
Article:
Liu, Z.-H., Jiang, L.-B., Wei, H.-L. orcid.org/0000-0002-4704-7346 et al. (2 more authors) (2021) Optimal transport based deep domain adaptation approach for fault diagnosis of rotating machine. IEEE Transactions on Instrumentation and Measurement, 70. 3508912. ISSN 0018-9456

https://doi.org/10.1109/tim.2021.3050173

Manuscript accepted for publication (29 December 2020) in IEEE Transactions on Instrumentation and Measurement

Optimal Transport Based Deep Domain Adaptation Approach for Fault Diagnosis of Rotating Machine

Zhao-Hua Liu, *Member, IEEE*, Lin-Bo Jiang, Hua-Liang Wei, Lei Chen, and Xiao-Hua Li

***Abstract*—Rotating machinery working under changing operation conditions is prone to failure. In recent years, domain adaptation has been successfully used for fault diagnosis. However, the existing fault diagnosis methods based on domain adaptation have two main disadvantages: 1) With these methods, it is difficult to precisely measure and estimate the differences between the source and target domains; 2) They only consider the discrepancies in the feature space, but not in the label space. In this paper, a new optimal transport based deep domain adaptation model is proposed for rotating machine fault diagnosis. The framework of the proposed method comprises three main components. Firstly, an autoencoder network is designed to extract compact and class discriminative features from the raw data. Secondly, the domain-invariant representation features are trained by searching an optimal transport plan with a predefined cost function between source and target domains and by minimizing the discrepancies of a joint distribution of the feature and label spaces based on optimal transport. Finally, the classifier trained with data in the source domain is directly used to perform the classification task in the target domain. In addition, the optimal selection of the model hyper-parameters is verified through empirical analysis, and the transfer ability of the proposed model is visually illustrated in a reduced feature space. The experimental results show that the proposed method outperforms the existing machine learning and domain adaptation fault diagnosis methods, in terms of, e.g., classification accuracy and generalization ability.**

***Index Terms*—Autoencoder, deep learning, domain adaptation, fault diagnosis, optimal transport, rotating machine, transfer learning.**

I. Introduction

ROTATING machine plays an important role in industrial application, and it is an integral part of many industrial systems (e.g., wind turbines, aircraft engines, and alternators). In any rotating machine application, fault diagnosis is crucial for ensuring safe operation of the system, reducing machine downtime and saving maintenance costs [1], [2]. Therefore, rotating machine fault diagnosis has attracted extensive attention, and many fault diagnosis methods have been proposed.

Fault diagnosis techniques can be roughly categorized into two types: 1) model-based methods and 2) data-driven methods. Early fault diagnosis is mainly based on physical models, which can accurately describe how faults are linked to the associated industrial system [3], [4]. But this kind of fault diagnosis methods have two main disadvantages: 1) they are highly dependent on the priori knowledge of the system; 2) factors such as disturbance in the industrial operation process and some assumptions about the system (e.g., the form of noise and working conditions of the system) may be inappropriate, and these can result in uncertainty and misdiagnosis.

Data driven fault diagnosis methods [5], [6], where analysis is directly performed on the collected data using techniques such as signal processing and machine learning, can reduce the dependence on prior knowledge of the system and are more suitable for modern industrial application. For instance, Wang *et al.* [7] used the Hilbert transform to analyze fault features in the frequency domain. Singh *et al.* [8] proposed a fault diagnosis method based on wavelet analysis. In addition to traditional signal processing techniques, new fault diagnosis methods based on statistical learning have also been investigated. For example, Cao *et al.* [9] proposed a coupled hidden Markov model to identify the bearing fault stages. Wang *et al.* [10] presented a classifier algorithm combining wavelet packet decomposition with random forests, which can extract the fault features and reduce the influence of vibration signal noise. Although these methods can achieve good performance on fault diagnosis tasks, they still face a challenge to automatically extract the fault features from incipient fault signals. In reality, it is time-consuming to extract fault features manually. Therefore, there is a need to develop an efficient intelligent fault diagnosis algorithm, which can efficiently detect the fault diagnosis and automatically discover the fault features.

In recent years, great progress has been made in applying

Manuscript received September 15, 2020; revised October 29, 2020, December 11, 2020, and accepted December 29, 2020. This work was supported in part by the National Key Research and Development Project under Grant 2019YFE0105300, in part by the National Natural Science Foundation of China under Grant 61972443, in part by the Hunan Provincial Hu-Xiang Young Talents Project of China under Grant 2018RS3095, in part by the Hunan Provincial Natural Science Foundation of China under Grant 2020JJ5199.

H.-L. Zhao, L.-B. Jiang, L. Chen, and X.-H Li are with the School of Information and Electrical Engineering, Hunan University of Science and Technology, Xiangtan 411201, China (e-mail: zhaohualiu2009@hotmail.com; 3065649450@qq.com; chenlei@hnust.edu.cn; lixiaohua_0227@163.com).

H.-L. Wei is with the Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield S1 3JD, U.K. (e-mail: w.hualiang@sheffield.ac.uk).

deep learning in the field of fault diagnosis. Deep neural networks, including deep stacking networks and convolution neural networks (CNN), due to their excellent automatic feature extraction ability, have been proposed to solve the fault diagnosis problem. Sun *et al.* [11] proposed a sparse deep stacking network to overcome the feature extraction problem of conventional deep learning by adding a sparse regularization term to enforce the feature to be sparse; this can efficiently train the model and achieve better performance. Liu *et al.* [12] proposed a dislocated time series CNN architecture that can deal with the time series data in industrial application. Sun *et al.* [13] presented a convolutional discriminative feature learning networks based on CNN to automatically learn robust and invariant fault features. Chen *et al.* [14] proposed a deep learning scheme based on sparse autoencoder (SAE) and deep belief network (DBN), which combined time-domain and frequency-domain features to enhance the bearing fault diagnosis reliability. Shao *et al*. [15] proposed a deep learning-based multi-signal fault diagnosis method which showed robust performance by extracting features from multiple types of sensor signals. Ma *et al.* [16] proposed a deep residual convolutional network based on separable convolution to learn multiscale information from vibration signals and obtained satisfactory diagnostic results. Most of the existing approaches can precisely detect the fault and can automatically learn hierarchical representation from data based on deep learning network, but the success of these methods is based on two assumptions: 1) there is a large amount of labeled data for training; 2) the training data from the source domain and the test data from the target domain follow the same distribution. When one or both of the two assumptions are violated, the performance of these algorithms may degrade significantly. In practice, these assumptions tend to be violated due to the variable working environment and unstable load torque in real industrial applications such as wind power system. It is known that data collection is time-consuming, so it is often unrealistic to collect a large amount of labeled training data. It is also commonly aware that models trained from scratch data may be less reliable. Even if labeled data can be obtained under certain working conditions, the data distribution may change with other new working conditions. To overcome these issues, a good solution is to transfer knowledge from source domain with a large amount of label training data to target domain with relatively smaller or fewer number of labeled data.

Domain adaptation is an efficient method to solve data imbalance or data scarce problem and has attracted increasing attention. It has proven that with the domain adaption algorithms [17]-[19], the performance of a learning approach in the target domain can be bounded by the performance of the learning algorithms in source domain, and the discrepancies between source domain and target domain can be reduced or minimized. Methods and algorithms for reducing the discrepancies between the two domains have been well studied in machine learning community [20], [21]. As an effective approach, domain adaptation technology can overcome the weakness of deep learning. For example, in fault diagnosis application, when the working condition varies, a traditional deep learning model trained from data collected under a specific scratch working condition usually does not work for other working conditions. In order to obtain a good model that works for different working conditions, the model must be trained using a massively large amount of data collected under different conditions; this implies a large amount of cost spent on data collection. With domain adaption approach, however, a model trained from data collected under certain working conditions does not need to be retrained when it is applied to data collected under some new working conditions, so as to reduce the cost of data collection and the corresponding training time and computational cost. Due to these reasons, domain adaptation has been widely used to solve the fault diagnosis problem under different working conditions [22]. In [23], a fault diagnosis model based on deep neural network was presented to extract transferable features by utilizing an autoencoder (AE) network with a maximum mean discrepancy (MMD) term. In [24], an AE model, together with a joint distribution adaptation (JDA) term, was proposed to extract domain-invariant and discriminative features. To transfer the diagnosis model from the source domain to the target domain, Wang *et al.* [25] adopted a stacked denoising AE by replacing the correlation alignment (CORAL) distance with the differences of the two data distributions. Sohaib *et al.* [26] used bispectrum analysis and a CNN model to identify bearing fault under inconsistent working conditions. Most of the existing work can effectively mitigate the impact of data distribution differences and significantly improve the performance of the associated classifiers. However, the existing fault diagnosis methods based on domain adaptation theory have two main disadvantages: 1) when the distribution between the source domain and the target domain is multimodal, the probability metrics used in these methods cannot represent the differences between the two domains; 2) the differences between the source and target domains are only considered in high-dimensional feature space, but the differences in the label space are usually ignored, leading to inadequate adaptation. In order to solve the above problem, a more effective probability metric needs to be proposed to precisely measure the discrepancies.

