Deep Learning-Based Partial Domain Adaptation Method on Intelligent Machinery Fault Diagnostics

Xiang Li and Wei Zhang

***Abstract*—In the past years, deep learning-based machinery fault diagnosis methods have been successfully developed, and the basic diagnostic problems have been well addressed where the training and testing data are collected under the same operating conditions. When the training and testing data are from different distributions, domain adaptation approaches have been introduced. However, the existing methods generally assume the availability of the target-domain data in all the health conditions during training, which is not in accordance with the real industrial scenarios. This article proposes a deep learning-based fault diagnosis method to address the partial domain adaptation problems, where the unsupervised target-domain training data do not cover the full machine health state label space. The conditional data alignment and unsupervised prediction consistency schemes are proposed to achieve partial domain adaptation. The experimental results on two rotating machinery datasets suggest the proposed method offers a promising tool for this practical industrial problem.**

***Index Terms*—Deep learning, fault diagnosis, partial domain adaptation, rotating machines, transfer learning.**

I. Introduction

IN THE recent years, as the industrial machines are becoming more and more complex, the traditional fault diagnosis methods based on physical models and signal processing techniques have been less effective in machinery health condition monitoring. Meanwhile, with the rapid development of the computing power and algorithms, the intelligent data-driven fault diagnostic approaches have been promising in capturing the relationship between the measured machinery data and the corresponding health states. Many industrial processes have largely benefited from the applications of the machine learning-based diagnostic methods [1]–[6], including intelligent manufacturing, aerospace industry, automotive, etc.

Despite the successful development, the main assumption of the data-driven fault diagnosis methods lies in that, the training data and the testing data must be subject to the same distribution. That indicates the training and testing data are supposed to be collected from the same machine under identical operating condition. However, the assumption is difficult and mostly impossible to meet in the real industrial scenarios. For instance, the tested machine in the production line generally cannot be used for collecting training data in different faulty states, and the rotating machines typically have multiple working conditions in practice. Consequently, the distribution discrepancy exists between the training and testing data in most cases, that leads to the fault diagnostic performance deteriorations in the real applications [7]. This challenging issue is known as the domain shift problem [8].

To enhance the model generalization ability in different testing scenarios, transfer learning techniques have been successfully developed in the recent years [9], which aim to transfer the knowledge learned from the supervised training data, denoted as source domain, to the unsupervised testing data, denoted as target domain. Specifically, as a specialized form of transfer learning, the domain adaptation methods have been popularly adopted in the fault diagnosis problems, and the learned diagnostic knowledge is expected to generalize well in different machinery operating conditions.

In the current literature, many domain adaptation approaches have been proposed and promising cross-domain diagnostic performance has been obtained [10]–[15]. However, the existing methods mostly aim to extract the domain-invariant features for fault diagnosis through minimizing the marginal data distribution discrepancy between the source and target domains. That indicates the target-domain data in the full machine health state label space are supposed to be available during the model training process.

Manuscript received August 22, 2019; revised November 22, 2019, January 9, 2020, February 11, 2020, and March 4, 2020; accepted March 23, 2020. Date of publication April 7, 2020; date of current version January 27, 2021. This was supported in part by the Fundamental Research Funds for the Central Universities under Grant N2005010, Grant N180703018, Grant N180708009, and Grant N170308028, in part by the National Natural Science Foundation of China under Grant 11902202, in part by the Key Laboratory of Vibration and Control of Aero-Propulsion System Ministry of Education, Northeastern University under Grant VCAME201906, and in part by the Liaoning Provincial Department of Science and Technology under Grant 2019-BS-184. *(Corresponding author: Wei Zhang).*

Xiang Li is with the College of Sciences, and Key Laboratory of Vibration and Control of Aero-Propulsion System Ministry of Education, Northeastern University, Shenyang 110819, China, and also with the Department of Mechanical and Materials Engineering, University of Cincinnati, Cincinnati, OH 45221 USA (e-mail: xiangli@mail.neu.edu.cn).

Wei Zhang is with the School of Aerospace Engineering, Shenyang Aerospace University, Shenyang 110136, China, and also with the Key Laboratory of Vibration and Control of Aero-Propulsion System Ministry of Education, Northeastern University, Shenyang 110136, China (e-mail: zw_7126257@163.com).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIE.2020.2984968

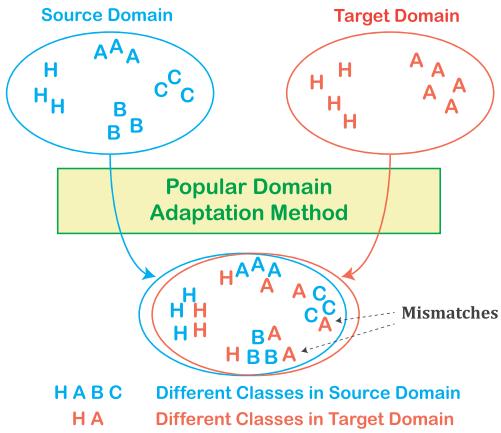

Fig. 1. Illustration of the partial domain adaptation problem, which cannot be readily addressed by the existing methods.

In the real industrial scenarios, while the source data in different machinery health conditions can be usually obtained for training, it is difficult and mostly impossible to collect the testing data in all the health states in advance. Consequently, the popular domain adaptation methods focusing on the marginal distributions generally cannot be directly applied in the real fault diagnosis cases. However, it is noted that the unsupervised target-domain data in the partial label space can be usually collected in advance. For instance, it is basically feasible to collect the testing data in machinery healthy or incipient faulty conditions for model training, despite the absence of the health state labels. Therefore, the partial domain adaptation problem is in better accordance with the real industrial scenarios, where the available unsupervised target-domain data only cover a subset of all the concerned health conditions.

Fig. 1 shows the illustration of the partial domain adaptation problem. In general, poor cross-domain fault diagnosis performance can be mostly achieved if the existing methods are directly applied in the partial domain adaptation cases due to the outlier source classes. Therefore, it is challenging and important to address this problem in real applications. This article proposes a deep learning-based partial domain adaptation method for rotating machinery fault diagnostics. Conditional data alignments are achieved through minimization of the MMD between domains. A prediction consistency scheme is proposed to leverage the performances of multiple classification modules by using adversarial learning.

The main advantage of the proposed method lies in that an effective and practical solution is provided for the partial domain adaptation problem in the real industrial fault diagnosis problems. While the partial domain adaptation issue is common in practice, few solutions can be found in the current literature, and this article offers one of the first successful attempts on this challenging problem. Moreover, using the proposed method, the unsupervised data can be efficiently explored for model establishment which are easy to collect in the industrial scenarios. Therefore, the proposed method is promising for real industrial applications. The effectiveness is validated by the experiments on two real datasets in this study.

The remainder of this article starts with the related works in Section II. The proposed fault diagnosis method is presented in Section III, and experimentally validated and investigated in Section IV. Section V concludes this article.

II. Related Works

The intelligent data-driven machinery fault diagnosis problems have been attracting increasing attention in the past years, and deep learning has been popularly used in related studies [16], [17]. The basic fault diagnosis problems have been well addressed, where the training and testing data are collected from the same distribution. Zhang *et al.* [18] proposed a deep residual learning-based method for rolling bearing fault diagnosis, and a stacked denoising auto-encoder network architecture was proposed by Lu *et al.* [19] for machinery health state identification. In general, fairly high testing accuracies have been achieved in the basic fault diagnosis problems.

Recently, due to rising industrial demands on the generalization ability and applicability of the fault diagnosis models in different operating scenarios, transfer learning techniques have been successfully used to transfer the learned diagnostic knowledge across different working conditions. Specifically, the domain shift problem has been successfully addressed using the domain adaptation methods [20], [21]. Xu *et al.* [22] proposed an online fault diagnosis method based on a deep transfer convolutional neural network model, which includes both online and offline networks, and achieves significant improvement on the fault diagnosis tasks. A transfer learning framework was proposed by Han *et al.* [23] based on pre-trained CNN, which well exploits the knowledge learned from the training data to facilitate fault diagnosis in similar tasks.

