Hindawi
Shock and Vibration
Volume 2020, Article ID 1274380, 18 pages
https://doi.org/10.1155/2020/1274380

Research Article

Fault Diagnosis under Variable Working Conditions Based on STFT and Transfer Deep Residual Network

Yan Du,[1] **Aiming Wang**,[1] **Shuai Wang**,[1] **Baomei He**,[2] **and Guoying Meng**[1]

[1]*School of Mechanical Electronic and Information Engineering, China University of Mining and Technology, Beijing 100083, China*
[2]*Graduate School, China University of Mining and Technology, Beijing 100083, China*

Correspondence should be addressed to Yan Du; duyan042188@163.com

Received 19 February 2020; Accepted 15 April 2020; Published 4 May 2020

Academic Editor: Anil Kumar

Fault diagnosis plays a very important role in ensuring the safe and reliable operations of machines. Currently, the deep learning-based fault diagnosis is attracting increasing attention. However, fault diagnosis under variable working conditions has been a significant challenge due to the domain discrepancy problem. This problem is also unavoidable in deep learning-based fault diagnosis methods. This paper contributes to the ongoing investigation by proposing a new approach for the fault diagnosis under variable working conditions based on STFT and transfer deep residual network (TDRN). The STFT was employed to convert vibration signal to time-frequency image as the input of the TDRN. To address the domain discrepancy problem, the TDRN was developed in this paper. Unlike traditional deep convolutional neural network (DCNN) methods, by combining with transfer learning, the TDRN can make a bridge between two different working conditions, thereby using the knowledge learned from a working condition to achieve a high classification accuracy in another working condition. Moreover, since the residual learning is introducing, the TDRN can overcome the problems of training difficulty and performance degradation existing in traditional DCNN methods, thus further improving the classification accuracy. Experiments are conducted on the popular CWRU bearing dataset to validate the effectiveness and superiority of the proposed approach. The results show that the developed TDRN outperforms those methods without transfer learning and/or residual learning in terms of the accuracy and feature learning ability for the fault diagnosis under variable working conditions.

1. Introduction

Mechanical equipment is widely used in various industrial fields, and their reliability is directly related to the economic benefits of enterprises and even the safety of personnel [1, 2]. Since machine fault diagnosis methods can identify the health condition of equipment and provide a basis for equipment maintenance, it has important practical significance [3, 4]. Along with the modern machines becoming increasingly complex and sophisticated, fault diagnosis plays a more and more important role in ensuring the safe and reliable operations of machines. And many researchers have done a lot in this field in the decades [5–7].

Traditionally, machine fault diagnosis includes three main steps: signal acquisition, feature extraction, and fault pattern recognition. In the signal acquisition step, vibrational signals are commonly used because they carry tremendous information and can be easily measured. In the second step, many signal-processing methods, including time domain, frequency domain, and time-frequency domain methods, are employed to analyze vibrational signals and extract fault features. Finally, machine learning models are trained using the extracted features to conduct fault pattern recognition, such as random forest (RF) [8], support vector machines (SVMs) [9], artificial neural networks (ANNs) [10, 11], fuzzy inference, and other improved models [12, 13].

Although these traditional fault diagnosis methods have made many achievements, some drawbacks still exist [14–16]. First, in the feature extraction step, features are

manually selected, which require a lot of specialist knowledge and experience. Furthermore, handcrafted features are often task-specific and may only apply to accurately make predictions under certain circumstances. It is difficult to design a set of features that are effective among all conditions. Second, in the pattern recognition step, machine learning models are often used as classifiers and cannot dig out more useful information. As a result, their performances are affected by the handcrafted features to a large extent.

Deep learning, as a relatively new and rapidly developing machine learning methods, has the ability to overcome the above drawback [17]. It has a powerful feature learning ability and can automatically learn the representation features of raw data. Since deep architectures consist of multiple hidden layers, deep learning can learn multiscale/multilevel/hierarchical representation directly from the input data. As a result, more useful information can be extracted. By automatically learning features from the input data, deep learning can reduce the effects of the handcrafted features used in traditional methods. Through model training, deep learning can automatically pick out more discriminative representation features according to the training data, which are helpful to make accurate predictions in the subsequent pattern recognition steps. Deep learning has been paid more attentions and been successfully applied in various areas including natural language processing (NLP), speech recognition, computer vision, and bioinformatics [6]. Unsurprisingly, deep learning also has been used in the machine fault diagnosis field widely. Various deep learning models have been attempted by many researchers [6, 12, 18], such as sparse autoencoder (SAE), deep belief network (DBN), deep Boltzmann machine (DBM), and convolutional neural network (CNN).

As one of the most popular models, CNN has been paid more attentions, since it has some unique structures such as local receptive field, shared weight filter, and pooled subsampling. Recently, many CNN-based methods have shown their effectiveness in machine fault diagnosis applications [19, 20]. Sun et al. [21] proposed a convolutional discriminative feature learning method, which uses convolutional pooling architecture to extract the discriminative and invariant features. Experiments indicate that it is effective and efficient for induction motor fault diagnosis. Jing et al. [22] developed a CNN-based feature learning and fault diagnosis method for gearboxes using frequency data of vibration signals as the input. Experiments on two gearbox datasets validated its effectiveness and demonstrated that feature learning with CNN provides better results than manual feature extraction. Min et al. [23] investigated the CNN with multiple sensors for fault diagnosis. The results showed the proposed CNN-based method was more accurate and reliable than traditional approaches using manual feature extraction. These CNN-based methods are superior to the traditional fault diagnosis methods based on shallow machine learning.

Although CNN has achieved its great success in many machine fault diagnosis tasks, there still exist two problems associated with CNN, which may also exist in other deep learning models [24]. First, it is difficult to train a deeper CNN. The problem of vanishing/exploding gradients may often occur in the process of training deeper CNNs as the layer went deeper, for the gradient is calculated by back-propagation according to the chain rule. Second, a degradation problem has been exposed when deeper CNNs are able to start converging, which leads to a higher training error. Both of the problems limit the further development of CNN in the field of fault diagnosis to a large extent.

Recently, deep residual CNNs (DRNs) have emerged as a state-of-the-art deep learning method [25]. By introducing a residual learning structure with identity shortcuts, data information can be allowed to propagate directly in the whole network, and thus training parameters can be optimized more easily. Therefore, it is easier to train a DRN than a classical CNN constructed by simply stacking more layers. Usually, the deeper the network, the better the features can be learned. Moreover, the identity shortcut realizes identity mapping between the input and output, which can address the degradation problem. Therefore, DRN may have more potential than classical CNNs for machine fault diagnosis. Zhang et al. [26] constructed a 1D DRN for rotating machinery fault diagnosis. Zhao et al. [27] proposed a DRN with dynamically weighted wavelet coefficients for planetary gearbox fault diagnosis. Peng et al. [28] developed a novel deeper 1D CNN with residual learning and used it for fault diagnosis of wheelset bearings in high-speed trains. Experiments showed that all the residual learning-based methods mentioned above obtained better performance than those based on classical CNNs. In this study, a DRN was also constructed for machine fault diagnosis.

However, there still exist several problems associated with deep learning-based fault diagnosis methods [29–31], which also exist in fault diagnosis methods based on the DRN. First, these works are mainly carried out under the assumption that training data and test data share the same distribution. However, in the real world, the working conditions of machines, especially bearings, are not fixed. When the training and test data are collected from different working conditions, their feature distributions could also be different, which would lead to a significant decrease in the diagnosis ability. This is the domain discrepancy problem. Moreover, training deep learning models usually requires a lot of data, while the labeled fault samples are usually scarce in actual fault diagnosis tasks. We cannot get enough samples from each working condition for the training of deeper models under all working conditions. And in many working conditions, only a few samples can be collected.

Recently, aiming at solving the problem of transferring the generalization knowledge from the related tasks to the target tasks, transfer learning is developed [32, 33]. And many transfer learning methods have been widely studied in many areas, such as NLP, text classification, image classification, and biometrics [31, 34, 35]. In order to address the problems mentioned above, transfer learning is introduced into deep learning methods.

In this paper, by combining the DRN with transfer learning, a novel CNN model, named as transfer DRN (TDRN), is proposed to make full use of the knowledge in different working conditions. First, a DRN model is trained

from scratch by using massive data collected from a certain working condition (source domain). Then, the structure and the parameters are transferred to construct a TDRN model in which the structure is altered according to a few labeled data collected from another working condition (target domain). Finally, the TDRN model can be used to conduct fault diagnosis in the target domain.

In addition, it should be noted that the structure of input data also affects the final performance. In essence, TDRN is a kind of CNN model, which is more suitable for processing two-dimensional (2D) data [36]. Therefore, in this study, we convert the one-dimensional (1D) vibration signals into 2D images, which we call vibration images. There are four types of commonly used vibration images, including 2D rearrangement image of 1D signal [16], time-domain waveform image [37], spectrum image [38, 39], and time-frequency image (TFI) [14, 15, 40]. Since TFI can better uncover the dynamic properties of nonstationary vibration signals, it is used as the input data of the deep network models in this paper. Time-frequency images (TFIs) can be obtained by conducting a time-frequency analysis of vibration signals. At present, there are many kinds of time-frequency (TFA) methods [41, 42], such as short-time Fourier transform (STFT) [43], wavelet transform (WT) [44], bilinear/quadratic TFA, and sparse time-frequency analysis (STFA) [45]. To be specific, in this paper, STFT is adopted to convert vibration signals into TFIs since STFT is a simple and easy-to-apply TFA method.

To sum up, a new machine fault diagnosis approach under variable working conditions is proposed based on STFT and TDRN. The novelty is that TDRN, which is a new CNN model based on residual learning and transfer learning, is developed to make full use of the knowledge of different working conditions. By introducing the TDRN, the proposed fault diagnosis approach can utilize data in a working condition (source domain) to obtain better feature learning ability and higher classification accuracy with a small amount of labeled data in another working condition (target domain).

The rest of this paper is organized as follows: Section 2 presents the proposed approach, including a brief introduction of the STFT and the developed TDRN. In Section 3, data description and parameter setup are briefly introduced. In Section 4, experiments on a bearing dataset are conducted to validate the effectiveness and superiority of the proposed approach. Finally, the conclusions of this investigation and future works are presented in Section 5.

2. Proposed Approach

A new approach for machine fault diagnosis under variable working conditions was proposed based on STFT and TDRN. STFT was adopted to convert vibration signals into 2D TFIs. TDRN, which is a new CNN model based on residual learning and transfer learning, was developed to make a bridge between the source domain and target domain. The framework of the proposed approach is shown in Figure 1. It mainly comprises four major stages. The first one is the pretraining stage, in which a DRN model is pretrained with a large number of source domain data from working condition 1. The second stage is the model transfer. In this stage, the structure and parameters of the pretrained DRN are transferred to construct a TDRN model. After that, the TDRN is fine-tuned by a few target domain data from working condition 2, which is the third stage. The fourth and last stage is the fault diagnosis stage. In this stage, the classification for the test samples from working condition 2 can be achieved by the well-trained TDRN. In short, the proposed approach can use data in a working condition (source domain) to obtain better learning ability and higher classification accuracy with a small amount of labeled data in another working condition (target domain).

