IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 21, 2019, accepted September 2, 2019, date of publication September 6, 2019,
date of current version September 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2939876

Cross-Domain Fault Diagnosis Using Knowledge Transfer Strategy: A Review

HUAILIANG ZHENG[1], RIXIN WANG[1], YUANTAO YANG[1], JIANCHENG YIN[1], YONGBO LI[2], (Member, IEEE), YUQING LI[1], AND MINQIANG XU[1]

[1]Deep Space Exploration Research Center, Harbin Institute of Technology, Harbin 150001, China
[2]MIIT Key Laboratory of Dynamics and Control of Complex Systems, Northwestern Polytechnical University, Xi'an 710072, China

Corresponding author: Rixin Wang (wangrx@hit.edu.cn)

ABSTRACT Data-driven fault diagnosis has been a hot topic in recent years with the development of machine learning techniques. However, the prerequisite that the training data and the test data should follow an identical distribution prevents the conventional data-driven diagnosis methods from being applied to the engineering diagnosis problems. To tackle this dilemma, cross-domain fault diagnosis using knowledge transfer strategy is becoming popular in the past five years. The diagnosis methods based on transfer learning aim to build models that can perform well on target tasks by leveraging knowledge from semantic related but distribution different source domains. This paper for the first time summarizes the state-of-art cross-domain fault diagnosis research works. The literatures are introduced from three different viewpoints: research motivations, cross-domain strategies, and application objects. In addition, the corresponding open-source fault datasets and several future directions are also presented. The survey provides readers a framework for better understanding and identifying the research status, challenges and future directions of cross-domain fault diagnosis.

INDEX TERMS Cross-domain, domain adaptation, fault diagnosis, review, transfer learning.

NOMENCLATURE

C	The number of classes in diagnosis tasks
C_s	The number of classes of the source task
C_t	The number of classes of the target task
C_{svm}	The regularization parameter of SVM
conv	Convolutional layer
$\mathcal{D}$	Domain
$\mathcal{D}_s$	Source domain
$\mathcal{D}_t$	Target domain
D	Dimension of feature space $\mathcal{X}$
d	Dimension of subspace
D	Discriminative model of GAN
Dic	Shared dictionary matrix in (14)
Dic_c	Sub-dictionary corresponding to class c
$\mathbf{E}_s$	Noise matrix of source domain with respect to Dic
$\mathbf{E}_t$	Noise matrix of target domain with respect to Dic
$\mathbf{F} = [\mathbf{F}_s, \mathbf{F}_t]^{\text{T}}$	Embedded matrix of the source domain samples and the target domain samples in the common subspace computed from (19)
$f(\cdot)$	Prediction function of task $\mathcal{T}$
$f_s(\cdot)$	Prediction function of source task $\mathcal{T}_s$
$f_t(\cdot)$	Prediction function of target task $\mathcal{T}_t$
$\Delta f(\cdot)$	Bias term between $f_s(\cdot)$ and $f_t(\cdot)$
fc	Fully-connected layer
G	Generative model of GAN
G_f,G_c,G_d	Feature extractor, label predictor, and domain discriminator in DANN
g_i	Ground-truth domain label of $\boldsymbol{x}_i$
$\hat{g}(\boldsymbol{x}_i)$	Output of domain classifier with respect to $\boldsymbol{x}_i$
$\mathcal{H}$	Reproducing Kernel Hilbert Space
$\mathcal{I} = \begin{bmatrix} 0 & \mathbf{I} \\ \mathbf{I} & 0 \end{bmatrix}$	Inter-set correspondence between samples of the source domain and the target domain datasets in (19)
$\mathbf{I}$	Identity matrix
$I[\cdot]$	Indicator function

The associate editor coordinating the review of this manuscript and approving it for publication was Kezhi Li.

$\mathbf{K}$	Kernel matrix
$K_{i,j} = K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)$	The (i,j) entry of kernel matrix $\mathbf{K}$, and $K(\cdot,\cdot)$ denotes a kernel function
K_{poly}	Polynomial kernel function
K_{rbf}	RBF kernel function
$\mathcal{L}_{\text{ae}}$	Cost function of auto-encoder
$\mathcal{L}_c$	Supervised learning (classification) cost of deep neural networks
$\mathcal{L}_{\text{cluster}} = -\mathcal{L}_{\text{inter}} + \eta\mathcal{L}_{\text{intra}}$	Representation clustering cost in (30), where $\mathcal{L}_{\text{inter}}$ is the inter-class separability cost, and $\mathcal{L}_{\text{intra}}$ is the intra-class compactness cost
$\mathcal{L}_{\text{D}}$	Distribution distance cost in deep domain adaptation
$\mathcal{L}_d$	Domain classification cost
$\mathcal{L}_{\text{grad}}$	Gradient penalty in (40)
$\mathcal{L}_{\text{SF}}$	Cost function of sparse filtering
$\mathcal{L}_{\text{weight}}$	Weight regularization term in (33)
L	Layer number of deep neural networks
$\ell(\boldsymbol{x}, y, \theta)$	Prediction loss function with respect to sample $(\boldsymbol{x}, y)$ under model parameter θ
$\ell_d(\cdot)$	Domain prediction loss function
$\mathbf{M}$	MMD matrix with each entry $M_{i,j}$
$\mathbf{M}_0$	MMD matrix of marginal distribution distance
$\mathbf{M}_c$	MMD matrix of the distribution distance between the samples of c-th category from two different domains
m	Batch size when training deep neural networks based on mini-batch gradient descent algorithm
n_a	The number of labeled samples of the source and target domains
n_s	Sample size of dataset X_s
n_t	Sample size of dataset X_t
n_s^c	Sample size of X_s^c
n_t^c	Sample size of X_t^c
$\mathcal{P}(X)$	Marginal probability distribution of domain $\mathcal{D}$
$\mathcal{P}_{data}$	Data distribution of GAN model
$\mathcal{P}_s(X_s)$	Marginal probability distribution of $\mathcal{D}_s$
$\tilde{\mathcal{P}}_s(\cdot)$	Empirical estimation of $\mathcal{P}_s(\cdot)$
$\mathcal{P}_t(X_t)$	Marginal probability distribution of $\mathcal{D}_t$
$\mathcal{P}_z(z)$	Input noise distribution of GAN model
pool	Pooling layer
$\mathcal{Q}(Y\|X)$	Conditional probability distribution of task $\mathcal{T}$
$\mathcal{Q}_s(Y_s\|X_s)$	Conditional probability distribution of $\mathcal{T}_s$
$\mathcal{Q}_t(Y_t\|X_t)$	Conditional probability distribution of $\mathcal{T}_t$
$\mathbf{Q}_s$	Ideal representation coefficient matrix of source domain in supervised dictionary-based transfer subspace learning
$\mathbf{Q}_t$	Ideal representation coefficient matrix of target domain in supervised dictionary-based transfer subspace learning
$\boldsymbol{\mathcal{R}}_s$	Reconstruction coefficient matrix of source domain based on LRE
$\boldsymbol{\mathcal{R}}_t$	Reconstruction coefficient matrix of target domain based on LRE
$\boldsymbol{\mathcal{R}} = \begin{bmatrix} \boldsymbol{\mathcal{R}}_s & 0 \\ 0 & \boldsymbol{\mathcal{R}}_t \end{bmatrix}$	Block reconstruction coefficient matrix in (19)
$\mathbf{R}_s$	Representation coefficient matrix of source domain with respect to Dic
$\mathbf{R}_t$	Representation coefficient matrix of target domain with respect to Dic
$\boldsymbol{S}_w$	Within-class scatter matrix
$\boldsymbol{S}_b$	Between-class scatter matrix
$\mathcal{T}$	Task
$\mathcal{T}_s$	Source task
$\mathcal{T}_t$	Target task
$\mathbf{W}$	Transformation matrix from the original space or RKHS to the new d-dimensional subspace
$W(\mathcal{P}_s, \mathcal{P}_t)$	Wasserstein distance between $\mathcal{P}_s$ and $\mathcal{P}_t$
$\mathcal{X}$	Feature space of domain $\mathcal{D}$
$\mathcal{X}_s$	Feature space of source domain $\mathcal{D}_s$
$\mathcal{X}_t$	Feature space of target domain $\mathcal{D}_t$
$\mathbf{X}$	Data matrix in original feature space
$\mathbf{X}_s \in \mathbb{R}^{\text{D}\times n_s}$	Data matrix of source domain
$\mathbf{X}_t \in \mathbb{R}^{\text{D}\times n_t}$	Data matrix of target domain
$\mathbf{X}_{\text{D}} = [\mathbf{X}_s, \mathbf{X}_t]$	Data matrix of training dataset, including the samples of source and target domains.
X	A dataset sampled from domain $\mathcal{D}$
X_s	A dataset sampled from source domain $\mathcal{D}_s$
X_t	A dataset sampled from target domain $\mathcal{D}_t$
X_t^l	Labeled sample set from target domain $\mathcal{D}_t$
X_s^c	Sample set of c-th class of the source domain
X_t^c	Sample set of c-th class of the target domain

$\boldsymbol{x}_i$	A sample in X
$\bar{\boldsymbol{x}}_c$	The mean of the c-th samples
$\bar{\boldsymbol{x}}_0$	The mean of all samples
$\boldsymbol{x}_i^s$	A sample in X_s
$\boldsymbol{x}_i^t$	A sample in X_t
$\mathcal{Y}$	Label space of task $\mathcal{T}$
$\mathcal{Y}_s$	Label space of source task $\mathcal{T}_s$
$\mathcal{Y}_t$	Label space of target task $\mathcal{T}_t$
$\mathbf{Y}_{\mathrm{D}}$	virtual label matrix of the dictionary Dic
Y	Label vector corresponding to X
Y_s	Label vector corresponding to X_s
Y_t	Label vector corresponding to X_t
y_i	Label corresponding to sample $\boldsymbol{x}_i$
y_i^s	Label corresponding to sample $\boldsymbol{x}_i^s$
$\mathbf{Z}_a$	Aligned subspace
$\mathbf{Z}_s \in \mathbb{R}^{\mathrm{D}\times d}$	d dimension subspace of source domain
$\mathbf{Z}_t \in \mathbb{R}^{\mathrm{D}\times d}$	d dimension subspace of target domain

GREEK LETTERS

α	Multi-kernel coefficient
β	Regularization parameter that trades off the model complexity, used in (10), (33)
γ	Penalty parameter of domain classifier cost, used in (29), (38)
ζ	Joint probability distribution
η	Trade-off parameter between inter-class separability and intra-class compactness in $\mathcal{L}_{\mathrm{cluster}}$
$\theta \in \Theta$	A model parameter family for seeking the optimal solution, or the parameters of deep neural networks
θ_{com}	The common parts between θ_s and θ_t
θ_s,θ_t	Network parameters for the source task and the target task, respectively
θ_s',θ_t'	Specific parameters for the source task and the target task, respectively
θ_f,θ_c,θ_d	Parameters of the feature extractor, health condition classifier (label predictor), and the domain classifier (domain discriminator), respectively.
θ^*	The optimal model parameter
ϑ	Penalty parameter of representation clustering cost $\mathcal{L}_{\mathrm{cluster}}$ in (30)
Λ	Indexes of layers in (32)
λ	Penalty parameter of distribution distance cost, used in (8), (21), (29), (30), (31), (33), (34), (35)
μ	Trade-off parameter that dominates the importance of the local geometry in (19)
ξ_i	Penalizing variable of SVM model
$\prod(\mathcal{P}_s, \mathcal{P}_t)$	The set $\Psi \times \Psi$ of all joint distributions
ρ	A coefficient that balances domain critic loss and gradient penalty in (40)
τ	A punishment factor in $\mathcal{L}_{\mathrm{weight}}$ term
υ	Penalty parameter of the sparsity of noise matrices in (15)
Φ^l	The l-th layer representation of deep neural network
Φ_s^l	The l-th layer representation of deep neural network with respect to source domain samples
Φ_t^l	The l-th layer representation of deep neural network with respect to target domain samples
$\hat{\varphi} = \{\hat{\varphi}_s, \hat{\varphi}_t\}$	The combination of normalized feature matrix of the source domain data $\hat{\varphi}_s$ and the target domain data $\hat{\varphi}_t$
Ψ	Compact metric set
$\phi(\boldsymbol{x})$	Map from $\mathcal{X}$ to $\mathcal{H}$

ABBREVIATIONS

AdaBN	Adaptive Batch Normalization
ADDA	Adversarial Discriminative Domain Adaptation
AE	Auto-encoder
A-SVM	Adaptive Support Vector Machines
A2CNN	Adversarial Adaptive 1-D CNN
CAN	Convolutional Adaptation Network
CBM	Condition Based Monitoring
CNN	Convolutional Neural Network
CORAL	Correlation alignment
CWRU	Case Western Reserve University
DA-DCGAN	Domain Adaptation combined with Deep Convolutional Generative Adversarial Network
DAFD	**D**eep neural network for domain **A**daptation in **F**ault **D**iagnosis
DAFTL	Domain Adaptation by using Feature Transfer Learning
DAN	Deep Adaptation Network
DANN	Domain Adversarial Neural Network
DATF	Domain Adaptation using Transferable Features
DBN	Deep Belief Network
DCTLN	Deep Convolutional Transfer Learning Network
DIRG	Dynamic and Identification Research Group
DOF	Degree of Freedom
FTNN	Feature-based Transfer Neural Network
GAN	Generative Adversarial Networks
HKL	High-order Kullback-Leibler
IMS	Intelligent Maintenance System
JDA	Joint Distribution Adaptation
KL	Kullback-Leibler divergence
KNN	k-Nearest Neighbor
LRE	Low Rank Embedding
LSSVM	Least Square Support Vector Machine

MFPT	Society for Machinery Failure Prevention Technology
MMD	Maximum Mean Discrepancy
MK-MMD	Multiple Kernel MMD
PCA	Principal Components Analysis
PHM	Prognostic and Health Management
RBF	Radial Basis Function
RBM	Restricted Boltzmann Machines
RF	Random Forest
RKHS	Reproducing Kernel Hilbert Space
RL	Railway Locomotive
RNN	Recurrent Neural Networks
RUL	Remaining Useful Life
SAE	Sparse auto-encoder
sAE	Stacked Auto-encoders
SF	Sparse Filtering
SSTCA	Semi-supervised TCA
STPN	Spatiotemporal Pattern Network
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TCA	Transfer Component Analysis
TICNN	Convolution Neural Networks with Training Interference
WDMAN	Wasserstein Distance Guided Multi-Adversarial Networks
WGAN	Wasserstein Generative Adversarial Networks

I. INTRODUCTION

With the development of modern industries, both the demands for mechanical systems that provide higher reliability and safety and the new challenge of maintenance management are raised. To meet these demands, it is a promising means developing Prognostic and Health Management (PHM) systems that aim to reasonably allocate maintenance resources through monitoring the real-time health condition and the trend of performance degradation [1]–[3]. As one of the essential components of PHM, fault diagnosis that focuses on detecting and identifying faults is crucial to guarantee safe operation and avoid economic loss in industry applications [4].

In recent years, data-driven fault diagnosis methods have been a hot topic due to the accumulation of industrial big data and the rapid development of machine learning especially deep learning [5]–[7]. In general, the conventional machine learning algorithms such as Support Vector Machine (SVM) [8], Random Forest (RF) [9], and k-Nearest Neighbor (KNN) [10] or the deep neural networks such as Stacked Auto-encoders (sAE) [11], Deep Belief Network (DBN) [12], and Convolutional Neural Network (CNN) [13], [14] are employed to learn the fault characteristics and identification models from massive amounts of historical data. The general implementation procedures of data-driven diagnosis methods based on conventional machine learning and deep learning are illustrated in Fig.1 (a) and Fig.1 (b), respectively.

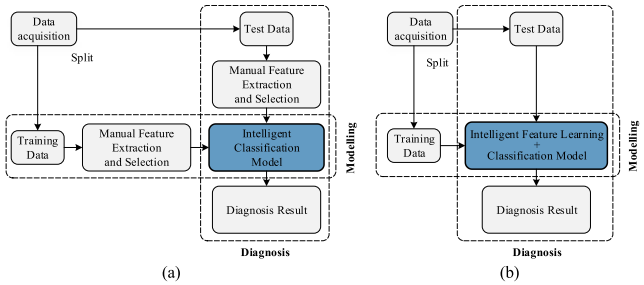


FIGURE 1. Framework showing general implementation procedures of data-driven fault diagnosis. (a) procedures of the conventional machine learning based methods, (b) procedures of the deep learning based methods.

The step of feature extraction is necessary for most of the methods that use conventional machine learning, such as time-domain statistics, frequency-domain analysis, and time-frequency analysis [15]. Differently, the diagnosis methods based on deep learning can automatically learn discriminative features from raw monitoring signals without manual feature extraction and selection, as shown in Fig.1 (b).

Data is the carrier of diagnosis knowledge and dominates the performance of data-driven diagnosis models. To ensure the robustness and the generalization performance on test data, two prerequisites should be satisfied in the stage of training diagnosis models: (1) massive amounts of high-quality annotated data are available, (2) the data to be tested should be drawn from the same distribution with the training data. As a matter of fact, as shown in Fig.1, the commonly used validation manner of data-driven diagnosis methods [8]–[14] guaranteed those two prerequisites through splitting one dataset into the training set and the test set. However, in practical diagnosis scenario, this validation manner is impractical, and to satisfy the prerequisites is very difficult due to the follows two issues:

(1) Generally, it is hard or even impossible to obtain a training dataset with the same distribution as the test dataset before building the diagnosis model, because it means that we need to collect data of each fault category under the same machine and even the same operating conditions with the target one.

(2) For in-service machines, scarce labeled fault data can be obtained, because it may not be allowed to work continuously under faulty conditions.

These two obstacles prevent the diagnosis methods based on conventional machine learning and deep learning from being applied to the engineering fault diagnosis.

Actually, in practical scenario, the available fault data for training identification models are usually collected from different operating conditions, other same-type machines, or fault simulation experiments in the laboratory. These data from multiple different sources may follow different distributions from the test data we interested, due to the differences existed in physical space. But there are two underlying probabilities for building effective diagnosis models by using these data. First, similar fault characteristics

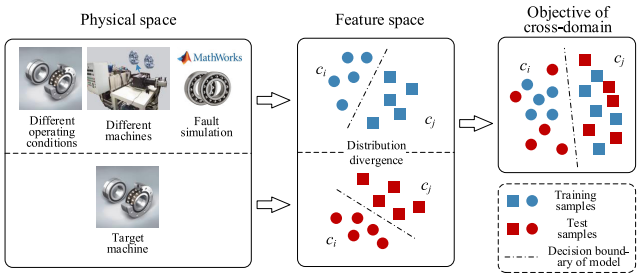


FIGURE 2. Intuitive illustration of cross-domain fault diagnosis.

should be contained in these data from multiple sources because of the same working principle and the similar failure mechanism of the machines from which these data are generated. Second, the data-driven fault diagnosis that is called intelligent fault diagnosis aims to imitate the diagnosticians using machine learning techniques, whereas diagnosticians can diagnose faults by extending the knowledge learned from other same-type machines, but not only the knowledge learned from the machine they interested. That is to say, leveraging knowledge from related datasets when building diagnosis models is feasible. However, when the distribution divergence exists, the conventional machine learning and deep learning techniques all cannot be used directly. To make the best use of previous multiple source data, cross-domain fault diagnosis is a new attempt that holds the potential to overcome the obstacles in the current data-driven fault diagnosis. In the context of this paper, cross-domain diagnosis means that the training data and the test data can be drawn from different potential distributions, and its objective is to construct diagnosis models with considerable generalization performance on the test data, as illustrated in Fig.2.

In the past five years, many papers studied the cross-domain fault diagnosis problem and most of them employed the knowledge transfer strategy. The objective of this paper is to review the related state-of-art cross-domain fault diagnosis research works. Currently, there have been other surveys on data-driven fault diagnosis techniques during the past few years [5]–[7], [16], but they only introduced the diagnosis methods based on conventional machine learning or deep learning and none of them considered the cross-domain diagnosis techniques. For example, the survey written by Zhao *et al.* [6] categorized and reviewed the deep learning-based diagnosis methods according to the network architectures, including Auto-encoder (AE) and its variants, Restricted Boltzmann Machines (RBM) and its variants, CNN and Recurrent Neural Networks (RNN). In addition, there are also reviews on transfer learning or domain adaptation [17]–[21], but none of them contain the methods applied in machinery fault diagnosis field.

In this paper, the knowledge transfer strategies for machinery cross-domain fault diagnosis are mainly focused. Specifically, the key contributions of this review are as follows: (1) For the first time, we present a systematic introduction of the research works about cross-domain fault diagnosis according to research motivations, cross-domain strategies, and application objects. (2) In this review, all of the traditional transfer approaches, deep transfer approaches, and adversarial-based transfer approaches are included, while some cross-domain diagnosis approaches without transfer are also summarized. (3) We give a comprehensive summary of the open-source datasets for facilitating readers to start studies of cross-domain fault diagnosis. (4) Several future research directions are discussed on cross-domain fault diagnosis.

The remainder of this paper is organized as follows. We start with an introduction of transfer learning in Section II. Then, Section III reviews the cross-domain fault diagnosis according to research motivations and problem settings, cross-domain approaches, and applications. Section IV provides a comprehensive summary of the open-source fault datasets. In Section V, the discussions and future directions are presented. Finally, the paper is concluded in Section VI.

II. OVERVIEW OF TRANSFER LEARNING

Before reviewing the research works about cross-domain fault diagnosis most of which employed transfer learning methods, a brief overview to transfer learning is given in this section. The basic definitions about transfer learning are given firstly, and then the basic transfer ideas according to ''what to transfer'' are introduced briefly. These contents will help readers understand what is transfer learning and what are basic existing strategies to implement knowledge transfer.

A. DEFINITION OF TRANSFER LEARNING

Transfer learning aims to address the learning problems between two or multiple domains. A Domain $\mathcal{D}$, as defined by ***Definition 1***, is a mathematical description of the characteristics of the corresponding subjects or systems, such as the characteristics of images in image classification, the characteristics of vibration signals in bearing and gear fault diagnosis. The feature space $\mathcal{X}$ describes the characteristics of the subjects through D features, meanwhile $\mathcal{P}(X)$ describes the specific distribution state of the considered problems. Corresponding to a domain $\mathcal{D}$, the Task $\mathcal{T}$, as given by ***Definition 2***, defines the learning objective, that is to say the mapping relation between Y and X.

