A Novel Transfer Learning Method for Fault Diagnosis Using Maximum Classifier Discrepancy With Marginal Probability Distribution Adaptation

SXIN JIA, JINRUI WANG, BAOKUN HAN, GUOWEI ZHANG, XIAOYU WANG, AND JINGTAO HE
College of Mechanical and Electronic Engineering, Shandong University of Science and Technology, Qingdao 266590, China
Corresponding author: Baokun Han (bk_han@163.com)

ABSTRACT Effective fault diagnosis is essential to ensure the safe and reliable operation of equipment. In recent years, several transfer learning-based methods for diagnosing faults under variable working conditions have been developed. However, these models are designed to completely match the feature distributions between different domains, which is difficult to accomplish because each domain has unique characteristics. To solve this problem, we propose a framework based on the maximum classifier discrepancy with marginal probability distribution adaptation that focuses on task-specific decision boundaries. Specifically, this method captures ambiguous target samples through the predicted discrepancy between two classifiers for the target samples. Furthermore, marginal probability distribution adaptation facilitates the capture of target samples located far from the source domain, and these target samples are brought closer to the source domain through adversarial training. Experimental results indicate that the proposed method demonstrates higher performance and generalization ability than existing fault diagnosis methods.

INDEX TERMS Unsupervised learning, mechanical engineering, vibration measurement, signal analysis, artificial intelligence, artificial neural networks, data processing, failure analysis.

I. INTRODUCTION
During the operation and maintenance of rotating mechanical equipment, fault diagnosis plays a significant role in the condition monitoring, health maintenance, and life prediction of such equipment [1]–[4]. Furthermore, the amount of data collected by sensors has exponentially increased with the development of fault detection systems for modern machinery [5], [6]. With its extensive use in target detection, semantic segmentation, autonomous driving [7]–[9], and other fields, deep learning has become the most widely used tool for processing massive amounts of data. The generalized deep learning method employs rich internal information to acquire depth features [10], [11]. Therefore, deep learning-based methods have been widely used to realize fault diagnosis [12]–[14].

In the last few years, many fault diagnosis methods have been presented based on deep learning [15], [16]. For instance, Wang et al. [17] proposed batch-normalized deep neural networks (BNDNN) to solve the internal covariate shift problem in stacked autoencoders (SAE), and an improved training speed and accuracy were realized. Jia et al. [18] built a local connection network by using a normalized sparse autoencoder (NSAE) for intelligent fault diagnosis, in which shift-invariant features could be produced and the mechanical health conditions could be effectively recognized. Jia et al. [19] developed a deep normalized convolutional neural network (DNCNN) for imbalanced fault diagnosis and improved the training process by using normalized layers. In addition, sparse filtering [20], [21] and support vector machines (SVM) [22], [23] have been widely applied.
to realize fault diagnosis. Given that the training and test data are sampled in the same distribution, the above mentioned methods can easily exhibit a favorable performance. However, these methods are not applicable when the test samples are collected from similar yet different operating conditions. In practice, due to the limitations of the working environment, the distributions of the training and test data may differ. Therefore, the working conditions affect the collected vibration signals, which is directly reflected in the deviation and blurring of deep features. Specifically, if the training and testing samples are not collected at the same time, the speed and load of the test data may change, and the working condition information may be modified to a certain extent. The diagnostic performance declines when the classification knowledge is learned from the training sample by using the aforementioned methods.

Therefore, a novel fault diagnosis framework must be developed to apply the previous knowledge of labeled samples to the classification of unlabeled samples under new working conditions. Huang et al. [24] applied a nuisance attribute projection (NAP) to extract the features that are irrelevant to the working conditions and contain fault information. Yang et al. [25] proposed a novel triaxial signal information fusion model, and the C-means method was used to classify the fused signals. Furthermore, in several methods for fault diagnosis, the concept of transfer learning [26], [27] has been adopted to achieve a favorable diagnosis performance. For instance, Wen et al. [28] introduced the maximum average difference error into the sparse autoencoder and demonstrated the achievement of a satisfactory performance. An et al. [29] developed a transfer model by using the maximum mean discrepancy (MMD) based on multiple kernels, in which the features from two different domains were involved in the domain adaptation. Xu et al. [30] proposed a renewable fusion fault diagnosis network that can achieve a high diagnostic accuracy and extract the domain-invariant features at different speeds from unbalanced training samples. Cheng et al. [31] developed a Wasserstein distance-based deep transfer learning (WD-DTL) fault diagnosis framework that can learn the domain-invariant features and minimize the difference in the distributions of the source and target domains.

