Applied Soft Computing Journal 103 (2021) 107150

Contents lists available at ScienceDirect

Applied Soft Computing Journal

journal homepage: www.elsevier.com/locate/asoc

Deep transfer learning with limited data for machinery fault diagnosis

Te Han [a,b], Chao Liu [a,c,*], Rui Wu [a,b], Dongxiang Jiang [a,b]

[a] Department of Energy and Power Engineering, Tsinghua University, Beijing 100084, China
[b] State Key Laboratory of Control and Simulation of Power System and Generation Equipment, Tsinghua University, Beijing 100084, China
[c] Key Laboratory for Thermal Science and Power Engineering of Ministry of Education, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Article history:
Received 2 June 2020
Received in revised form 18 December 2020
Accepted 25 January 2021
Available online 1 February 2021

Keywords:
Transfer learning
Adversarial domain adaptation
Limited data
Fault diagnosis
Rotating machinery

ABSTRACT

Investigation of deep transfer learning on machinery fault diagnosis is helpful to overcome the limitations of a large volume of training data, and accelerate the practical applications of diagnostic algorithms. However, previous reported methods, mainly including parameter transfer and domain adaptation, still require a few labeled or massive unlabeled fault samples, which are not always available. In general, only extremely limited fault data, namely sparse data (single or several samples), can be obtained, and the labeling is also easy to be processed. This paper presents a novel framework for disposing the problem of transfer diagnosis with sparse target data. In consideration of the unclear data distribution described by the sparse data, the main idea is to pair the source and target data with the same machine condition and conduct individual domain adaptation so as to alleviate the lack of target data, diminish the distribution discrepancy as well as avoid negative transfer. More impressive, the issue of label space mismatching can be appropriately addressed in our network. The extensive experiments on two case studies are used to verify the proposed method. Comprehensive transfer scenarios, i.e., diverse working conditions and diverse machines, are considered. The thorough evaluation shows that the proposed method presents superior performance with respect to traditional transfer learning methods.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

In modern manufacturing and industries, fault diagnosis technique is the crucial mean of enhancing the safety and availability of equipment [1]. With the rapid development of the Internet of Things (IoT), massive amounts of data are generated by a multitude of devices in a real-time manner. In this context, advanced fault diagnosis technique expects to exploit the shared expert knowledge, online analyze and dispose the monitoring data, as well as reducing the manual effort required for any failure analysis [2]. The past decade has witnessed the rise and development of data-driven fault diagnosis researches, due to their high data processing efficiency and advantages in terms of automation and intelligence.

The traditional researches about data-driven fault diagnosis tend to learn a classification model from large amounts of labeled fault data to have sufficient generalization capacity [3,4]. The reported works focus specifically on advanced signal processing methods for feature extraction, and machine learning algorithms for automatic diagnosis [5]. A variety of popular methods, such as empirical mode decomposition (EMD), time-frequency analysis, artificial neural network (ANN), support vector machine (SVM), deep learning, etc., have been widely studied and employed within this field [6–14]. Despite numerous success stories, most of these works are based on the assumption that the training and testing data follow the same distribution. This means that the deployment scenarios of the well-trained model should be identical to the ones where the training data are collected. However, the data to be diagnosed in industry tasks are generally monitored from different working conditions or even different-type machines [15–17]. For this reason, the data distributions of the training and testing sets are not consistent in most of actual diagnosis scenarios, and the distribution discrepancy will easily lead to the drop of generalization capacity of the well-trained model. Moreover, retraining a diagnostic model from scratch for the deployment scenarios is even impossible, since the labeled data on a large enough scale is too labor-consuming to collect and annotate [18,19].

Consequently, recent works start focusing on some emerging study directions and trying to address the above-mentioned issues. Zhu et al. used the capsule network structure to modify the standard convolutional neural network (CNN), and achieved strong generalization ability for bearing fault diagnosis [20]. Yang et al. presented a novel fault diagnosis scheme toward unseen working conditions by introducing the center loss to traditional

* Corresponding author at: Department of Energy and Power Engineering, Tsinghua University, Beijing 100084, China.
E-mail address: cliu5@tsinghua.edu.cn (C. Liu).

https://doi.org/10.1016/j.asoc.2021.107150
1568-4946/© 2021 Elsevier B.V. All rights reserved.

deep learning model [21]. Zheng et al. proposed a multi-source domain generalization method, which can learn the general representation of the diagnosis structure on Grassmann manifold [22]. The diagnostic experiments on unseen bearing faults demonstrated the generalization ability of the proposed method. Liao et al. developed a deep semi-supervised learning framework to adversarially train the labeled and unlabeled machine data [23]. The generalization diagnosis tasks of the transmission and bearing faults evaluated the effectiveness of the proposed method. Liu et al. used the auxiliary tasks to improve the diagnosis performance of target tasks under a multi-task learning framework [24]. The shared features for different tasks can be fully utilized in this manner.

Among the methods explored in machinery fault diagnosis, transfer learning, which reuses the learnt fault knowledge from the existing tasks (source domain) to facilitate the diagnosis of the new but similar tasks (target domain) [15,18], receives growing concern in nearly three years. Most of the reported methods about transfer fault diagnosis can be categorized into two branches: parameter transfer-based fine-tuning [25,26] and domain adaptation [27–33]. Parameter transfer-based fine-tuning feeds a few labeled target data to fine-tune a network, whose parameters are initialized with the values from the pre-trained network in source domain. Domain adaptation mainly considers to align the feature distributions between source and target domains, so that the trained classifier by the labeled source data can be generalized to target task. As domain adaptation approaches generally assume that massive fault data are unlabeled in target domain, most of the domain adaptation methods essentially belong to unsupervised transfer learning flow.

In real-world applications, the monitored machine data are exceedingly imbalanced, and most of them belong to the health category [33]. The limited data containing the mechanical fault information are scattered in the historical database with low density [34]. In a more extreme case, only sparse data (as low as one or several samples for each fault category) can be obtained and labeled. This is a challenging problem where the aim is to diagnose each fault category of target machine identified by a single or several target samples, along with the learnt knowledge in source domain. The popular transfer learning methods, whether parameter transfer-based fine-tuning or domain adaptation, still have their limitations in such engineering background. To solve this problem, the paper provides a novel fault diagnosis approach for addressing the problem of sparse data. Our approach embeds the transfer learning into an end-to-end trainable deep network. Both the plenteous source data and sparse target data are utilized so as to make it possible for adapting the network to target diagnostic tasks instead of overfitting. To exploit the discriminative information present in the category tags to avoid negative transfer in the process of domain adaptation, multiple domain discriminators are designed for learning domain-invariant features within each corresponding fault category.

The remaining parts of this paper are organized as follows. The related works are introduced in Section 2. Then Section 3 presents the preliminaries. The proposed method and the experiments are described in Sections 4 and 5 respectively. Section 6 gives the results. And Section 7 summarizes this paper.

2. Related works

The researches about machinery fault diagnosis have been significantly advanced by deep learning recently, due to the great ability of feature learning and classification. Various deep models, such as auto-encoder [26], deep belief network (DBN) [4], recurrent neural network (RNN) [3], and especially CNN [35,36], have been investigated and modified to achieve the state-of-the art performance for fault diagnosis cases. However, the impressive diagnostic results come only when the training and testing data follow the same data distribution.

Given the practical requirements of cross-domain fault diagnosis, the transfer learning based diagnosis researches are emerging, typically by leveraging the off-the-shelf data with rich supervision to facilitate target model fine-tuning or adaptation. Since this work is designed in the deep learning framework, only the deep transfer learning methods are reviewed. The well-known parameter transfer-based fine-tuning assumes that few labeled target data are available. Zhong et al. presented a deep transfer diagnosis method for gas turbine by using 340 target samples, which are composed of 268 health samples and 72 fault samples for all categories [25]. He et al. investigated to fine-tune a pre-trained deep auto-encoder network with 80 target fault samples in gearbox fault diagnosis [26]. To avoid overfitting with only scarce target data, the focus concerned in this process are mainly identified as two problems, i.e., reducing the number of trainable parameters in the deep model, and developing more efficient fine-tuning algorithms. However, it is still challenging or even prohibitive to simply fine-tune the deep model in the such cases where there are only extremely limited target data. Using one or several samples for each fault category to fine-tune the pre-trained model from source domain, especially deep neural network, will undoubtedly lead to overfitting.

In the presence of distribution discrepancy between source and target domains, domain adaptation methods aim at learning domain-invariant feature representation. Maximum mean discrepancy (MMD) and adversarial training are two most efficient approaches in deep domain adaptation [37]. MMD explicitly measures the domain discrepancy, which is appended to the loss function and minimized during network training. Han et al. calculated the MMD in both marginal distribution and conditional distribution for domain adaptation [15]. The extensive diagnosis cases in wind turbine, bearing and gearbox verify the effectiveness. Yang et al. conducted the MMD minimization in each convolutional layer and fully-connected layer of a deep CNN [31]. The multi-layer domain adaptation yields significant results in the transfer diagnosis case from laboratory bearings to real-case bearings. Adversarial training incorporates an additional domain discriminator to distinguish the source or target features, and encourages the domain confusion in a minimax game so that learning domain-invariant feature representation. Adversarial training is proved to be effective to align the marginal distribution of machinery fault data from sensors at different positions [38]. Zhang et al. presented an adversarial training induced domain adaptation method for the transfer diagnosis of bearing under different working conditions [39]. Despite the success, these reported works generally require a large number of target samples with abundant machine fault characteristics during the adaptation process, while ignoring the fact that the fault samples in target domain are sparse. The sparse data with very little of intra-class variations of machine fault condition could not exactly describe the data distribution of target domain. Moreover, the large data distribution shift and different label space across domains will easily cause negative transfer during unsupervised domain adaptation. Fig. 1 gives the illustrations of the two cases. (1) The discrepancy of working conditions and machine structures between domains may render a large data distribution shift. As a result, the unsupervised adaptation process is prone to the false alignment of the different fault categories, as intuitively illustrated in Fig. 1(a). (2) The source and target domains are usually assumed to share an identical label space in most of existing works. Nevertheless, the machine conditions of interest are unseen in target domain. The categories between the source task and the target task are considerably different,

