@gmail.com

Received 8 October 2021; Accepted 15 November 2021; Published 6 December 2021

Academic Editor: Haidong Shao

Nowadays, deep learning has made great achievements in the field of rotating machinery fault diagnosis. But in the practical engineering scenarios, when facing a large number of unlabeled data and variable operating conditions, only using a deep learning algorithm may reduce the performance. In order to solve the above problem, this paper uses a method of combining transfer learning with deep learning. First, the deep shrinkage residual network is constructed by adding soft thresholds to extract the characteristics of bearing vibration data under noise redundancy. Then, the joint maximum mean deviation (JMMD) criterion and conditional domain adversarial (CDA) learning domain adapting network are used to align the source and target domains. At the same time, adding transferable semantic augmentation (TSA) regular items improves alignment performance between classes. Finally, the proposed model is verified by three experiments: variable load, variable speed, and variable noise, which overcomes the shortcomings of traditional deep learning and shallow transfer learning algorithms.

1. Introduction

With the development of modern industry toward intelligence, the health management mode of industrial equipment based on big data has become a hot research field. To achieve the goal of real-time monitoring of mechanical health and performance, it is increasingly important to speed up the establishment of a stable and reliable Prognostic and Health Management (PHM) [1]. In an industrial system, all working elements are in a relatively coupled working state, and any failure may affect the normal operation of the whole mechanical system. Since the measured signals are usually transient and dynamic, it is difficult to achieve early diagnosis of monitoring and failure by using the traditional time-frequency analysis method [2]. In order to ensure the highest possible uptime, the way of system maintenance should change to the way of real-time monitoring and predictive prevention [3]. To achieve these purposes, the intelligent fault diagnosis method has become an important research field in recent years.

The intelligent fault diagnosis method is developed on the basis of traditional machine learning and deep learning. Different from the traditional method of extracting fault feature signals manually, the intelligent fault diagnosis method does not require much prior knowledge about signal processing but directly extracts useful information from the vibration data collected and realizes early fault diagnosis in a data-driven way [4]. Among them, artificial neural network (ANN), support vector machine (SVM), deep neural network (DNN), and other models are the most widely used models for intelligent fault diagnosis [5, 6]. Merainani et al. [7] used a self-organizing feature map (SOM) neural network to identify and classify gearbox faults automatically and used a self-organizing and adaptive algorithm to identify gearbox early faults effectively. Lu et al. [8] used the stacked denoising autoencoder (SDA) for greedy layer-wise training and achieved higher accuracy than ANN and SVM in diagnosing signals containing ambient noise and fluctuating working conditions. In the development process of intelligent fault diagnosis, algorithms and data are always the two most important cores. As the complexity of the system and the volume of data acquired increase, the cost of labeling data increases. When facing a large number of unlabeled data, it is difficult to guarantee an ideal accuracy simply by relying on the general deep learning network. At the same time, it is impossible for monitoring data to maintain the same spatial distribution throughout the survey period, considering the actual engineering conditions. The joint distribution of data changes with the change of mechanical speed, load, and noise. Therefore, in practical applications, the generalization performance based on intelligent fault diagnosis may be reduced.

Therefore, transfer learning as a new fault diagnosis tool solves the above problems well.Moreover, the theory of transfer learning has been continuously supplemented and perfected and has proved its applicability in various fields [9, 10]. The focus of transfer learning is how to solve new problems according to the knowledge that has been learned and reuse the learned knowledge through the similarity of the intrinsic characteristics of things [11]. Deep learning is superior in extracting the high-dimensional abstract features of data. It can map two groups of data with different distributions (source domain and target domain) into the same space. At this time, it can reduce the difference between features by transfer learning, which can not only accurately classify the source domain data but also achieve the purpose of domain adaptation. For some tasks with little difference in distribution, better results can be achieved in the target domain only by transferring the parameters of the pretrained network to the untrained network. However, in practice, the source and target domains have different feature spaces, but they can be aligned by minimizing the measurement differences between domains. Indicators for measuring the distribution differences between domains include KL divergence, maximum mean discrepancies (MMD), Wasserstein distance, and CORAL loss [12]. These indexes are added to the loss function, and then, the adaptive purpose is achieved through gradient descent. However, this shallow adaptive layer is still inadequate because it can only achieve the effect of domain adaptation globally, while overlapping confusion can occur in some domains with smaller discrimination. Deep transfer learning (DTL) based on deep network inherits the ability of the deep neural network to extract strong signal features. On the other hand, it overcomes the shortcomings of robustness and generalization of shallow transfer learning. Zheng et al. [13] summarized DTL into the following five methods: instance reweighting approach, feature transfer approach, classifier adaptation approach, deep learning-based approach, and adversarial-based approach. Han et al. [14] used the data of known working conditions to pretrain the CNN and realized the fault diagnosis under unknown working conditions based on CNN by fine-tuning the weight parameters. An et al. [15] used a multicore MMD domain adaptive framework to make the features of different domains approach each other in the reproducing kernel Hilbert space, which improved the stability and accuracy of the results. Wen et al. [16] used ResNet-50 combined with transfer learning to extract the characteristics of time domain fault signals converted to RGB images and had achieved the most advanced results on the test dataset. These studies show the validity of deep transfer learning in the diagnosis of mechanical variable conditions, but there are still some problems that need further study: (1) Most transfer learning methods do not take into account the joint distribution between the classifier output labels and the input data but only the marginal distribution of the data. (2) The effect of the nonlinear feature extraction capability of the deep learning framework on domain adaptation was not discussed in the process of innovation of the transfer learning algorithm.

In view of the problems above, this paper conducts related research through the following ideas.

(1) Two modules are constructed to achieve domain adaptation for the source domain and target domain. On the one hand, using the joint distribution differences of input features and output labels, domain adaptations are made in feature extraction and classification layers by JMMD. On the other hand, cross-entropy was used between feature and prediction labels to conduct domain adversarial training to reduce domain drift. The two modules not only realize the maximum distinction between classes but also realize domain adaptation under multimode conditions. Meanwhile, transferable semantic augmentation (TSA) regular terms are added to the loss function to enhance the implicit characteristics of the source domain and improve the effect of domain adaptation

(2) A deep shrinkage residual network is constructed as the main network for feature extraction of one-dimensional vibration signals. By setting a soft threshold in the residual block, the noise in the original signal is suppressed, so that the fault characteristics can be better adapted in the mapping space, and the robustness of the whole algorithm framework is enhanced

(3) The datasets used in the experiment are the CWRU bearing dataset and the Canadian-Ottawa bearing dataset. The validity of the deep denoising domain adaptive network proposed in this paper is verified by the three scenarios of staged operation, continuous operation, and antinoise, and the most advanced results are obtained

2. Transfer Learning Method

2.1. Preliminaries. In transfer learning, "domain" and "task" are the two most important concepts. The domain and task are separately divided into the source domain, target domain, source task, and target task [11]. This paper intends to solve the problem of unsupervised domain adaptation where training data has labels and test data does not have labels. Formally, we denote $D_k^s = (x_i^{s,k}, y_i^{s,k})_{i=1}^{k}$ as labeled source domain data and $D^t = \{x_i^{t,k}\}_{i=1}^{k}$ as unlabeled target domain data, where x_i and y_i are, respectively, the *i*th sample's features and the ith sample's labels. Among them, superscript s denotes the source domain, t denotes target domain, and K denotes the number of domains. Here also denotes a domain as $D = \{\mathcal{X}, P(X)\}$, where $\mathcal{X}$ is the d -dimensional feature space of the souce domain and target domain, and $P(X)$ is the marginal probability distribution, $X = \{x_1, x_2, \cdots, x_n\} \subset \mathcal{X}$. For a domain $D = \{\mathcal{X}, P(X)\}$, $T = \{\mathcal{Y}, P(Y|X)\}$ is used to represent a task of domain adaptation. Among them, $\mathcal{Y}$ is the label space and $P(Y|X)$ is the marginal probability distribution, that is, the marginal distribution relationship between feature vector X and label space $\mathcal{Y}$ under the mapping of prediction function $f(\cdot)$. In the network training period, only the source domain has label space, but the target domain does not have label space, so the training of prediction function $f(\cdot)$ can only rely on the source domain data. In the case of transfer, $D_s \neq D_t$ or $T_s \neq T_t$ is often present. Therefore, training the prediction function $f(\cdot)$ only through the source domain data will lead to limited generalization ability of the model. In order to achieve domain adaptation, it is necessary to integrate the differences between target domain data features and source domain data features into network training, so that the target domain data in the test set can be correctly mapped to its corresponding label space in the case of $P_s(X_s) \neq P_t(X_t)$.