The above mentioned methods attempt to match the distribution of the deep features without considering the sample category. Regardless of the measurement scale, general domain adaptation approaches reduce the domain distance by aligning the overall features. However, the feature extractor may extract some fuzzy features located far from the source domain, which may make the classification challenging, because a single classifier cannot capture the ambiguous target samples. The above domain adaptive methods do not consider task-specific decision boundaries in the adaptive process. Although these approaches completely align the characteristics of the two domains, they ignore the fact that each domain has unique characteristics. Inspired by the adversarial training technique of generative adversarial networks (GAN) [32], Saito et al. [33] proposed an unsupervised domain adaptive method based on the maximum classifier discrepancy (MCD). This method applies adversarial domain adaptation to the transfer learning framework and demonstrates a favorable performance in digit classification, object classification, and semantic segmentation. Based on this method, Lee et al. [34] proposed the use of the sliced Wasserstein discrepancy (SWD) to capture the discrepancy between the outputs of task-specific classifiers. Lin et al. [35] proposed a maximum classifier discrepancy model that considers the joint distribution of the source and target domain data to realize heterogeneous domain adaption. In this method, MMD is applied to adapt the data distributions between the source and target domains before adversarial training is performed. However, these methods are rarely used in the field of fault diagnosis.

Although MCD-based methods have been considerably improved, none of them consider the relationship between the classifier discrepancy and marginal probability distributions. When two domains exhibit large differences in their marginal probability distributions, the output discrepancy of the classifiers can be easily increased but cannot be easily reduced by updating the parameters of the generator. To improve the fault diagnosis performance under different working conditions, we develop a novel transfer learning model that uses a task-specific classifier for fault diagnosis. In the proposed method, the marginal probability distribution adaptation is introduced to the process of adversarial domain adaptation, and the classifier difference can be narrowed more easily to ensure the classification performance in the target domain. The main contributions of the proposed method can be summarized as follows:

1) The discrepancy between the outputs of two classifiers is measured to locate target samples that are located far from the source domain. This discrepancy is eliminated through adversarial training.
2) The difference in the marginal probability distributions of the training and testing data is reduced to accelerate the adversarial training and improve the stability of the model training.

II. THEORETICAL BACKGROUND
A. GENERAL DOMAIN ADAPTATION
For most traditional deep learning approaches, the model is trained with labeled samples in one domain and tested on the unlabeled samples in the same domain. In this work, the domain represents the working condition. However, optimizing a framework solely concerning the source information results in poor generalizability when the training and testing data are collected from different domains. In transfer learning-based models, the testing and training data are collected from different yet related domains, namely, the target ($D^T$) and source domains ($D^S$), respectively. Given the input data $x_s$ and the corresponding labels $y_s$ drawn from the source set $\{X_s, Y_s\}$ and the input data $x_t$ drawn from the target set $X_t$, the purpose of domain adaptation is to facilitate a knowledge transfer from the labeled source data to the
unlabeled target data. When the distributions of $X_s$ and $X_t$ are sufficiently similar, it is possible to simply concentrate on minimizing the empirical risk of probability distribution $P(X_s, Y_s)$. Therefore, domain adaptation-based methods can diagnose faults on the unlabeled testing data under a certain working condition by using the knowledge learned from the labeled data under another working condition. The problem to be solved in domain adaptation is that the source domain $D^s$ is not equal to the target domain $D^t$; that is, these domains have different data spaces and marginal distributions with $X_s \neq X_t$ and $P(X_i) \neq P(X_t)$. Given that fault samples with the same health status exist in the two domains, the label space of the source domain $Y_s$ is the same as that of the target domain $Y_t$. Furthermore, the labels in the target domain $D^t$ are not available during the model training.

**B. DOMAIN ADAPTATION BASED ON TASK-SPECIFIC DECISION BOUNDARIES**

Most of the existing transfer learning methods may fail to generate discriminative features because they do not consider the relationship between the target sample features and the task-specific decision boundary. For general domain adaptation methods, the generator can produce ambiguous features near the boundary by simply trying to establish the similarities between the two distributions, as shown in the upper part of Fig. 1. Although the classifier performs well in the source domain, the target samples near the class boundaries tend to be misclassified. Therefore, although the classifier exhibits a satisfactory performance in the source domain, the applicability of the classifier in the target domain remains questionable.

![Comparison of the domain adaptation method based on the task-specific decision boundary (the lower part) and the general domain adaptation method (the upper part).](image)

To overcome this problem, we develop a novel adversarial domain adaptation method to align the distributions of the features from the two domains by using the discrepancy of the classifiers. Through this method, we examine the task-specific decision boundaries while aligning the overall distribution. Specifically, we introduce a feature generator and two independent task-specific classifiers (C1 and C2). As shown in the lower part of Fig. 1, both classifiers attempt to accurately classify the source samples and may misclassify some samples in the target domain. Furthermore, the target samples outside the source space tend to be classified differently by the two distinct classifiers because these classifiers are independent. Therefore, the discrepancy region between the two classifiers should be identified and minimized by training the generator to avoid the extraction of target features outside the source space. This discrepancy can be expressed as $d (p_1 (y | x_t) , p_2 (y | x_t))$, where $d (\cdot)$ is a function that measures the discrepancy between the two probabilistic outputs of the classifiers and $p (y | x)$ denotes the probabilistic outputs for the input $x$ obtained by the classifier.