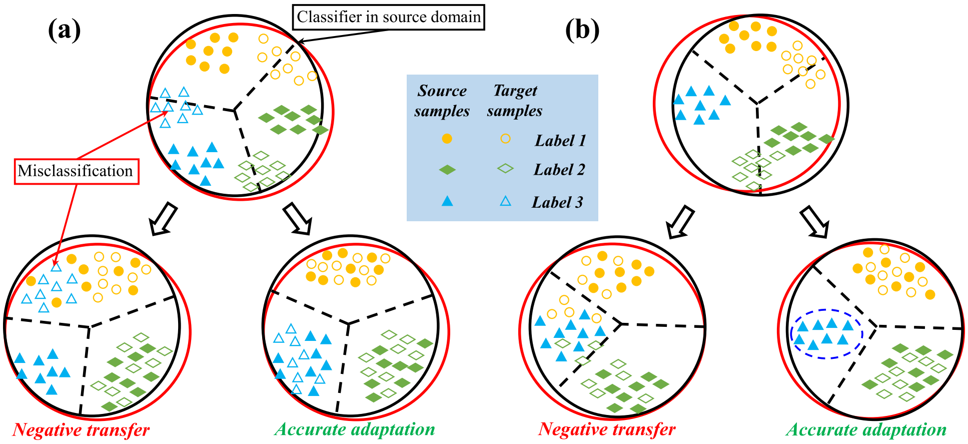


Fig. 1. Illustrations of the negative transfer and accurate feature distribution alignment for domain adaptation methods. In the first case (a), the target data of labels 3 are falsely aligned to the source data of label 1 due to the large distribution shift. In the second case (b), there are 3 fault categories in source domain while 2 fault categories in target domain. The outlier source labels (label 3) will cause negative transfer.

especially for cross-machine transfer diagnosis tasks. The false alignment caused by the outlier source category is shown in Fig. 1(b) [40,41]. Since most of domain adaptation approaches are designed in the unsupervised manner, these works restrict the use of the category information of the sparse data to find similar categories and induce a more accurate distribution adaptation.

Toward an efficient approach for the sparse target data, a hybrid method by simultaneously training the model with the little supervision and learning domain-invariant features with adversarial training is proposed to reduce the overfitting risk and hold the diagnosis accuracy. Compared with the previous works, the main novelties and contributions of this study are listed as follows:

(1) Most of transfer diagnosis works still require a few labeled target data or large amounts of unlabeled target data. This work moves forward to the scenario of sparse fault data (as low as single or several samples for each machine condition). The category information of the sparse data are fully exploited to boost the discrimination ability of classifier as well as learning transferable features from sufficient source data.
(2) A multiple adversarial domain adaptation method is presented to consider the two practical issues in machine fault diagnosis, i.e., improving the precision of feature distribution adaptation with a large dataset shift, and considering the partial transfer learning to cope with the issue of mismatching of diagnostic label spaces.
(3) More comprehensive transfer diagnosis experiments are designed to demonstrate the effectiveness of the proposed method, including the transfer tasks across different working conditions and machines. Especially, in the cross-machine diagnostic case, the working conditions, machines and fault label spaces of the two bearing datasets are all different, which are closer to the actual industrial applications.

3. Preliminaries

3.1. Problem definition

The machine fault diagnosis via transfer learning is investigated in this work, where limited data per machine condition are available. First, a large-scale labeled fault dataset $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ is assumed to be accessed. Herein, $\mathcal{D}_s$ is denoted as source domain. $x_i^s \in X_s$ is the source sample. $y_i^s \in Y_s$ is the label of corresponding machine condition. X_s denotes the feature space with a certain distribution and Y_s is the label space. Besides, the target dataset containing a tiny labeled sample set is defined as $\mathcal{D}_t = \{(x_i^t, y_i^t)\}_{i=1}^{N_t}$. Similarly, $x_i^t \in X_t$ and $y_i^t \in Y_t$. Due to the domain discrepancy $P_s(X_s) \neq P_t(X_t)$, the model trained on source dataset cannot be tailored for target domain. Under these settings, transfer learning aims to use the relevant fault information of $\mathcal{D}_s$ to assist in learning a diagnostic model for target task with the insufficient data of $\mathcal{D}_t$.

The problem defined thus far is generally considered as the parameter transfer-based fine-tuning or supervised domain adaptation. This study is especially concerned with the issue that the number of available fault data in target domain, N_t, is extremely sparse (as low as one or several samples per machine condition). Hereafter, the formulated problem is also referred to as sparse target data. Obviously, the challenge in this scenario is that the sparse samples capture very little of intra-class variations of each machine fault condition. Moreover, considering the practical applications, the machine health conditions, namely label spaces, are generally different for source and target domains. In this work, the source label space is assumed to be large enough to subsume the labels of target domain, i.e., $Y_t \subseteq Y_s$. It is reasonable since we can readily simulate various machine health conditions in historical diagnosis tasks or customized fault experiments.

3.2. Convolutional neural network

CNN is adopted as the base deep model in this work, following an end-to-end manner [42]. A typical CNN is composed of two modules: the feature extractor G_f and the label classifier G_y. The convolutional layers and pooling layers are generally used in G_f. The convolutional operation is to calculate the feature activation of the output of the previous layer by using a set of convolutional kernels. Then, the convolutional results are processed by the nonlinear activation function, i.e., rectified linear unit (ReLU) in this work. This process is described as:

$$O_j^k = ReLU(\sum_i^n W_{ij}^k * O_i^{k-1} + b_i^k) \tag{1}$$

where O_i^{k-1} is the ith feature map in $(k-1)$th layer, O_j^k is the jth feature map in kth layer, W_{ij}^k denotes the kernel connecting ith and jth feature maps, b_i^k denotes the basis and $(*)$ represents the convolutional operation [43].

After convolutional and pooling operations, the extracted features are flattened to a 1-D vector for nonlinearly mapping in classifier G_y. The G_y typically consists of fully-connected layers. This process is denoted as:

$$O^k = ReLU(W^k O^{k-1} + b^k) \tag{2}$$

where O^{k-1} and O^k are the features in the $(k-1)$th and kth layers, respectively, W^k denotes the weight matrix between the $(k-1)$th and kth layers, b^k denotes the basis in the kth layer.

To indicate the predicted probabilities for each machine health condition, the final outputs are further processed by softmax function. Overall, the training process of the CNN can be represented as the optimization task [15]:

$$\arg\min_{\theta} \sum_i \mathcal{L}_y(G_y(G_f(x_i)), y_i) \tag{3}$$

where $\{x_i, y_i\}$ is the input sample with corresponding label, θ is the parameters collection of a CNN, $\mathcal{L}_y$ represents the loss between the estimated label $G_y(G_f(x_i))$ by the trainable network and the real label y_i.

3.3. Adversarial domain adaptation

Adversarial domain adaptation [44] has been successfully adopted in many machine fault diagnosis researches recently. On the basis of traditional CNN, a domain discriminator G_d, similar to G_y, is added, so that a minmax two-player game is introduced between G_f and G_d. G_d is trained to distinguish source features or target features, while G_f tries to learn domain-invariant features for confusing the G_d. The objective function of adversarial domain adaptation typically contains label loss of source sample and the domain loss of samples in both source and target domains. It is formulated as:

$$\begin{aligned}\mathcal{L}(\theta_f, \theta_y, \theta_d) = & \frac{1}{N_s} \sum_{x_i \in \mathcal{D}_s} \mathcal{L}_y(G_y(G_f(x_i)), y_i) \\ & - \frac{\lambda}{N_s + N_t} \sum_{x_i \in \mathcal{D}_s \bigcup \mathcal{D}_t} \mathcal{L}_d(G_d(G_f(x_i)), d_i)\end{aligned} \tag{4}$$

where θ_f, θ_y, θ_d are the parameter collection of G_f, G_y, and G_d respectively, d_i means the domain label of the sample [41,44].

The θ_f is learned to maximize the domain loss, while the θ_y is trained to minimize it. Besides, the θ_f and θ_y are updated to minimize the label loss of source samples. The optimization with respect to (4) is described as follows:

$$(\widehat{\theta}_f, \widehat{\theta}_y) = \arg\min_{\theta_f, \theta_y} \mathcal{L}_0(\theta_f, \theta_y, \widehat{\theta}_d) \tag{5}$$

$$and \quad \widehat{\theta}_d = \arg\max_{\theta_d} \mathcal{L}_0(\widehat{\theta}_f, \widehat{\theta}_y, \theta_d) \tag{6}$$

Conventional adversarial network has been proved to be effective for learning domain-invariant features in many transfer diagnosis cases, where sufficient unlabeled target data are provided, the working conditions are different but the label spaces are the same across domains.

4. Proposed method

To extend the previous transfer diagnosis framework to the scenario of sparse target fault data, a novel transfer network (as shown in Fig. 2) by considering to pair the target data with source data is presented to make use of category information of target data. The sparse labeled data can not only be used for supervised fine-tuning, but also to induce a more convincing feature distribution alignment via a multiple adversarial domain adaptation [40]. The adversarial training within the data pair of same machine condition is conducive to utilize the transferable features from the relevant source data and learn the adapted feature representation for target tasks instead of overfitting. More importantly, the negative adaptation caused by the large distribution shift and label space mismatching is significantly prevented in this manner.

4.1. Proposed network

Motivated by conventional adversarial network, the proposed network consists of three main modules: the feature extractor G_f, the diagnostic classifier G_y and multiclass domain discriminators $G_d^k, k = 1, 2, \dots, K$. Derived from the Refs. [15,28,43], the detailed architecture and parameters of the three modules are shown in Fig. 3. Both the massive source samples $\{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ and sparse target samples $\{(x_i^t, y_i^t)\}_{i=1}^{N_t}$ are fed into the feature extractor G_f for achieving abstract feature representation $f = G_f(x)$. Then, the extracted features are sent to the classifier G_y and multiple domain discriminators $G_d^k, k = 1, 2, \dots, K$, respectively. Overall, the objective of the proposed network contains two parts: supervised fine-tuning and multiple adversarial domain adaptation.