The maximum mean discrepancy (MMD) is a popular metric in transfer learning, which measures the distance between source and target-domain data distributions. The high-level representations of different domain can be projected into the same region in the subspace through optimization of MMD metric, which facilitates extraction of generalized knowledge. Lu *et al.* [24] proposed a deep model-based domain adaptation method for machinery fault diagnosis, where the MMD metric is adopted to measure the domain distance. By minimizing MMD, promising domain fusion effect can be achieved. A sparse auto-encoder model was used by Wen *et al.* [25] for transfer learning problem in fault diagnosis, where the MMD term is also adopted to bridge the domain gap. The cross-domain diagnostic task was also addressed by Li *et al.* [26], where a deep metric learning algorithm is employed. The MMD is minimized to achieve better representation learning performance.

Besides the popular MMD metric, domain adversarial learning in deep neural networks has been emerging as an effective tool for extracting domain-invariant features [27], [28], and has achieved promising cross-domain fault diagnosis results in the recent years. Guo *et al.* [29] proposed a deep convolutional neural network model to transfer knowledge across domains, where the conditional recognition and domain adaptation modules are used. Both the adversarial learning and minimization of MMD between the source and target domains are utilized. The

experimental results show that generalized diagnostic features can be well extracted by the two domain adaptation methods. The adversarial learning scheme was also adopted by Li *et al.* [30] on the rotating machinery transfer learning problem. The shared features of different domains can be effectively learned, which have strong generalization ability. The recent advances on the cross-domain fault diagnostic problems have been well reviewed in [31], [32].

Despite the rapid development of domain adaptation in fault diagnosis, the existing approaches generally assume the availability of sufficient target-domain data in all the health conditions during model training, and the partial domain adaptation problem has received far less attention. In the image processing field, some initial attempts have been made on similar problems recently. Cao *et al.* [33] developed a selective adversarial network (SAN) for partial domain adaptation, where multiple discriminators are adopted. In [34], the unsupervised partial transfer learning problem was addressed using the importance-weighted adversarial network (IWAN). Weights are attached to the source samples to measure their similarities with the target domain. However, to the authors' knowledge, no study can be found on this practical industrial fault diagnosis problem in the current literature. This article aims to bridge the gap and address the partial domain adaptation fault diagnostic problem using deep learning.

III. Proposed Method

A. Problem Formulation

In this article, the partial domain adaptation problem in machinery fault diagnosis is investigated. Let $\mathbf{x} \in \mathrm{R}^{N_{\text{input}}}$ and y denote the data sample and the machine health condition label, respectively, where N_{input} represents the sample dimension. It is assumed the source-domain data X_S and the corresponding labels Y_S are drawn from the distribution $P_S(\mathbf{x}, y)$, and the target-domain data X_T and the corresponding labels Y_T are from the distribution $P_T(\mathbf{x}, y)$. Due to domain shift, $P_S(\mathbf{x}, y) \neq P_T(\mathbf{x}, y)$.

Let $\mathcal{D}_s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ and $\mathcal{D}_t = \{(\mathbf{x}_i^t, y_i^t)\}_{i=1}^{n_t}$ denote the source and target-domain datasets, respectively. $\mathbf{x}_i^s \in \mathrm{R}^{N_{\text{input}}}$ and $\mathbf{x}_i^t \in \mathrm{R}^{N_{\text{input}}}$ are the sample instances in the two domains, and $y_i^s \in Y^{\text{all}}$ and $y_i^t \in Y^{\text{all}}$ represent the corresponding machine health condition labels. Y^{all} denotes the general label space. n_s and n_t are the numbers of the source and target-domain samples, respectively.

In this study, the practical partial domain adaptation task is focused on, which indicates the available target-domain observation labels during training only cover a subset of all the labels in Y^{all}. Let $\mathcal{D}_t^{\text{train}} = \{(\mathbf{x}_i^{t,\text{train}}, y_i^{t,\text{train}})\}_{i=1}^{n_{t,\text{train}}}$ denote the available target-domain observations in model training. $\mathcal{D}_t^{\text{train}} \subseteq \mathcal{D}_t$ and $n_{t,\text{train}}$ is the number of the samples in $\mathcal{D}_t^{\text{train}}$. $y_i^{t,\text{train}} \in Y^{\text{partial}}$, where $Y^{\text{partial}} \subseteq Y^{\text{all}}$ is the corresponding partial label space. It should be noted that the source domain is assumed to be complete with all the machinery health conditions in this study. In practice, this assumption is reasonable since it is usually easy to collect data from different health conditions in the experimental environment.

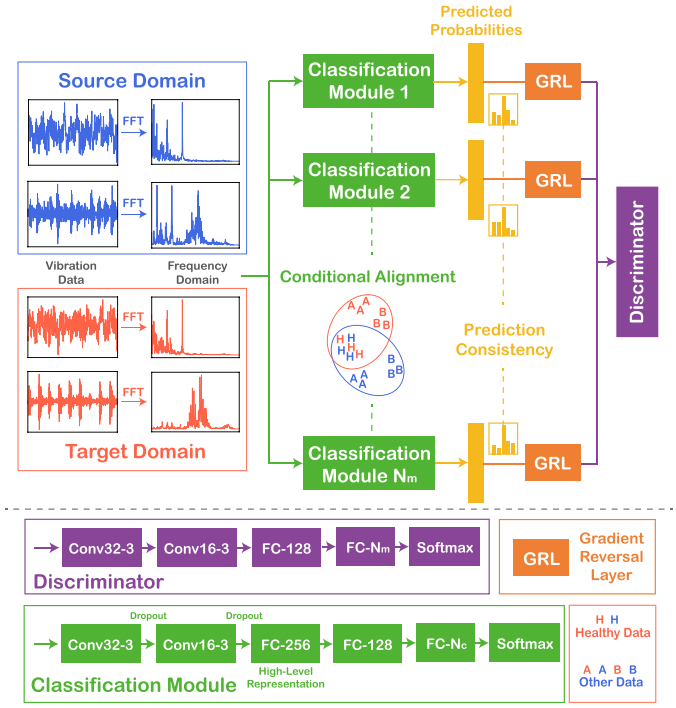

Fig. 2. Proposed deep neural network architecture. Conv32-3: convolutional layer with 32 filters of size 3. FC: fully connected layer. GRL: gradient reversal layer.

Furthermore, it is assumed the target-domain data in the machinery healthy condition are available for training, which are generally easy to collect in the real industrial scenarios. Let $\mathcal{D}_t^h = \{(\mathbf{x}_i^{t,h}, y = \text{“healthy”})\}_{i=1}^{n_{t,h}}$ denote the target-domain healthy state dataset where $n_{t,h}$ is the number of the samples. This article aims to build a cross-domain fault diagnosis classifier using labeled source data $\mathcal{D}_s$, target healthy data $\mathcal{D}_t^h$, and unlabeled partial target data $\mathcal{D}_t^{\text{train}}$, which holds on all the target-domain data $\mathcal{D}_t$. It should be noted that the whole target domain $\mathcal{D}_t$ includes data in all the machinery health conditions, while the target-domain training data, i.e., $\mathcal{D}_t^h$ and $\mathcal{D}_t^{\text{train}}$, only cover partial health condition label space.

B. Method Overview and Network Architecture

Fig. 2 presents the architecture of the proposed deep neural network on the partial domain adaptation fault diagnosis problem. In general, multiple classification modules are used to extract diverse features. Besides the typical supervised learning scheme, two techniques are proposed to address the domain shift phenomenon, i.e., conditional data alignment and unsupervised prediction consistency.