Definition 1 (Domain [17]): A Domain $\mathcal{D} = \{\mathcal{X}, \mathcal{P}(X)\}$ is composed of two components: a feature space $\mathcal{X}$ and a marginal probability distribution $\mathcal{D}(X)$, where $X = \{\boldsymbol{x}_i\}_{i=1}^{n} \in \mathcal{X}$ is a dataset with each $\boldsymbol{x}_i \in \mathbb{R}^{\mathrm{D}}$ sampled from this domain.

Definition 2 (Task [17]): Given a Domain $\mathcal{D} = \{\mathcal{X}, \mathcal{P}(X)\}$, a Task $\mathcal{T} = \{\mathcal{Y}, f(X)\}$ is also composed of two components: a label space $\mathcal{Y}$ and a prediction function $f(X) = \mathcal{Q}(Y|X)$, where $Y = \{y_i\}_{i=1}^{n}$ is the label vector of X with $y_i \in \mathcal{Y}$ is the label of $\boldsymbol{x}_i$, $\mathcal{Q}(Y|X)$ is the conditional probability distribution.

Based on domain $\mathcal{D}$ and task $\mathcal{T}$, the definition of Transfer Learning is given by ***Definition 3***. Generally, in transfer learning, there are a source domain $\mathcal{D}_s = \{\mathcal{X}_s, \mathcal{P}_s(X_s)\}$

and a target domain $\mathcal{D}_t = \{\mathcal{X}_t, \mathcal{P}_t(X_t)\}$, where $\mathcal{X}_s$ and $\mathcal{X}_t$ denote the feature spaces of the source and target domains respectively, $\mathcal{P}_s(X_s)$ and $\mathcal{P}_t(X_t)$ denote the marginal probability distributions of the source and target domains respectively. Correspondingly, there are a source task $\mathcal{T}_s = \{\mathcal{Y}_s, \mathcal{Q}_s(Y_s|X_s)\}$ and a target task $\mathcal{T}_t = \{\mathcal{Y}_t, \mathcal{Q}_t(Y_t|X_t)\}$, where $\mathcal{Y}_s$ and $\mathcal{Y}_t$ are the label spaces of the source and target tasks respectively, $\mathcal{Q}_s(Y_s|X_s)$ and $\mathcal{Q}_t(Y_t|X_t)$ are the conditional probability distributions of the source and target domains respectively. Let $X_s = \{\boldsymbol{x}_i^s \in \mathcal{X}_s\}_{i=1}^{n_s}$ denotes a dataset with n_s samples from the source domain, and $X_t = \{\boldsymbol{x}_i^t \in \mathcal{X}_t\}_{i=1}^{n_t}$ denotes a dataset with n_t samples from the target domain. Usually, the samples from the source domain are fully labeled and the corresponding label is $Y_s = \{y_i^s \in \mathcal{Y}_s\}_{i=1}^{n_s}$. But the samples from the target domain may be fully labeled ($n_t \ll n_s$), unlabeled or partially labeled in specific problem settings.

Domain adaptation, a concept related to transfer learning, has aroused wide concern recently and has been widely developed for tackling cross-domain learning tasks. It is a sub-problem of transfer learning in which the source task and the target task are the same. It means that $\mathcal{Y}_s = \mathcal{Y}_t$ and $\mathcal{Q}_s(Y_s|X_s) = \mathcal{Q}_t(Y_t|X_t)$. However, the second item is rather strong and does not always hold in real life applications. Therefore, the definition of domain adaptation is relaxed to the case where only the $\mathcal{Y}_s = \mathcal{Y}_t$ is required [22].

Definition 3 (Transfer Learning [17]): Given a source domain $\mathcal{D}_s$ and learning task $\mathcal{T}_s$, a target domain $\mathcal{D}_t$ and learning task $\mathcal{T}_t$, transfer learning aims to promote the performance of target predictive function $f_t(\cdot)$ in $\mathcal{D}_t$ through leveraging the knowledge in $\mathcal{D}_s$ and $\mathcal{T}_s$, where $\mathcal{D}_s \neq \mathcal{D}_t$ or $\mathcal{T}_s \neq \mathcal{T}_t$.

Definition 4 (Domain Adaptation [22]): Domain Adaptation is a sub-problem of transfer learning, where it is assumed that the source task $\mathcal{T}_s$ and the target task $\mathcal{T}_t$ are the same, *i.e.* $\mathcal{T}_s = \mathcal{T}_t$.

B. COMPARISON WITH TRADITIONAL MACHINE LEARNING

Traditional machine learning methods have made tremendous contributions to classification, regression, and clustering tasks in computer vision [23], natural language processing [24], as well as fault diagnosis [5]–[7]. However, these traditional machine learning algorithms which are under the framework of statistical learning theory follow a basic assumption that the training data and the test data are drawn from the same distribution. If this assumption is not hold, the generalization performance of these methods may drop dramatically. Unfortunately, the distribution discrepancy between datasets is a universal phenomenon in real world applications. For example, in fault diagnosis, the different operating conditions, loads, positions of sensors, and machine sizes etc. may cause the divergence of vibration signals and lead to the distribution discrepancy in the feature space. In visual recognition, different environments, lighting, background, resolutions, and view angles are potential factors that may affect the distribution of image data [18], [19].

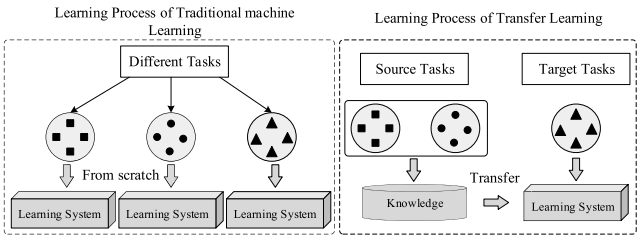


FIGURE 3. Different learning processes between traditional machine learning and transfer learning [17].

Usually, in order to guarantee the performance of traditional learning methods in new but similar tasks, a large amount of labeled samples under the target tasks are required for retraining the corresponding models. However, labeling a large number of target samples for any new tasks is labor-intensive and unrealistic for actual applications. Meanwhile, traditional machine learning approaches tend to break down when trained by the data from different conditions than that for test. Hence, developing learning algorithms that can construct robust models for current tasks by leveraging knowledge from other related datasets with sufficiently labeled samples but different distributions is an important and compelling problem.

Transfer Learning is a promising method to address this kind of cross-domain learning problems in which the distributions of training dataset and test dataset are allowed to be different. Transfer learning aims to leverage knowledge from one or multiple related datasets, which are called source domains, to improve the model's performance in the current dataset, which is called target domain. It is inspired by the capabilities of human beings that reusing the knowledge from some previous tasks without learning a new task from scratch. The learning processes of transfer learning and traditional machine learning are illustrated and compared in Fig.3 [17].

C. TRANSFER LEARNING METHODS

In the fields such as computer vision and natural language understanding, transfer learning has been a widely discussed topic in recent years. Several reviews about transfer learning and domain adaptation can be referred to in [17]–[22]. In general, transfer learning methods are divided into several categories according to the criterion of "what to transfer" [17]. Besides, deep learning-based and adversarial-based transfer methods are progressively investigated most recently, due to the powerful representation learning and end-to-end training capability [19]. Therefore, the following several transfer strategies are introduced briefly to help readers understand transfer learning.

1) INSTANCE REWEIGHTING APPROACH

Instance reweighting methods can be used to address the domain-shift problem in which estimated weights are incorporated into a loss function in an attempt to make the

weighted training distribution approximate the testing distribution. Actually, the goal of transfer learning is to learn a function $f_t(\cdot)$ that predicts the class label of test samples from the target domain. In general, the optimal parameters θ^* of $f_t(\cdot)$ is learned by minimizing the expected risk

$$\theta^* = \arg\min_{\theta\in\Theta} \sum_{(\boldsymbol{x},y)\in\mathcal{X}\times\mathcal{Y}} \mathcal{P}_t(X)\,\ell(\boldsymbol{x},y,\theta) \tag{1}$$

where $\ell(\boldsymbol{x},y,\theta)$ is a loss function, $\theta \in \Theta$ is a model parameter family from which we want to select an optimal parameter θ^*. In transfer learning, the training instances $X = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ are randomly sampled from the distribution of source domain $\mathcal{P}_s(X)$. Then we get

$$\begin{aligned}\theta_t^* &= \arg\min_{\theta\in\Theta} \sum_{(\boldsymbol{x},y)\in\mathcal{X}\times\mathcal{Y}} \frac{\mathcal{P}_t(X)}{\mathcal{P}_s(X)}\mathcal{P}_s(X) l(\boldsymbol{x},y,\theta) \\ &\approx \arg\min_{\theta\in\Theta} \sum_{(\boldsymbol{x},y)\in\mathcal{X}\times\mathcal{Y}} \frac{\mathcal{P}_t(X)}{\mathcal{P}_s(X)}\tilde{\mathcal{P}}_s(X) l(\boldsymbol{x},y,\theta)\end{aligned} \tag{2}$$

where $\tilde{\mathcal{P}}_s(X)$ is the empirical estimation of $\mathcal{P}_s(X)$. As we can see from (2), the solution to the transfer learning problem can be achieved by weighting the loss of the source domain samples by $\mathcal{P}_t(X)/\mathcal{P}_s(X)$.

It means that the model trained using the source domain data can be generalized to the target domain by estimating a weight $\mathcal{P}_t(X)/\mathcal{P}_s(X)$ for each training sample. There are many existing strategies designed for learning the weights [25]–[29]. A popular method, called TrAdaBoost, proposed by Dai *et al.* [27] has been applied to fault diagnosis field. TrAdaBoost attempted to iteratively reweight the source domain data under an ensemble learning architecture, AdaBoost. During each round of iteration, TrAdaBoost reweighted the source domain samples to reduce the effect of the "bad" source samples while encourage the "good" source samples to contribute more for the target domain based on the error computed on the target domain data. Sample re-weighting based domain adaptation methods mainly focus on the case where the difference between the source domain and the target domain is not too large [19].

2) FEATURE TRANSFER APPROACH

Another intuitive idea of transfer learning is to learn a new feature representation space, in which the source domain and target domain look "similar" and can be compared. The latent assumption under this kind of transfer methods is that a common subspace or higher-level representation exists for encoding the common characteristics between domains. In the new space supported by the domain-invariant features, the classifier trained using the labeled data from the source domain can be generalized to the target domain. Using different transferring criteria, the specific transfer strategies can be categorized into: (1) feature-transformation [30]–[32], (2) subspace-based [33]–[35], (3) sparse coding-based [36], [37], and (4) low-rank representation-based [38], [39].

Transfer Component Analysis (TCA), proposed by Pan *et al.* [30], is a representative feature transfer approach and has been successfully applied to fault diagnosis problems. The learning objective of TCA is to find a domain-invariant feature space in which the marginal distribution distance between two domains $\mathcal{P}_s(X_s)$ and $\mathcal{P}_t(X_t)$ is minimized. The distribution distance is measured using the Maximum Mean Discrepancy (MMD) criterion, as shown in (3) [40], and the objective function of TCA is defined by (4)

$$\mathrm{MMD}^2(X_s, X_t) = \left\| \frac{1}{n_s}\sum_{\boldsymbol{x}_i\in X_s}\phi(\boldsymbol{x}_i) - \frac{1}{n_t}\sum_{\boldsymbol{x}_j\in X_t}\phi(\boldsymbol{x}_j)\right\|_{\mathcal{H}}^2 \tag{3}$$

$$\min_{\mathbf{W}} \mathrm{tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{K}\mathbf{M}\mathbf{K}\mathbf{W}\right) + \mu\,\mathrm{tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{W}\right) \tag{4}$$

where $\phi(\boldsymbol{x})$ is the feature mapping from original space to Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$, $\mathbf{K} \in \mathbb{R}^{(n_s+n_t)\times(n_s+n_t)}$ is the kernel matrix with $K_{i,j} = \langle\phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j)\rangle$, $\mathbf{W}$ is the transformation matrix from RKHS to the new d-dimensional space. $\mathbf{M}$ is calculated using the following (5)

$$M_{i,j} = \begin{cases} 1/n_s^2 & \boldsymbol{x}_i, \boldsymbol{x}_j \in X_s \\ 1/n_t^2 & \boldsymbol{x}_i, \boldsymbol{x}_j \in X_t \\ -1/n_s n_t & \boldsymbol{x}_i \in X_t \wedge \boldsymbol{x}_j \in X_s \ or \\ & \boldsymbol{x}_i \in X_s \wedge \boldsymbol{x}_j \in X_t \end{cases} \tag{5}$$

3) CLASSIFIER ADAPTATION APPROACH

Instead of learning a domain-invariant feature space before constructing the classifier, classifier adaptation approaches aim to directly design an adaptive classifier for transfer learning tasks. It is also an effective strategy to handle the fundamental problem of mismatched distributions between the training and test datasets. According to reference [19], typical classifier adaptation approaches can be divided into: (1) kernel classifier-based [41]–[43], (2) manifold regularizer-based [44], and (3) Bayesian classifier-based [45], [46].

Adaptive Support Vector Machines (A-SVM), proposed by Yang *et al.* [41] for visual concept classification, is an intuitive and typical approach to understand this kind of transfer strategies. A-SVM aims to adapt the source domain classifier $f_s(\boldsymbol{x})$ trained on the labeled source data $\left\{\left(\boldsymbol{x}_i^s, y_i^s\right)\right\}_{i=1}^{n_s}$ to a new classifier $f_t(\boldsymbol{x})$ for the target task. This process is implemented by adding a bias term in the form of $\Delta f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x})$ on the basis of $f_s(\boldsymbol{x})$

$$f_t(\boldsymbol{x}) = f_s(\boldsymbol{x}) + \Delta f(\boldsymbol{x}) = f_s(\boldsymbol{x}) + \boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}) \tag{6}$$

where $\phi(\boldsymbol{x})$ is a feature map that projects $\boldsymbol{x}$ into a high-dimensional feature space. $\boldsymbol{w}$ is leaned using the labeled data of target domain $X_t^l = \left\{\left(\boldsymbol{x}_i^{tl}, y_i^{tl}\right)\right\}_{i=1}^{n_t^l}$, and the following optimization problem is solved

$$\begin{aligned}&\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C_{\mathrm{svm}}\sum\nolimits_{i=1}^{n_t^l}\xi_i \\ &\text{s.t. } \xi_i \geq 0 \\ &\quad y_i f_s(\boldsymbol{x}_i) + y_i\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i) \geq 1 - \xi_i, \\ &\quad \forall(\boldsymbol{x}_i, y_i) \in X_t^l\end{aligned} \tag{7}$$

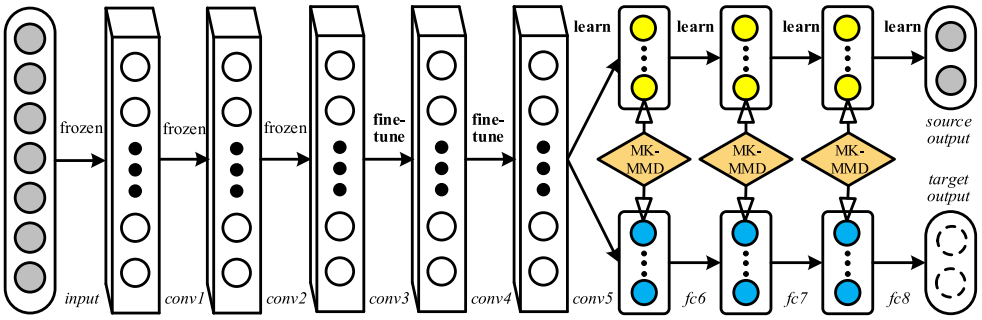


FIGURE 4. Architecture of Deep Adaptation Network (DAN) proposed by Long *et al.* for learning transferable features [50]. In the figure, *input* denotes the input layer, *conv* denotes convolutional layer, and *fc* denotes fully connected layer.

where ξ_i is the penalizing variable and C_{svm} is the regularization parameter. The above objective function seeks a new decision boundary that is close to the boundary of the source classifier and meanwhile separates the labeled samples of the target domain correctly.

4) DEEP LEARNING-BASED APPROACH

In recent years, deep learning has had tremendous success in achieving state-of-the-art performance in speech recognition, visual object recognition, drug discovery [47], and even machinery fault diagnosis. As end-to-end systems, deep neural networks learn representations of raw data with multiple levels of abstraction by multiple processing layers. The transfer learning methods based on deep neural networks aim to learn more transferable representations by embedding domain adaptation into the pipeline of deep learning [21].

Loosely speaking, deep learning-based transfer approaches can be divided into two categories: (1) parameter transfer [48], [49], (2) representation adaptation [50]–[53]. Parameter transfer is a commonly used strategy for training deep models under cross-domain scenarios in various applications. The intuitive idea of parameter transfer is to fine-tune a pre-trained deep neural network (model for the source domain) using a small amount of target data. Generally, the pre-trained deep neural network is trained on a source domain with massive amounts of labeled data.

The intuitive idea of representation adaptation is to embed representation adaptation goal into the process of deep learning. Usually, in order to learn representations that are both discriminative for faults and domain-invariant, a trade-off term that penalizes the representation distribution discrepancy between domains is added into the objective function of the deep neural network. With the domain-invariant representation, the generalization performance of the deep model on the target domain would be promoted.

To assist the readers understand the deep representation adaptation strategy, the Deep Adaptation Network (DAN) architecture, which is proposed by Long *et al.* [50] is introduced here. DAN is based on CNN, and its overview architecture is shown in Fig.4. In the objective function of DAN as defined by (8), a multiple kernel MMD (MK-MMD)-based adaptation regularization term is added to the CNN risk to approximate the distributions of the source and target domains under the hidden representations of the last three fully-connected layers ($l_1 \rightarrow l_3$). Finally, DAN enhanced the transferability of features from task-specific layers of the CNN.

$$\min_{\Theta} \frac{1}{n_a} \sum_{i=1}^{n_a} \ell\left(\theta\left(\boldsymbol{x}_i^a\right), y_i^a\right) + \lambda \sum_{l=l_1}^{l_3} \text{MMD}_k^2\left(\Phi_s^l, \Phi_t^l\right) \quad (8)$$

where the first term is the classification cost of CNN, and the second term is the MK-MMD adaptation term. n_a denotes the number of labeled samples of the source and target domains. $\lambda > 0$ is a penalty parameter of MK-MMD term.

5) ADVERSARIAL-BASED APPROACH

Recently, Generative Adversarial Networks (GAN) [54] has achieved great success in for generating feature-level representations by training robust deep neural networks via an adversarial learning process. GAN consists of two models: a generative model G that extracts the data distribution and a discriminative model D that distinguishes whether a sample is from G or training datasets by predicting a binary label. D is trained to maximize the probability of assigning the correct label to both training examples and samples generated by G. While G is trained to minimize $\log(1 - D(G(z)))$ simultaneously.

$$\min_G \max_D \mathcal{L}(D, G) = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{P}_{data}}\left[\log D(x)\right] + \mathbb{E}_{\mathbf{z} \sim \mathcal{P}_z(z)}\left[\log(1 - D(G(z)))\right] \quad (9)$$

where $\mathcal{P}_z(z)$ denotes the prior on input noise distribution, $\mathcal{P}_{data}$ denotes the data distribution.

Inspired by the adversarial learning process, adversarial-based transfer learning approaches have been widely researched as an increasing popular idea. According to different strategies, adversarial-based approaches can be divided into two categories. The first one is the generative-based strategy [55]–[57], and the core idea is to generate synthetic target data with ground-truth annotations with the help of source data and then enable the cross-domain tasks by using

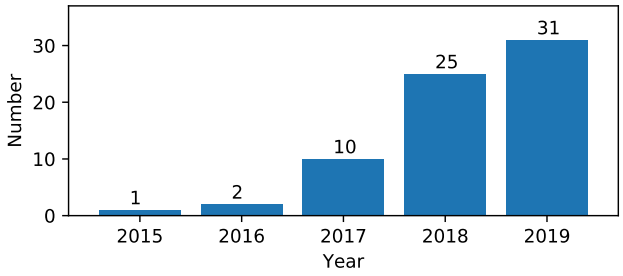


FIGURE 5. Literature statistics of cross-domain fault diagnosis articles according to published year.

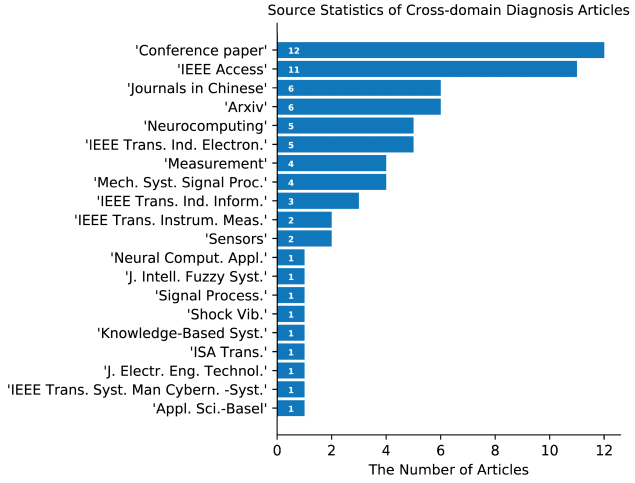


FIGURE 6. Source statistics of cross-domain fault diagnosis articles.

synthesized target data. The second one is the adversarial adaptation-based strategy [58], [59], which aims to adapt the representation distributions of the source and target domains through employing a domain discriminator. During the adversarial learning process, the similarity of the representations learned by deep neural networks would be ensured when the domain discriminator cannot distinguish between the source and target domains.

III. CROSS-DOMAIN FAULT DIAGNOSIS RESEARCH WORKS

A. LITERATURE STATISTIC

Transfer learning, as a promising approach to handle domain-shift issue, has been widely applied in cross-domain fault diagnosis during the last few years. To systematically introduce the existing articles of this topic, the scope of the investigation in this paper covers the period between 2015 and 2019. Literature statistics about cross-domain fault diagnosis according to years are shown in Fig.5. From the figure, it is found that this topic becomes popular in the last two years. These articles consist of journal articles, conference papers, and articles in preprint, all of which are directly related to cross-domain fault diagnosis using data-driven approaches. We collected these research articles using electronic databases including: Web of Science, IEEE Xplorer, Science Direct, arXiv.org, and CNKI. The primary retrieval keywords were "fault diagnosis", "transfer learning", and "domain adaptation".

There are 69 articles in total including 51 journal articles, 12 conference papers, and 6 articles in preprint. It should be noted that 6 journal articles in Chinese are contained. Fig.6 shows the detailed source statistics of these articles. Through the investigation, the authors found that there are no review papers on cross-domain fault diagnosis. This provides a straightforward motivation for the authors to introduce the related publications of this topic.

