Unsupervised Cross-Domain Fault Diagnosis Using Feature Representation Alignment Networks for Rotating Machinery

Jiahong Chen, *Member, IEEE*, Jing Wang, Jianxin Zhu, Tong Heng Lee, *Member, IEEE*, and Clarence W. de Silva, *Life Fellow, IEEE*

***Abstract*—In this article, the problem of the cross-domain fault diagnosis of rotating machinery is considered. In a practical setting of this approach, the operating platform of the machine may have a different setup and conditions compared to the experimental platform that is used to collect the training data. This can lead to significant data variations, specifically domain shifts. Conventional data-driven approaches are known to adapt poorly to these domain shifts, resulting in a significant drop in the diagnosis accuracy when the pretrained model is applied in the actual operating situation. In this article, an unsupervised domain adaptation approach is developed to mitigate the domain shifts between the data gathered from the experimental platform (the source domain) and the operating platform (the target domain) by aligning the features extracted from the two data domains. The mutual information between the target feature space and the entire feature space is maximized to improve the knowledge transferability of the labeled data in the source domain. Furthermore, the feature-level discrepancy between the two domains is minimized to further improve diagnosis accuracy. The experiments using public datasets and real-world adaptation scenarios demonstrate the feasibility and the superior performance of the proposed method.**

***Index Terms*—Information theory, machine fault diagnosis, neural network, unsupervised domain adaptation (UDA).**

Manuscript received August 21, 2020; revised November 4, 2020; accepted December 17, 2020. Date of publication December 21, 2020; date of current version October 14, 2021. Recommended by Technical Editor A. Alanis and Senior Editor R. Gao. This work was supported by the MITACS under Grant 11R11370. *(Corresponding author: Jiahong Chen.)*

Jiahong Chen and Clarence W. de Silva are with the Department of Mechanical Engineering, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: jiahong.chen@ieee.org; desilva@mech.ubc.ca).

Jing Wang is with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: jing@ece.ubc.ca).

Jianxin Zhu is with the Hefei General Machinery Research Institute, Hefei 230031, China (e-mail: zhujianxin@hgmri.com).

Tong Heng Lee is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119260, Singapore (e-mail: eleleeth@nus.edu.sg).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMECH.2020.3046277.

Digital Object Identifier 10.1109/TMECH.2020.3046277

I. Introduction

FAULT diagnosis of rotating machinery plays an important role in the safety and the reliability of these machinery and estimating their residual life [1], [2]. Rotating machinery usually works with varying heavy loads and under extreme environments, which accelerates the failure of their components, such as gears and bearings. The increasing fault rate of rotating machinery also boosts their cost of operation and maintenance. The collection and processing of various signals is the common approach to diagnose the faults of rotating machinery, with the goals of recovering from malfunctions and faults and preventing future failures, in a timely manner.

In the approach of fault diagnosis by the neural network for different conditions of the machine, a significant amount of training data are collected and analyzed to extract the sufficient features of the system variables for classification [3]. Different supervised learning approaches have been utilized to automatically detect machinery faults. Linear regression (LR) is utilized to find the linear projections for the feature embeddings that can optimize fault diagnosis [4], [5]. K-nearest neighbors (KNN) is used to improve the diagnosis accuracy by introducing the nonlinearity to those projections [6], [7]. Moreover, researchers utilize support vector machines (SVMs) to further improve the diagnosis accuracy by extracting the high-dimensional features [8], [9]. However, the performance of the system based on the supervised data-driven approaches heavily relies on the handcrafted features that will be harder to extract owing the increase in the complexity of the input data distribution.

To solve this problem, deep learning (DL) techniques have been proposed to extract features automatically from the input data without any human guidance [10]–[12]. DL-based methods manage to significantly improve the performance of fault diagnosis by not using the handcrafted features [13], [14]. Recurrent neural networks (RNNs) have been utilized for detecting the faults of induction motors by taking advantage of the temporal dynamics of RNNs [15]. Convolutional neural networks have been used to generate nonhandcrafted features for machine health monitoring [16]. Autoencoders are also utilized to produce more robust features from the input signals, which can reduce the model complexity and improve the fault diagnosis accuracy [17]. Nevertheless, DL-based approaches make a strong assumption that the discriminative features in the training data

(in the source domain) and the testing data (in the target domain) are the same. In real-world scenarios, the variations between the training data distributions and the testing data distributions can significantly degrade the model generalization. For example, the operating conditions (such as the mounting conditions of the experimental platform, humidity, and temperature) while collecting the training data could be completely different from those of the actual operating platform. These variations (domain shifts) could result in a huge divergence between the features extracted from the data collected from the experimental settings and the data observed in the actual operating situation. Therefore, the current single-domain DL methods cannot guarantee the generalization of the fault diagnosis model, which leads to poor performance on the actual operating platforms.

In response to this shortcoming, unsupervised domain adaptation (UDA) has been proposed to extract the domain-invariant features from different data domains; thereby, mitigating the impact of data variation on the pretrained model [18]. A decision boundary can be learned from the labels for the source input samples and can be applied to the target domain by taking advantage of the domain-invariant features that are shared by the two data domains. Therefore, in real-world applications, a set of vibration signals can be sampled on the target-domain platform without knowing its health condition. Then, these unlabeled target-domain samples are trained with the source data to mitigate domain divergence. Once the model is trained, the target-domain knowledge is captured and can be deployed to diagnose target-domain machine faults. In general, UDA approaches fall into two main categories: the adversarial UDA and the nonadversarial UDA.

The adversarial UDA methods attempt to align the features extracted from the two domains using a generative adversarial network that uses two networks to compete with each other to improve its generalization [19], [20]. Adversarial UDA models have been proposed for the cross-domain bearing fault diagnosis. Liu *et al.* [21] combine the standard adversarial UDA model with the stacked autoencoder to improve the domain-invariant feature extraction for the cross-domain diagnosis. Li *et al.* [22] utilize an adversarial UDA model to transfer the knowledge learned from the data of multiple machines to a target machine, which achieves relatively good performance on the fault diagnosis of the target machine. Furthermore, adversarial UDA methods are also utilized to diagnose machinery faults across sensors at different places [23]. Nevertheless, the traditional adversarial approaches only discriminate the features as originated from the source domain or the target domain, and the extracted features are ambiguous near the class boundaries. The maximum classifier discrepancy (MCD) solves this issue by explicitly considering the class-specific decision boundaries while aligning the features from the two domains [24]. The MCD avoids generating features near the class boundaries and extracts the target features under the support of the source discriminative features.

The adversarial UDA approaches may not fully optimize the features that are important for the classification tasks in the target domain. Moreover, adversarial training is hard to converge in the case of large datasets. The nonadversarial UDA avoids this issue by quantifying domain shifts with some statistical distances. Methods based on the maximum mean discrepancy (MMD) are proposed to minimize the variance between the feature spaces of the two domains to mitigate the domain shifts [25]. The nonadversarial UDA methods have been utilized to solve the cross-domain fault diagnosis as well. Wen *et al.* [26] propose a deep transfer learning (DTL) model based on the sparse autoencoder to extract features from the source domain and minimize the domain shift by the MMD. However, the utilization of the sparse autoencoder has its inherent limitation in extracting the spatial features. Domain adaptation for fault diagnosis (DAFD) is proposed to strengthen the representative information of the input data to improve the knowledge transferability [27]. Zheng *et al.* [28] designed an intelligent fault identification approach based on the transfer locality preserving projection (TLPPIFI). TLPPIFI projects the input data from the source domain into a low-dimensional subspace to preserve the intrinsic geometry and the local structure of the source data distribution, while the MMD between the two data domains is minimized to mitigate the domain shifts. However, both DAFD and TLPPIFI utilized SVMs as their label classifiers that rely on the handcrafted features, and thus, the generalization of their models is generally poor. Furthermore, stepwise adaptive feature norm (SAFN) improves MMD-based approaches by placing the target-domain features far away from the small-norm regions, which is less informative, to improve the model transferability [29]. Lately, Wang *et al.* [30] demonstrated that the knowledge learned from the source domain can be better transferred to the target domain by projecting the features of the two domains to the same prior distribution space.

Currently, most UDA approaches for fault diagnosis do not consider the inherent correlations between the source features and the target features. The input data for machinery fault diagnosis involve processing time-series signals, which contain redundant information in the input level and are less effective for solving the UDA problem [31]. Therefore, fault-diagnosis-specific features should be extracted and analyzed to mitigate the domain shift. In this article, a feature representation alignment network (FRAN) is proposed to minimize the domain shift in the fault-diagnosis-specific feature level and maximize the knowledge transfer for the cross-domain machinery fault diagnosis. First, the features are extracted from the two data domains using a unified feature extractor. Then, two approaches are utilized to align the source feature representation and the target feature representation: 1) the mutual information (MI) between the target feature space and the entire feature space is maximized to reduce the uncertainty in the unlabeled domain and improve the model transferability; and 2) a feature-level MMD is developed to align the fault-diagnosis-specific features from the two domains. Last, when the training of the model converges, it can be directly deployed to diagnosis the unlabeled input signals in the target domain.

