Knowledge Transfer for Rotary Machine Fault Diagnosis

Ruqiang Yan#, *Senior Member, IEEE,* Fei Shen#, Chuang Sun, and Xuefeng Chen

***Abstract*—This paper intends to provide an overview on recent development of knowledge transfer for rotary machine fault diagnosis (RMFD) by using different transfer learning techniques. After brief introduction of parameter-based, instance-based, feature-based and relevance-based knowledge transfer, the applications of knowledge transfer in RMFD are summarized from four categories: transfer between multiple working conditions, transfer between multiple locations, transfer between multiple machines, and transfer between multiple fault types. Case studies on four datasets including gears, bearing, and motor faults verified effectiveness of knowledge transfer on improving diagnostic accuracy. Meanwhile, research trends on transfer learning in the field of RMFD are discussed.**

***Index Terms*—Transfer learning, rotary machine fault diagnosis, multiple working conditions, multiple locations, multiple machines, multiple fault types.**

ABBREVIATIONS

2SW-MDA	Two Stage Weighting framework for Multi-source Domain Adaptation
ADDA	Adversarial Discriminative Domain Adaptation
AFS	Angle-Frequency domain Synchronous analysis
AKT-MIL	Adaptive Knowledge Transfer for Multiple Instances Learning
ARC-t	Asymmetric Regularized Cross-domain transformation
ARM	Association Rule Mining
ARTL	Adaptation Regularization based Transfer Learning
BBR	Build-in Bowed Rotor
BE	Bearing End
BF	Ball Fault
BFB	Build-in Fault Bearing
BRB	Broken Rotor Bar
BRU	Build-in Rotor Unbalance
CDA	Conditional Distribution Adaptation
CNNs	Convolutional Neural Networks
CORAL	Correlation Alignment
CPD	Conditional Probability Distribution
CP-MDA	Conditional Probability based Multi-source Domain Adaptation
CTF	Chipped Tooth Fault
CWT-BSS	Continuous Wavelet Transform-based Blind Source Separation
DAFD	Domain Adaptation Fault Diagnosis
DAM	Domain Adaptation Machine
DAN	Deep Adaptation Networks
DASVM	Domain Adaptation Support Vector Machine
DATF	Domain Adaptation using Transferable Features
DCTLN	Deep Convolutional Transfer Learning Network
DE	Drive End
DISVM	Domain Independent Support Vector Machine
DSM	Domain Selection Machine
DTM	Deep Transfer via Markov logic
DTNs	Domain Transfer Networks
EKF	Extended Kalman Filter
FE	Fan End
FTNN	Feature-based Transfer Neural Network
GFK	Geodesic Flow Kernel
HDDA	Hierarchical Deep Domain Adaptation
HEA	Healthy
HSIC	Hilbert-Schmidt Independence Criterion
IF	Inner race Fault
JDA	Joint Domain Adaptation
JHDA	Joint Hierarchical Domain Adaptation
KNN	K Nearest Neighbor
LSDT	Latent Sparse Domain Transfer
LSSVM	Least Square Support Vector Machine
MACNN	Multi-layer Adaptation CNN
MDA	Marginal Distribution Adaptation
MDL	Minimum Description Length
ME	Motor End
MK-MMD	Multiple Kernel variant of Maximum Mean Difference
MLNs	Markov Logic Networks
MMKT	Multi-Model Knowledge Transfer
MITL	Multiple Instance Transfer Learning
ML	Manifold Learning
MR	Manifold Regularization
MSST	Matching Synchronous Squeezing Transform
MTF	Missing Tooth Fault

Manuscript received -; revised -; accepted. Date of publication -; date of current version -. This work was supported by the National Natural Science Foundation of China under Grant 51575102 and the Scientific Research Foundation of Graduate School of Southeast University under Grant YBPY1887 (Corresponding author: Ruqiang Yan.)

Ruqiang Yan, Chuang Sun and Xuefeng Chen are with the School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, 710049 China (yanruqiang@xjtu.edu.cn, ch.sun@xjtu.edu.cn, and chenxf@xjtu.edu.cn).

Fei Shen is with the School of Instrument Science and Engineering, Southeast University, Nanjing, 210096 China (sfseu0311@163.com).

#These authors contributed equally to this work.

NC	Normal Condition
NMTF	Nonnegative Matrix Tri-factorizations
NSAE-LCN	Local Connection Network with Normalized Sparse Auto-Encoder
OF	Outer race Fault
PG	Planetary Gearbox
PSO	Particle Swarm Optimization
RAP	Relational Adaptive Bootstrapping
RG	Reduction Gearbox
RMFD	Rotary Machine Fault Diagnosis
RTNs	Residual Transfer Networks
SA	Subspace Alignment
SAE	Sparse Auto-Encoder
SDA	Stacked Denoising Autoencoder
SEMG	Surface Electromyography
SFA	Spectral Feature Alignment
SGD	Stochastic Gradient Descent
SMITL	Selective Multiple Instance Transfer Learning
SPF	Selective Parameter Freezing
SST	Synchronous Squeezing Transform
SVD	Singular Value Decomposition
SVM	Support Vector Machine
SWDF	Stator Winding Fault
SWF	Surface Wear Fault
TAMAR	Transfer via Automatic Mapping And Revision
TCA	Transfer Component Analysis
TCCHC	Transfer Compact Coding for Hyper plane Classifier
TCF	Tooth Crack Fault
TCNN	Transfer Convolutional Neural Networks
TFA	Transfer Factor Analysis
TFM	Target Function Modeling
TICNN	Convolution Neural Networks with Training Interference
TJM	Transfer Joint Matching
TL	Transfer Learning
VBHMM	Variational Bayesian Hidden Markov Model
WD-DTL	Wasserstein Distance based Deep Transfer Learning

I. Introduction

TRANSFER learning (TL), which is motivated by the fact that people can flexibly apply the knowledge learned before to solve new problem faster or better [1-7], has been increasingly applied in the rotary machine fault diagnosis (RMFD) in recent years. The integration of knowledge transfer into machine fault diagnosis [8-13] has become an important and hot topic to address due to the following reasons:

- ❖ Over the last 10 years, TL techniques have achieved successes in some fields of engineering, such as image/text processing [14,15], biological recognition [16,17];
- ❖ There is an increasing need for high-accuracy, real-time, low-cost RMFD products, whose function will not be affected by external factors, such as varying working conditions of the rotary machine [18,19];
- ❖ The application of deep neural network models in RMFD is more and more popular, such as convolutional neural networks (CNNs) [13,20].

Although Pan [21] and Weiss [4] provided a detailed survey on transfer learning in 2010 and 2016, respectively, they focused on the general introduction and engineering applications of knowledge transfer. Taylor [22] also summarized transfer learning in reinforcement learning domains in 2009. In addition, Zheng [23] presented a systematic introduction of the research works about cross-domain fault diagnosis. Meanwhile, they gave a comprehensive summary and several future research directions of the open-source datasets for facilitating readers to start studies of cross-domain fault diagnosis. On the whole, TL-based RMFD is almost to be summarized up to now because it only lasts about five years from its first application in this area [24,25]. In fact, some specific engineering questions need to be solved before transfer learning can be applied in RMFD, e.g., the applicability of transfer strategy when facing with a specific fault diagnosis problem. In this case, we observe that more and more literatures begin to pay close attention to a certain RMFD problem with TL algorithms, but without providing a systematic guidance. Therefore, this paper attempts to provide an overview of the latest applications of knowledge transfer on rotary machine fault diagnosis, and guide researchers in this field to carry out relevant studies. In this paper, some key questions will be answered, including: 1) The motivation of knowledge transfer in RMFD; 2) The applicability of knowledge transfer in RMFD; 3) The effectiveness of knowledge transfer in RMFD; and 4) Other issues.