**C. WASSERSTEIN DISTANCE**

The Wasserstein distance has recently been applied to loss functions giving its superiority over other probability measures. Compared with other probability measures such as the total variation distance, Kullback–Leibler (KL) divergence, and Jensen–Shannon (JS) divergence, the Wasserstein distance has much better properties because it takes into account the underlying geometry of the probability space [36]. As shown in Eq. (1), this distance can be mathematically written as

$$W(p_1, p_2) = \inf_{\gamma \in \Pi(p_1, p_2)} \mathbb{E}_{(x, y) \sim \gamma} [\|x - y\|],$$

where $\Pi(p_1, p_2)$ is the set of all the distributions $\gamma(x, y)$ whose marginals are $p_1$ and $p_2$, respectively. $\gamma(x, y)$ denotes how far $x$ should be from $y$ to transform the distribution $p_1$ to distribution $p_2$. The Wasserstein distance can be understood as the cost of an optimal transportation plan.

However, the infimum in (1) is highly intractable. According to the dual principle of Kantorovich–Rubinstein [37], the following form of the Wasserstein distance can be obtained:

$$W(p_1, p_2) = \sup_{f : \|f\|_L \leq 1} E_{x \sim p_1} [f(x)] - E_{x \sim p_2} [f(x)],$$

where the supremum is over all 1-Lipschitz functions $f : x \to \mathbb{R}$.

**III. PROPOSED METHOD**

This section describes the novel framework for transfer fault diagnosis. In addition, the structure of the proposed model is discussed in detail.

**A. PROPOSED FRAMEWORK**

The proposed framework comprises a generator $G$ and two independent classifiers $C_1$ and $C_2$. The generator is a network with three fully connected layers, and each classifier is a network with two fully connected layers. The framework is illustrated in Fig. 2. The spectra of the vibration signals are input into the feature generator, and high dimensional features are output. The two classifiers produce the logits.
\[ p_1(y|x_t^1), p_2(y|x_t^2) \]. To improve the fault diagnosis performance, the following optimization objects are used across the framework:

1) CLASSIFICATION LOSS

The first optimization objective of the framework is to minimize the classification error in the source domain. By doing so, the classifiers can obtain the classification knowledge from the labeled dataset \( X_s = \{x_i\}_{i=1}^{n_s} \) in the source domain. Therefore, \( x_i^t \) is translated to \( o_i^t \) through generator \( G \). The objective function can be formulated as the following standard softmax regression loss:

\[
L_c(o^c, y^i) = -\frac{1}{n^s} \sum_{i=1}^{n^s} \sum_{c=1}^{C} I[y_i^c = c] \log \frac{\exp(o_i^c)}{\sum_{j=1}^{C} \exp(o_i^j)},
\]

where \( I[\cdot] \) is an indicator function returning a value representing the probability that the situation is true, and \( n^s \) is the number of source samples. \( L_{c1} \) and \( L_{c2} \) are minimized for both classifiers to learn the classifiable information from the source dataset.

2) DOMAIN DISTANCE LOSS

As in the general domain adaptation method, after the generator \( G \) extracts the high-dimensional features, the overall difference between the two domains is minimized. The Wasserstein distance is used to determine the difference in the marginal probability distributions. This distance can be approximated by maximizing the domain generator loss \( L_{wd} \), as follows:

\[
L_{wd}(o_s, o_t) = \left| \frac{1}{n_s} \sum_{i=1}^{n_s} G(x_{si}) - \frac{1}{n_t} \sum_{i=1}^{n_t} G(x_{ti}) \right|.
\]

where \( o_s \) and \( o_t \) are the features drawn from the source \( X_s \) and target domains \( X_t \) through generator \( G \), respectively.

3) DISCREPANCY LOSS

The discrepancy loss used in Ref. [33] is applied in the proposed method. As shown in Eq. (5), the discrepancy loss can be expressed as the absolute difference between the probabilistic outputs of the two classifiers:

\[
d(p_1, p_2) = \frac{1}{C} \sum_{c=1}^{C} |p_{1c} - p_{2c}|,
\]

where \( p_{1c} \) and \( p_{2c} \) denote the probability outputs of \( p_1 \) and \( p_2 \) for class \( c \), respectively.