2.2. JMMD and CDA. Borgwardt et al. [17] first proposed the maximum mean discrepancies (MMD) method to measure the difference between the two distributions in the statistical sense. Given the data characteristic distributions X_s and X_t of the source domain and target domain, MMD can be defined as follows:

$$\mathrm{MMD}_{\mathscr{H}}(X_s, X_t) = \left\| \frac{1}{n_S} \sum_{i=1}^{n_S} \phi(x_i^s) - \frac{1}{n_T} \sum_{j=1}^{n_T} \phi\left(x_j^T\right) \right\|_{\mathscr{H}}^2, \quad (1)$$

where $\mathscr{H}$ represents the reproducing kernel Hilbert space (RKHS), and the data is mapped from high-dimensional feature space to low-dimensional space through kernel function ϕ. In practical application, the domain adaptation of data through MMD under complex multimodal conditions is very limited, and the kernel parameters are difficult to optimize. Gretton et al. [18] proposed a convex combination of multiple cores for effective mapping estimation to achieve depth domain adaptation. However, when Multikernel MMD (MK-MMD) is used for depth domain adaptation, the feature can only be transferred at the top layer by deepening the number of network layers, and the transfer of label distribution $P(Y_S)$ and $Q(Y_t)$ still stays at the classification layer. In order to fully consider the joint distribution of feature space and label space in the field, Long et al. [19] proposed the joint maximum mean deviation (JMMD) method, which is defined as

$$\mathscr{L}_{\mathrm{JMMD}}(P, Q) = \left\| \mathbb{E}_P\left(\otimes_{l=1}^{|L|} \phi^l\left(z^{sl}\right)\right) - \mathbb{E}_Q\left(\otimes_{l=1}^{|L|} \phi^l\left(z^{tl}\right)\right) \right\|_{\otimes_{l=1}^{|L|} \mathscr{H}'}^2, \quad (2)$$

Table 1: Parameters of the GRL.

Layers	Parameters
Fc1	out_features = 2048
Dropout1	$P = 0.5$
Fc2	out_features = 1024
Dropout2	$P = 0.5$
Fc3	out_features = 2

where Z^{sl} represents the output of the activation function of the lth level network, $\otimes_{l=1}^{|L|} \phi^l(x^l) = \phi^1(x^1) \otimes \cdots \otimes \phi^{|L|}(x^{|L|})$. Compared with formula (1), JMMD calculates the mapping of each layer of feature space in tensor product Hilbert space when measuring distance. The feature samples are mapped to a fixed diameter hypersphere through the activation function, so that the samples with similar features gather more closely in the feature space; that is, the distance between classes is expanded and the distance between classes is reduced, so as to balance the training difficulty between different distributed data [20]. In order to enhance domain adaptation, this paper adopts the idea of the domain adversarial neural network for reference and forms a depth adversarial domain adaptation network by adding the gradient reversal layer (GRL) after the feature extraction layer. The structural parameters of GRL are the same as those in literature [21], which are all three fully connected layers. The specific parameters are shown in Table 1.

Unlike the MMD method that takes the space metric distance, adversarial-based domain training follows the idea of a game in the generative adversarial network, so that the source domain and the target domain can be aligned in the network training. The domain adversarial network is generally divided into feature extraction layer G_f, classification layer G_c, and domain identification layer G_d, and the parameters of each layer are represented by θ_f, θ_c, and θ_d, respectively. GRL is also called the domain discriminator. Its function is to maximize the classification loss between source domain and target domain and confuse target domain data with source domain data. The classifier in the network realizes the accurate classification of data by minimizing the classification loss. Different from the generative adversarial network, the domain adversarial network does not need a generator. In order to carry out adversarial training, we multiply the error of the gradient inversion layer by a negative parameter $-\lambda$, so that the network training objectives before and after the GRL layer are opposite, achieving adversarial training [22]. At the end of the adversarial training, it shows that the loss of the domain discriminator has reached the maximum, so the domain discriminator has aligned the source domain and the target domain to the greatest extent. However, domain adversarial training still has the same defect as MMD. It only calculates the marginal distribution of $P(X)$ and $Q(X)$ and ignores its joint distribution. Similar to JMMD, in order to solve the problem $P(X_s, Y_s) \neq Q(X_t, Y_t)$, it is necessary to consider the joint distribution of the sign extraction layer and classification layer.

The multidimensional and multifeature data in the feature layer and classification layer of the domain adversarial network are matrix operated by means of mean mapping $x \otimes y$. In the GRL layer, the source field label is set to 0, and the target field label is set to 1. In order to prevent the loss function value of individual samples from tending to infinity due to nontransfer, so that the domain adversarial training cannot converge effectively, before the training, the entropy criterion is applied to the label prediction probability corresponding to the feature [21]. The loss function formed by the above form is a conditional domain adversarial (CDA) loss function, which is defined as

$$w(H(p)) = 1 + e^{-H(p)},\ H(p) = -\sum_{c=0}^{C-1} p_c \log p_c, \tag{3}$$

$$\begin{aligned} L_{\mathrm{CDA}}\left(\theta_f, \theta_d\right) = &-\frac{1}{n_s} w\left(H\left(p_i^s\right)\right) \sum_{i=1}^{n_s} \log\left[1 - D\left(F\left(x_i^s ; \theta_f\right) ; \theta_d\right)\right] \\ &- \frac{1}{n_t} w\left(H\left(p_i^t\right)\right) \sum_{i=1}^{n_t} \log\left[D\left(F\left(x_j^t ; \theta_f\right) ; \theta_d\right)\right]. \end{aligned} \tag{4}$$

2.3. Transferable Cross-Entropy Loss Learning. The methods of CDA and JMMD use the depth transferable features of the source domain and target domain to achieve domain adaptation. Therefore, superimposing the two in a transfer learning module can theoretically achieve the effect of complementary advantages, as JMMD not only reduces the difference of marginal distribution but also reduces the difference of joint distribution, and domain confrontation helps to reduce the phenomenon of domain drift in the process of domain adaptation. Transfer learning is widely applied in image recognition, and many transfer learning methods have achieved good results in the open dataset. However, compared with some image public datasets, the bearing fault signals also have the characteristics of high coupling, nonlinearity, and nonstationary. Therefore, in order to further enhance the domain adaptability of bearing fault diagnosis, we should make full use of the labeled samples in the source domain. As shown in Figure 1, Li et al. [23] proposed a transferable semantic augmentation (TSA) method to enhance the adaptability of the classifier by implicitly generating source features to target semantics.