Once the network parameters are initialized, the source samples along with the sparse target samples are used to train the classifier G_y. The discriminative ability of the classifier for diagnosing the machine health conditions is captured in this stage. The cross-entropy is utilized as the loss function of G_y in this work. The classification loss of the G_y is denoted as [44]:

$$\mathcal{L}_y = \frac{1}{N_s} \sum_{x_i \in \mathcal{D}_s} \mathcal{L}_y(G_y(G_f(x_i)), y_i) + \frac{1}{N_t} \sum_{x_j \in \mathcal{D}_t} \mathcal{L}_y(G_y(G_f(x_j)), y_j) \tag{7}$$

To avoid overfitting with the very little target supervision, the adversarial training via feature extractor G_f and multiple domain discriminators $G_d^k, k = 1, 2, \dots, K$, is conducted in the second stage. Pairing the data from different domains but same machine condition, essentially compensates for the indistinct feature distribution described by the sparse target data in the process of distribution alignment. The domain-invariant feature representation for the same machine condition can be learned with the induction of adversarial training within the data pair. Consequently, trained classifier with both the source and target supervision possesses enough generalization capability in target data. The domain loss of all K domain discriminators is defined as:

$$\mathcal{L}_d = \frac{1}{N_s + N_t} \sum_{k=1}^{K} \sum_{x_i \in \mathcal{D}_s \bigcup \mathcal{D}_t} \mathcal{L}_d^k(G_d^k(G_f(x_i)), d_i) \tag{8}$$

where G_d^k is the kth domain discriminator associated with label k, $\mathcal{L}_d^k$ denotes the binary cross entropy (BCE) loss for G_d^k, and d_i is the domain label (the domain label 0 for source sample while the domain label 1 for target sample). The BCE loss $\mathcal{L}_d^k$ for G_d^k is formulated as:

$$\mathcal{L}_d^k(G_d^k(G_f(x_i)), d_i) = -d_i log(G_d^k(G_f(x_i))) - (1 - d_i) log(1 - G_d^k(G_f(x_i))) \tag{9}$$

One point worth emphasizing is that, for the issue of mismatching of diagnostic labels between domains, i.e., $Y_t \subseteq Y_s$, the domain discriminators are only built for the shared labels while the outlier source labels are filtered out.

In the training process, the feature extractor G_f and classifier G_y are trained to minimize the classification loss $\mathcal{L}_y$ for accurately diagnosing machine conditions. Meanwhile, the G_f is trained to maximize the domain loss $\mathcal{L}_d$ for generating more shared feature representation, while domain discriminators $G_d^k, k = 1, 2, \dots, K$ are updated to minimize the $\mathcal{L}_d$. Integrating the above two parts,

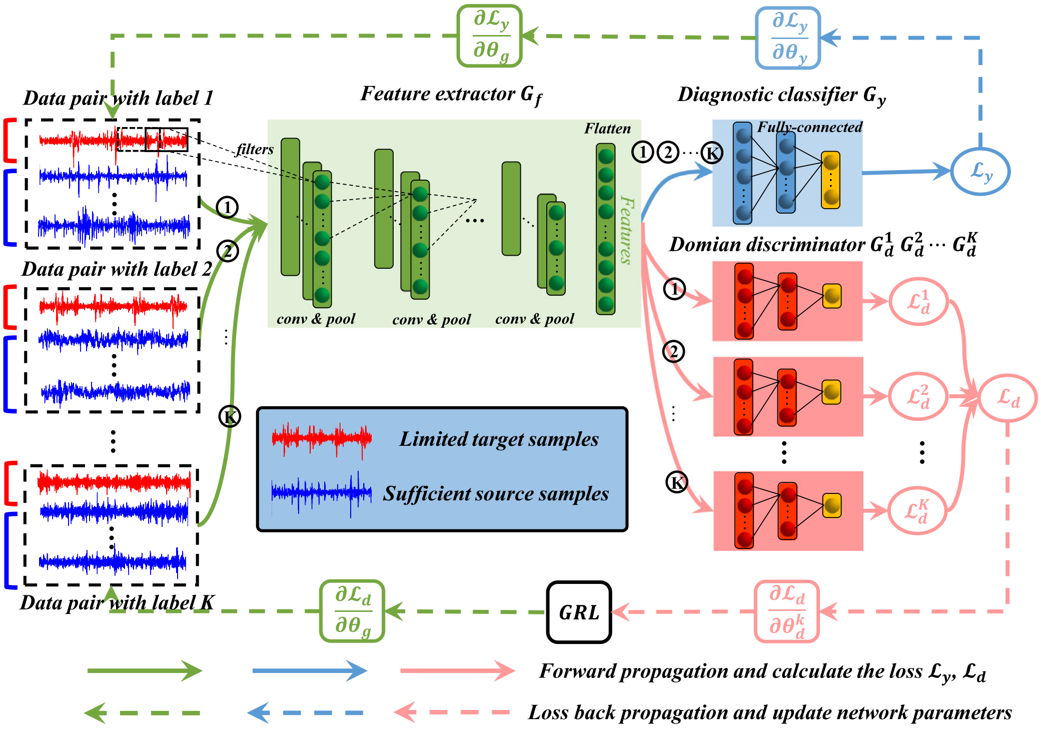

$\frac{\partial \mathcal{L}_y}{\partial \theta_g}$
$\frac{\partial \mathcal{L}_y}{\partial \theta_y}$
Data pair with label 1
Feature extractor G_f
Diagnostic classifier G_y
filters
Flatten
Fully-connected
$\mathcal{L}_y$
Features
conv & pool
conv & pool
conv & pool
Data pair with label 2
Domian discriminator G_d^1 $G_d^2 \cdots G_d^K$
$\mathcal{L}_d^1$
$\mathcal{L}_d^2$
$\mathcal{L}_d^K$
$\mathcal{L}_d$
Limited target samples
Sufficient source samples
Data pair with label K
$\frac{\partial \mathcal{L}_d}{\partial \theta_g}$
GRL
$\frac{\partial \mathcal{L}_d}{\partial \theta_d^k}$
Forward propagation and calculate the loss $\mathcal{L}_y$, $\mathcal{L}_d$
Loss back propagation and update network parameters

Fig. 2. The illustration of the proposed network. The data pairs are fed to network for forward propagation and calculating the loss $\mathcal{L}_y$ and $\mathcal{L}_d$. The domain discriminator G_d^k only processes the extracted features of data pair k, $k = 1, \dots, K$. Then, the each loss is back-propagated to update the parameters of corresponding module. GRL is the gradient reversal layer to take the gradient from the domain discriminators, but changes the sign before back-propagating to the feature extractor G_f [44].

the entire optimization objective is to seek the $\widehat{\theta}_f$, $\widehat{\theta}_y$ and $(\widehat{\theta}_d^1 \dots \widehat{\theta}_d^K)$ so that

$$\widehat{\theta_f} = \arg\{\min_{\theta_f} \mathcal{L}_y(\theta_f, \widehat{\theta_y}), \max_{\theta_f} \mathcal{L}_d(\theta_f, \widehat{\theta}_d^k|_{k=1}^K)\} \tag{10}$$

$$\widehat{\theta_y} = \arg\min_{\theta_y} \mathcal{L}(\widehat{\theta_f}.\theta_y) \tag{11}$$

$$\text{and} \quad (\widehat{\theta}_d^1, \dots, \widehat{\theta}_d^K) = \arg\min_{\theta_d^1,\dots,\theta_d^K} \mathcal{L}_d(\widehat{\theta_f}, \theta_d^k|_{k=1}^K) \tag{12}$$

where the θ_f, θ_y and $(\theta_d^1, \dots, \theta_d^K)$ are the network parameters of G_f, G_y and G_d^k, $k = 1, \dots, K$, respectively.

4.2. Main procedure of proposed method

The main procedure for the proposed transfer diagnosis method involves: (1) pre-training, (2) data augmentation, (3) transfer learning and (4) deployment of the adapted model. The source samples $\{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ are firstly used to train the baseline network, namely, the feature extractor G_f and the diagnostic classifier G_y for parameter initialization. Before executing the transfer learning, data augmentation is another efficient way to tackle the challenge of sparse data. Two approaches, i.e., adding Gaussian noise and amplitude shifting, are implemented to the target machine samples. In this manner, more target samples, which are similar but different to the limited target samples, can be artificially generated to facilitate the transfer learning between the sufficient source samples and the extremely limited target samples, instead of overfitting. Mathematically, it is formulated as:

$$\widetilde{x}_i = \alpha_{scal} x_i + \alpha_{gaus} G, G \sim N(0, 1) \tag{13}$$

where x_i is the original sample, $\widetilde{x}_i$ is the augmented sample, the factor α_{scal} is to scale the amplitude of the sample and the factor α_{gaus} is to adjust the power of Gaussian noise G [34].

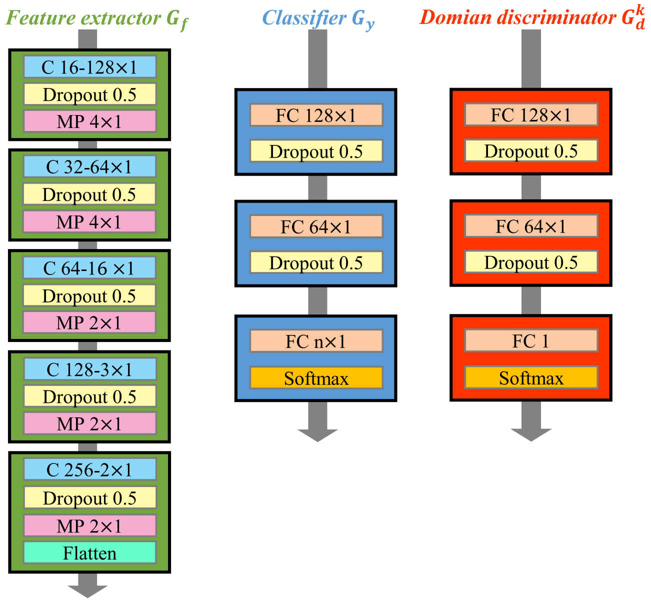


Fig. 3. The designed architecture and key parameters of the proposed network. Therein, C denotes convolutional layer. For instance 16-128 × 1 means 16 kernels with the size of 128 × 1. MP means the max-pooling layer and FC represents the fully-connected layer. The other parameters, such as stride and padding, are all set to default in PyTorch platform.