In the following sections, we summarize these related research works from three viewpoints: (1) research motivations, (2) cross-domain diagnosis approaches (most of them are based on transfer learning), (3) application objects. The research motivations discussed here mean the different scenarios in which the cross-domain diagnosis tasks would be considered, such as the diagnosis tasks under variations of operating condition or fault degree, leveraging knowledge from different but related machines etc. Through this aspect, the common cross-domain scenarios in fault diagnosis field are summarized. In the second aspect, those research works would be reviewed according to different cross-domain strategies. The corresponding methods are divided into traditional transfer approaches, deep transfer approaches, adversarial-based approaches, and other strategies. Finally, the main application subjects of those research works are introduced. It should be noted that some research works that do not use transfer methods but consider the cross-domain diagnosis tasks are also included in this review.

Based on the definitions in Section II.A, we use $\mathcal{D}_s$ and $\mathcal{D}_t$ to denote source domain and target domain, respectively. Let $X_s = \{(\boldsymbol{x}_i^s, y_i^s)\}|_{i=1}^{n_s}$ and $X_t = \{(\boldsymbol{x}_i^t, y_i^t)\}|_{i=1}^{n_t}$ be the datasets sampled from the source domain and the target domain, where n_s and n_t denote the number of samples correspondingly. $y_i^s \in \mathcal{Y}_s$ is the label of $\boldsymbol{x}_i^s$, where $\mathcal{Y}_s$ denotes the label space of the source task. Similarly, $y_i^t \in \mathcal{Y}_t$ denotes the label of $\boldsymbol{x}_i^t$, and $\mathcal{Y}_t$ denotes the label space of the target task. In fault diagnosis problem, there may be C different health conditions, and we denote each of the health condition using $1, \cdots, \mathrm{C}$. Let C_s and C_t be the health condition numbers of the source task and the target task respectively, then $\mathcal{Y}_s = \{1, \cdots, \mathrm{C}_s\}$ and $\mathcal{Y}_t = \{1, \cdots, \mathrm{C}_t\}$. Without loss of generality, let $y_i^s = 1$ and $y_i^t = 1$ denote the normal condition in the following discussions.

B. RESEARCH MOTIVATIONS AND PROBLEM SETTINGS

As discussed above, the main motivation of transfer learning is borrowing knowledge from source domains to facilitate learning in a target domain. From this perspective, the following cross-domain diagnosis research works to be discussed only leverage the idea of knowledge transfer to fault diagnosis which is usually treated as a pattern recognition problem. The main difference between cross-domain fault

diagnosis and the cross-domain tasks in computer vision and natural language processing *etc.* is that the distribution discrepancy between domains is caused by variations of operating condition (load and rotating speed), working environment, fault degree, or data source (simulation model and different machines). According to different research motivations, that is to say different domain-shift scenarios, these related works are summarized into five items in Section III.B.1).

1) RESEARCH MOTIVATIONS

a: MOTIVATION 1: ADDRESSING CROSS-DOMAIN FAULT DIAGNOSIS BETWEEN DIFFERENT OPERATING CONDITIONS

Distribution discrepancy between training dataset and test dataset is the essential cause that affects the generalization performance of data-driven fault diagnosis methods. The variation of the operating condition of machinery system is a major factor that may lead to the distribution discrepancy between datasets. Therefore, it is the first motivation for employing transfer learning to construct effective fault identification models for the current operating condition using the historical data collected from other operating conditions of the same machine.

In the diagnosis tasks of rotating machine, such as rotor system, bearing, and gearbox, vibration signals are most widely used to infer machines' health conditions, because rich fault information can be easily measured with low-cost vibration sensors. The variations of rotating speed will influence the vibration frequency and amplitude of measured signals, and further influence the probability distribution of the data in the feature space. Furthermore, the working load is changing according to actual requirements in real-world industrial applications. It is very significant to address the performance degradation of fault diagnosis methods under such scenarios. For example, in [60]–[78], several transfer strategies were proposed to improve the performance of diagnosis models in target operating condition through reusing the data from different rotating speeds and loads.

The noise of the working environment is inevitable and unpredictable in industrial production, which may also alter the distribution state of data in feature space. Aiming at this issue, [73] and [74] proposed two different deep learning based methods which have good anti-noise and domain adaptation abilities.

In [79], Wang *et al.* presented a deep transfer network to tackle the diagnosis problem of power equipment under different environment temperatures. In addition, in many research works [80]–[89], the diagnosis tasks between different operating conditions were only employed to simulate the transfer scenarios for verifying the effectiveness of their methods. But the ultimate objectives of those research works were to address general cross-domain fault diagnosis problems and were not limited to the discrepancy of operating conditions.

b: MOTIVATION 2: ADDRESSING CROSS-DOMAIN FAULT DIAGNOSIS BETWEEN DIFFERENT FAULT DEGREES

Except for the motivation under different operating conditions, there are some research works in which the diagnosis tasks between different fault degrees were considered [85], [90]–[93]. Zhang *et al.* [90] validated their proposed bearing diagnosis method using data with different fault diameters and data with different fault diameters while different loads. References [85], [92], [93] also studied the performances of their methods on the diagnosis tasks between different fault degrees. Besides, diagnosis of incipient fault is a very important and difficult issue. Usually, very limited incipient fault data are available for training robust diagnosis model, especially deep learning based model. Chen *et al.* proposed a parameter transfer learning method based on deep auto-encoder in [91]. The proposed method can facilitate incipient fault diagnosis using fault samples with significant fault characteristics.

c: MOTIVATION 3: PROMOTING DIAGNOSIS PERFORMANCE BY LEVERAGING KNOWLEDGE FROM DIFFERENT BUT RELATED MACHINES OR SIMULATION MODELS

In the above-mentioned research works, although the source domain dataset for training models and the target domain dataset to be identified come from different operating conditions, working environments, or fault degrees, the machines or systems from which the monitoring signals are collected are the same ones. However, this diagnosis scenario may be laborious to implement in practical applications, because the probability of occurring all fault modes is low during the past use of the machine to be diagnosed and usually a limited number of fault samples can be collected for training the diagnosis model.

Usually, collecting or generating historical fault data from other same-type machines, the simulation experiments in the laboratory, or mathematical simulation models are more feasible and easier means. From the aspect of feasibility, these data also contain the inherent fault information of this type of machine or system, and using these data to train diagnosis model is more consistent with engineering requirements for data-driven fault diagnosis.

Under this motivation, several research works organized the cross-domain diagnosis tasks that the training dataset and the test dataset were acquired from different machines to verify the corresponding transfer diagnosis methods [94]–[96]. In [94], Guo *et al.* proposed an intelligent method, Deep Convolutional Transfer Learning Network (DCTLN), for machinery fault diagnosis. Six cross-domain tasks based on three different datasets, Case Western Reserve University (CWRU) bearing dataset, Intelligent Maintenance System (IMS) bearing dataset, and Railway Locomotive (RL) bearing dataset, were used to verify the effectiveness of the proposed method. In [95], Zhang *et al.* proposed a supervised dictionary-based transfer subspace learning method to diagnose suck rod pumping systems by using the monitoring data from different wells. Zheng *et al.* proposed a fault diagnosis

method through considering multiple source domains in [96], the proposed method built the diagnosis model using the data of one bearing and diagnosed another bearing of different models.

Besides, there are several research works that used fault simulation data from the laboratory (or artificial fault) to facilitate diagnosing the fault modes of real machines (or natural fault) [92], [97]–[100]. The intuitive motivation of these research works is that simulating the fault modes in the laboratory is easier than collecting fault data in actual engineering. In [97] and [98], Yang *et al.* proposed two deep transfer approaches under the framework of CNN to diagnose the real locomotive bearings by leveraging the diagnosis knowledge from the bearing simulation data of the laboratory. A similar diagnosis case can be found in [100] for rotor system fault diagnosis. Also focusing on bearing fault diagnosis, Kim andYoun [92] evaluated their diagnosis method based on deep parameter transfer under the tasks from artificial fault data to actual damage data collected in life test.

In addition, in actual applications, the availability of the historical fault data generated by physical machines may be pretty limited. The simulation models, which can describe inherent behaviors and rules of physical machines or systems, can generate massive data under different health conditions and even different operating conditions. Leveraging knowledge from the virtual simulation data that can also provide insight into the fault characteristics is also one motivation for using transfer learning in fault diagnosis application. Xu *et al.* [101] presented a digital-twin-assisted fault diagnosis method combined with deep transfer learning. The fault data of source domain are generated by a digital shop floor, established by Process Designer & Process Simulate, while the target domain is the corresponding physical shop floor. Sobie *et al.* [102] proposed a simulation-driven intelligent diagnosis method for race fault classification of bearing. In their method, a one-dimensional 3-DOF (degree of freedom) dynamic model of bearing, which was implemented using Siemens LMS Imagine, Lab Amesim simulation software, was employed to create fault simulation data. After the preprocessing steps, the classification models, CNN and nearest-neighbor dynamic time warping, were trained using generated simulation data, and then were used to diagnose seeded-fault experiment data (CWRU data, SpectraQuest data, and Society for Machinery Failure Prevention Technology data) and industrial wind turbine bearing data.

d: MOTIVATION 4: USING VISUAL IMAGES TO FACILITATE FAULT DIAGNOSIS TASKS

In general, although deep neural networks can extract high-quality characteristics by multi-layer transformation, training deep learning models for fault diagnosis requires a large number of labeled samples and considerable computational resources. However, the limited fault data could not support the training of robust multi-layer neural networks. Motivated by the success of deep learning in computer vision, several research works transferred some parameters of the deep networks pre-trained using ImageNet dataset (a widely used image dataset in computer vision, available at http://www.image-net.org/) to accelerate the training and promote the accuracy of the networks for machinery fault diagnosis [103]–[106].

In [103], Cao *et al.* presented a deep convolutional neural network-based transfer learning approach for gearbox fault diagnosis. The first 21 layers of a pre-trained deep network using massive image data (1.2 million) from the ImageNet dataset were transferred to a new network for fault diagnosis. The new network, which consists of the same first 21 layers with pre-trained network and three newly added layers, was fine-tuned utilizing limited data specific to gearbox fault diagnosis tasks. Similarly, Shao *et al.* developed a deep learning framework to achieve highly-accurate machine fault diagnosis in [104]. The pre-trained network, VGG-16, was also trained using the ImageNet dataset. However, in fine-tune stage, both the last two convolution blocks of the pre-trained network and the newly added fully-connected layers were trainable. In addition, [105] and [106] also presented two similar deep transfer methods for machinery fault diagnosis. The pre-trained networks were also trained using the ImageNet dataset, but the structures of the deep networks were different from those in [103] and [104].

e: MOTIVATION 5: LEVERAGING KNOWLEDGE FROM THE SOURCE WITH INCOMPLETE INFORMATION

There may be such diagnosis scenarios that massive training data with incomplete information are available. Several research works discussed how to utilize these incomplete samples of source domain to facilitate the target diagnosis task through transfer learning methods [90]–[93], [107], [108].

In [90], a deep CNN was trained firstly using the data of source domain which consists of the samples of three different health conditions, then a new CNN was fine-tuned with small amount of target data which consists of the samples of five classes after transferring partial parameters from the pre-trained CNN.

Zhong *et al.* presented a feature mapping method to extract the feature representations for fault dataset by reusing the internal layers of CNN trained on the normal dataset [108]. This work showed how feature representations learned by CNN on large-scale annotated gas turbine normal dataset can be efficiently transferred to fault diagnosis task with limited fault data.

Besides, reference [107] proposed a fault diagnosis framework that uses structurally incomplete samples to facilitate the model training of the target domain. They declared that a large number of incomplete samples also contain useful information, and transferring them to target diagnosis task is helpful.

2) DIFFERENT PROBLEM SETTINGS

With respect to the above-mentioned research works, the settings of the diagnosis problem can be mainly divided into

TABLE 1. The Taxonomy of research works about cross-domain fault diagnosis according to different research motivations and problem settings.

Motivations	Problem settings	Descriptions of the training dataset	References
Motivation 1	Setting 1	Source domain: single source, fully labeled samples Target domain: limited labeled samples	[66] [67] [68] [69] [70] [80] [81] [83] [87] [109] [110] [111]
	Setting 2	Source domain: single source, fully labeled samples Target domain: Unlabeled samples, but contains the samples of each category	[60] [61] [62] [63] [64] [65] [66] [71] [72] [75] [76] [77] [78] [79] [84] [85] [86] [88] [112] [113]
	Setting 3	Source domain: single source, fully labeled samples Target domain: only labeled normal samples are available	[82] [89] [114]
	Setting 4	Source domain: single source, fully labeled samples Target domain: no target domain samples are used	[73] [74] [116] [117] [118] [119] [120] [121] [122]
Motivation 2	Setting 1	Source domain: single source, fully labeled samples Target domain: limited labeled samples	[90] [91] [92] [93]
	Setting 2	Source domain: single source, fully labeled samples Target domain: Unlabeled samples, but contains the samples of each category	[85]
Motivation 3	Setting 1	Source domain: single source, fully labeled samples Target domain: limited labeled samples	[101] [100] [92]
	Setting 2	Source domain: single source, fully labeled samples Target domain: Unlabeled samples, but contains the samples of each category	[94] [97] [98]
	Setting 3	Source domain: single source, fully labeled samples Target domain: only labeled normal samples are available	[115]
	Setting 4	Source domain: single source, fully labeled samples Target domain: no target domain samples are used	[99] [102]
Motivation 4	——	Source domain: **fully labeled visual images** Target domain: limited labeled samples for fault diagnosis task	[103] [104] [105] [106]
Motivation 5	Setting (a)	Source domain: single source, fully labeled samples **with less categories** Target domain: limited labeled samples	[90] [91] [92] [93] [108]
	——	Source domain: large number of incomplete samples Target domain: limited number of complete samples	[107]

Note: Setting (a): $X_s = \left\{\left(\boldsymbol{x}_i^s, y_i^s\right)\right\}\Big|_{i=1}^{n_s}$; $X_t = \left\{\left(\boldsymbol{x}_i^t, y_i^t\right)\right\}\Big|_{i=1}^{n_t}$; $\mathrm{C}_s < \mathrm{C}_t$

four categories. Although the motivations may be different, the essential idea of these research works is leveraging diagnosis knowledge from a related task to help the learning of current diagnosis task. Usually, in the training stage of diagnosis models, only one source domain with massive fully labeled samples is always available, but the cases of available samples from the target domain are different. Specifically, there are the following four main cases:

a: PROBLEM SETTING 1: $X_S = \left\{\left(\boldsymbol{x}_i^s, y_i^s\right)\right\}\Big|_{i=1}^{n_s}$, $y_i^s \in \{1, \cdots, C_s\}$; $X_t = \left\{\left(\boldsymbol{x}_i^t, y_i^t\right)\right\}\Big|_{i=1}^{n_t}$, $y_i^t \in \{1, \cdots, C_t\}$; *and* $n_t < n_s$; $C_s = C_t$

In the training stage, the dataset from the source domain is fully labeled and contains the samples of C_s classes, the dataset from the target domain is also labeled and contains the samples of C_t classes. But the samples of the target domain for training are very limit, that is $n_t < n_s$, where n_s and n_t are the sample size of X_s and X_t respectively. Besides, the numbers of health conditions in the source domain and the target domain are the same, that is $C_s = C_t$. References [66]–[70], [80], [81], [83], [87], [90], [100], [101], [109]–[111] follow this setting.

b: PROBLEM SETTING 2: $X_S = \left\{\left(\boldsymbol{x}_i^s, y_i^s\right)\right\}\Big|_{i=1}^{n_s}$, $y_i^s \in \{1, \cdots, C_s\}$; $X_t = \left\{\boldsymbol{x}_i^t\right\}\Big|_{i=1}^{n_t}$; *and* $C_s = C_t$

Similarly, in this setting, massive labeled samples with C_s classes from the source domain are also available in training stage. But the samples from target domain are unlabeled. The cross-domain learning tasks under this setting are usually called unsupervised domain adaptation [20]. References [60]–[66], [71]–[73], [75]–[79], [84]–[86], [88], [94], [97], [98], [112], [113] follow this setting.

c: PROBLEM SETTING 3: $X_S = \left\{\left(\boldsymbol{x}_i^s, y_i^s\right)\right\}\Big|_{i=1}^{n_s}$, $y_i^s \in \{1, \cdots, C_s\}$; $X_t = \left\{\left(\boldsymbol{x}_i^t, y_i^t = 1\right)\right\}\Big|_{i=1}^{n_t}$; *and* $C_s = C_t$.

In fault diagnosis field, usually only massive labeled samples under normal condition can be collected for target machine before building diagnosis models. That is to say, in training stage, apart from massive labeled samples from the source domain, only normal samples of the target domain are available. Several research works discussed the cross-domain diagnosis problem under this setting [82], [89], [96], [114], [115].

d: PROBLEM SETTING 4: $X_S = \left\{\left(\boldsymbol{x}_i^s, y_i^s\right)\right\}\Big|_{i=1}^{n_s}$, $y_i^s \in \{1, \cdots, C_s\}$

There are also several research works that only use the labeled samples from the source domain in models' training stage [74], [99], [102], [116]–[122]. There is no available data to describe the distribution information of the target domain, so the diagnosis tasks with the relative smaller discrepancy between domains were usually tackled under this setting, such as the diagnosis tasks among different operating conditions [74], [116]–[122].

According to different motivations and problem settings, the taxonomy of related research works is summarized in Table 1.

TABLE 2. Summary of the references that use traditional transfer learning approaches.

Categories	Sub-categories	References
Instance reweighting	——	[81] [83]
Feature-based	Feature transformation	[60] [61] [62] [64] [67] [71] [72] [77] [123]
	Subspace based	[75] [95]
	Manifold alignment	[111]
Classifier adaptation	——	[80]

C. CROSS-DOMAIN APPROACHES APPLIED IN FAULT DIAGNOSIS

In this section, we mainly introduce the specific approaches applied in cross-domain fault diagnosis by dividing them into four categories: (1) traditional transfer approaches, (2) deep transfer approaches, (3) adversarial-based approaches, and (4) other approaches. Section III.C.1) summarizes the transfer approaches with shallow structures. Section III.C.2) introduces the specific transfer approaches based on deep neural networks, such as CNN and AE. The specific approaches based on adversarial strategy will be depicted in Section III.C.3).The detailed strategies of several research works that do not use transfer learning during cross-domain fault diagnosis are introduced in Section III.C.4) as well. Focusing on the inputs of cross-domain diagnosis approaches, we provide a summary and discussion about different types of input data in Section III.C.5).

1) TRADITIONAL TRANSFER APPROACHES

The research works which applied traditional transfer approaches during cross-domain fault diagnosis are summarized in Table 2. They are categorized into: 1) instance reweighting based approaches, 2) feature based approaches, and 3) classifier adaptation approaches.

In [81], Shen *et al.* proposed a Singular Value Decomposition (SVD) + TrAdaboost based bearing fault diagnosis method. In their method, the eigenvalue vector of autocorrelation matrix of vibration signal was extracted using SVD as diagnosis features, the instance based transfer algorithm, TrAdaboost proposed by Dai *et al.* [27], was applied to reweight the source domain data from different operating conditions. This approach was also used to the fault diagnosis of induction motor in [83].

Feature-based transfer learning methods were widely employed and developed in cross-domain fault diagnosis. Several research works intended to learn a new space through feature transformation, in which the distribution discrepancy between the source and target domains was reduced. TCA, which is proposed by Pan *et al.* [30] has been employed to address cross-domain fault diagnosis of gear by Xie *et al.* [60], [61], rolling element bearing by Chen *et al.* [72], and delta 3D printer by Guo *et al.* [123]. Similarly, Kang *et al.* [67] utilized the Semi-supervised TCA (SSTCA) to diagnose bearing fault under variations of operating conditions. A multi-kernel kernel function is constructed for SSTCA by combining Polynomial kernel K_{poly} and Radial Basis Function (RBF) kernel K_{rbf}, that is $K_{i,j} = \alpha K_{\text{poly}} + (1-\alpha) K_{\text{rbf}}$ where $0 \leq \alpha \leq 1$ is a multi-kernel coefficient.

In [77], Tong *et al.* presented a bearing fault diagnosis method called domain adaptation using transferable features (DATF). DATF aimed to learn a feature space by $\mathbf{W}^{\mathrm{T}}\mathbf{X}_{\mathrm{D}}$ in which the marginal distribution and conditional distribution between domains were simultaneously minimized. $\mathbf{X}_{\mathrm{D}} = [\mathbf{X}_s, \mathbf{X}_t] \in \mathbb{R}^{\mathrm{D}\times(n_s+n_t)}$ is the data matrix of training dataset with $\mathbf{X}_s$ is labeled and $\mathbf{X}_t$ is unlabeled. The cost function of DAFD was

$$\begin{aligned} \arg\min_{\mathbf{W}} \ & (1-\beta)\left[\operatorname{tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{X}_{\mathrm{D}}\mathbf{M}_0\mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{W}\right)\right. \\ & \left. + \sum_{c=1}^{\mathrm{C}} \operatorname{tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{X}_{\mathrm{D}}\mathbf{M}_c\mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{W}\right)\right] \\ & + \beta \|\mathbf{W}\|_F^2 \\ s.t. \ & \mathbf{W}^{\mathrm{T}}\mathbf{X}_{\mathrm{D}}\mathbf{H}\mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{W} = \mathbf{I} \end{aligned} \tag{10}$$

$\operatorname{tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{X}_{\mathrm{D}}\mathbf{M}_0\mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{W}\right)$ is the marginal distribution distance between the source domain and the target domain computed using MMD, and $\sum_{c=1}^{\mathrm{C}} \operatorname{tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{X}_{\mathrm{D}}\mathbf{M}_c\mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{W}\right)$ is the corresponding conditional distribution distance, where $\mathbf{M}_0$, $\mathbf{M}_c$ were calculated according to

$$(\mathbf{M}_0)_{i,j} = \begin{cases} \dfrac{1}{n_s n_s} & \boldsymbol{x}_i, \boldsymbol{x}_j \in X_s \\ \dfrac{1}{n_t n_t} & \boldsymbol{x}_i, \boldsymbol{x}_j \in X_t \\ -\dfrac{1}{n_s n_t} & \text{otherwise} \end{cases};$$

$$(\mathbf{M}_c)_{i,j} = \begin{cases} \dfrac{1}{n_s^c n_s^c} & \boldsymbol{x}_i, \boldsymbol{x}_j \in X_s^c \\ \dfrac{1}{n_t^c n_t^c} & \boldsymbol{x}_i, \boldsymbol{x}_j \in X_t^c \\ \dfrac{-1}{n_s^c n_t^c} & \begin{cases} x_i \in X_s^c, & x_j \in X_t^c \\ x_i \in X_t^c, & x_j \in X_s^c \end{cases} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

$\|\mathbf{W}\|_F^2$ is the regularization term, where $\|\cdot\|_F$ is the Frobenius norm of transformation matrix $\mathbf{W}$. β is the regularization parameter that trades off the model complexity. X_s^c and X_t^c denote the sample sets of c-th class of the source domain and the target domain, respectively. $n_s^c = \left|X_s^c\right|$ and $n_t^c = \left|X_t^c\right|$. For computing the conditional MMD, the pseudo labels of unlabeled samples of the target domain were obtained by a base classifier, nearest-neighbor classifier, iteratively.