2.1. Short-Time Fourier Transform. The short-time Fourier transform is one of the most mature and widely used time-frequency analysis methods. The main idea of the STFT is summarized as described below. A windowed signal can be extracted from the desired signal by adding a short-time window, and then the Fourier spectrum of the windowed signal is calculated. By sliding the window along with the time axis, the time-frequency representation of the signal can be obtained [41, 46]. The STFT of the continuous-time signal $\mathbf{x}(t)$ can be expressed as follows:

$$\mathbf{X}(t,\omega) = \int_{-\infty}^{+\infty} \mathbf{x}(\tau) w(\tau - t) \exp(-j\omega\tau) d\tau, \tag{1}$$

where $w(t)$ is the sliding window function.

Correspondingly, the discrete STFT of a discrete signal $\mathbf{x}(n)$ of period N, where $n = 0, 1, ..., N-1$, can be defined as follows:

$$\mathbf{X}(m,l) = \sum_{n=0}^{N-1} \mathbf{x}(n) w(n-m) W_N^{nl}, \tag{2}$$

where $l = 0, 1, ..., N-1$ and $W_N = \exp(-j2\pi/N)$.

Through the STFT, the vibration signal can be transformed from the time domain to the time-frequency domain, and the corresponding time-frequency representation matrix is obtained. In order to train and test the deep networks, the time-frequency representation matrix is needed to be converted into an RGB image, which is a 2D image with 3 channels.

In this paper, imagesec, getframe, and imwrite functions in MATLAB were used to conduct this conversion. Imagesec function was used to display a time-frequency representation matrix as an RGB image. After that, the getframe function captured the axes or figure as a movie frame. Finally, the movie frame was saved as an RGB image by the imwrite function.

2.2. Transfer Deep Residual Network

2.2.1. Deep Residual Network (DRN). A DRN is a type of CNN. Different from the traditional CNN model, the DRN has residual block structures. For deeper CNN architectures, the parameters such as weights and biases are usually not easy to optimize. The residual block structure is helpful for backpropagation of gradients, so as to update the weights

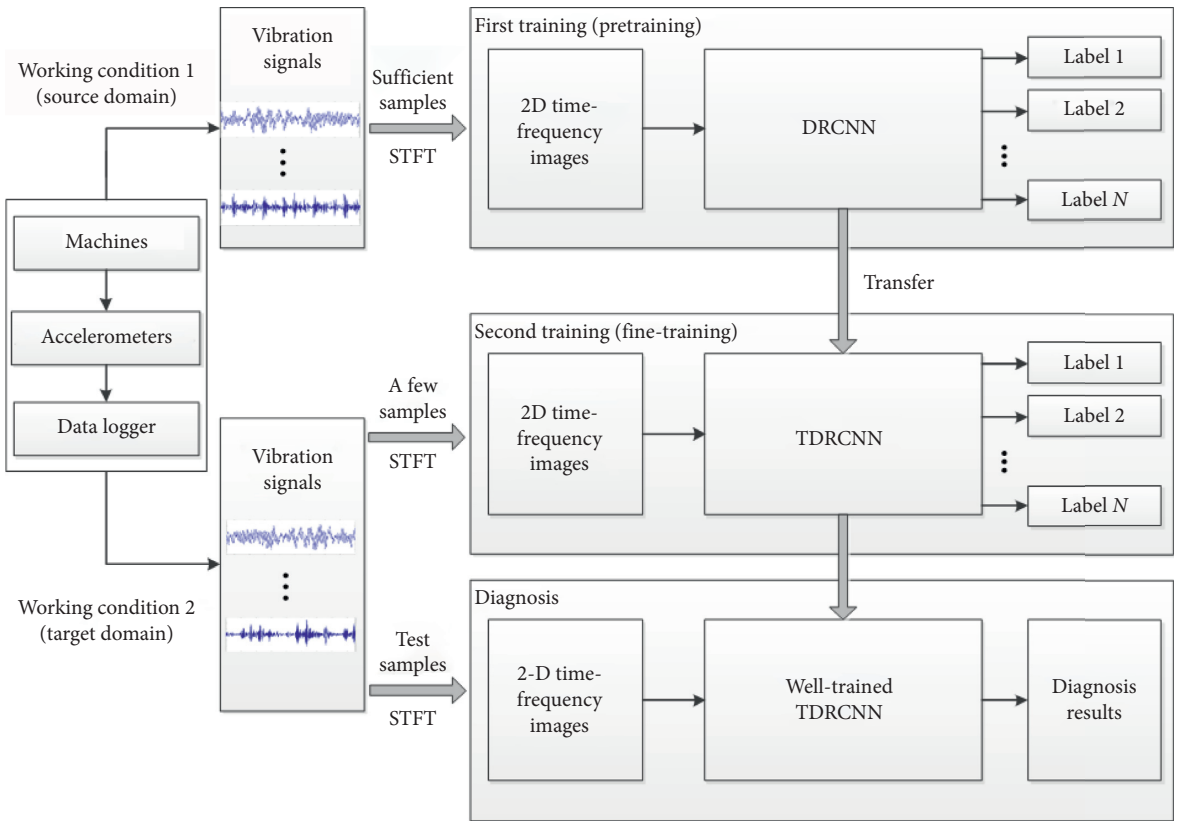


Figure 1: Framework of the proposed approach.

and biases efficiently. Therefore, a DRN was constructed in this study. Figure 2 shows the architecture of the DRN, which uses the TFI of vibration signal as input and consists of convolutional layers, ReLU activation functions, batch normalizations (BNs), residual blocks, a global average pooling layer, a fully connected output layer, and so forth. In Figure 2, we use 'SRB-128' to denote a subsampling residual block with 128 convolution kernels in each convolution layer and 'IRB-128' to denote an identity residual block with 128 convolution kernels in each convolution layer. The others are represented in the same way. More details about the components of the DRN are given as follows.

(1) Convolutional layer

In a convolutional layer, one or more convolution kernels with a scale significantly smaller than the input feature map are used to extract local features, so as to establish the sparse connectivity between two adjacent convolutional layers. Moreover, the weights of the convolution kernels are shared since each kernel slides on the input feature map. Therefore, by means of a convolution operation, the standard convolutional layer introduces the strategies of sparse connectivity and weights sharing, which can reduce the number of parameters and computational complexity compared with the traditional fully connected layer. The convolution layer can be expressed as follows:

$$\mathbf{x}_j^l = \sum_{i \in M_j} \mathbf{x}_i^{l-1} * \mathbf{k}_{ij}^l + b_j^l, \tag{3}$$

where $\mathbf{x}_i^{l-1}$ is the ith channel of the feature map at the $l-1$ th layer, $\mathbf{x}_j^l$ is the j th channel of the feature map at the lth layer, M_j is the selection of channels used for calculating the lth output channel, $\mathbf{k}_{ij}^l$ is the convolution kernel at the lth layer, and b_j^l is the bias corresponding to the jth channel of the feature map at the lth convolution layer. At a convolution layer, each channel of the output feature map corresponds to a convolution kernel, and the convolution kernels corresponding to different channels are different.

In this paper, convolution kernels of size 3×3 were used, as they have high computational efficiency, and are large enough to detect basic local features, including local maxima. As shown in Figure 2, the input image is fed into a convolution layer, which includes 64 convolution kernels of size 3×3 with a stride of 2.

(2) ReLU activation function

After the convolution layer, an activation function is essential. The rectified linear unit (ReLU) activation function [47] is employed in this paper. It can be expressed as follows:

$$f(\mathbf{x}) = \max(0, \mathbf{x}). \tag{4}$$

By forcing the negative feature values to be zero, the ReLU activation function can achieve nonlinear transformations. Compared with the classical sigmoid and tanh activation functions, the ReLU activation function is more effective for avoiding the problem of vanishing/exploding gradients, which can make convergence faster, greatly

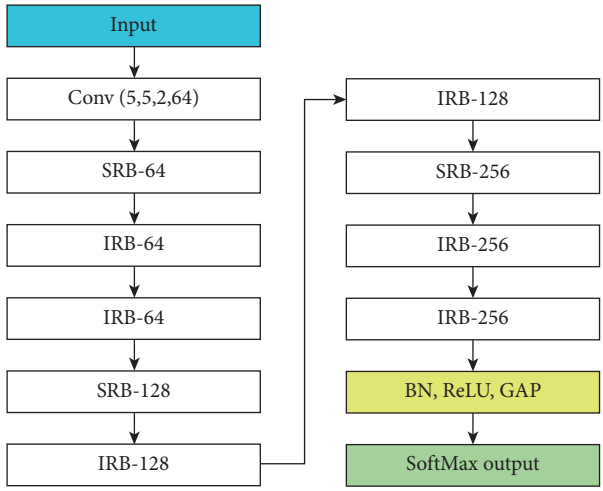


Figure 2: Architecture of the deep residual network.

accelerate training processing, and then further improve the performance of the network.

(3) Batch normalization

In this paper, batch normalization (BN) is used to address the internal covariance shift problem [48]. In each training iteration, the distribution of features learned by the DRN from a randomly selected small batch of training samples often continuously changes. In this case, to adapt to the changed distributions, the weights and biases must be continuously updated, which causes more training difficulty of deep networks. The BN technology forces the inputs to have a similar distribution, which is similar to a standardized operation. Therefore, the BN can address the internal covariance shift problem, so as to improve the training efficiency and enhance the generalization ability of the networks. In this study, BN is deployed before ReLU activation and after each convolutional layer.

(4) Residual blocks

Residual block is based on the idea of skipping one or more convolutional layers by using shortcut connections, which can make the gradients easily backpropagate through a deep network. Therefore, by introducing residual blocks, the weights and biases in a DRN can be updated more effectively than those in a traditional CNN without shortcut connections, and thus, higher accuracies can be yielded by the DRN.