The remainder of this paper is organized as follows. In Section 2, the theoretical background of transfer learning is introduced, and the main knowledge transfer strategies are categorized and discussed in detail. Section 3 reviews and summarizes applications of transfer learning in rotary machine fault diagnosis, and offers some case studies. After that, Section 4 discusses the future research trends of TL-based RMFD from several aspects, such as applicability, effectiveness, and real-time capability.

II. Theoretical Background of Knowledge Transfer

Transfer learning is known as a tool to solve the basic problem of unlabeled and insufficient data in target domain, by utilizing the available knowledge from source domain. To begin with, we refer to Pan's [21] notations and definitions of transfer learning, which will be used in our work. Given source domain $\mathcal{D}_s$, source task $\mathcal{T}_s$, target domain $\mathcal{D}_t$ and target task $\mathcal{T}_t$, transfer learning aims to optimize predictive function f_t in target domain using the knowledge from predictive function f_s in source domain. Here $\mathcal{D} = \{\mathcal{X}, \mathrm{P}(\boldsymbol{x})\}$ is composed of the feature space $\mathcal{X}$ and its marginal probability distribution (MPD) $\mathrm{P}(\boldsymbol{x})$ [26-34], $\boldsymbol{x} = \{x_1, \cdots, x_n\} \in \mathcal{X}$. $\mathcal{T} = \{\mathcal{Y}, f\}$ is composed of the label space $\mathcal{Y}$, and the predictive function f. In non-transfer learning, $\mathcal{D}_\mathrm{s} = \mathcal{D}_\mathrm{t}$ and $\mathcal{T}_\mathrm{s} = \mathcal{T}_\mathrm{t}$, but in transfer learning, $\mathcal{D}_\mathrm{s} \neq \mathcal{D}_\mathrm{t}$ or $\mathcal{T}_\mathrm{s} \neq \mathcal{T}_\mathrm{t}$.

- ❖ $\mathcal{D}_\mathrm{s} \neq \mathcal{D}_\mathrm{t}$ means that $\mathcal{X}_s \neq \mathcal{X}_t$ and/or $\mathrm{P}(\boldsymbol{x}_s) \neq \mathrm{P}(\boldsymbol{x}_t)$.

- $\mathcal{T}_s \neq \mathcal{T}_t$ means that $\mathcal{Y}_s \neq \mathcal{Y}_t$ and/or $P(\boldsymbol{y}_s|\boldsymbol{x}_s) \neq P(\boldsymbol{y}_t|\boldsymbol{x}_t)$. where $P(\boldsymbol{y}|\boldsymbol{x})$ represents the conditional probability distribution (CPD) [35-37], $\boldsymbol{y} = \{y_1, \cdots, y_n\} \in \mathcal{Y}$.

According to Weiss [4], existing models can be divided into four categories, including parameter-based, instance-based, feature-based, and relevance-based knowledge transfer strategies, which will be introduced in the following subsections in detail.

A. Parameter-based Knowledge Transfer

In parameter-based knowledge transfer model, some special parameters are considered as the carriers of transfer learning. Although Rozantsev [38] listed only one transfer case, his idea can be extended to other parameter-based knowledge transfer models. In this model, it is assumed that f_s and f_t share parameter spaces partly [39]. Therefore, part of parameters of f_t can be replaced using corresponding parameters in f_s after appropriate adjustment. Basically, the representation of predictive function f_t in target domain can be defined as

$$f_t(\boldsymbol{x}_t, \boldsymbol{w}, \theta) = \langle \boldsymbol{w}, \psi(\boldsymbol{x}_t, \theta)\rangle \tag{1}$$

where ψ is the parametric mapping function from space $\mathcal{X}_t$ to $\mathbb{R}^{\mathrm{m}}$, which contains the parameter $\theta, (\theta \in \Theta_t)$. Here Θ_t is the subset of a normed space. Another parameter $\boldsymbol{w}, (\boldsymbol{w} \in \mathcal{W}_t)$ is also introduced to have a convolution operation with ψ to obtain the expression of f_t. Here $\mathcal{W}_t$ is a subset of $\mathbb{R}^{\mathrm{m}}$. So, the whole sets $\mathcal{F}_t$ of predictive function in target domain can be parameterized using θ and $\boldsymbol{w}$ by

$$\mathcal{F}_t = \{f_t(\boldsymbol{w}, \theta) | \theta \in \Theta_t, \boldsymbol{w} \in \mathcal{W}_t\} \tag{2}$$

In a similar way, the whole sets $\mathcal{F}_s$ of predictive function in source domain can also be parameterized using θ and $\boldsymbol{w}$ by

$$\mathcal{F}_s = \{f_s(\boldsymbol{w}, \theta) | \theta \in \Theta_s, \boldsymbol{w} \in \mathcal{W}_s\} \tag{3}$$

where $\Theta_s \cap \Theta_t \neq \phi$ or $\mathcal{W}_s \cap \mathcal{W}_t \neq \phi$, which means that a part of common space exists between Θ_s and Θ_t, or between $\mathcal{W}_s$ and $\mathcal{W}_t$.

Let $\theta_s^* \in \Theta_s$ and $\boldsymbol{w}_s^* \in \mathcal{W}_s$ represent trained parameters which possess the strong predictive ability in $\mathcal{D}_s$. After giving two risk functions $\mathcal{R}_1$ and $\mathcal{R}_2$, the high-quality parameters $\theta_t^* \in \Theta_t$, and $\boldsymbol{w}_t^* \in \mathcal{W}_t$ can be calculated and adjusted using the processes from Eq.(4) to Eq.(6):

Training in $\mathcal{D}_s$:

$$[f_s(\boldsymbol{x}_s, \boldsymbol{w}, \theta), \mathcal{Y}_s] \rightarrow [\boldsymbol{w}_s^*, \theta_s^*] \tag{4}$$

Transferring parameters:

$$\theta_t^* = \underset{\theta \in \Theta_t}{\operatorname{argmin}} \mathcal{R}_1[f_t\{\boldsymbol{w}_s^*, \theta\}] = \underset{\theta \in \Theta_t}{\operatorname{argmin}} \mathcal{Q}_1(\boldsymbol{w}_s^*, \theta)$$

$$\boldsymbol{w}_t^* = \underset{\boldsymbol{w} \in \mathcal{W}_t}{\operatorname{argmin}} \mathcal{R}_2[f_t\{\boldsymbol{w}, \theta_s^*\}] = \underset{\boldsymbol{w} \in \mathcal{W}_t}{\operatorname{argmin}} \mathcal{Q}_2(\boldsymbol{w}, \theta_s^*) \tag{5}$$

Model application in $\mathcal{D}_t$:

$$[\boldsymbol{w}_t^*, \theta_t^*] \rightarrow [f_t(\boldsymbol{x}_t, \boldsymbol{w}, \theta), \mathcal{Y}_t] \tag{6}$$

where $\mathcal{Q}_1 = \mathcal{R}_1(f_t)$, $\mathcal{Q}_2(\cdot) = \mathcal{R}_2(f_t)$.

For the implementation of parameter-based knowledge transfer, many algorithms have demonstrated impressive performance empirically. One of the major methods is to explore two key parameters in support vector machine (SVM) [40-42]: the direction parameter $\omega = (\omega_1; \omega_2; \ldots; \omega_{\mathrm{d}})$ and the distance parameter $\boldsymbol{b} = (b_1; b_2; \ldots; b_{\mathrm{d}})$ of the hyper plane for classification, and to extend them to the domain adaptation framework. For example, Tommasi [36,43] proposed a multi-model knowledge transfer (MMKT) target function, which used a group of coefficients from multiple prior models to replace single coefficient in classical least square support vector machine (LSSVM) [44,45], and made LSSVM possess the transfer ability. Bruzzone's [46] presented a domain adaptation support vector machine (DASVM) that extended SVM to the domain adaptation framework by exploiting labeled samples in $\mathcal{D}_{\mathrm{s}}$ and unlabeled samples in $\mathcal{D}_{\mathrm{t}}$ during the iteration of the algorithm. Different from DASVM, the domain independent support vector machine (DISVM) model [47] simultaneously minimized the SVM empirical risk and the dependence via Hilbert-Schmidt independence criterion (HSIC) [48] to realize domain adaptation. Other extended algorithms include domain adaptation machine (DAM) [49] and domain selection machine (DSM) [50]. The advantage of these models is that SVM parameters have clear physical meanings, and can be easily utilized.