The main contributions of this article are the following:

1) A feature representation alignment framework is proposed to improve the knowledge transfer and the model generalization in the cross-domain fault diagnosis for rotating machinery.
2) An MI-based method is developed to better align the fault-diagnosis-specific features from different data domains.

3) A feature-level discrepancy is proposed to mitigate the domain shift in the cross-domain fault diagnosis.
4) The developed framework, FRAN, is shown to outperform state-of-the-art algorithms in several public datasets and the real-world adaptation scenarios.

The rest of this article is organized as follows. Section II presents the preliminaries and formulates the problem. Section III gives details about the methodology and the proposed framework. The experiments to evaluate the proposed framework are conducted and analyzed in Section IV. Finally, Section V concludes this article.

II. Problem Formulation

A. Preliminaries

1) Mutual Information: MI estimates the nonlinear relationship between two random variables, which is viewed as their true dependence, in probability and statistics. Let (X, Y) be a pair of random variables with values over the space $\mathcal{X} \times \mathcal{Y}$. Their MI is defined as

$$\mathrm{I}(X;Y) = \mathrm{H}(X) - \mathrm{H}(X|Y) \tag{1}$$

where $\mathrm{I}(\cdot)$ denotes the MI between the two distribution spaces and $\mathrm{H}(\cdot)$ denotes the Shannon entropy. The Shannon entropy may be explicitly written as

$$\mathrm{H}(X) = -\sum_{x\in X} \mathrm{P}(x)\log \mathrm{P}(x). \tag{2}$$

The conditional entropy $\mathrm{H}(X|Y)$ is given by

$$\mathrm{H}(X|Y) = -\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(y)}. \tag{3}$$

2) Maximum Mean Discrepancy: The MMD is a commonly used nonparametric statistical approach to estimate the discrepancy between two data domains [32]. The basic idea of MMD is that all the statistical parameters of two probability distributions should be the same if they are statistically identical. Therefore, the MMD is a kernel-based test that either accepts or rejects a null hypothesis based on the observation (whether or not the distribution p equals the distribution q). The objective of MMD is to minimize the discrepancy between the source input space and the target input space in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$. Given source-domain samples $\mathbf{x}_s^{(i)}$ and target-domain samples $\mathbf{x}_t^{(i)}$, the empirical estimation of MMD between the source input distribution and the target input distribution can be given by

$$d_{\mathcal{H}}(p,q) = \sup_{f\in\mathcal{H}}(\mathbb{E}_{\mathbf{x}_s}(f(\mathbf{x}_s)) - \mathbb{E}_{\mathbf{x}_t}(f(\mathbf{x}_t)) \tag{4}$$

where $\mathcal{H}$ is a class of functions.

B. Problem Formulation

In the setting of cross-domain fault diagnosis, a source domain $D_S = \{(\mathbf{x}_s^{(i)}, \mathbf{y}_s^{(i)})\}_{i=1}^{N}$ is given, which consists of n labeled signals from the source input space $\{X_S, Y_S\}$. $\mathbf{x}_s^{(i)} \in X_S$ denotes a sample from the source input space that has the probability distribution function $p_s(X_S)$, and $\mathbf{y}_s^{(i)} \in Y_S$ is the corresponding label that identifies the fault type for classification. Similarly, a target domain $D_T = \{(\mathbf{x}_t^{(j)})\}_{j=1}^{M}$ is provided with m unlabeled signals that are sampled from the target input space $\{X_T, Y_T\}$. Target samples have the probability distribution function $p_t(X_T)$. Commonly, the source domain and the target domain are not the same due to the domain shift, and hence, the probability distributions of the source and the target domains are different, $p_s(X_S) \neq p_t(X_T)$. The operating condition of the machine and the types of possible faults are assumed to be known. In this article, the type of faults for classification in the target domain is considered to be the same as that in the source domain. The objective of cross-domain fault diagnosis is to learn a generalized model that performs well in unlabeled target-domain classification, for fault diagnosis, with the help of labeled source domain

$$\min \|\mathbf{y}_t, \hat{\mathbf{y}}_t\| \tag{5}$$

where $\hat{\mathbf{y}}_t$ denotes the predicted fault type of the target-domain inputs.

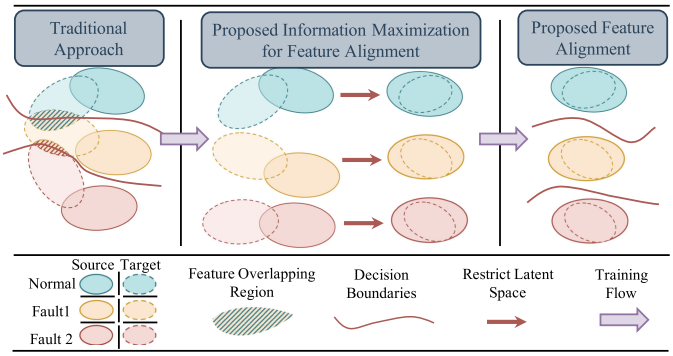


Fig. 1. Overall concept for the proposed feature representation alignment among multiple classes.

III. Methodology

A. Overall Idea for Feature Representation Alignment

The objective of UDA, in the present context, is to produce a generalized feature extractor G that can produce feature representations for both the source domain and the target domain. The classifier F can then accurately predict the unlabeled target samples with the assistance of knowledge in source-domain classification. Fig. 1 presents the overall concept for the proposed feature representation alignment among multiple classes. For simplicity, only the normal operating condition and two fault conditions are used for illustration. Traditional domain adaptation approaches, as shown in the leftmost segment of Fig. 1, utilize a model that can correctly classify the labeled source data, and then, the corresponding decision boundary can be established among source features. However, as there exist domain shifts between the source and the target domains, target features that are not supported by the source knowledge may have overlapping feature regions between different classes, which prohibits the classifier from generating a correct prediction.

This article proposes to regularize feature-level discrepancy during the training, which maximizes the knowledge transferability. This process facilitates the alignment of the feature distributions of the source and the target domains and also enables the target feature to get support from source knowledge. In this manner, as shown in the rightmost segment of Fig. 1, the domain shift is reduced, and hence, more labeled source knowledge can be used to support the target-domain diagnosis. This helps the classifier to produce more accurate target-domain predictions.

B. MI Maximization for Feature Alignment

It is difficult to directly produce a good feature representation due to the lack of labels in the target-domain data. Under these conditions, the classifier cannot predict the labels of the target data with high confidence. It has been found that maximizing the MI between the two domains can reduce the uncertainty in the unlabeled domain. Denoting the feature extractor as G, the source feature representation and target representation can be expressed as $Z_S = G(X_S)$ and $Z_T = G(X_T)$, respectively. Denoting the classifier as F, the fault type predictions for the source and target-domain data can be represented as $\hat{\mathbf{y}}_s^{(i)} = F(G(\mathbf{x}_s^{(i)}))$ and $\hat{\mathbf{y}}_t^{(j)} = F(G(\mathbf{x}_t^{(j)}))$, respectively.