While the popular marginal domain adaptation method is not feasible in this problem, the conditional data alignment of the healthy state is straightforward to be implemented. The healthy data of the source and target domains can be projected into the same region in the new high-level subspace through deep learning, which helps bridge the domain gap and facilitates knowledge transfer. This technique is supposed to be applied for all the classification models.

Furthermore, despite the domain fusion of the healthy data, the alignment of the rest classes in the two domains can not be guaranteed. To address this issue, the prediction consistency scheme is introduced, where the predictions on the unsupervised target data by different classification modules are supposed to be consistent. In this way, more generalized features can be learned, and the overfitting risks can be reduced.

Specifically, the machinery vibration acceleration signals are used for fault diagnosis in this study. The temporal data are transformed to the frequency domain using fast Fourier transformation (FFT) first, and the frequency-domain samples are used as the model inputs. Let N_m denote the number of the classification modules. In this study, all the models share identical network structures for simplicity. However, different structures can be readily used for different classification modules in the proposed framework.

The typical deep convolutional neural network is adopted in each classification module, which consists of two one-dimensional convolutional layers with 32 and 16 filters, a flattened layer, and three fully connected layers with 256, 128, and N_c neurons, respectively. N_c denotes the number of the machine health conditions. The softmax function is used for classification, and the rectified linear units (ReLU) activation function is employed throughout the network.

C. Network Optimization

In general, three components are included in the network optimization objective, i.e., source supervision, conditional data alignment, and unsupervised prediction consistency.

1) Source Supervision: First, the typical supervised learning scheme is used, and the empirical machinery health state classification error of the labeled data samples is supposed to be minimized for all the classification modules. The cross-entropy loss function L_s is adopted, that is defined as

$$L_s = \frac{1}{N_m}\sum_{i=1}^{N_m} \times \left[-\frac{1}{n_s+n_{t,h}}\sum_{j=1}^{n_s+n_{t,h}}\sum_{k=1}^{N_c} 1\{y_j=k\}\log\frac{e^{x_{j,k}^i}}{\sum_{m=1}^{N_c} e^{x_{j,m}^i}}\right], \tag{1}$$

where $x_{j,k}^i$ denotes the kth element of output vector in the ith classification module, taking as input the jth labeled sample in $\mathcal{D}_s \cup \mathcal{D}_t^h$.

2) Conditional Data Alignment: Since the data in the machine healthy state are generally easy to collect in the real industrial scenarios, the conditional data alignment between the source and target domains can be readily implemented for domain adaptation. Specifically, the MMD metric [35] is adopted in this study, which is defined as the squared distance between the kernel embeddings of data distributions P_S and P_T in the reproducing kernel Hilbert space (RKHS).

$$\mathrm{MMD}_{\mathbf{k}}(P_S,P_T)^2 \triangleq \left\|\mathbf{E}_{P_S}\left[\phi\left(\mathbf{x}^s\right)\right]-\mathbf{E}_{P_T}\left[\phi\left(\mathbf{x}^t\right)\right]\right\|_{\mathcal{H}_{\mathbf{k}}}^2 \tag{2}$$

where ϕ is the feature map, $\mathbf{k}$ is the kernel function induced by ϕ, and $\mathcal{H}_{\mathbf{k}}$ represents the RKHS with the characteristic kernel $\mathbf{k}$.

In the practical implementation, the unbiased empirical estimation of $\mathbf{E}_{P_S}[\phi(\mathbf{x}^s)] = \frac{1}{n_s}\sum_{i=1}^{n_s}\phi(\mathbf{x}_i^s)$ can be used for computation, and the unbiased empirical estimation of MMD between the source and target domains can be formulated as [36]

$$\begin{aligned}\mathrm{MMD}_{\mathbf{k}}(P_S,P_T)^2 &= \left\|\frac{1}{n_s}\sum_{i=1}^{n_s}\phi\left(\mathbf{x}_i^s\right)-\frac{1}{n_t}\sum_{i=1}^{n_t}\phi\left(\mathbf{x}_i^t\right)\right\|_{\mathcal{H}_{\mathbf{k}}}^2 \\ &= \frac{1}{n_s^2}\sum_{i=1}^{n_s}\sum_{j=1}^{n_s}\mathbf{k}\left(\mathbf{x}_i^s,\mathbf{x}_j^s\right) \\ &\quad+\frac{1}{n_t^2}\sum_{i=1}^{n_t}\sum_{j=1}^{n_t}\mathbf{k}\left(\mathbf{x}_i^t,\mathbf{x}_j^t\right) \\ &\quad-\frac{2}{n_s n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}\mathbf{k}\left(\mathbf{x}_i^s,\mathbf{x}_j^t\right).\end{aligned} \tag{3}$$

Based on previous studies [37], [38], multiple kernels of MMD are suggested to be used for better performance, and five Gaussian kernels with bandwidth parameters of 1, 2, 4, 8, and 16 are adopted.

In this study, the healthy data distribution discrepancy of the two domains is supposed to be minimized for all the classification modules, and the loss function can be defined as

$$L_c = \sum_{i=1}^{N_m}\mathrm{MMD}_{\mathbf{k}}(P_S^{h,i},P_T^{h,i}) \tag{4}$$

where $P_S^{h,i}$ and $P_T^{h,i}$ denote the distributions of the high-level representations of the source and target-domain healthy data in the ith classification module, respectively.

3) Unsupervised Prediction Consistency: In this study, the multiple classification modules generally learn different features for diagnostics. Better domain adaptation is expected to be achieved if identical prediction results on the unsupervised target-domain data can be obtained by different classification modules. The adversarial learning approach is introduced to achieve the unsupervised prediction consistency [39].

As shown in Fig. 2, the predicted class probabilities by different models are fed into a discriminator module D, which consists of two convolutional layers with 32 and 16 filters, a flattened layer, two fully connected layers with 256 and N_m neurons, and a softmax function. The discriminator aims to accurately identify the sources of the input data, i.e., the classification modules from which the predicted probability vectors are obtained, and output the identification results. On the contrary, the multiple classification modules are optimized to generate probability vectors that cannot be distinguished by the discriminator. Through adversarial training between the classification modules and the discriminator, similar prediction results can be obtained by different models, thus achieving prediction consistency.

The optimization problem can be formulated as

$$\begin{aligned}\{\hat{\theta}_{f,i}\}_{i=1}^{N_m} &= \arg\max_{\{\boldsymbol{\theta}_{f,i}\}_{i=1}^{N_m}} L_d(D(\{F_i(\mathbf{x};\boldsymbol{\theta}_{f,i})\}_{i=1}^{N_m};\hat{\theta}_d)) \\ \hat{\theta}_d &= \arg\min_{\boldsymbol{\theta}_d} L_d(D(\{F_i(\mathbf{x};\hat{\theta}_{f,i})\}_{i=1}^{N_m};\boldsymbol{\theta}_d)) \\ \mathbf{x} &\in \mathcal{D}_t^{\text{train}},\end{aligned} \tag{5}$$

where F_i denotes the ith classification module with parameter of $\theta_{f,i}$, and $\boldsymbol{\theta}_d$ is the parameter of D. $\hat{\theta}_{f,i}$ and $\hat{\theta}_d$ represent the optimal values of $\boldsymbol{\theta}_{f,i}$ and $\boldsymbol{\theta}_d$, respectively. L_d is the cross-entropy loss function used to quantify the identification error by the discriminator as

$$L_d = \sum_{i=1}^{N_m} \left[-\frac{1}{n_t} \sum_{j=1}^{n_t} \sum_{k=1}^{N_m} 1\{i=k\} \log \frac{e^{x_{d,j,k}^{t,i}}}{\sum_{m=1}^{N_m} e^{x_{d,j,m}^{t,i}}} \right], \quad (6)$$

where $x_{d,j,k}^{t,i}$ denotes the kth element of the output vector in the discriminator module, taking the representation of the jth target-domain sample in the ith classification module as input.