In this paper, two kinds of residual blocks are employed, as shown in Figure 3. Figure 3(a) shows the architecture of identity residual block-m (IRB-m), where m refers to the number of convolution kernels. The IRB-m consists of two branches, where one branch includes two convolution layers, two ReLU functions, and two BNs, and the other is an identity shortcut connection. In each convolution layer, m convolution kernels of size 3×3 with a stride of 1 are employed. It should be noted that the input and output feature maps of the IRB-m must have the same dimensions; otherwise, the addition operation of the two feature maps cannot be implemented. As shown in Figure 3(b), the subsampling residual block-m (SRB-m) consists of three convolution layers, one of which is located on the shortcut connection branch and has the convolution kernels of size 1×1. Unlike the IRB-m, the first convolutional layer of the SRB-m adopts the convolution kernel of size 3×3 with a stride of 2, which can reduce the size of the feature map, thereby reducing the amount of calculation during the model training. Consequently, to match the dimensions, a stride of 2 is employed in the kernels of the convolutional layer located on the shortcut connection branch.

The main difference between the residual blocks at different depths of the network is the number of convolution kernels. As shown in Figure 2, with the increase in depth, the number of convolution kernels of residual blocks increases gradually, from the initial 64 to 128 and then to 256.

(5) Global average pooling (GAP)

In the DRN model, rather than the fully connected layer commonly used in a traditional CNN, a GAP layer is used before the last fully connected output layer. The GAP can greatly reduce the number of parameters, so as to effectively avoid the overfitting problem occurred in the full connection layer and improve the generalization ability of the whole network. In the GAP, the average value of each channel of the input feature map is calculated out, which can be expressed as follows:

$$O_G(c) = \underset{i,j}{\text{average}}\, I_G(i, j, c), \tag{5}$$

where I_G and O_G are the input and output feature maps of the GAP, respectively. Then, the output feature map O_G is sent to the fully connected output layer to obtain the classification results. Compared with the fully connected layer, the GAP is more suitable for the convolution structure by enhancing the correspondence between feature maps and classes. In addition, there are no parameters that need to be optimized in the GAP, so the overfitting problem can be avoided in this layer.

(6) Fully connected output layer

Obviously, the machine fault diagnosis is a multi-classification task. So a SoftMax function is used as the activation function of the fully connected output layer [49]. The SoftMax function is generally used as a classifier to estimate the probability distribution of a sample belonging to different classes, since it can map the output of multiple neurons to the range of (0, 1) and sum up to 1. Assuming that K is the total number of classes, the SoftMax function can be expressed as follows:

$$q_j(\mathbf{x}) = \frac{e^{\mathbf{x}_i}}{\sum_{i=1}^{K} e^{\mathbf{x}_i}}, \tag{6}$$

where x_i is the output feature of the ith neuron in the output layer and $q_j(\mathbf{x})$ is the predicted probability of the input sample $\mathbf{x}$ belonging to the jth class.

Correspondingly, the cross-entropy loss function is used to measure the error between the true values and the outputs of the SoftMax function, which is defined as follows:

$$L = E(p(\mathbf{x}), q(\mathbf{x})) = -\sum_{j=1}^{K} p_j(\mathbf{x}) \log\left(q_j(\mathbf{x})\right), \tag{7}$$

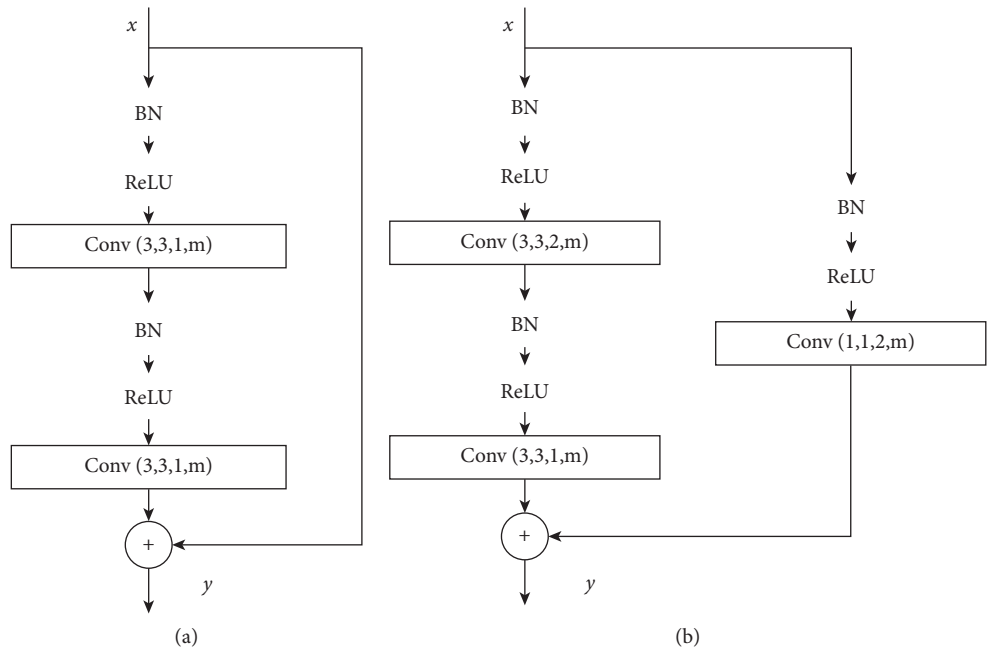


Figure 3: Architectures of two kinds of residual blocks: (a) identity residual block-m (IRB-m) and (b) subsampling residual block-m (SRB-m), where m refers to the number of the convolution kernel.

where $p(\mathbf{x})$ and $q(\mathbf{x})$ are the real and predicted probability distribution of the input sample $(\mathbf{x})$, respectively, and $p_j(\mathbf{x})$ is the predicted probability of the input sample $(\mathbf{x})$ belonging to the jth class.

Finally, an optimization algorithm is used to reduce the cross-entropy value during training so that the predicted distribution gets closer to the true distribution. Thereby, the prediction accuracy of the model can be gradually improved.

2.2.2. TDRN Based on Transfer Learning. The fault diagnosis methods based on DRN have achieved great success in various machine fault diagnosis tasks. However, when the training and testing data are collected from different working conditions, respectively, their feature distributions become different, which leads to a decrease in classification accuracy. Moreover, training deep learning models usually requires a lot of labeled data, but we cannot get enough samples from all working conditions in reality. The two problems lead to a significant decrease in the ability of fault diagnosis. The two problems limit the application of the DRN-based fault diagnosis methods.

Transfer learning can make a bridge between the source domain and the target domain by transferring knowledge, which is very suitable for situations when the source data and the target data are in different feature spaces or distributions. Therefore, transfer learning is introduced into the area of deep learning. Various transfer learning-based methods have been developed and widely studied in NLP, text classification, image classification, and biometrics.

In this study, transfer learning technology is employed to address the two problems mentioned above. By combining DRN with transfer learning, a novel network model, named as transfer DRN (TDRN), is proposed to make full use of the knowledge in different working conditions and thus improve the performance of fault diagnosis. Figure 4 shows the transfer learning strategy of building the TDRN. The detailed steps are as follows:

(1) A DRN model is pretrained from scratch by using massive source domain data collecting from a certain working condition.

(2) The structure and the parameters of the pretrained DRN model are transferred to construct a TDRN model. In terms of transfer learning method, the structure should be altered according to the target domain data collecting from another working condition. In this study, the target domain and the source domain have the same feature space, so there is no need to modify the structure.

(3) All the layers are set to be trainable by setting their trainable attribute to truth for fine-tuning the network.

In short, the TDRN has the same structure as the DRN, while the parameters are not initialized randomly, but pretrained by the source domain data. It should be noted that an optimizer with a very low learning rate should be used when the TDRN is fine-tuned with target domain data. The reason is that too large weight updates can result in the unrestricted magnitude of the modifications for the representations which may harm these representations.

3. Data Description and Parameter Setup

The experimental data used in this paper were obtained from the Bearing Data Center of the Electrical Engineering Laboratory at the Case Western Reserve University (CWRU)

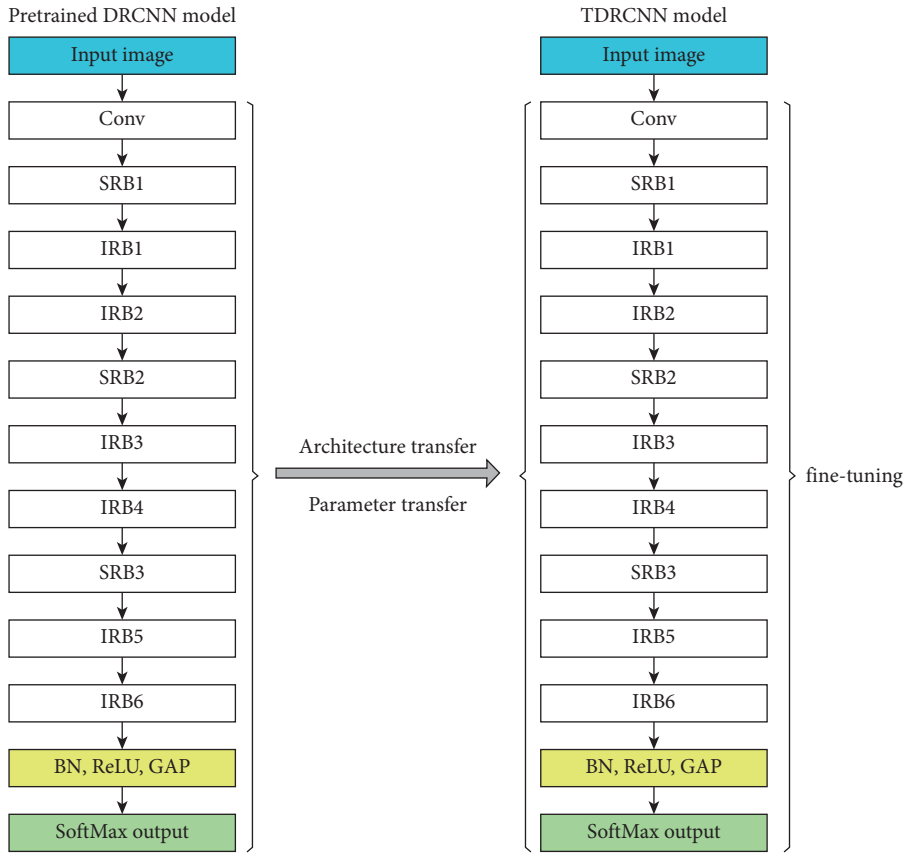


Figure 4: The transfer learning strategy of building the TDRN.

[50–52]. Figure 5 shows the bearing test stand and its structural diagram. The test stand contained a 2 horsepower (HP) motor, a torque transducer/encoder, a dynamometer, and control electronics. Single-point faults with different severity degrees ranging from 0.007 to 0.040 inches in diameter were seeded separately at the inner race, outer race, and ball of motor bearings using electrodischarge machining (EDM). Two accelerometers were used to collect vibration signals. They were installed on the housing using magnetic bases and placed at a 12 o'clock position on the drive end and fan end of the motor housing, respectively. The sampling frequency was set to 12 kHz.