Furtherly, Tong *et al.* presented a similar approach, called Domain Adaptation by using Feature Transfer Learning (DAFTL), in [62]. Apart from minimizing the marginal and conditional distribution distances between domains during learning the new representation $V = \mathbf{W}^{\mathrm{T}}\mathbf{X}_{\mathrm{D}}$, the local geometric structure of the data was maximally preserved through minimizing the within-class distance and maximizing the

between-class distance simultaneously. The objective function of DAFTL was defined as follows

$$\begin{aligned} &\arg\min_{\mathbf{W}} \sum\nolimits_{c=0}^{\mathrm{C}} \mathrm{tr}\left[\mathbf{W}^{\mathrm{T}}\mathbf{X}_{\mathrm{D}}\mathbf{M}_c\mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{W} + \boldsymbol{S}_w\right] + \lambda \left\|\mathbf{W}\right\|_F^2 \\ &s.t.\ \mathbf{W}^{\mathrm{T}}\boldsymbol{S}_b\mathbf{W} = \mathbf{I} \end{aligned} \tag{12}$$

where $\boldsymbol{S}_w = \sum_{\forall c\in C}\sum_{\boldsymbol{x}_i\in X_s^c}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_c)^{\mathrm{T}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_c)$ is the within-class scatter matrix, $\bar{\boldsymbol{x}}_c$ denotes the mean of the c-th samples. $\boldsymbol{S}_b = \sum n_s^c(\bar{\boldsymbol{x}}_c - \bar{\boldsymbol{x}}_0)^{\mathrm{T}}(\bar{\boldsymbol{x}}_c - \bar{\boldsymbol{x}}_0)$ is the between-class scatter matrix, and $\bar{\boldsymbol{x}}_0$ denotes the mean of all samples.

Besides, several research works proposed subspace based transfer learning approaches to address cross-domain fault diagnosis. In [75], an unsupervised domain adaptation approach based on subspace alignment was proposed for bearing fault diagnosis across different operating conditions by Zhang *et al.* First, the subspace of the source domain $\mathbf{Z}_s \in \mathbb{R}^{\mathrm{D}\times d}$ and the subspace of the target domain $\mathbf{Z}_t \in \mathbb{R}^{\mathrm{D}\times d}$ were generated using Principal Components Analysis (PCA). The basis vectors of $\mathbf{Z}_s$ and $\mathbf{Z}_t$ were d eigenvectors corresponding to the d largest eigenvalues of PCA transformation, and using the PCA transformation the data of source domain and target domain (D dimension) were projected to d dimension subspace. The proposed subspace alignment method aimed to align the two subspace by a transformation matrix $\mathbf{W}$. That is to solve the following optimization problem.

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} \left\|\mathbf{Z}_s\mathbf{W} - \mathbf{Z}_t\right\|_F^2 \tag{13}$$

The transformation matrix $\mathbf{W}$ defined a movement to make $\mathbf{Z}_s$ and $\mathbf{Z}_t$ close to each other, and then domain-shift was corrected. After obtaining $\mathbf{W}^*$, the labeled samples from the source domain were projected to the aligned subspace denoted by $\mathbf{Z}_a = \mathbf{Z}_s\mathbf{W}^*$. In the new subspace $\mathbf{Z}_a$, a classification model (SVM) was trained using labeled source domain samples, and the unlabeled target domain samples were predicted using the trained model.

In [95], Zhang *et al.* proposed a supervised dictionary-based transfer subspace learning method for fault diagnosis of sucker rod pumping systems. A transformation matrix was learned to transfer both the source domain data and target domain data into a common subspace, in which they can be represented by the same dictionary. Specifically, suppose a source domain data matrix $\mathbf{X}_s \in \mathbb{R}^{\mathrm{D}\times n_s}$ of C_s classes, a target domain data matrix $\mathbf{X}_t \in \mathbb{R}^{\mathrm{D}\times n_t}$ of C_t classes, and $\mathrm{C}_s > \mathrm{C}_t$. The proposed method assumed that there is a transformation matrix $\mathbf{W}$ to project the samples of the source domain and the target domain into a common subspace with lower dimension d. In the new subspace, the samples of the source domain and the target domain can be well reconstructed by a shared dictionary matrix Dic.

$$\begin{aligned} \mathbf{W}^{\mathrm{T}}\mathbf{X}_s &= \mathrm{Dic}\cdot\mathbf{R}_s + \mathbf{E}_s \\ \mathbf{W}^{\mathrm{T}}\mathbf{X}_t &= \mathrm{Dic}\cdot\mathbf{R}_t + \mathbf{E}_t \end{aligned} \tag{14}$$

where $d < \mathrm{D}$ and the dictionary $\mathrm{Dic} = \left[\mathrm{Dic}_1, \cdots, \mathrm{Dic}_{\mathrm{C}_s}\right]$ has C_s sub-dictionaries where Dic_c corresponding to class c. $\mathbf{R}_s$ and $\mathbf{R}_t$ are the representation coefficient matrices with respect to Dic, $\mathbf{E}_s$ and $\mathbf{E}_t$ are the noise matrices. The low-rank constraint was introduced which aims at enforcing $\mathbf{R}_s$ and $\mathbf{R}_t$ to have a block-wise structure. To further explore the data structure and utilize the label information, the proposed method introduced two ideal regularization terms $\mathbf{R}_s = \mathbf{Q}_s$; $\mathbf{R}_t = \mathbf{Q}_t$, where $\mathbf{Q}_s$ and $\mathbf{Q}_t$ are ideal representation coefficient matrices. $\mathbf{Q}_s = [q_1^s, \cdots, q_{n_s}^s] \in \mathbb{R}^{d\times n_s}$, $\mathbf{Q}_t = [q_1^t, \cdots, q_{n_t}^t] \in \mathbb{R}^{d\times n_t}$. If $\boldsymbol{x}_i^s$ belong to class c, then the coefficients in q_i^s for Dic_c are all **1**s, while the others are all **0**s. The final objective function was

$$\begin{aligned} &\min\ \left\|\mathbf{R}_s\right\|_* + \left\|\mathbf{R}_t\right\|_* + \upsilon\left(\left\|\mathbf{E}_s\right\|_1 + \left\|\mathbf{E}_t\right\|_1\right) \\ &s.t.\ \mathbf{W}^{\mathrm{T}}\mathbf{X}_s = \mathrm{Dic}\cdot\mathbf{R}_s + \mathbf{E}_s \\ &\quad \mathbf{W}^{\mathrm{T}}\mathbf{X}_t = \mathrm{Dic}\cdot\mathbf{R}_t + \mathbf{E}_t \\ &\quad \mathbf{R}_s = \mathbf{Q}_s \\ &\quad \mathbf{R}_t = \mathbf{Q}_t \end{aligned} \tag{15}$$

where $\left\|\mathbf{R}_s\right\|_*$ is the nuclear norm of $\mathbf{R}_s$, and $\left\|\cdot\right\|_1$ means ℓ_1-norm. υ dominates the sparsity of the noise matrices. After obtaining the optimal transform $\mathbf{W}$ and the shared dictionary matrix Dic, the representation coefficients $\mathbf{R}$ of the new test samples $\mathbf{X}$ can be calculated by solving the following optimization problem

$$\begin{aligned} &\min\ \left\|\mathbf{R}\right\|_* + \upsilon\left\|\mathbf{E}\right\|_1 \\ &s.t.\ \mathbf{W}^{\mathrm{T}}\mathbf{X} = \mathrm{Dic}\cdot\mathbf{R} + \mathbf{E} \end{aligned} \tag{16}$$

The label of $\mathbf{X}$ was predicted by $\mathbf{Y} = \mathbf{Y}_{\mathrm{D}}\mathbf{R}$, where $\mathbf{Y}_{\mathrm{D}} = \left[\mathbf{Y}_1, \cdots, \mathbf{Y}_{\mathrm{C}_s}\right]$ is the virtual label matrix of the dictionary Dic. $\mathbf{Y}_i$ corresponds to the i-th sub-dictionary. Elements of $\mathbf{Y}_i$ were set to 0 except the i-th row, of which the elements were set to 1.

In [111], Mahyari *et al.* used a Low Rank Embedding (LRE) which is belong to manifold alignment framework to address the anomaly detection of industrial robot. The main challenge with the existing detection algorithms was that when the task of the robot changes, the extracted features differ from those of the normal behavior and lead to false alarm. To eliminate the false alarm, the source domain data (normal data from task A) and the target domain data (normal and anomaly data from task B) were projected into a common subspace through LRE. Given source domain data matrix $\mathbf{X}_s \in \mathbb{R}^{\mathrm{D}\times n_s}$ and target domain data matrix $\mathbf{X}_t \in \mathbb{R}^{\mathrm{D}\times n_t}$, the LRE was calculated through minimizing the following loss function

$$\min_{\mathcal{R}_s} \frac{1}{2}\left\|\mathbf{X}_s^{\mathrm{T}} - \mathbf{X}_s^{\mathrm{T}}\mathcal{R}_s\right\|_F^2 + \tau\left\|\mathcal{R}_s\right\|_* \tag{17}$$

$$\min_{\mathcal{R}_t} \frac{1}{2}\left\|\mathbf{X}_t^{\mathrm{T}} - \mathbf{X}_t^{\mathrm{T}}\mathcal{R}_t\right\|_F^2 + \tau\left\|\mathcal{R}_t\right\|_* \tag{18}$$

where $\left\|\cdot\right\|_F$ and $\left\|\cdot\right\|_*$ are Frobenius and spectral norms, respectively. $\mathbf{X}_s^{\mathrm{T}}\mathcal{R}_s$ and $\mathbf{X}_t^{\mathrm{T}}\mathcal{R}_t$ are the low rank maps of $\mathbf{X}_s^{\mathrm{T}}$ and $\mathbf{X}_t^{\mathrm{T}}$, $\mathcal{R}_s$ and $\mathcal{R}_t$ are their reconstruction coefficient matrices. After finding LRE of the source and target samples, the projection matrices from the source and the target space

TABLE 3. Summary of the references that use deep transfer methods.

Categories	Sub-categories	References
Representation adaptation	Top-layer adaptation	CNN: [79] [85] [94] [97] [112] [129] AE: [88] [101]
	Multiple layers adaptation	CNN: [84] [98] AE: [65] [89] [113] SF: [63]
Parameter transfer	——	CNN: [103] [90] [87] [69] [70] [93] [106] [104] [105] [108] [92] [109] AE: [91] [107]
Others deep strategies	AdaBN	CNN: [73] AE: [124]
	——	[82] (deep generative neural network) [126] (CNN + mSDA) [68] (TrAdaBoost + CNN)

into the common subspace were calculated by minimizing the following cost function

$$(1-\mu)\left\|\mathbf{F}-\boldsymbol{\mathcal{R}}\mathbf{F}\right\|_F^2+\mu\sum\nolimits_{i,j=1}^{n_s}\left\|\mathbf{F}_i-\mathbf{F}_j\right\|^2\mathcal{I}(i,j) \quad (19)$$

where $\mu \in [0,1]$ determines the importance of the local geometry. The block reconstruction coefficient matrix is $\boldsymbol{\mathcal{R}}=\begin{bmatrix}\boldsymbol{\mathcal{R}}_s & 0\\ 0 & \boldsymbol{\mathcal{R}}_t\end{bmatrix}$. $\mathcal{I}=\begin{bmatrix}0 & \mathbf{I}\\ \mathbf{I} & 0\end{bmatrix}$ represents the inter-set correspondence between samples of the source domain and the target domain datasets, $\mathbf{I}$ is the identity matrix. $\mathbf{F}=[\mathbf{F}_s,\mathbf{F}_t]^{\mathrm{T}}\in\mathbb{R}^{(n_s+n_t)\times d}$ is the embedded matrix of the source domain samples and the target domain samples in the common subspace. In the new d dimensional subspace, the embedded test sample was compared to that of the source domain healthy data using Euclidean distance for anomaly detection.

Besides, Shen *et al.* [80] proposed a modified Least Square Support Vector Machine (LSSVM) method through adding penalty term and constraint condition of the source domain data to the original objective function of LSSVM. The proposed method, which belongs to classifier adaptation, achieved the cross-domain fault diagnosis of rolling element bearing.

2) DEEP TRANSFER APPROACHES

In the past few years, deep learning based transfer learning methods for cross-domain fault diagnosis have been intensively studied, which, in our taxonomy, can be categorized into: (1) representation adaptation, (2) parameter transfer, and (3) other deep transfer strategies. The summary of related references can be found in Table 3. In the table, the specific network architectures employed in those research works are also presented, including CNN, AE and its variants, and Sparse Filtering (SF).

a: REPRESENTATION ADAPTATION BASED APPROACHES

From the perspective of representation learning, deep networks can learn high-level abstract representations by multiple layers of non-linear transformations. Generally, the representations in lower layers are more general and those in higher layers are more specific to learning objectives. Due to the discrepancy of original signals between domains in cross-domain fault diagnosis, the network trained using the source domain data tends to break down when applied to the target domain data. It means that the representations of the source domain and the target domain learned by the same deep network are also different.

Representation adaptation based approaches aim to learn domain agnostic representation in the top layer or several intermediate layers, then the trained network using massive samples from the source domain may perform well on target tasks. Usually, this kind of approaches align the statistical distributions of the representations through adding a trade-off term which punishes the distribution discrepancy between domains in the learning process. The distribution distance statistics, such as MMD [65], [79], [84], [85], [88], [89], [94], [97], [98], [101], [112], correlation alignment (CORAL) distance [113], and Kullback-Leibler (KL) divergence [63], are commonly used for comparing the distribution shift between domains in fault diagnosis. We divide these approaches into: (a) top-layer adaptation and (b) multiple layer adaptation.

i) TOP-LAYER ADAPTATION

Intuitively, a deep neural network consists of the feature extractor and the classifier (or label predictor). The top-layer representation of the feature extractor is the most abstract one across all layers, and it is directly connected to the classifier, such as Softmax layer or SVM, corresponding to specific classification or regression tasks. To adapt the shift between domains, several research works proposed to align the distributions of the source domain and the target domain in the top-layer of the feature extractor. By this means, the domain-invariant representation across the two domains was learned.

Under the CNN architecture, Han *et al.* [85] proposed a deep transfer network with joint distribution adaptation. The architecture of the proposed method is illustrated in Fig.7. Step 1, a CNN model was pre-trained on the sufficient labeled data of source domain $X_s=\left\{\left(\boldsymbol{x}_i^s,y_i^s\right)\right\}_{i=1}^{n_s}$ from scratch with minimizing the following optimization objective $\mathcal{L}_c$

$$\mathcal{L}_c=\sum\nolimits_j\ell\left[y_j,f\left(\boldsymbol{x}_j:\{W_i,b_i\}_{i=1}^{L}\right)\right] \quad (20)$$

where ℓ denotes the loss function between the true label y_j and the predicted label by CNN model $f\left(\boldsymbol{x}_j:\{W_i,b_i\}_{i=1}^{L}\right)$. The pseudo labels $\hat{Y}_0=\left\{\hat{y}_i^t\right\}_{i=1}^{n_t}$ of the unlabeled target samples $X_t=\left\{\boldsymbol{x}_i^t\right\}_{i=1}^{n_t}$ for training was predicted using the pre-trained CNN. Step 2, domain adaptation training was implemented based on a new cost function which integrates the $\mathcal{L}_c$ and a regularization term of joint distribution adaptation (JDA).

$$\mathcal{L}(\Theta)=\mathcal{L}_c+\lambda\mathcal{L}_{\mathrm{D}}(J_s,J_t) \quad (21)$$

where $\Theta=\{W_i,b_i\}_{i=1}^{L}$ is the parameter collection of the CNN with L layers. $\mathcal{L}_{\mathrm{D}}(J_s,J_t)$ is the JDA term that simultaneously measures the discrepancy of the marginal and

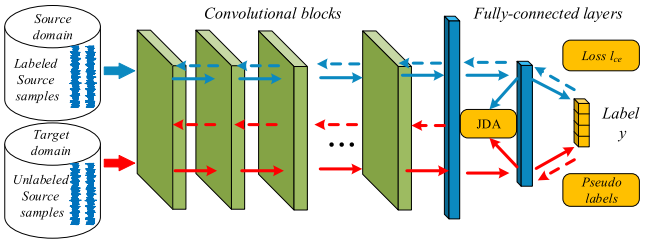


FIGURE 7. Architecture of deep transfer netwrok for unsupervised domain adaptation [85].

condition distributions between two domains.

$$\mathcal{L}_{\mathrm{D}}\left(J_s, J_t\right) = \mathrm{MMD}^2_{\mathcal{H}}\left(X_s, X_t\right) + \sum\nolimits_{c=1}^{\mathrm{C}} \mathrm{MMD}^2_{\mathcal{H}}\left(X_s^c, X_t^c\right) \tag{22}$$

$\mathrm{MMD}^2_{\mathcal{H}}\left(X_s, X_t\right)$ is the MMD distance between the marginal distributions $\mathcal{P}_s$ and $\mathcal{P}_t$. $\sum_{c=1}^{\mathrm{C}} \mathrm{MMD}^2_{\mathcal{H}}\left(X_s^c, X_t^c\right)$ is the MMD distance between the conditional distributions $\mathcal{Q}_s$ and $\mathcal{Q}_t$. It was computed through combining the MMDs corresponding to C common categories between domains. That is to say they were computed by the following formulas (23) and (24)

$$\begin{aligned}&\mathrm{MMD}^2_{\mathcal{H}}\left(X_s, X_t\right)\\ &= \left\| \frac{1}{n_s} \sum\nolimits_{i=1}^{n_s} \phi\left(\boldsymbol{x}_i^s\right) - \frac{1}{n_t} \sum\nolimits_{i=1}^{n_t} \phi\left(\boldsymbol{x}_i^t\right) \right\|^2_{\mathcal{H}}\end{aligned} \tag{23}$$

$$\begin{aligned}&\mathrm{MMD}^2_{\mathcal{H}}\left(X_s^c, X_t^c\right)\\ &= \left\| \frac{1}{n_s^c} \sum_{\boldsymbol{x}_i^s \in X_s^c} \phi\left(\boldsymbol{x}_i^s\right) - \frac{1}{n_t^c} \sum_{\boldsymbol{x}_j^t \in X_t^c} \phi\left(\boldsymbol{x}_j^t\right) \right\|^2_{\mathcal{H}}\end{aligned} \tag{24}$$

where $\phi : \mathcal{X} \to \mathcal{H}$ is the nonlinear mapping from original space to RKHS. In the process of domain adaptation training, the pseudo labels of unlabeled target samples were iteratively updated to obtain the optimal prediction accuracy. Step 3, the CNN after domain adaptation training were applied to identify the unseen samples of the target domain. Extensive cross-domain diagnosis tasks of rolling bearing and gearbox validated the applicability and practicability of the proposed method.

A similar research work can be found in [79] for power equipment fault diagnosis. Besides, Yang *et al.* also proposed a transfer learning method named convolutional adaptation network (CAN) in [97], the main differences between CAN and the method proposed in [85] are that CAN just reduced the marginal distribution discrepancy between the top-layer representations of two domains and the distribution distance was estimated by a MK-MMD.

Also under the CNN architecture, Guo *et al.* [94] proposed an intelligent fault diagnosis method for machinery, named deep convolutional transfer learning network (DCTLN), which also follows the top-layer adaptation strategy. In their problem setting, unlabeled samples of the target domain were available in the training stage. The DCTLN included a condition recognition module and a domain adaptation module, as shown in Fig.8. Condition recognition was achieved by a 1-D CNN with 16 layers among which the last layer is seen as a health condition classifier. Domain adaptation was implemented by a domain classifier and a distribution distance metrics. The adaptation module was connected to the feature extractor to help the CNN learn domain-invariant representations. In DCTLN, the first optimization object was to minimize the health condition classification error on the source domain data for learning features that are able to distinguish different health conditions. For the source domain dataset with C categories, the objective function of condition recognition was defined as a standard Softmax regression loss

$$\mathcal{L}_c = \frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{\mathrm{C}} I[y_i = k] \log \frac{e^{\left((\mathbf{w}_j)^{\mathrm{T}} \Phi^{fc2} + b\right)}}{\sum_{l=1}^{k} e^{\left((\mathbf{w}_j)^{\mathrm{T}} \Phi^{fc2} + b\right)}}\right] \tag{25}$$

where m is the batch size of the training samples, $I[\cdot]$ is an indicator function, Φ^{fc2} is the output of layer $fc2$. The second optimization object of the DCTLN was to maximize the domain classification error of the domain classifier which was connected with the feature extractor on the source and the target data. The domain classification loss $\mathcal{L}_d$ was

$$\mathcal{L}_d = \frac{1}{n_s}\sum_{i=1}^{n_s} \ell_d\left(\Phi_{si}^{fc2}\right) + \frac{1}{n_t}\sum_{j=1}^{n_t} \ell_d\left(\Phi_{tj}^{fc2}\right) \tag{26}$$

where Φ_s^{fc2} and Φ_t^{fc2} are the output of $fc2$ (high-level representations) from the source domain data and the target domain data, respectively. n_s and n_t are the numbers of samples of the source domain and the target domain respectively. $\ell_d(\cdot)$ is the empirical risk of domain classifier that is defined as (27)

$$\ell_d = \frac{1}{m}\sum_{i=1}^{m}\left[g_i \log \hat{g}\left(\boldsymbol{x}_i\right) + \left(1 - g_i\right) \log\left(1 - \hat{g}\left(\boldsymbol{x}_i\right)\right)\right] \tag{27}$$

where g_i is the ground-truth domain label, and $\hat{g}\left(\boldsymbol{x}_i\right)$ denotes the output domain label for i-th sample. The third optimization object of the DCTLN was to minimize the distribution distance between the source and target domains, and the MMD distance between high-level representations of the source domain $fc2$s and the target domain $fc2$t was estimated

$$\begin{aligned}\mathcal{L}_{\mathrm{D}} = &\frac{1}{n_s^2}\sum_{i=1}^{n_s}\sum_{j=1}^{n_s} K\left(\Phi_{si}^{fc2}, \Phi_{sj}^{fc2}\right)\\ &+ \frac{1}{n_t^2}\sum_{i=1}^{n_t}\sum_{j=1}^{n_t} K\left(\Phi_{ti}^{fc2}, \Phi_{tj}^{fc2}\right)\\ &- \frac{2}{n_s n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t} K\left(\Phi_{si}^{fc2}, \Phi_{tj}^{fc2}\right)\end{aligned} \tag{28}$$

where $K(\cdot, \cdot)$ is a kernel function.