In general, the network optimization in (5) cannot be directly implemented using the popular stochastic gradient descent (SGD) method. In the literature, the gradient reversal layer (GRL) is basically adopted for adversarial training [34], [39], [40]. In the forward propagation, the GRL serves as identify mapping. In the backpropagation (BP) process, the GRL flips the sign of the gradient received from the following layer, and passes it to the preceding layer afterwards. Therefore, the maximization and minimization problems can be unified. This technique can be readily implemented with the current programming methods.

Specifically in this study, the GRLs are placed between the classification modules and the discriminator. They can be ignored when calculating the network output due to the identity mapping feature, and they are only considered during BPs. In this way, the optimization objective in (5) can be readily solved by the popular SGD method using the GRL. Correspondingly, the diverse diagnostic features learned by different models can be leveraged, and better domain fusion can be achieved with lower risk of overfitting under the prediction consistency.

4) General Optimization Objective: In summary, by integrating the optimization objectives in (1), (4), and (5), the parameters of the network can be updated in each training epoch as

$$\begin{aligned} \boldsymbol{\theta}_{f,i} &\leftarrow \boldsymbol{\theta}_{f,i} - \delta \left(\varepsilon_s \frac{\partial L_s}{\partial \boldsymbol{\theta}_{f,i}} + \varepsilon_c \frac{\partial L_c}{\partial \boldsymbol{\theta}_{f,i}} - \varepsilon_d \frac{\partial L_d}{\partial \boldsymbol{\theta}_{f,i}} \right) \\ \boldsymbol{\theta}_d &\leftarrow \boldsymbol{\theta}_d - \delta \varepsilon_d \frac{\partial L_d}{\partial \boldsymbol{\theta}_d} \\ i &= 1, 2, \ldots, N_m \end{aligned} \quad (7)$$

where δ denotes the learning rate. ε_s, ε_c, and ε_d represent the penalty coefficients for the losses L_s, L_c, and L_d, respectively.

After training, the testing samples can be fed into the network to evaluate the fault diagnosis performance, and N_m classification results can be obtained. In this study, the majority of the estimations is used as the final fault diagnosis result.

IV. Experimental Study

A. Dataset Descriptions

1) *Case Western Reserve University (CWRU):* This popular rotating machinery dataset is provided by the Bearing Data Center of Case Western Reserve University [41]. Extensive studies have been carried out on this rolling bearing fault diagnosis dataset. The vibration acceleration data used in this study were collected from the motor drive end under four rotating speeds, i.e., 1797, 1772, 1750, and 1730 r/min, and on four health conditions: 1) healthy (H), 2) outer race fault (OF), 3) inner race fault (IF), and 4) ball fault (BF). The three kinds of faults are manually created with diameters of 7, 14, and 21 mils, respectively. In summary, 10 health states in four rotating speed domains are considered. It should be noted that the data in the CWRU dataset are relatively clean with weak noise, that makes it basically easy for fault diagnosis. However, since it has been popularly explored in the literature and many existing studies are available for comparisons, it is used in this study to generally examine the proposed method.

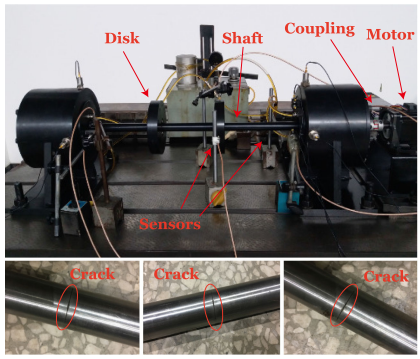

Fig. 3. Test rig and shaft cracks in the Crack dataset [10].

TABLE I
INFORMATION OF THE TWO DATASETS

Dataset	Class Label	1	2	3	4	5	6	7	8	9	10
CWRU	Fault Location	N/A (H)	IF	IF	IF	BF	BF	BF	OF	OF	OF
	Fault Size (mil)	0	7	14	21	7	14	21	7	14	21
Crack	Crack Depth (mm)	N/A	2.5	3.5	5						

2) *Crack:* This rotating machinery dataset is collected from a test rig supported on two rolling bearings as shown in Fig. 3. A 400-W dc motor drives the rotor operating through a flexible coupling. A shaft that is 672 mils long and 24 mils in diameter, and two disks with thicks of 30 mils are included. The vibration acceleration signals used in this article are transformed from the data obtained with the eddy current sensor recording the disk displacement along the vertical direction. The sampling frequency is 2 kHz, and four rotating speeds are implemented as 938, 1055, 1289, and 1523 r/min. Four shaft health conditions are considered with different levels of cracks. The detailed information of the two datasets is presented in Table I.

B. Implementation Details and Tasks

In this study, multiple partial domain adaptation tasks are investigated on each rotating machinery dataset, which covers unsupervised samples of different classes randomly selected on the target domain. The detailed information of the fault diagnosis tasks is presented in Table II. In network optimization, the BP algorithm is applied for the parameter updates, and the Adam optimization method is used [42]. By default, it is assumed 200 labeled samples are available for training with respect to each health condition in the source domain, and 200 unlabeled target-domain samples are available for each of the classes that

TABLE II
PARTIAL DOMAIN ADAPTATION FAULT DIAGNOSIS TASKS IN THIS ARTICLE

	CWRU			Crack	
Task Name	Source -> Target (Rotating Speed)	Target Classes	Task Name	Source -> Target (Rotating Speed)	Target Classes
C_1	1797− > 1730	Non-Partial	T_1	1523− > 938	Non-Partial
C_2	1772− > 1750	1,2,4,5,6,7,9,10	T_2	1289− > 1055	1,2,3
C_3	1750− > 1730	1,3,5,6,8,9,10	T_3	1055− > 938	1,3,4
C_4	1772− > 1730	1,2,4,5,6,8,9	T_4	1523− > 1289	1,2,4
C_5	1797− > 1750	1,2,3,6,8,10	T_5	1289− > 938	1,2
C_6	1730− > 1797	1,3,5,8,9	T_6	938− > 1523	1,3
C_7	1750− > 1772	1,2,4,6	T_7	1055− > 1289	1,4
C_8	1730− > 1750	1,4,8	T_8	938− > 1055	1,3,4
C_9	1730− > 1772	1,3,9	T_9	1289− > 1523	1,2,4
C_{10}	1750− > 1797	1,6	T_{10}	938− > 1289	1,4
C_{11}	1750− > 1730	1	T_{11}	938− > 1055	1

TABLE III
PARAMETERS USED IN THIS ARTICLE

Parameter	Value	Parameter	Value
Epochs	40000	ε_s	1
δ	10^{-5}	ε_c	1
batch size for L_s	64	ε_d	1
batch size for L_c	128	N_{input}	512
batch size for L_d	256		

are included in different partial domain adaptation tasks. In each training epoch, mini-batches of samples are selected for computing the optimization gradients. Specifically, 64, 128, and 256 samples are used for estimating the supervised learning loss L_s, conditional alignment loss L_c, and prediction consistency loss L_d, respectively. The reported experimental results in this study are generally averaged by ten trials to reduce the effect of randomness. The model parameters are presented in Table III. They are mostly determined from the CWRU validation task where the source and target domains are under 1797 and 1730 r/min, respectively, and the target training data include the classes 1, 2, 3, 4, and 5. The penalty coefficients ε_s, ε_c, and ε_d are kept equal in this study for simplicity. However, they can be readily set different in order to attach different importance to different objectives in practice. Based on the experiments, the influence of the parameters on the model performance is not significant as long as they are in a reasonable range, due to the convergence of the algorithm. It should be noted that based on the experiments, the model performance on the partial domain adaptation tasks is generally influenced by the optimization objective. The effect of the network structures on the diagnosis results is not significant under the condition that sufficient learning capacity is provided by the model. By default, three classification modules $N_m = 3$ are used in the experiments.

C. Compared Approaches

In this study, different approaches on the partial domain adaptation problem are implemented for comparisons. Specifically, the following methods are carried out, which have similar experimental settings with the proposed method.