To quantitatively measure the power of the additional Gaussian noise, the signal noise ratio (SNR) is used, which can be defined as follows:

$$SNR_{dB} = 10 \log_{10}(\frac{P_{signal}}{P_{noise}}) \tag{14}$$

where P_{signal} and P_{noise} are the power of original signal and additional Gaussian noise, respectively [43]. In this work, the α_{gaus} is set to control the SNR in the range of [5 dB, 20 dB]. And the factor α_{scal} is randomly selected in the range of [0.9, 1.1]. After data augmentation, the number of target data, N_t, increases to match the number of source data, N_s. It is worth noting that the useful information hidden in the original target data is still limited, and the data augmentation scheme could not generate additional but meaningful fault features.

After preparations, the proposed network is trained with both the source samples and the augmented target data. With respect to the optimization objective in (9), the parameters of network are updated as follows:

$$\theta_f \leftarrow \theta_f - \mu\left(\frac{\partial \mathcal{L}_y}{\partial \theta_f} - \frac{\partial \mathcal{L}_d}{\partial \theta_f}\right) \tag{15}$$

$$\theta_y \leftarrow \theta_y - \mu \frac{\partial \mathcal{L}_y}{\partial \theta_y} \tag{16}$$

$$\theta_d^k \leftarrow \theta_d^k - \mu \frac{\partial \mathcal{L}_d}{\partial \theta_d^k} \tag{17}$$

where μ is the learning rate. The classical stochastic gradient descent (SGD) algorithm is employed for network training. And finally, the trained feature extractor G_f and the diagnostic classifier G_y are extracted and deployed in target tasks. The performance of trained model is evaluated by other testing samples in target domain. Overall, the training process is summarized in Algorithm 1.

Algorithm 1 Training of the proposed transfer diagnosis method

```
Input: Sufficient source data $\mathcal{D}_s = \{x_i^s, y_i^s\}_{i=1}^{N_s}$, sparse target data $\mathcal{D}_t = \{x_i^t, y_i^t\}_{i=1}^{N_t}$, the baseline network $G_f$ and $G_y$.
Output: Learnt network $G_f$ and $G_y$.
  Pre-train the baseline network with $\mathcal{D}_s$ to initialize $\theta_f$ and $\theta_y$.
  Augment the target data $\mathcal{D}_t$ according to (13)
  repeat
    Randomly partition mini-batch $\mathcal{D}_s^m$ and $\mathcal{D}_t^m$ from $\mathcal{D}_s$ and $\mathcal{D}_s$.
    for each mini-batch $\mathcal{D}_s^m$ and $\mathcal{D}_t^m$ do.
      Calculate the classification loss $\mathcal{L}_y$ by (7).
      Calculate the multiple domain loss $\mathcal{L}_d$ by (8).
      Update the network parameters $\theta_f$ and $\theta_y$ according to (10)–(11).
      Update the domain classifiers $(\theta_d^1, \ldots \theta_d^K)$ according to (12).
    end for
  until convergence or reach maximum iterations
```

5. Experiments descriptions

Two sets of transfer experiments are conducted in this study. The diagnosis of rotor-related faults and bearing faults are two concerns within this field. A wind turbine dataset containing rotor-related faults is firstly used to design the transfer experiments across diverse working conditions. Then, two bearing datasets are employed for verifying the effectiveness of proposed method in cross-machine transfer diagnosis. The partial transfer learning is considered for the problem of label space mismatching in this case. Meanwhile, to demonstrate the advantages of proposed method in the scenario of sparse fault data, the popular transfer learning approaches, i.e., supervised fine-tuning and traditional domain adaptation, are also introduced for comparison studies.

Table 1
List of fault bearing selected in PU dataset for method verification.

Code[a]	KI04	KI14	KI16	KA04	KA15	KA22
Location	IF	IF	IF	OF	OF	OF
Damage	Pitting	Pitting	Pitting	Pitting	Indentations	Pitting
Extent[b]	1	1	3	1	1	1

[a] The bearing codes are used in [48].
[b] The extent is determined by length of damage. Details can be found in [48].

5.1. Experiments description

5.1.1. Wind turbine fault dataset

The experiments are carried out in a wind turbine platform, whose diagram can be seen in Fig. 4. The experimental platform mainly consists of three parts: the mechanical system of a direct drive wind turbine, the wind tunnel and the signal acquisition system. According to the surveys about the health monitoring of wind turbine [45–47], the faults in rotor-systems, such as misalignment, bearing faults, faults of rotor hub and blade, etc., contribute to the majority of downtime. And thus, ten machine states (labels 0–9) are simulated in the rotor systems, including health, front bearing pedestal loosening, back bearing pedestal loosening, rolling element fault of front bearing, inner race fault of front bearing, outer race fault of front bearing, misalignment in horizontal direction, misalignment in vertical direction, variation in airfoil of blades and yaw fault, respectively. These experiments can basically contain the common failure modes of a wind turbine. The detailed experiment schemes for the fault conditions are illustrated in Fig. 5.

The wind turbine is driven by the wind source generated by a wind tunnel. The wind speed ranges from 5.8 m/s to 11.5 m/s. The monitored data under wind speeds 5.8 m/s, 6.9 m/s, 8.0 m/s and 11.5 m/s (load conditions 0–3) are used to design transfer experiments. Fig. 6 shows the time waveform of vibration data under load condition 3. Specifically, the data under each load condition corresponds to a domain, and three transfer experiments are considered, i.e., load 0→load 3, load 1→load 3, load 2→load 3. For instance, in load 0→load 3, the load 0 is in source domain, where sufficient samples are available, and load 3 is in target domain, where only sparse samples are provided.

As shown in Fig. 4, in this case study, the vibration data from the bearing pedestal are used to monitor the health condition of the machine. The acceleration sensors are both installed in the front and back bearing pedestals. The sampling frequencies of the acceleration sensors are 20 kHz. The collected signals are segmented into a set of samples with the length of 4096 data points. Generally, the size of the samples should be set to the length of several periods of impulse or interest in the raw signal. Furthermore, to fuse the multi-source data, each of the samples from the front bearing pedestal and the back bearing pedestal are connected as a single sample. And thus, each sample in the first case study contains 8196 data points. In the three transfer tasks, there are 200 samples for each category in source domain. The number of target samples are set to 1, 2 and 4 for each category in turn. Another 100 testing samples for each category in target domain (totally 1000) are used for performance evaluation.

5.1.2. Bearing fault datasets

The first bearing fault dataset is from Case Western Reserve University (CWRU) [5]. Artificial faults by using electric spark machining are introduced to outer race, inner race and ball. Four bearing states, namely, health (H), inner race fault (IF), outer race fault (OF), ball fault (BF) (labels 0–3) are simulated in the experiments. The fault diameter is 0.007 inches. The data under four operation conditions, i.e., 0 HP, 1 HP, 2 HP, and 3 HP are served

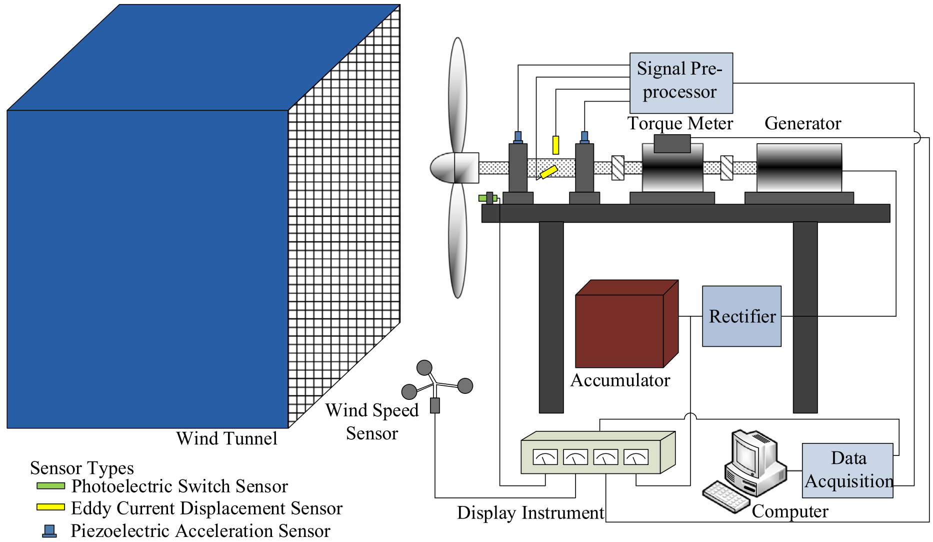
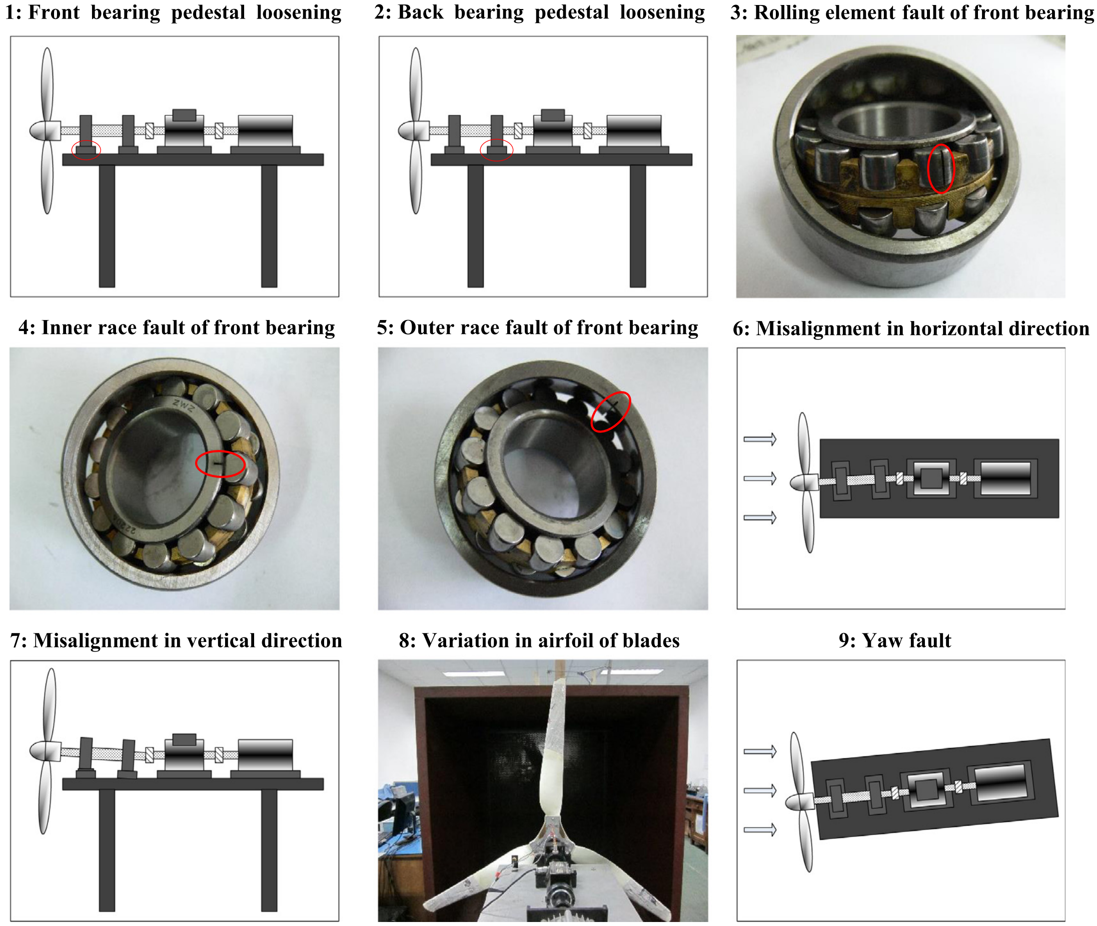