Another method is to extend the convolutional neural networks (CNNs) to the domain adaptation framework. At first, the parameters of CNNs are obtained by the pre-training process in source domain $\mathcal{D}_s$. Then, they are transferred to the CNNs in target domain $\mathcal{D}_{\mathrm{t}}$ after fine-tuning. Shin [1] evaluated the transfer capability of different CNN architectures, with different widths and depths, as well as different dataset scales. The most critical issue in CNNs-based transfer model is to select transferred parameters because there are too many parameters. For example, Krizhevsky's deep CNN structure [51] is composed of 5 convolutional layers and 3 fully-connected layers, and has more than 60 million parameters. Other 19 layers [52] and 22 layers [53] deep CNN structures contain even more parameters. Therefore, to simply the parameter model and obtain better transfer ability, Han [54] only transferred the useful parameters in max pooling layers and achieved outstanding performance in image classification. In addition, abstract CNN parameters have better learning ability but higher complexity than SVM parameters [55].

Beside two methods described above, other parameter-based knowledge transfer models were also studied, such as hidden parameter Markov decision process [56,57], domain transfer multiple kernel learning [26,58], sparse coding [59-61], etc. Overall, such a model has wide range of application due to small cost of only transferring parameters.

B. Instance-based Knowledge Transfer

Instance-based knowledge transfer strategy is able to solve the TL problem when both low quality and high quality samples exist in source domain $\mathcal{D}_s$. Here low quality samples hamper the modeling of f_t while high quality samples boost the modeling of f_t. Its main idea is to improve the predictive performance of f_t by weakening the weight of low quality samples and enhancing the weight of high quality instances in $\mathcal{D}_s$. Weight adjustment [62] and domain separation [63] are two major means for implementation of instance-based knowledge transfer. The later can be regarded as the particular case of the former where the weights of some samples are set as one and others are set as zero. Mathematically, in instance-based knowledge transfer model, f_s and f_t can be defined as

$$f_s(\boldsymbol{x}_s) = \langle \boldsymbol{w}_s, \psi_s(\boldsymbol{x}_s) \rangle$$
$$f_t(\boldsymbol{x}_t) = \langle \boldsymbol{w}_t, \psi_t(\boldsymbol{x}_t) \rangle \quad (7)$$

where ψ_s and ψ_t are the mapping functions from spaces $\mathcal{X}_s$ and $\mathcal{X}_t$ to $\mathbb{R}^{\mathrm{m}}$, respectively. $\boldsymbol{w}_s$ and $\boldsymbol{w}_t$ represent the vectors which have a convolution with ψ_s and ψ_t.

In such a model, source domain $\mathcal{D}_s$ is separated into positive domain $\mathcal{D}_s^p = \mathcal{D}_s \cap \mathcal{D}_t$ and negative domain $\mathcal{D}_s^n = \mathcal{D}_s - \mathcal{D}_s^p$. Therefore, the weight trend factors $\{m_s^+, m_s^-\}$ can be calculated during iterations as

$$m_s^+(\boldsymbol{x}_s^i) = \frac{1}{2}\left[\mathcal{U}(\boldsymbol{x}_s^i, \mathcal{D}_s^p) + 1 - \mathcal{U}(\boldsymbol{x}_s^i, \mathcal{D}_s^n)\right]$$
$$m_s^-(\boldsymbol{x}_s^i) = \frac{1}{2}\left[\mathcal{U}(\boldsymbol{x}_s^i, \mathcal{D}_s^n) + 1 - \mathcal{U}(\boldsymbol{x}_s^i, \mathcal{D}_s^p)\right] \quad (8)$$

where $\boldsymbol{x}_s^i$ represents the i-th sample in $\mathcal{D}_s$; $m_s^+(\boldsymbol{x}_s^i)$ and $m_s^-(\boldsymbol{x}_s^i)$ represent the positive and negative weight trend, respectively; $\mathcal{U}(\boldsymbol{x}, \mathcal{D}) \subset [0,1]$ defines the normalized discrimination function between sample $\boldsymbol{x}$ and domain $\mathcal{D}$. As an example, $\mathcal{U}(\boldsymbol{x}, \mathcal{D})$ can be defined as [64]

$$\mathcal{U}(\boldsymbol{x}, \mathcal{D}) = \frac{1}{2}\sum_{\boldsymbol{x}_j \in \mathcal{D}} e^{-\|f_t(x) - f_t(x_j)\|} \quad (9)$$

where f_t represents the predictive function in the r-th iteration. The (r+1)-th weight of sample $\boldsymbol{x}_s^i$ can be updated by

$$w_{r+1}(\boldsymbol{x}_s^i) = w_r(\boldsymbol{x}_s^i) + \epsilon[m_s^+(\boldsymbol{x}_s^i) - m_s^-(\boldsymbol{x}_s^i)] \quad (10)$$

where w_{r+1} and w_r represent the r-th and (r+1)-th weights, respectively; $\epsilon \subset [0,1]$ represents the weight regulator. According to description above, the general process of instance-based knowledge transfer strategy is generalized as [65]:

Input:
- ❖ A small quantity of labeled samples from $\mathcal{D}_t^1$;
- ❖ A large quantity of unlabeled samples from $\mathcal{D}_t^2$;
- ❖ A large quantity of labeled samples from $\mathcal{D}_s$;

Iteration:
- ❖ $w_1(\boldsymbol{x}_s^i)$=1; $\mathcal{D}_s^p$ =random samples in $\mathcal{D}_s$; $\mathcal{D}_s^n = \mathcal{D}_s - \mathcal{D}_s^p$;
- ❖ **While** $r \le ite_{\max}$, **Do**
- ❖ Update the weight trend factors $\{m_s^+, m_s^-\}$ using Eq.(8);
- ❖ Update the weight of samples in $\mathcal{D}_s$ using Eq.(10);
- ❖ A basic classier f_r is carried out in domain $\{\mathcal{D}_t^1; \mathcal{D}_s\}$, and the effectiveness of $\mathcal{D}_t^1$ is calculated using evaluation function e($\mathcal{D}_t^1$). If qualified then **Break**;
- ❖ $\mathcal{D}_s^p$ ←properly classified samples in $\mathcal{D}_s$; $\mathcal{D}_s^p$ ←wrongly classified samples in $\mathcal{D}_s$; $r \leftarrow r + 1$;

  End While

Output:
- ❖ $f_t = f_r$; $\boldsymbol{y}_t = f_t(\boldsymbol{x}_t)$;

Although such a model is limited by the characteristics of samples in $\mathcal{D}_s$, some algorithms have been studied in recent years. Liu [64] proposed a selective multiple instance transfer learning (SMITL) model, which used the bi-memberships factor to solve the instance transfer problem for text classification. Chattopadhyay [66] successfully applied the weight vector in conditional probability based multi-source domain adaptation (CP-MDA) strategy to surface electromyography (SEMG), by minimizing the difference in predicted labels between two nearby points in $\mathcal{D}_t$. Cheng [67] designed a weighted multi-source TrAdaboost strategy to take full advantage of valuable information from multiple domains, which enhanced the generalization ability of classical TrAdaboost [68]. Other related algorithms include Chattopadhyay’s two stage weighting framework for multi-source domain adaptation (2SW-MDA) [66], Kotzias’s deep multi-instance transfer [69], Xie’s semantic instance annotation-based transfer [70], Hasse’s active appearance model-based transfer [71], Huang’s adaptive instance normalization [72], and Babenko’s multiple instance transfer learning (MITL) [73]. Particularly, to overcome the drawback that most existing MITL methods cannot solve the problem of insufficient samples, Tan [74] presented an adaptive knowledge transfer for multiple instances learning (AKT-MIL), which was successfully applied to image classification. Overall, the advantage of these models is that they are helpful to solve the transfer problem with more than one source domains because they focus more on characteristics of samples than domains.