B. TRAINING STEPS

The goal of generator \( G \) is to minimize the discrepancy between the classifiers and to reduce the Wasserstein distance between the two distributions. Moreover, the purpose of classifiers \( C1 \) and \( C2 \) is to accurately classify the different types of samples and to maximize the discrepancy. As shown in Fig. 3, the training process is divided into the following steps to achieve these goals:

**Step 1:** As shown in Fig. 3, the whole model is trained in the source domain \( \{X_s, Y_s\} \). Consequently, both the generator \( G \) and the classifiers in the source domain can accurately classify all the samples. However, increasing the discrepancy by merely updating the parameters \( \theta_G \) of generator \( G \) may be difficult. Furthermore, the classifiers tend to produce similar outputs when the marginal distributions of the two domains are similar. As the domain distance decreases, the discrepancy can be easily narrowed. Therefore, we introduce the domain distance loss in this step. The objective is formulated as

\[
\min_{\theta_G, C1, C2} L_C(X_s, Y_s) + \lambda L_{wd}(X_s, X_t),
\]

\[
L_C(X_s, Y_s) = L_{c1}(o^s, y^s) + L_{c2}(o^t, y^t)
\]

where \( L_{wd}(X_s, X_t) \) is the domain distance loss.

**Step 2:** The parameters \( \theta_G \) of generator \( G \) are frozen, and the classifiers are updated to maximize the discrepancy. By training these classifiers to increase the discrepancy, the target samples that have been excluded by the source domain can be detected. Furthermore, to maintain the same classification performance on the source domain, the classification loss on the source domain is added in this step.
FIGURE 3. Model training steps. We obtain two classifiers that maximize the discrepancy on the target domain and a generator that minimizes this discrepancy.

The objective can be formulated as follows:

\[
\min_{C_1, C_2} L(X_s, Y_s) - \lambda L_{adv}(X_t),
\]

\[
L_{adv}(X_t) = E_{x_t \sim X_t} [d(p_1(y|x_t), p_2(y|x_t))]
\]

where \( L_{adv}(X_t) \) is the discrepancy of the classifiers for the target samples \( X_t \).

**Step 3:** The parameters \( \theta_{C_1} \) and \( \theta_{C_2} \) of classifiers \( C_1 \) and \( C_2 \) are frozen, and generator \( G \) is updated to minimize the discrepancy. The objective is formulated as follows:

\[
\min_G L_{adv}(X_t)
\]

These three steps are repeated in the framework. As shown in Fig. 3, a generator that can create features within a closer space and two classifiers that can classify the samples of the two domains accurately are eventually obtained.

**IV. EXPERIMENTAL VERIFICATION**

**A. EXPERIMENT 1: FAULT DIAGNOSIS ON A DATASET WITH DIFFERENT LOADS**

The famous motor bearing fault data provided by Case Western Reserve University are used to validate the proposed method. The test bed comprises a driving motor, a torque sensor, a power meter, accelerometers, and an electronic control unit. The bearing (SKF6205) is damaged by electrical discharge machining (EDM) at a single point. Single faults and combined faults can be implemented. The frequency of sampling is set to 12 kHz. All datasets are divided into four health conditions, namely, normal, inner ring fault, outer ring fault, and rolling body fault. Each fault has 3 types of damage dimensions, namely, 0.18, 0.36, and 0.54 mm, for 10 healthy types of bearing datasets, namely, NC, IF1, IF2, IF3, OF1, OF2, OF3, RF1, RF2, and RF3. Each health condition consists of 500 samples, each sample includes 2400 data points, and the fast Fourier transformation (FFT) is performed on the time-domain signals to obtain the frequency-domain samples containing 1200 Fourier coefficients. For each health condition of the two domains, half the samples are used for training, and the other half is used for testing.

Four datasets under different working conditions are considered, namely, L0 (0 horsepower), L1 (1 horsepower), L2 (2 horsepower), and L3 (3 horsepower). All fault diagnosis frameworks are trained on the labeled training samples under a single load to classify the unlabeled test samples under another load. Therefore, a total of 12 tasks are performed.

The structure of the generator \( G \) is \([1200, 600, 300, 150]\), whereas that of the classifiers is \([150, 50, 10]\). Fifteen trials are carried out for each task to avoid random factors. Every training batch contains 100 source samples and 100 target samples. The learning rate is set to 0.001, and the Adam optimization method is adopted.

To validate the effectiveness of the proposed method, it is compared with other approaches involving similar network architectures and experimental settings. These methods include the traditional fault diagnosis method, the batch-normalized stacked autoencoder (BNSAE) [15], which does not apply domain adaptation, the transfer component analysis (TCA) technique [28], which is a traditional transfer learning method, and the Wasserstein distance-based deep transfer learning (WD-DTL) method [31]. As shown in Table 1, the proposed method achieves the highest accuracy and lowest standard deviations among all methods for most tasks, whereas BNSAE demonstrates the worst performance due to the lack of domain adaptation. As a classic migration learning algorithm, TCA obtains a considerably higher accuracy than BNSAE; however, this accuracy is slightly lower than that of the other two methods, thereby suggesting that the deep domain adaptation method can effectively apply the classification knowledge learned in the source domain to the target domain. The proposed method also outperforms WD-DTL in most tasks. For instance, in L3 \(\rightarrow\) L0 and L3 \(\rightarrow\) L1, the proposed method outperforms WD-DTL by 5.33% and 4.42%, respectively. In other words, when the two domains have different data distributions, the proposed
TABLE 1. Accuracy of the considered methods in case 1.