Optimal transport (OT) theory [27], [28], as a powerful tool, can compute the distance between the probability distributions. OT distance is also called Wasserstein distance, Monge-Kantorovich distance, or Earth Mover distance. The OT distance can be directly computed based on the samples of the distributions without performing density estimation or other non-parametric methods. As described in [29], OT outperforms traditional probability metrics such as Kullback-Leibler (KL) divergence and Jensen-Shannon (JS) divergence. For example, KL divergence may be infinite, and JS divergence can have a sudden step when there is no overlap between the distributions, but OT distance can still provide useful results for parameter update with gradient descent method. In addition, OT can also be used in other metric spaces. However, computing the OT distance involves the use

of linear programming, which requires high computational costs. Several efficient algorithms [30], [31] have been proposed to solve the problem. OT theory has been widely used in the fields of image processing [32], [33], domain adaptation [34], [35], and signal processing [36], [37].

In this paper, a novel method called optimal transport based deep domain adaptation (OTDDA) is proposed to simultaneously align the distribution of the feature space and label space of the source domain and the target domain. First, an AE network is constructed for unsupervised feature learning in the proposed model, which can extract the class discriminative features from the input data. Then, the discrepancies between the joint learning representations and labels is minimized based on OT. The main idea is to search a transport plan between the feature and label spaces of the source and target domains and retain the label information of the source domain. Experimental results show that our method can precisely detect faults under different working conditions. To the best of our knowledge, this is the first time to apply OT theory to solve fault diagnosis problems using domain adaptation learning. The main contributions of the paper are summarized as follows:

1) An optimal transport based deep domain adaptation (OTDDA) framework is proposed for rotating machine fault diagnosis under different working conditions. The method can automatically extract the domain-invariant and discriminative features from the raw data and precisely detect faults under different working conditions.

2) In order to measure the discrepancies between the labels and features in the source and target domains, the OT distance is introduced to characterize the discrepancies between the source and target domains.

3) To improve the performance of the fault diagnosis, an AE network is designed as a feature extractor to discover representative features in the raw data. Then the transport plan is learned in the hidden layer of the AE network to align the representation in the source and target domains. In addition, the mini-batch method is used to solve the high computing cost problem in computing OT distance.

The rest of the paper is organized as follows. In Section II, the theory and framework of the proposed method, referred to as the optimal transport based deep domain adaptation (OTDDA), is depicted. In Section III, the implementation algorithm of OTDDA, used for fault diagnosis, is presented. The experimental results are described in Section IV. Finally, the main work is summarized in Section V.

II. The Proposed Optimal Transport Based Deep Domain Adaptation

A. Problem Formulation in Fault Diagnosis

The problem of domain drift and domain adaptation is shown in Fig. 1. Domain drift can occur once the working conditions change. Therefore, a classifier trained using data in the source domain cannot be directly applied to the target domain as the direct use of the classifier can lead to misclassification. In this work, the fault diagnosis setting is as follows: the labeled training data from the source domain are collected under a specific working condition, and the unlabeled target domain data are collected under another working condition. This study is mainly concerned with unsupervised domain adaptation problem. The goal of the domain adaptation is to align the distribution of the source and target domains and train a classifier using the source domain data, and the classifier is then applied to the target domain.

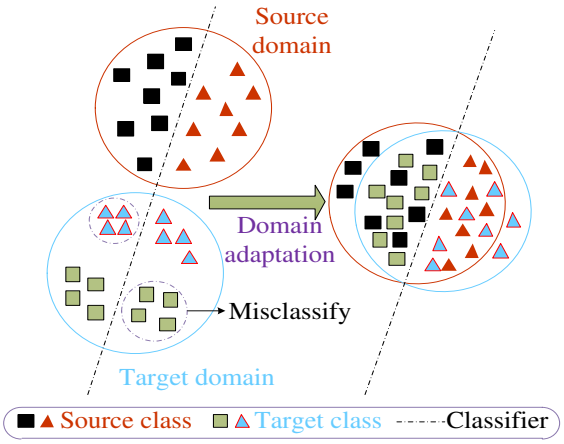


Fig. 1. The graphical representation of the problem of the domain drift and domain adaptation.

Definition 1 (Domain): A domain D is composed of a m-dimensional feature space X and a marginal probability distribution $P(x)$, i.e., $D=\{X,P(x)\}$, where $x \in X$.

Definition 2 (Task): Given a domain D, a task T is composed of a C-cardinality label set Y with a distribution $P(y)$ and a function map $f(x)$, i.e., $T=\{f(x),Y\}$, where $y \in Y$, and $f(x)=Q(y \mid x)$, which can be regarded as the conditional probability distribution.

Unsupervised domain adaptation: Given a labeled source domain $D_s=\{X_s,P_s(x_s)\}$ with label Y_s and unlabeled target domain $D_t=\{X_t,Q_t(x_t)\}$, where $P_s(x_s,y_s) \neq Q_t(x_t,y_t)$, and $X_s=X_t$, the goal of unsupervised domain adaptation is to utilize the labeled source data to learn a mapping $f(x): X \to Y$, which has a good performance on the target domain.

B. Optimal Transport

1) Original optimal transport problem

The original OT problem, called Monge problem, was first studied by the French mathematician Gaspard Monge in the middle of the 19th century. The goal of the OT is to find a lowest-cost way to transport large amounts of sands into a given hole, as shown in Fig. 2. The formal definition is introduced as follows.

Let $\Omega \in R^d$ be an input measurement space. Assume that a mapping T needs to be found to minimize the cost $C(T)$ as follows:

$$C(T)=\int_{\Omega} c(x,T(x))d\mu(x) \tag{1}$$

where the cost function $c:\Omega \times \Omega \to R^+$ is a distance function

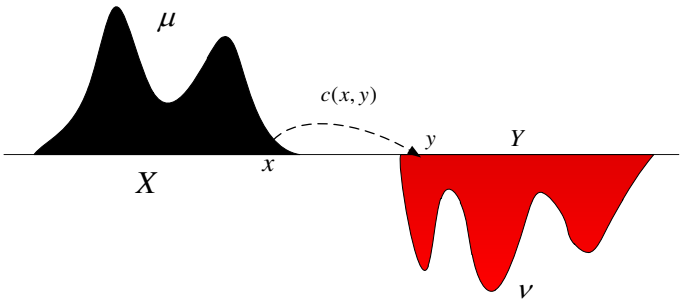


Fig. 2. The graphical representation of the OT problem.

over the metric space Ω. The solution to the Monge problem can be formulated as:

$$T_0 = \arg\min_T \int_{\Omega_s} c(x, T(x)) d\mu_s(x) \quad subject\ to\ T\#\mu_s = \mu_t \tag{2}$$

where $T\#\mu_s = \mu_s(T^{-1}(x))$, which is said to be a transport map or push-forward from μ_s to μ_t.

Kantorovitch reformulated the Monge problem and extended it to the more general case where a large amount of sands can be split into several parts [38]. Let $\prod$ be the set of all the probabilistic couplings in $P(\Omega_s \times \Omega_t)$ with marginal μ_s and μ_t. The Kantorovitch problem aims to find a coupling that satisfies the following equation:

$$\gamma_0 = \arg\min_{\gamma \in \prod} \int_{\Omega_s \times \Omega_t} c(x_s, x_t) d\gamma(x_s, x_t) \tag{3}$$

where γ_0 is a transportation plan.