1) *Supervised Only (SupOnly):* First, the SupOnly method is used where only the supervised learning scheme is considered. That means the fault diagnosis model is trained with minimization of only the objective in (1). This approach follows the conventional machine learning pattern with supervised learning and with no transfer learning technique, thus providing a baseline for comparison.
2) *Marginal Domain Adaptation (MargDA):* As a popular way for domain adaptation, the MargDA approach aims to achieve marginal distribution alignment between the source and target domains in the high-level data representation space [24], [26], despite the fact that the available unsupervised target data only cover the partial label space. Therefore, this method offers a fair comparison with the existing popular transfer learning approaches.
3) *No Prediction Consistency (NoPredConsist):* In order to evaluate the benefits of the unsupervised prediction consistency scheme, the NoPredConsist method is implemented where that scheme is removed from the proposed framework. Specifically, the loss function in (5) of the proposed method is not considered.
4) *No Conditional Alignment (NoCondiAlign):* Similar with the NoPredConsist method, the NoCondiAlign approach does not implement the conditional data alignment with the loss function in (4). In this way, the effect of the healthy data fusion scheme can be examined, that shows the benefits of the utilization of the healthy state data.

D. Experimental Results

The experimental results on the partial domain adaptation fault diagnosis tasks on the CWRU dataset are presented in Table IV. Specifically, both the testing accuracies on the unsupervised partial target training data $\mathcal{D}_t^{\text{train}}$ and the whole target domain $\mathcal{D}_t$ are provided for better evaluations. It can be observed that since the CWRU data are relatively clean with less noise, high testing accuracies are mostly obtained. Specifically, when no transfer learning technique is used, lower accuracies are generally obtained by the baseline method SupOnly, which indicates the distribution discrepancy between different domains significantly deteriorates the model performance. When the popular marginal domain adaptation method is used, fairly high testing accuracies are achieved by the MargDA method in tasks C_1–C_3, where most of the health conditions are included in the target-domain training data. However, when less target classes are covered such as tasks C_9 and C_{10}, the MargDA method generally fails with low testing accuracies. That is because drawing the source domain with full label space and the target domain with partial label space into the same region leads to mismatches of the classes as illustrated in Fig. 1. Therefore, the existing popular transfer learning methods cannot be readily used in the practical partial domain adaptation problems.

The effects of the two proposed techniques are also evaluated. When the unsupervised prediction consistency scheme is removed, significant testing performance drops in different scenarios are observed by the NoPredConsist approach compared with the proposed method. Similarly, despite the improved testing accuracies compared with the baseline approach, the network performance without the conditional data alignment scheme, i.e., the NoCondiAlign method, is less competitive than the proposed method. The results validate the benefits and necessities of the two proposed domain adaptation techniques.

TABLE IV
TESTING ACCURACIES ON THE PARTIAL DOMAIN ADAPTATION FAULT DIAGNOSIS TASKS ON THE CWRU DATASET (%)

Method	SupOnly		MargDA		NoPredConsist		NoCondiAlign		SAN		IWAN		Proposed	
Test Data	$\mathcal{D}_t^{train}$	$\mathcal{D}_t$	$\mathcal{D}_t^{train}$	$\mathcal{D}_t$	$\mathcal{D}_t^{train}$	$\mathcal{D}_t$	$\mathcal{D}_t^{train}$	$\mathcal{D}_t$	$\mathcal{D}_t^{train}$	$\mathcal{D}_t$	$\mathcal{D}_t^{train}$	$\mathcal{D}_t$	$\mathcal{D}_t^{train}$	$\mathcal{D}_t$
C_1	86.4	86.4	99.9	99.9	92.3	92.3	98.7	98.7	99.9	99.9	99.9	99.9	**99.9**	**99.9**
C_2	89.3	88.9	98.8	98.3	93.4	92.8	98.4	98.2	99.8	99.6	99.7	99.7	**99.7**	**99.6**
C_3	92.3	91.7	96.9	96.4	94.3	93.2	97.3	96.2	98.9	97.9	98.8	97.1	**98.8**	**98.4**
C_4	87.6	88.3	94.7	93.1	88.9	88.9	95.4	94.9	99.0	98.4	98.5	97.0	**98.9**	**99.0**
C_5	88.3	88.9	97.0	94.7	91.2	89.8	97.3	97.0	98.4	97.1	99.0	96.8	**98.7**	**97.4**
C_6	93.5	91.5	94.5	95.1	93.4	92.8	96.4	96.5	97.8	97.0	98.3	97.1	**98.1**	**97.5**
C_7	92.4	91.9	87.0	90.7	94.3	92.9	97.8	95.4	98.0	94.5	97.4	95.0	**97.2**	**95.8**
C_8	95.7	92.7	81.7	83.3	96.7	95.3	96.5	95.8	96.5	96.2	98.1	96.1	**97.4**	**96.7**
C_9	92.6	90.3	84.5	83.0	94.3	93.2	94.6	93.1	97.8	92.3	96.4	92.4	**97.4**	**94.0**
C_{10}	93.4	89.0	72.3	82.0	96.9	89.3	95.9	87.5	95.7	85.3	96.4	82.8	**96.5**	**89.5**
C_{11}	92.2	87.8	68.6	80.4	96.6	90.0	94.8	89.1	96.4	89.4	96.1	88.9	**96.2**	**90.1**
Average	91.3	89.8	88.7	90.7	93.9	91.8	96.6	94.3	98.0	95.2	98.1	94.8	**98.1**	**96.2**

TABLE V
TESTING ACCURACIES ON THE PARTIAL DOMAIN ADAPTATION FAULT DIAGNOSIS TASKS ON THE CRACK DATASET (%)

Method	SupOnly		MargDA		NoPredConsist		NoCondiAlign		SAN		IWAN		Proposed	
Test Data	$\mathcal{D}_t^{train}$	$\mathcal{D}_t$	$\mathcal{D}_t^{train}$	$\mathcal{D}_t$	$\mathcal{D}_t^{train}$	$\mathcal{D}_t$	$\mathcal{D}_t^{train}$	$\mathcal{D}_t$	$\mathcal{D}_t^{train}$	$\mathcal{D}_t$	$\mathcal{D}_t^{train}$	$\mathcal{D}_t$	$\mathcal{D}_t^{train}$	$\mathcal{D}_t$
T_1	72.4	72.4	97.6	97.6	82.5	82.5	88.6	88.6	96.7	96.7	96.3	96.3	**96.5**	**96.5**
T_2	81.5	77.2	92.5	86.3	85.6	82.0	92.3	86.5	92.4	87.4	92.0	88.2	**92.1**	**88.3**
T_3	76.3	72.4	85.5	80.3	82.4	78.2	86.8	84.5	87.9	83.1	88.6	81.3	**88.3**	**83.5**
T_4	75.6	75.8	80.5	81.3	85.4	79.0	84.7	81.5	85.8	83.2	84.8	83.1	**85.2**	**84.0**
T_5	78.3	77.9	81.4	80.2	82.3	77.6	91.4	89.2	92.7	86.8	91.3	87.4	**92.0**	**88.5**
T_6	82.0	78.2	85.7	79.5	86.3	80.5	91.6	90.8	91.7	90.2	91.2	87.7	**92.3**	**89.6**
T_7	68.4	66.1	75.4	77.2	74.3	71.6	86.3	82.0	88.6	79.3	86.4	78.2	**88.7**	**81.4**
T_8	73.6	74.2	79.3	77.3	78.1	75.4	86.2	84.3	86.0	80.2	85.1	79.3	**85.3**	**82.6**
T_9	77.5	76.7	74.6	81.4	81.0	79.8	87.3	84.7	87.4	83.5	86.6	83.2	**88.3**	**86.4**
T_{10}	73.6	75.1	77.4	80.2	78.5	77.8	90.7	87.3	92.1	83.5	93.2	84.6	**93.4**	**87.5**
T_{11}	72.9	75.5	75.1	78.7	79.0	78.3	89.9	85.8	91.5	79.9	89.3	81.6	**93.1**	**85.8**
Average	75.6	74.7	82.3	81.8	81.4	78.4	88.7	85.9	90.3	84.9	89.5	84.6	**90.5**	**86.7**

Moreover, it is noted that the testing accuracies on the target training data $\mathcal{D}_t^{\text{train}}$ are generally higher than those on the whole target domain $\mathcal{D}_t$ for different methods. This suggests that the unsupervised target-domain data help improve the generalization ability of the data-driven fault diagnosis model. The proposed method is still able to achieve the best testing performance on the whole target domain $\mathcal{D}_t$, with generally higher than 94% accuracies.