For bearings of the same structure, the high-dimensional characteristics of vibration signals will migrate in all directions in the transfer learning, and the high-frequency characteristics induced by resonance will transfer along a certain direction of bearing frequency. Therefore, a quantitative measurement method is needed to highlight the direction that has the greatest impact on domain adaptation. Use μ_s and μ_t to represent the mean value of the feature space of the source domain and the target domain, respectively, and use $\Delta\mu^c = \mu_t^c - \mu_s^c$ to represent the difference of the mean value of a class c sample in the source domain and the target domain. The greater the difference, the greater the overall deviation, so the domain drift can be reduced by means of the value of $\Delta\mu^c$. However, the value of $\Delta\mu^c$ is relatively extensive. In order to more accurately measure the distribution difference, it is also necessary to calculate the covariance $\sum_t^c$ in the target class and measure the difference of all offset directions of the target domain relative to the source domain at the highest level of the network. Finally, the multivariate distribution difference $N(\Delta\mu^c, \sum_t^c)$ is composed of interdomain mean difference $\Delta\mu^c$ and intratarget covariance $\sum_t^c$. It should be noted that the TSA method focuses on using the characteristics of the source domain to approach the target domain as much as possible. Here, the characteristics of the highest layer of the network refer to the output matrix $\mathbf{f}_{si} \sim N(\boldsymbol{\Delta\mu}^{y_{si}}, \boldsymbol{\Sigma}_t^{y_{si}})$ of the last full connection layer, which will be reflected in the following formula. In network training, the loss of transfer after M iterations is given by using cross-entropy:

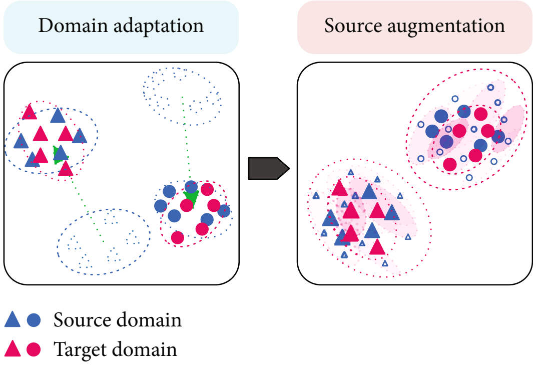


Figure 1: Illustration of TSA method.

$$\mathscr{L}_M = \frac{1}{n_s} \sum_{i=1}^{n_s} \mathbb{E}_{\tilde{\mathbf{f}}_{si}} \left[\log\left(\sum_{c=1}^{C} e^{\left(\mathbf{w}_c^\top - \mathbf{w}_{y_{si}}^\top\right)\tilde{\mathbf{f}}_{si} + \left(b_c - b_{y_{si}}\right)}\right)\right], \tag{5}$$

where $\mathbf{W}$ and b represent the weight matrix and offset vector of the last layer of the network and the full connection layer, respectively. Similarly, in order to improve the domain adaptability under unsupervised learning, it is necessary to use the joint distribution probability of label space and feature space. At this time, the target domain space lacks annotation, so we need to use the pseudolabel method and target features to form the mutual information value $\mathscr{L}_{M_I}$. Therefore, TSA is defined as follows:

$$\mathscr{L}_{\mathrm{TSA}} = \mathscr{L}_M + \beta \mathscr{L}_{MI}, \tag{6}$$

where β is an empirical parameter, and the value needs to be compared and explored in the experimental part [23]. The TSA loss function based on the interdomain characteristic mean deviation and class conditional covariance has less computation than the data generation antidomain adaptation, and its lightweight advantage can be embedded in other domain adaptation algorithms. Therefore, combined with the above domain adaptation methods, the loss function of unsupervised bearing fault transfer learning is constructed as follows:

$$\mathscr{L}_{\mathrm{LOSS}} = \mathscr{L}_{\mathrm{JMMD}} + \mathscr{L}_{\mathrm{CDA}} + \mathscr{L}_{\mathrm{TSA}}. \tag{7}$$

In the process of domain adaptation, due to the random initialization of the network, the network parameters cannot reflect the real domain feature distribution in the initial stage, so the rich labels in the bearing source domain data are used for pretraining before transfer. After the set epoch, the classification layer can achieve better classification effect on the source domain and then start domain adaptation.

2.4. Deep Residual Shrinkage Networks. In the deep domain adaptation, the network backbone plays an important role in feature transfer. To some extent, the appropriate backbone network is more important than the advanced transfer algorithm [21]. In many experiments on domain adaptation, such as CNN [24], ResNet [25], VGG [26], and AE [27] show excellent feature extraction ability in the application of image transfer, semantic transfer, and signal transfer. However, there is no relevant research on which appropriate backbone network should be selected for specific transfer objects. The purpose of this paper is to carry out the transfer learning for bearing faults under complex and variable working conditions. It is hoped that the fault can be diagnosed early by the vibration data and labels collected from the bearings under unknown working conditions. Considering that in the process of actual vibration signal acquisition, the sensor collects not only the actual vibration signal of the tested bearing but also other noise signals, such as bearing vibration interference of other parts, noise interference of working environment, and noise interference of transmission parts. Suppressing noise interference is always a difficult and hot issue to extract weak signals of bearing early fault by signal processing methods [28]. Among them, various improved algorithms based on wavelet transform are widely used in bearing signal noise filtering, but the premise is to master the prior knowledge of the signal, and the design of the filter and the selection of wavelet parameters need continuous experiments to obtain the optimal value. In addition, the existence of noise will reduce the ability of the neural network to extract weak early fault feature signals and make the boundary of high-dimensional feature clusters blurred in clustering, so that the effect of domain adaptation becomes worse in the process of transfer. Therefore, in order to overcome the influence of noise on the domain adaptation of the bearing fault signal, a network layer similar to the filtering algorithm needs to be embedded in the backbone network to adaptively reduce the influence of noise on feature extraction. Zhao et al. [29] proposed adding a soft threshold to the residual network to automatically learn the noise threshold, reduce the noise interference, and realize the bearing fault diagnosis under high noise. Based on the deep residual shrinkage network, this paper will improve part of the network structure to build the backbone network in transfer learning.

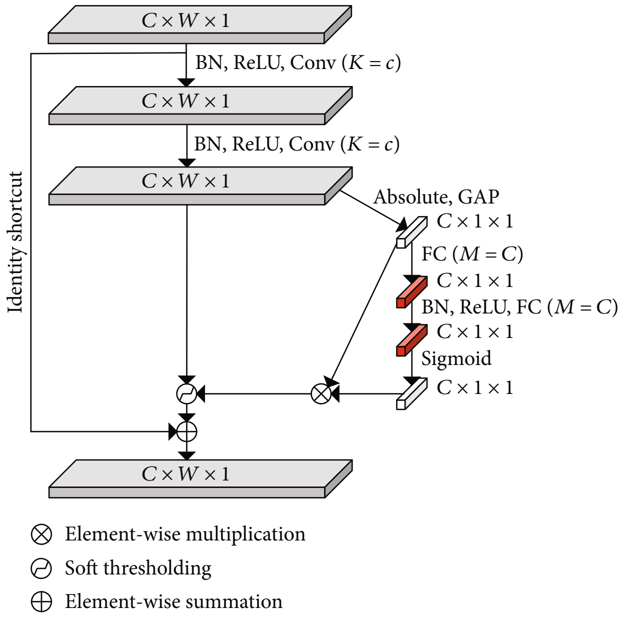


Figure 2: A building unit entitled RSBU-CW.

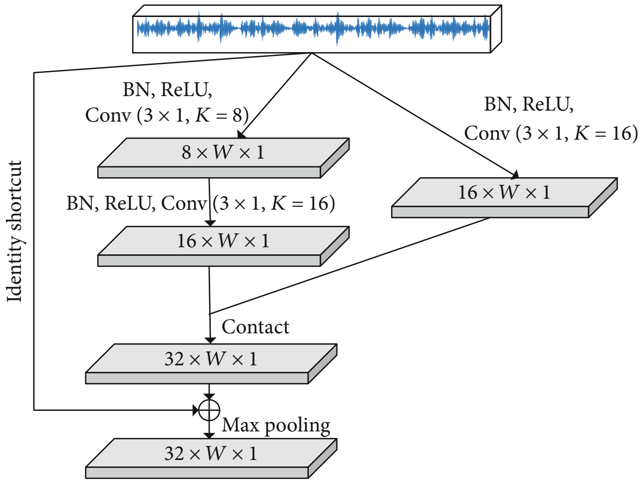


Figure 3: A building unit of improved pooling layer.