The Wasserstein distance of order p between μ_s and μ_t is defined as:

$$W_p(\mu_s, \mu_t) = (\inf_{\gamma \in \prod} \int_{\Omega_s \times \Omega_t} c(x_s, x_t)^p d\gamma(x_s, x_t))^{\frac{1}{p}} \tag{4}$$

2) Optimal transport for domain adaptation

Assumed that the source and target domains have n_s and n_t samples, respectively. Suppose μ_s and μ_t are their corresponding marginals. The goal is to find a transformation $T: \mu_s \to \mu_t$ and retain the knowledge of source domain that is highly related to the OT problem. In this paper, only the discrete OT problem is concerned, and the domain adaptation problem can be regarded as a special case of the discrete OT problem. Therefore, equation (3) can be rewritten as:

$$\mu_s = \sum_{i=1}^{n_s} p_i^s \delta_{x_i^s} \tag{5}$$

$$\mu_t = \sum_{i=1}^{n_t} p_i^t \delta_{x_i^t} \tag{6}$$

$$\sum_{i=1}^{n_s} p_i^s = \sum_{i=1}^{n_t} p_i^t = 1 \tag{7}$$

$$\prod = \left\{ \gamma \in (R^+)^{n_s \times n_t} \mid \gamma 1_{n_t} = \mu_s, \gamma^T 1_{n_s} = \mu_t \right\} \tag{8}$$

$$\gamma_0 = \arg\min_{\gamma \in \prod} \langle C, \gamma \rangle_F \tag{9}$$

where $\delta_{x_i^s}$ and $\delta_{x_i^t}$ are the Dirac function at locations x_i^s and x_i^t, respectively, 1_d is a d-dimensional vector of ones, $\langle \cdot, \cdot \rangle_F$ is the Frobenius dot product, C is the cost matrix, and $C(i, j) = c(x_i^s, x_j^t)$. Once a transport plan γ_0 is obtained, the barycentric mapping can be used to transport the source samples to target samples or target samples to source samples, which can be described as

$$X_{st} = diag(\gamma_0 1_{n_t})^{-1} \gamma_0 X_t \tag{10}$$

$$X_{tt} = diag(\gamma_0^T 1_{n_s})^{-1} \gamma_0^T X_s \tag{11}$$

where X_{st} and X_{tt} are transformed source samples and target samples, respectively. X_s and X_t are data matrix that represent the source data and target data.

3) Joint distribution optimal transport loss

Courty *et al.* [35] proposed a joint distribution transportation using the discrepancies between the feature and label space when the domain changes, and introduced it for solving the transport problem. Ideally, for adapting the feature space and the label space, a metric that can be applied to both the feature and label spaces should be used. For feature space, the most commonly used metric is the l_2 distance, while the classification loss is usually used to measure the discrepancies in the label space. In this paper, the cross-entropy loss is calculated as the classification loss. The cost function is defined as

$$c(x_i^s, y_i^s, x_j^t, y_j^t) = \alpha d(x_i^s, x_j^t) + L(y_i^s, y_j^t) \tag{12}$$

where $d(\cdot, \cdot)$ represents the l_2 distance, $L(\cdot, \cdot)$ is the cross-entropy function, and α is the trade-off parameter. However, the target domain is unlabeled, which means that y_j^t cannot be directly used. Therefore, the classifier $g(x)$ trained in source domain is utilized to represent the target labels. The optimization objective function can be written as:

$$\min_{g, \gamma \in \prod} \langle C, \gamma \rangle = \sum_{i,j} \gamma_{i,j} (\alpha d(x_i^s, x_j^t) + L(y_i^s, g(x_j^t))) \tag{13}$$

C. The Proposed Optimal Transport Based Deep Domain Adaptation

Although joint distribution transportation can reduce the domain drift in the feature and label spaces, it has two drawbacks: 1) useful features cannot be automatically extracted from the raw data; 2) it is extremely difficult and intractable to solve γ when the dataset is large. To solve these two problems, a new OTDDA framework is proposed in this paper, which comprises three parts: 1) a feature extractor f, 2) a label predictor g, and 3) an OT solver. An overview of the proposed framework is shown in Fig. 3. Firstly, the feature extractor $f: x \to z$ is trained to map the input data to the latent space Z, where useful features can be extracted from the input data. Second, the label classifier $g: z \to y$ maps the feature space to the label space to classify the obtained features generated from the feature extractor. Finally, based on the

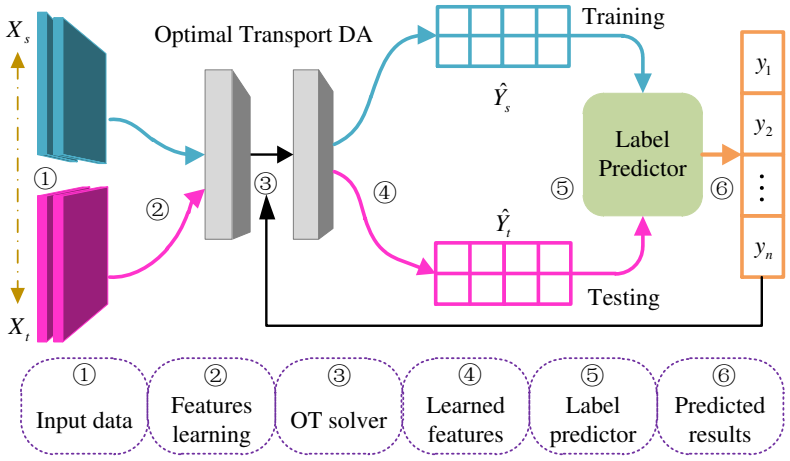


Fig. 3. The overview of the proposed optimal transport based deep domain adaptation framework.

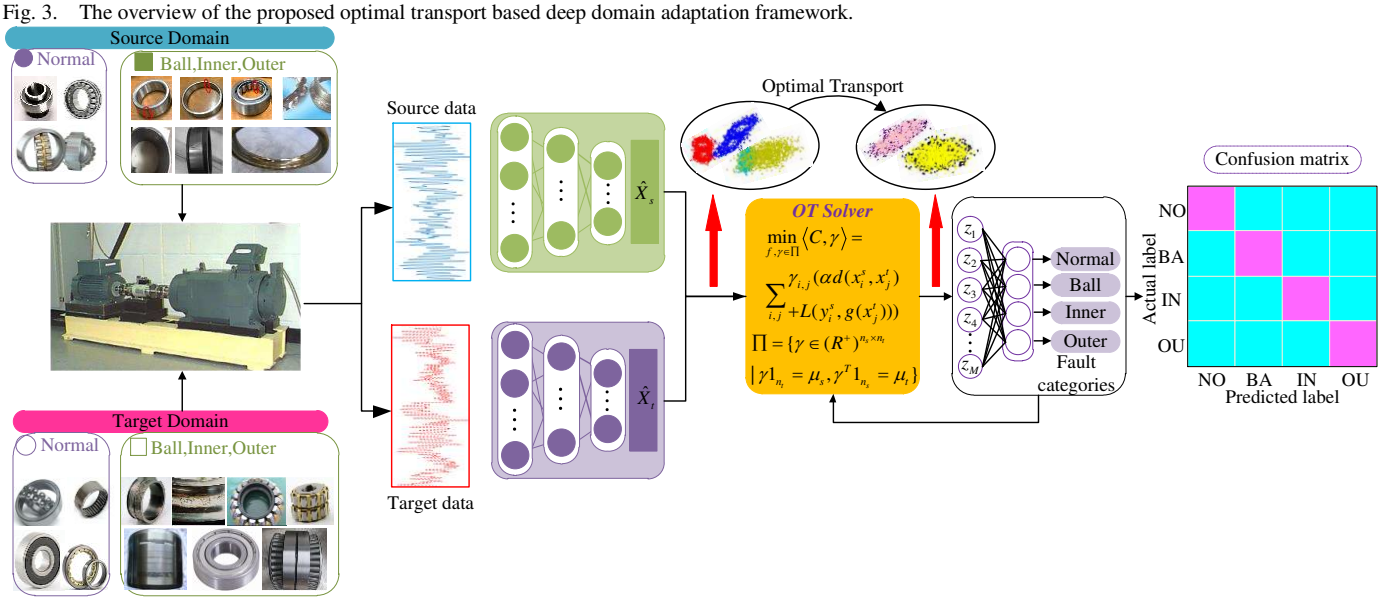


Fig. 4. The flow chart of the proposed OTDDA algorithm for fault diagnosis.

joint distribution loss, the OT solver is constructed to search a transport plan to align the distribution of the feature and label spaces relating to the source and target domains, respectively. The goal can be achieved by minimizing the following objective function:

$$\begin{aligned}\min_{f,g,\gamma\in\prod}\mathcal{L}(\gamma,f,g)=&\frac{1}{n_s}\sum_i L_s(y_i^s,g(f(x_i^s)))\\&+(\sum_{i,j}\gamma_{i,j}(\alpha d(f(x_i^s),f(x_j^t)))\\&+\sum_{i,j}\gamma_{i,j}(L_t(y_i^s,g(f(x_j^t)))))\end{aligned}\tag{14}$$

where the first term is to ensure that the proposed model can obtain a good performance in the source domain. The second term is to align the distribution of the feature space and the label space, corresponding to the source domain and target domain, respectively, and make the classifier trained in the source domain be directly applicable in the target domain.