For the purpose of aligning the source and the target feature distributions, this article proposes to maximize the MI between the target feature representation and the entire feature space, specifically

$$\max \mathrm{I}(Z_T; Z) = \max \mathrm{H}(Z_T) - \mathrm{H}(Z_T|Z). \tag{6}$$

Here, Z denotes the entire feature representation, which is the union of the source feature representation set and the target feature representation set

$$Z = Z_S \cup Z_T. \tag{7}$$

The Shannon entropy, $\mathrm{H}(\cdot)$, can be represented in terms of the expected value corresponding to its probability distribution according to (2). It is maximized as

$$\max \mathrm{H}(Z) = -\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[\log q(\mathbf{z})]. \tag{8}$$

The substitution of (8) into (6) provides a particular structure, as seen from

$$\begin{aligned}&\max \mathsf{I}(Z_T; Z) = \max \mathsf{H}(Z_T) - \mathsf{H}(Z_T|Z)\\ &= \max \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[\log q(Z_T|Z)] - \mathbb{E}_{\mathbf{z}_t\sim q(\mathbf{z}_t)}[\log q(Z_T)].\end{aligned} \tag{9}$$

Equation (9) can be simplified. According to (7), $\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[\log q(Z_T|Z)]$ can be expanded from the entire feature space to the union source-domain feature representation and the target-domain representation

$$\begin{aligned}&\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[\log q(Z_T|Z)]\\ &= \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[\log q(Z_T|(Z_S \cup Z_T))]\\ &= \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[\log \frac{q(Z|Z_T)\cdot q(Z_T)}{q(Z)}]\\ &= \mathbb{E}_{\mathbf{z}_t\sim q(\mathbf{z}_t)}[\log q(Z_T)] - \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[\log q(Z)].\end{aligned} \tag{10}$$

By combining (9) and (10), the MI between the entire feature space and the target feature representation can be simplified as

$$\begin{aligned}&\max \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[\log q(Z_T|Z)] - \mathbb{E}_{\mathbf{z}_t\sim q(\mathbf{z}_t)}[\log q(Z_T)]\\ &= \max \mathbb{E}_{\mathbf{z}_t\sim q(\mathbf{z}_t)}[\log q(Z_T)] - \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[\log q(Z)]\\ &\quad - \mathbb{E}_{\mathbf{z}_t\sim q(\mathbf{z}_t)}[\log q(Z_T)]\\ &= \max -\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[\log q(Z)]\\ &= \max \mathsf{H}(Z).\end{aligned} \tag{11}$$

It is seen that maximizing the MI between the entire feature space and the target feature representations is equivalent to the entropy maximization of the entire feature space.

Equation (11) is expanded for easier computation, as

$$\begin{aligned}\max \mathsf{H}(Z) &= \max -\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[\log q(Z)]\\ &= \max -\mathbb{E}_{\mathbf{z}_t\sim q(\mathbf{z}_t),\mathbf{z}_s\sim q(\mathbf{z}_s)}[\log q(Z_T \cup Z_S)].\end{aligned} \tag{12}$$

The left-hand side of (12) is expanded as

$$\begin{aligned}&-\mathbb{E}_{\mathbf{z}_t\sim q(\mathbf{z}_t),\mathbf{z}_s\sim q(\mathbf{z}_s)}[\log q(Z_T \cup Z_S)]\\ &= -\mathbb{E}_{\mathbf{z}_t\sim q(\mathbf{z}_t)}[\log q(Z_T)] - \mathbb{E}_{\mathbf{z}_s\sim q(\mathbf{z}_s)}[\log q(Z_S)]\\ &\quad + \mathbb{E}_{\mathbf{z}_t\sim q(\mathbf{z}_t),\mathbf{z}_s\sim q(\mathbf{z}_s)}[\log q(Z_T, Z_S)]\\ &= -(\mathbb{E}_{\mathbf{z}_t\sim q(\mathbf{z}_t)}[\log q(Z_T)] - \mathbb{E}_{\mathbf{z}_s\sim q(\mathbf{z}_s)}[\log q(Z_S)]).\end{aligned} \tag{13}$$

Then, substitution of (13) into (12) gives

$$\begin{aligned}&\max \mathsf{H}(Z)\\ &= \min \mathbb{E}_{\mathbf{z}_t\sim q(\mathbf{z}_t)}[\log q(Z_T)] + \mathbb{E}_{\mathbf{z}_s\sim q(\mathbf{z}_s)}[\log q(Z_S)]\\ &= \max \mathsf{H}(Z_S) + \mathsf{H}(Z_T).\end{aligned} \tag{14}$$

From this development, it can be concluded that simultaneously maximizing the entropy of Z_S and Z_T will maximize the knowledge transfer from the source to the target. Here, the distributions of the source and the target domains can be aligned, which enables the source knowledge supporting the unlabeled target domain, leading to improved data classification. In this manner, the classifier is able to make more accurate predictions, as illustrated in Fig. 1.

C. Maximum Mean Feature Discrepancy for Feature Alignment

Apart from the maximization of the MI, distribution discrepancy between the source and the target domains should be analyzed to improve the knowledge transfer. The use of the RKHS is a popular statistical learning approach for calculating the distribution discrepancy. Denote a random variable $\mathbf{x} \in X$ in domain D, whose distribution is $p(\mathbf{x})$. A real-valued RKHS $\mathcal{H}$ on domain D with kernel $k(x, x^*)$ represents a real-valued Hilbert space of the real-valued function $f : D \mapsto \mathbb{R}$, where the evaluation functional over the Hilbert space of functions $\mathcal{H}$ is $\mu_{\mathbf{x}} : f \mapsto f(\mathbf{x})$. The kernel of $\mathcal{H}$ satisfies the reproducing property

$$\langle f(\cdot), k(\mathbf{x}, \cdot)\rangle_{\mathcal{H}} = f(\mathbf{x}). \tag{15}$$

In the case of UDA, the kernel function $k(\mathbf{x}, \cdot)$ can be regarded as the feature map $\phi(\mathbf{x})$, where $k(\mathbf{x}, \mathbf{x}^*) = \langle \phi(\mathbf{x}), \phi(\mathbf{x}^*)\rangle_{\mathcal{H}}$.

Now, denote a probability distribution p as an element in RKHS through kernel embedding

$$\mu_{\mathbf{x}}(p) = \mathbb{E}_{\mathbf{x}}(\phi(\mathbf{x})) \tag{16}$$

where $\mathbb{E}_{\mathbf{x}}(k(\mathbf{x}, \mathbf{x}^*))$ is finite. The mean embedding function $\mu_{\mathbf{x}}$ can evaluate the expectation (expected value) of any RKHS function f, using an inner product in Hilbert space $\mathcal{H}$ [33], [34]

$$\langle \mu_{\mathbf{x}}, f \rangle_{\mathcal{H}} = \mathbb{E}_{\mathbf{x}}(f(\mathbf{x})) \quad \forall f \in \mathcal{H}. \tag{17}$$

The kernel mean embedding process $\mu_{\mathbf{x}}$, which is a nonparametric approach, helps to modify distributions by drawing data from its domain. Therefore, a kernel $k(\cdot, \cdot)$ with an embedding $\mu_{\mathbf{x}}(p)$ to RKHS can retrieve discriminative features of any distribution. This process does not require the estimation of the probability density of the intractable true distribution p [35]. Therefore, the embedding of the intractable true posterior $p(\mathbf{x})$ can be approximated by using a finite set of data samples $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}\}$. The resulting empirical kernel embedding can be represented as $\hat{\mu}_{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \phi(\mathbf{x}_i)$, which converges to $\|\mu_{\mathbf{x}} - \hat{\mu}_{\mathbf{x}}\|_{\mathcal{H}}$ in RKHS norm.

However, traditional RKHS kernel embedding focuses only on signal domain. In the case of UDA, data samples are drawn from the source distribution $p(\mathbf{x}_s)$ and target distribution $q(\mathbf{x}_t)$, where $X_S = \{\mathbf{x}_s^{(1)}, \mathbf{x}_s^{(2)}, \ldots, \mathbf{x}_s^{(n)}\}$ and $\{\mathbf{x}_t^{(1)}, \mathbf{x}_t^{(2)}, \ldots, \mathbf{x}_t^{(m)}\}$ are the set of data samples from the source and the target domains. The MMD solves the associated problem by introducing a two-sample test that rejects/accepts the null hypothesis of $p = q$ [32]. MMD makes the assumption that if the probability distributions are identical, the statistical features should be the same

$$d_{\mathcal{H}}(p, q) = \sup_{f \in \mathcal{H}} (\mathbb{E}_{\mathbf{x}_s}(f(\mathbf{x}_s)) - \mathbb{E}_{\mathbf{x}_t}(f(\mathbf{x}_t)) \tag{18}$$

where $d_{\mathcal{H}}(p, q) = 0 \iff p = q$. An unbiased estimator of $d_{\mathcal{H}}(p, q)$ may be represented as

$$\begin{aligned} &\hat{d}_{\mathcal{H}}(p, q) \\ &= \frac{1}{n_s^2}\sum_{i=1}^{n_s}\sum_{j=1}^{n_s} k(\mathbf{x}_s^{(i)}, \mathbf{x}_s^{(i)}) + \frac{1}{n_t^2}\sum_{i=1}^{n_t}\sum_{j=1}^{n_t} k(\mathbf{x}_t^{(i)}, \mathbf{x}_t^{(i)}) \\ &\quad - \frac{2}{n_s n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t} k(\mathbf{x}_s^{(i)}, \mathbf{x}_t^{(i)}). \end{aligned} \tag{19}$$