The 6205-2RS JEM SKF deep groove ball bearings located in the drive end were chosen as study objects. The vibration signals used in this paper were collected under four working conditions of load motor loads of 0 to 3 horsepower (motor speeds of 1797 to 1720 RPM). Table 1 lists the description of the four working conditions, denoted A, B, C, and D. There are four classes of health type in each working condition, including normal, inner race fault, outer race fault, and ball fault, which are labeled 1, 2, 3, and 4, respectively.

In the following section, some experiments are conducted to evaluate the performance of the proposed approach. For comparison, through the same transfer strategy, we constructed a TDCNN which is based on a traditional DCNN. This traditional DCNN did not have shortcut connections, while it was the same as the DRN for other architecture and parameters. In addition to the above two models, the DRN and DCNN were also tested. It should be noted that the parameters of DRN and DCNN were randomly initialized.

For the proposed approach, we adopted the Adam optimization algorithm with a learning rate of 0.0001 in the pretraining stage and 0.00001 in the finetuning stage. In each of these two stages, the mini-batch size was set to 32 and 50 epochs were conducted for training. For the comparative method based on TDCNN, we used the same parameters. For the comparative methods without transfer, which are based on DRN and DCNN, respectively, we adopted the Adam optimization algorithm with a learning rate of 0.0001, and the mini-batch size and epoch number for training were same as those for the proposed approach.

Ten trials in each experiment were conducted. All the experiments were carried out on a ThinkPad T470p laptop with Windows 10 operation system, Intel Core i5-7300HQ CPU, 16 GB RAM, and NVIDIA GeForce 940MX GPU. The time-frequency images were constructed using Matlab 2017a. The network models were implemented by Python 3.6 in the popular Keras framework using TensorFlow as a backend.

We adopt classification accuracy for performance evaluation and comparison. The classification accuracy is the ratio of the number of correctly classified testing samples to the total number of testing samples. It can be expressed as follows:

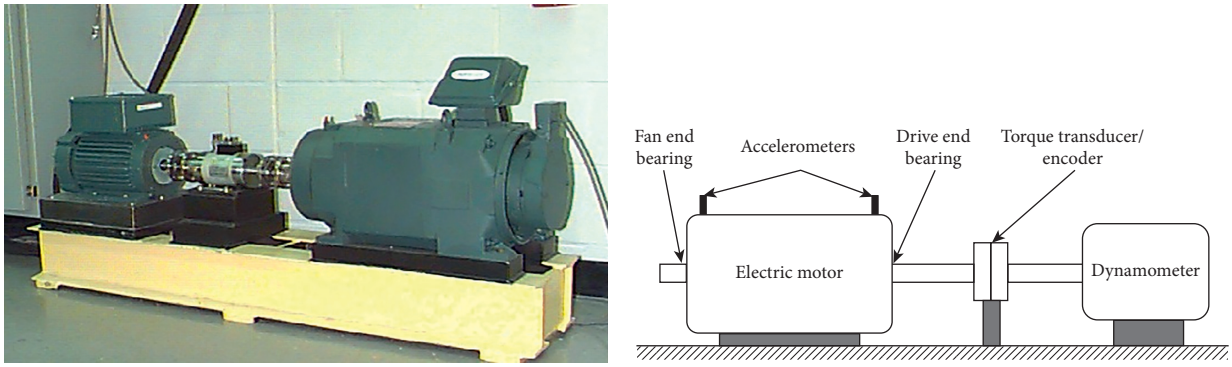


Figure 5: (a) Bearing test stand; (b) its structural diagram.

Table 1: Description of the four working conditions.

Working condition	Load (hp)	Speed (*r*/min)	Class no.	Health type (no. of labels)
A	0	1797	4	Normal (1), inner race fault (2), outer race fault (3), ball fault (4)
B	1	1772	4	
C	2	1750	4	
D	3	1730	4	

$$\text{accuracy} = \frac{N_{CT}}{N_{AT}} \times 100\%, \tag{8}$$

where N_{CT} and N_{AT} are the number of correctly classified testing samples and the total number of testing samples, respectively.

4. Results and Discussion

4.1. Diagnosis Results on a Transfer Task. To validate the effectiveness and superiority of the proposed approach, a transfer task was employed in this section. Table 2 lists the setting details of the transfer task A ⟶ C. The data of source domain and target domain were randomly selected from working condition A and working condition C, respectively. There were sufficient training samples (120 per class) in the source domain, while only a few training samples (12 per class) in the target domain. Another 120 samples of each class in the target domain were selected for testing. Each sample contained 1024 data points. It should be noted that samples of each fault type consisted of the same number of samples with different fault severity degrees (0.007, 0.014, and 0.021 inches).

Experiments were conducted on the transfer task A ⟶ C. Besides the developed TDRN, other methods based on TDCNN, DRN, and DCNN were also used to conduct comparative experiments. It is worth noting that, for comparison, the DRN was trained, respectively, using samples from the source domain (working condition A) and the target domain (working condition C), denoted by DRN-S and DRN-T, and so was the DCNN. The testing results of different models are shown in Table 3.

Obviously, we can see that the TDRN obtained the best performance in terms of the average accuracy since it integrates transfer learning and residual learning. And the second-best performance was obtained by the TDCNN method; that is, the methods with transfer learning obtained higher accuracy than those without transfer learning. The results indicated that the transfer learning method can significantly improve the diagnosis performance under variable working conditions. The results were reasonable because the prior knowledge learned from sufficient source domain data was transferred to the target domain, and the domain discrepancy problem was addressed by fine-tuning the TDRN or TDCNN using a few target domain data. Moreover, the fact that the performance of TDRN was better than that of TDCNN demonstrated that the residual learning structure was benefit for fault classification.

On the contrary, the classification accuracies obtained by them decreased the DRN-S and DCNN-S obviously. The reason is that when the training and test data were collected from different working conditions, respectively, their feature distributions became different, which led to a decrease in classification accuracy. Likewise, the classification accuracies obtained by DRN-T and DCNN-T showed varying degrees of decline, indicating that a few training data were not enough to efficiently train a deep network with a large number of parameters because the overfitting problem may easily occur in such a case. These results showed the urgency and importance to adopt transfer learning in these situations.

In addition, it is worth mentioning that the residual structure can more effectively learn the feature representation related to fault diagnosis in the case of small samples. As we can see in Table 3, the average accuracy of DRN-T was 95.67% substantially higher than 69.54% average accuracy of DCNN-T.

To further present the effectiveness and superiority of the TDRN on diagnosis performance, we used the t-SNE technology to visualize the high-dimensional feature distribution of the test samples extracted from the GAPs of

Table 2: Setting details of transfer task A⟶C.

Transfer task	Domain	Working condition	Classes no.	Sample no. of each class	
				Training	Testing
A ⟶ C	Source	A	4	120	-
	Target	C	4	12	120

Table 3: Testing results of different methods for transfer task A⟶C.

Method	Working conditions for training	Accuracy (%)
TDRN	A, C	99.54
DRN-S	A	91.38
DRN-T	C	95.67
TDCNN	A, C	97.79
DCNN-S	A	92.83
DCNN-T	C	69.54

these models in a low-dimensional space. The 2D representations are shown in Figure 6, where the different health types of rolling bearings are denoted by different colors. The corresponding confusion matrixes are shown in Figure 7.

It can be seen that the feature distributions obtained by different methods were different from each other in terms of intraclass compactness and interclass separability [53].

As displayed in Figure 6(a), the feature distribution obtained by the TDRN had the best intraclass compactness and interclass separability; that is, the features in the same health type were clustered together well and the features of different health types were completely separable. The result indicated that the TDRN can learn more distinguishable fault features from complex TFIs. As we can see from the corresponding confusion matrix of the TDRN shown in Figure 7(a), there were only two samples being misclassified. From Figure 6(b), it can be found that, although the interclass separability slightly decreased, the distributions obtained by the TDCNN had better intraclass compactness. Correspondingly, only a small number of samples were misclassified, as shown in Figure 7(b).

However, as displayed in Figures 6(c) and 6(d), the distributions obtained by DRN-S and DCNN-S suffered from different degrees of decline in intraclass compactness and interclass separability. This indicated that the fault features learned from sufficient data in a working condition cannot be directly used to classify the samples in another working condition. Likewise, we can see from Figures 6(e) and 6(f) that the distributions obtained by DRN-T and DCNN-T had worse intraclass compactness and interclass separability, which meant that a deep network with a large number of parameters cannot be trained efficiently by a few target domain data. The corresponding confusion matrixes shown in Figures 7(c)–7(f) quantitatively confirmed the results mentioned above.

All these results clearly demonstrated that, by combining transfer learning and residual learning together, the TDRN can obtain the best ability of feature learning and classification.

In addition, the t-SNE was employed to visualize the high-dimensional feature distribution of the test samples extracted from different layers of TDRN. Figure 8 shows the 2D visualization results corresponding to 6 layers.

First, it is obvious from Figure 8(a) that the features of different health types in the input layer are very chaotic. Although the features of health type 1 (normal) can be well separable, which showed that STFT had a strong ability to distinguish normal signals and fault signals, the features of the other three health types (fault types) were seriously overlapped. Second, as the layers went deeper, the features of different health types became more and more separable; meanwhile, the features of the same health type were gradually clustered together. For instance, it can be observed that many features of type 2 and type 3 were inseparable in Figure 8(a), most features of type 2 and type 3 were easy to be separable in Figure 8(d), whereas the features were completely separated in Figure 8(e). These phenomena suggested that the network can extract more abstract and higher-level fault features with the increase in the network layer, distinguishing different health types easier. Finally, as shown in Figure 8(f), most features of different health types were completely separated and those of the same health type was well clustered together in the GAP layer, which meant that the TDRN can obtain high accuracy when there was only a small amount of training data in the target domain. All the results demonstrated the intelligence and effectiveness of the developed TDRN.

4.2. Performance under Variable Working Condition Differences. In practical applications, it is often necessary to implement transfer tasks with variable working condition differences. In this section, to test the performance of the proposed approach under variable working condition differences, another two transfer tasks were built using samples from different working conditions as the target domain data. Table 4 lists the setting details of the two transfer tasks. Obviously, transfer task A ⟶ B has less working condition differences than transfer task A ⟶ C, whereas transfer task A ⟶ D has more working condition differences than transfer task A ⟶ C. Similarly, samples of each fault type consisted of the same number of samples with different fault severity degrees (0.007, 0.014, and 0.021 inches). Experiments were conducted on the two transfer tasks. The performance comparison of different methods in the three transfer tasks is displayed in Figure 9.

First, from Figure 9, it can be found that, like the testing results in task A ⟶ C, the developed TDRN also obtained the highest testing accuracy in task A ⟶ B and A ⟶ D, respectively. This further validates the effectiveness and superiority of TDRN under variable working condition differences, and the advantages of transfer learning and residual learning are also confirmed.