Fig. 4. The illustration of wind turbine platform.


1: Front bearing pedestal loosening

2: Back bearing pedestal loosening

3: Rolling element fault of front bearing

4: Inner race fault of front bearing

5: Outer race fault of front bearing

6: Misalignment in horizontal direction

7: Misalignment in vertical direction

8: Variation in airfoil of blades

9: Yaw fault

Fig. 5. The illustrations of the faults in different positions.

as source dataset. There are 1000 samples for each category. The sampling frequency is 12 kHz and each sample contains 4096 points.

The second dataset is collected from the modular test rig in Paderborn University (PU), as shown in Fig. 7. The accelerated lifetime tests are performed to generate real damage in bearing. Only the IF and OF are observed at the end of these tests. The data of three bearings with inner ring fault (KI04, KI14 and KI16), three bearings with outer ring fault (KA04, KA15 and KA22) and three health bearings (K001, K003 and K004) are selected as the target data. The descriptions of used fault bearings are listed in Table 1. The four operating conditions are all considered, and the details can be found in [48]. The sampling frequency in the experiments is 64 kHz. To accord with the sampling frequency in

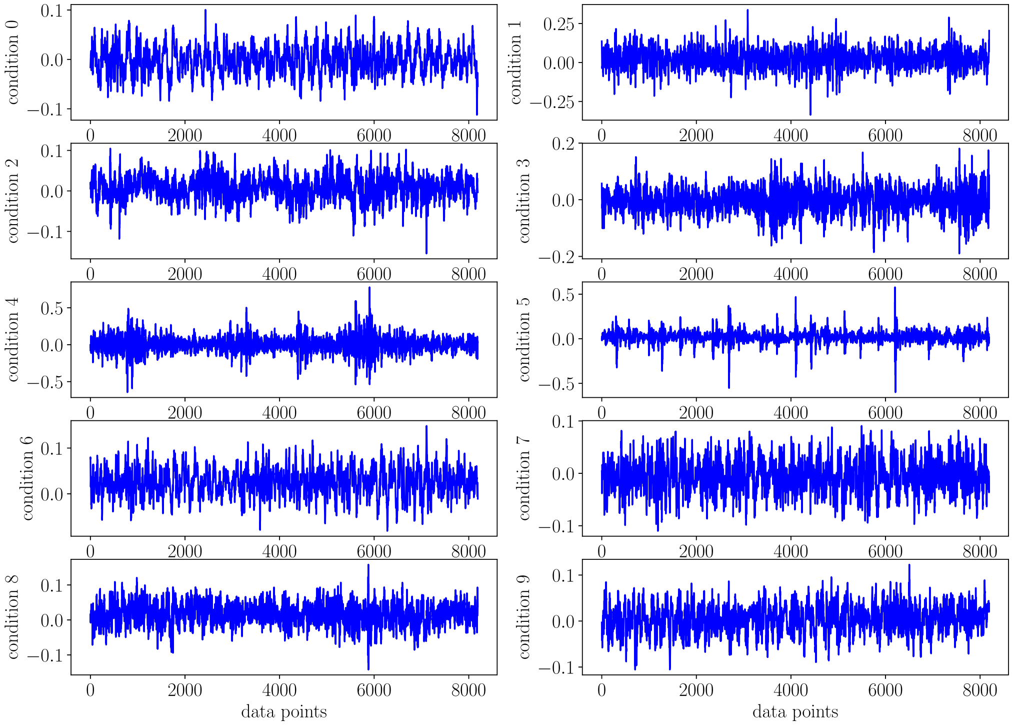


Fig. 6. The time waveform of vibration data under load condition 3.

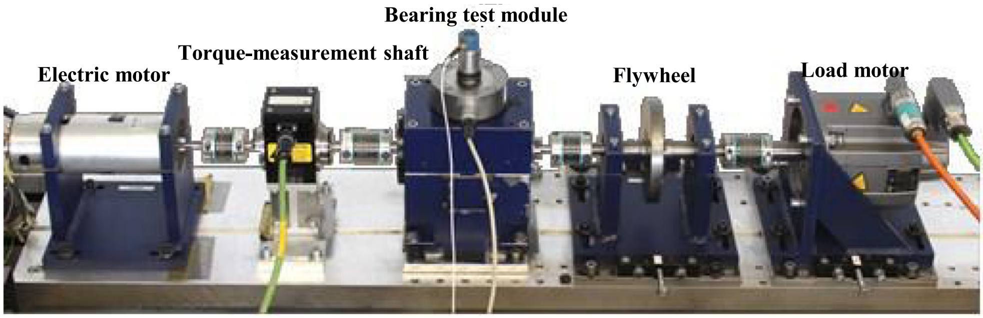


Fig. 7. The modular test rig in Paderborn University.

Table 2
Descriptions of designed bearing fault datasets.

Dataset	Bearing types	Bearing states	No. of samples
A(CWRU)	SKF 6205-2RS	H,IF,OF,BF	1000 × 4
B(PU:K001,KI04,KA04)	FAG, MTK and IBU/IBB 6203	H,IF,OF	1×3, 2×3 and 4 × 3
C(PU:K003,KI14,KA15)			
D(PU:K004,KI16,KA22)			

source dataset, the collected signals are down-sampled with the frequency of 12 kHz. Each sample consists of 4096 points.

The vibration signals of two bearing datasets are shown in Fig. 8. As shown in Table 2, the bearing data in PU are further partitioned into three datasets, and three transfer tasks are designed, i.e., A→B, A→C and A→D. Similar to the first case, the number of labeled target sample is 1×3, 2×3 and 4 × 3 in turn. After implementing the transfer learning procedure, the trained model is verified by other 100 testing samples for each bearing state. More notably, although CWRU bearing data has been a benchmark for algorithm validation in recent years, the fault characteristics in this dataset are distinct and easily diagnosable [5,49]. Different from the researches whose training and validation processes are both performed in the CWRU bearing data, our goal is to build a cross-machine transfer diagnosis framework by utilizing the typical fault data in CWRU dataset to facilitate the diagnosis of PU bearing data. This framework contributes to the diagnostic model learning on the condition of lacking typical fault data in practical industry issues.

5.2. Comparison methods

This study focuses on the issue of machine fault diagnosis with sparse fault data. And it is entirely impossible to only use single or several samples to train a deep network from scratch. Consequently, the methods in deep transfer learning framework are employed for convincing comparison studies. According to the previous works, two parameter transfer-based fine-tuning methods [25,26] and two unsupervised domain adaptation methods, i.e., deep transfer network (DTN) [15,50] and domain adversarial neural network (DANN) [27,44] are selected. The first two methods use the target data to fine-tune the network trained with sufficient source data. And the two domain adaptation methods send both the labeled source data and unlabeled target data into the network for aligning the feature distributions across domains via MMD and adversarial training respectively. All the methods use the same baseline network architectures, as shown in Fig. 3.

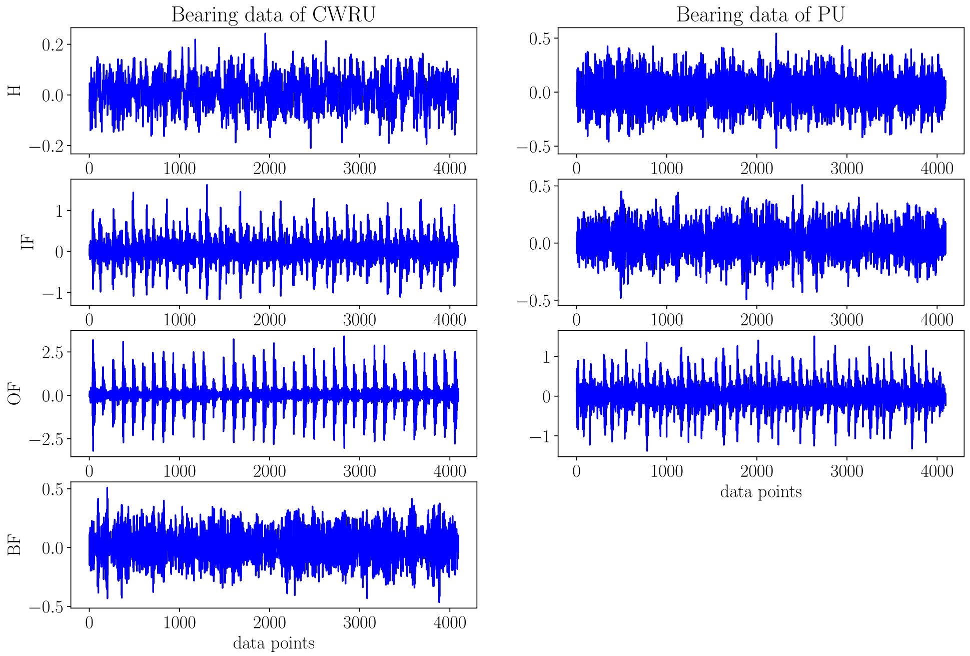


Fig. 8. The time waveform of vibration signal from two bearing datasets.