2.4.1. Deep Residual Shrinkage Module. The residual shrinkage module is the basic unit of the deep residual shrinkage network, which embodies the idea of an attention mechanism: by eliminating the data features with a low contribution ratio, the important features are more prominent in the overall data features. Although this approach may eliminate the features conducive to transfer learning, it will remain in the network through the identity mapping in the residual module, but its proportion will be reduced after the transfer of the module. A residual shrinkage building unit with channel-wise thresholds (RSBU-CW) is shown in Figure 2.

C and W, respectively, represent the width and channel of the feature. Each channel of RSBU-CW has an independent threshold. The features are reduced to a one-dimensional vector through absolute value and global average pooling, and then, the one-dimensional vector is transmitted to the fully connected network with two-layer FC. Each channel can have an independent threshold by making

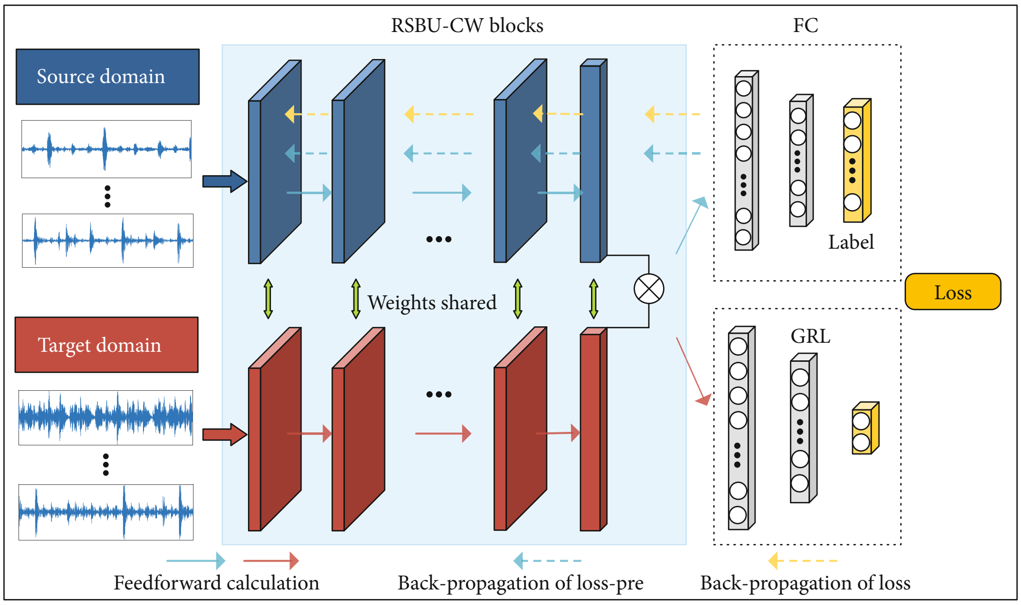


Figure 4: Domain transfer network based on deep residual shrinkage residual module.

Table 2: The description of class labels of CWRU.

Task	Speed (rpm)	Fault location	Fault size (mils)	Class label
0/1 2/3	1797/1772 1750/1730	OF	7	1
		OF	14	2
		OF	21	3
		BF	7	4
		BF	14	5
		BF	21	6
		IF	7	7
		IF	14	8
		IF	21	9
		NA	0	10

Table 3: The description of time-varying speed dataset.

Task	Speed-varying conditions	Fault location	Class label
0	Increasing speed	Healthy	1
1	Decreasing speed		
2	Increasing then decreasing speed	Inner race fault	2
3	Decreasing then increasing speed	Outer race fault	3

the number of neurons in the second layer consistent with the number of channels of the input feature map. The threshold $\boldsymbol{\tau}_c$ can be defined as

$$\boldsymbol{\tau}_c = \boldsymbol{\sigma}_c \cdot \text{average}\left|\mathbf{X}_{i,j,c}\right|, \tag{8}$$

where $\boldsymbol{\sigma}_c$ is the parameter of the cth layer scaled to $(0, 1)$ and i, j, and c are the indexes of width, height, and channel of the feature map X. In the network iterative training, the threshold of each channel will change with time. When the feature is in the range of $[-\boldsymbol{\tau}_c, -\boldsymbol{\tau}_c]$, the channel threshold will be set to 0, and those features $\mathbf{X}$ far from 0 will approach 0.

2.4.2. Improved Pooling Layer Based on Inception Module. The inception module is proposed to solve the problem of performance saturation and light weight when the number of layers of the google net network is deepened. From inception-V1 to V4, the model is constantly improved and the performance is also continuously improved. The main idea of inception is to transform large convolution blocks into small convolution blocks through series and stacking. Because the collected bearing data is one-dimensional vibration data, convolution pooling needs to be carried out before being transmitted to RSBU-CW. If the traditional large convolution block 7 × 1 is adopted, it will not be suitable for fault diagnosis of large bearing data in industry. In order to ensure the effect of feature extraction and reduce the volume of network calculation, for the one-dimensional time-domain input signal of bearing, an improved data pooling layer is shown in Figure 3.

The improved data pooling layer in the figure adopts 3 small convolution layers instead of 7 × 1 convolution layers. The number of channels is set to 8, 16, and 16, respectively, and the residual connection structure is added. Finally, the extracted data feature information is output through the maximum pooling layer.

2.4.3. Residual Block-Based Dilated Convolution. Dilated convolution has the same convolution operation as ordinary convolution, but dilated convolution uses a specific step to read data in a jumping way, which can obtain a larger receptive field while keeping the parameters unchanged, so that each convolution output contains more information. Therefore, this paper replaces the ordinary convolution in RSBU-

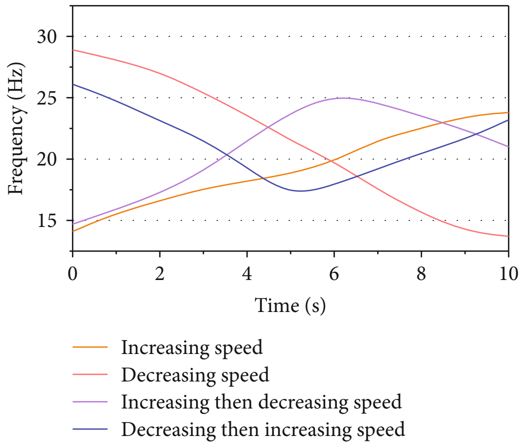


Figure 5: The change of frequency under time-varying rotational speed conditions.

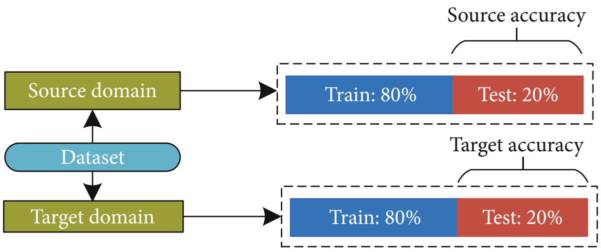


Figure 6: Division of input data.

CW with dilated convolution and increases the receptive field by setting the dilated rate. It is assumed that the kernel width of ordinary convolution is w. When the dilated convolution with dilated rate d is introduced, the width of the dilated convolution kernel becomes $w + (w - 1)(d - 1)$. And one-dimensional convolution is used; the height of the convolution kernel is always 1. Dilated convolution improves the sparsity of bearing signal characteristics, but at the same time, in order to ensure the continuity of vibration signal after convolution operation, the dilated rate should not be set too large. It is verified by relevant experiments that the bearing fault accuracy is higher when the dilated rate $d = 2$ is adopted. At this time, the receptive field obtained by the 3×3 convolution kernel in RSBU-CW is equivalent to the receptive field brought by the 5×5 convolution kernel.

2.5. Framework of Network Training. As shown in Figure 4, it is the network training framework proposed in this paper. Aiming at the problem of the generalization ability of traditional deep learning for bearing fault diagnosis under variable conditions, a domain transfer training network based on the deep residual shrinkage residual module is proposed. The whole network is composed of a backbone network using the RSBU-CW module and deep domain transfer algorithm. In order to improve the effect of domain transfer, the loss function consists of explicit JMMD loss and CDA loss and implicit TSA loss. The network is pretrained through the bearing vibration data with known labels in the source domain. After updating the network parameters, the unlabeled target domain data are transferred to accelerate the speed of the domain adaptation.