Simply calculating the domain shifts between $p(\mathbf{x}_s)$ and $q(\mathbf{x}_t)$ may introduce noise from the input space. As there exists domain shift between the source and the target domains, the samples in the source domain and the target domain may not be the same. Therefore, some features in the input space may not help the cross-domain classification. Unlike input space samples, feature representations are produced by a feature extractor, where the discriminative features are maximized [35]. Directly optimizing the divergence between the source and the target feature representations may improve the performance of RKHS kernel embedding in cross-domain classification. Following the structure of a traditional MMD, the Hilbert space embedding of the source feature distribution $p(\mathbf{z}_s)$ and the target feature distribution $q(\mathbf{z}_t)$ can be used to measure the discrepancy of the two feature representation distributions. The resulting optimal divergence measurement is called maximum mean feature discrepancy (MMFD), which is defined as

$$d_{\mathcal{F}}(p, q) = \sup_{f \in \mathcal{H}} (\mathbb{E}_{\mathbf{z}_s}(f(\mathbf{z}_s)) - \mathbb{E}_{\mathbf{z}_t}(f(\mathbf{z}_t)) \tag{20}$$

where samples are drawn from the source feature distribution $p(\mathbf{z}_s)$ and the target feature distribution $q(\mathbf{z}_t)$. $Z_S = \{\mathbf{z}_s^{(1)}, \mathbf{z}_s^{(2)}, \ldots, \mathbf{z}_s^{(n)}\}$ and $Z_T = \{\mathbf{z}_t^{(1)}, \mathbf{z}_t^{(2)}, \ldots, \mathbf{z}_t^{(m)}\}$ are the sets of data samples from the source feature representation and the target feature representation, respectively.

According to the kernel two-sample test theory of MMD and (18), $q(\mathbf{z}_t)$ can approximate the intractable true posterior $p(\mathbf{z}_s)$ if $p(\mathbf{z}_t) = q(\mathbf{z}_s)$, which happens if and only if $d_{\mathcal{F}}(p, q) = 0$. Similar to (19), the unbiased estimator of $d_{\mathcal{F}}(p, q)$ can be represented as the sum of the squared distance between the kernel mean embeddings

$$\begin{aligned} &\hat{d}_{\mathcal{F}}(p, q) \\ &= \frac{1}{n_s^2}\sum_{i=1}^{n_s}\sum_{j=1}^{n_s} k(\mathbf{z}_s^{(i)}, \mathbf{z}_s^{(i)}) + \frac{1}{n_t^2}\sum_{i=1}^{n_t}\sum_{j=1}^{n_t} k(\mathbf{z}_t^{(i)}, \mathbf{z}_t^{(i)}) \\ &\quad - \frac{2}{n_s n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t} k(\mathbf{z}_s^{(i)}, \mathbf{z}_t^{(i)}) \end{aligned} \tag{21}$$

where the default kernel $k(\cdot, \cdot)$ is the Gaussian kernel

$$k(\mathbf{z}, \mathbf{z}') = \exp(-(\mathbf{z} - \mathbf{z}')^2/\gamma) \tag{22}$$

where γ is a scalar bandwidth parameter.

D. System Framework of FRAN

Based on results obtained in Sections III-B and III-C, the system framework of FRAN is presented in this section. In the present work, the objective function consists of three parts: 1) labeled source-domain classification; 2) MI maximization for feature-level knowledge transfer; and 3) MMFD.

Following the standard UDA protocol, the softmax cross-entropy loss is used to evaluate the labeled source-domain classification. During the training, both the feature extractor G and the classifier F are trained to minimize the loss function of the source-domain classification

$$\begin{aligned} &\mathcal{L}_{cls}(X_S, Y_S) \\ &= -\frac{1}{N}\sum_{i=1}^{N} \delta(\mathbf{y_s}^{(i)}, \hat{\mathbf{y}}_{\mathbf{s}}^{(i)}) \log\left[\sigma \circ F \circ G(\mathbf{x}_s^{(i)})\right] \end{aligned} \tag{23}$$

where $\sigma(\cdot)$ is the softmax function. $\sigma(\cdot)$ interprets the model outputs as nonnegative probabilities that add up to 1. G is the feature extractor, which encodes the input diagnosis signals from the source domain to their feature representation, $\mathbf{z_s}^{(i)} = G(\mathbf{x_s}^{(i)})$. F is the classifier, which calculates the likelihood for each fault types from the feature representation generated by G. The output of F is then normalized by the softmax function σ and produces a probability distribution over the predicted fault types $\hat{\mathbf{y}}_{\mathbf{s}}^{(i)} = \sigma \circ F(\mathbf{z_s}^{(i)})$. The binary indicator, $\delta(\mathbf{y_s}^{(i)}, \hat{\mathbf{y}}_{\mathbf{s}}^{(i)})$, outputs

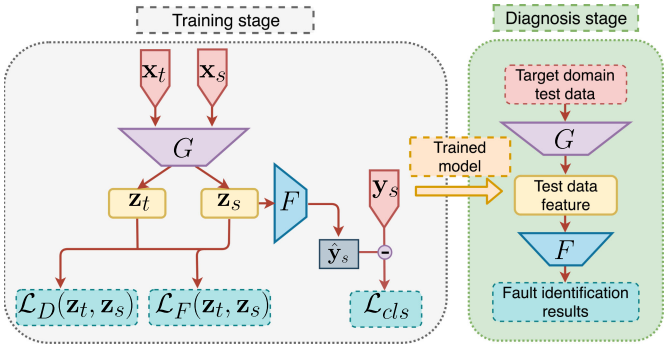


Fig. 2. System framework of FRAN.

1 if the prediction $\hat{\mathbf{y}}_\mathbf{s}^{(i)}$ matches its corresponding class label ${\mathbf{y}_\mathbf{s}}^{(i)}$.

In the MI maximization for feature-level knowledge transfer, the loss function can be obtained according to (14)

$$\begin{aligned}&\mathcal{L}_F(X_S, X_T)\\&= \frac{1}{N}\sum_{i=1}^{N}[G(\mathbf{x}_s^{(i)})\cdot\log(G(\mathbf{x}_s^{(i)})) + G(\mathbf{x}_t^{(i)})\cdot\log(G(\mathbf{x}_t^{(i)}))]\\&= \frac{1}{N}\sum_{i=1}^{N}[\mathbf{z}_s^{(i)}\cdot\log(\mathbf{z}_s^{(i)}) + \mathbf{z}_t^{(i)}\cdot\log(\mathbf{z}_t^{(i)})].\end{aligned} \tag{24}$$

The loss function for MMFD is provided from (21) as

$$\begin{aligned}&\mathcal{L}_D(Z_S, Z_T)\\&= \frac{1}{n_s^2}\sum_{i=1}^{n_s}\sum_{j=1}^{n_s}k(\mathbf{z}_s^{(i)}, \mathbf{z}_s^{(i)}) + \frac{1}{n_t^2}\sum_{i=1}^{n_t}\sum_{j=1}^{n_t}k(\mathbf{z}_t^{(i)}, \mathbf{z}_t^{(i)})\\&\quad - \frac{2}{n_s n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}k(\mathbf{z}_s^{(i)}, \mathbf{z}_t^{(i)}).\end{aligned} \tag{25}$$

Summarizing (23)–(25), the objective function for the entire system framework can be presented as

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha\mathcal{L}_F + \beta\mathcal{L}_D \tag{26}$$

where α and β are the weighting coefficients.

Fig. 2 presents the detailed system framework of FRAN, as developed in this work. The proposed system framework includes the training stage and the diagnosis stage. The model training stage incorporates the UDA protocol, which only takes source-domain samples ($\mathbf{x}_s$), source-domain labels ($\mathbf{y}_s$), and target-domain samples ($\mathbf{x}_t$) as inputs. In the beginning, the source-domain samples and target-domain samples are passed through the feature extractor F to produce feature representations $\mathbf{z}_s$ and $\mathbf{z}_t$. Both source and target feature representations are utilized to calculate the feature alignment losses $\mathcal{L}_F$ and $\mathcal{L}_D$. Meanwhile, only the source feature representation $\mathbf{z}_s$ is passed through the classifier F to generate predictions and calculate the classification loss $\mathcal{L}_{\text{cls}}$. Last, all the losses are summed according to (26) and the optimizer backpropagates the errors to improve the model performance.

Algorithm 1: The Implementation of FRAN.

Input: X_S, X_T, Y_S, and hyper-parameters α and β.