The final optimization object of the DCTLN was

$$\begin{aligned}\mathcal{L}\left(\theta_f^*, \theta_c^*, \theta_d^*\right) = \min_{\theta_f, \theta_c, \theta_d} &\ \mathcal{L}_c\left(\theta_f, \theta_c\right) - \gamma \mathcal{L}_d\left(\theta_f, \theta_d\right)\\ &+ \lambda \mathcal{L}_{\mathrm{D}}\left(\theta_f\right)\end{aligned} \tag{29}$$

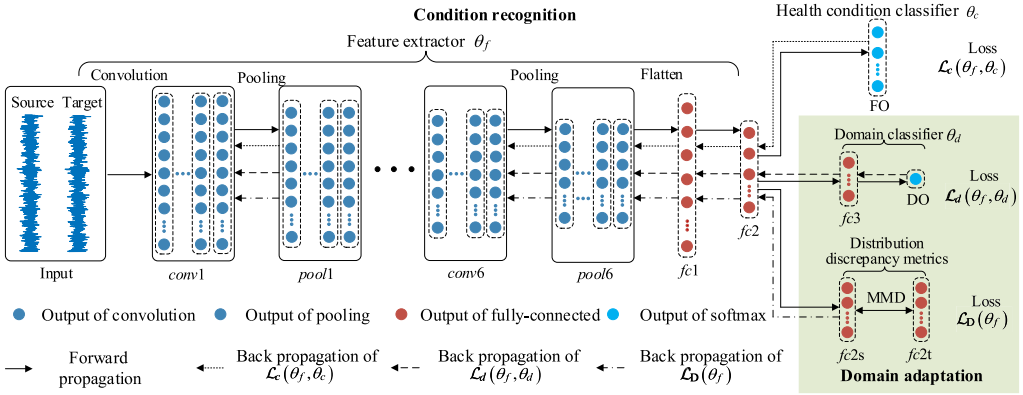


FIGURE 8. Structure illustration of deep convolutional transfer learning network (DCTLN) [94]. *conv1*, *conv6* denote convolutional layers, *pool1*, *pool6* denote pooling layers, *fc*1, *fc*2, and *fc*3 denote fully-connected layers. FO denotes the output of health condition classifier, and DO denotes the output of domain classifier.

where γ, λ are two trade-off parameters for domain adaptation. θ_f, θ_c, and θ_d are the parameters of the feature extractor, health condition classifier, and the domain classifier, respectively.

In [112], Li *et al.* proposed a deep distance metric learning method for rolling bearing fault diagnosis under the CNN architecture as well. The objective function of their method was

$$\min \mathcal{L} = \mathcal{L}_c + \lambda \mathcal{L}_{\mathrm{D}}^{fc} + \vartheta \mathcal{L}_{\mathrm{cluster}}^{fc} \tag{30}$$

where $\mathcal{L}_c$ is the cross-entropy loss function of Softmax classifier, $\mathcal{L}_{\mathrm{D}}^{fc} = \mathrm{MMD}_k\left(\Phi_s^{fc}, \Phi_t^{fc}\right)$ represents the MK-MMD between the representation Φ_s^{fc} and Φ_t^{fc} of the fully-connected layer *fc* of the source and target domains. λ and ϑ are the regularization parameters. In addition to these two objectives, this method considered a representation clustering term $\mathcal{L}_{\mathrm{cluster}}^{fc} = -\mathcal{L}_{\mathrm{inter}} + \eta \mathcal{L}_{\mathrm{intra}}$ that is favorable to fault classification. $\mathcal{L}_{\mathrm{inter}}$ and $\mathcal{L}_{\mathrm{intra}}$ measure the inter-class separability and intra-class compactness of the fully-connected layer representation respectively, and η is a scaling coefficient.

Under AE architecture, Wen *et al.* [88] proposed a top-layer deep adaptation method which used a three-layer sparse auto-encoder (SAE) to extract the features of power spectrum, and applied the MMD term to adapt the distribution discrepancy between the features from the data of the source and target domains. The architecture of the proposed method is shown in Fig.9. First, the three-layer sparse auto-encoder network was pre-trained by the labeled source data and the unlabeled target data. Then, the whole network was fine-tuned through optimizing the following objective function

$$\mathcal{L}_{\mathrm{DTL}}(\theta) = \mathcal{L}_c\left(y_s, \hat{y}_s\right) + \lambda \mathcal{L}_{\mathrm{D}}\left(\Phi_s^3, \Phi_t^3\right) \tag{31}$$

where $\mathcal{L}_c\left(y_s, \hat{y}_s\right)$ is the classification loss on source dataset, and $\mathcal{L}_{\mathrm{D}}\left(\Phi_s^3, \Phi_t^3\right)$ is the discrepancy penalty between the features of source dataset and target dataset. Φ_s^3 and Φ_t^3 denote the output of three-layer SAE corresponding to the source data and the target data, respectively.

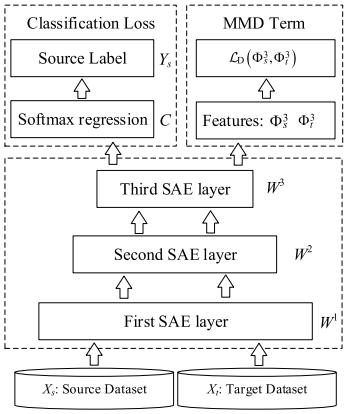


FIGURE 9. Architecture of the method proposed by Wen *et al.* [88].

Xu *et al.* [101] also proposed a similar method named two-phase digital-twin-assisted fault diagnosis method using deep transfer learning, which was also based on the SAE architecture.

ii) MULTIPLE LAYERS ADAPTATION

To achieve distribution alignment in the top layer of the feature extractor is the ultimate objective of deep domain adaptation. However, the top-layer representation is learned through multiple abstraction process of different levels corresponding to multiple intermediate representations. The distribution discrepancies of the representations in multiple intermediate layers between the source and target domains may influence the final adaptation result in the top layer. Some research works have implemented the domain adaptation in the top layer and multiple intermediate layers simultaneously [63], [65], [84], [89], [98], [113].

Under CNN architecture, Yang *et al.* proposed a feature-based transfer neural network (FTNN) to diagnose an actual locomotive bearing by leveraging knowledge from the data of laboratory bearings [98]. This FTNN model, whose architecture is shown in Fig.10, employed a domain-shared CNN to

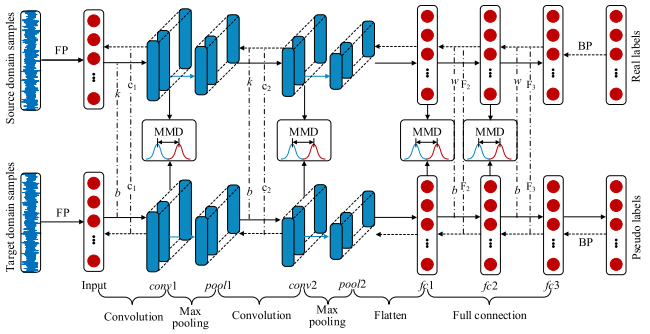


FIGURE 10. Architecture of the feature-based transfer neural network (FTNN) model [98].

extract transferable features from the raw vibration data both in the source and target domains. In the domain adaptation process, the distributions of the learned features in two convolutional layers *conv*1, *conv*2 and two fully-connected layers *fc*1, *fc*2 were adapted. The multi-layer domain adaptation term $\mathcal{L}_{\mathrm{D}}\left(\Phi_s^{\Lambda}, \Phi_t^{\Lambda}\right)$ in the objective function of FTNN was defined as follows

$$\begin{aligned}\mathcal{L}_{\mathrm{D}}\left(\Phi_s^{\Lambda}, \Phi_t^{\Lambda}\right) = &\frac{1}{n_s^2}\sum_{i=1}^{n_s}\sum_{j=1}^{n_s}\sum_{l=\Lambda}\kappa_l \cdot K\left(\Phi_{si}^l, \Phi_{sj}^l\right) \\ &-\frac{2}{n_s n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}\sum_{l=\Lambda}\kappa_l \cdot K\left(\Phi_{si}^l, \Phi_{tj}^l\right) \\ &+\frac{1}{n_t^2}\sum_{i=1}^{n_t}\sum_{j=1}^{n_t}\sum_{l=\Lambda}\kappa_l \cdot K\left(\Phi_{ti}^l, \Phi_{tj}^l\right) \\ \kappa_l = &\,1-\mathcal{L}_{\mathrm{D}}\left(\Phi_s^l, \Phi_t^l\right)\Big/\mathcal{L}_{\mathrm{D}}\left(\Phi_s^{\Lambda}, \Phi_t^{\Lambda}\right) \end{aligned} \tag{32}$$

where $\Phi_s^{\Lambda} = \cup_{l\in\Lambda}\Phi_s^l$ and $\Phi_t^{\Lambda} = \cup_{l\in\Lambda}\Phi_t^l$ represent the multi-layer features of the source domain and target domain, respectively. $\Lambda = \{conv1, conv2, fc1, fc2\}$ is the indexes of adaptation layers.

In [84], Li *et al.* proposed a similar multi-layer domain adaptation method under CNN architecture. A MK-MMD term which measures the distribution discrepancy between the two domains in multiple layers was integrated with the classification loss of CNN for learning domain-invariant features.

Under the AE architecture, Lu *et al.* [89] proposed a **D**eep neural network for domain **A**daptation in **F**ault **D**iagnosis (DAFD). DAFD followed a different domain adaptation strategy from [84], [98] that the feature distribution between the source domain and the target domain was aligned **layer-by-layer**. In each layer, a MMD term $\mathcal{L}_{\mathrm{D}}$ for reducing the discrepancy between distributions of the source and target domains and a weight regularization term $\mathcal{L}_{\mathrm{weight}}$ for reinforcing the representative features were added to the loss function of auto-encoder $\mathcal{L}_{ae}$. The final objective function of DAFD model was

$$\mathcal{L}_{\mathrm{DAFD}} = \mathcal{L}_{ae} + \lambda\mathcal{L}_{\mathrm{D}} + \frac{\beta}{2}\mathcal{L}_{\mathrm{weight}} \tag{33}$$

where $\lambda > 0$ and $\beta > 0$ control the tradeoff among three terms. $\mathcal{L}_{ae} = \|\mathbf{R}-\mathbf{X}\|_F^2/2n$, and $\mathbf{R}$ is the output of the decoder for reconstructing the input data. $\|\cdot\|_F$ denotes Frobenius norm. The MMD was employed to measure the discrepancy of labeled samples according to the categories and domains they belong to. The weight regularization term was $\mathcal{L}_{\mathrm{weight}} = \sum_{\mathrm{W}_k\in\{\mathrm{W}_e,\mathrm{W}_d\}}\exp\left(-\|\mathrm{W}_k\|_F^2/\tau\right)$, where $\{\mathrm{W}_e, \mathrm{W}_d\}$ is the weight matrix set of AE, τ is a punishment factor.

Comparing [89] with [84], [98], it is found that the domain adaptation process of DAFD [89] is implemented in unsupervised representation learning stage layer-by-layer, but the process in [84] and [98] is implemented during the supervised learning stage through one step.

Similar to [89], Qian *et al.* proposed a deep transfer learning method for fault diagnosis of rotating machinery based on SF architecture which was also an unsupervised feature learning method [63]. During feature extraction, a distribution distance penalty term $\mathcal{L}_{\mathrm{D}}\left(\hat{\varphi}_s, \hat{\varphi}_t\right)$ that was measured by the high-order Kullback-Leibler (HKL) was integrated with the loss of SF. Then, the objective function of feature extractor was

$$\mathcal{L}_1 = \mathcal{L}_{\mathrm{SF}}\left(\hat{\varphi}\right) + \lambda_1\mathcal{L}_{\mathrm{D}}\left(\hat{\varphi}_s, \hat{\varphi}_t\right) \tag{34}$$

where $\hat{\varphi}_s$ and $\hat{\varphi}_t$ are the normalized feature matrix of the source domain data and the target domain data respectively. $\hat{\varphi} = \left\{\hat{\varphi}_s, \hat{\varphi}_t\right\}$ is the combination of $\hat{\varphi}_s$ and $\hat{\varphi}_t$. $\mathcal{L}_{\mathrm{SF}}\left(\hat{\varphi}\right)$ is the objective function of SF. Besides the adaptation process in unsupervised feature extraction, this method aligned the two domains in feature classification stage at the same time by optimizing the following objective function

$$\mathcal{L}_2 = \mathcal{L}_c\left(\varphi_s, Y_s\right) + \lambda_3\mathcal{L}_{\mathrm{D}}\left(\Phi_s, \Phi_t\right) \tag{35}$$

where $\mathcal{L}_c\left(\varphi_s, Y_s\right)$ is the loss function of classifier (Softmax layer). $\mathcal{L}_{\mathrm{D}}\left(\Phi_s, \Phi_t\right)$ was also measured by HKL, and $\Phi_s = |\varphi_s|$, $\Phi_t = |\varphi_t|$.

b: PARAMETER TRANSFER BASED APPROACHES

From the perspective of feature transformation, a deep neural network transforms the input signals into a new space by its multi-layer parameters learned from the training dataset. Therefore, the implicit relation model between the input and the output of the network is determined by the parameters among layers. With respect to the cross-domain scenario, the parameter transfer approaches assume that the individual models for related tasks should share a part of parameters θ_{com} [17], [90]. Formally, let θ_s and θ_t denote the network parameters for the source task and the target task, respectively. Then,

$$\theta_s = \theta_{\mathrm{com}} + \theta_s' \tag{36}$$

$$\theta_t = \theta_{\mathrm{com}} + \theta_t' \tag{37}$$

where θ_{com} is the common parameters while θ_s' and θ_t' are specific parameters for the source task and the target task respectively.

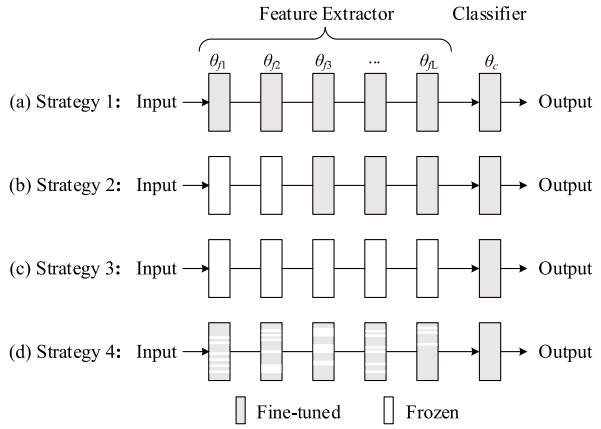


FIGURE 11. Different fine-tune strategies of parameter transfer based approaches for cross-domain fault diagnosis.

Based on this idea, some research works proposed to inherit a part of parameters from the networks for source tasks when training the network for the target task [69], [70], [87], [90]–[93], [103]–[109]. The basic procedures of these parameter transfer approaches under deep neural network architecture include three main steps:

Step1: A pre-trained network **A** is trained using the source domain data. Usually, the data from the source domain are fully labeled, and the amount of the data is large enough to train a robust identification model.

Step2: Construct a new network **B** for the target task and transfer a portion of the parameters from the pre-trained network **A** for initializing the network **B**.

Step3: Fine-tune the new network **B** using a small amount of data from the target domain.

In most cases, the new constructed network **B** has the same architecture of feature extractor with the network **A**, and the parameters of the feature extractor are initialized using the parameters of the network **A** in corresponding layers. But the classifier layers of the network **B** are usually constructed in a new structure for the target task and its parameters are randomly initialized.

The fine-tune strategies of those parameter transfer approaches for cross-domain fault diagnosis can be summarized into four categories, as demonstrated in Fig.11:

(1) Fine-tune all layers (including feature extractor and classifier) using the data of the target domain [69], [70], [87], [90], [93], [103], [106], [109];
(2) Fine-tune a portion of layers of the feature extractor, and the classifier layers [104], [109];
(3) Just fine-tune the classifier layers [105], [108], [109];
(4) Selective parameter fine-tune [92].

In [109], Han *et al.* presented a parameter transfer framework based on pre-trained CNN. In their research work, three different parameter transfer strategies were discussed and compared to investigate the applicability as well as the significance of feature transferability from the different levels of a deep structure. They considered that two factors, the dataset size and similarity, would guide the selection of parameter transfer strategy. When the dataset of the target domain is large, fine-tuning the entire network is a good choice. When the target dataset is small and similar to the source dataset, it will be better to fix the parameters of feature extractor and just fine-tune the classification layer. In the last case, the target dataset is both small and dissimilar to the source dataset, retraining the front convolutional blocks and the classification layer may be efficient.

Kim and Youn [92] proposed a selective parameter freezing approach, which can retrain of only unnecessary parameters to the target data while remain the important parameters from the source network. The proposed method offered an option for adjusting the freezing and fine-tuning inside a layer.

c: OTHER DEEP TRANSFER STRATEGIES

In addition to the above two strategies, several other transfer strategies based on deep neural networks are also proposed to address cross-domain fault diagnosis tasks.

In [73] and [124], a simple domain adaptation method, called Adaptive Batch Normalization (AdaBN) [125], which modulates the statistics from the source domain to the target domain in all Batch Normalization layers across the network, was employed for address cross-domain fault diagnosis. In AdaBN, the standardization of each layer by domain ensures that each layer receives data complying with similar distribution, regardless of the source domain or target domain. Given a deep neural network model pre-trained using the data of source domain, the AdaBN algorithm is as follows

Algorithm Adaptive Batch Normalization (AdaBN)

For: neuron j in deep neural network do

Concatenate neuron response on all samples of target domain

$$t : x_j = \left[\cdots, x_j(m), \cdots\right]$$

Compute the mean and variance of the target domain:

$$\bar{x}_j^t = \mathbb{E}\left(x_j^t\right), \sigma_j^t = \sqrt{Var\left(x_j^t\right)}$$

End for

For: neuron j in DNN, testing sample m in target domain **do**

Compute Batch Normalization output

$$y_j(m) := \gamma_j \frac{\left(x_j(m) - \bar{x}_j^t\right)}{\sigma_j^t} + \beta_j$$

End for

In [82], Li *et al.* presented a cross-domain fault diagnosis method based on deep generative neural networks. First, different classes of fake fault samples of the target domain were generated through training C_s-1 generators which record the relations between normal samples and fault samples of each category. Second, a top-layer deep adaptation neural network was trained for cross-domain classification. In [126], CNN was combined with marginalized stacked denoising autoencoder to learn fault sensitive features and eliminate data distribution differences between different conditions. In [68],

Xiao *et al.* proposed a fault diagnosis framework through combining TrAdaBoost and CNN.

3) ADVERSARIAL-BASED APPROACHES

Ultimately, the objective of deep domain adaptation is learning domain-invariant representations across the source and target domains. To achieve this purpose, a statistic that measures the distribution discrepancy between the representations of the source and target domains is usually employed. Then, the distribution distance between domains is penalized during the representation learning process through adding its trade-off term into the original objective function of deep neural network. That is to say, the similarity of the source domain and the target domain is determined by a specific distance measurement statistic, and the representations with low distance statistic value are considered to be domain-invariant. Differently, inspired by GAN [54], adversarial-based domain adaptation determines the similarity between domains through an adversarial objective with respect to a domain discriminator.

Domain-adversarial neural network (DANN) is a representative adversarial-based domain adaptation method (its architecture is shown in Fig.12), and has been successfully implemented in computer vision application [58]. Wang *et al.* introduced DANN to cross-domain fault diagnosis in [86]. In order to learn discriminative and domain-invariant representations using a deep neural network, DANN used a rather different way to measure the disparity between distributions based on their separability by a discriminatively-trained classifier. DANN included three components: a feature extractor G_f with parameter θ_f, a label predictor G_c with parameter θ_c, and a domain discriminator G_d with parameter θ_d. During the learning process, in order to ensure the discriminativeness of the representation, DANN first to minimize the label prediction loss $\mathcal{L}_c\left(\theta_f, \theta_c\right)$ through optimizing the parameters of both the feature extractor and the label predictor. At the same time, in order to obtain domain-invariant features, DANN optimized the parameters of feature extractor to maximize the loss of domain prediction $\mathcal{L}_d\left(\theta_f, \theta_d\right)$ which means that the feature distributions of the source and target domains are similar. Simultaneously, the parameter θ_d of the domain discriminator would be optimized to minimize the loss of domain prediction. Formally, DANN can be considered to be the following min-max problem

$$\begin{aligned} \mathcal{L}\left(\theta_f, \theta_c, \theta_d\right) &= \mathcal{L}_c\left(\theta_f, \theta_c\right) - \gamma \mathcal{L}_d\left(\theta_f, \theta_d\right) \\ \left(\theta_f^*, \theta_c^*\right) &= \arg\min_{\theta_f, \theta_c} \mathcal{L}\left(\theta_f, \theta_c, \theta_d^*\right) \\ \theta_d^* &= \arg\max_{\theta_d} \mathcal{L}\left(\theta_f^*, \theta_c^*, \theta_d\right) \end{aligned} \tag{38}$$

In [120], the same strategy was also employed to address the diagnosis problems of wind turbine and gearbox. But, at the training stage no target domain data were used, and instead the training dataset was randomly divided into two parts for implementing adversarial training. Due to the training dataset contained the data from multiple sources the domain-invariant features or common diagnosis knowledge across these domains may be learned during the adversarial process.

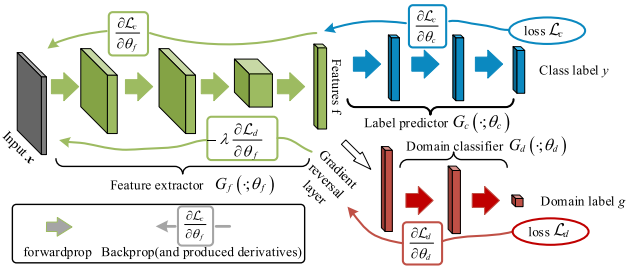


FIGURE 12. Architecture of domain-adversarial neural network (DANN) [58].