| Source domain | Method    | L0          | L1          | L2          | L3          |
|---------------|-----------|-------------|-------------|-------------|-------------|
| L0            | BNSAE     | 78.23% ± 2.56% | 75.50% ± 2.35% | 69.05% ± 3.23% |
|               | TCA       | 95.67% ± 0.51% | 94.21% ± 1.13% | 93.23% ± 1.60% |
|               | WD-DTL    | 100% ± 0.00% | 98.43% ± 1.32% | 95.42% ± 1.15% |
|               | WD-MCD (ours) | 100% ± 0.00% | 100% ± 0.00% | 99.80% ± 0.05% |
| L1            | BNSAE     | 78.15% ± 1.68% | 74.84% ± 1.52% | 68.23% ± 2.85% |
|               | TCA       | 96.05% ± 0.76% | 95.85% ± 1.05% | 95.02% ± 1.13% |
|               | WD-DTL    | 97.80% ± 0.80% | 99.65% ± 0.33% | 97.51% ± 1.57% |
|               | WD-MCD (ours) | 100% ± 0.00% | 100% ± 0.00% | 99.89% ± 0.10% |
| L2            | BNSAE     | 69.53% ± 1.50% | 71.21% ± 1.21% | 72.05% ± 0.87% |
|               | TCA       | 95.21% ± 0.87% | 96.23% ± 1.14% | 94.13% ± 0.75% |
|               | WD-DTL    | 98.16% ± 0.41% | 97.08% ± 2.20% | 97.56% ± 1.51% |
|               | WD-MCD (ours) | 99.32% ± 0.28% | 100% ± 0.00% | 99.88% ± 0.12% |
| L3            | BNSAE     | 61.52% ± 2.68% | 60.21% ± 2.05% | 68.60% ± 2.88% |
|               | TCA       | 84.15% ± 3.66% | 90.05% ± 1.23% | 92.52% ± 1.05% |
|               | WD-DTL    | 92.82% ± 1.41% | 94.16% ± 1.86% | 98.22% ± 0.65% |
|               | WD-MCD (ours) | 97.95% ± 0.73% | 98.58% ± 0.55% | 99.87% ± 0.10% |

FIGURE 4. Features extracted by each hidden layer of the generator for the source and target domains: (a) first, (b) second, and (c) third hidden layers.

The method significantly outperforms the other methods. The general domain adaptation methods may learn the incorrect classification knowledge if the data distribution difference between the two domains is relatively large. Therefore, task L3 → L0 is used to investigate how the proposed method achieves transfer learning in this task-specific scenario.
To explore the feature extraction process of the generator, OF1 samples are used to compare the source samples with the target samples in three hidden layers of a 3D diagram. As shown in Fig. 4, the source-domain features are similar to the target-domain features in each hidden layer. Furthermore, the dimensions of the sample decrease, and the features become sparser with an increase in the number of layers, thereby improving the accuracy of the classifier.

To verify the effectiveness of the marginal probability distribution adaptation, the proposed model is trained without $L_{wd}$. The performances of the considered models for $L3\rightarrow L1$ are compared in Figs. 5 and 6. Given that the classification accuracies of the two classifiers in the source domain are almost identical [33], the accuracy curve of only one classifier is presented. Fig. 5 shows that when the model with $L_{wd}$ is trained, its testing accuracy is higher than that of the model without $L_{wd}$. Furthermore, the testing accuracy of the trained model without $L_{wd}$ is unstable, although the accuracy increases gradually. Fig. 6 shows that the discrepancy loss decreases and converges faster when the marginal probability distribution adaptation is incorporated. Therefore, the use of the marginal probability distribution adaptation can rapidly reduce the discrepancy of the classifiers in an adversarial process, thereby rapidly increasing the classification accuracy in the target domain.

To visually explore the changes in the features generated in the different layers, t-SNE [38] is applied to embed the high-dimensional vectors onto a 2D image. The results for $L2\rightarrow L1$ are plotted in Fig. 7. As shown in Fig. 7(a), the overall distributions of the target and source domains differ, and some types exhibit different degrees of confusion on the edges, which shows the necessity of using the domain adaptation. The first hidden layer of generator $G$ shown in Fig. 7(b) indicates that the distance between the classes corresponding to the source and target domains gradually narrows. Furthermore, Fig. 7(c) shows that the features with the same health conditions are classified close to one another, and the corresponding categories of the two domains overlap, except for fault classes IF1, OF3, and RF2. Certain discrepancies may remain between the predictions of the two classifiers because some target samples located far from the source domain are not detected. Only when the two domains are aligned at the class level can the classification knowledge from the source domain be fully applied to the target domain, as shown in Fig. 7(d). Therefore, the proposed method can be used to render the target samples discriminative.
B. EXPERIMENT 2: FAULT DIAGNOSIS ON A DATASET WITH DIFFERENT SPEEDS