III. Optimal Transport Based Deep Domain Adaptation for Fault Diagnosis

In order to improve the accuracy of fault diagnosis in real industrial production, the proposed OTDDA framework in section II is applied to detect the fault category of rolling bearings. To better deal with the fault data, the AE network is constructed as a feature extractor to extract the features of the fault data from the source and target domains. Then, a Softmax classifier is trained as a label predictor to predict the labels of these extracted features. The feature space and label space are aligned by the OT solver. The flow chart of the proposed OTDDA algorithm for fault diagnosis is shown in Fig. 4. The details of the algorithm are described below.

A. Autoencoder for Feature Learning

Feature extraction is crucial for improving the classification performance and accelerating the convergence for fault diagnosis. However, manual feature annotation is time-consuming and inefficient. Therefore, an efficient feature extractor is needed to automatically extract features and improve the convergence rate of the model. In this paper, an AE network is designed as a feature extractor to learn more training process, which can avoid extracting features from incipient fault signals. The architecture of the AE network is shown in Fig. 5.

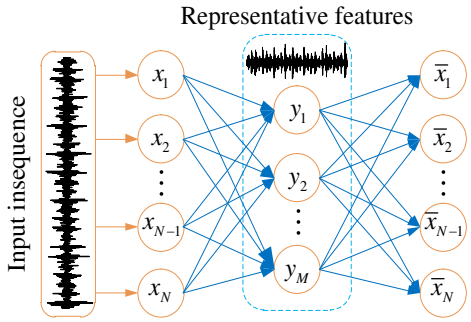


Fig. 5. The architecture of the AE network.

The AE network maps the input data $x \in \mathbb{R}^n$ to the representative features $y \in \mathbb{R}^m (m < n)$ by a function h_θ with parameters $\theta \triangleq \{W, b\}$, which is written as follows:

$$y = h_\theta(x) = f(Wx + b) \tag{15}$$

where W represents a $m \times n$ weight matrix, b is a bias vector, and f is an activation function.

Correspondingly, the decoding part is to reconstruct the input data with the representative features y, and the process is described as follows:

$$\bar{x} = g_{\theta'}(y) = f(W'y + b') \tag{16}$$

where $\theta' \triangleq \{W', b'\}$, W' represents a $n \times m$ weight matrix, and b' is a bias vector. Then, the training of the AE network updates W, W', b, and b' through the following loss function:

$$\frac{1}{2m}\sum_{i=1}^{m} || \bar{x}_i - x_i ||_2 \tag{17}$$

where $\bar{x}_i$ and x_i represent the i th reconstructed data and input data, respectively.

B. Softmax Classifier

Softmax classifier is the most commonly used algorithm for multiclass classification. Given input data $\{x^{(i)}\}_{i=1}^m$ and a label set $\{y^{(i)}\}_{i=1}^m$ consisting of k types of labels, where $x^{(i)} \in \mathbb{R}^n$, and $y^{(i)} \in \{1, 2, \cdots, k\}$, the main function of Softmax is to estimate the probability that each sample belongs to each category, and take the category with the highest probability as the category of the sample. This probability is given by:

$$\begin{aligned} p(y_{i=j}^{(i)} | x^{(i)}; \theta) &= \left[p(y^{(i)} = 1 | x^{(i)}; \theta), \ldots, p(y^{(i)} = k | x^{(i)}; \theta) \right]^T \\ &= \frac{1}{\sum_{l=1}^{k} e^{\theta_l^T x^{(i)}}} \begin{bmatrix} e^{\theta_1^T x^{(i)}} \\ e^{\theta_2^T x^{(i)}} \\ \cdots \\ e^{\theta_k^T x^{(i)}} \end{bmatrix} \end{aligned} \tag{18}$$

The cross-entropy of the predicted labels of Softmax and true labels is chosen as the cost function.

C. Training Strategy

In this section, the training strategy of the proposed model is described in detail as follows.

1) Initialization: It has been proven that the initialization strategy can critically affect the quality of the training. In this paper, the Xavier initialization [39] is adopted to initiate the parameters of the AE network and the cost matrix $C = d(x^s, x^t)$, where $C(i, j) = d(x_i^s, x_j^t)$.

2) Training procedure of the OTDDA model: Equation (14) involves two groups of variables to be optimized: 1) the OT coupling γ and 2) the parameters of feature extractor f and Softmax classifier g. The main task of the training is to search the transport plan; this is a linear programming problem which needs high computational cost. In this paper, the block coordinate descent algorithm [35] is used to optimize the cost function. In addition, a mini-batch method is applied to solve the computational cost problem of OT distance.

When the parameters of feature extractor f and Softmax classifier g are fixed, equation (14) can be rewritten as:

$$\begin{aligned} \min_{\gamma \in \Pi} \mathcal{L}(\gamma, f, g) = &\sum_{i,j} \gamma_{i,j} (\alpha d(f(x_i^s), f(x_j^t))) \\ &+ \sum_{i,j} \gamma_{i,j} (L_t(y_i^s, g(f(x_j^t)))) \end{aligned} \tag{19}$$

This is a standard linear programming problem, which can be solved by using the network simplex flow algorithm.

When the parameter γ is fixed, equation (14) can be rewritten as follows:

$$\begin{aligned} \min_{f,\, g} \mathcal{L}(\gamma, f, g) = &\frac{1}{n_s} \sum_i L_s(y_i^s, g(f(x_i^s))) \\ &+ (\sum_{i,j} \gamma_{i,j} (\alpha d(f(x_i^s), f(x_j^t))) \\ &+ \sum_{i,j} \gamma_{i,j} L_t(y_i^s, g(f(x_j^t)))) \end{aligned} \tag{20}$$

This is a classical deep learning problem that can be solved by using the gradient descent algorithm. In this paper, the Adam algorithm is adopted to optimize (20). The update rules are described as follows:

$$\theta_f \leftarrow \theta_f - \lambda \frac{\partial \mathcal{L}}{\partial \theta_f} \tag{21}$$

$$\theta_g \leftarrow \theta_g - \lambda \frac{\partial \mathcal{L}}{\partial \theta_g} \tag{22}$$

where λ represents the learning rate.

The detailed procedure of the proposed OTDDA algorithm for fault diagnosis is summarized in Algorithm 1.

IV. Experiment

A. Data Preparation

The proposed method is evaluated on the rolling bearing dataset collected under different working conditions. The bearing test rig is shown in Fig. 6 [40].

1) Bearing dataset: The rolling bearing data were collected from the single-point drive end of the bearing at room

Algorithm 1: OTDDA algorithm for fault diagnosis

Input: The labeled source data $D_s = \{x_1^s, x_2^s, ..., x_m^s\}$ corresponding to the label $Y_s = \{y_1^s, y_2^s, ..., y_m^s\}$ and the unlabeled target data $D_t = \{x_1^t, x_2^t, ..., x_n^t\}$.

Output: The feature extractor f and classifier g.

extract the useful features from raw fault data.

1: Set the f parameters:

- activation = 'sigmoid', loss = 'mean squared error', optimizer = 'adam', epochs = 600, batch size = 400.
- #hidden layer units = 400.

2: Initialize the cost matrix C, where $C(i, j) = d(x_i^s, x_j^t)$.

#minimize the discrepancies between the source and target domains.

3: **while** not convergence **do**

fix f and g to solve γ in (19)

fix γ to solve f and g in (20)

End while

4: **return** f and g.

Fig. 6. Bearing test rig [40].