3. Experimental Results

In this section, two open-source bearing datasets will be used to verify the effectiveness of the proposed method in bearing fault diagnosis. The main framework is written by Python. All experiments were run on a computer equipped with i7-9300h CPU and NVIDIA GeForce GTX 1050 GPU.

3.1. Datasets

3.1.1. Case Western Reserve University (CWRU) Dataset. The CWRU [30] bearing dataset is an open-source dataset of the Case Western Reserve University Laboratory which is widely used in the research of bearing fault diagnosis. In the experiment, the amplitude data of SKF6205 motor bearing are collected by the acceleration sensors installed at the motor driving end and fan end. The data consists of normal bearing operation data and fault bearing operation data. The fault location and damage size are different. Detailed data description is shown in Table 2.

The bearing transfer tasks are $\{0, 1, 2, 3\}$, corresponding to four different speeds, respectively. The load of the bearing is also different at each speed. At a certain constant speed, it is divided into 10 data types. The locations of bearing faults are inner fault (IF), ball fault (BF), and outer fault (OF), where NA represents normal bearing.

3.1.2. University of Ottawa Bearing Dataset. The dataset is collected from the University of Ottawa laboratory [31]. Each sample of this dataset is collected under time-varying rotational speed conditions, which is different from the CWRU dataset. Detailed data description is shown in Table 3.

The collection time of each sample is 10 s in total, and the sampling frequency is 200 kHz. During the sampling time, the running speed of the bearing will change, which can be divided into four types: acceleration, deceleration, acceleration before deceleration, and deceleration before acceleration. As shown in Figure 5, the changes of bearing operating speed under four operating conditions are shown, respectively, and the speed is represented by bearing rotation frequency. The bearing health status is divided into three conditions: normal, inner ring fault, and outer ring fault. Among them, the transfer task $0 \longrightarrow 1$ indicates that the source domain is the fault data under the accelerated running condition, and the target domain is under the condition of bearing deceleration.

3.2. Implementation Details

3.2.1. Division of Input Data. The two datasets are slightly different in sample balance. Among them, the number of samples under normal working conditions of the CWRU dataset is more than that under other working conditions, while the sample number of Ottawa bearing data is well balanced. This experiment does not deal with the sample balance but makes the number of samples in each source

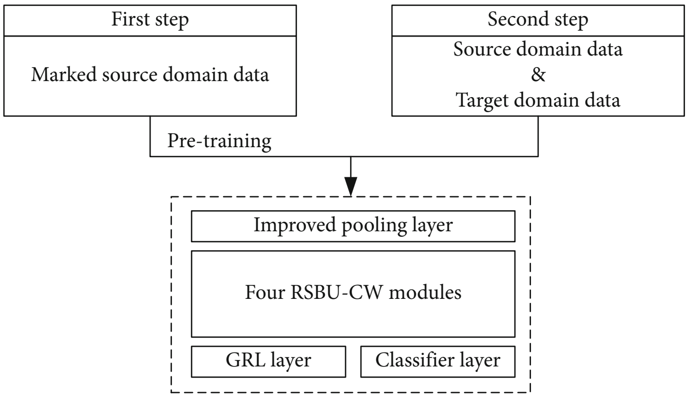


Figure 7: Training method.

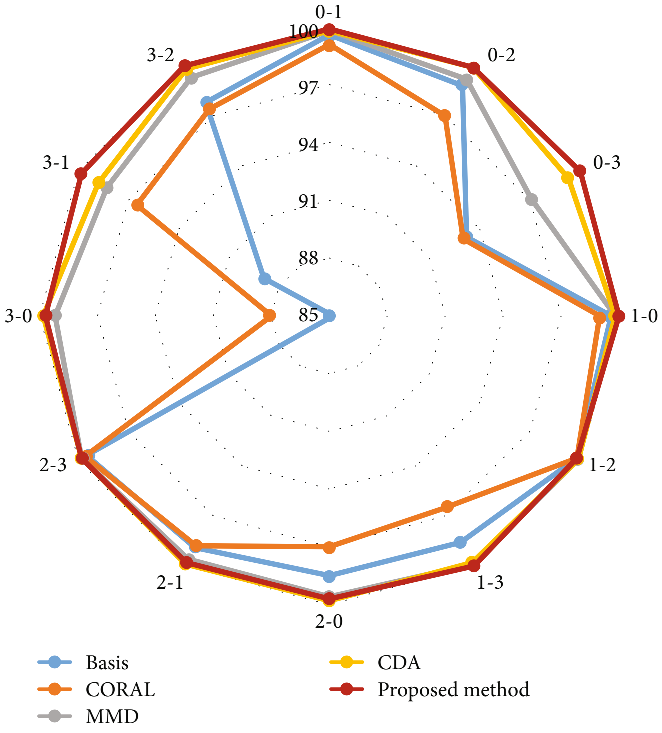


Figure 8: Comparison of accuracy of five methods under CWRU dataset.

domain and target domain the same. When segmenting the bearing timing vibration signal, the method of enhanced data will not be used, because the enhanced data may overlap the training data and test data in a certain period of time, resulting in unreliable test accuracy. As shown in Figure 6, it is a schematic diagram of the division of source domain and target domain samples. Among them, the training sets of source domain samples and target domain samples account for 80%, and the test sets account for 20%. Considering the difference in sampling frequency between the two datasets, 1024 and 8192 are taken as the sample length of the CWRU dataset and Ottawa dataset, respectively.

3.2.2. Training Method. After the reasonable division of samples, the data of the source domain and target domain are sent to the network for training. The one-dimensional bearing vibration data sample first passes through the improved pooling layer proposed in this paper, then passes through four RSBU-CW modules, and finally calculates the domain adaptive loss value through the GRL layer and classifier layer. In network training, the updating of parameters is divided into two stages. As shown in Figure 7, the first is the pretraining process using the marked source domain data, in which the target domain data does not participate in the training process. In the second stage, the source domain and target domain data are sent to the network at the same time for domain adaptation, and the loss value of domain adaptation is used for back propagation. Among them, 50 epochs are set for pretraining nodes and 200 epochs are set for domain adaptation nodes. The gradient descent algorithm adopts Adam, the momentum value is set to 0.9, and the batch size is 64.