The implement details are determined according to original references to achieve the optimal performances. The details are introduced as follows:

Parameter transfer I: The pre-trained network parameters in source domain are transferred to a target network, which has the same architecture as the source network. The top classification layer is customized to target task. And the parameters of the whole network are fine-tuned with the augmented target samples.

Parameter transfer II: On the basis of the strategy of parameter transfer I, the parameters of preceding convolutional layers, namely θ_f, are fixed (also referred as frozen layers) to reduce the number of trainable parameters so that avoid overfitting.

DTN: The domain discrepancy of the last hidden fully-connected layer is measured by MMD. And a joint distribution adaptation term is appended to loss function for minimizing the discrepancy. It is an unsupervised method without using the label information of augmented target data.

DANN: The unsupervised adversarial training is applied for domain adaptation. However, there is only one discriminator to distinguish the source data from augmented target data.

6. Results

6.1. Results of wind turbine experiments

The extensive results with respect to diverse methods are listed in Table 3, which reveal the following observations.

(1) All the methods perform better when the working conditions are closer across domains. For instance, in the first task (load 0→load 3), DANN achieves the accuracy of 78.1% via 4 target samples, whereas the corresponding accuracy is 90.0% in the third task (load 2→load 3). The results are reasonable, since the data under closer working conditions are more transferable.

(2) The proposed method outperforms the comparative methods in all the experiments, proving that pairing the data across domains by exploiting the class information present in sparse target data can effectively boost the performance during transfer process. The superiority of proposed method with respect to simple parameter transfer-based fine-tuning and unsupervised domain adaptation is fully demonstrated in the scenario of sparse target data.

(3) For a fair comparison, the parameter transfer-based fine-tuning methods and the proposed method all utilize the labels of sparse target data. Differing from the strategy of fine-tuning, the proposed method adapts the pre-trained network to target domain by jointly using the source and target data, which largely prevents the impact of overfitting to some degree. For instance, in the first task with single target sample, compared with the accuracy 58% of fine-tuning I, the proposed method yields a surprising 23.7% improvement.

(4) The DTN and DANN ignore the use of the labels of sparse target data. Accordingly, this unsupervised manner may suffer from two problems in this case, i.e., unclear target distribution described by sparse data and large feature distribution shift across domains. As a reference, the results with more target samples are shown in Fig. 9. In the third transfer task, the interesting fact is that the performance of DANN is improved from 86.4% to 93.6%, when the number of target samples of each category increases from 1 to 200. Similar phenomenon can be found for DTN. These results indicate that the unsupervised domain adaptation requires

Table 3
Results (%) of wind turbine fault experiments.

Tasks	No. of samples	Parameter transfer I	Parameter transfer II	DTN	DANN	Proposed method
load0→load3	1[a]	58.0±8.6[b]	73.5±2.6	73.6±1.1	76.8±2.2	81.7±1.1
	2	78.5±4.3	76.6±2.7	73.8±0.6	77.8±1.9	85.1±2.2
	4	85.1±3.3	79.8±1.7	73.8±1.6	78.1±2.9	90.6±2.1
load1→load3	1	61.2±4.6	78.8±3.3	79.1±2.5	78.7±3.5	83.4±2.9
	2	76.8±5.3	80.7±2.9	80.3±2.9	80.8±1.6	88.4±2.3
	4	84.8±4.3	81.1±2.7	81.2±2.5	80.9±2.4	92.1±1.3
load2→load3	1	73.5±5.9	81.0±2.7	87.8±4.0	86.4±4.8	91.4±2.6
	2	75.8±5.7	82.6±2.9	91.3±3.9	89.1±2.3	94.8±1.4
	4	86.1±4.5	85.6±1.2	91.6±2.7	90.0±1.5	95.3±1.1

[a] It is the number of target samples for each machine category during training.
[b] The results are the average of ten tests, where the training and testing samples are randomly selected from the dataset. The average accuracies and standard deviations are presented.

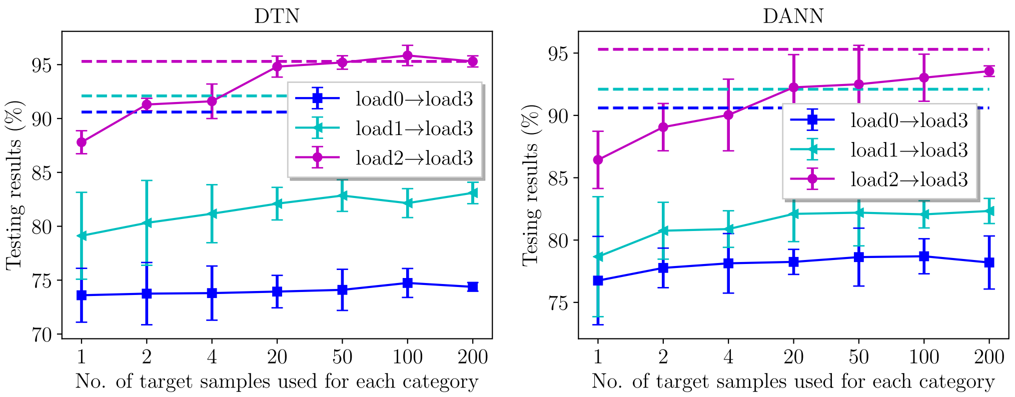


Fig. 9. The results of DTN and DANN versus number of target samples used for each category. The dotted lines represents the reference accuracies of proposed method by using 4 samples for each machine health category.

Table 4
Average computing time of different methods in task load 0→load 3 with 4 target samples.

Methods	Computing time (s/epoch)	
	Training stage	Testing stage
Parameter transfer I	0.43	0.08
Parameter transfer II	0.37	0.09
DTN	7.38	0.08
DANN	3.55	0.07
Proposed method	10.78	0.07

Table 5
Results (%) of bearing fault experiments.

Tasks	No. of samples	Parameter transfer I	Parameter transfer II	DTN	DANN	Proposed method
A→B	1	58.8±9.6	46.0±5.3	41.1±11.8	29.0±7.0	79.9±8.6
	2	64.1±7.7	52.0±7.0	49.8±1.3	35.0±7.6	84.7±7.6
	4	78.2±7.6	58.7±4.5	49.8±0.6	46.3±9.0	92.4±7.6
A→C	1	63.8±9.7	50.1±7.3	15.5±3.5	26.4±4.7	79.6±5.1
	2	72.5±8.7	56.8±9.0	20.5±4.2	27.8±4.7	86.7±6.7
	4	78.0±5.7	62.0±3.3	33.4±0.1	23.5±2.3	93.8±5.4
A→D	1	68.7±14.8	57.0±13.3	50.2±11.6	41.8±11.8	88.2±3.1
	2	79.9±11.7	63.3±10.9	53.7±6.6	42.9±3.7	89.8±3.5
	4	90.3±1.1	69.3±6.1	58.1±0.6	46.8±6.9	96.2±4.2

sufficient target data for feature distribution alignment. Furthermore, in the first task, there are no salient increases of accuracy when providing more target samples. The working condition load 0 in source domain is far away from the target load 4, as a consequence, there will be a large distribution discrepancy between the source and target data, implying that the false alignments, namely the negative transfer may easily occur.

To further observe the separability of the learnt features, the t-SNE algorithm is applied to the outputs of the last hidden fully-connected layer. Due to the pages limitation, the results of parameter transfer I, DTN, DANN and the proposed method are shown in Fig. 10. The parameter transfer method only utilizes the augmented target samples to train the pre-trained network, whose fault characteristics are essentially limited. As a result, the fine-tuned network is easily prone to overfitting and the discriminative boundary is obscure. For the two unsupervised methods, it is clear that the target samples of categories 2 and 8 are almost entirely aligned to the false category in source domain. The results intuitively illustrate the negative transfer due to the large discrepancy across working conditions. For comparison, the proposed method can achieve a relatively accurate distribution adaptation while discriminate different machine conditions.

The deep learning models used in this work have the multilayer structure and over millions of trainable parameters, and thus have special requirement in terms of computing resources. All the experimental tests are executed in the software environment of Windows 10 Edu (64 bits) and PyTorch 0.3.1. The hardware of the used computer mainly includes the Intel quad-core CPU i7-4790 3.6 GHz, the DDR3-1600 16G memory and the GPU of Nvidia GeForce GTX 1080 Ti with 11G RAM. Table 4 lists the average computing times of the different methods in task load 0→load 3. It should be emphasized that, as the guidance to stop training before overfitting, the early stopping is used in the training stage of all the methods. Herein, to make a comparative analysis, Table 4 gives the computing time of one epoch. It can be observed that, accompanied by superior diagnostic performance, the computing complexity is also clearly increased in the proposed method during training stage. After transfer learning, all the transferred models are deployed for diagnosing the testing samples. The same benchmark model, namely the feature extractor G_f and classifier G_y, are used in different methods. The testing times for the different methods are much the same, which are less than 0.1 s. Once a reliable deep diagnosis model is built, the deployment and real-time surveillance are promising in industry application.