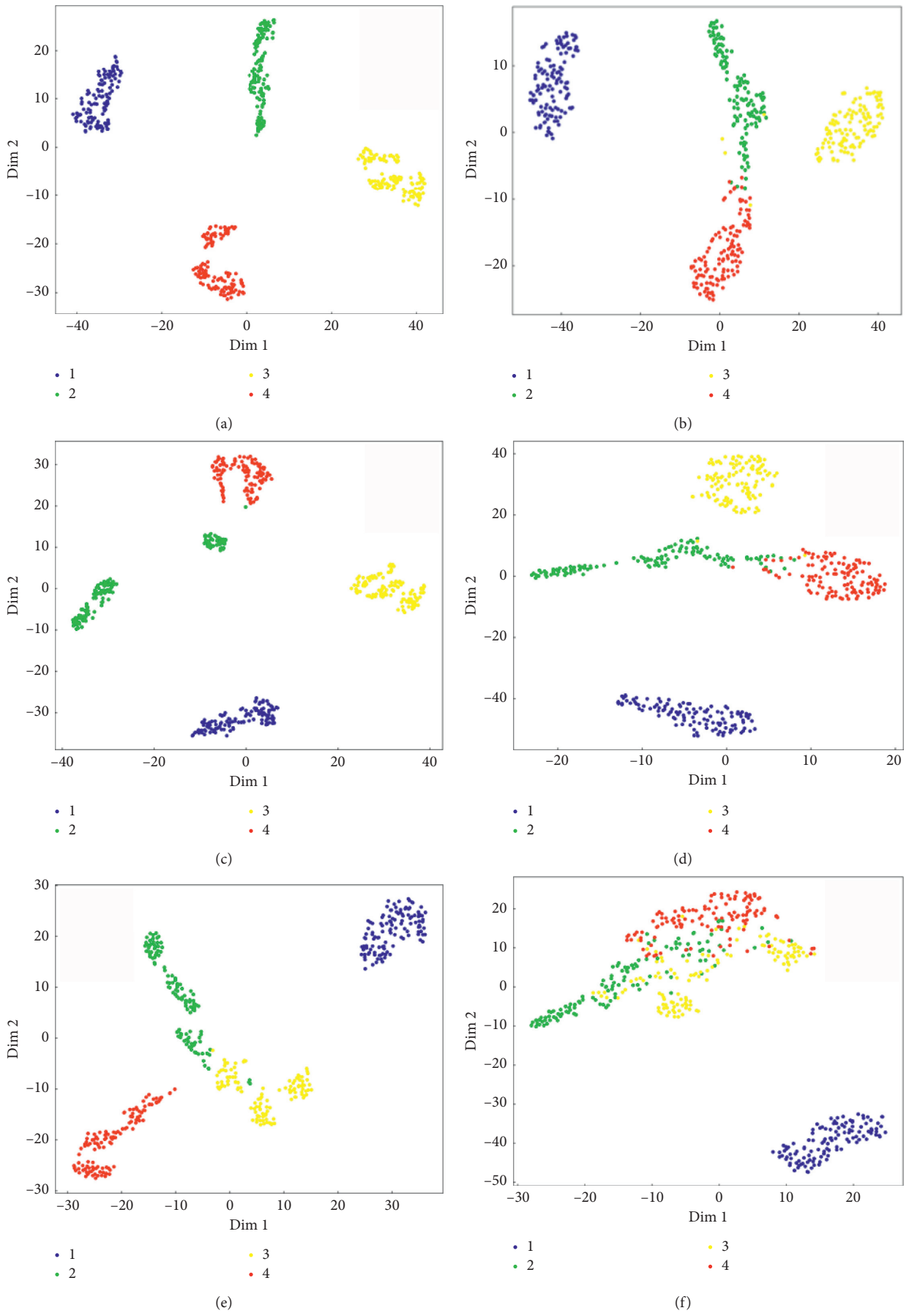


Figure 6: Visualization of the learned features using different methods: (a) TDRN; (b) TDCNN; (c) DRN-S; (d) DCNN-S; (e) DRN-T; (f) DCNN-T.

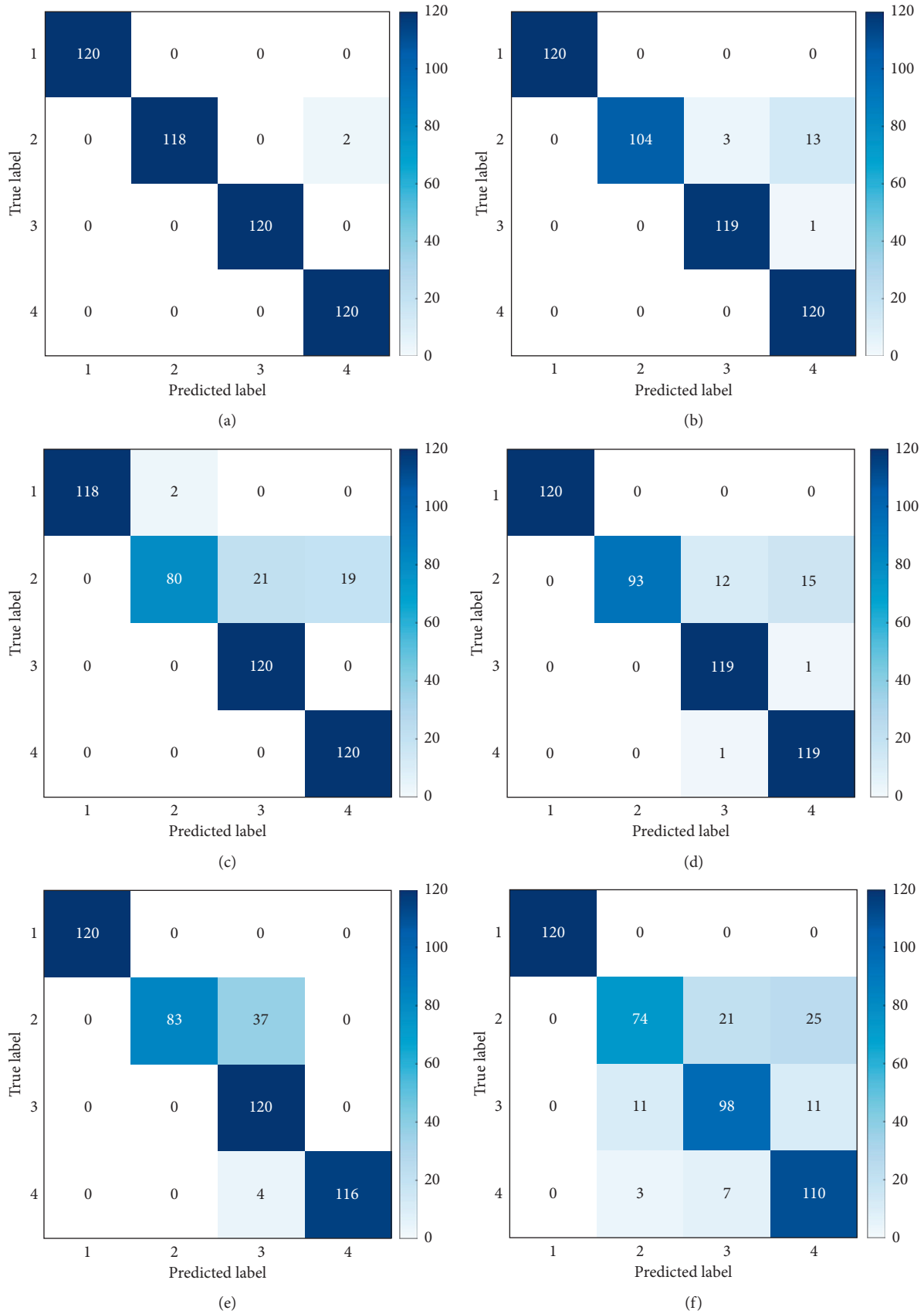


Figure 7: Confusion matrixes using different methods: (a) TDRN; (b) TDCNN; (c) DRN-S; (d) DCNN-S; (e) DRN-T; (f) DCNN-T.

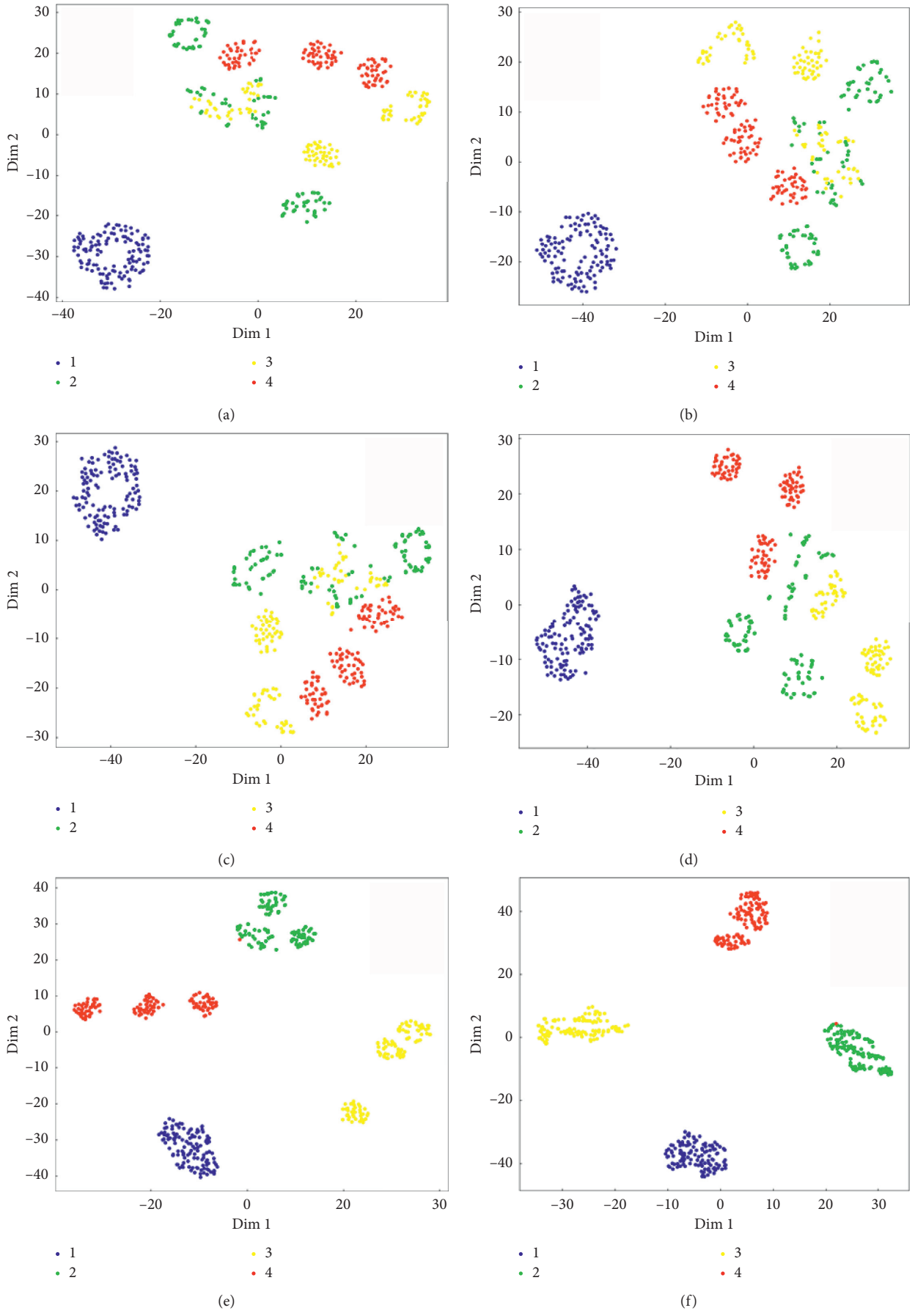


Figure 8: Visualization of the learned features from different layers of TDRN: (a) input; (b) SRB1; (c) SRB2; (d) SRB3; (e) IRB6; (f) GAP.