C. Feature-based Knowledge Transfer

Different from instance-based knowledge transfer, feature-based knowledge transfer pays attention to the different feature spaces between source domain and target domain. The marginal distribution difference and conditional distribution difference exist between $\mathcal{D}_s$ and $\mathcal{D}_t$ [75]. Therefore, the purpose of such a model is to close their feature spaces through some means such as feature mapping. Two strategies are usually used in this kind of model, including symmetric and asymmetric feature transfer, which are compared in Fig. 1.

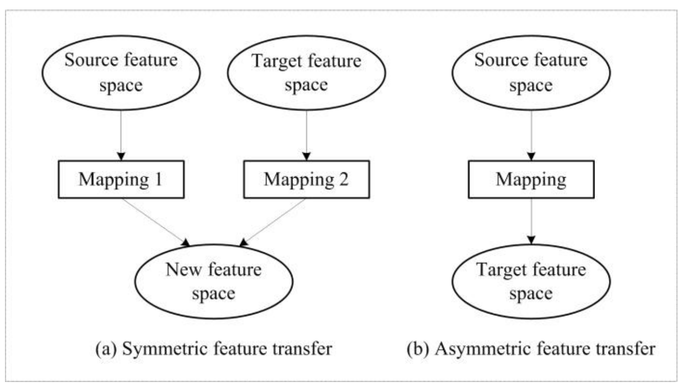

Fig. 1. The comparison between symmetric and asymmetric feature transfer.

Among symmetric feature transfer methods, both $\mathcal{D}_s$ and $\mathcal{D}_t$ are mapped to two new feature spaces $\mathcal{D}_s'$ and $\mathcal{D}_t'$, where the differences between source domain and target domain are weakened. The general process of symmetric feature-based knowledge transfer strategy is generalized as

Input:

- A large quantity of unlabeled samples from $\mathcal{D}_t$;
- A large quantity of labeled samples from $\mathcal{D}_s$;

Feature mapping: $\mathcal{D}_s' \leftarrow \mathcal{D}_s$; $\mathcal{D}_t' \leftarrow \mathcal{D}_t$;

$$\min_{\boldsymbol{V}_1,\boldsymbol{V}_2} \begin{array}{l} D_f[\mathrm{P}(\boldsymbol{x}_{m1}), \mathrm{P}(\boldsymbol{x}_{m2})] \\ + \lambda D_f[\mathrm{P}(\widehat{\boldsymbol{y}}_{m1}|\boldsymbol{x}_{m1}), \mathrm{P}(\widehat{\boldsymbol{y}}_{m2}|\boldsymbol{x}_{m2})] \end{array} \tag{11}$$

Training in $\mathcal{D}_s'$:

$$[\boldsymbol{x}_s, \boldsymbol{y}_s] \rightarrow [\boldsymbol{x}_s', \boldsymbol{y}_s] \rightarrow f_s' = f_t' \tag{12}$$

Model application in $\mathcal{D}_t'$:

$$[\boldsymbol{x}_t, f_t'] \rightarrow [\boldsymbol{x}_t', f_t'] \rightarrow \boldsymbol{y}_t \tag{13}$$

Inverse mapping: $\mathcal{D}_t \leftarrow \mathcal{D}_t'$

$$[\boldsymbol{x}_t', \boldsymbol{y}_t] \rightarrow [\boldsymbol{x}_t, \boldsymbol{y}_t] \tag{14}$$

where $\boldsymbol{x}_{m1} \in \mathcal{D}_{m1}(\boldsymbol{V}_1, \mathcal{D}_s)$, $\mathcal{D}_{m1}(\boldsymbol{V}_1, \mathcal{D}_s)$ represents the mapped feature space of $\mathcal{D}_s$ using the projected vector $\boldsymbol{V}_1$; $\boldsymbol{x}_{m2} \in \mathcal{D}_{m2}(\boldsymbol{V}_2, \mathcal{D}_t)$, $\mathcal{D}_{m2}(\boldsymbol{V}_2, \mathcal{D}_t)$ represents the mapped feature space of $\mathcal{D}_s$ using the projected vector $\boldsymbol{V}_2$; D_f defines the divergence function; $\mathrm{P}(\boldsymbol{x})$ represents the marginal distribution; $\mathrm{P}(\widehat{\boldsymbol{y}}|\boldsymbol{x})$ represents the likelihood distribution where $\widehat{\boldsymbol{y}}$ is the estimation of $\boldsymbol{y}$; λ represents the weight factor to adjust two divergence values.

As a typical symmetric feature transfer model, the goal of transfer component analysis (TCA) [33,76,77] is to discover common latent features that have the same marginal distribution between $\mathcal{D}_s'$ and $\mathcal{D}_t'$. Meanwhile, the intrinsic structures before and after mapping are the same. Some projection methods are often used to optimize the target function of TCA, such as maximum mean discrepancy [78,79], manifold regularization [80,81]. Another typical symmetric feature transfer model is the deep CNNs [82,83]. Different from deep CNNs in parameter-based knowledge transfer strategy, the middle neurons of CNNs are considered as transferred features here. Therefore, the middle layers of deep CNNs can be seen as new feature space, where the feature differences between source and target domains are weakened. In Long's deep adaptation networks (DAN) [84], the neurons in the 1-st to the 3-rd convolutional layers are general thus being frozen, and the neurons in the 4-th and 5-th convolutional layers are slightly less transferable, thus being learned via fine-tuning, and the neurons in the 6-th to the 8-th fully connected layers are tailored to fit specific tasks, thus being adapted with multiple kernel variant of MMD (MK-MMD) [85]. In addition, many other algorithms were also studied. For example, Pan's [32] spectral feature alignment (SFA) used the domain-independent features as a bridge to build a relationship between $\mathcal{D}_s$ and $\mathcal{D}_t$. Glorot [27] and Kandaswamy [86] proposed the stacked denoising autoencoder (SDA) model to discover the common invariant latent feature space. Gong [29] used geodesic flow kernel (GFK) to map two original spaces to a new low-dimensional feature space for reducing the marginal distribution differences. Rajesh [87] presented the joint nonnegative matrix tri-factorizations (NMTF) to adapt the differences in both marginal and conditional distributions of feature spaces.

Among asymmetric feature transfer methods, $\mathcal{D}_s$ is directly mapped to $\mathcal{D}_s'$, which has the same characteristic with $\mathcal{D}_t$. The general process of asymmetric feature-based knowledge transfer strategy is generalized as

Input:

- A large quantity of unlabeled samples from $\mathcal{D}_t$;
- A large quantity of labeled samples from $\mathcal{D}_s$;

Feature mapping: $\mathcal{D}_s' \leftarrow \mathcal{D}_s$;

$$\min_{\boldsymbol{V}} \begin{array}{l} D_f[\mathrm{P}(\boldsymbol{x}_m), \mathrm{P}(\boldsymbol{x}_t)] \\ + \lambda D_f[\mathrm{P}(\widehat{\boldsymbol{y}}_m|\boldsymbol{x}_m), \mathrm{P}(\widehat{\boldsymbol{y}}_t|\boldsymbol{x}_t)] \end{array} \tag{15}$$

Training in $\mathcal{D}_s'$:

$$[\boldsymbol{x}_s, \boldsymbol{y}_s] \rightarrow [\boldsymbol{x}_s', \boldsymbol{y}_s] \rightarrow f_s' = f_t \tag{16}$$

Model application in $\mathcal{D}_t$:

$$[\boldsymbol{x}_t, f_t] \rightarrow [\boldsymbol{x}_t', f_t] \rightarrow \boldsymbol{y}_t \tag{17}$$

where $\boldsymbol{x}_m \in \mathcal{D}_m(\boldsymbol{V}, \mathcal{D}_s)$, $\mathcal{D}_m(\boldsymbol{V}, \mathcal{D}_s)$ represents the mapped feature space of $\mathcal{D}_s$ using the projected vector $\boldsymbol{V}$.