We use a bearing fault dataset collected at different speeds to verify the transferability of the proposed model. Fig. 8 shows the employed experimental equipment. The vibration signal of the bearing base is collected by an LMS data acquisition instrument with a vibration sensor, and the sampling frequency is 25.6 kHz. The engine speed is 2000 r/min, and the cylindrical roller bearing type is N205EU. All the datasets are divided into five health conditions, namely, normal, inner ring fault, outer ring fault, rolling element fault, and combined fault of the rolling element and outer ring. Each type of fault has two types of damage dimensions, specifically, 0.2 mm and 0.4 mm, resulting in 9 healthy types, namely, N, I1, I2, O1, O2, R1, R2, RO1, and RO2. Each condition includes 500 samples, and the dataset has 4500 samples. Three datasets are obtained under variable speeds, specifically, S1 (1800 r/min), S2 (1500 r/min), and S3 (1300 r/min). Therefore, six transfer learning tests are carried out in the experiment.

The structure of the proposed model is the same as that in case 1. A total of 15 trials are carried out for each task. As shown in Table 2, the proposed method outperforms the other approaches in all tasks. Three transfer models significantly outperform BNSAE, in which domain adaptation is not introduced. The testing accuracies of TCA are 90.15% ± 0.85% and 93.85% ± 0.15% for tasks S3 → S1 and S3 → S2, respectively. For WD-DTL, the accuracies are 97.55% ± 0.45% and 98.14% ± 0.35% for the transfer fault diagnosis experiments S3 → S1 and S3 → S2, which means that the proposed method is feasible if the training and test data are collected at different speeds.

The accuracies of the three transfer methods for S3 → S1 are lower than those for S1 → S3. Specifically, the accuracies decrease by 3.38%, 3.45%, and 1.10%. The same phenomenon can be observed in other tasks when the source and target domains swap identities, thereby suggesting that
TABLE 2. Accuracy of the considered methods in case 2.

| Source domain | Method     | S1           | S2           | S3           |
|---------------|------------|--------------|--------------|--------------|
| S1            | BNSAE      | 65.25% ± 3.05% | 59.58% ± 4.82% |              |
|               | TCA        | 93.75% ± 1.65% | 89.64% ± 2.05% |              |
|               | WD-DTL     | 96.95% ± 0.53% | 93.60% ± 0.80% |              |
|               | WD-MCD (ours) | 99.74% ± 0.16% | 99.05% ± 0.35% |              |
| S2            | BNSAE      | 63.14% ± 2.75% | 60.75% ± 3.62% |              |
|               | TCA        | 91.04% ± 1.58% | 90.05% ± 1.65% |              |
|               | WD-DTL     | 95.53% ± 0.66% | 96.58% ± 0.50% |              |
|               | WD-MCD (ours) | 98.95% ± 0.20% | 98.88% ± 0.24% |              |
| S3            | BNSAE      | 53.50% ± 4.60% | 60.28% ± 3.26% |              |
|               | TCA        | 86.26% ± 1.34% | 90.23% ± 1.14% |              |
|               | WD-DTL     | 90.15% ± 0.85% | 94.85% ± 0.15% |              |
|               | WD-MCD (ours) | 97.95% ± 0.75% | 98.54% ± 0.35% |              |

FIGURE 9. Confusion matrix of the prediction results obtained in S3→S1.

FIGURE 10. t-SNE visualization of features in the output layer for S3 → S1.

the data collected at low speeds contain less useful information than the data collected at high speeds. In this case, the accuracy is influenced if the test data are collected at high speeds. The accuracy of the proposed method is the least affected compared with that of the other methods, which highlights that the proposed method has a strong generalization ability when performing fault diagnosis under variable speeds.
The confusion matrix of the proposed method for $S_3 \rightarrow S_1$ is shown in Fig. 9. The method achieves a high diagnostic accuracy in the separation of samples containing a single fault. However, the two types of compound faults can be easily confused with each other. To visualize the classification results, t-SNE is used to draw the features extracted from the output layer of the generator, as shown in Fig. 10. A degree of confusion can be observed between the two compound fault samples, which agrees with the descriptions shown in the confusion matrix.

V. CONCLUSION

This paper proposes a novel transfer learning model to realize fault diagnosis based on a task-specific decision boundary for dealing with varying operating conditions. The discrepancy in the predicted output of the two classifiers is used to capture the target samples located far from the source domain support, and adversarial training is applied to minimize the discrepancy. By reducing the difference between the marginal probability distributions of two domains, the adversarial training is accelerated, and the stability of the model training is improved. The proposed framework is verified by using datasets collected under different loads and speeds. The diagnostic result accuracy and train loss curve indicate that the marginal probability distribution adaptation makes it easier for the classifier to capture the target samples located far from the source domain. Furthermore, the changes in the activation vectors across the hidden layers of the generator are explored by using the t-SNE. As these layers deepen, the features of the same classes in the two domains become similar. The results show that the proposed method outperforms the related methods and exhibits a high generalization ability in dealing with this problem.