Table 4: Setting details of transfer task A ⟶ B and A ⟶ D.

Transfer task	Domain	Working condition	Classes no.	Sample no. of each class	
				Training	Testing
A ⟶ B	Source	A	4	120	—
	Target	B	4	12	120
A ⟶ D	Source	A	4	120	—
	Target	D	4	12	120

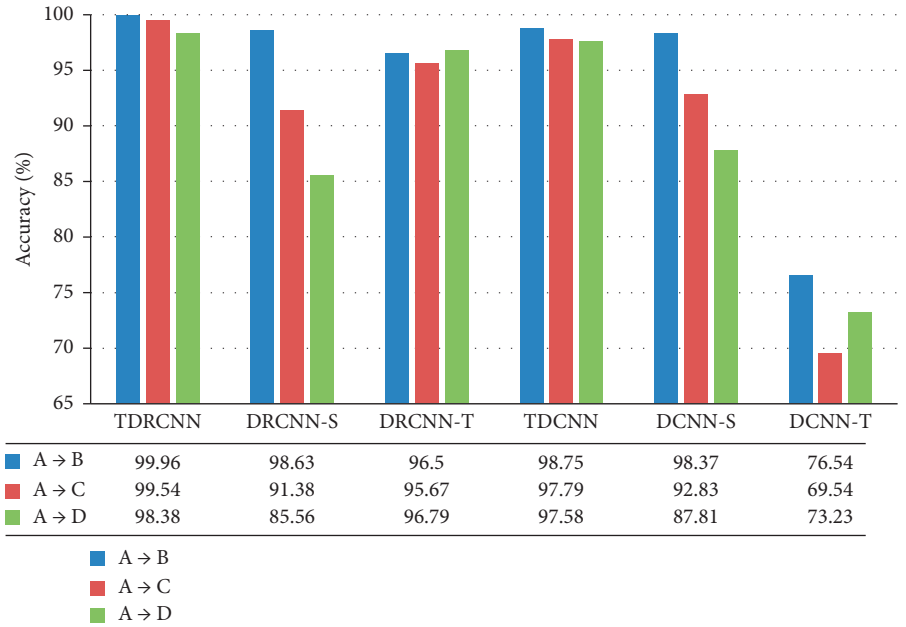


Figure 9: Performance comparison of different methods in the three transfer tasks.

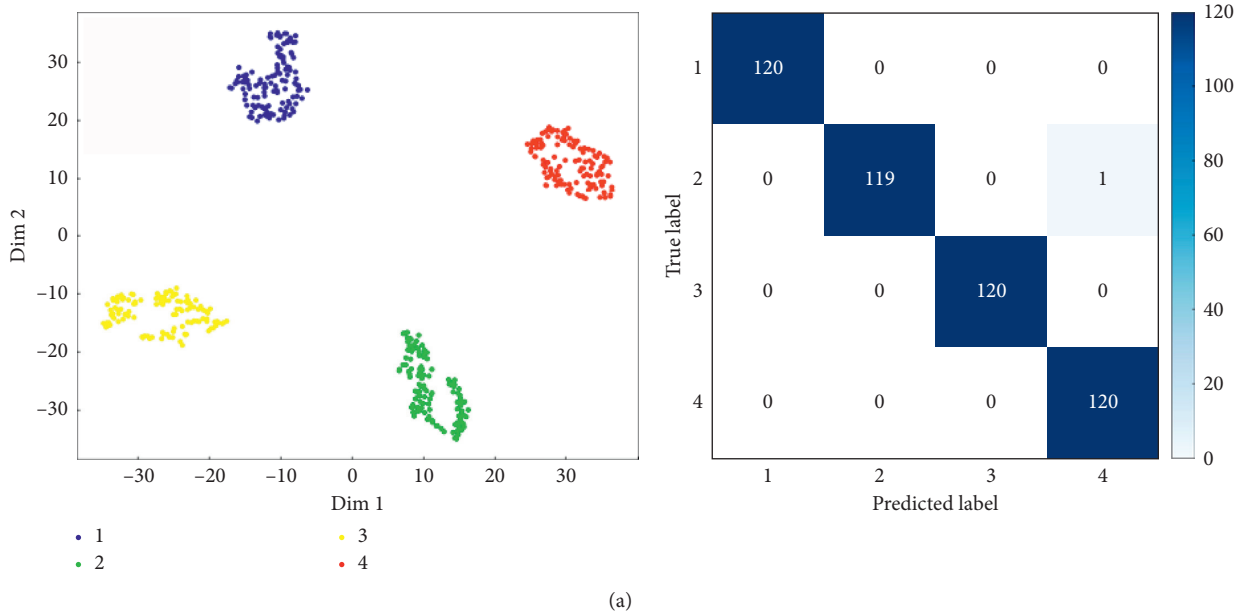
Second, compared with the test results in task A ⟶ C, the accuracies obtained by TDRN, TDCNN, DRN-S, and DCNN-S in task A ⟶ B had different degrees of improvement, whereas those in task A ⟶ D dropped in different degrees. To be specific, as the working condition difference increased, the classification accuracies obtained by those methods mentioned above decreased. This is because the increase in working condition difference makes the domain discrepancy enlarge, resulting in the degeneration of diagnosis performance. Nevertheless, the TDRN was able to maintain very high accuracies in the three tasks. Even in the transfer task with the most working condition difference, i.e., task A ⟶ D, the TDRN could still archive desirable results with an accuracy of 98.38%. This result demonstrated that the advantage of TDRN was more significant in the case of more working condition differences than that in the case of small condition differences.

Third, we can observe that the testing results of DRN-T and DCNN-T in the three tasks basically kept constant, since the training samples and testing samples were selected from the same working condition. And the DRN always achieved better testing results than DCNN in such a case, which showed the advantage of residual learning.

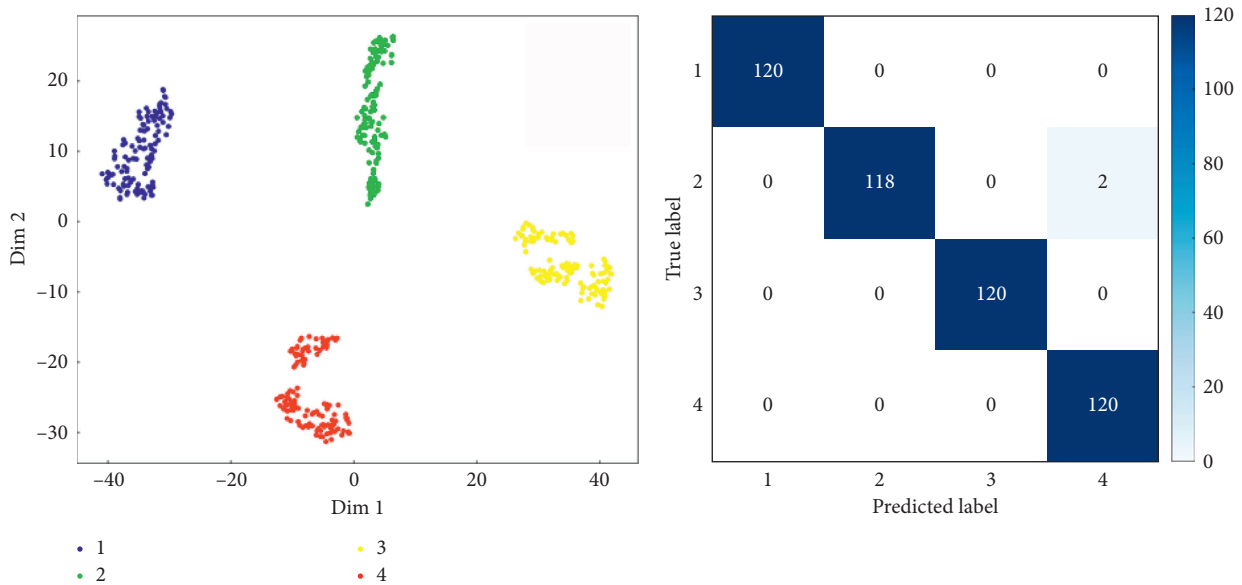
To show more clearly the diagnosis performance of the developed TDRN under variable working condition differences, t-SNE was used to visualize the high-dimensional feature maps/distribution of the test samples extracted from the GAP of TDRN. Figure 10 shows the 2D visualization results and the corresponding confusion matrixes in the three transfer tasks.

As shown in Figure 10(a), the features of testing samples corresponding to task A ⟶ B had the best intraclass compactness and interclass separability, and only one sample of health type 2 was misclassified. The intraclass compactness of testing samples in task A ⟶ C was slightly worse than that corresponding to task A ⟶ B, and there was one more misclassified sample, as observed in Figure 10(b). Task A ⟶ D had the worst intraclass compactness and interclass separability. Although most features of testing samples were very good in terms of intraclass compactness and interclass separability in Figure 10(c), it can be seen that the features of a few samples of three fault types were overlapped. And there were many misclassified samples in the corresponding confusion matrix. These results intuitively demonstrated that the diagnosis performance of the TDRN decreased with the increase in the working condition difference.