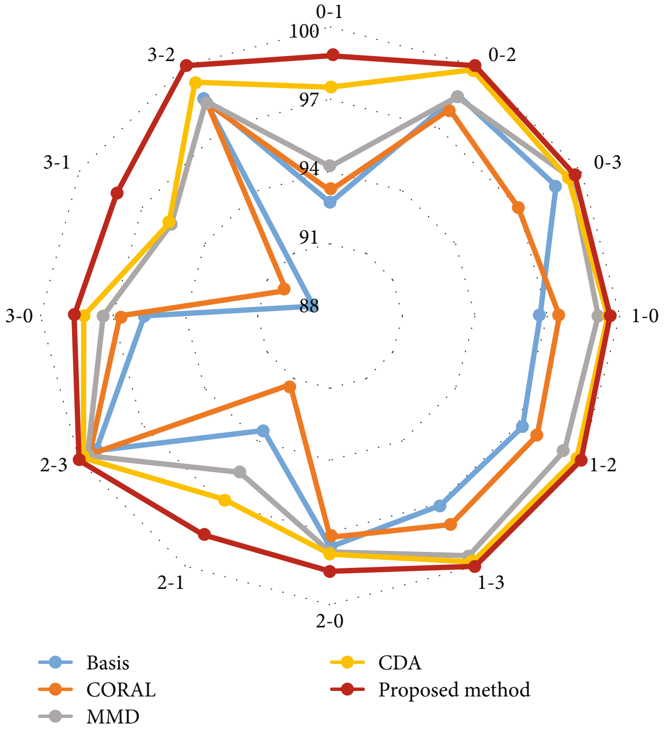
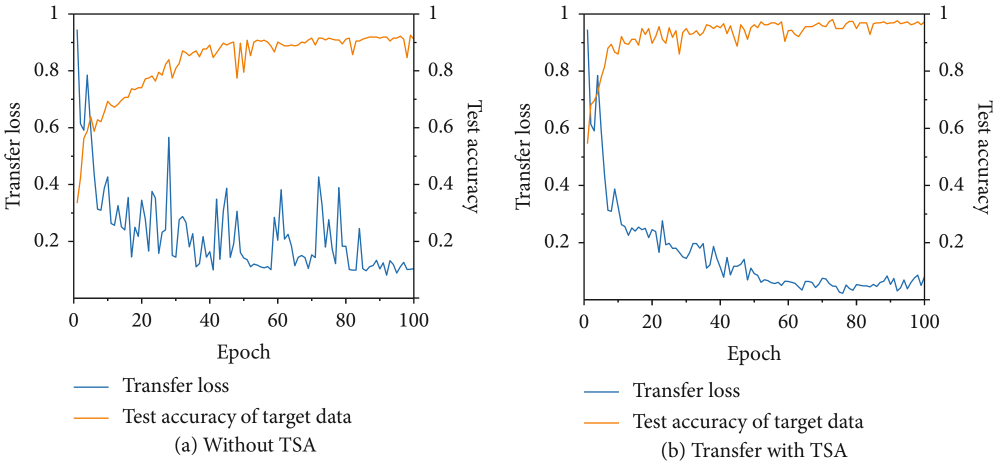


Figure 9: Comparison of accuracy of five methods under Ottawa dataset.



(a) Without TSA

(b) Transfer with TSA

Figure 10: Effect of TSA regular term on transfer loss and test accuracy of target data.

3.3. Evaluation Results. In this paper, two open-source datasets are used to verify the effectiveness of the proposed model for bearing fault diagnosis under variable working conditions. In order to fit the industrial application scenario, the applicability of the model under high noise will also be discussed, and the results will be represented by visual charts.

3.3.1. Results of Models. As shown in Figures 8 and 9, the test accuracy of two datasets under five methods is shown, respectively. The five methods are Basis, CORAL, MMD, CDA, and the method proposed in this paper. Among them, Basis means that it does not use any domain adaptation method and only uses the network trained by the source domain data to test the test set of the target domain directly. The other three methods are common domain adaptation methods. In order to ensure the reliability of the results, 10 experiments were carried out for each method, and the average value of the test results of the last epoch was taken as the final result.

It can be seen from the figure that in addition to the CORAL method, adopting other domain adaptation methods can greatly improve the accuracy of fault diagnosis under variable working conditions; especially when the

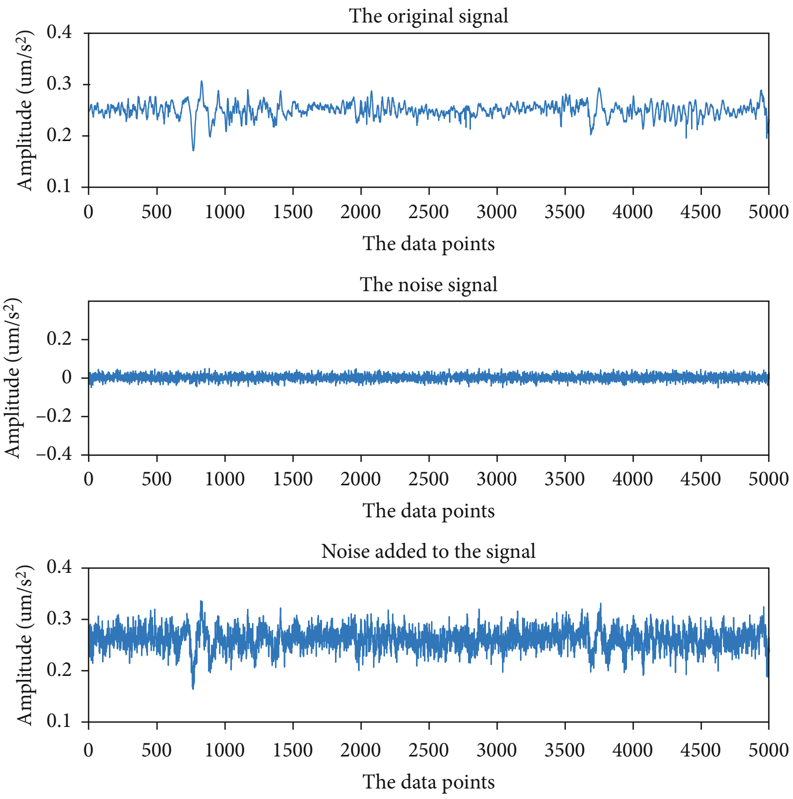


Figure 11: Schematic diagram of adding Gaussian white noise to the signal.

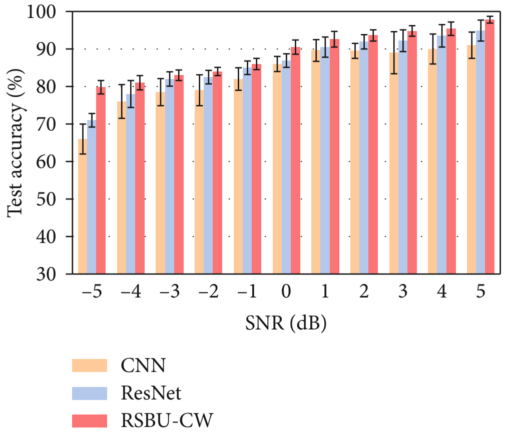


Figure 12: Fault diagnosis accuracy of test dataset under Gaussian noise.

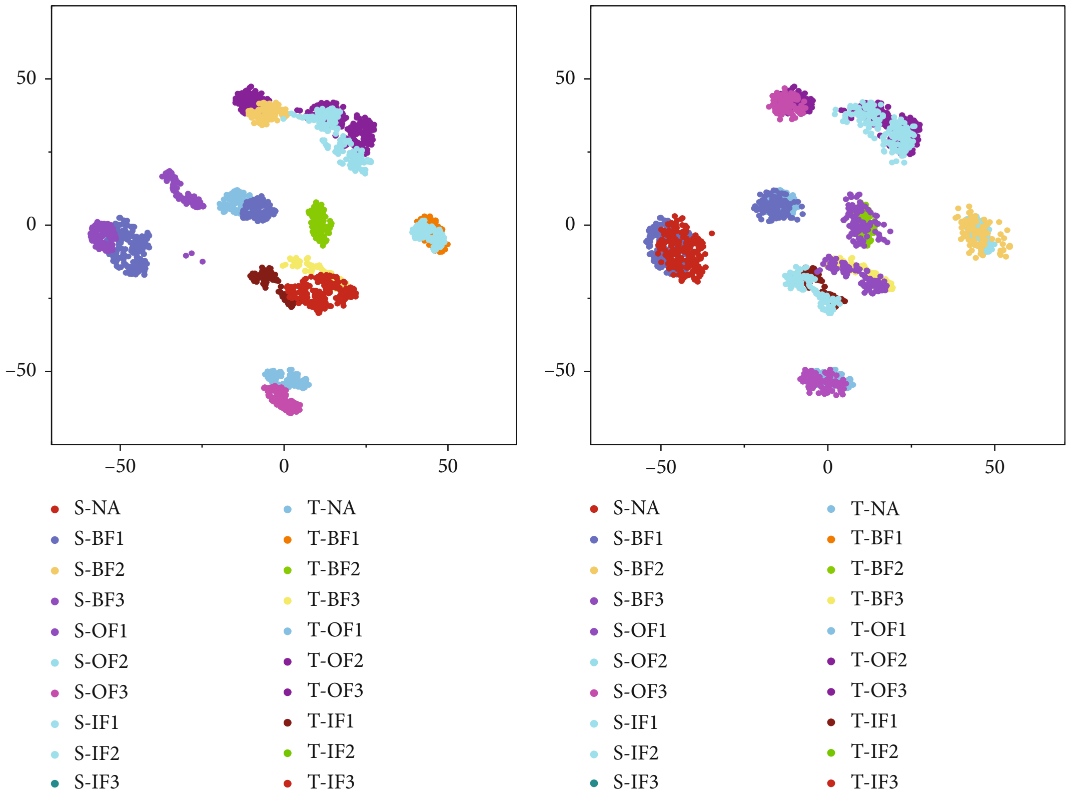
Table 4: The time cost of three backbone networks.