In practical applications, the source and target domains may share only partial categories. The categories unique to the source domain may cause the framework to learn useless classification knowledge and may even lead to a negative transfer. Therefore, the proposed model cannot be suitably applied. This problem is considered as a direction for future research and is expected to be solved by weighting all the categories in the source domain to assign small weights to the unique categories and force the framework to focus on useful categories.

REFERENCES

[1] H. Zhao, J. Zheng, J. Xu, and W. Deng, “Fault diagnosis method based on principal component analysis and broad learning system,” IEEE Access, vol. 7, pp. 90263–90272, 2019.

[2] Z. Wang, J. Zhou, J. Wang, W. Du, J. Wang, X. Han, and G. He, “A novel fault diagnosis method of gearbox based on maximum kurtosis spectral entropy deconvolution.” IEEE Access, vol. 7, pp. 29520–29532, 2019.

[3] C. Li, J. Zheng, H. Pan, J. Tong, and Y. Zhang, “Refined composite multivariate multiscale dispersion entropy and its application to fault diagnosis of rolling bearing.” IEEE Access, vol. 7, pp. 47663–47673, 2019.

[4] S. Z. Feng and X. Han, “A novel multi-grid based reanalysis approach for efficient prediction of fatigue crack propagation.” Comput. Methods Appl. Mech. Eng., vol. 353, pp. 107–122, Aug. 2019.

[5] Z. Li, F. Liu, S. Sun, T. Sarkodie-Gyan, and W. Li, “Decoupling of multiple concurrent faults for diagnosing coal cutter gearboxes: An extensive experimental investigation with multichannel sensor measurements.” J. Nondestruct. Eval., Diag. Prognostics Eng. Syst., vol. 2, no. 4, p. 1, Nov. 2019.

[6] J. K. Sinha and E. Elbhbah, “A fault resistance of vibration based condition monitoring of rotating machines,” Mech. Syst. Signal Process., vol. 34, nos. 1–2, pp. 231–240, Jan. 2013.

[7] K. Chen, X. Song, X. Zhai, B. Zhang, B. Hou, and Y. Wang, “An integrated deep learning framework for occluded pedestrian tracking.” IEEE Access, vol. 7, pp. 26060–26072, 2019.

[8] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, “ENet: A deep neural network architecture for real-time semantic segmentation,” 2016, condition X: 1606.02147. [Online]. Available: http://arxiv.org/abs/1606.02147.

[9] A. Salah, M. Abdou, E. Perot, and S. Yogamani, “Deep reinforcement learning framework for autonomous driving,” Electron. Imag., vol. 2017, no. 19, pp. 70–76, Jan. 2017.

[10] Y. Lei, “Opportunities and challenges of machinery intelligent fault diagnosis in big data era.” J. Mech. Eng., vol. 54, no. 5, pp. 94–104, 2018.

[11] Y. Xu, Y. Sun, J. Wan, X. Liu, and Z. Song, “Industrial big data for fault diagnosis: Taxonomy, review, and applications.” IEEE Access, vol. 5, pp. 17368–17380, 2017.

[12] J. Dong, M. Wang, X. Zhang, L. Ma, and K. Peng, “Joint data-driven fault diagnosis integrating causality graph with statistical process monitoring for complex industrial processes.” IEEE Access, vol. 5, pp. 25217–25225, 2017.

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

[14] D. Cabrera, F. Sancho, J. Long, R.-V. Sanchez, S. Zhang, M. Cerrada, and C. Li, “Generative adversarial networks selection approach for extremely imbalanced fault diagnosis of reciprocating machinery.” IEEE Access, vol. 7, pp. 70643–70653, 2019.

[15] Z. Chen and W. 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.

[16] H. Zhao, J. Zheng, J. Xu, and W. Deng, “Fault diagnosis method based on principal component analysis and broad learning system.” IEEE Access, vol. 7, pp. 26060–26072, 2019.

[17] Z. Zhang, Q. Yang, and W. Zhou, “A feature extraction method based on sparse filtering with local structure preserved and its applications to bearing fault diagnosis.” IEEE Access, vol. 7, pp. 160559–160572, 2019.

[18] W. Qian, S. Li, J. Wang, and Q. Wu, “A novel supervised sparse feature extraction method and its application on rotating machine fault diagnosis,” Neurocomputing, vol. 329, pp. 53–65, Feb. 2019.

[19] F. Jia, Y. Lei, L. Guo, J. Lin, and S. Xing, “A neural network constructed by deep learning technique and its application to intelligent fault diagnosis of machines.” Neurocomputing, vol. 272, pp. 619–628, Jan. 2018.

[20] F. Jia, Y. Lei, L. Guo, 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, Sep. 2018.

[21] Z. Zhang, Q. Yang, and W. Zhou, “A feature extraction method based on sparse filtering with local structure preserved and its applications to bearing fault diagnosis.” IEEE Access, vol. 7, pp. 160559–160572, 2019.