In addition, the partial domain adaptation methods in the current literature are also implemented for evaluation, i.e., the SANs [33] and IWANs [34]. It can be observed that competitive results are obtained in different cases. However, the proposed method generally outperforms the compared approaches, especially in the scenarios with fewer target labels and on the whole health condition space. The performance is mostly due to the fact that the conditional data alignment of the machinery health state is implemented by the proposed method. That is one of the main characteristics of the industrial fault diagnostic problem where the health state data are basically easy to obtain in different domains, thus facilitating knowledge transfer.

The experimental results on different partial domain adaptation tasks on the Crack dataset are presented in Table V. In general, similar display patterns are observed with those on the CWRU dataset as shown in Table IV, and the proposed method still achieves the highest testing accuracies in most scenarios.

It should be noted that due to the high complexity of the deep neural network structure and the training procedure, many factors introduce strong randomness to the model and have influence on the performance, such as network initialization and batch selection of training data. Therefore, the proposed method does not obtain the highest testing accuracy in every task. The numerical results in this article aim to provide a general evaluation on the effectiveness of different methods. Despite the lower accuracies in limited cases, the proposed method is still more effective for the partial domain adaptation problems in general. For instance, on the Crack dataset, 11 partial domain adaptation tasks are evaluated. While the NoCondiAlign method achieves slightly higher accuracy than the proposed method in one problem setting $\mathcal{D}_t^{\text{train}}$ of T_2 and T_8 (92.3%→92.1% in task T_2), it performs noticeably worse in all the other tasks. Especially, on task T_1, the performance gap is remarkable (88.6%→96.5%), that is, due to the characteristics of the NoCondiAlign method, where the conditional alignment of the healthy data is not well exploited. In fact, as one of the main unique features of the industrial fault diagnosis task, the healthy data are mostly easy to obtain, and proper utilization of the healthy data is expected to largely improve the model performance. Hence, in spite of some randomness and fluctuations, the effectiveness and suitability of the proposed method on the partial domain adaptation problem can be still validated.

Furthermore, Fig. 4 shows the confusion matrices of the proposed method in different tasks, where the predictions and the ground truths of different classes are presented. It is noted that when the health conditions are covered in the unsupervised target-domain training data, relatively higher testing accuracies can be obtained. The proposed method still achieves promising performance on the machinery health conditions that are not included during training. The effectiveness and superiority of the proposed method on the partial domain adaptation tasks have thus been validated.

1) Performance Analysis: First, the parameters of the proposed model are evaluated in this section. Fig. 5 shows the accuracies in the validation task, where multiple network structure parameters are examined. It can be observed that in general, the influence of the network architecture on the model performance is not significant. When the typical settings of convolutional neural network are applied, promising results can be obtained in different cases. However, when the network

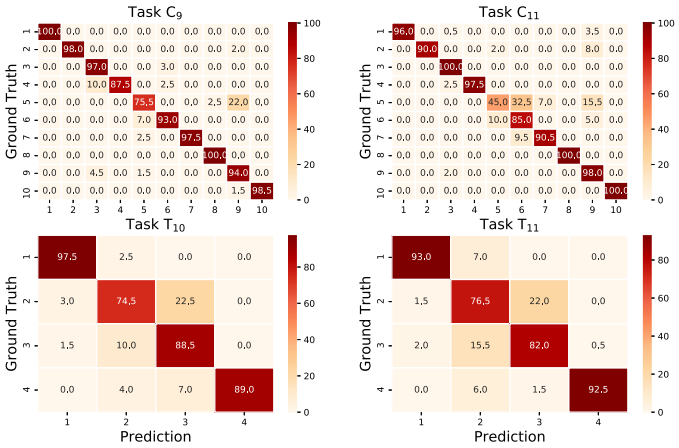

Fig. 4. Confusion matrices in different partial domain adaptation tasks with the proposed method. The numerical testing accuracies are presented (%).

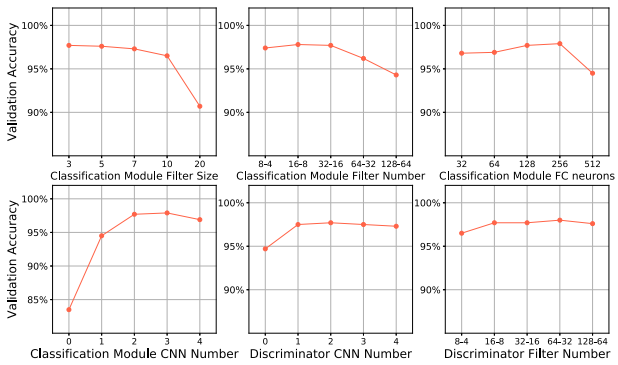

Fig. 5. Influence of network parameters on model validation performance. The labels in filter number denote the convolutional filters in the first and second layers, respectively.

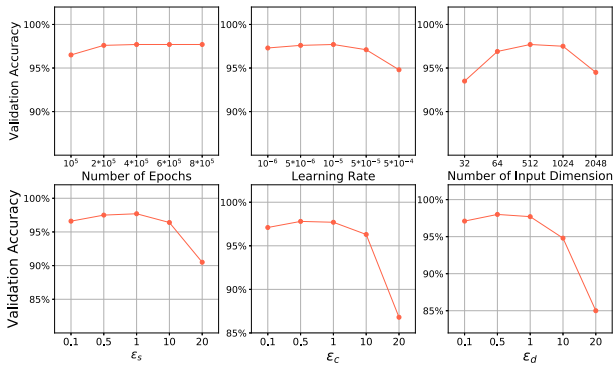

Fig. 6. Influence of model parameters on the validation performance.

capacity is insufficient, such as zero convolutional layer, the model performance remarkably deteriorates. If extremely large network is adopted, such as filter size of 20 and filter number of 128-64, the performance also decays due to overfitting. In practical implementations, the network structure can be properly determined using the validation dataset.

Next, the influence of the model parameters is investigated, and the results are shown in Fig. 6. It can be observed that similar

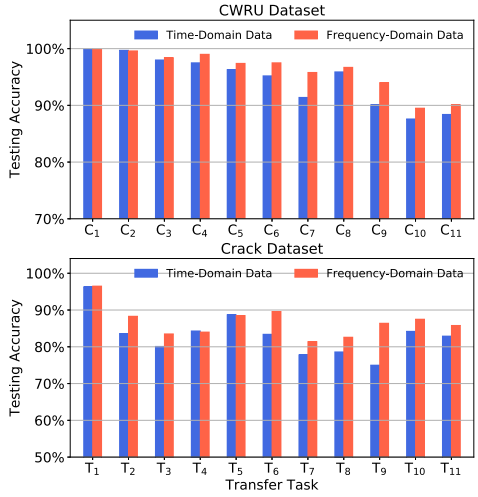

Fig. 7. Testing accuracies in different partial domain adaptation tasks by the proposed method, where the time and frequency-domain data samples are used as the model inputs.

with the network structure, promising validation results can be obtained when typical parameters in deep learning are applied in the proposed method. For instance, the number of epochs does not show noticeable influence on model performance, since the model is stable in the final training periods. When relatively small learning rate is applied, the model can be well trained. The network input dimension should be of medium size. The three coefficients ϵ_s, ϵ_c, and ϵ_d also do not show significant influence of the model performance, under the condition that they are in a reasonable range.