As two typical asymmetric feature transfer models, the joint domain adaptation (JDA) [77] and the adaptation regularization based transfer learning (ARTL) [88] were proposed by Long. Here, the marginal distribution adaptation (MDA), the conditional distribution adaptation (CDA) and manifold regularization (MR) were carried out in proper order in $\mathcal{D}_s$ to match the feature space of $\mathcal{D}_t$. Finally, $\mathcal{D}_s$ and $\mathcal{D}_t$ had the same feature distribution for successfully image classification. Later, Nguyen [89] modified JDA to joint hierarchical domain adaptation (JHDA), which prevented excessively rapid growth of data dimension when the number of layer increases. Kulis [90] proposed an asymmetric regularized cross-domain transformation (ARC-t) to adjust the heterogeneous feature space, where a regularization term was contained in the objective function to the learned transformation matrix from $\mathcal{D}_s$ to $\mathcal{D}_t$. Excellent learning ability of ARC-t model was proved by Kulis in solving the problem of multiple object recognition. Kandemir [91] achieved the aim of asymmetric transfer by the deep gaussian process, which performed knowledge transfer not only by projecting $\mathcal{D}_s$ onto $\mathcal{D}_t$ but also projecting $\mathcal{D}_t$ onto $\mathcal{D}_s$, thus possessing higher transfer performance than single transfer from $\mathcal{D}_s$ to $\mathcal{D}_t$. Especially, multiple source domains and one target domain existed in Harel's transfer method [92], where all source feature spaces were transferred to the target feature space. Here, each individual source domain was paired with the corresponding target domain, and the singular value decomposition (SVD) [93,94] was performed to obtain the transformation matrix of each source domain. Wang [95] improved Harel's method later, which replaced the SVD matrix with sparse feature transformation matrix, and further enhanced the domain similarity. Finally, the idea of descending dimension was also introduced into asymmetric feature-based transfer to select more similar features between $\mathcal{D}_s$ and $\mathcal{D}_t$, such as, gain ratio [96], Chi square [97], relief-F [98], and decision tree [99]. Although directly mapping from source domain to target domain is faster, its classification accuracy is lower than symmetric feature transfer methods, because its performance depends on the quality of target feature space.

D. Relevance-based Knowledge Transfer

Generally, more than one object exists in the source domain and target domain. In some situations, the relevance between two objects in source domain can be transferred to the relevance between two similar objects in target domain. For example, the relation between student and teacher can be transferred to the relation between underling and leader, called relevance-based knowledge transfer model. If the label space $\boldsymbol{y}$ is seen as the object space, the general process of relevance-based knowledge transfer strategy will be generalized as

Input:

- ❖ A large quantity of unlabeled samples from $\mathcal{D}_t$;
- ❖ A large quantity of labeled samples from $\mathcal{D}_s$;

Obtain the relevance between labels in $\boldsymbol{\mathcal{D}_s}$:

$$r_{ij}^s = R\left[\boldsymbol{x}_s^i, \boldsymbol{x}_s^j\right], \text{if } f_s(\boldsymbol{x}_s^i) = y_i^s \;\&\; f_s\left(\boldsymbol{x}_s^j\right) = y_j^s \tag{18}$$

Relevance mapping: $r_{ij}^s \to r_{ij}^t$

$$r_{ij}^t = \varphi\left(r_{ij}^s\right) \tag{19}$$

Relevance application in $\boldsymbol{\mathcal{D}_t}$:

$$f_t\left(\boldsymbol{x}_t^i\right) = y_i^t \;\&\; f_t\left(\boldsymbol{x}_t^j\right) = y_j^t, \text{ if } \min\left\{R\left[\boldsymbol{x}_t^i, \boldsymbol{x}_t^j\right] - r_{ij}^t\right\} \tag{20}$$

where the $\mathcal{Y}_s = \{y_1^s, \cdots, y_n^s\}$ and $\mathcal{Y}_t = \{y_1^t, \cdots, y_n^t\}$ represent the label space in source task and target task, respectively; $\mathcal{R}_s = \left\{r_{ij}^s\right\}$ and $\mathcal{R}_t = \left\{r_{ij}^t\right\}, i \in [1, n], j \in [1, n], i \neq j$ represent the relational space in source and target tasks. $\boldsymbol{x}_s^i \in \mathcal{D}_s^i$ and $\mathcal{D}_s^i$ represents the source domain whose label is y_i; R represents the relation function between two datasets; φ defines an adjust function from r_{ij}^s to r_{ij}^t, which depends on the known knowledge.

Of all important transfer strategies, relevance-based knowledge transfer has been the least studied because it has to meet the conditions of multiple objects. Based on the work of Qiu [100] and Jakob [101], Li [30] proposed a relational adaptive bootstrapping (RAP) method to extract the key words from a text document, which was applied to analyzing the cross domains of sentiment and topic words. In Li's model, all domains were equally treated, which could be replaced by other cross domain algorithms, such as the domain transfer networks (DTNs) [102], the residual transfer networks (RTNs) [103], the latent sparse domain transfer (LSDT) [104], etc. Mihalkova [105] and Torrey [106] proposed the transfer via automatic mapping and revision (TAMAR) model that transferred relevance with markov logic networks (MLNs) and was applied to transferring between three real-world domains: movie relationship domain (IMDB dataset) [107], academic relationship domain (UW-CSE dataset) [108] and education relationship domain (WebKB dataset) [109]. Davis [110] modified Mihalkova's approach to further expand application, called the deep transfer via markov logic (DTM). This method had a successful application in the transfer between the molecular biology domain and the Web domain. Overall, relevance-based knowledge transfer strategy possesses the outstanding advantages in transfer between two very different domains as compared with other three models.

III. Applications of Knowledge Transfer in Rotary Machine Fault Diagnosis

The purpose of rotary machine fault diagnosis is to recognize the mechanical faults by analyzing signals collected from key rotating units, such as gears [111-114], bearings [115-120], motor [20,121-123], etc. In traditional RMFD, the signals for training dataset $\boldsymbol{T}$ and testing dataset $\boldsymbol{S}$ are from the same condition. Therefore, their marginal distributions $\mathrm{P}(\boldsymbol{x})$ and conditional distributions $\mathrm{P}(\boldsymbol{y}|\boldsymbol{x})$ are the same. However, varying working conditions and other factors always happen during the actual signal acquisition process, which affects final diagnostic results. Generally, there are three major strategies to reduce these influences:

- ❖ Fusing multiple algorithms to overcome the drawback when single algorithm is difficult to deal with unknown working conditions. For example, Liu [124] fused the Hilbert-Huang transform and singular value decomposition (SVD), and Wang [125] fused the wavelet packet transform and manifold learning (ML) for rolling element bearing fault diagnosis;
- ❖ Improving single algorithm to adapt the change of working conditions. For example, Valencia [126] promoted the Fourier transform to short-frequency Fourier transform for fault diagnosis of induction machines working in transient regime. Wang [127] modified the synchronous squeezing transform (SST) to the matching SST (MSST) for machine fault diagnosis under fast varying instantaneous frequency.
- ❖ Building a robust RMFD model to resist the change of external factors. For example, Mulumba [128] developed a model-based fault diagnosis method by applying SVM techniques to model parameters calculated online.