1 Input signal normalization;
2 Initialize network parameters θ ;
3 **while** *epoch* $\leq$ *max epoch* **do**
4 **for** *batch* $\leftarrow 1$ **to** M **do**
5 Draw k minibatch samples $\mathbf{x}_s, \mathbf{x}_t$ from source domain X_S and target domain X_T;
6 Draw k source labels $\mathbf{y}_s$ from Y_S;
7 $\mathbf{z}_s \leftarrow G(\mathbf{x}_s;\theta)$, $\mathbf{z}_t \leftarrow G(\mathbf{z}_t;\theta)$;
8 $\hat{\mathbf{y}}_s \leftarrow F(\mathbf{z}_s)$;
9 $\mathcal{L}_{cls} = -\frac{1}{k}\sum_{i=1}^{k}\delta({\mathbf{y}_\mathbf{s}}^{(i)}, \hat{\mathbf{y}}_\mathbf{s}^{(i)})\log(\hat{\mathbf{y}}_s^{(i)})$;
10 $\mathcal{L}_F = -\frac{1}{k}\sum_{i=1}^{k}[\mathbf{z}_s^{(i)}\cdot\log(\mathbf{z}_s^{(i)}) + \mathbf{z}_t^{(i)}\cdot\log(\mathbf{z}_t^{(i)})]$;
11 $\mathcal{L}_D = \frac{1}{n_s^2}\sum_{i=1}^{n_s}\sum_{j=1}^{n_s}k(\mathbf{z}_s^{(i)}, \mathbf{z}_s^{(i)}) + \frac{1}{n_t^2}\sum_{i=1}^{n_t}\sum_{j=1}^{n_t}k(\mathbf{z}_t^{(i)}, \mathbf{z}_t^{(i)}) - \frac{2}{n_s n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}k(\mathbf{z}_s^{(i)}, \mathbf{z}_t^{(i)})$;
12 Optimize G and F to $\min_{G,F}[\mathcal{L}_{cls} + \alpha\mathcal{L}_F + \beta\mathcal{L}_D]$, and update θ.
13 **end**
14 **end**

Once the model is trained, it can be deployed to test the adaptation performance. The diagnosis stage takes the target-domain data samples as the input and pass them through the trained feature extractor and classifier to produce the predictions on fault identification $\hat{\mathbf{y}}_t$.

The detailed implementation of the proposed system framework, FRAN, is summarized in Algorithm 1. The input of the algorithm constitutes X_S, X_T, Y_S, and hyperparameters. In the beginning, the input signals are normalized, and the network parameters are initialized. Then, the minibatch training is conducted from line 3 to line 14, to optimize the model for *epoch* times. In each iteration, k minibatch data samples and labels are drawn from the input data in a random manner. Then, the source samples and the target samples are passed through the feature extractor G to produce feature representations in line 7. The source-domain prediction results are obtained in line 8 via the classifier F. From line 9 to line 11, the loss functions are computed according to (23)–(25). Last, in line 12, the feature extractor and classifier is optimized to minimize the loss function, and the network parameters θ are updated for the next iteration.

E. Computational Complexity Analysis

In the proposed system framework, $\mathcal{L}_{\text{cls}}$ is the main objective that evaluates the classification error on the labeled source domain, which requires the most computational resources. For a fair comparison, the feature extractor and classifier that produces $\mathcal{L}_{\text{cls}}$ are considered the same as the benchmark algorithms. Denote the computational complexity of F and G as $\mathcal{O}(G)$ and $\mathcal{O}(F)$, respectively. The domain adaptation process occurs when minimizing the objectives $\mathcal{L}_F$ and $\mathcal{L}_D$, which have the complexity, $O(N)$ and $O(k^2)$, respectively. Therefore, the computational

complexity of the proposed framework is $\mathcal{O}(G + F + N + k^2)$. Normally, $G + F \gg N + k^2$, and the computational complexity of the proposed framework can be reduced $\mathcal{O}(G + F)$, which is the same as $F \circ G$. Thus, the system framework proposed in this article should be able to solve cross-domain fault diagnosis problems efficiently.

IV. Experimental Validation

This section presents the results from experiments carried out using the system framework developed in this article. The performance of the proposed system framework is compared with that using state-of-the-art benchmark algorithms on both public and real-world datasets, with the objective of validating the developed system framework.

A. Experimental Setup

In all the experiments carried out for the present purpose, the standard training protocol for UDA [20], [36], [37] is followed, which utilizes all labeled source data samples and unlabeled target data samples. The evaluation metric is the accuracy of the target-domain diagnosis, $\frac{1}{M}\sum_{i=1}^{M}\delta(\mathbf{y_t}^{(i)}, \hat{\mathbf{y}}_\mathbf{t}^{(i)})$. The Adam optimizer [38], which backpropagates through the proposed network to minimize the objective function, is used for the optimization with minibatch to train the model. The Adam optimizer is an efficient stochastic optimizer that requires less memory to compute and can obtain faster convergence. A unified network architecture is used: the feature extractor comprises two convolutional layers with each layer followed by a max-pooling layer. The kernel sizes for both convolutional layers are 2, and the output channels are 32 and 64, respectively. Two fully connected layers, whose output channels are 1000 and 3, respectively, are placed behind for calculating the classification scores. In this work, the hyperparameters are tuned through the grid search to consider all parameter combinations. The learning rate is tuned from 1e-5 to 0.1, and both α and β are tuned from 1e-7 to 1.0. The reported performance is obtained by setting the learning rate to 1e-4, α to 1e-6, β to 1.0, and the minibatch size to 64. Each transfer task is repeated ten times to obtain the average accuracy. All experiments are implemented on the *PyTorch* platform. The conditional entropy of the softmax predictions is minimized for the target samples, which encourages the model to make a more confident prediction on the unlabeled target samples [39], $\mathcal{L}_{\text{ent}} = \frac{1}{|X_T|}\sum_{\mathbf{x}_t \in X_T} -F(G(\mathbf{x}_t)) \log F(G(\mathbf{x}_t))$.

B. Experiment Design and Datasets

According to [26] and [28], four cross-domain fault diagnosis scenarios are conducted in this work to evaluate the proposed framework. First, the performance on two common rotating machine faults (bearing fault and gearbox fault) is tested. The most vulnerable components for a bearing are its inner race, outer race, and ball, which can wear out easily due to metal-to-metal contact under high load and high running speed [40]. The gear tooth is the most vulnerable component of a gearbox, whose frequent fault types are the chipped tooth (CT) and the missing tooth (MT). In real-world scenarios, these faults are inspected by checking if the system vibration and temperature are increased and if there are visible defects on the surface of bearings and gears. Then, complex fault diagnosis scenarios and a real-world bearing fault diagnosis are conducted to further test the performance of the proposed framework.

1) Bearing Fault Diagnosis: Three domains are used in this scenario: two from the CWRU dataset and one from the MFPT dataset. The *CWRU* bearing dataset is provided by the bearing data center in the Case Western Reserve University.[1] Deep grove ball bearings are used for the experiment, which is driven by a three-phase induction motor [40]. Two different settings on the *CWRU* dataset are used in this test scenario, which shares the same experiment platform and sampling rate (12 kHz). The vibration signals in *CWRUA* are collected only from the driven end with the motor load at 0 hp. The vibration signals in *CWRUB* are collected with the motor load at 3 hp. For both *CWRUA* and *CWRUB*, four health conditions are considered: ball fault (B), inner race fault (IR), outer race fault (OR), and normal condition (N). The *MFPT* dataset is provided by MFPT Society,[2] which is gathered from the setups with different deep grove ball bearings and motors. *MFPT* includes the vibration signals under three machine health conditions (OR, IR, and N), which are collected under different motor loads, from 250 lb to 300 lb, with 48-kHz sampling rate.

2) Gearbox Fault Diagnosis: Gearbox data are provided by PHM09 Data Challenge.[3] Vibration signals are collected at five different shaft speeds (30, 35, 40, 45, and 50 rev/s) under high (H) and low (L) load conditions. There are three machine health conditions in this test scenario: CT, MT, and normal. For a fair comparison, the transfer tasks are set as in [27] and [28] to examine the model transfer performance under different shaft speeds and loading. The domains are denoted as "**xxH**" or "**xxL**," where **xx** represents the corresponding shaft speed, and **H** and **L** denotes the load conditions.

3) Fault Severity Diagnosis: To further test the performance of the proposed system framework in handling complex adaptation scenarios, fault severity diagnosis is conducted. The full CWRU bearing dataset is used in this test scenario, where the data samples are vibration signals collected at both drive-end (DE) and fan-end (FE) with three severity levels (fault diameter: 0.007, 0.014, and 0.021 in). These faults are introduced separately at the inner raceway (IR) and outer raceway (OR). Besides, data samples from each severity level are collected at the sampling rate of 12 kHz, under various motor loads from 0 to 3 hp. As for the adaptation scenario, different domains are denoted by "**DExxx**" or "**FExxx**," and **FE** and **DE** denote the different data domains that contain all types of severity level.