Furthermore, Fig. 7 shows the testing accuracies in different partial domain adaptation tasks by the proposed method, where the time and frequency-domain data are used as the model inputs. Specifically, corresponding with the frequency-domain data, $N_{\text{input}} = 1024$ is adopted for the time-domain samples. It can be observed that in most cases, the frequency-domain inputs lead to noticeable better performance than the time-domain inputs. It is also noted that in the relatively balanced tasks where the classes in the source and target domains mostly match, such as tasks C_1, C_2, and T_1, the performances of both the inputs are basically the same. That shows the time-domain data are able to achieve promising results in the typical fault diagnosis problems. However, in more difficult partial tasks where the target-domain classes are largely biased, such as tasks C_7, C_9, and T_9, the frequency-domain data generally achieve significant higher testing accuracies. The reason lies in that the frequency-domain data can be basically considered the higher level features of the raw time-domain vibration acceleration data. That is a well-established approach for conventional signal processing. Compared with the time-domain data, it is easier for the network to automatically learn the discriminative features from the frequency-domain data. Therefore, the frequency-domain signals are preferred in this study.

In the proposed method, the diverse fault diagnosis knowledge learned by multiple classification modules is leveraged for the partial domain adaptation problem. The effects of the number

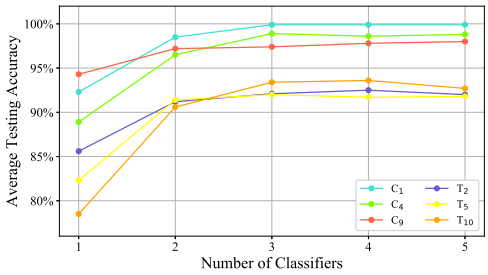

Fig. 8. Testing accuracies of the proposed method using different number of classification modules in different tasks.

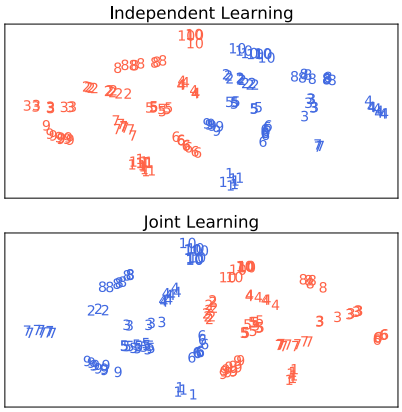

Fig. 9. Visualizations of the high-level representations of the source-domain data in classification modules 1 (blue) and 2 (red) on task C_6. Different numbers denote the machine health condition classes.

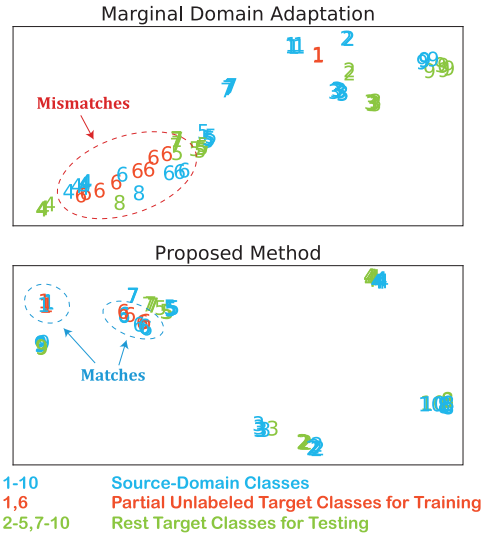

Fig. 10. Visualizations of the learned high-level representations of the source and target domains on task C_{10}. The MargDA and proposed methods are evaluated.

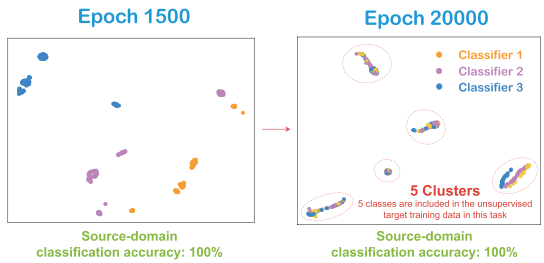

Fig. 11. Visualizations of the predicted probabilities of the unsupervised target training data by different classification modules on task C_6. The results at different training epochs are presented.

of the classification modules on the testing performance are investigated and shown in Fig. 8.

It can be observed that generally more classification modules lead to higher testing accuracies in different fault diagnosis tasks, and the performance is mostly stable with $N_m \geq 3$. It should be noted that when only one classification module is adopted, the proposed method becomes the NoPredConsist approach. Considering more classification modules also results in larger model size and heavier computational load, three classification modules are used by default in this study.

2) Feature Visualization: In this section, the popular t-SNE approach is used to visualize the learned diagnostic features by different methods. Fig. 9 shows the visualization results of the high-level representations of the source-domain data in classification modules 1 and 2 on task C_6. Both the scenarios where the classification modules are jointly trained using the proposed method and independently trained using the SupOnly method are presented. It can be observed that diverse fault diagnosis features are learned by different classification modules in both the cases, even in the presence of joint training. This suggests that unique knowledge can be learned by different modules, and their joint training facilitates knowledge exchange to achieve better generalization performance. It should be pointed out that while different modules share the same network structure in this study, the learned features are not the same in practice. This is because in typical deep neural networks with large amounts of parameters, the loss surface is quite complex with lots of local minima and saddle points. During training, limited modules generally cannot reach the same optimization area. Therefore, as validated in the visualizations, different features can be learned by different modules for diagnostics.

Fig. 10 shows the high-level feature visualization results of the classification module 1 on task C_{10}. Using the popular marginal domain adaptation method, i.e., the MargDA approach, the limited target training data of classes 1 and 6 are forced to match the whole source domain. Consequently, class-level mismatches are obtained. Using the proposed method, promising conditional data alignments between domains are achieved, which indicates the proposed method is well suited for the partial domain adaptation problems.

Furthermore, the prediction results of the proposed method on the unsupervised target-domain training data by the three classification modules are shown in Fig. 11, where task C_6 is investigated. It is noted that in the early period, while the source-domain samples can be accurately classified, the prediction results on the unsupervised target data are not consistent for different classification modules, that results in low cross-domain diagnosis performance. Through optimization iterations, the target prediction consistency can be achieved, enhancing the model generalization ability.

V. Conclusion

This article proposed a deep learning-based partial domain adaptation method for machinery fault diagnosis. Multiple classification modules were used to learn diverse diagnostic knowledge, and conditional data alignment scheme was adopted to extract domain-invariant features for the healthy state data in different domains. For the unsupervised partial target-domain training data, the prediction consistency method was proposed to achieve similar prediction probabilities. Experiments were carried out for validation.

It should be pointed out that the sample size has noticeable influence on the model performance. In general, larger sample contains more information for diagnostics. However, higher risk of overfitting also occurs. Therefore, a medium size was suggested in the applications. Because mini-batch optimization was used in the proposed method, the influence of the ratio of target samples to source samples was not significant on the model performance. Moreover, it was assumed that the target-domain classes were included in the source-domain classes in this study. In practice, it is possible that some additional faults may occur in the target condition, which do not belong to the source classes. This challenging problem was denoted as the open-set domain adaptation [43], which was an extension of the partial domain adaptation problem and will be focused on in the following research works.

References

[1] R. Razavi-Far, E. Hallaji, M. Farajzadeh-Zanjani, and M. Saif, "A semi-supervised diagnostic framework based on the surface estimation of faulty distributions," *IEEE Trans. Ind. Informat.*, vol. 15, no. 3, pp. 1277–1286, Mar. 2019.