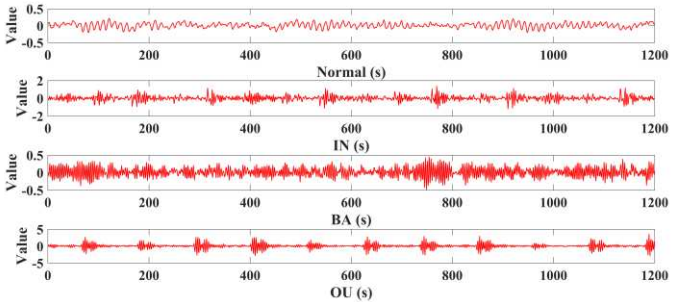


Fig. 7. The visualization of a sample for each type of data in the dataset.

temperature, where the accelerometer was used to get the normal and faulty data. Motor bearings were seeded with faults using electro-discharge machining (EDM). The faults mainly occur in three places: the inner race (IN), the outer race (OU), and the ball (BA). Each type of faults has four fault diameters, namely, 0.007 inches, 0.014 inches, 0.021 inches, and 0.028 inches. In addition, the motor loader was in one of the four classes: 0hp, 1hp, 2hp, and 3hp. The bearing data were collected at a frequency of 12,000 data points/second. In the setting, each 1200 data points are treated as a training/test sample segment, and each 800 sample segments are regarded as a unit of data. Therefore, there are totally 3200 normal sample segments in four working conditions (i.e, 0hp, 1hp, 2hp, and 3hp), and there are a total of 38400 fault sample segments involving all the four fault diameters and the four working conditions. As discussed in section II, the cross-domain fault data are deliberately created in this study to verify the performance of the proposed method. Based on the different types of motor loader, six cross-domain data are obtained, i.e., 0-1hp, 0-2hp, 0-3hp, 1-2hp, 1-3hp, and 2-3hp. The visualization of a normal bearing sample and three fault bearing sample segments with a size of 0.007 inches under 0hp working condition is shown in Fig. 7. The six domain shift cases are summarized in Table I. Taking 0-1hp as an example, the form of the domain adaptation problem is detailed as follows.

TABLE I
THE DESCRIPTION OF THE DOMAIN ADAPTATION TASKS

Task	Domain shift	Source classes	Target classes
D1	0hp->1hp	1, 2, 3, 4, 5, 6, 7	1, 2, 3, 4, 5, 6, 7
D2	0hp->2hp	1, 2, 3, 4, 5, 6, 7	1, 2, 3, 4, 5, 6, 7
D3	0hp->3hp	1, 2, 3, 4, 5, 6, 7	1, 2, 3, 4, 5, 6, 7
D4	1hp->2hp	1, 2, 3, 4, 5, 6, 7	1, 2, 3, 4, 5, 6, 7
D5	1hp->3hp	1, 2, 3, 4, 5, 6, 7	1, 2, 3, 4, 5, 6, 7
D6	2hp->3hp	1, 2, 3, 4, 5, 6, 7	1, 2, 3, 4, 5, 6, 7

Source domain: The source domain consists of the normal and faulty data collected without the motor loader $D_s = \{normal, IN, OU, BA\}$; the diameters of the fault selected are 0.007 inches and 0.014 inches. Therefore, there are totally 5600 sample segments in the source domain.

Target domain: Similar to the source domain, the target domain consists of the data collected from the 1hp motor loader without giving any label information.

Task: The task is to utilize the labeled source data to train a model that can categorize the unlabeled target data into four classes, namely, $\{normal, IN, OU, BA\}$. The number of sample segments in the target domain is the same as the source domain.

2) Data preprocessing: Firstly, for each cross-domain task, 5600 sample segments are chosen from the source and target domains with 80% overlap, respectively. The vibration data are noise contaminated. In order to reduce the effects of noise, the fast Fourier transform (FFT) is used to denoise the vibration data, and the value of the data after FFT is magnified 10 times because the FFT values are too small. The single-side frequency amplitude calculated in the last step is taken as the final input for training the model, and the input dimension is 600.

B. Experiment Results

1) Comparison with other methods: In order to verify the effectiveness of the OTDDA method, several supervised learning algorithms and domain adaptation methods are applied to the same data; these methods include SVM, K-nearest neighbor (KNN), Softmax, back-propagation neural network (BP), transfer component analysis (TCA) [20], JDA [21], CORAL [41], sparse autoencoder (SAE) with OT (SAE+OT), denoising autoencoder (DAE) [42] with OT (DAE+OT), and domain-specific batch normalization (DSBN) [43] The performance of the proposed method is compared

TABLE II
CLASSIFICATION ACCURACY OF ALL THE METHODS BASED ON THE ROLLING BEARING DATA

Without domain adaptation techniques							
Trail number	1	2	3	4	5	6	
Methods	0-1hp	0-2hp	0-3hp	1-2hp	1-3hp	2-3hp	Avg
Softnax	73.2%	71.4%	75.7%	69.5%	70.7%	66.1%	71.1%
KNN	82.3%	69.3%	77.4%	82.8%	85.6%	84.7%	80.3%
BP	75.2%	72.4%	76.4%	57.8%	55.3%	63.4%	66.7%
SVM	89.4%	83.2%	82.0%	70.3%	73.6%	86.2%	80.7%
With domain adaptation techniques							
TCA	82.8%	77.0%	87.2%	75.4%	84.8%	74.4%	80.3%
JDA	95.9%	92.0%	98.1%	89.7%	95.2%	93.7%	94.1%
CORAL	85.7%	82.9%	87.6%	74.9%	79.5%	70.1%	80.1%
DSBN	**97.2%**	**98.6%**	**98.8%**	**98.4%**	**98.9%**	**98.5%**	**98.4%**
SAE+OT	**95.6%**	**96.9%**	**97.2%**	**96.5%**	**97.4%**	**98.6%**	**97.0%**
DAE+OT	**96.2%**	**98.1%**	**98.9%**	**99.1%**	**98.6%**	**97.8%**	**98.1%**
OTDDA	**98.0%**	**99.8%**	**98.7%**	**98.9%**	**99.2%**	**98.4%**	**98.8%**

with that of these methods. The first four methods are classical supervised classification methods without using domain adaptation technique; they have been successful in many fault diagnosis applications. The remaining methods are domain adaptation methods, which have been widely used in computer vision.

2) Implementation details: Before training, the pretrained AE network is utilized as a feature extractor to extract useful features to preprocess the source and target data. For SVM, KNN, Softmax, and BP, the classifiers are trained on the labeled samples from source domain D_s and then used to predict the labels of the target samples from target domain D_t. The remaining methods based on domain adaptation use all the samples of D_s and D_t to learn a feature representation to align the distribution of the D_s and D_t. In addition, the Softmax is chosen as the base classifier of the domain adaptation methods.

For a fair comparison, the compared methods are evaluated by empirically searching the parameter space to find a set of optimal parameter settings and choose the best result for each method.

In this work, the SVM algorithm, provided by LIBSVM [44], is used. The bandwidth of the radical basis function (RBF) kernel is selected by searching from the set {1, 2, 4, 8, 16, 32}. Other methods are implemented in MATLAB environment (version R2016a). For KNN, the optimal nearest neighbor number is chosen by searching in the set {1, 2, 4, 8, 16, 32, 64}. For Softmax, the parameter of the regularization term is chosen from {0.0001, 0.001, 0.01, 0.1, 1, 10}. For BP, the number of hidden layers is set to 2, and the number of hidden neurons of each layer is set to 1000; the weight decay is searched from {0.002, 0.02, 0.2, 2}. For the domain adaptation based methods, the parameters of the base classifier are chosen using the same scheme as mentioned above for Softmax. For TCA and JDA, the optimal subspace dimensions are set to be {2, 4, 8, 16, 32, 64, 128, 256}, and the trade-off

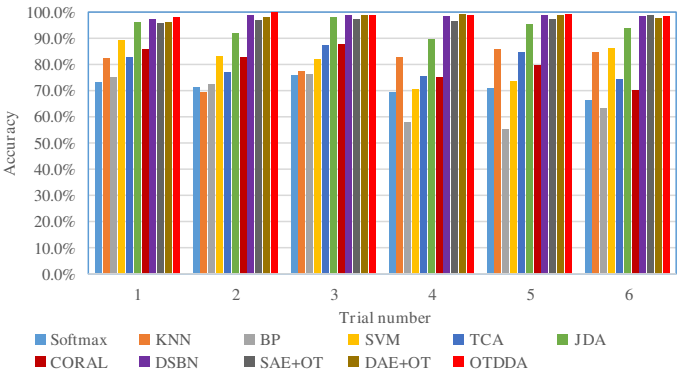


Fig. 8. The fault diagnosis accuracy of the 11 methods on bearing data.

parameter is chosen from {0.01, 0.1, 1, 10}. In addition, in order to determine which kinds of AE networks are best for the proposed framework, another two types of AE networks (i.e, SAE and DAE) are considered and compared with the models presented in section III. All the networks considered and compared have the same structure. For SAE, the KL divergence is added as a regularization term to the loss function to enforce the encoder feature to be sparse, and the sparse parameters are chosen from {0.001, 0.01, 0.1, 1, 10}; for DAE, some nodes of the input layer are set to be 0 randomly and are regarded as the input of AE network for training. For DSBN [43], the structure of CNN consists of two convolutional layers and a Softmax layer, the number of the convolution kernels is 10, and the size of convolution kernels is 5.