4) Real-World Run-to-Failure Stage Diagnosis: In this test scenario, the knowledge transfer performance from public datasets in real-world applications is examined. The public datasets *CWRUA*, *CWRUB*, and *MFPT* are used as the source domain, while a real-world bearing fault scenario is used as

[1][Online]. Available: http://csegroups.case.edu/bearingdatacenter/home
[2][Online]. Available: https://www.mfpt.org/fault-data-sets/
[3][Online]. Available: https://www.phmsociety.org/competition/PHM/09/apparatus

TABLE I
DOMAIN ADAPTATION RESULTS IN BEARING FAULT DIAGNOSIS

Methods	CWRUA→CWRUB	CWRUA→MFPT	MFPT→CWRUA	CWRUB→CWRUA	CWRUB→MFPT	MFPT→CWRUB	Average (%)
SVM [8]	84.6	33.3	55.6	90.8	33.3	55.2	58.8
KNN [6]	83.3	33.7	33.7	90.6	33.3	33.3	51.3
LR [4]	88.3	66.7	55.6	90.8	50.0	55.6	67.8
TCA [41]	88.9	51.3	66.7	86.9	49.3	66.7	68.3
DAFD [27]	48.2	33.4	41.4	68.2	41.7	39.7	45.4
TLPPIFI [28]	90.0	83.7	69.3	91.7	82.0	**70.7**	81.2
DTL [26]	91.1	80.0	40.8	91.4	72.3	42.5	69.7
MCD [24]	73.1	81.0	50.6	87.2	62.3	48.6	67.1
SAFN [29]	71.1	71.0	65.3	81.4	79.7	66.1	72.4
FRAN (proposed)	**91.3**	**93.1**	**71.2**	**91.7**	**93.4**	70.5	**85.2**

the target domain. To better examine the knowledge transfer performance under complex real-world bearing faults, two types of target domains are considered: early degradation stage (*Early*) and serious fault dataset (*Serious*). Early degradation is inobvious and does not affect the operation of the machine, but may develop into serious faults, which may cause damage to the entire machine. Therefore, in the early degradation stage, the maintenance strategy can be adjusted to prevent further degradation. Moreover, the identification of serious fault requires the repair or replacement of the failed components, which prevents damage to the entire machine. The real-world run to failure dataset is collected by the Hangzhou Bearing Test Research Center.[4] In this dataset, deep grove ball bearings are mounted on a shaft that is coupled by a rubber belt and driven by an ac motor. The experimental platform is different to *CWRU* and *MFPT*. In this experiment, four experimental bearings are mounted on a shaft that is driven by a motor running constantly at 3000 r/min. As for the data samples, 10 s of vibration signals are collected at all four experiment bearings, every 5 min at the sampling rate of 24 kHz. During the experiment, radial loads of 10.25 kN are applied to each bearing until the failure occurs, which finally causes an inner race fault in bearing no. 4. The sampling data in the no. 4 bearing are then segmented into three health conditions: normal (N), early degradation (IR_E), and serious fault (IR_S).

C. Experimental Results

The training of the model was conducted on a PC with 4.0-GHz Quad-Core CPU, 16-GB memory, and an Nvidia 1080 GPU. For the fault severity diagnosis, the training time for 100 epochs is 705.3 s, and the test time for all target samples is 0.16 s. For other datasets, the training time is around 7.5 s, and the test time is 0.01 s.

1) Results for Bearing Fault Diagnosis: The identification results of the bearing fault diagnosis are presented in Table I, where → denotes the adaptation is conducted from left-side dataset (source) to right-side dataset (target). It is seen that the proposed framework outperforms benchmark algorithms by a significant margin in all adaptation scenarios except *MFPT→CWRUB*. The proposed framework also achieves the highest average diagnosis accuracy of over 85.0%, suggesting that aligning feature representations can help to transfer the discriminative features from the labeled source domain to the unlabeled target domain. Notably, the adaptation scenarios using *MFPT* as the source domain have the lower diagnosis accuracies (around 70%), which is because *CWRU*-related datasets contain more health condition types than *MFPT*. The *MFPT* dataset only collects three health conditions (OR, IR, and N), while *CWRU*-related datasets have one more health condition: ball faults ("B"). The lack of knowledge in the ball faults results in poor diagnosis accuracy when using *MFPT* as the source domain. Moreover, the proposed framework achieves more significant improvements in the adaptation scenarios with larger domain shifts. For example, the improvements in *CWRUA↔CWRUB* are less significant when compared to other adaptation scenarios. A larger domain shift requires the model to extract more representative features that are shared by the two domains, which can be better analyzed by the FRAN. Therefore, the FRAN is more effective when the domain shift is larger, which makes FRAN suitable for complex adaptation scenarios and real-world applications.

2) Results for Gearbox Fault Diagnosis: The results of gearbox fault diagnosis are presented in Table II. The proposed framework achieves 100% fault identification accuracy in all adaptation scenarios, outperforming the benchmark algorithms by a significant margin. It is seen that all the traditional single-domain classification approaches cannot perform well in cross-domain gearbox fault diagnosis, and the best prediction accuracy is only 62.8%, which is nearly 40.0% less than the proposed framework. In comparison, most DA-based benchmark approaches achieve high accuracy in this test scenario, with at least 95.8% accuracy. However, DA-based benchmark approaches cannot perform well in the adaptation scenarios with large domain shifts, such as the scenarios with the different shaft speeds under high load (**30H→40H** and **45H→35H**). Shaft speed variance under high load may lead to significant changes in the feature representation, which leads to larger domain shifts. The DA-based benchmark approaches only focus on the input-level features that failed to align features representations. Then, when the domain shift is large, the fault-diagnosis-specific features may not be captured, which negatively impacts the diagnosis performance. The improvement achieved by the FRAN suggests that it can successfully extract the fault-diagnosis-specific features on the domain scenarios with large domain shifts, which validates the effectiveness of aligning feature representations using MI and MMFD.

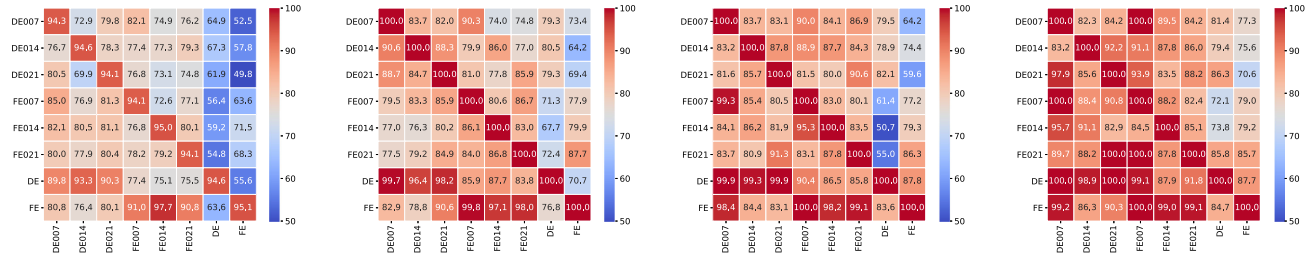
The results for bearing fault diagnosis and gearbox fault diagnosis illustrate the superior performance of DA-based approaches in cross-domain fault diagnosis. Compared to traditional approaches, DA-based approaches successfully transfer the knowledge from the source domain to the target domain.

[4][Online]. Available: https://data.mendeley.com/datasets/z4s9bx4wrn

TABLE II
DOMAIN ADAPTATION RESULTS IN GEARBOX FAULT DIAGNOSIS

Method	30L→50H	40L→50H	50L→35H	30H→40H	40L→50L	45H→35H	35L→35H	40L→40H	50H→50L	Average (%)
SVM	33.3	65.7	33.3	100.0	65.0	56.3	33.3	66.7	66.7	57.8
KNN	33.3	66.7	33.3	66.7	66.3	40.3	33.3	66.7	66.3	52.5
LR	66.7	66.7	33.3	100.0	66.7	66.7	33.3	65.3	66.7	62.8
TCA	33.3	33.3	33.3	66.7	66.7	66.7	66.7	66.7	66.7	55.6
DAFD	61.5	61.8	45.0	67.6	68.4	61.7	48.0	64.9	62.0	60.1
TLPPIFI	100.0	100.0	98.7	100.0	100.0	100.0	100.0	96.7	66.7	95.8
DTL	100.0	100.0	100.0	96.5	100.0	96.7	99.0	100.0	99.2	99.0
MCD	100.0	100.0	100.0	96.7	100.0	100.0	99.3	100.0	100.0	99.6
SAFN	100.0	100.0	96.7	98.4	100.0	95.6	100.0	100.0	99.3	98.9
FRAN (proposed)	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**

"**Xxh**" or "**Xxl**" denotes the gearbox ran on "**Xx**" shaft speed under **H** or **L** load condition.