TABLE I
Similarities in The Signal Space of Mechanical System

No.	Similarities	Transfer style
1	Similar signal's singular values under two different rotating speeds	Transfer between multiple working conditions
2	Similar vibrational frequency in two meshed gears	Transfer between multiple locations
3	The same fault types of gear installed in two different machines	Transfer between multiple machines
4	Wear and crack both grow in approximately exponential style	Transfer between multiple fault types

Alternatively, the TL theory has been introduced into the field of RMFD [129,130]. The idea of knowledge transfer can be applied to this field due to the fact that some similarity characteristics exist in the signal space of mechanical system, as listed in Table I. Taking advantage of these similarity characteristics, four types of knowledge transfer applications can be realized for rotary machine fault diagnosis, including transfer between multiple working conditions, transfer between multiple locations, transfer between multiple machines, and transfer between multiple fault types.

It should be noted that TL-based rotary machine fault diagnosis has the following advantages: 1) The useful information in training dataset can be utilized to assist the learning of testing datasets when number of samples in the latter is insufficient, thus improving the diagnostic performance; 2) The pre-training in source domain can save the diagnostic time in target domain, even they have different distributions, thus offering an effective way for real-time fault diagnosis.

A. *Transfer between Multiple Working Conditions*

1) Transfer goals. As shown in Fig. 2, the principle of this type of knowledge transfer is that model trained using data under one working condition will be applied to process the data under another working condition. Its goal is to obtain effective diagnostic results when the machine runs in a new working environment while using the model built in the old working environment. In fact, most of TL-based RMFD literatures focus on this kind of studies, which includes domain selection, varying rotating speed, and varying load studies.

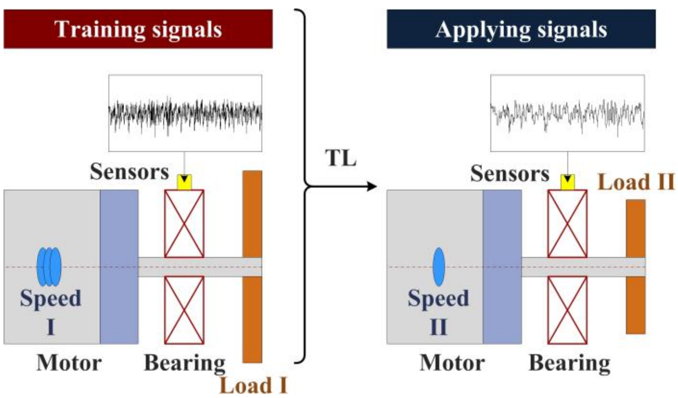

Fig. 2. The transfer style between multiple working conditions.

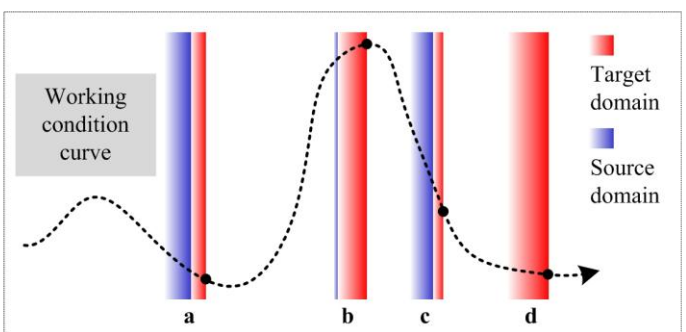

Fig. 3. The working condition scan method for domain selection.

2) Domain selection studies. Although transfer learning models have better domain adaption capacity than traditional models, the data quality in source domain still has direct influence on the final diagnostic performance. Therefore, selecting suitable data in source working condition is one of the main topics for discussion. The first measure is to extract high quality features from the measured signals, which are less affected by varying working conditions. Cao [130], Zhang [131], Cheng [132], and Han [133] made use of deep CNNs as the feature extractor from signal's time-frequency distribution, where the middle layer neurons of CNNs are considered as features for transfer learning. Other features have also been utilized to weaken the difference between $\mathcal{D}_s$ and $\mathcal{D}_t$, e.g., the sparse coding features [134], the principal components features [135] and the singular value features [65]. The second measure is to design the indexes to assess the domain similarity between $\mathcal{D}_s$ and $\mathcal{D}_t$, and then select suitable $\mathcal{D}_s$ for analysis. For example, Cao [136] defined a joint task kernel $\boldsymbol{C}$ to automatically model the cross task-link similarity by

$$\boldsymbol{C} = \boldsymbol{T} \otimes \boldsymbol{K} + \sigma^2 \boldsymbol{I} \tag{21}$$

where $\boldsymbol{T}$ is the positive semi-definite (PSD) matrix that specifies the inter-task similarities; $\boldsymbol{K}$ is the kernel matrix for link modeling using kernel function k; σ represents the Gaussian variance; $\boldsymbol{I}$ represents the unit matrix; $\otimes$ represents the operation of multiplying elements. Other assessment indexes include the Kullback-Leibler divergence [137], L1-distance [138], Jensen-Shannon divergence [139], and Jaccard distance [140], etc. In this study, a working condition scan method is used as an example for domain selection, as shown in Fig. 3. With time passing by, measured signals in current moment are considered as the target domain, and the signals in its nearest historical moment are considered as the source domain, whose length depends on the rate of working conditions. Long target domain signals are selected in slow change, e.g. $l(d) > l(b) > l(a) > l(c)$ in Fig.3, where l represents the length of target domain.

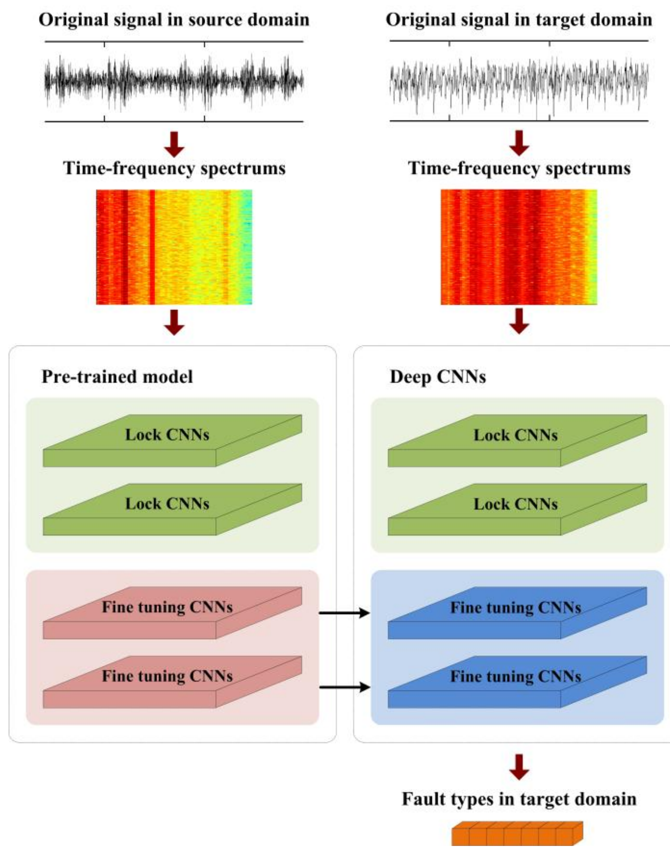

Fig. 4. Deep CNNs model for RMFD.