Nevertheless, we can see that the feature distributions of testing samples in the three tasks had good intraclass compactness and interclass separability. Moreover, from the corresponding confusion matrixes, it can be seen that the classification results were very well. There were only 8 misclassified samples even in task A ⟶ D with the most

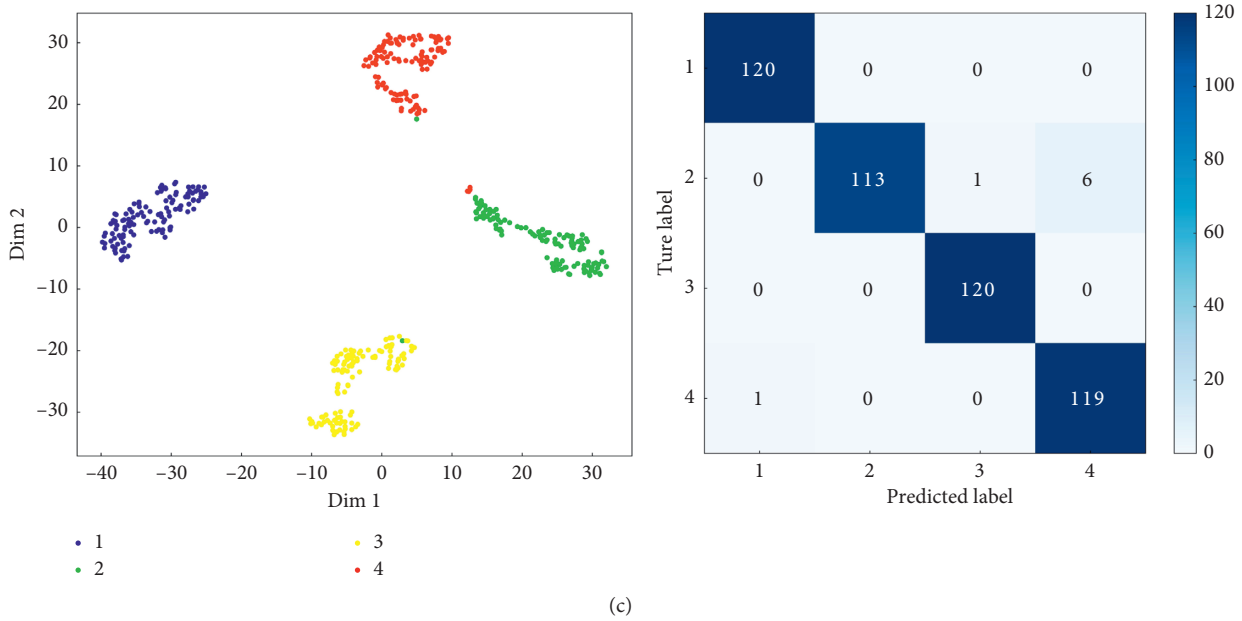


(a)



(b)

Figure 10: Continued.



(c)

Figure 10: Visualization of the learned features and confusion matrixes in three transfer tasks: (a) A⟶B; (b) A⟶C; (c) A⟶D.

working condition difference. These results showed the effectiveness of the developed TDRN.

4.3. Effect of the Number of Training Samples in Target Domain. In this section, the effect of the number of training samples in the target domain on the classification performance was investigated. In the following experiment on transfer task A ⟶ C, the performance of different methods was investigated using 6, 12, 24, 36, and 60 training samples of each class in the target domain, respectively. Figure 11 shows the testing accuracies obtained by different methods with the different number of training samples in the target domain for task A ⟶ C.

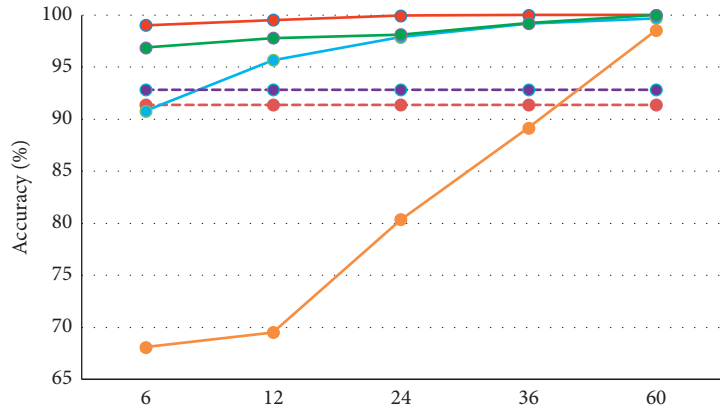
First, it is obvious that the developed TDRN always achieved the highest testing accuracy with identical training samples in the target domain. This further validated the effectiveness and superiority of TDRN with the different number of training samples in the target domain, and the performance improvements achieved by adopting the transfer learning and residual learning were also confirmed.

Second, we can also find that the testing accuracy increased with the rise of the number of training samples in the target domain for all the methods except DRN-S and DCNN-S. In general, the more the training samples, the higher the classification accuracy. If the number of training samples of each class was more than 36, the testing accuracy of 100% can be achieved by the TDRN. And even the DCNN-T can reach more than 98% accuracy when 60 training samples were used.

In addition, the developed TDRN was able to achieve very good performances even with a small number of training samples in the target domain. For instance, the testing accuracy obtained by the TDRN with only 6 training samples of each class was 99.03%. However, the methods without transfer learning were not able to diagnosis bearing health types well with a small number of training samples. Take the DCNN-T as an example, although it performed well with a large number of training samples, its performance degenerated rapidly when the number of training samples became smaller. The testing accuracy obtained by the DCNN-T was only 68.12% when the number of training samples of each class dropped to 6. For the TDCNN, its performance decline was lower than those of the DCNN-T and DRN-T. But the testing accuracy obtained by it was 96.88% when the number of training samples of each class was 6, which was still not good enough. Therefore, the developed TDRN was more robust on the quantity of training data in the target domain.

In short, all these results indicate that, compared with other methods, the performance improvement achieved by the developed TDRN with a limited number of training samples in the target domain was more significant.

4.4. Discussion: Comparison with Other Methods. For conventional intelligent fault diagnosis approaches, feature extraction from the vibration signals is a very important step. FFT (fast Fourier transform) spectral analysis is a one of the commonly used methods for signal preprocessing and feature extraction [38]. There are many feature extraction methods involving FFT spectral analysis. Glowacz et al. [1, 54–57] proposed a series of feature extraction methods based on FFT spectra for mechanical fault diagnosis, such as MSAF-5, MSAF-17-MULTIEXPANDED-FILTER-14, and MSAF-RATIO-24-MULTIEXPANDED-FILTER-8, which have been applied for the fault diagnosis of different kinds of mechanical equipment and obtained very good classification results. This kind of methods need to manually design corresponding features for specific mechanical equipment, and the feature vectors extracted for different equipment are



	6	12	24	36	60
TDRCNN	99.03	99.54	99.95	100.00	100.00
DRCNN-S	91.38	91.38	91.38	91.38	91.38
DRCNN-T	90.78	95.67	97.90	99.19	99.70
TDCNN	96.88	97.79	98.13	99.22	100.00
DCNN-S	92.83	92.83	92.83	92.83	92.83
DCNN-T	68.12	69.54	80.36	89.17	98.54

Figure 11: Testing accuracies obtained by different networks with different numbers of training samples in the target domain for transfer task A⟶C.

not the same. It means that the performance heavily depends on the extracted features. However, the proposed approach can automatically extract features through the deep network model. The advantage is that not only there is no need to extract sensitive handcrafted features for specific equipment but also the classification accuracy is very high. And this approach can effectively use the data in another working conditions to improve the diagnosis performance when the training samples in the target domain are insufficient.

5. Conclusions

In this paper, a new approach for the bearing fault diagnosis under variable working conditions based on STFT and TDRN was proposed. The STFT was employed to obtain TFIs of vibration signals. The TDRN was developed to make a bridge among data from different working conditions. Thus, the proposed approach can realize the machine fault diagnosis under variable working conditions.

The effectiveness and superiority of the proposed approach was validated by experiments conducted on the popular CWRU bearing dataset. The results showed that, by introducing the transfer learning method, the developed TDRN can overcome the domain discrepancy problem. Moreover, it was found that the classification performance of TDRN is better than TDCNN constructed based on the traditional DCNN since the residual learning structure can address the problems of training difficulty and performance degradation in traditional DCNN. Therefore, the proposed approach can obtain better learning ability and higher classification accuracy than those without transfer learning and/or residual learning. Additional experiments were conducted to investigate the effects of some influencing factors, which further verify the effectiveness and superiority of the proposed approach. It was found that the developed TDRN still obtained high classification accuracy under more working condition differences. And even with a very small number of training samples in the target domain, the TDRN still had high classification performance. These results showed that the proposed approach was very suitable for diagnosis under variable working conditions. To sum up, this study clearly demonstrated that the proposed approach has significant potential to be a powerful tool for the machine fault diagnosis under variable working conditions.

However, there are some limitations to the proposed approach. One is that a few labeled target domain data are still needed by the developed TDRN. Future work will focus on achieving fault diagnosis under variable working conditions without labeled data from the target domain. Domain adaptation would be a promising tool to achieve this goal. We can introduce it to the proposed approach. Another limitation of the proposed approach, which is also the problem existing in many deep learning-based methods, is that the diagnosis performance degenerates on imbalanced datasets. In the future, we intend to integrate a generative adversarial network (GAN) into the proposed approach. The GAN can be used to artificially generate fake samples, such that the class distributions can be balanced. In addition, due to the complex working conditions in the industrial scene, there are a lot of noise and other interferences in practical applications. Therefore, in the future, we will also consider collecting the industrial field equipment data and carry out experiments to further verify, analyze, and improve the method proposed in this paper.

Data Availability

Data used in this paper were acquired from the bearing data center of Case Western Reserve University (CWRU) and

web page http://csegroups.case.edu/bearingdatacenter/home (accessed on October 2019).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research work was supported by the National Key Research and Development Program of China (Grant no. 2016YFC0600900), the Yue Qi Distinguished Scholar Project of China University of Mining & Technology (Beijing) (Grant no. 800015Z1145), the National Natural Science Foundation of China (Grant no. U1361127), and the Fundamental Research Funds for the Central Universities (Grant no. 00/800015HJ).

References

[1] A. Glowacz, W. Glowacz, Z. Glowacz, and J. Kozik, "Early fault diagnosis of bearing and stator faults of the single-phase induction motor using acoustic signals," *Measurement*, vol. 113, pp. 1–9, 2018.

[2] T. Berredjem and M. Benidir, "Bearing faults diagnosis using fuzzy expert system relying on an improved range overlaps and similarity method," *Expert Systems with Applications*, vol. 108, pp. 134–142, 2018.

[3] Z. Duan, T. Wu, S. Guo, T. Shao, R. Malekian, and Z. Li, "Development and trend of condition monitoring and fault diagnosis of multi-sensors information fusion for rolling bearings: a review," *International Journal of Advanced Manufacturing Technology*, vol. 96, no. 1–4, pp. 803–819, 2018.