(a) (b) (c) (d)

Fig. 3. Domain adaptation results in fault severity diagnosis (domain adaptation tasks were conducted from the row labels to the column labels; "**xxx**" in **DExxx** or **FExxx** denotes the fault diameter; **FE** and **DE** contains all severity levels). (a) DTL. (b) MCD. (c) SAFN. (d) FRAN (proposed).

In the following experiments, both complex adaptation scenarios and real-world adaptation tasks are utilized to further examine the performance of the proposed work. Note that only the best performing DA approaches are tested in the following complex scenarios as they significantly outperform traditional approaches.

3) Results for Fault Severity Diagnosis: The results for diagnosing fault severity of a bearing are presented in Fig. 3. The results for each algorithm are summarized using a square matrix for better visualization, where domain adaptation tasks are conducted from row label domains to column label domains. The main diagonal of the result matrix can be regarded as the signal-domain classification since the source domain and the target domain are the same here. FRAN, SAFN, and MCD achieve 100% accuracy in all single-domain classification tasks. Overall, the proposed framework achieves the highest average diagnosis accuracy at 89.9%, while DTL provides the worst average diagnosis accuracy at 77.5%. The SAFN and MCD achieve second-best average diagnosis accuracy at 86.1% and 84.9%, respectively. The worst performance for each algorithm is obtained in the last two columns of the evaluation matrices, which are the most complex adaptation scenarios. These adaptation scenarios only utilize a small part of the dataset to train a model used to diagnose the full dataset. The knowledge of the severity level in each source domain is very limited, which makes it difficult to produce accurate diagnosis on the target domain with more data variations. Nevertheless, the FRAN achieves significantly higher accuracy than the benchmark algorithms on these complex adaptation scenarios by progressively aligning the feature representations using MI and MMFD. Apart from the adaptation scenarios in the last two columns, the FRAN also outperforms the benchmark algorithms significantly on other simple adaptation scenarios. Overall, the superior performance of the FRAN on the fault severity diagnosis suggests the effectiveness of the proposed feature alignment on the knowledge transfer.

4) Results for Real-World Run-to-Failure Stage Diagnosis: The experiments on diagnosing a run-to-failure bearing in different fault stages can validate the generalization performance of the proposed system framework. The results of the real-world run-to-failure stage diagnosis are presented in Fig. 4. Similar to the gearbox fault diagnosis, the FRAN achieves 100% fault identification accuracy in all adaptation scenarios. It is seen that most benchmark algorithms achieve higher diagnosis accuracy when adapting to *Serious* than adapting to *Early*. This is because there is only one fault (inner race fault) that occurred in the test bearing (target domain) in the serious fault stage, which is identical to "IR" feature in the source domains. Besides, the source domains contain more types of data variations, which makes the adaptations be conducted from the complex domains to a relatively simple domain. Therefore, state-of-the-art approaches, such as SAFN, can also successfully detect these adaptation scenarios. In comparison, the machine fault in the early degradation stage is not obvious, and the features of such fault are not similar to all the fault types that occur in the source domain. Therefore, the domain shift between the source domain and the target domain is large. Experimental results show that the FRAN can successfully mitigate the domain shift and obtains 100% diagnosis accuracy, which outperforms the benchmark algorithms significantly. The superior performance of the FRAN

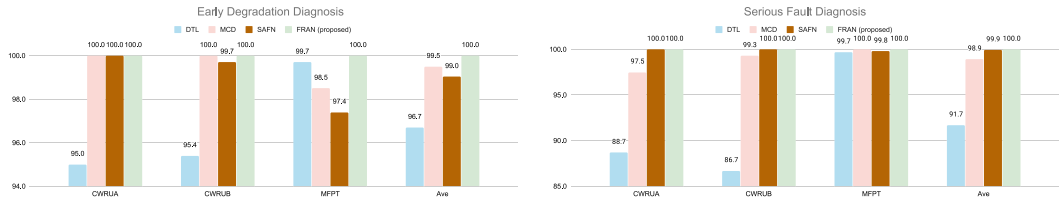


Fig. 4. Domain adaptation results in real-world run-to-failure stage diagnosis (domain adaptation tasks were conducted from x-axis label domains to early/serious stages). (a) Early degradation diagnosis. (b) Serious fault diagnosis.

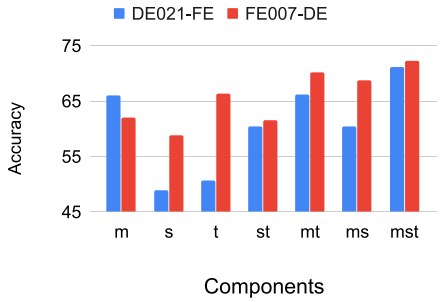


Fig. 5. Ablation study for FRAN components.

on both the early-stage diagnosis and the serious-stage diagnosis suggests its capability in handling both the complex and the simple real-world adaptation scenarios. Furthermore, the ability in diagnosing early-stage degradation can help to give early warnings in real-world applications, thereby preventing further loss.

D. Ablation Study

Fig. 5 presents the contribution of each component of the FRAN: the entropy of the target feature representation, i.e., $\mathsf{H}(\mathbf{z}_t)$, the entropy of the source feature representation, i.e., $\mathsf{H}(\mathbf{z}_s)$, and the MMFD penalty between $\mathbf{z}_s$ and $\mathbf{z}_t$, to the overall performance while keeping all hyperparameters the same. For $x \in \{\mathrm{m}, \mathrm{s}, \mathrm{t}, \mathrm{st}, \mathrm{mt}, \mathrm{ms}, \mathrm{rmk}\}$, FRAN-$x$ denotes the proposed model with only the components x enabled. For this component analysis, m denotes MMFD, s denotes the entropy of the source feature representation, and t denotes the target feature representation. FRAN-s performs the worst suggesting that merely maximizing the entropy of the source feature representation has little help in transferring the knowledge from the source domain to the target domain. FRAN-m and FRAN-t slightly improve the performance from FRAN-s indicating that both maximizing the entropy of the target feature representation and minimizing the feature-level discrepancy can improve the knowledge transferability for the problem of the machine fault diagnosis. Then, further improvements are achieved by FRAN-st, FRAN-mt, and FRAN-ms. The performance of FRAN-m (MMFD) is better than FRAN-st (MI), which suggests that MMFD is more effective than MI if they were applied solely. Overall, the best performance is achieved by jointly using all components of FRAN, i.e., FRAN-mst, which indicates the effectiveness of optimizing MI and MMFD simultaneously.

V. Conclusion

In this article, a novel framework was developed for the cross-domain fault diagnosis as applied to rotating machinery. The proposed approach provided a unified network structure for aligning feature representations to improve the knowledge transfer from the labeled vibration signals (source domain) to the unlabeled vibration signals (target domain). A novel deep neural network was proposed to maximize the MI between the feature space of the source domain and the feature space of the entire input domain while minimizing the discrepancy between the two domains. Through the proposed alignment, more features extracted from the source domain could be used to support the diagnosis in the target domain. Therefore, a more generalized model could be produced to handle complex adaptation scenarios. Furthermore, the process of the MI maximization was simplified to simultaneously maximize the entropy for the feature space of each domain. Experiments were conducted on both the public datasets and the real-world adaptation scenarios, which validated the feasibility and the superior performance of the proposed framework on the cross-domain fault diagnosis.

For the rotating machinery fault diagnosis, the temporal information of the entire vibration signals might reveal the potential of the machine faults and help improve the accuracy of the fault diagnosis. In this article, the proposed approach only focuses on the diagnosis of the input samples but did not consider the temporal correlations among them. More analysis regarding the temporal correlation for the fault diagnosis will be conducted in future works.

References

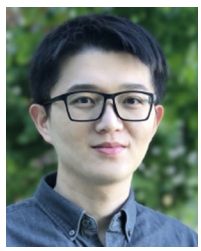
[1] S. Chen, Y. Meng, H. Tang, Y. Tian, N. He, and C. Shao, "Robust deep learning-based diagnosis of mixed faults in rotating machinery," *IEEE/ASME Trans. Mechatronics*, vol. 25, no. 5, pp. 2167–2176, Oct. 2020.

[2] A. Abid, M. T. Khan, H. Lang, and C. W. de Silva, "Adaptive system identification and severity index-based fault diagnosis in motors," *IEEE/ASME Trans. Mechatronics*, vol. 24, no. 4, pp. 1628–1639, Aug. 2019.

[3] X. Zhao, M. Jia, P. Ding, C. Yang, D. She, and Z. Liu, "Intelligent fault diagnosis of multichannel motor-rotor system based on multimanifold deep extreme learning machine," *IEEE/ASME Trans. Mechatronics*, vol. 25, no. 5, pp. 2177–2187, Oct. 2020.