Backbone network	SNR (dB)	Training time (s)
CNN	-5 dB	1256 s
ResNet	-5 dB	1302 s
RSBU-CW	-5 dB	1109 s
CNN	5 dB	1145 s
ResNet	5 dB	1204 s
RSBU-CW	5 dB	1003 s

working conditions are greatly different, the more obvious the migration effect is. For example, for the mutual migration under the two working conditions $\{0, 3\}$ in Figure 8, the lifting effect is the most obvious under the domain adaptation method. For CWRU datasets, although the proposed method does not have much chance for improvement, the overall accuracy is still slightly higher than other methods. This is closely related to the dataset itself, because the faults of the CWRU dataset are artificially set, and the fault characteristics are obvious. In addition, the migration of the CWRU dataset is mainly the transfer under different loads, and the bearing data is measured under a uniform speed condition. In comparison, the working condition of the Ottawa dataset is more complex, and the speed difference between the migrated datasets is more obvious. The change of speed will lead to the change of fault characteristics, so the accuracy of Figure 9 is lower than that of Figure 8 as a whole. On the whole, the proposed method combines the advantages of domain confrontation migration and joint distribution migration. By embedding TSA loss, it solves the problem of domain drift in the traditional domain adaptation methods and enhances the adaptability of the classifier.

In order to further illustrate the influence of adding the TSA regular term on the training convergence results, Figure 10 draws the loss curve and test accuracy before and after adding the TSA regular term. The transfer task of the curve shown in the figure is (0, 1) in the Ottawa dataset.

It can be clearly seen that after adding the TSA regular term, the fluctuation of the transfer loss decline curve is



(a) Basis method

(b) Proposed method

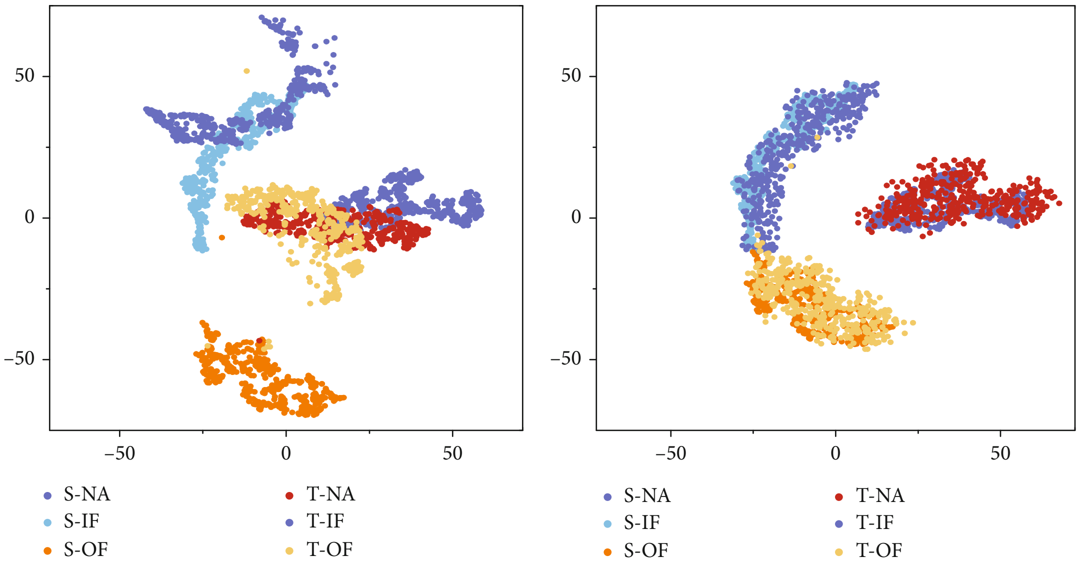
Figure 13: Network visualization of CWRU dataset transfer task 3 ⟶ 1.

improved, and the accuracy is significantly improved after 10 epochs. This is because before adding the TSA regular term, the domain migration method based on confrontation will cause the fluctuation of loss value, which will affect the accuracy of the test set. TSA can implicitly strengthen the migration of data features from the source domain to the target domain, enhance the ability of the classifier to adapt to the domain, and reduce the fluctuation of classification effect caused by adversarial domain training.

3.3.2. Robustness of Backbone Network. Backbone networks also have a great impact on domain adaptation. In order to fairly compare the effects of transfer learning, comparative experiments need to be carried out under the same backbone network, so the discussion of backbone networks has been ignored. In this paper, we choose a deep residual shrinkage network with antinoise effect, one is because of the need for industrial actual conditions, two is to suppress the effect of noise on domain migration. The actual measured bearing vibration signals contain rich noise signals, which can cause redundancy in data-intensive places, and denoising will help domain migration. As shown in Figure 11, in order to simulate the actual working environment, Gaussian white noise is added to the target in the experiment.

The experimental object is Ottawa bearing data. Three different network structures CNN, ResNet, and RSBU-CW are adopted to carry out the experiment according to the set domain migration task. The noise intensity is −5 dB ~ 5 dB, 10 experiments are carried out for each migration diagnosis task, and finally, the average value is taken as the result. As shown in Figure 12, the fault diagnosis accuracy of the test set in Gaussian noise is the average of all migration tasks. Among them, CNN and ResNet adopt convolution blocks of the same size as RSBU-CW. From the final results, it can be seen that the domain migration diagnosis effect of RSBU-CW in high noise environment is better than that of traditional CNN and ResNet and can maintain strong robustness. This is because the backbone network adopts the soft threshold as the shrinkage function, which effectively suppresses the redundant noise features in the bearing fault features, so as to give full play to the effect of the domain adaptation method. Although the introduction of a soft threshold will increase the amount of calculation of the network, higher fault diagnosis accuracy under variable working conditions of bearing is obtained. In order to further compare the influence of soft threshold on model complexity, Table 4 compares the training time of the three models under the noise intensity of -5 dB and 5 dB, respectively. It can be seen that the improvement of the data pooling layer in this paper has offset the influence brought by the introduction of a soft threshold algorithm to a certain extent.

3.3.3. Network Visualization. Figures 13 and 14 are the network visualization results embedded in the full connection layer using *t*-distributed stochastic neighbor embedding (t-SNE), where *S* represents the source domain sample, *T* represents the target domain sample, and *S*-Na represents the



(a) Basis method

(b) Proposed method

Figure 14: Network visualization of Ottawa dataset transfer task 3 ⟶ 1.

normal sample in the source domain. It can be seen from the clustering results that when no domain adaptation method is used, classes in the same domain can be effectively distinguished, but classes with the same source domain and target domain cannot get the correct mapping relationship, and a large number of overlapping regions can be found in the graph. Under the framework proposed in this paper, the same category between source and target domains is aligned well. Therefore, both the accuracy of domain adaptation and the results of network visualization prove the effectiveness of the network framework in this paper.