[2] H. Ma, J. Zeng, R. Feng, X. Pang, Q. Wang, and B. Wen, "Review on dynamics of cracked gear systems," *Eng. Failure Anal.*, vol. 55, pp. 224–245, 2015.

[3] K. Yu, T. R. Lin, H. Ma, H. Li, and J. Zeng, "A combined polynomial chirplet transform and synchroextracting technique for analyzing nonstationary signals of rotating machinery," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 4, pp. 1505–1518, Apr. 2020.

[4] W. Zhang, X. Li, X.-D. Jia, H. Ma, Z. Luo, and X. Li, "Machinery fault diagnosis with imbalanced data using deep generative adversarial networks," *Measurement*, vol. 152, 2020, Art. no. 107377.

[5] Z. Luo, J. Wang, R. Tang, and D. Wang, "Research on vibration performance of the nonlinear combined support-flexible rotor system," *Nonlinear Dyn.*, vol. 98, no. 1, pp. 113–128, 2019.

[6] Y. Qin, X. Wang, and J. Zou, "The optimized deep belief networks with improved logistic sigmoid units and their application in fault diagnosis for planetary gearboxes of wind turbines," *IEEE Trans. Ind. Electron.*, vol. 66, no. 5, pp. 3814–3824, May 2019.

[7] Z. Chen, K. Gryllias, and W. Li, "Intelligent fault diagnosis for rotary machinery using transferable convolutional neural network," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 339–349, Jan. 2020.

[8] G. Csurka, "Domain adaptation for visual applications: A comprehensive survey," 2017, *arXiv:1702.05374*.

[9] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, vol. 37, pp. 97–105.

[10] X. Li, W. Zhang, N. Xu, and Q. Ding, "Deep learning-based machinery fault diagnostics with domain adaptation across sensors at different places," *IEEE Trans. Ind. Electron.*, vol. 67, no. 8, pp. 6785–6794, 2020.

[11] S. Shao, S. McAleer, R. Yan, and P. Baldi, "Highly-accurate machine fault diagnosis using deep transfer learning," *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2446–2455, Apr. 2019.

[12] F. Jia, Y. Lei, N. Lu, and S. Xing, "Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization," *Mech. Syst. Signal Process.*, vol. 110, pp. 349–367, 2018.

[13] X. Li, W. Zhang, and Q. Ding, "Cross-domain fault diagnosis of rolling element bearings using deep generative neural networks," *IEEE Trans. Ind. Electron.*, vol. 66, no. 7, pp. 5525–5534, Jul. 2019.

[14] B. Yang, Y. Lei, F. Jia, and S. Xing, "An intelligent fault diagnosis approach based on transfer learning from laboratory bearings to locomotive bearings," *Mech. Syst. Signal Process.*, vol. 122, pp. 692–706, 2019.

[15] M. He and D. He, "Deep learning based approach for bearing fault diagnosis," *IEEE Trans. Ind. Appl.*, vol. 53, no. 3, pp. 3057–3065, May–Jun. 2017.

[16] X. Jin, M. Zhao, T. W. S. Chow, and M. Pecht, "Motor bearing fault diagnosis using trace ratio linear discriminant analysis," *IEEE Trans. Ind. Electron.*, vol. 61, no. 5, pp. 2441–2451, May 2014.

[17] W. Sun, R. Zhao, R. Yan, S. Shao, and X. Chen, "Convolutional discriminative feature learning for induction motor fault diagnosis," *IEEE Trans. Ind. Informat.*, vol. 13, no. 3, pp. 1350–1359, Jun. 2017.

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

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

[20] X. Wang, H. He, and L. Li, "A hierarchical deep domain adaptation approach for fault diagnosis of power plant thermal system," *IEEE Trans. Ind. Informat.*, vol. 15, no. 9, pp. 5139–5148, Sep. 2019.

[21] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.

[22] G. Xu, M. Liu, Z. Jiang, W. Shen, and C. Huang, "Online fault diagnosis method based on transfer convolutional neural networks," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 2, pp. 509–520, Feb. 2020.

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

[24] W. Lu, B. Liang, Y. Cheng, D. Meng, J. Yang, and T. Zhang, "Deep model based domain adaptation for fault diagnosis," *IEEE Trans. Ind. Electron.*, vol. 64, no. 3, pp. 2296–2305, Mar. 2017.

[25] L. Wen, L. Gao, and X. Li, "A new deep transfer learning based on sparse auto-encoder for fault diagnosis," *IEEE Trans. Syst. Man Cybern.: Syst.*, vol. 49, no. 1, pp. 136–144, Jan. 2019.

[26] X. Li, W. Zhang, and Q. Ding, "A robust intelligent fault diagnosis method for rolling element bearings based on deep distance metric learning," *Neurocomputing*, vol. 310, pp. 77–95, 2018.

[27] A. Zhang and X. Gao, "Supervised dictionary-based transfer subspace learning and applications for fault diagnosis of sucker rod pumping systems," *Neurocomputing*, vol. 338, pp. 293–306, 2019.

[28] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 2962–2971.

[29] L. Guo, Y. Lei, S. Xing, T. Yan, and N. Li, "Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data," *IEEE Trans. Ind. Electron.*, vol. 66, no. 9, pp. 7316–7325, Sep. 2019.

[30] X. Li, W. Zhang, Q. Ding, and X. Li, "Diagnosing rotating machines with weakly supervised data using deep transfer learning," *IEEE Trans. Ind. Informat.*, vol. 16, no. 3, pp. 1688–1697, Mar. 2020.

[31] R. Yan, F. Shen, C. Sun, and X. Chen, "Knowledge transfer for rotary machine fault diagnosis," *IEEE Sensors J.*, to be published.

[32] H. Zheng *et al.*, "Cross-domain fault diagnosis using knowledge transfer strategy: A review," *IEEE Access*, vol. 7, pp. 129260–129290, 2019.

[33] Z. Cao, M. Long, J. Wang, and M. Jordan, "Partial transfer learning with selective adversarial networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 2724–2732.

[34] J. Zhang, Z. Ding, W. Li, and P. Ogunbona, "Importance weighted adversarial nets for partial domain adaptation," 2018, *arXiv:1803.09210*.

[35] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, 2012.

[36] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 5400–5409.

[37] Y. Li, K. Swersky, and R. Zemel, "Generative moment matching networks," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 1718–1727.

[38] A. Gretton *et al.*, *Optimal Kernel Choice for Large-Scale Two-Sample Tests*. Curran Associates, Inc., 2012.

[39] Y. Ganinand V. Lempitsky, "Unsupervised domain adaptation by backpropagation," 2015, *arXiv:1409.7495*.

[40] Y. Ganin*et al.*, "Domain-adversarial training of neural networks," 2015, *arXiv:1505.07818*.

[41] W. A. Smith and R. B. Randall, "Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study," *Mech. Syst. Signal Process.*, vol. 64–65, pp. 100–131, 2015.
[42] D. Kingmaand J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[43] P. P. Busto, A. Iqbal, and J. Gall, "Open set domain adaptation for image and action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 413–429, Feb. 2020.

Xiang Li received the B.S. degree in engineering mechanics and engineering management and the Ph.D. degree in mechanics from Tianjin University in 2012 and 2017, respectively.

He was a Visiting Scholar with the University of California at Merced, USA, in 2015. He is currently a Postdoctoral Fellow with the Department of Mechanical Engineering, University of Cincinnati, USA, and an Assistant Professor with the College of Sciences, Northeastern University, China. His research interests include deep learning, fault diagnosis, prognostics and health management, multiobjective optimization algorithm, and intelligent transportation system.

Wei Zhang received the Ph.D. degree in mechanics from Tianjin University in 2017.

She was a Visiting Scholar with Texas A&M University, USA, in 2015. She is currently an Assistant Professor with the School of Aerospace Engineering, Shenyang Aerospace University, China. Her research interests include rotor dynamics, multiobjective optimization algorithm, and squeeze film damper.