In the research works mentioned above [86], [120], the source domain and the target domain used a shared feature extractor during the adversarial learning process, as shown in Fig.12. Based on two different feature extractors, Zhang *et al.* [76] proposed an adversarial adaptive 1-D CNN (A2CNN) method, which is very similar to the adversarial discriminative domain adaptation (ADDA) proposed in [59]. In A2CNN, two CNNs were separately learned for the source and target domains which are called source feature extractor G_f^s and target feature extractor G_f^t respectively, as depicted in Fig.13. The target feature extractor G_f^t has the same structure with G_f^s, and partial layers of the target feature extractor were initialized by the pre-trained source one before domain adversarial training. This manner may be more flexible because of allowing more domain-specific features to be learned. A2CNN has been successfully applied to cross-domain fault diagnosis of bearing under variations of operating conditions.

In order to mitigate the gradient vanishing of DANN loss which amounts to minimizing the Jensen-Shannon divergence between distributions of the source and target domains, Wasseratein distance based adversarial networks have been proposed in [66] and [127] for fault diagnosis problem inspired by Wasserstein Generative Adversarial Networks (WGAN) [128].

In [66], Cheng *et al.* proposed a Wasserstein distance based deep transfer learning method. In their method, Wasserstein-1 distance, defined as $W\left(\mathcal{P}_s, \mathcal{P}_t\right) = \inf_{\zeta \in \prod\left(\mathcal{P}_s, \mathcal{P}_t\right)} \mathbb{E}_{\left(h^s, h^t\right) \sim \zeta}\left[\left\|h^s - h^t\right\|\right]$, was employed to measure the loss of domain discriminator. Among the definition, Ψ denotes a compact metric set, Prob (Ψ) represents the space of probability measures on set Ψ, $\mathcal{P}_s, \mathcal{P}_t \in$ Prob (Ψ). ζ is a joint probability distribution and $\prod\left(\mathcal{P}_s, \mathcal{P}_t\right)$ denotes the set $\Psi \times \Psi$ of all joint distributions $\zeta\left(h^s, h^t\right)$ whose marginal are $\mathcal{P}_s$ and $\mathcal{P}_t$ respectively. Wasserstein-1 distance can be viewed as an optimal transport problem, it aims to find an optimal transport plan $\zeta\left(h^s, h^t\right)$, which indicates how much of 'mass' randomly transported from one place h^s over the domain of h^t, with the aim of transporting the distribution $\mathcal{P}_s$into the distribution $\mathcal{P}_t$. To estimate the Wasserstein-1

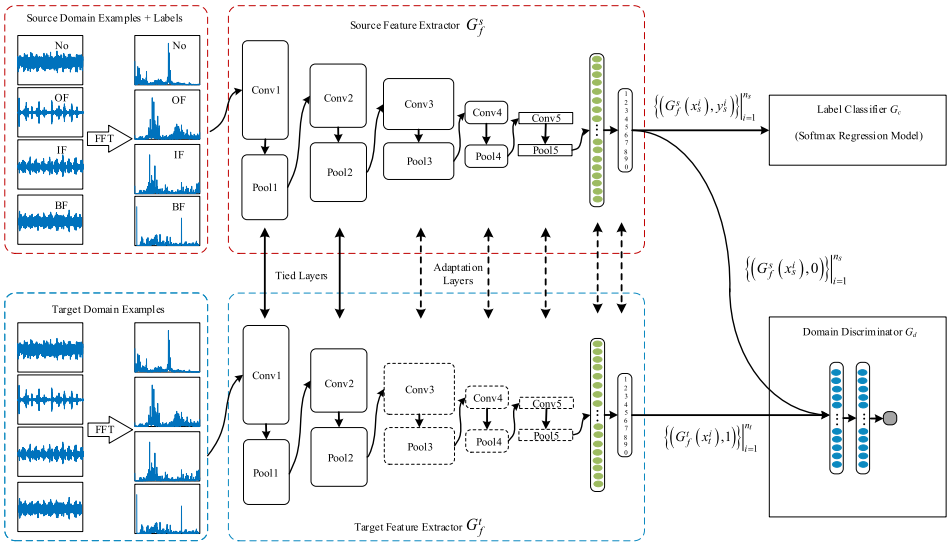


FIGURE 13. Architecture of adversarial adaptive 1-D CNN (A2CNN) [76].

distance between the features of domains, a domain critic was introduced to learn a solution $G_c : \Psi \to \mathbb{R}$ with corresponding parameters θ_c that maps the source and target features to a real number. The empirical Wasserstein-1 distance can be approximately computed as follow:

$$\mathcal{L}_d = \frac{1}{n_s} \sum_{\boldsymbol{x}_s \in X_s} G_c\left(G_f\left(\boldsymbol{x}_s\right)\right) - \frac{1}{n_t} \sum_{\boldsymbol{x}_t \in X_t} G_c\left(G_f\left(\boldsymbol{x}_t\right)\right) \quad (39)$$

where $\mathcal{L}_d$ denotes the domain critic loss between the source data X_s and the target data X_t. As the fact that the Wasserstein-1 distance is differentiable and continuous almost everywhere. The domain critic objective was trained by solving the following optimization problem:

$$\max_{\theta_c} \left\{\mathcal{L}_d - \rho \mathcal{L}_{grad}\right\} \quad (40)$$

where $\mathcal{L}_{grad} = \left(\|\nabla_{\mathbf{h}} G_c\left(\mathbf{h}\right)\|_2 - 1\right)^2$ is a gradient penalty, ρ is the balancing coefficient.

The final optimization problem of Wasserstein distance based deep transfer learning was

$$\min_{\theta_d, \theta_f} \left\{\mathcal{L}_c + \lambda \max_{\theta_c} \left[\mathcal{L}_{wd} - \rho \mathcal{L}_{grad}\right]\right\} \quad (41)$$

where $\mathcal{L}_c = \frac{1}{n_s} \sum_{i=1}^{n_s} -y_i^s \log \tilde{y}_i^s - \left(1 - y_i^s\right) \log\left(1 - \tilde{y}_i^s\right)$ is the classification loss of source domain. θ_d is the corresponding parameters of domain classifier.

In [127], a deep model named Wasserstein Distance Guided Multi-Adversarial Networks (WDMAN), which also used Wasserstein distance to measure the distribution discrepancy between domains during domain discrimination, was also proposed by Zhang *et al.* In adversarial training process, WDMAN adapted the distribution between domains in multiple layers based on multiple domain critic networks to promote the transfer capacity over previous single layer adaptation strategy.

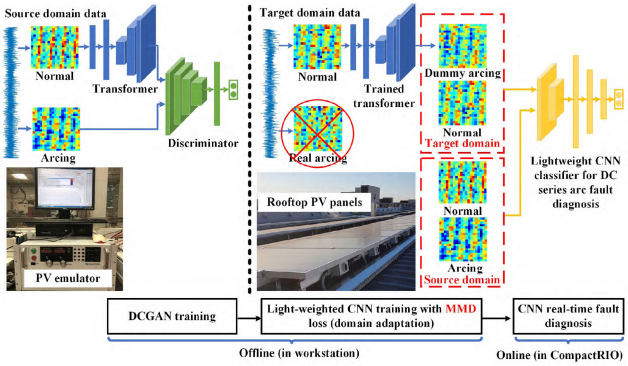


FIGURE 14. Framework of DA-DCGAN proposed by Lu *et al.* [115].

Actually, lacking of the fault data from the target domain is the main reason that prevents the generalization performance of fault diagnosis model. Generating unlimited quantities of synthetic target data using GAN is an appealing alternative to address this issue. In [115], Lu *et al.* proposed a diagnosis method for DC series arc fault, called **D**omain **A**daptation combined with **D**eep **C**onvolutional Generative **A**dversarial **N**etwork (DA-DCGAN), and its framework is depicted in Fig.14. DA-DCGAN first learns an intelligent normal-to-arcing transformation from the source-domain data. Then by generating dummy arcing data with the learned transformation using the normal data from the target domain and employing domain adaptation, a robust and reliable fault diagnosis scheme was achieved for the target domain.

4) OTHER CROSS-DOMAIN FAULT DIAGNOSIS STRATEGIES

Besides the methods that use knowledge transfer strategies to tackle the cross-domain fault diagnosis problems, there are some research works in which no transfer learning

TABLE 4. Summary of the references on cross-domain fault diagnosis with- out using transfer strategy.

Strategies	References
Based on specific deep network or training strategy	[74] [99] [116] [117] [118][121]
Preprocessing	[100] [102] [119] [122]

algorithms are employed but cross-domain diagnosis tasks are considered. These approaches may be also feasible manners to mitigate the data dilemma of traditional data-driven diagnosis methods. Most of them are deep learning based diagnosis methods, they learn abstract representations for fault identification from raw monitoring signals, such as vibration signals [74], [99], or vibration images [116]–[118], [121]. Based on the strategy of reducing the gap between domains, we divide these approaches into two sub-categories: (1) based on specific deep network structure or training strategy, (2) based on preprocessing. A summary of these approaches can be found in Table 4.

The first category is to mitigate the domain discrepancy through devising appropriate network structures or training strategies of diagnosis models [74], [99], [116]–[118], [121]. For example, Zhang *et al.* [74] presented a Convolution Neural Networks with Training Interference (TICNN) method which can achieve superior accuracy under noisy environment and variations of working load without any domain adaptation algorithm. TICNN enhanced the anti-noise and domain adaptation ability based on three tricks: (1) dropout was used in the first-layer to add noise to raw input, (2) very small batch training was used for better generalization ability of the model, and (3) ensemble learning was used to enhance the stability of the algorithm. Other methods, such as Deep Inception Net with Atrous Convolution [99], convolutional neural network based on capsule network [116], Snapshot Ensemble Convolutional Neural Network [118], and Noise Deep Convolution Neural Model [121] etc. were also employed or proposed to promote the generalization ability of the models under cross-domain diagnosis scenarios.

The second category is to mitigate the discrepancy between domains by some preprocessing approaches and then traditional deep learning algorithms were applied for constructing diagnosis models [100], [102], [119], [122].

In [100], Hyunseok *et al.* proposed a vibration image generation approach based on Stacking Omnidirectional Regenerated Signals, and then the images were processed by a Histogram of Oriented Gradients before fed into a DBN for fault identification. Based on the preprocessing step, the proposed method achieved accurately diagnosis of the rotor system of a real 500 MW steam turbine by training the diagnosis model using the data from a small testbed. It is worth noting that a pair of proximity sensors are necessary for this method.

In [102], Cameron *et al.* proposed to train the CNN network using the data generated by high resolution bearing dynamics simulation and to diagnose actual bearings. In their method, the vibration signals was preprocessed using the following three steps: (1) the accelerometer signal envelope was computed, (2) angle synchronous averaging was taken over each characteristic defect signal period, (3) signal normalization. The envelope signal contains the diagnostic information about bearing faults, the angle synchronous averaging was to eliminate the differences of the rotating speed and bearing model between the training and test datasets, and signal normalization was to eliminate the possible influence of vibration amplitude.

Wei *et al.* [122] also proposed a rotating speed normalization approach to address the cross domain learning problem caused by rotating speed fluctuation. Han *et al.* [119] presented a diagnosis framework that combined the spatiotemporal pattern network (STPN) approach with CNN which can diagnose unseen operating condition and fault severities. The STPN, which was built on the formulation of transition probabilities among the states generated by symbolic dynamics filtering, can extract spatial (between measurements) and temporal features (for each measurement) that were robust to operating condition and fault severity.

TABLE 5. Inputs of different cross-domain diagnosis approaches.

Types of input data		References
Hand-crafted features		[60] [61] [62] [67] [71] [75] [77] [80] [81] [83] [96] [111] [123]
Time series	Raw or preprocessed signal	[68] [73] [74] [84] [90] [92] [94] [97] [98] [99] [101] [102] [108] [109] [110] [113] [119] [120] [122] [127] [129]
	Spectrum	[63] [64] [65] [66] [76] [78] [82] [86] [88] [89] [93] [112] [114] [124] [126]
Images	Signal-segment-stack	[70] [87] [100] [105] [115] [121]
	Time-frequency representations	[69] [104] [106] [116] [117] [118]
	Others	[95] [103]

5) INPUTS OF CROSS-DOMAIN DIAGNOSIS APPROACHES

In the above four parts, we discussed the specific schemes of different diagnosis approaches applied in cross-domain scenarios. Another point the readers may interested is the types of input data of these approaches. Actually, different inputs mean different data types acceptable by different approaches and also mean the required levels of characteristic abstraction of these approaches. In the part, we summarize the different input types of cross-domain diagnosis approaches by Table 5.

Mostly, the inputs of the cross-domain diagnosis approaches based on traditional transfer learning were hand-crafted features, such as statistical parameters of vibration signal [60], [61], [67], [71], [80], [96], [123], SVD eigenvalues [81], [83] etc. The features used in these approaches were usually elaborately extracted and selected from massive candidates, and in this procedure the expert knowledge that which features are more discriminative has been considered. This kind of approaches may not be able to achieve superior performance when feeding into raw monitoring signals.

TABLE 6. Summary of application objects of cross-domain fault diagnosis.

Application objects	References
Bearing	[62] [63] [64] [65] [66] [67] [69] [70] [72] [73] [74] [75] [76] [77] [78] [80] [81] [82] [83] [84] [85] [86] [87] [88] [89] [90] [91] [92] [93] [94] [96] [97] [98] [99] [102] [104] [105] [106] [110] [112] [114] [116] [117] [118] [119] [124] [127] [129]
Gearbox	[60] [61] [63] [64] [71] [78] [85] [89] [103] [104] [107] [109] [120] [124]
Induction motor	[68] [83] [104]
Centrifugal pump	[105] [118] [87]
Wind turbine	[85] [119] [120]
Power equipment	[79] [113] [130]
Rotor system	[100] [122]
Other systems	[126] (Reciprocating compressor), [108] (Gas turbine) [101] (Production line), [95] (sucker rod pumping system) [123] (3D printer), [121] (RV reducer), [111] (Robot), [115] (Photovoltaic System)

On the contrary, the inputs of the cross-domain diagnosis approaches based on deep transfer learning, adversarial strategy, and other deep neural networks were raw monitoring signals or features with low levels of abstraction. According to different data types, we divide the inputs of these approaches into two categories: time series data and image data. Among them, 1D time series were the most common input type, such as raw or preprocessed vibration signals [68], [73], [74], [84], [90], [92], [94], [97]–[99], [102], [109], [110], [120], [122], [127], [129] and frequency spectra [63]–[66], [76], [78], [82], [86], [88], [89], [93], [112], [114], [124], [126]. Other approaches used 2D images as inputs, and the images were mostly generated by signal-segment-stack [70], [87], [100], [105], [115], [121] and time-frequency representations (including short-time Fourier transform [116], wavelet transform [104], [106], S-transform [69], [117], [118]). These approaches, most of which are based on deep neural networks, mixed the procedures of fault characteristic learning, pattern identification, and knowledge transfer together. And this is a new exploration for achieving a higher level of intelligent fault diagnosis.

D. APPLICATIONS

In this section, the applications of the cross-domain fault diagnosis methods are summarized. The summary of application objects about cross-domain fault diagnosis can be found in Table 6. The corresponding statistics on these research works is shown in Fig.15. Bearing and Gearbox are the two most widely research and validation objects of current cross-domain diagnosis literatures. The reason may be that some open-source fault data from different machines and different operating conditions are available, and the implementation of cross-domain experiments of them is relatively easier.

Shen *et al.* were the first to apply transfer learning to bearing fault diagnosis [81]. In their research, different domains were simulated by the data from different operating conditions (including rotating speed and load), and TrAdaboost method was applied to achieve knowledge transfer. After that, Zhang and Peng *et al.* [73], [74], [116], Li, Zhang, and Ding *et al.* [82], [84], [112], Zhang, Li and Tong *et al.* [62], [75]–[78], Wen and Gao *et al.* [88], [105], [117], [118], Qian and Li *et al.* [63]–[65], [124], Han *et al.* [85], Xu *et al.* [87], Cheng *et al.* [66], M.J.Hasan and Kim [69], [70], and others [67], [86], [110], [127] also focused on the cross-domain diagnosis problem of bearing under variation of operating conditions. Aiming at the data discrepancy caused by different fault severities, references [90], [91] and [93] achieved the cross-domain diagnosis between different bearing fault sizes using corresponding transfer learning methods. Besides, [92], [94], [97], [98], [102], [129] presented the feasibility that improving the bearing diagnosis accuracy by leveraging knowledge from other same-type machines or simulation models. Usually, the normal condition, inner race fault, outer race fault, and ball fault were considered in these cross-domain diagnosis tasks of bearing.

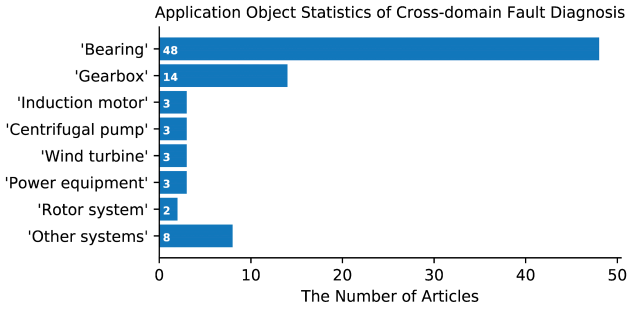


FIGURE 15. Statistics of application objects of cross-domain fault diagnosis.

Gearbox cross-domain diagnosis problem under different operating conditions was first discussed by Xie *et al.* in [60], [61], [71]. A traditional feature-based transfer learning method, TCA, was employed in their research work. Lu *et al.* [89], Zhang *et al.* [78], Han and Liu *et al.* [85], and Qian and Li *et al.* [63], [64], [124] validated their deep transfer methods both on gear and bearing diagnosis. In addition, Cao *et al.* [103] and Shao *et al.* [104] proposed to improve the performance of gear fault diagnosis by pre-training deep neural networks using massive visual images.

Besides the bearing and gearbox, knowledge transfer strategies were also applied to other industrial equipment for promoting diagnosis models' generalization performance. The cross-domain fault diagnosis of induction motor has been discussed in [68], [83], [104], and all of them validated their methods on the diagnosis tasks of induction motor across different operating conditions. In [83], the diagnosis tasks were four-class classification problem with respect to four health conditions, normal condition, unbalanced rotor, bowed rotor, and broken bar. In [68] and [104], the normal condition and five fault conditions, stator winding defect (voltage imbalance in [68]), unbalanced rotor, defective bearing, broken bar, and bowed rotor, were considered. In [87], [105] and [118], the corresponding methods were verified on the cross-domain tasks of self-priming centrifugal pump. In [85], [119],

[120], Han *et al.* applied their methods to wind turbine fault diagnosis under different operating conditions. Ten health conditions [85], [120] and twelve health conditions [119] were included in the cross-domain tasks, respectively. Hyunseok *et al.* [100] and Wei [122] discussed the cross-domain diagnosis problem of rotor system. In [100], normal condition, misalignment, oil whirl, and unknown fault of a 500 MW steam turbine were identified by training a DBN model with the help of the fault data from a small testbed. The transfer learning strategy was also applied to power equipment in [79], [113] and [130]. Wang [79] applied a deep transfer network with top-layer representation adaptation on power data for the first time. In [113], a hierarchical deep domain adaptation approach was proposed for diagnosing a high-pressure heater system in a 600 MW power plant under varying loading conditions.

There are also several preliminary studies that aim to tackle cross-domain diagnosis tasks of other equipment or systems apart from the ones mentioned above, such as reciprocating compressor [126], gas turbine [108], production line of smart manufacturing [101], sucker rod pumping system [95], 3D printer [123], RV reducer [121], robot [111], and photovoltaic system [115].

IV. OPEN-SOURCE FAULT DATASETS

As demonstrated in Section III.D, bearing and gear fault diagnosis are the top-two widely application subjects among the research works of cross-domain fault diagnosis. A possible reason is that there are several available open-source datasets for bearing and gear fault diagnosis, and based on them the cross-domain tasks can be easily organized. In this section, open source datasets for machinery fault diagnosis are introduced to facilitate readers to evaluate and compare their transfer diagnosis methods. Totally, 7 datasets including 6 bearing datasets: CWRU [131], MFPT [132], Paderborn University Dataset [133], DIRG Dataset [134], IMS Dataset [135], PHM12 Data Challenge Dataset [136], and 1 gearbox dataset, PHM09 Data Challenge Dataset [137] are summarized.

A. CWRU DATASET

CWRU dataset is a popular benchmark for rolling element fault diagnosis, which includes 4 health conditions of samples drawn from 4 different operating conditions. The test apparatus is illustrated in Fig.16, which consists of a driven motor, a torque transducer, and a dynamometer. For the tests, drive-end bearing (SKF: 6205-2RS JEM) and fan-end bearing (SKF: 6203-2RS JEM) of the motor were seeded with faults using electro-discharge machining. The faults were seeded on the rolling elements, inner races, and outer races with diameters 0.007, 0.014, 0.021, and 0.028 inches, respectively. Faulty bearings were reinstalled into the test motor and vibration data was recorded under motor loads of 0 to 3 horsepower (motor speeds of 1797 to 1720 rpm). Two accelerometers were placed at both the drive-end and fan-end of the motor housing (another an accelerometer was attached to the motor supporting base plate during some experiments), and the vibration signal was measured with sampling rate 12 kHz and 48 kHz.

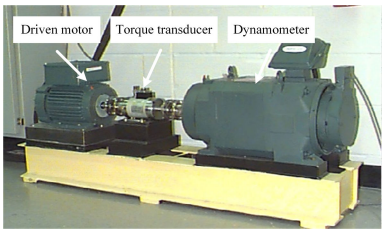


FIGURE 16. Test experimental apparatus of CWRU dataset.

References [62], [63], [75]–[78], [80]–[85], [64], [86]–[94], [96], [65], [97], [98], [102], [104], [105], [110], [112], [114], [116], [117], [66], [118], [119], [124], [127], [129], [67], [69], [72]–[74] organized cross-domain diagnosis tasks based on this dataset to validate their diagnosis methods.