4. Conclusion

Based on the deep residual shrinkage network, this paper uses the combination of conditional domain adversarial domain adaptation and joint distribution domain adaptation to solve the problems of low fault diagnosis accuracy, weak antinoise performance, and weak generalization ability caused by load change and noise interference in the actual operating environment of rolling bearing. According to the experiments on two open-source datasets, the conclusions are as follows:

(1) The transfer method proposed in this paper integrates the advantages of adversarial domain transfer and joint distributed transfer. At the same time, by adding the TSA regular term, it effectively solves the problem of domain drift under unsupervised domain adaptation. Compared with other traditional preadaptation methods, the accuracy is increased. At the same time, it improves the performance of transfer between different fields and expands the application scope of intelligent fault diagnosis. It provides a new idea to solve the problem of facing a large number of unmarked data in bearing fault diagnosis

(2) Adding a soft threshold in the backbone network improves the robustness of the whole network framework. At the same time, in the antinoise experiment, the performance of the deep residual shrinkage network using a soft threshold is about 3% and 6% higher than that of the traditional CNN and ResNet networks, respectively, which realizes the antinoise function of bearing fault diagnosis in industry. In addition, by improving the pooling layer based on the concept module, the feature information in the original data is effectively extracted, so that this method can transfer the feature information in the data efficiently

(3) The two datasets, respectively, contain the variable load and variable speed operation of the bearing. From the diagnosis accuracy of the final test set, the greater the change difference between the source domain and the target domain, the more difficult the transfer. However, the dataset collection used in this paper is carried out in the ideal experimental environment, and there is still a gap with the actual industrial production environment. Therefore, it is still a challenge to carry out early fault prediction on the bearing data without labels under complex variable working conditions. At the same time, there is a lack of relevant comparative experiments on the setting of network training parameters in this paper. In the next research, it will be combined with other intelligent algorithms for relevant optimization

Data Availability

The data used to support the findings of this study are available from the Case Western Reserve University Bearing Data Center Website (https://csegroups.case.edu/bearingdatacenter/pages/welcome-case-western-reserve-university-bearing-data-center-website).

Conflicts of Interest

We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

References

[1] J. Lee, F. J. Wu, and W. Y. Zhao, “Prognostics and health management design for rotary machinery systems– reviews, methodology and applications,” *Mechanical Systems and Signal Processing*, vol. 42, no. 1-2, pp. 314–334, 2014.

[2] Z. W. Gao, C. Cecati, and S. X. Ding, “A survey of fault diagnosis and fault-tolerant techniques-part I: fault diagnosis with model-based and signal-based approaches,” *IEEE Transactions on Industrial Electronics*, vol. 62, no. 6, pp. 3757–3767, 2015.

[3] Z. He, H. Shao, Z. Ding, H. Jiang, and J. Cheng, “Modified deep auto-encoder driven by multi-source parameters for fault transfer prognosis of aero-engine,” *IEEE Transactions on Industrial Electronics*, vol. 69, no. 1, pp. 845–855, 2022.

[4] X. W. Dai and Z. W. Gao, “From model, signal to knowledge: a data-driven perspective of fault detection and diagnosis,” *IEEE Transactions on Industrial Informatics*, vol. 9, no. 4, pp. 2226–2238, 2013.

[5] S. Zhang, S. Zhang, B. Wang, and T. G. Habetler, “Deep learning algorithms for bearing fault diagnostics—a comprehensive review,” *Access*, vol. 8, pp. 29857–29881, 2020.

[6] X. Li, Y. Yang, H. Shao, X. Zhong, J. Cheng, and J. Cheng, “Symplectic weighted sparse support matrix machine for gear fault diagnosis,” *Measurement*, vol. 168, p. 108392, 2021.

[7] B. Merainani, C. Rahmoune, and D. Benazzouz, “A novel gearbox fault feature extraction and classification using Hilbert empirical wavelet transform, singular value decomposition, and SOM neural network,” *Journal of Vibration and Control*, vol. 24, no. 12, pp. 2512–2531, 2018.

[8] C. Lu, Z. Y. Wang, W. L. Qin, and J. Ma, “Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification,” *Signal Processing*, vol. 130, pp. 377–388, 2017.

[9] S. J. Pan, I. W. Tsang, and J. T. Kwok, “Domain adaptation via transfer component analysis,” *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2011.

[10] H. C. Shin, H. R. Roth, and M. C. Gao, “Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning,” *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[11] S. J. Pan and Q. A. Yang, “A survey on transfer learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[12] X. Li and W. Zhang, “Deep learning-based partial domain adaptation method on intelligent machinery fault diagnostics,” *IEEE Transactions on Industrial Electronics*, vol. 68, no. 5, pp. 4351–4361, 2021.

[13] H. Zheng, R. Wang, Y. Yang et al., “Cross-domain fault diagnosis using knowledge transfer strategy: a review,” *Access*, vol. 7, pp. 129260–129290, 2019.

[14] T. Han, C. Liu, and W. Yang, “Learning transferable features in deep convolutional neural networks for diagnosing unseen machine conditions,” *ISA Transactions*, vol. 93, pp. 341–353, 2019.

[15] Z. An, S. Li, J. Wang, Y. Xin, and K. Xu, “Generalization of deep neural network for bearing fault diagnosis under different working conditions using multiple kernel method,” *Neurocomputing*, vol. 352, pp. 42–53, 2019.

[16] L. Wen, X. Y. Li, and L. Gao, “A transfer convolutional neural network for fault diagnosis based on ResNet-50,” *Neural Computing and Applications*, vol. 32, no. 10, pp. 6111–6124, 2020.

[17] K. M. Borgwardt, A. Gretton, M. J. Rasch, H. P. Kriegel, B. Schölkopf, and A. J. Smola, “Integrating structured biological data by kernel maximum mean discrepancy,” *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[18] A. Gretton, D. Sejdinovic, H. Strathmann et al., “Optimal kernel choice for large-scale two-sample tests,” *Advances in Neural Information Processing Systems*, pp. 1205–1213, 2012.

[19] M. Long, H. Zhu, J. Wang, and M. I. Jordan, “Deep transfer learning with joint adaptation networks,” *International Conference on Machine Learning*, pp. 2208–2217, 2017.

[20] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?,” 2014, https://arxiv.org/abs/1411.1792.

[21] Z. Zhao, Q. Zhang, X. Yu et al., “Unsupervised deep transfer learning for intelligent fault diagnosis: an open source and comparative study,” 2019, https://arxiv.org/abs/1912.12528.

[22] Y. Ganin, E. Ustinova, H. Ajakan et al., “Domain-adversarial training of neural networks,” *The journal of machine learning research*, vol. 17, no. 1, p. 2096, 2016.

[23] S. Li, M. Xie, K. Gong, C. H. Liu, Y. Wang, and W. Li, “Transferable semantic augmentation for domain adaptation,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11516–11525, 2021.

[24] W. P. Wang, Z. R. Wang, Z. F. Zhou et al., “Anomaly detection of industrial control systems based on transfer learning,” *Tsinghua Science and Technology*, vol. 26, no. 6, pp. 821–832, 2021.

[25] A. Alhudhaif, K. Polat, and O. Karaman, “Determination of COVID-19 pneumonia based on generalized convolutional neural network model from chest X-ray images,” vol. 180, Expert Systems with Applications, Oct, 2021, 180.

[26] E. Jangam, A. A. D. Barreto, and C. S. R. Annavarapu, “Automatic detection of COVID-19 from chest CT scan and chest X-rays images using deep learning, transfer learning and stacking,” *Applied Intelligence*, pp. 1–17, 2021.

[27] X. Liu, Z. M. Cao, and Z. J. Zhang, “Short-term predictions of multiple wind turbine power outputs based on deep neural networks with transfer learning,” *Energy*, vol. 217, p. 119356, 2021.

[28] A. Rai and S. H. Upadhyay, “A review on signal processing techniques utilized in the fault diagnosis of rolling element bearings,” *Tribology International*, vol. 96, pp. 289–306, 2016.

[29] M. Zhao, S. Zhong, X. Fu, B. Tang, and M. Pecht, “Deep residual shrinkage networks for fault diagnosis,” *IEEE Transactions on Industrial Informatics*, vol. 16, no. 7, pp. 4681–4690, 2020.

[30] “Case Western Reserve University,” http://csegroups.case.edu/bearingdatacenter/home.

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