[4] D. W. Hosmer Jr., S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, vol. 398. New York, NY, USA: Wiley, 2013.

[5] B. Li and Y. Zhang, "Supervised locally linear embedding projection (SLLEP) for machinery fault diagnosis," *Mech. Syst. Signal Process.*, vol. 25, no. 8, pp. 3125–3134, 2011.
[6] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy k-nearest neighbor algorithm," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-15, no. 4, pp. 580–585, Jul./Aug. 1985.
[7] D. Wang, "K-nearest neighbors based methods for identification of different gear crack levels under different motor speeds and loads: Revisited," *Mech. Syst. Signal Process.*, vol. 70, pp. 201–208, 2016.
[8] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, 1999.
[9] J. Zheng, H. Pan, and J. Cheng, "Rolling bearing fault detection and diagnosis based on composite multiscale fuzzy entropy and ensemble support vector machines," *Mech. Syst. Signal Process.*, vol. 85, pp. 746–759, 2017.
[10] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.
[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[12] J. Chen, T. Li, J. Wang, and C. W. de Silva, "WSN sampling optimization for signal reconstruction using spatiotemporal autoencoder," *IEEE Sens. J.*, vol. 20, no. 23, pp. 14290–14301, Dec. 2020.
[13] P. Tamilselvan and P. Wang, "Failure diagnosis using deep belief learning based health state classification," *Rel. Eng. Syst. Saf.*, vol. 115, pp. 124–135, 2013.
[14] M. Gan, C. Wang, and C. Zhu, "Construction of hierarchical diagnosis network based on deep learning and its application in the fault pattern recognition of rolling element bearings," *Mech. Syst. Signal Process.*, vol. 72, pp. 92–104, 2016.
[15] H. C. Cho, J. Knowles, M. S. Fadali, and K. S. Lee, "Fault detection and isolation of induction motors using recurrent neural networks and dynamic Bayesian modeling," *IEEE Trans. Control Syst. Technol.*, vol. 18, no. 2, pp. 430–437, Mar. 2010.
[16] M. Xia, T. Li, L. Xu, L. Liu, and C. W. De Silva, "Fault diagnosis for rotating machinery using multiple sensors and convolutional neural networks," *IEEE/ASME Trans. Mechatronics*, vol. 23, no. 1, pp. 101–110, Feb. 2018.
[17] Y. Lei, F. Jia, J. Lin, S. Xing, and S. X. Ding, "An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data," *IEEE Trans. Ind. Electron.*, vol. 63, no. 5, pp. 3137–3147, May 2016.
[18] J. Hoffman, E. Rodner, J. Donahue, T. Darrell, and K. Saenko, "Efficient learning of domain-invariant image representations," in *Proc. Int. Conf. Learn. Representations*, 2013.
[19] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[20] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Mach. Learn. Res.*, PMLR, vol. 37, Jul. 7-9, 2015, pp. 1180–1189.
[21] Z. Liu, B. Lu, H. Wei, L. Chen, X. Li, and M. Rtsch, "Deep adversarial domain adaptation model for bearing fault diagnosis," *IEEE Trans. Syst., Man, Cybern. Syst.*, to be published, doi: 10.1109/TSMC.2019.2932000.
[22] X. Li, W. Zhang, Q. Ding, and X. Li, "Diagnosing rotating machines with weakly supervised data using deep transfer learning," *IEEE Trans. Ind. Inform.*, vol. 16, no. 3, pp. 1688–1697, Mar. 2020.
[23] X. Li, W. Zhang, N. Xu, and Q. Ding, "Deep learning-based machinery fault diagnostics with domain adaptation across sensors at different places," *IEEE Trans. Ind. Electron.*, vol. 67, no. 8, pp. 6785–6794, Aug. 2020.
[24] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3723–3732.
[25] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e 49–e57, 2006.
[26] 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.
[27] W. Lu, B. Liang, Y. Cheng, D. Meng, J. Yang, and T. Zhang, "Deep model based domain adaptation for fault diagnosis," *IEEE Trans. Ind. Electron.*, vol. 64, no. 3, pp. 2296–2305, Mar. 2017.
[28] H. Zheng, R. Wang, J. Yin, Y. Li, H. Lu, and M. Xu, "A new intelligent fault identification method based on transfer locality preserving projection for actual diagnosis scenario of rotating machinery," *Mech. Syst. Signal Process.*, vol. 135, 2020, Art. no. 106344.
[29] R. Xu, G. Li, J. Yang, and L. Lin, "Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1426–1435.
[30] J. Wang, J. Chen, J. Lin, L. Sigal, and C. W. de Silva, "Discriminative feature alignment: Improving transferability of unsupervised domain adaptation by Gaussian-guided latent alignment," 2020, *arXiv:2006.12770*.
[31] W. M. Kouw, L. J. Van Der Maaten, J. H. Krijthe, and M. Loog, "Feature-level domain adaptation," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 5943–5974, 2016.
[32] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, 2012.
[33] A. Smola, A. Gretton, L. Song, and B. Schölkopf, "A Hilbert space embedding for distributions," in *Proc. Int. Conf. Algorithmic Learn. Theory*, 2007, pp. 13–31.
[34] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet, "Hilbert space embeddings and metrics on probability measures," *J. Mach. Learn. Res.*, vol. 11, pp. 1517–1561, Apr. 2010.
[35] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017,vol. 70, pp. 2208–2217.
[36] Z. Cao, M. Long, J. Wang, and M. I. Jordan, "Partial transfer learning with selective adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2724–2732.
[37] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7167–7176.
[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, San Diego, CA, USA, May 7-9, 2015.
[39] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2005, pp. 529–536.
[40] X. Lou and K. A. Loparo, "Bearing fault diagnosis based on wavelet transform and fuzzy inference," *Mech. Syst. Signal Process.*, vol. 18, no. 5, pp. 1077–1095, 2004.
[41] J. Xie, L. Zhang, L. Duan, and J. Wang, "On cross-domain feature fusion in gearbox fault diagnosis under various operating conditions based on transfer component analysis," in *Proc. IEEE Int. Conf. Prognostics Health Manage.*, 2016, pp. 1–6.



Jiahong Chen (Member, IEEE) received the Ph.D. degree in mechanical engineering from the University of British Columbia, Vancouver, BC, Canada, in 2019.

He is currently a Postdoctoral Fellow with the Department of Mechanical Engineering, The University of British Columbia. His current research interests include signal processing, wireless sensor networks, and machine learning.

Jing Wang received the B.A.Sc. degree in electrical and computer engineering in 2018 from the University of British Columbia, Vancouver, BC, Canada, where he is currently working toward the M.A.Sc. degree.

His research interests include transfer learning, reinforcement learning, and silicon photonics.

Jianxin Zhu received the Ph.D. degree in engineering thermal physics from Zhejiang University, Hangzhou, China, in 2004.

He is currently a Research Fellow with the Hefei General Machinery Research Institute, Hefei, China. His main research interests include failure diagnosis and integrity management of mechanical equipment in the process industry.

Tong Heng Lee (Member, IEEE) received the B.A. degree (with first class Hons.) in the engineering tripos from Cambridge University, Cambridge, U.K., in 1980, the M.Eng. degree in electrical engineering from the National University of Singapore (NUS), Singapore, in 1985, and the Ph.D. degree in electrical engineering from Yale University, New Haven, CT, USA, in 1987.

He is currently a Professor with the Department of Electrical and Computer Engineering, NUS, and also a Professor with the NUS Graduate School. He was a Past Vice-President (Research) of NUS. His research interests include adaptive systems, knowledge-based control, intelligent mechatronics, and computational intelligence.

Dr. Lee is an Associate Editor for the IEEE TRANSACTIONS IN SYSTEMS, MAN, AND CYBERNETICS, *Control Engineering Practice*, and the *International Journal of Systems Science*. He is also the Deputy Editor-in-Chief of *Mechatronics*.

Clarence W. de Silva (Life Fellow, IEEE) received the Ph.D. degree in dynamics and control of automated guided vehicle systems from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1978, and the University of Cambridge, Cambridge, U.K., in 1998, and the honorary D.Eng. degree in dynamic systems and control from the University of Waterloo, Waterloo, ON, Canada, in 2008.

Since 1988, he has been a Professor of mechanical engineering, a Senior Canada Research Chair, and an NSERC-BC Packers Chair in industrial automation with the University of British Columbia, Vancouver, BC, Canada. He has authored 24 books and more than 550 papers, approximately half of which are in journals.

Prof. de Silva is also a Fellow of the American Society of Mechanical Engineers, the Canadian Academy of Engineering, and the Royal Society of Canada.