B. MFPT DATASET

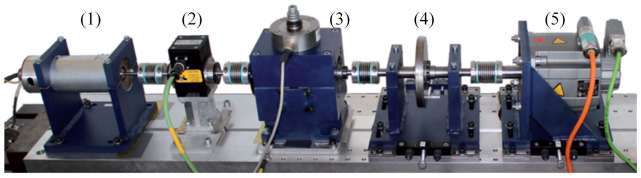
MFPT dataset is also for the fault diagnosis of the rolling element bearing, and it is provided by the Society for Machinery Failure Prevention Technology (MFPT). This dataset includes the vibration signals collected from three different health conditions: normal condition, outer race fault and inner race fault. In addition, the data from three real-world faults are also included. The experiment object of this dataset is also deep groove ball bearing, and its specific structural parameters are as follows: pitch diameter is 31.62 mm, ball diameter is 5.97 mm, contact angle is 0° and the number of element is 8.

This dataset has been employed to verify the cross-domain diagnosis methods in [102] and [118].

C. PADERBORN UNIVERSITY DATASET

This dataset is also for bearing fault diagnosis and is provided by KAT datacenter in Paderborn University. The mechanical setup of the test rig is shown in Fig.17, and the basic components of the test rig are a drive motor, a torque measurement shaft, a test modules and a load motor. The experimental bearing is FAG-6203 ball bearing. This dataset consists of the motor currents and vibration signals from 32 different bearing experiments, which belong to three main groups: 6 healthy bearings, 12 artificially damaged bearings, and 14 bearings with real damages. The faulty bearings are with inner race damage, outer race damage, or multiple damages. Artificially damaged bearings were damaged by electric discharge machining, drilling, and manual electric engraving. The real bearing damages were generated in an accelerated lifetime test rig by pre-defined continuous loads. The experiments were operated under different operating conditions as listed in Table 7.

References [99], [116], and [105] used this dataset in their research works.



(1) electric motor; (2) torque-measurement shaft; (3) rolling bearing test module; (4) flywheel; (5) load motor

FIGURE 17. The test rig of the Paderborn university dataset.

TABLE 7. Experimental operating conditions of Paderborn University dataset.

No.	Rotational speed [rpm]	Load Torque [Nm]	Radial force [N]
0	1500	0.7	1000
1	900	0.7	1000
2	1500	0.1	1000
3	1500	0.7	400

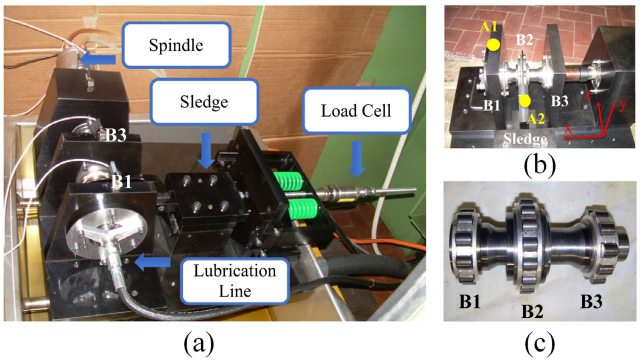


FIGURE 18. The test rig corresponding to DIRG dataset. (a) the overall architecture of the test rig, (b) the positions of three bearings (B1, B2, and B3) and two accelerometers (A1, A2), (c) the shaft with its three roller bearings.

D. DIRG DATASET

The DIRG dataset is about the high-speed aeronautical roller bearing, which is provided by the Dynamic and Identification Research Group (DIRG), in the Department of Mechanical and Aerospace Engineering at Politecnico di Torino. The corresponding test rig and its main parts are shown in Fig.18. The rotational speed of the spindle was set through the control panel of an inverter, and the speed was higher than 6000 rpm in the experiments. Two accelerometers were placed in points A1 and A2, as shown in Fig.18(b). The data samples corresponding to two different experimental sessions are included in DIRG dataset. The first one is the test under the B1 bearing with different damages (localized fault on inner race ring or a single roller), running at different speeds and under different loads. The second one is about a single damaged bearing undergoing a long (about 330h) test at a constant speed and load. The sampling rate of vibration signals is 51,200 Hz.

E. IMS DATASET

This is an accelerated life test dataset of bearings which is provided by the Center for Intelligent Maintenance Systems (IMS) in University of Cincinnati. Four test bearings were

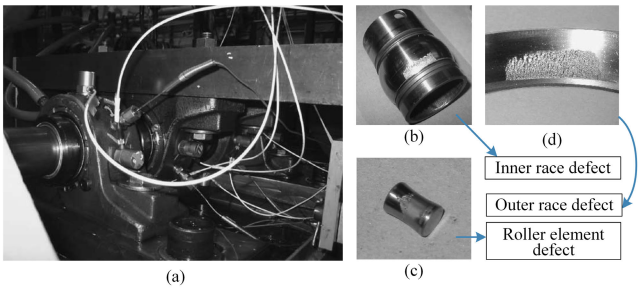


FIGURE 19. The test rig and damage modes of IMS datasets. (a) overview of the test rig. (b) inner race defect in bearing 3, test 1. (c) roller element defect in bearing 4, test 1. (d) outer race defect in bearing 1, test 2.

mounted on one shaft driven by an AC motor and coupled by rub belts. The rotational speed was kept constant at 2000rpm. In this experiment, Rexnord ZA-2115 double row bearings were tested. A radial load of 6000 lbs was added to the shaft and bearings by a spring mechanism. Vibration data were collected every 20 min. The sample rate was 20 kHz and the data length was 20480 points. Three sets of tests were carried out, and the damage modes are illustrated in Fig.19. Guo *et al.* used this dataset in [94] to verify their deep transfer method under the cross-domain scenarios between different bearings.

F. FEMTO DATASET

FEMTO dataset also contains the real data related to accelerated degradation of bearings, which is provided by FEMTO-ST Institute and is generated using an experimental platform called PRONOSTIA. PRONOSTIA testbed is composed of three main parts: a rotating part, a degradation generation part (with a radial force applied on the tested bearing) and a measurement part, as shown in Fig.20. The rotating part includes the asynchronous motor with a gearbox and its two shafts. The loading part is mounted in an aluminum plate, which supports a pneumatic jack, a vertical axis and its lever arm, a force sensor, a clamping ring of the test bearing, a support test bearing shaft, two pillow blocks and their large oversized bearings. The force issued from the pneumatic jack is first amplified by a lever arm, and is then indirectly applied on the external ring of the test ball bearing through its clamping ring.

The measurement part can record two types of signals: vibration (with horizontal and vertical accelerometers) and temperature for monitoring the health of the test bearings. The vibration signals were recorded each 10 seconds with sampling frequency 25.6 kHz. Totally 17 run-to-failure data under 3 different operating conditions were included in FEMTO dataset, but the specific faulty mode of the failure bearing under each test is not declared.

G. PHM09 GEARBOX DATASET

This dataset is for the fault diagnosis of gearbox and is provided by the Prognostics and Health Management Society for the 2009 data challenge competition. The experimental gearbox is two-stage and contains 3 shafts, 4 gears, and 6 bearings. An inside view of the gearbox (with helical

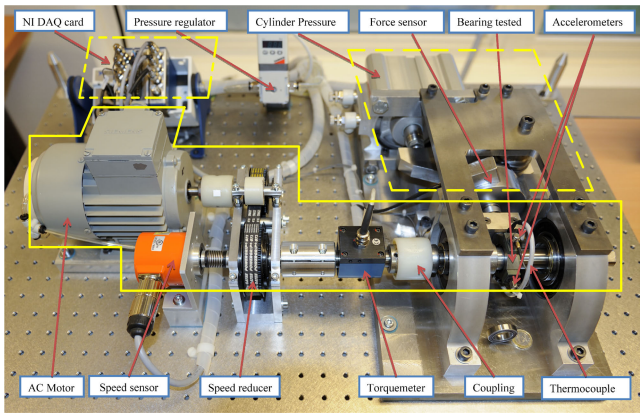


FIGURE 20. Overview of PRONOSTIA experimental platform.

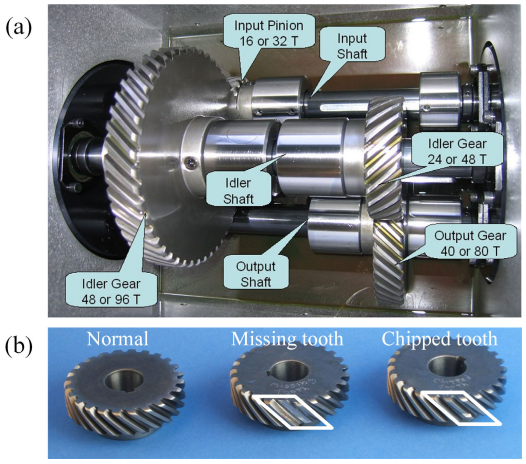


FIGURE 21. Inside view of the gearbox and experimental helical gears. (a) Inside view of the gearbox (with helical gears). (b) the normal and faulty helical gears.

gears) is shown in Fig.21 (a). The dataset provided consists of 560 data sets, in which the gearbox was tested under 5 different speeds (30, 35, 40, 45 and 50 Hz), 2 different loads (Low and High load), and two different gear types (spur gear and helical gear). The measured signals in the experiments consisted of two accelerometer signals along with a tachometer signal. The sampling frequency of the vibration signal was 66.6667 kHz. The experimental data under gear faults (as illustrated in Fig.21 (b)), bearing faults, and combined faults were included.

In [78], [89], this dataset was employed to verify the corresponding transfer methods for gear fault diagnosis.

V. DISCUSSION AND FUTURE DIRECTIONS

In Section III, the research works about cross-domain fault diagnosis have been introduced and summarized. Most of these research works employed transfer learning methods for tackling cross-domain learning problem. Ultimately, for industrial applications, the data-driven fault diagnosis methods should achieve comparable generalization performance under cross-domain scenarios in which the training data and the test data come from different operating conditions or even different same-type machines. Therefore, developing accurate cross-domain fault diagnosis methods is very significant for implementing industrial applications of data-driven fault diagnosis. In the future, the following potential research directions could be considered and studied furtherly:

A. ADDRESSING FAULT DETECTION AND RUL PROGNOSIS TASKS USING TRANSFER LEARNING METHODS

Fault detection, diagnosis, and Remaining Useful Life (RUL) prognosis are three main modules that support a Condition Based Monitoring (CBM) system. The domain-shift, that is the training data and the test data follow different distributions, is a common problem in both of the fault detection, diagnosis, and RUL prognosis tasks. For fault detection, the detection threshold estimated using historical data from other operating conditions or other same-type machines cannot adapt well to the current operating condition or the machine to be detected, and as a result, false alarm occurs. For RUL prognosis, in general, the degradation process of a machine has close relations with operating condition and working environment, such as the life of rolling element bearing is related to its rotational speed, radial load, and lubrication quality. However, obtaining a large amount of life-cycle data for a specific machine under the same operating condition is very difficult, and usually the life-cycle data from other same-type machines or different operating conditions are available for constructing RUL prognostic models. However, the models trained using these life-cycle data cannot perform well when directly applied to the current machine for RUL prognosis because of the potential data discrepancy. Transfer learning may be a feasible way to handle the cross-dataset detection and prognosis tasks that are also domain-shift problems in essence.

However, from the overview of the research works about machinery health monitoring, most of them just focus on addressing fault identification problem using proposed transfer learning or domain adaptation methods. Although the effectiveness of transfer learning has been preliminarily discussed for anomaly detection in [111] and for RUL prognosis in [138], it is meaningful to pay more attention to the fault detection and RUL prognosis tasks of more machines.

B. FOCUSING ON THE CROSS-DOMAIN DIAGNOSIS BETWEEN DIFFERENT BUT SAME-TYPE MACHINES

From the perspective of data acquisition, transfer learning mitigate the data dilemma of the data-driven fault diagnosis methods based on traditional machine learning algorithms. Using transfer learning, the data for training can be collected from a distribution different from the test data, which makes it possible that to build data-driven diagnosis models under practical diagnosis scenarios.

Currently, most of the research works about cross-domain fault diagnosis discuss and validate the performance of their transfer methods on the cross-domain tasks that the source domain and the target domain are from different operating

conditions. However, obtaining data of various faults from the working process of other same-type machines or from the laboratory by fault simulation are more feasible manners than from different operating conditions. So it is more significant that applying the proposed transfer learning methods to the fault diagnosis tasks between the same-type but different machines. Under this circumstance, there are more factors that may influence the distribution discrepancy between domains except for the operating conditions, and it is more difficult to eliminate the domain discrepancy using transfer learning methods.

C. COMBINING THE TRANSFER LEARNING METHOD WITH A PRIORI DIAGNOSTIC KNOWLEDGE

The essence of transfer learning is to mitigate the discrepancy between the training data and the test data caused by the differences of machine models, operating conditions, or other factors in fault diagnosis problem. Transfer learning is a feasible manner but not the only one. In [100], [102], [119] and [122], several diagnosis schemes verified the possibility of tackling cross-domain diagnosis problem through signal preprocessing [100], [102], [122] or extracting domain agnostic features manually [119]. For example, the vibration signals were processed by envelope, angle synchronous average, and normalization in sequence before feeding into the CNN in [102], and the *a priori* knowledge about bearing fault characteristics (envelope) and eliminating differences between the training and test signals (angle synchronous average and normalization) were considered during these steps.

Therefore, combining the transfer learning method with *a priori* diagnostic knowledge may achieve superior generalization performance on cross-domain diagnosis tasks. In addition, this may be an effective manner to address the cross-domain diagnosis tasks when the discrepancy between the source domain and the target domain is pretty large. Meanwhile, in our opinion, the signal pre-processing with considering *a priori* diagnostic knowledge is very necessary for deep learning based diagnosis methods, because it is very difficult that learning consistent abstract concept (such as fault characteristic frequency of rolling bearings) through deep neural networks from the original monitoring signals (such as with different rotational speeds, loads, and sampling frequencies).

D. IMPLEMENTING CROSS-DOMAIN FAULT DIAGNOSIS UNDER MULTIPLE SOURCE SCENARIO

Currently, most of these diagnosis methods based on transfer learning mentioned above only consider one single source in learning process. However, there may be more available data from multiple different operating conditions or other same-type machines in engineering. The fusion of data from multiple sources has two advantages, one is that the description to fault characteristics would be more comprehensive, the other one is that the risk of over-fitting during model training is reduced with more training data. Therefore, learning the general diagnostic knowledge from multiple related source domains and transferring the knowledge to facilitate the target tasks is also a crucial issue.

VI. CONCLUSION

In this paper, we have provided a systematic overview of cross-domain fault diagnosis, most of which tackle the cross-domain diagnosis tasks using transfer learning methods. Transfer learning aims to leverage knowledge from a source domain to promote the generalization performance in a related target domain. Facing the diagnosis tasks in which the training data and the test data may be drawn from different potential distributions, the transfer learning methods were employed in most recently. The research works about cross-domain diagnosis were summarized from three viewpoints. First, the related research works were introduced by dividing them into five different research motivations and four different problem settings. Second, the specific cross-domain diagnosis strategies were classified into four different categories: traditional transfer approaches, deep transfer approaches, adversarial-based approaches, and other approaches. Third, the cross-domain diagnosis applications were summarized. It was found that bearing and gearbox were the two most widely research objects among these research works.

In addition, open-source datasets for machinery fault diagnosis were also introduced to help readers to start the study of cross-domain fault diagnosis. Finally, some potential research trends were also given: 1) addressing fault detection and RUL prognosis tasks using transfer learning methods, 2) focusing on the cross-domain diagnosis between different but same-type machines, 3) combining the transfer learning method with *a priori* diagnostic knowledge, and 4) implementing cross-domain fault diagnosis under multiple source scenario.

REFERENCES

[1] J. Lee, F. J. Wu, W. Y. Zhao, M. Ghaffari, L. X. Liao, and D. Siegel, "Prognostics and health management design for rotary machinery systems-reviews, methodology and applications," *Mech. Syst. Signal Process.*, vol. 42, nos. 1–2, pp. 314–334, Jan. 2014.

[2] K.-L. Tsui, N. Chen, Q. Zhou, Y. Hai, and W. Wang, "Prognostics and health management: A review on data driven approaches," *Math. Problems Eng.*, vol. 2015, May 2015, Art. no. 793161.

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

[4] A. K. S. Jardine, D. Lin, and D. Banjevic, "A review on machinery diagnostics and prognostics implementing condition-based maintenance," *Mech. Syst. Signal Process.*, vol. 20, no. 7, pp. 1483–1510, 2006.

[5] R. Liu, B. Yang, E. Zio, and X. Chen, "Artificial intelligence for fault diagnosis of rotating machinery: A review," *Mech. Syst. Signal Process.*, vol. 108, pp. 33–47, Aug. 2018.

[6] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao, "Deep learning and its applications to machine health monitoring," *Mech. Syst. Signal Process.*, vol. 115, pp. 213–237, Jan. 2019.

[7] S. Zhang, S. Zhang, B. Wang, and T. G. Habetler, "Machine learning and deep learning algorithms for bearing fault diagnostics—A comprehensive review," Jan. 2019, *arXiv:1901.08247*. [Online]. Available: https://arxiv.org/abs/1901.08247

[8] X. Zhang, W. Chen, B. Wang, and X. Chen, "Intelligent fault diagnosis of rotating machinery using support vector machine with ant colony algorithm for synchronous feature selection and parameter optimization," *Neurocomputing*, vol. 167, pp. 260–279, Nov. 2015.

[9] Z. Wang, Q. Zhang, J. Xiong, M. Xiao, G. Sun, and J. He, "Fault diagnosis of a rolling bearing using wavelet packet denoising and random forests," *IEEE Sensors J.*, vol. 17, no. 17, pp. 5581–5588, Sep. 2017.

[10] Z. Su, B. Tang, J. Ma, and L. Deng, "Fault diagnosis method based on incremental enhanced supervised locally linear embedding and adaptive nearest neighbor classifier," *Measurement*, vol. 48, pp. 136–148, Feb. 2014.

[11] F. Jia, Y. G. Lei, J. Lin, X. Zhou, and N. Lu, "Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data," *Mech. Syst. Signal Process.*, vols. 72–73, pp. 303–315, May 2016.

[12] H. Shao, H. Jiang, F. Wang, and Y. Wang, "Rolling bearing fault diagnosis using adaptive deep belief network with dual-tree complex wavelet packet," *ISA Trans.*, vol. 69, pp. 187–201, Jul. 2017.

[13] O. Janssens, V. Slavkovikj, B. Vervisch, K. Stockman, M. Loccufier, S. Verstockt, R. Van de Walle, and S. Van Hoecke, "Convolutional neural network based fault detection for rotating machinery," *J. Sound Vib.*, vol. 377, pp. 331–345, Sep. 2016.

[14] L. Wen, X. Li, L. Gao, and Y. Zhang, "A new convolutional neural network-based data-driven fault diagnosis method," *IEEE Trans. Ind. Electron.*, vol. 65, no. 7, pp. 5990–5998, Jul. 2018.

[15] Z. Feng, M. Liang, and F. Chu, "Recent advances in time–frequency analysis methods for machinery fault diagnosis: A review with application examples," *Mech. Syst. Signal Process.*, vol. 38, no. 1, pp. 165–205, Jul. 2013.

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

[17] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[18] J. Zhang, W. Li, P. Ogunbona, and D. Xu, "Recent advances in transfer learning for cross-dataset visual recognition: A problem-oriented perspective," *ACM Comput. Surv.*, vol. 52, no. 1, pp. 1–38, Feb. 2019.

[19] L. Zhang, "Transfer adaptation learning: A decade survey," Mar. 2019, *arXiv:1903.0468*. [Online]. Available: https://arxiv.org/abs/1903.0468

[20] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 53–69, May 2015.

[21] W. Mei and D. Weihong, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, Jul. 2018.

[22] G. Csurka, "Domain adaptation for visual applications: A comprehensive survey," Feb. 2017, *arXiv:1702.05374*. [Online]. Available: https://arxiv.org/abs/1702.05374

[23] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.*, vol. 2018, pp. 1–13, Jul. 2018.

[24] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, Aug. 2018.

[25] J. Huang, A. Smola, A. Gretton, K. M. Borgwardt, and B. Schölkopf, "Correcting sample selection bias by unlabeled data," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2006, pp. 601–608.

[26] J. Jiang and C. Zhai, "Instance weighting for domain adaptation in NLP," in *Proc. ACL*, Prague, Czech Republic, Jun. 2007, pp. 264–271.

[27] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for transfer learning," in *Proc. Int. Conf. Mach. Learn.*, Corvallis, OR, USA, 2007, pp. 193–200.

[28] R. Wang, M. Utiyama, L. Liu, K. Chen, and E. Sumita, "Instance weighting for neural machine translation domain adaptation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Copenhagen, Denmark, Sep. 2017, pp. 1482–1488.

[29] E. Zhong, W. Fan, J. Peng, K. Zhang, J. Ren, D. Turaga, and O. Verscheure, "Cross domain distribution adaptation via kernel mapping," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Paris, France, Jun. 2009, pp. 1027–1036.

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

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

[32] S. Herath, M. Harandi, and F. Porikli, "Learning an invariant Hilbert space for domain adaptation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 3956–3965.

[33] B. Sun and K. Saenko, "Subspace distribution alignment for unsupervised domain adaptation," in *Proc. Brit. Mach. Vis. Conf.*, Swansea, U.K., Sep. 2015, pp. 24.1–24.10.

[34] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2066–2073.

[35] R. Gopalan, R. Li, and R. Chellappa, "Domain adaptation for object recognition: An unsupervised approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 999–1006.

[36] Q. Qiu, V. M. Patel, P. Turaga, and R. Chellappa, "Domain adaptive dictionary learning," in *Proc. Eur. Conf. Comput. Vis.*, Florence, Italy, Oct. 2012, pp. 631–645.

[37] Q. Qiu and R. Chellappa, "Compositional dictionaries for domain adaptive face recognition," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5152–5165, Dec. 2015.

[38] I.-H. Jhuo, D. Liu, D. T. Lee, and S.-F. Chang, "Robust visual domain adaptation with low-rank reconstruction," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2168–2175.

[39] M. Shao, D. Kit, and Y. Fu, "Generalized transfer subspace learning through low-rank constraint," *Int. J. Comput. Vis.*, vol. 109, nos. 1–2, pp. 74–93, 2014.

[40] 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, Mar. 2012.

[41] J. Yang, R. Yan, and A. G. Hauptmann, "Cross-domain video concept detection using adaptive SVMs," in *Proc. ACM Int. Conf. Multimedia*, Augsburg, Germany, Sep. 2007, pp. 188–197.

[42] L. Duan, I. W. Tsang, and D. Xu, "Domain transfer multiple kernel learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 465–479, Mar. 2012.

[43] L. Bruzzone and M. Marconcini, "Domain adaptation problems: A DASVM classification technique and a circular validation strategy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 770–787, May 2010.