[4] L. Yang and H. Chen, "Fault diagnosis of gearbox based on RBF-PF and particle swarm optimization wavelet neural network," *Neural Computing and Applications*, vol. 31, no. 9, pp. 4463–4478, 2019.

[5] Y. Lei, Z. He, and Y. Zi, "Application of an intelligent classification method to mechanical fault diagnosis," *Expert Systems with Applications*, vol. 36, no. 6, pp. 9941–9948, 2009.

[6] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Gao, and R. X. Gao, "Deep learning and its applications to machine health monitoring," *Mechanical Systems and Signal Processing*, vol. 115, pp. 213–237, 2019.

[7] R. Liu, B. Yang, E. Zio, and X. Chen, "Artificial intelligence for fault diagnosis of rotating machinery: a review," *Mechanical Systems and Signal Processing*, vol. 108, pp. 33–47, 2018.

[8] M. Deepak Sonje, P. Kundu, and A. Chowdhury, "A novel approach for multi class fault diagnosis in induction machine based on statistical time features and random forest classifier," *IOP Conference Series: Materials Science and Engineering*, vol. 225, Article ID 012141, 2017.

[9] X. Zhang, W. Chen, B. Wang, and X. Chen, "Intelligent fault diagnosis of rotating machinery using support vector machine with ant colony algorithm for synchronous feature selection and parameter optimization," *Neurocomputing*, vol. 167, pp. 260–279, 2015.

[10] B. Samanta and C. Nataraj, "Use of particle swarm optimization for machinery fault detection," *Engineering Applications of Artificial Intelligence*, vol. 22, no. 2, pp. 308–316, 2009.

[11] V. T. Tran, B.-S. Yang, F. Gu, and A. Ball, "Thermal image enhancement using bi-dimensional empirical mode decomposition in combination with relevance vector machine for rotating machinery fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 38, no. 2, pp. 601–614, 2013.

[12] A. Youssef, C. Delpha, and D. Diallo, "An optimal fault detection threshold for early detection using Kullback-Leibler Divergence for unknown distribution data," *Signal Processing*, vol. 120, pp. 266–279, 2016.

[13] Z. Li, H. Fang, and M. Huang, "Diversified learning for continuous hidden Markov models with application to fault diagnosis," *Expert Systems with Applications*, vol. 42, no. 23, pp. 9165–9173, 2015.

[14] S. Shao, S. McAleer, R. Yan, and P. Baldi, "Highly accurate machine fault diagnosis using deep transfer learning," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2446–2455, 2019.

[15] J. Wang, Z. Mo, H. Zhang, and Q. Miao, "A deep learning method for bearing fault diagnosis based on time-frequency image," *IEEE Access*, vol. 7, pp. 42373–42383, 2019.

[16] L. Wen, X. Li, and L. Gao, "A transfer convolutional neural network for fault diagnosis based on ResNet-50," *Neural Computing and Applications*, 2019.

[17] X. Guo, L. Chen, and C. Shen, "Hierarchical adaptive deep convolution neural network and its application to bearing fault diagnosis," *Measurement*, vol. 93, pp. 490–502, 2016.

[18] J. Jiao, M. Zhao, J. Lin, and C. Ding, "Deep coupled dense convolutional network with complementary data for intelligent fault diagnosis," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9858–9867, 2019.

[19] D.-T. Hoang and H.-J. Kang, "Rolling element bearing fault diagnosis using convolutional neural network and vibration image," *Cognitive Systems Research*, vol. 53, pp. 42–50, 2019.

[20] G. Jiang, H. He, J. Xie, and X. Ping, "Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 4, pp. 3196–3207, 2019.

[21] W. Sun, R. Zhao, R. Yan, S. Shao, and X. Chen, "Convolutional discriminative feature learning for induction motor fault diagnosis," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 3, pp. 1350–1359, 2017.

[22] L. Jing, M. Zhao, P. Li, and X. Xu, "A convolutional neural network based feature learning and fault diagnosis method for the condition monitoring of gearbox," *Measurement*, vol. 111, pp. 1–10, 2017.

[23] X. Min, L. Teng, X. Lin, L. Liu, and C. W. D. Silva, "fault diagnosis for rotating machinery using multiple sensors and convolutional neural networks," *IEEE/ASME Transactions on Mechatronics*, vol. 23, no. 1, pp. 101–110, 2018.

[24] Z. Minghang, K. Myeongsu, T. Baoping, and P. Michael, "Multi-wavelet coefficients fusion in deep residual networks for fault diagnosis," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 6, pp. 4696–4706, 2019.

[25] K. He, X. Zhang, S. Ren, and S. Jian, "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 2016.

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

[27] M. Zhao, M. Kang, B. Tang, and M. Pecht, "Deep residual networks with dynamically weighted wavelet coefficients for fault diagnosis of planetary gearboxes," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 5, pp. 4290–4300, 2018.

[28] D. Peng, Z. Liu, H. Wang, Y. Qin, and L. Jia, "A novel deeper one-dimensional CNN with residual learning for fault

diagnosis of wheelset bearings in high-speed trains," *IEEE Access*, vol. 7, pp. 10278–10293, 2019.

[29] P. Ma, H. Zhang, W. Fan, C. Wang, G. Wen, and X. Zhang, "A novel bearing fault diagnosis method based on 2-D image representation and transfer learning–convolutional neural network," *Measurement Science and Technology*, vol. 30, no. 5, 10 pages, 2019.

[30] M. J. Hasan and J.-M. Kim, "bearing fault diagnosis under variable rotational speeds using stockwell transform-based vibration imaging and transfer learning," *Applied Sciences*, vol. 8, no. 12, 2018.

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

[32] R. Zhang, H. Tao, L. Wu, and Y. Guan, "Transfer learning with neural networks for bearing fault diagnosis in changing working conditions," *IEEE Access*, vol. 5, pp. 14347–14357, 2017.

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

[34] Z. Bei, H. Bo, and Y. Zhong, "Transfer learning with fully pretrained deep convolution networks for land-use classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 9, pp. 1436–1440, 2017.

[35] H. Kim and B. D. Youn, "A new parameter repurposing method for parameter transfer with small dataset and its application in fault diagnosis of rolling element bearings," *IEEE Access*, vol. 7, pp. 46917–46930, 2019.

[36] X. Ding and Q. He, "Energy-fluctuated multiscale feature learning with deep ConvNet for intelligent spindle bearing fault diagnosis," *IEEE Transactions on Instrumentation and Measurement*, vol. 66, no. 8, pp. 1926–1935, 2017.

[37] P. Cao, S. Zhang, and J. Tang, "Pre-processing-free gear fault diagnosis using small datasets with deep convolutional neural network-based transfer learning," *IEEE Access*, vol. 6, no. 99, pp. 26241–26253, 2017.

[38] M. Qiu, W. Li, Z. Zhu, F. Jiang, and G. Zhou, "fault diagnosis of bearings with adjusted vibration spectrum images," *Shock and Vibration*, vol. 2018, Article ID 6981760, 17 pages, 2018.

[39] W. Li, M. Qiu, Z. Zhu, Bo Wu, and G. Zhou, "Bearing fault diagnosis based on spectrum images of vibration signals," *Measurement Science and Technology*, vol. 27, no. 3, p. 10, 2016.

[40] H. Liu, L. Li, Ma, and Jian, "Rolling bearing fault diagnosis based on STFT-deep learning and sound signals," *Shock and Vibration*, vol. 2016, Article ID 6127479, 12 pages, 2016.

[41] Z. Feng, M. Liang, and F. Chu, "Recent advances in time-frequency analysis methods for machinery fault diagnosis: a review with application examples," *Mechanical Systems and Signal Processing*, vol. 38, no. 1, pp. 165–205, 2013.

[42] Y.-P. Chen, Z.-M. Peng, Z.-H. He, L. Tian, and D.-J. Zhang, "The optimal fractional Gabor transform based on the adaptive window function and its application," *Applied Geophysics*, vol. 10, no. 3, pp. 305–313, 2013.

[43] L. Durak and O. Arikan, "Short-time Fourier transform: two fundamental properties and an optimal implementation," *IEEE Transactions on Signal Processing*, vol. 51, no. 5, pp. 1231–1242, 2003.

[44] Z. K. Peng and F. L. Chu, "Application of the wavelet transform in machine condition monitoring and fault diagnostics: a review with bibliography," *Mechanical Systems and Signal Processing*, vol. 18, no. 2, pp. 199–221, 2004.

[45] Z. Liu, P. You, X. Wei, D. Liao, and X. Li, "High resolution time-frequency distribution based on short-time sparse representation," *Circuits, Systems, and Signal Processing*, vol. 33, no. 12, pp. 3949–3965, 2014.

[46] S. H. Nawab, "Short-time fourier transform," *Advanced Topics in Signal Processing*, vol. 32, no. 2, pp. 289–337, 1988.

[47] L. Xu, C. S. Choy, and Y. W. Li, "Deep sparse rectifier neural networks for speech denoising," in *Proceedings of the 2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, Xi'an, China, September 2016.

[48] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37, pp. 448–456, 2015.

[49] P. Zhou and J. Austin, "Learning criteria for training neural network classifiers," *Neural Computing & Applications*, vol. 7, no. 4, pp. 334–342, 1998.

[50] W. A. Smith and R. B. Randall, "Rolling element bearing diagnostics using the Case Western Reserve University data: a benchmark study," *Mechanical Systems and Signal Processing*, vol. 64-65, pp. 100–131, 2015.

[51] K. A. Loparo, *Bearings Vibration Data Set*, Case Western Reserve University, Cleveland, OH, USA, 2003, http://csegroups.case.edu/bearingdatacenter/pages/download-data-file.

[52] B. Pang, G. Tang, T. Tian, and C. Zhou, "Rolling bearing fault diagnosis based on an improved HTT transform," *Sensors*, vol. 18, no. 4, p. 1203, 2018.

[53] Y. Luo, Y. Wong, M. Kankanhalli, and Q. Zhao, "G-softmax: improving intra-class compactness and inter-class separability of features," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 2, pp. 685–699, 2019.

[54] A. Glowacz, "Acoustic fault analysis of three commutator motors," *Mechanical Systems and Signal Processing*, vol. 133, Article ID 106226, 2019.

[55] A. Glowacz, "Fault detection of electric impact drills and coffee grinders using acoustic signals," *Sensors*, vol. 19, no. 2, p. 269, 2019.

[56] A. Glowacz and W. Glowacz, "Vibration-based fault diagnosis of commutator motor," *Shock and Vibration*, vol. 2018, Article ID 7460419, 10 pages, 2018.

[57] A. Glowacz, "Recognition of acoustic signals of loaded synchronous motor using FFT, MSAF-5 and LSVM," *Archives of Acoustics*, vol. 40, no. 2, pp. 197–203, 2015.