[22] W. Qian, S. Li, J. Wang, and Q. Wu, “A novel supervised sparse feature extraction method and its application on rotating machine fault diagnosis,” Neurocomputing, vol. 320, pp. 129–140, Dec. 2018.

[23] Y. Zhang, X. Li, H. Zheng, H. Yao, J. Liu, C. Zhang, H. Peng, and J. Jiao, “A fault diagnosis model of power transformers based on dissolved gas analysis features selection and improved krigl herd algorithm optimized support vector machine,” IEEE Access, vol. 7, pp. 102803–102811, 2019.

[24] J. Guo, J. Wu, Z. Sun, J. Long, and S. Zhang, “Fault diagnosis of delta 3D printers using transfer support vector machine with attitude signals.” IEEE Access, vol. 7, pp. 40359–40368, 2019.

[25] W. Huang, J. Cheng, and Y. Yang, “Rolling bearing fault diagnosis and performance degradation assessment under variable operation conditions based on nuisance attribute projection,” Mech. Syst. Signal Process., vol. 114, pp. 165–188, Jan. 2019.

[26] J. Yang, Y. Bai, J. Wang, and Y. Zhao, “Tri-axial vibration information fusion model and its application to gear fault diagnosis in variable working conditions,” Meas. Sci. Technol., vol. 30, no. 9, Sep. 2019, Art. no. 095009.

[27] R. Zhang, H. Tao, L. Wu, and Y. Guan, “Transfer learning with neural networks for bearing fault diagnosis in changing working conditions,” IEEE Access, vol. 5, pp. 14347–14357, 2017.
[27] Y. Xu, Y. Sun, X. Liu, and Y. Zheng, “A digital-twin-assisted fault diagnosis using deep transfer learning,” IEEE Access, vol. 7, pp. 19990–19999, 2019.

[28] 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.

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

[30] K. Xu, S. Li, X. Jiian, Z. An, J. Wang, and T. Yu, “A renewable fusion fault diagnosis network for the variable speed conditions under unbalanced samples,” Neurocomputing, vol. 379, pp. 12–29, Feb. 2020.

[31] C. Cheng, B. Zhou, G. Ma, D. Wu, and Y. Yuan, “Wasserstein distance based deep adversarial transfer learning for intelligent fault diagnosis,” 2019, arXiv:1903.06753. [Online]. Available: http://arxiv.org/abs/1903.06753

[32] I. Goodfellow, J. Pouget-Abadie, and M. Mirza, “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 2672–2680.

[33] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, “Maximum classifier discrepancy for unsupervised domain adaptation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 3723–3732.

[34] C.-Y. Lee, T. Batra, M. H. Baig, and D. Ulbricht, “Sliced wasserstein discrepancy for unsupervised domain adaptation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 10285–10295.

[35] Z. Lin, Z. Yang, R. Situ, F. Huang, J. Lv, Q. Li, and W. Liu, “Improving maximum classifier discrepancy by considering joint distribution for domain adaptation,” in Proc. Int. Conf. Web Inf. Syst. Eng. Cham, Switzerland: Springer, 2018, pp. 253–268.

[36] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein GAN,” 2017, arXiv:1701.07875. [Online]. Available: http://arxiv.org/abs/1701.07875

[37] C. Villani, Optimal Transport: Old and New. Springer, 2008.

[38] L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” J. Mach. Learn. Res., vol. 9, pp. 2579–2605, Nov. 2008.

SIXIANG JIA received the B.S. degree from the Shandong University of Science and Technology, Qingdao, China, in 2018. He is currently a Graduate Student with the Shandong University of Science and Technology, Qingdao. His current research interests include intelligent fault diagnosis of machines, noise and vibration analysis, and control.

BAOKUN HAN received the Ph.D. degree from the School of Mechanical Engineering, Beijing Institute of Technology, Beijing, China, in 2003. He is currently a Professor with the Shandong University of Science and Technology (SDUST). His current research interests include signal processing and fault diagnosis, vehicle dynamics, noise and vibration analysis, and control.

GUOWEI ZHANG received the B.S. degree from Zaozhuang University, Zaozhuang, China, in 2018. He is currently a Graduate Student with the Shandong University of Science and Technology, Qingdao, China. His current research interests include intelligent fault diagnosis of machines, noise and vibration analysis, and control.

XIAOYU WANG received the B.S. degree from the Liren College, Yanshan University, Qinhuangdao, China, in 2017. He is currently a Graduate Student with the Shandong University of Science and Technology, Qingdao, China. His current research interests include intelligent fault diagnosis of machines, noise and vibration analysis, and control.

JINGTAO HE received the B.S. degree from Zaozhuang University, Zaozhuang, China, in 2018. He is currently a Graduate Student with the Shandong University of Science and Technology, Qingdao, China. His current research interests include intelligent fault diagnosis of machines, noise and vibration analysis, and control.

VOLUME 8, 2020

71485