3) The performance metrics: In the experiment, the accuracy on the target data is calculated by (23) to measure the performance of all the methods.

$$accuracy = \frac{|x : x \in D_t \wedge f(x) = label(x)|}{|x : x \in D|} \tag{23}$$

The accuracy values generated from the 10 compared methods and the proposed method for the rolling bearing data using (23) are shown in Table II and Fig. 8.

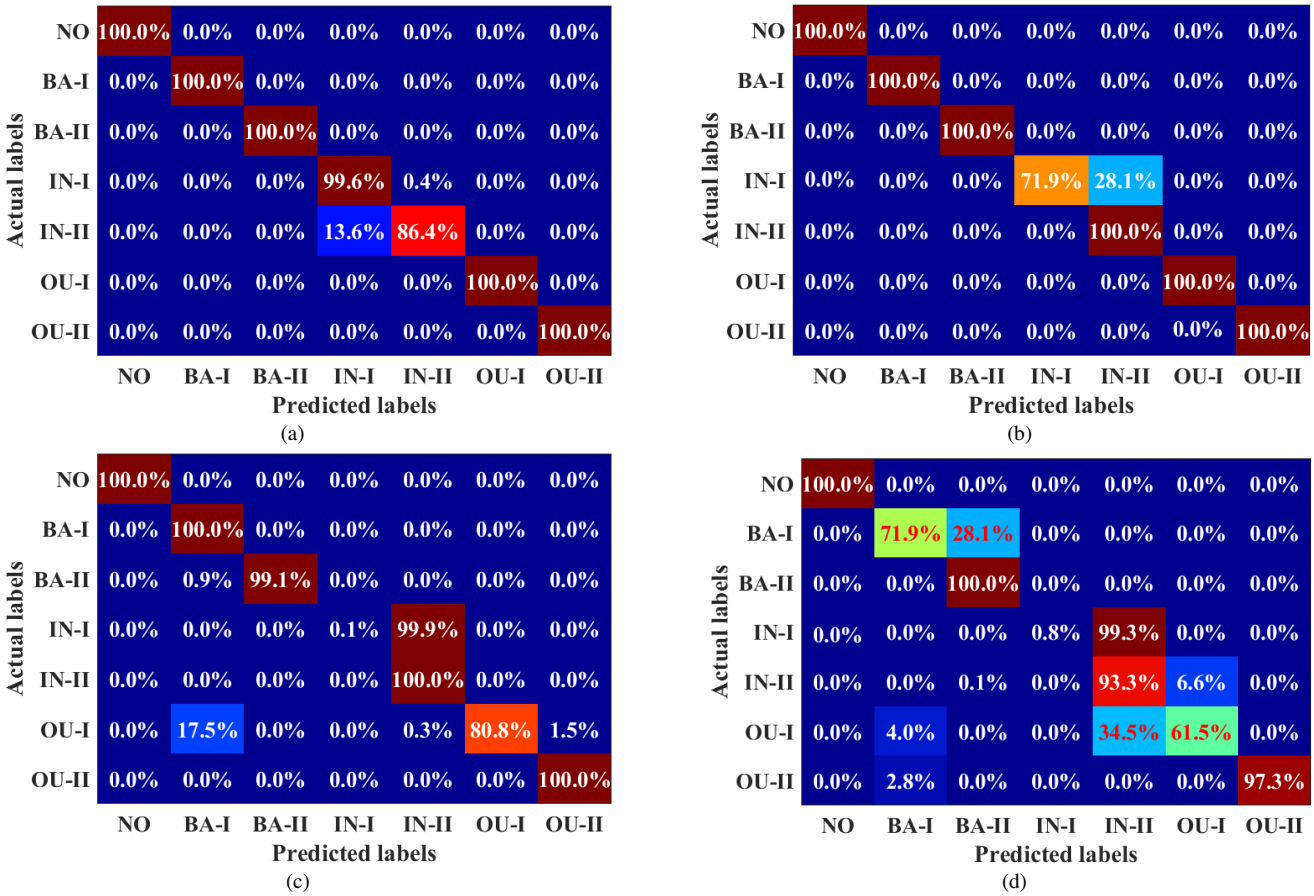


Fig. 9. The confusion matrices of domain adaptation methods for the trial number 1. (a) OTDDA. (b) JDA. (c) TCA. (d) CORAL.

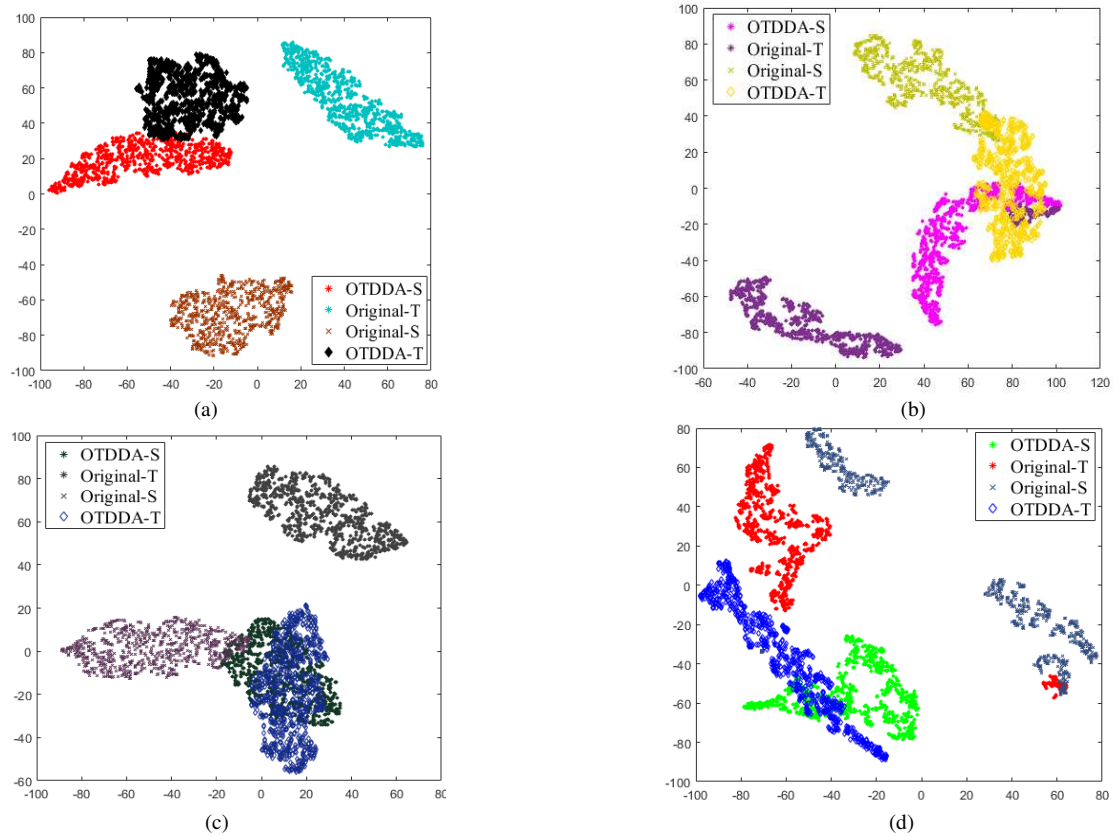


Fig. 10. The visualization results of the original features and OTDDA features in a reduced dimensional space by t-SNE, where the symbol S and T represent the source domain and target domain, respectively. In particular, the original features represent no domain adaptation. (a) Normal features. (b) IN features. (c) OU features. (d) Ball features.