6.2. Results of bearing experiments

Table 5 reports the results of the bearing experiments. The similar phenomenon to the first case study can be observed here. The proposed method works well in all experiments. Conversely, the results of DTN and DANN are quite unacceptable, indicating that the large discrepancy in this cross-machine diagnostic task causes severe negative transfer. Furthermore, the feature visualization in Fig. 11 illustrates the occurred negative transfer. The target samples of category 2 (OF) and source samples of category 3 (BF) are incorrectly aligned together in DTN. Otherwise, the adaptation results for the samples of category 1 and 2 indicate the negative transfer in DANN. Owing to the multiple domain adaptation scheme, the distributions of source and target samples are correctly adapted, and the proposed method can effectively solve the problem of label spaces mismatching. For parameter transfer-based fine-tuning methods, the performance is improved significantly, when the number of target samples per category increases from 1 to 4. Compared with our method, the fine-tuning strategies still require more samples to achieve more excellent results, meaning that this simple strategy with feeding data is sub-optimal. Table 6 gives the computing times of different methods in task A→D. By revisiting the analysis of wind turbine

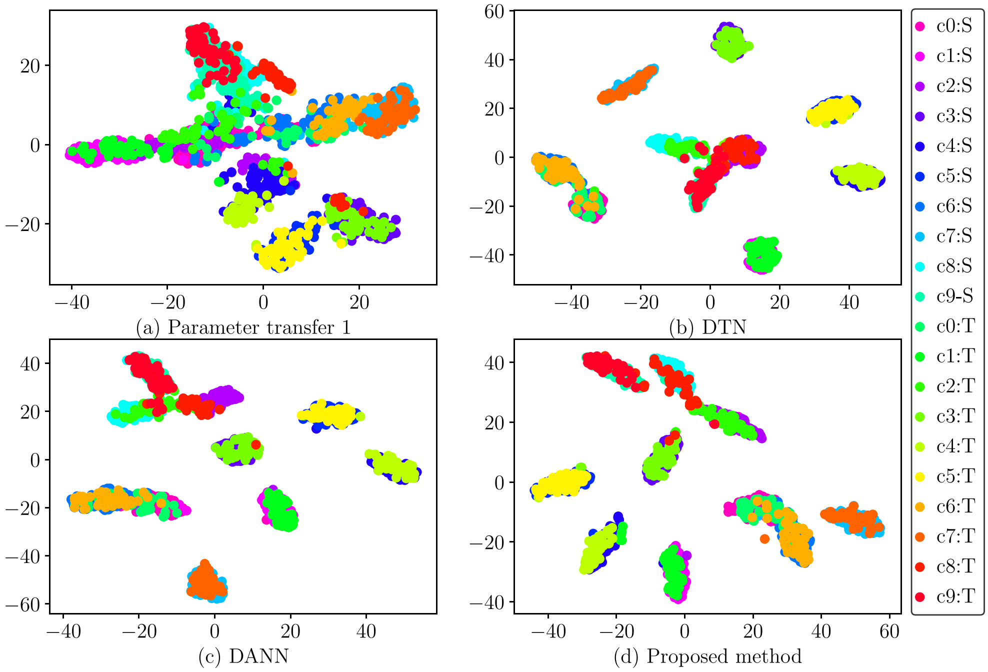


Fig. 10. The feature visualization in the task load 0→load 3 with 4 target samples. For instance, c0:S represents the samples of category 0 in source domain, and c0:S corresponds to the ones in target domain.

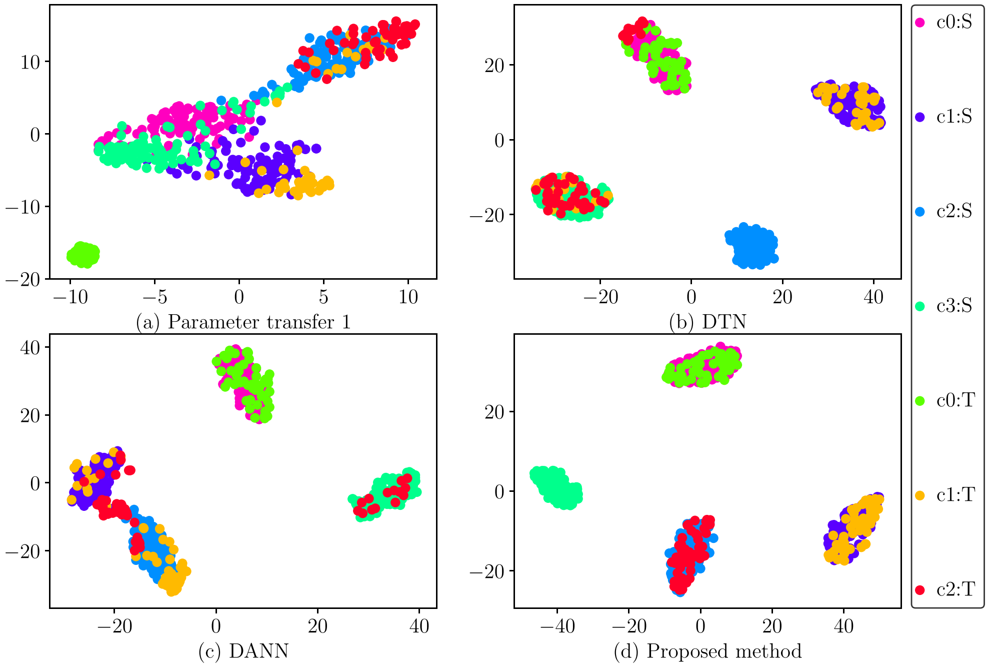


Fig. 11. The feature visualization in the task A→D with 4 target samples.

transfer experiments, similar conclusions can be reached from these results.

7. Conclusions and future work

Considering the practical issue of extremely limited fault data (single or several samples) in machinery fault diagnosis, this paper proposed a novel network by simultaneously conducting the supervised classification and multiple adversarial domain adaptation to improve the performance of deep transfer learning. Specifically, the adversarial training using the data pair of the same category is capable of generating a more precise distribution adaptation regardless of massive target data. Moreover, the challenge of large data shift caused by domain discrepancy and label space mismatching can be effectively tackled in the proposed method. In the first set of wind turbine diagnostic experiments, compared with the parameter transfer I, parameter transfer II, DTN and DANN, the proposed method achieves the 13.7%, 9.2%, 7.8%, 7.1% performance enhancements for the average diagnostic accuracy of all the tests. Similarly, in the transfer diagnostic experiments of two sets of bearings, the proposed method significantly outperforms the four comparative

Table 6
Average computing time of different methods in task A→D with 4 target samples.

Methods	Computing time (s/epoch)	
	Training stage	Testing stage
Parameter transfer I	0.60	0.10
Parameter transfer II	0.55	0.15
DTN	5.65	0.11
DANN	5.25	0.10
Proposed method	6.77	0.10

methods. The improvements of the average diagnostic accuracy are 15.2%, 30.7%, 46.6% and 52.4%, respectively. The extensive analysis demonstrates the effectiveness and superiority of the proposed network as well as the limitations of traditional parameter transfer-based fine-tuning approaches and unsupervised domain adaptation approaches.

Beyond the transfer learning between sufficient source data and limited target data, another common challenge is that no any fault data can be obtained in the target domain, but a precise diagnostic model is still needed to be built for unseen target tasks. Consequently, how to only leverage the mechanical fault data in source domain to learn a domain-invariant and generalized feature representation is significant in this application scenario. And our future work will be devoted to investigate this new research topic, i.e., domain generalization methods, in machinery fault diagnosis.

CRediT authorship contribution statement

Te Han: Conceptualization, Methodology, Software, Writing - original draft. **Chao Liu:** Data curation, Software, Validation. **Rui Wu:** Investigation, Writing - review & editing. **Dongxiang Jiang:** Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 11572167, 51174273 and 71731008), the National Key R&D Program of China (Grant No. 2018YFB1306100) and the Shuimu Tsinghua Scholar project. The authors would like to express their sincere gratitude to Mr. Shaohua Li for his contributions on the acquisition of experimental data. All authors have read and agreed to the published version of the manuscript.

References

[1] M. Cerrada, R.-V. Sánchez, C. Li, F. Pacheco, D. Cabrera, J.V. de Oliveira, R.E. Vásquez, A review on data-driven fault severity assessment in rolling bearings, Mech. Syst. Signal Process. 99 (2018) 169–196.

[2] S. Khan, T. Yairi, A review on the application of deep learning in system health management, Mech. Syst. Signal Process. 107 (2018) 241–265.

[3] Z. An, S. Li, J. Wang, X. Jiang, A novel bearing intelligent fault diagnosis framework under time-varying working conditions using recurrent neural network, ISA Trans. 100 (2020) 155–170.

[4] H. Shao, H. Jiang, F. Wang, Y. Wang, Rolling bearing fault diagnosis using adaptive deep belief network with dual-tree complex wavelet packet, ISA Trans. 69 (2017) 187–201.

[5] Y. Li, X. Wang, S. Si, S. Huang, Entropy based fault classification using the case western reserve university data: A benchmark study, IEEE Trans. Reliab. 69 (2) (2020) 754–767.

[6] T. Han, D. Jiang, Q. Zhao, L. Wang, K. Yin, Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery, Trans. Inst. Meas. Control 40 (2018) 2681–2693.

[7] W. Mao, W. Feng, X. Liang, A novel deep output kernel learning method for bearing fault structural diagnosis, Mech. Syst. Signal Process. 117 (2019) 293–318.

[8] K. Kaplan, Y. Kaya, M. Kuncan, M.R. Minaz, H.M. Ertunç, An improved feature extraction method using texture analysis with lbp for bearing fault diagnosis, Appl. Soft Comput. 87 (2020) 106019.

[9] L. Bai, Z. Han, J. Ren, X. Qin, Research on feature selection for rotating machinery based on supervision kernel entropy component analysis with whale optimization algorithm, Appl. Soft Comput. 92 (2020) 106245.

[10] Y. Chen, G. Peng, Z. Zhu, S. Li, A novel deep learning method based on attention mechanism for bearing remaining useful life prediction, Appl. Soft Comput. 86 (2020) 105919.

[11] H. Zhu, J. Cheng, C. Zhang, J. Wu, X. Shao, Stacked pruning sparse denoising autoencoder based intelligent fault diagnosis of rolling bearings, Appl. Soft Comput. 88 (2020) 106060.