3) Varying rotating speed transfer models. Change of rotating speed affects the corresponding frequency information. In this case, the role of transfer learning is to weaken the frequency difference between source domain and target domain when facing with varying rotating speed. Deep CNNs model [129] is one of the solutions for RMFD, as shown in Fig.4. Cao [130] used this model to transfer between different rotating speeds, and it was consisted of two processes: Extracting the features with a pre-trained deep neural network and designing a fully connected stage to classify these CNNs features. Although Cao's method was tested against other methods such as the angle-frequency domain synchronous analysis (AFS)-SVM using gear fault data[141], the limitations still exist because it selected both $\mathcal{D}_{\mathrm{s}}$ and $\mathcal{D}_{\mathrm{t}}$ randomly. As the improvement of Cao' method, Cheng [132] presented a Wasserstein distance based deep transfer learning (WD-DTL), which also treated the deep CNNs as the feature extractor, but further minimized the frequency difference through the adversarial training process. The diagnostic accuracy of WD-DTL was proved to be higher than TCA [33,76], JDA[77], Correlation alignment (CORAL) [142], CNNs [13,20], and deep adaptation network (DAN) [143,144]. When mixing more than one rotating speed, instance-based knowledge transfer is a good solution. Shen [65] utilized the TrAdaboost algorithm to transfer from multiple source rotating speeds (from 20Hz to 50Hz) to the target rotating speed (30Hz), and those similar signals (around 30Hz) in source domain were assigned high weights during iteration, leading to improved diagnostic performance for bearings. It was also concluded that the instance-based transfer method is superior to feature-based transfer method when the speed changes quickly.

4) Varying load transfer models. Different from rotating speed, varying load tends to change the time domain features of measured signals, such as the waveform, pulse, skewness [145] and kurtosis [146]. Han [133] proposed a DTN with JDA framework to solve the varying load problem. For CNNs, the features always change with the increase of layer depth. The upper layers tend to be more abstract and the lower layers are more specific [147]. Therefore, the JDA regularization term was designed by Han [133] in the lowest hidden fully-connected layer to minimize the discrepancy between $\mathcal{D}_s$ and $\mathcal{D}_t$. Using the wind turbine dataset, the DTN with JDA framework was proved to be superior to other algorithms, such as transfer joint matching (TJM) [148], TCA, JDA, and DTN with marginal distribution adaptation (MDA) [149], with high accuracy and robustness under varying load conditions. Furthermore, Shao [129] presented a fine-tuning strategy to overcome the difference of motor vibration when the load began to change. The features from the first three convolution blocks were transferred, and the last two convolution blocks as well as the fully connected layers were fine-tuned to match the target domain. All these works, including Li's deep distance metric learning[150], Pan's deep belief network[151], Wen's negative correlation ensemble transfer learning based on CNNs[152], Zhang's convolution neural networks with training interference (TICNN) [131], proved that deep transfer learning (DTL) provided an effective approach for machine fault diagnosis resulting from varying load conditions.

In addition, Jia [132,153,154] presented a local connection network with normalized sparse auto-encoder (NSAE-LCN) to overcome the limitations of classical auto-encoder when rotating speed and load change together. In NSAE-LCN, the weight matrices of local layers and feature layers were trained separately. Then, they were simultaneously updated using a fine-tuning strategy to improve final fault recognition accuracy. However, in Jia's method, the parameters required to be given in advance, unless some optimization techniques were introduced, such as the particle swarm optimization (PSO) [155]. Overall, the DTL technique is easier to be implemented for machine diagnosis than other algorithms because parameters can be automatically adjusted with change of load.

B. Transfer between Multiple Locations

1) Transfer goals. As shown in Fig. 5, the principle of this type of knowledge transfer is that model trained using data in one location can be applied to process the data in other locations. Its goal is to obtain effective diagnostic results when the sensors cannot be installed in target location due to physical limitation. Since signal coupling always exists in different locations of the gear drive system shown in Fig. 5, especially in two adjacent gearboxes, this offers the possibility of knowledge transfer. Generally, the locations near fault have stronger signal response and are more important than those far from fault location.

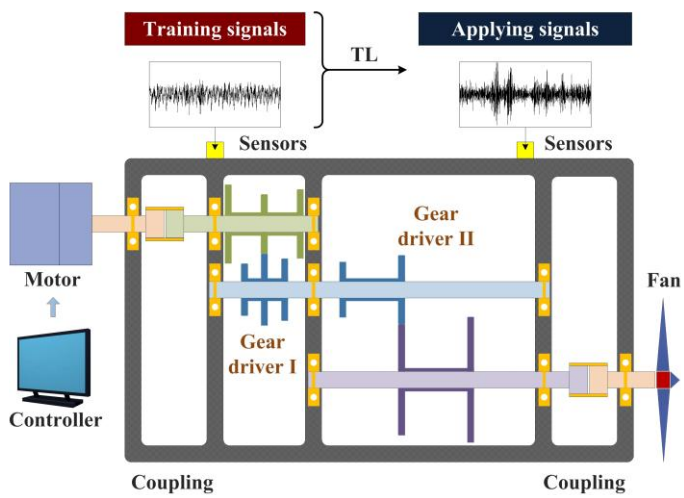

Fig. 5. The transfer style between multiple locations.

2) Domain selection studies. Although some distance or similarity measures were mentioned in existing literatures, such as Euclidean distance [6,156], n-dimensional Manhattan distance [157], and Minkowski distance [158], the relationship between signals at different locations are often ignored. In fact, signals measured from the machine system contain different components related to fault, environmental noise, and machine structure itself. Therefore, a signal separation method is recommended in Fig. 6 for domain selection, where the signals from all locations are separated into multiple vibration components by the blind source separation (BSS) [159] technique, such as the variational Bayesian hidden Markov model (VBHMM) [160], and continuous wavelet transform-based BSS (CWT-BSS) [161]. The signal which contains the strongest fault component will be considered in the target domain and other signals will be considered in the source domains according to the separation matrixes.

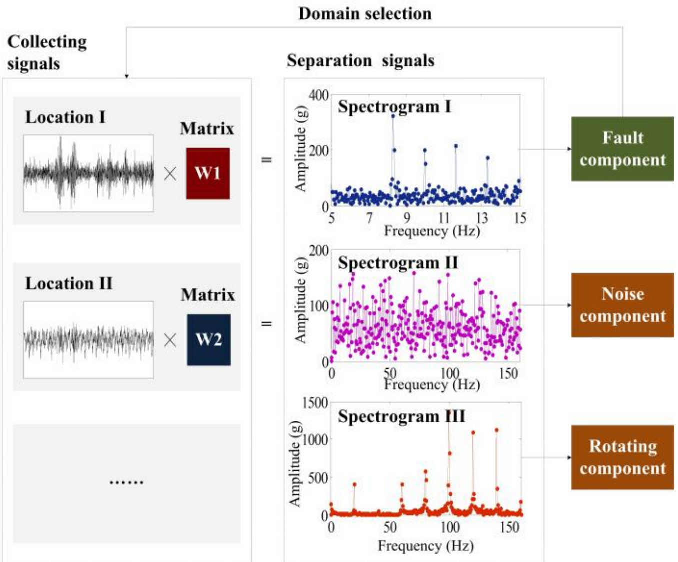

Fig. 6. The signal separation method for domain selection.