4) Result analysis: As shown in Table II and Fig. 8, the performance of these methods without domain adaptation is in general lower than that with domain adaptation. It can be seen that the proposed OTDDA method can significantly improve the performance of the base classifier and is much better than other methods. This can be explained that the introduction of domain adaptation can improve the performance of the classifier on the target domain and mitigate the effect of the domain drift. For example, in trial number 2, the accuracy of Softmax is improved by 28.4% (from 74.4% to 99.8%) due to the use of domain adaption in OTDDA. More importantly, the domain adaptation method enables the model trained on source domain to be directly reused in the target domain, and reduces the labeled data collection cost and computational cost. Fig. 9 shows the confusion matrices of the proposed method and some other domain adaptation techniques for the first domain shift situation (i.e., trial number 1). For the choice of AE network, it can be seen that the original AE network in the proposed model is better than DAE and SAE. Although DAE and SAE can provide more powerful representation than original AE network, they reduce the transferability of model.

C. Empirical Analysis

The experimental results show that the proposed method is significantly better than the compared methods. In order to explore the reason of the better performance shown by OTDDA, the t-distributed stochastic neighbor embedding (t-SNE) [45] approach, as a dimension reduction visualization method, is used to reduce the dimension of features involved in the original feature space used by OTDDA. The two cases of 0hp and 1hp are selected as the source and target domain respectively, and the case of '0.007 inches' is selected as the diameter of the fault. In the experiment setting, the output of the second hidden layer of the AE network is designed to conduct the dimension reduction. The visualization result generated by t-SNE is shown in Fig. 10, where the Original-S and Original-T represent the features extracted from the source and target domains, respectively, by AE network without domain adaptation. Similarly, OTDDA-S and OTDDA-T represent the source and target features, respectively, extracted by the proposed method. The feature visualization of the four states (i.e., Normal, IN, OU, and Ball) is also shown in Fig. 10. It can be seen that the distance of the features between the source and target domains extracted by the OTDDA method becomes closer than that between the original features, and each source domain feature exactly matches the target domain feature. This shows that the proposed method can reduce the discrepancies of cross domain and align the distribution of source and target domain.

D. Parameter Analysis

In this section, the effects of the model parameters on the proposed method are analyzed. For illustration purpose, the four domain shift cases (i.e., trial number 1-6) are chosen as the input data of the experiment. The five key parameters considered are summarized in Table III.

TABLE III
THE KEY STRUCTURAL PARAMETERS OF OTDDA

Parameter	Value
Trade-off parameter α	0.5
Learning rate λ	0.01
Inputs size	600
Number of hidden layers	2
Number of hidden units	400

1) The trade-off parameter: The trade-off parameter α can crucially affect the performance of the proposed algorithm. In the experiment, the parameter α mentioned in (13) is mainly used to control the trade-off between the feature and label space. In addition, the classifier of the proposed OTDDA method is Softmax. For each domain shift case, ten different values are considered for α, which are from 0 to 0.9 in steps of 0.1; these values are used to test the classification accuracy of the proposed method. The experimental results are shown in Fig. 11, from which it is observed that the accuracy of the proposed method is at least 92.4%, and the best trade-off parameter is in the range of 0.4 to 0.6.

2) The number of hidden layers: In order to fully explore the potential of the proposed method, it is essential to analyze the impact of the number of hidden units in hidden layer. Fig. 12 shows the accuracy of the proposed method with different numbers of hidden units. The experimental results show that the proposed method is sensitive to the number of hidden units. It shows that the optimal number of hidden units is between 400 to 500. Therefore, 400 neurons are used in the hidden layer in this study.

V. CONCLUSION

In this paper, an optimal transport based deep domain

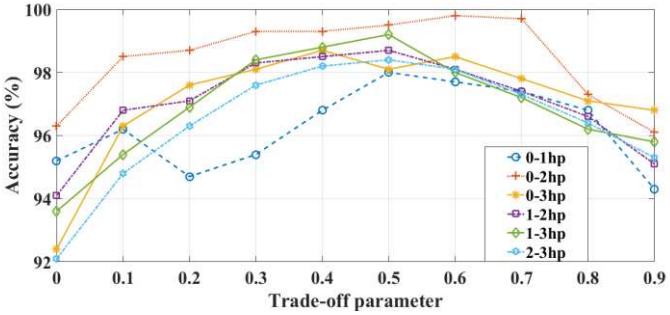


Fig. 11. The influence of the trade-off parameter for the accuracy of OTDDA (%).

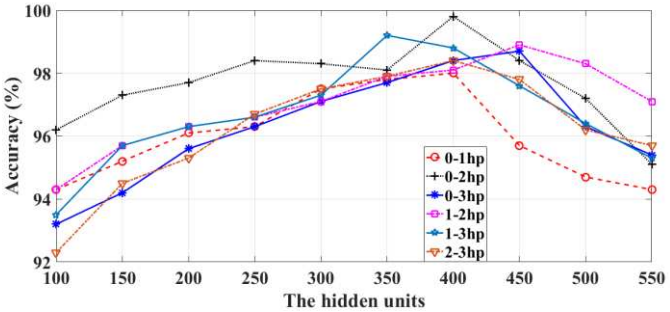


Fig. 12. The influence of the number of hidden units for the accuracy of OTDDA (%).

adaptation (OTDDA) framework was proposed for rotating machine fault diagnosis under different working conditions. The proposed method can automatically extract the domain-invariant and discriminative features from the vibration signals, and precisely detect faults under different working conditions. In order to reduce the heavy load of computing the OT distance, a mini-batch algorithm was applied to facilitate processing large-scale data. The main reason of the good performance of the OTDDA model was explored using a visualization analysis approach. the effects of model hyper-parameters were analyzed, and the experimental results show that the proposed model is sensitive to the change of the five key hyper-parameters considered.

In the future work, we will consider extending the proposed framework to other industrial systems, and develop an approach for automatically determining the hyper-parameters. Moreover, the existing domain adaptation fault diagnosis methods assume that the source and target domains have the same class of labels. Such an assumption, however, may not be valid in many real applications due to the existence of outliers and other factors. Therefore, the detection of outliers and the recognition of unknown labels become an important research question. These will be investigated in the future.

REFERENCES

[1] Z. W. Gao, C. Cecati, and S. X. Ding, "A survey of fault diagnosis and fault-tolerant techniques—Part II: fault diagnosis with knowledge-based and hybrid/active approaches," *IEEE Trans. Ind. Electron.*, vol. 62, no. 6, pp. 3768-3774, Jun. 2015.

[2] Z. W. Gao, C. Cecati, and S. X. Ding, "A survey of fault diagnosis and fault-tolerant techniques—Part I: fault diagnosis with model-based and signal-based approaches," *IEEE Trans. Ind. Electron.*, vol. 62, no. 6, pp. 3757-3767, Jun. 2015.

[3] D. U. Campos-Delgado and D. R. Espinoza-Trejo, "An observer-based diagnosis scheme for single and simultaneous open-switch faults in induction motor drives," *IEEE Trans. Ind. Electron.*, vol. 58, no. 2, pp. 671-679, Feb. 2011.

[4] H. K. Li, X. T. Lian, C. Guo, and P. S. Zhao, "Investigation on early fault classification for rolling element bearing based on the optimal frequency band determination," *J. Intell. Manuf.*, vol. 26, no. 1, pp. 189-198, Feb. 2015.

[5] S. Yin, X. W. Li, H. J. Gao, and O. Kaynak, "Data-based techniques focused on modern industry: an overview," *IEEE Trans. Ind. Electron.*, vol. 62, no. 1, pp. 657-667, Jan. 2015.

[6] S. Yin, S. X. Ding, X. C. Xie, and H. Luo, "A review on basic data-driven approaches for industrial process monitoring," *IEEE Trans. Ind. Electron.*, vol. 61, no. 11, pp. 6418-6428, Nov. 2014.

[7] J. H. Wang, L. Y. Qiao, Y. Q. Ye, and Y. Q. Chen, "Fractional envelope analysis for rolling element bearing weak fault feature extraction," *IEEE/CAA J. Autom. Sinica*, vol. 4, no. 2, pp. 353-360, Apr. 2017.

[8] D. S. Singh and Q. Zhao, "Pseudo-fault signal assisted EMD for fault detection and isolation in rotating machines," *Mech. Syst. Signal Process.*, vol. 81, pp: 202-218, 2016.

[9] L. Cao, Y. Shen, T. M. Shan, Y. B. Xia, J. L. Wang, and Z. L. Lin, "Bearing fault diagnosis method based on GMM and coupled hidden Markov model," in *Proc. 2018 Prognostics Syst. Health Manage. Conf. (PHM-Chongqing)*, Chongqing, China, Oct. 2018, pp. 932-936.

[10] Z. W. Wang, Q. H. Zhang, J. B. Xiong, M. Xiao, G. X. Sun, and J. He, "Fault diagnosis of a rolling bearing using wavelet packet denoising and random forests," *IEEE Sensors J.*, vol. 17, no. 17, pp. 5581-5588, Sept. 2017.