[44] M. Long, J. Wang, G. Ding, S. J. Pan, and P. S. Yu, "Adaptation regularization: A general framework for transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 5, pp. 1076–1089, May 2014.

[45] A. Ramachandran, S. Gupta, S. Rana, and S. Venkatesh, "Information-theoretic transfer learning framework for Bayesian optimisation," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, Dublin, Ireland, Sep. 2018, pp. 827–842.

[46] M. Gönen and A. A. Margolin, "Kernelized Bayesian transfer learning," in *Proc. Conf. AAAI Artif. Intell.*, Québec City, QC, Canada, Jul. 2014, pp. 1831–1839.

[47] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[48] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 3908–3916.

[49] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 645–657, Feb. 2017.

[50] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, Lille, France, Jul. 2015, pp. 97–105.

[51] M. Long, Y. Cao, Z. Cao, J. Wang, and M. I. Jordan, "Transferable representation learning with deep adaptation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published. doi: 10.1109/TPAMI.2018.2868685.

[52] F. Zhuang, X. Cheng, P. Luo, S. J. Pan, and Q. He, "Supervised representation learning: Transfer learning with deep autoencoders," in *Proc. Int. Conf. Artif. Intell.*, Buenos Aires, Argentina, Jul. 2015, pp. 4119–4125.

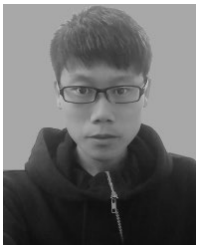
[53] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Unified deep supervised domain adaptation and generalization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, Oct. 2017, pp. 5716–5726.

[54] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2014, pp. 2672–2680.

[55] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 3722–3731.

[56] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," Nov. 2016, *arXiv:1611.02200*. [Online]. Available: https://arxiv.org/abs/1611.02200
[57] W. Hong, Z. Wang, M. Yang, and J. Yuan, "Conditional generative adversarial network for structured domain adaptation," in *Proc. IEEE/CVF Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 1335–1344.
[58] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 59, pp. 1–35, 2016.
[59] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 2962–2971.
[60] 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.*, Ottawa, ON, Canada, Jun. 2016, pp. 1–6.
[61] L. Duan, J. Xie, K. Wang, and J. Wang, "Gearbox diagnosis based on auxiliary monitoring datasets of different working conditions," (in Chinese), *J. Vib. Shock*, vol. 36, no. 10, pp. 104–116, 2017.
[62] Z. Tong, W. Li, B. Zhang, F. Jiang, and G. B. Zhou, "Bearing fault diagnosis under variable working conditions based on domain adaptation using feature transfer learning," *IEEE Access*, vol. 6, pp. 76187–76197, Nov. 2018.
[63] W. Qian, S. Li, and J. Wang, "A new transfer learning method and its application on rotating machine fault diagnosis under variant working conditions," *IEEE Access*, vol. 6, pp. 69907–69917, 2018.
[64] W. Qian, S. Li, P. Yi, and K. Zhang, "A novel transfer learning method for robust fault diagnosis of rotating machines under variable working conditions," *Measurement*, vol. 138, pp. 514–525, May 2019.
[65] 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.
[66] 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: https://arxiv.org/abs/1903.06753
[67] S. Kang, M. Hu, Y. Wang, J. Xie, and V. I. Mikulovich, "Fault diagnosis method of a rolling bearing under variable working conditions based on feature transfer learning," (in Chinese), *Proc. CSEE*, vol. 39, no. 3, pp. 764–773, Feb. 2019.
[68] D. Xiao, Y. Huang, C. Qin, Z. Liu, Y. Li, and C. Liu, "Transfer learning with convolutional neural networks for small sample size problem in machinery fault diagnosis," *Proc. Inst. Mech. Eng. C, J. Mech. Eng. Sci.*, vol. 233, no. 14, pp. 5131–5143, Jul. 2019.
[69] M. J. Hasan and J.-M. Kim, "Bearing fault diagnosis under variable rotational speeds using stockwell transform-based vibration imaging and transfer learning," *Appl. Sci.*, vol. 8, no. 12, p. 2357, Nov. 2018.
[70] J. J. Hasan, M. M. M. Islam, and J.-M. Kim, "Acoustic spectral imaging and transfer learning for reliable bearing fault diagnosis under variable speed conditions," *Measurement*, vol. 138, pp. 620–631, May 2019.
[71] J. Xie, J. Wang, R. Zhao, L. Duan, and K. Wang, "Application of transfer factor analysis in gearbox fault diagnosis under various working condition," (in Chinese), *J. Electron. Meas. Instrum.*, vol. 30, no. 4, pp. 534–541, Apr. 2016.
[72] C. Chen, Z. Li, J. Yang, and B. Liang, "A cross domain feature extraction method based on transfer component analysis for rolling bearing fault diagnosis," in *Proc. CCDC*, Chongqing, China, May 2017, pp. 5622–5626.
[73] W. Zhang, G. Peng, C. Li, Y. Chen, and Z. Zhang, "A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals," *Sensors*, vol. 17, no. 2, p. 425, 2017.
[74] W. Zhang, C. Li, G. Peng, Y. Chen, and Z. Zhang, "A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load," *Mech. Syst. Signal Process.*, vol. 100, pp. 439–453, Feb. 2018.
[75] B. Zhang, W. Li, Z. Tong, and M. Zhang, "Bearing fault diagnosis under varying working condition based on domain adaptation," Jul. 2017, *arXiv:1707.09890*. [Online]. Available: https://arxiv.org/abs/1707.09890
[76] B. Zhang, W. Li, J. Hao, X.-L. Li, and M. Zhang, "Adversarial adaptive 1-D convolutional neural networks for bearing fault diagnosis under varying working condition," May 2018, *arXiv:1805.00778*. [Online]. Available: https://arxiv.org/abs/1805.00778
[77] Z. Tong, W. Li, B. Zhang, and M. Zhang, "Bearing fault diagnosis based on domain adaptation using transferable features under different working conditions," *Shock Vib.*, vol. 2018, Jun. 2018, Art. no. 6714520.
[78] B. Zhang, W. Li, X.-L. Li, and S.-K. Ng, "Intelligent fault diagnosis under varying working conditions based on domain adaptive convolutional neural networks," *IEEE Access*, vol. 6, pp. 66367–66384, 2018.
[79] K. Wang and B. Wu, "Power equipment fault diagnosis model based on deep transfer learning with balanced distribution adaptation," in *Proc. Int. Conf. Adv. Data Mining Appl.*, Nanjing, China, Nov. 2018, pp. 178–188.
[80] C. Chen, F. Shen, and Y. R. Qiang, "Enhanced least squares support vector machine-based transfer learning strategy for bearing fault diagnosis," (in Chinese), *Chin. J. Sci. Instrum.*, vol. 38, no. 01, pp. 33–40, Jan. 2017.
[81] F. Shen, C. Chen, R. Yan, and R. X. Gao, "Bearing fault diagnosis based on SVD feature extraction and transfer learning classification," in *Proc. IEEE Int. Conf. Prognostics Health Manage.*, Beijing, China, Oct. 2015, pp. 1–6.
[82] X. Li, W. Zhang, and Q. Ding, "Cross-domain fault diagnosis of rolling element bearings using deep generative neural networks," *IEEE Trans. Ind. Electron.*, vol. 66, no. 7, pp. 5525–5534, Jul. 2019.
[83] F. Shen, C. Chen, and R. Yan, "Application of SVD and transfer learning strategy on motor fault diagnosis," (in Chinese), *J. Vib. Eng.*, vol. 30, no. 1, pp. 118–126, Feb. 2017.
[84] X. Li, W. Zhang, Q. Ding, and J.-Q. Sun, "Multi-layer domain adaptation method for rolling bearing fault diagnosis," *Signal Process.*, vol. 157, pp. 180–197, Apr. 2019.
[85] T. Han, C. Liu, W. Yang, and D. Jiang, "Deep transfer network with joint distribution adaptation: A new intelligent fault diagnosis framework for industry application," Apr. 2018, *arXiv:1804.07265*. [Online]. Available: https://arxiv.org/abs/1804.07265
[86] Q. Wang, G. Michau, and O. Fink, "Domain adaptive transfer learning for fault diagnosis," May 2019, *arXiv:1905.06004*. [Online]. Available: https://arxiv.org/abs/1905.06004
[87] G. Xu, M. Liu, Z. Jiang, W. Shen, and C. Huang, "Online fault diagnosis method based on transfer convolutional neural networks," *IEEE Trans. Instrum. Meas.*, to be published. doi: 10.1109/TIM.2019.2902003.
[88] 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.
[89] 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.
[90] 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.
[91] D. Chen, S. Yang, and F. Zhou, "Incipient fault diagnosis based on DNN with transfer learning," in *Proc. ICCAIS*, Hangzhou, China, Oct. 2018, pp. 1–6.
[92] H. Kim and B. D. Youn, "A new parameter repurposing method for parameter transfer with small dataset and its application in fault diagnosis of rolling element bearings," *IEEE Access*, vol. 7, pp. 46917–46930, Mar. 2019.
[93] M. J. Hasan, M. Sohaib, and J.-M. Kim, "1D CNN-based transfer learning model for bearing fault diagnosis under variable working conditions," in *Proc. CIIS*, Gadong, Brunei, Nov. 2018, pp. 13–23.
[94] L. Guo, Y. Lei, S. Xing, T. Yan, and N. Li, "Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data," *IEEE Trans. Ind. Electron.*, vol. 66, no. 9, pp. 7316–7325, Sep. 2018.
[95] A. Zhang and X. Gao, "Supervised dictionary-based transfer subspace learning and applications for fault diagnosis of sucker rod pumping systems," *Neurocomputing*, vol. 338, pp. 293–306, Apr. 2019.
[96] H. Zheng, R. Wang, Y. Yang, Y. Li, and M. Xu, "Intelligent fault identification based on multi-source domain generalization towards actual diagnosis scenario," *IEEE Trans. Ind. Electron.*, to be published. doi: 10.1109/TIE.2019.2898619.
[97] B. Yang, Y. Lei, F. Jia, and S. Xing, "A transfer learning method for intelligent fault diagnosis from laboratory machines to real-case machines," in *Proc. SDPC*, Xi'an, China, Aug. 2018, pp. 35–40.
[98] B. Yang, Y. Lei, F. Jia, and S. Xing, "An intelligent fault diagnosis approach based on transfer learning from laboratory bearings to locomotive bearings," *Mech. Syst. Signal Process.*, vol. 122, pp. 692–706, May 2019.

[99] C. Yuanhang, P. Gaoliang, X. Chaohao, Z. Wei, L. Chuanhao, and L. Shaohui, "ACDIN: Bridging the gap between artificial and real bearing damages for bearing fault diagnosis," *Neurocomputing*, vol. 294, pp. 61–71, Jun. 2018.
[100] H. Oh, J. H. Jung, B. C. Jeon, and B. D. Youn, "Scalable and unsupervised feature engineering using vibration-imaging and deep learning for rotor system diagnosis," *IEEE Trans. Ind. Electron.*, vol. 65, no. 4, pp. 3539–3549, Apr. 2018.
[101] 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, Jan. 2019.
[102] C. Sobie, C. Freitas, and M. Nicolai, "Simulation-driven machine learning: Bearing fault classification," *Mech. Syst. Signal Process.*, vol. 99, pp. 403–419, Jan. 2018.
[103] P. Cao, S. Zhang, and J. Tang, "Preprocessing-free gear fault diagnosis using small datasets with deep convolutional neural network-based transfer learning," *IEEE Access*, vol. 6, pp. 26241–26253, May 2018.
[104] S. Shao, S. McAleer, R. Yan, and P. Baldi, "Highly accurate machine fault diagnosis using deep transfer learning," *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2446–2455, Apr. 2019.
[105] L. Wen, X. Li, and L. Gao, "A transfer convolutional neural network for fault diagnosis based on ResNet-50," *Neural Comput. Appl.*, to be published. doi: 10.1007/s00521-019-04097-w.
[106] M. Hemmer, H. Van Khang, K. Robbersmyr, T. Waag, and T. Meyer, "Fault classification of axial and radial roller bearings using transfer learning through a pretrained convolutional neural network," *Designs*, vol. 2, no. 4, p. 56, Dec. 2018.
[107] D. Chen, S. Yang, and F. Zhou, "Transfer learning based fault diagnosis with missing data due to multi-rate sampling," *Sensors*, vol. 19, no. 8, p. 1826, Apr. 2019.
[108] Z. Shi-Sheng, F. Song, and L. Lin, "A novel gas turbine fault diagnosis method based on transfer learning with CNN," *Measurement*, vol. 137, pp. 435–453, Apr. 2019.
[109] T. Han, C. Liu, W. Yang, and D. Jiang, "Learning transferable features in deep convolutional neural networks for diagnosing unseen machine conditions," *ISA Trans.*, to be published. doi: 10.1016/j.isatra.2019.03.017.
[110] W. Chunfeng, L. Zheng, Z. Jun, and W. Wei, "Heterogeneous transfer learning based on stack sparse auto-encoders for fault diagnosis," in *Proc. CAC*, Xi'an, China, Nov. 2018, pp. 4277–4281.
[111] A. G. Mahyari, "Domain adaptation in robot fault diagnostic systems," Sep. 2018, *arXiv:1809.08626*. [Online]. Available: https://arxiv.org/abs/1809.08626
[112] X. Li, W. Zhang, and Q. Ding, "A robust intelligent fault diagnosis method for rolling element bearings based on deep distance metric learning," *Neurocomputing*, vol. 310, pp. 77–95, Oct. 2018.
[113] X. Wang, H. He, and L. Li, "A hierarchical deep domain adaptation approach for fault diagnosis of power plant thermal system," *IEEE Trans. Ind. Informat.*, to be published. doi: 10.1109/TII.2019.2899118.
[114] Y. Xie and T. Zhang, "A transfer learning strategy for rotation machinery fault diagnosis based on cycle-consistent generative adversarial networks," in *Proc. CAC*, Xi'an, China, Nov./Dec. 2018, pp. 1309–1313.
[115] S. Lu, T. Sirojan, B. T. Phung, D. Zhang, and E. Ambikairajah, "DA-DCGAN: An effective methodology for DC series arc fault diagnosis in photovoltaic systems," *IEEE Access*, vol. 7, pp. 45831–45840, Apr. 2019.
[116] Z. Zhu, G. Peng, Y. Chen, and H. Gao, "A convolutional neural network based on a capsule network with strong generalization for bearing fault diagnosis," *Neurocomputing*, vol. 323, pp. 62–75, Jan. 2019.
[117] L. Wen, X. Li, and L. Gao, "A new two-level hierarchical diagnosis network based on convolutional neural network," *IEEE Trans. Instrum. Meas.*, to be published. doi: 10.1109/TIM.2019.2896370.
[118] L. Wen, L. Gao, and X. Li, "A new snapshot ensemble convolutional neural network for fault diagnosis," *IEEE Access*, vol. 7, pp. 32037–32047, Mar. 2019.
[119] T. Han, C. Liu, L. Wu, S. Sarkar, and D. Jiang, "An adaptive spatiotemporal feature learning approach for fault diagnosis in complex systems," *Mech. Syst. Signal Process.*, vol. 117, pp. 170–187, Feb. 2019.
[120] T. Han, C. Liu, W. Yang, and D. Jiang, "A novel adversarial learning framework in deep convolutional neural network for intelligent diagnosis of mechanical faults," *Knowl.-Based Syst.*, vol. 165, pp. 474–487, Feb. 2019.
[121] P. Peng and J. Wang, "NOSCNN: A robust method for fault diagnosis of RV reducer," *Measurement*, vol. 138, pp. 652–658, May 2019.
[122] D. Wei, K. Wang, S. Heyns, and M. J. Zuo, "Convolutional neural networks for fault diagnosis using rotating speed normalized vibration," in *Proc. Int. Conf. Condition Monitor. Mach. Non-Stationary Oper.*, Santander, Spain, Jun. 2018, pp. 67–76.
[123] 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, Mar. 2019.
[124] W. Qian, S. Li, J. Wang, Y. Xin, and H. Ma, "A new deep transfer learning network for fault diagnosis of rotating machine under variable working conditions," in *Proc. PHM-Chongqing*, Oct. 2018, pp. 1010–1016.
[125] Y. Li, N. Wang, J. Shi, X. Hou, and J. Liu, "Adaptive batch normalization for practical domain adaptation," *Pattern Recognit.*, vol. 80, pp. 109–117, Aug. 2018.
[126] L. Duan, X. Wang, M. Xie, Z. Yuan, and J. Wang, "Auxiliary-model-based domain adaptation for reciprocating compressor diagnosis under variable conditions," *J. Intell. Fuzzy Syst.*, vol. 34, no. 6, pp. 3595–3604, Jun. 2018.
[127] M. Zhang, D. Wang, W. Lu, J. Yang, Z. Li, and B. Liang, "A deep transfer model with wasserstein distance guided multi-adversarial networks for bearing fault diagnosis under different working conditions," *IEEE Access*, vol. 7, pp. 65303–65318, May 2019.
[128] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.
[129] Y. Lei, B. Yang, Z. Du, and N. Lu, "Deep transfer diagnosis method for machinery in big data era," (in Chinese), *J. Mech. Eng.*, vol. 55, no. 7, pp. 1–8, Apr. 2019.
[130] Y. Pan, F. Mei, H. Miao, J. Zheng, K. Zhu, and H. Sha, "An approach for HVCB mechanical fault diagnosis based on a deep belief network and a transfer learning strategy," *J. Elect. Eng. Technol.*, vol. 14, no. 1, pp. 407–419, Jan. 2019.
[131] Case Western Reserve University Bearing Data Center. *CWRU Dataset*. Accessed: May 2019. [Online]. Available: http://csegroups.case.edu/bearingdatacenter/pages/download-data-file
[132] Society For Machinery Failure Prevention Technology. *MFPT Dataset*. Accessed: May 2019. [Online]. Available: https://mfpt.org/fault-data-sets/
[133] C. Lessmeier, J. Kimotho, D. Zimmer, and W. Sextro, "Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification," in *Proc. Eur. Conf. Prognostics Health Manage. Soc.*, Bilbao, Spain, 2016, pp. 1–17.
[134] A. P. Daga, A. Fasana, S. Marchesiello, and L. Garibaldi, "The politecnico di torino rolling bearing test rig: Description and analysis of open access data," *Mech. Syst. Signal Process.*, vol. 120, pp. 252–273, Apr. 2019.
[135] H. Qiu, J. Lee, J. Lin, and G. Yu, "Robust performance degradation assessment methods for enhanced rolling element bearing prognostics," *Adv. Eng. Inform.*, vol. 17, nos. 3–4, pp. 127–140, Jul. 2003.
[136] P. Nectoux, R. Gouriveau, K. Medjaher, E. Ramasso, B. Chebel-Morello, N. Zerhouni, and C. Varnier, "PRONOSTIA: An experimental platform for bearings accelerated degradation tests," in *Proc. IEEE Int. Conf. Prognostics Health Manage.*, CO, USA, Jun. 2012, pp. 1–8.
[137] PHM Society. *PHM09 Gearbox Dataset*. Acessed: May 2019. [Online]. Available: https://www.phmsociety.org/competition/PHM/09
[138] C. Sun, M. Ma, Z. Zhao, S. Tian, R. Yan, and X. Chen, "Deep transfer learning based on sparse autoencoder for remaining useful life prediction of tool in manufacturing," *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2416–2425, Apr. 2019.

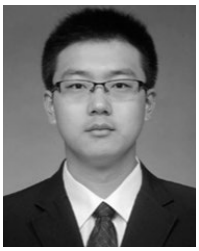


HUAILIANG ZHENG received the B.S. and M.S. degrees in mechanics from the Harbin Institute of Technology, Harbin, China, in 2013 and 2015, respectively, where he is currently pursuing the Ph.D. degree with Deep Space Exploration Research Center.

His research interests include fault diagnosis of machinery, intelligent fault diagnosis method, and transfer learning.

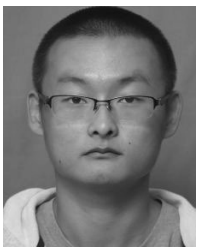
RIXIN WANG received the B.E. degree in computer science from the Harbin University of Science and Technology, Harbin, China, in 1985, and the M.E. degree in computer science and the Ph.D. degree in spacecraft design from the Harbin Institute of Technology, Harbin, in 1991 and 2003, respectively, where he is currently an Associate Professor with the Department of Engineering Mechanics.

His research interests include fault detection and diagnosis for machinery and spacecraft.



YUANTAO YANG received the B.S. and M.S. degrees in mechanics from the Harbin Institute of Technology, Harbin, China, in 2013 and 2015, respectively, where he is currently pursuing the Ph.D. degree with Deep Space Exploration Research Center.

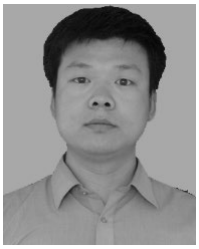
His research interests include fault diagnosis of machinery, intelligent fault diagnosis method, and deep learning.



JIANCHENG YIN received the B.S. degree in mechanics from the Harbin Institute of Technology, Harbin, China, in 2014, where he is currently pursuing the Ph.D. degree in mechanics with Deep Space Exploration Research Center.

His research interests include fault diagnosis of machinery, residual life prediction method, and signal processing.

YONGBO LI (M'16) received the master's degree from Harbin Engineering University, Harbin, China, in 2012, and the Ph.D. degree in general mechanics from the Harbin Institute of Technology, Harbin, in 2017. He is currently an Associate Professor with the School of Aeronautics, Northwestern Polytechnical University, China. Prior to joining Northwestern Polytechnical University, in 2017, he was a Visiting Student with the University of Alberta, Edmonton, AB, Canada. He was the Session Chair at the international conference of PHM 2018. He also served as a Guest Editor for the *Advances in Mechanical Engineering*.



YUQING LI received the B.E. degree in mechanical design manufacturing and automation and the M.E. and Ph.D. degrees in general mechanics from the Harbin Institute of Technology, Harbin, China, in 2002, 2004, and 2008, respectively, where he is currently an Associate Professor.

His main research interests include planning and scheduling of spacecraft, and spacecraft fault detection and diagnosis.

MINQIANG XU received the B.E. degree in electronics from Peking University, Beijing, China, in 1983, the M.E. degree in nuclear physics from Northeast Normal University, Changchun, China, in 1989, and the Ph.D. degree in general and fundamental mechanics from the Harbin Institute of Technology, Harbin, China, in 1999, where he has been a Professor, since 2000.

His research interests include machinery and spacecraft fault detection and diagnosis, signal processing, and space debris modeling.