[12] F.B. Abid, M. Sallem, A. Braham, Robust interpretable deep learning for intelligent fault diagnosis of induction motors, IEEE Trans. Instrum. Meas. 69 (6) (2020) 3506–3515.

[13] S.R. Saufi, Z.A.B. Ahmad, M.S. Leong, M.H. Lim, Gearbox fault diagnosis using a deep learning model with limited data sample, IEEE Trans. Ind. Inf. 16 (10) (2020) 6263–6271.

[14] H. Shao, J. Cheng, H. Jiang, Y. Yang, Z. Wu, Enhanced deep gated recurrent unit and complex wavelet packet energy moment entropy for early fault prognosis of bearing, Knowl.-Based Syst. 188 (2020) 105022.

[15] T. Han, C. Liu, W. Yang, D. Jiang, Deep transfer network with joint distribution adaptation: A new intelligent fault diagnosis framework for industry application, ISA Trans. 97 (2020) 269–281.

[16] T. Han, C. Liu, L. Wu, S. Sarkar, D. Jiang, An adaptive spatiotemporal feature learning approach for fault diagnosis in complex systems, Mech. Syst. Signal Process. 117 (2019) 170–187.

[17] Z. He, H. Shao, L. Jing, J. Cheng, Y. Yang, Transfer fault diagnosis of bearing installed in different machines using enhanced deep auto-encoder, Measurement 152 (2020) 107393.

[18] X. Li, Y. Hu, M. Li, J. Zheng, Fault diagnostics between different type of components: A transfer learning approach, Appl. Soft Comput. 86 (2020) 105950.

[19] H. Lv, J. Chen, T. Pan, Z. Zhou, Hybrid attribute conditional adversarial denoising autoencoder for zero-shot classification of mechanical intelligent fault diagnosis, Appl. Soft Comput. 95 (2020) 106577.

[20] Z. Zhu, G. Peng, Y. Chen, H. Gao, A convolutional neural network based on a capsule network with strong generalization for bearing fault diagnosis, Neurocomputing 323 (2019) 62–75.

[21] Y. Yang, J. Yin, H. Zheng, Y. Li, M. Xu, Y. Chen, Learn generalization feature via convolutional neural network: A fault diagnosis scheme toward unseen operating conditions, IEEE Access 8 (2020) 91103–91115.

[22] H. Zheng, R. Wang, Y. Yang, Y. Li, M. Xu, Intelligent fault identification based on multisource domain generalization towards actual diagnosis scenario, IEEE Trans. Ind. Electron. 67 (2) (2020) 1293–1304.

[23] Y. Liao, R. Huang, J. Li, Z. Chen, W. Li, Deep semisupervised domain generalization network for rotary machinery fault diagnosis under variable speed, IEEE Trans. Instrum. Meas. 69 (10) (2020) 8064–8075.

[24] Z. Liu, H. Wang, J. Liu, Y. Qin, D. Peng, Multi-task learning based on lightweight 1dcnn for fault diagnosis of wheelset bearings, IEEE Trans. Instrum. Meas. (2020) http://dx.doi.org/10.1109/TIM.2020.3017900.

[25] S. Zhong, S. Fu, L. Lin, A novel gas turbine fault diagnosis method based on transfer learning with cnn, Measurement 137 (2019) 435–453.

[26] Z. He, H. Shao, X. Zhang, J. Cheng, Y. Yang, Improved deep transfer auto-encoder for fault diagnosis of gearbox under variable working conditions with small training samples, IEEE Access 7 (2019) 115368–115377.

[27] X. Li, W. Zhang, Q. Ding, X. Li, Diagnosing rotating machines with weakly supervised data using deep transfer learning, IEEE Trans. Ind. Inf. 16 (3) (2020) 1688–1697.

[28] J. Jiao, M. Zhao, J. Lin, C. Ding, Classifier inconsistency based domain adaptation network for partial transfer intelligent diagnosis, IEEE Trans. Ind. Inf. 16 (9) (2020) 5965–5974.

[29] W. Mao, Y. Liu, L. Ding, A. Safian, X. Liang, A new structured domain adversarial neural network for transfer fault diagnosis of rolling bearings under different working conditions, IEEE Trans. Instrum. Meas. (2020) http://dx.doi.org/10.1109/TIM.2020.3038596.

[30] X. Li, W. Zhang, Q. Ding, J.-Q. Sun, Multi-layer domain adaptation method for rolling bearing fault diagnosis, Signal Process. 157 (2019) 180–197.

[31] B. Yang, Y. Lei, F. Jia, S. Xing, An intelligent fault diagnosis approach based on transfer learning from laboratory bearings to locomotive bearings, Mech. Syst. Signal Process. 122 (2019) 692–706.

[32] Q. Wang, G. Michau, O. Fink, Domain adaptive transfer learning for fault diagnosis, in: 2019 Prognostics and System Health Management Conference (PHM-Paris), 2019, pp. 279–285.

[33] W. Lu, B. Liang, C. Yu, D. Meng, Z. Tao, Deep model based domain adaptation for fault diagnosis, IEEE Trans. Ind. Electron. 64 (3) (2017) 2296–2305.

[34] X. Li, W. Zhang, Q. Ding, J.-Q. Sun, Intelligent rotating machinery fault diagnosis based on deep learning using data augmentation, J. Intell. Manuf. 31 (2020) 433–452.

[35] H. Li, W. Zhao, Y. Zhang, E. Zio, Remaining useful life prediction using multi-scale deep convolutional neural network, Appl. Soft Comput. 89 (2020) 106113.

[36] T. Han, C. Liu, W. Yang, D. Jiang, A novel adversarial learning framework in deep convolutional neural network for intelligent diagnosis of mechanical faults, Knowl.-Based Syst. 165 (2019) 474–487.

[37] J. Jiao, M. Zhao, J. Lin, K. Liang, Residual joint adaptation adversarial network for intelligent transfer fault diagnosis, Mech. Syst. Signal Process. 145 (2020) 106962.

[38] X. Li, W. Zhang, N. Xu, Q. Ding, Deep learning-based machinery fault diagnostics with domain adaptation across sensors at different places, IEEE Trans. Ind. Electron. 67 (8) (2020) 6785–6794.

[39] M. Zhang, D. Wang, W. Lu, J. Yang, Z. Li, B. Liang, A deep transfer model with wasserstein distance guided multi-adversarial networks for bearing fault diagnosis under different working conditions, IEEE Access 7 (2019) 65303–65318.

[40] Z. Cao, M. Long, J. Wang, M.I. Jordan, Partial transfer learning with selective adversarial networks, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2724–2732.

[41] Z. Cao, L. Ma, M. Long, J. Wang, Partial adversarial domain adaptation, 2018, arXiv preprint arXiv:1808.04205.

[42] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.

[43] W. Zhang, G. Peng, C. Li, Y. Chen, Z. Zhang, A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals, Sensors 17 (2017).

[44] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks, 2016, arXiv preprint arXiv:1505.07818.

[45] W. Qiao, D. Lu, A survey on wind turbine condition monitoring and fault diagnosis—Part i: Components and subsystems, IEEE Trans. Ind. Electron. 62 (10) (2015) 6536–6545.

[46] W. Qiao, D. Lu, A survey on wind turbine condition monitoring and fault diagnosis—Part II: Signals and signal processing methods, IEEE Trans. Ind. Electron. 62 (10) (2015) 6546–6557.

[47] Z. Liu, L. Zhang, A review of failure modes, condition monitoring and fault diagnosis methods for large-scale wind turbine bearings, Measurement 149 (2020) 107002.

[48] C. Lessmeier, J. Kimotho, D. Zimmer, W. Sextro, Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification, in: 2016 European Conference of the Prognostics and Health Management Society, 2016, pp. 05–08.07.

[49] W.A. Smith, R.B. Randall, Rolling element bearing diagnostics using the case western reserve university data: A benchmark study, Mech. Syst. Signal Process. 64–65 (2015) 100–131.

[50] X. Zhang, F.X. Yu, S.-F. Chang, S. Wang, Deep transfer network: Unsupervised domain adaptation, 2015, arXiv preprint arXiv:1503.00591.

Te Han received the B.S. and Ph.D. degree in energy and power engineering from Tsinghua University, Beijing, P.R. China, in 2015 and 2020, respectively. He is currently a Postdoctoral Research Fellow with Department of Industrial Engineering, Tsinghua University.

His research interests include machinery condition monitoring, intelligent fault diagnosis and prognostics, and reliability engineering.

Chao Liu received the B.S. degree in energy and power engineering from Huazhong University of Science and Technology, Wuhan, P.R. China, in 2008, and the Ph.D. degree in energy and power engineering from Tsinghua University, Beijing, P.R. China, in 2013. He is currently a Research Assistant Professor in the Department of Energy and Power Engineering, Tsinghua University. Prior to joining Tsinghua University in 2017, he was a Postdoctoral Research Fellow with Department of Mechanical Engineering, Iowa State University, Ames, IA.

His research interests focus on machine learning, prognostics, and health monitoring.

Rui Wu is currently working toward the Ph.D. degree in energy and power engineering at Tsinghua University, Beijing, P.R. China. He received the B.S. degree in energy and power engineering from Tsinghua University in 2018.

His research interests include machinery condition monitoring, signal processing, and remaining useful life prediction of rotating machinery.

Dongxiang Jiang received the B.S. degree in electronic engineering from the Shenyang University of Technology, Shenyang, P.R. China, in 1983, the M.S. degree in electrical engineering from Harbin Institute of Technology, Harbin, P.R. China, in 1989, and the Ph.D. degree in astronautics and mechanics from Harbin Institute of Technology in 1994. Currently, he is a Full Professor in the Department of Energy and Power Engineering, Tsinghua University. Prior to joining Tsinghua University in 1996, he worked as an assistant engineer and an engineer at Harbin Research Institute of Electrical Instrumentation for six years. He was also a Postdoctoral Research Fellow with the Department of Energy and Power Engineering, Tsinghua University.

His research interests include rotor dynamics, finite element analysis, condition monitoring and diagnostics for rotating machinery, and wind power technology. He is a member of ASME, Machinery Fault Diagnostic Division of the Chinese Society for Vibration Engineering, and the Chinese Wind Energy Association.