[11] C. Sun, M. Ma, Z. B. Zhao, and X. F. Chen, "Sparse deep stacking network for fault diagnosis of motor," *IEEE Trans. Ind. Informat.*, vol. 14, no. 7, pp. 3261-3270, Jul. 2018.

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

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

[14] Z. Y. Chen and W. H. Li, "Multisensor feature fusion for bearing fault diagnosis using sparse autoencoder and deep belief network," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 7, pp. 1693-1702, Jul. 2017.

[15] S. Shao, R. Yan, Y. Lu, P. Wang, and R. X. Gao, "DCNN-based multi-signal induction motor fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 6, pp. 2658-2669, Jun. 2020.

[16] S. Ma, W. Liu, W. Cai, Z. Shang, and G. Liu, "Lightweight deep residual CNN for fault diagnosis of rotating machinery based on depthwise separable convolutions," *IEEE Access*, vol. 7, pp. 57023-57036, 2019.

[17] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *Proc. Advances Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2007, pp.137-144.

[18] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Mach. Learn.*, vol. 79, no. 1-2, pp. 151-175, May 2010.

[19] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain adaptation: learning bounds and algorithms," in *Proc. 22nd Annual Conf. Learn. Theory*, Montreal, QC, Canada, Jun. 2009, pp. 34-47.

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

[21] M. S. Long, J. M. Wang, G. G. Ding, J. G. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *2013 IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 2200-2207.

[22] H. L. Zheng, R. X. Wang, Y. T. Yang, J. C. Yin, Y. B. Li, Y. Q. Li, and M. Q. Xu, "Cross-domain fault diagnosis using knowledge transfer strategy: a review," *IEEE Access*, vol. 7, pp. 129260-129290, Sept. 2019.

[23] W. N. Lu, B. Liang, Y. Cheng, D. S. 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.

[24] Z.-H. Liu, B.-L. Lu, H.-L. Wei, X.-H. Li, and L. Chen, "Fault diagnosis for electromechanical drivetrains using a joint distribution optimal deep domain adaptation approach," *IEEE Sensors J.*, vol. 19, no. 24, pp. 12261-12270, Dec. 2019.

[25] X. X. Wang, H. B. He, and L. S. 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, Sept. 2019.

[26] M. Sohaib and J.-M. Kim, "Fault diagnosis of rotary machine bearings under inconsistent working conditions," *IEEE Trans. Instrum. Meas.,* vol. 69, no. 6, pp. 3334-3347, Jun. 2020.

[27] F. Santambrogio, *Optimal Transport for Applied Mathematicians*. New York, NY, USA: Birk€ auser, 2015.

[28] C. Villani, *Optimal Transport: Old and New*. Berlin, Germany: Springer, 2009.

[29] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. the 34 th International Conference on Machine Learning,* Sydney, Australia, 2017, vol. 70, pp: 214-223.

[30] J. Altschuler, J. Weed, and P. Rigollet, "Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration," in *Proc. 31st Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 1964-1974.

[31] Y. J. Xie, X. F. Wang, R. J. Wang, and H. Y. Zha, "A fast proximal point method for computing exact Wasserstein distance," in *Proc. 35th Conf. Uncertain. Artif. Intell.*, Tel Aviv, Isreal, Jul. 2019, pp. 158.

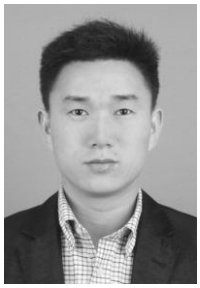
[32] Z. Y. Su, Y. L. Wang, R. Shi, W. Zeng, J. Sun, F. Luo, and X. F. Gu, "Optimal mass transport for shape matching and comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 11, pp. 2246-2259, Nov. 2015.

[33] S. Ferradans, N. Papadakis, J. Rabin, G. Peyré, and J.-F. Aujol, "Regularized discrete optimal transport," in *Proc. 4th Int. Conf. Scale Space Variational Meth. Comput. Vis.*, Leibnitz, Austria, Jun. 2013, pp. 428-439.

[34] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy, "Optimal transport for domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 9, pp. 1853-1865, Sept. 2017.

[35] N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy, "Joint distribution optimal transportation for domain adaptation," in *Proc. 31st*

Conf. Neural Inf. Process. Syst., Long Beach, CA, USA, Dec. 2017, pp. 3730-3739.
[36] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde, "Optimal mass transport: signal processing and machine-learning applications," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 43-59, Jul. 2017.
[37] J. B. Ye, P. R. Wu, J. Z. Wang, and J. Li, "Fast discrete distribution clustering using Wasserstein barycenter with sparse support," *IEEE Trans. Signal Process.*, vol. 65, no. 9, pp. 2317-2332, May 2017.
[38] L. V. Kantorovich, "On the translocation of masses," *J. Math. Sci.*, vol. 133, no 4, pp. 1381-1382, Mar. 2016.
[39] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, Sardinia, Italy, May 2010, pp. 249-256.
[40] K. A. Loparo. (2013). Case western reserve university bearing data center. [Online]. Available: http://csegroups.case.edu/bearingdatacenter/pages/12k-drive-end-bearing-fault-data
[41] B. C. Sun, J. S. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proc. Thirtieth AAAI Conf. Artif. Intell.*, Phoenix, AZ, USA, Feb. 2016, pp. 2058-2065.
[42] P. Vincent, H. Larochelle, I. Lajoie, *et al.*, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, no. 12, pp. 371-3408, 2010.
[43] W.-G. Chang, T. You, S. Seo, S. Kwak, and B. Han, "Domain-specific batch normalization for unsupervised domain adaptation," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7354-7362.
[44] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," IEEE Trans. Neural Netw., vol. 13, no. 2, pp. 415-425, Mar. 2002.
[45] L. Van Der Maaten and G. Hinton, "Visualizing data using t-SNE," J. *Mach. Learn. Res.*, vol. 9, pp. 2579-2605, Nov. 2008.



Zhao-Hua Liu (M'16) received the M.Sc. degree in computer science and engineering, and the Ph.D. degree in automatic control and electrical engineering from the Hunan University, China, in 2010 and 2012, respectively. He worked as a visiting researcher in the Department of Automatic Control and Systems Engineering at the University of Sheffield, United Kingdom, from 2015 to 2016.

He is currently an Associate Professor with the School of Information and Electrical Engineering, Hunan University of Science and Technology, Xiangtan, China. His current research interests include artificial intelligence and machine learning algorithm design, parameter estimation and control of permanent-magnet synchronous machine drives, and condition monitoring and fault diagnosis for electric power equipment.

Dr. Liu has published a monograph in the field of *Biological immune system inspired hybrid intelligent algorithm and its applications*, and published more than 30 research papers in refereed journals and conferences, including IEEE TRANSACTIONS/JOURNAL/MAGAZINE. He is a regular reviewer for several international journals and conferences.

Lin-Bo Jiang received the B.Eng. degree in measurement and control technology and instrument from Hunan University of Science and Technology, Xiangtan, China in 2018, where he is currently pursuing the M.Sc. degree in automatic control and electrical engineering.

His current research interests include deep learning algorithm design, transfer learning, and fault diagnosis for electric power equipment.

Hua-Liang Wei received the Ph.D. degree in automatic control from the University of Sheffield, Sheffield, U.K., in 2004.

He is currently a senior lecturer with the Department of Automatic Control and Systems Engineering, the University of Sheffield, Sheffield, UK. His research focuses on evolutionary algorithms, identification and modelling for complex nonlinear systems, applications and developments of signal processing, system identification and data modelling to control engineering.

Lei Chen received the M.S. degree in computer science and engineering, and the Ph.D. degree in automatic control and electrical engineering from the Hunan University, China, in 2012 and 2017, respectively.

He is currently a Lecturer with the School of Information and Electrical Engineering, Hunan University of Science and Technology, Xiangtan, China. His current research interests include deep learning, network representation learning, information security of industrial control system and big data analysis.

Xiao-Hua Li received the B.Eng. degree in computer science and technology from Hunan University of Science and Engineering, Yong zhou, China, in 2007 and the M.Sc. degree in computer science from Hunan University, Changsha, China, in 2010. Currently, She is currently a lecturer in the School of Information and Electrical Engineering, Hunan University of Science and Technology, Xiangtan, China. Her interests are in evolutionary computation.