3) Specific transfer models. Compared with varying working conditions, knowledge transfer between multiple locations seems more complex because the mechanical structure between two sensors may be unknown. Meanwhile, the noise always exists between them. But some links still exist in the signal space between $\mathcal{D}_s$ and $\mathcal{D}_t$. For example, $f_t = \gamma f_s$, where f_t represents a certain fault frequency in target location; f_s represents corresponding fault frequency in source location; γ represents the transmission ratio between two sensors. Also, the amplitude trend is approximately the same between two signal acquisition locations. Tong [162] proposed a domain adaptation using transferable features (DATF), where difference of marginal and conditional distributions was reduced simultaneously across domains based on maximum mean discrepancy (MMD) in feature space by refining pseudo test labels. The general domain adaptation method [88] includes three processes: marginal distribution adaptation (MDA), conditional distribution adaptation (CDA) and target function modeling (TFM). Particularly, Tong and Long's methods are both suitable for varying working conditions and varying locations because they adjusted the feature space of source domain to match the target domain. Furthermore, another two dimension reduction techniques, including the transfer component analysis (TCA) [163] and the transfer factor analysis (TFA) [164], were applied for fault diagnosis, respectively, by mapping the original high-dimensional space to a new low-dimensional space. In TCA, a distance estimator $\mathrm{Dist}(\boldsymbol{x}_s, \boldsymbol{x}_t)$ between two domains was defined and minimized, by

$$\mathrm{Dist}(\boldsymbol{x}_s, \boldsymbol{x}_t) = \left\| \frac{1}{m} \sum_{i=1}^{m} \emptyset\left(\boldsymbol{x}_{s_i}\right) + \frac{1}{n} \sum_{i=1}^{n} \emptyset\left(\boldsymbol{x}_{t_i}\right) \right\|_H^2 \quad (22)$$

where $\emptyset$ represents the kernel transformation function; H is the subspace; m and n represent the number of samples in source and target domains, respectively. In TFA, two different diagonal covariance matrixes $\emptyset_1$ and $\emptyset_2$ were introduced in $\boldsymbol{x}_s$ and $\boldsymbol{x}_t$, which were linked by a weight matrix W_x. Then, parameters $\emptyset_1$, $\emptyset_2$ and W_x were updated to obtain the approximate log likelihood functions: $\log \mathrm{P}(\boldsymbol{x}_s) \approx \log \mathrm{P}(\boldsymbol{x}_t)$. Although Wang's diagnostic results proved that the performance of TFA outperforms TCA, TFA costs more time than TCA to adjust the transfer parameters. In Wen's work[165], a three-layer sparse auto-encoder (SAE) [166] was utilized to extract the features from raw data, and then the MMD term was applied to minimizing the discrepancy penalty of two domains. The SAE-MMD method is effective on transferring between multiple locations. In addition, some other algorithms were also carried out, such as neural networks [167]. Particularly, Zhang [168] presented an unsupervised domain adaptation using subspace alignment (SA) to overcome the drawback that the classifier trained in one location cannot be directly used to other locations. This method can distinguish not only bearing faults categories, but also fault severities.

C. Transfer between Multiple Machines

1) Transfer goals. As shown in Fig. 7, the principle of this type of knowledge transfer is that model trained using data measured on the machine in laboratory can be applied to process data measured on another machine in industrial field. Its goal is to obtain effective diagnostic results when signals of target machine are not available in time. Once the model built in one machine can be transferred to another, the diagnostic cost of the target machine will be greatly saved. Generally, there are two kind of strategies for transfer between multiple machines: Abstract feature-based transfer and relevance-based transfer.

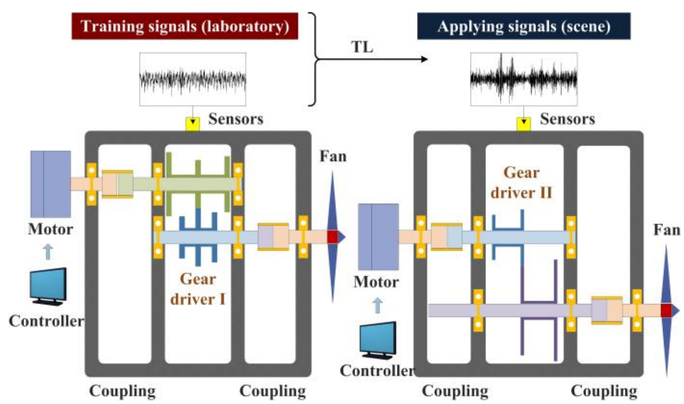

Fig. 7. The transfer style between multiple machines.

2) Abstract feature-based transfer. For abstract feature-based transfer, Yang [24] proposed a feature-based transfer neural network (FTNN) model to transfer from laboratory bearings to locomotive bearings. In Yang's method, a domain-shared CNN was employed to simultaneously extract the abstract vibrational features from both laboratory and real-case machines. Then the model was trained by jointly minimizing the error between the predicted and real labels in source domain, and the error between the predicted and pseudo labels in target domain. From experiment results, this method was superior to other algorithms, including CNN, TCA, multi-layer adaptation CNN (MACNN) [169], domain adaptation fault diagnosis (DAFD) [165], etc. Meanwhile, Guo [170] presented a deep convolutional transfer learning network (DCTLN), which consists of condition recognition and domain adaptation processes. The former was constructed by a 1D CNN to automatically recognize health conditions of machines. The latter facilitated the CNN to learn domain invariant features by maximizing domain recognition errors and minimizing probability distribution distance. Finally, This approach was utilized to transfer three datasets, including the CWRU bearing dataset [171], the IMS bearing dataset [172], and the RL bearing dataset [173]. In addition to deep CNNs, other algorithms can also be utilized for abstract features extraction, such as fast self-organizing feature mapper [174].

3**) Relevance-based transfer.** For relevance-based transfer, the link between two kinds of faults in source machine can be extracted and applied in the target machine. Fig. 8 gives an example of gearbox fault diagnosis, where the categories of testing gears include healthy (HEA), surface wear fault (SWF), tooth crack fault (TCF), chipped tooth fault (CTF) and missing tooth fault (MTF) in both two devices. The relevance vectors $\{r_1, r_2, r_3, r_4\}$ and $\{R_1, R_2, R_3, R_4\}$ can be learned using some association rule mining (ARM) [175] algorithms, such as fuzzy threshold [176,177] and neutrosophic theory [178,179]. After substituting these algorithms into Eq. (18) to (20), the fault gears in target machine can be recognized, thus offering an effective way for transferring between multiple machines.

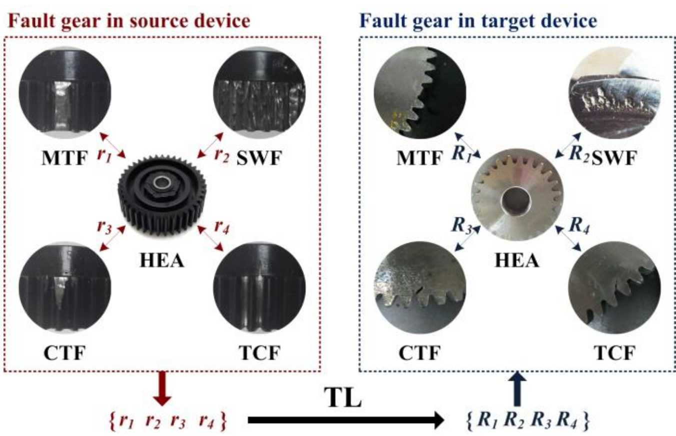

Fig. 8. An example of relevance-based transfer for gear fault diagnosis.

D. *Transfer between Multiple Fault Types*

1) Transfer goals. Transfer models described in previous sections attempt to solve the RMFD problem when $\mathrm{P}(\boldsymbol{x}_s) \neq \mathrm{P}(\boldsymbol{x}_t)$, or $\mathrm{P}(\boldsymbol{y}_s|\boldsymbol{x}_s) \neq \mathrm{P}(\boldsymbol{y}_t|\boldsymbol{x}_t)$ resulting from $\mathrm{P}(\boldsymbol{x}_s) \neq \mathrm{P}(\boldsymbol{x}_t)$. However, the label difference (i.e., different fault types) may exist in practical RMFD. The principle of knowledge transfer between multiple fault types is that model trained using data from some types of faults in source domain can be applied to process data from other types of faults in target domain. Its goal is to obtain effective diagnostic results when unknown fault types appear in target domain. Based on the gear fault diagnosis examples shown in Fig.9, three knowledge transfer model can be summarized: the similar tasks transfer model, the over-tasks transfer model, and the under-tasks transfer model.

2) The similar tasks transfer model. In this model, the fault types in source and target domains are not exactly the same. Han [133] gave an example, where two types (outer race fault, gear root crack) were involved in source domain but other two types (roller fault; tooth surface spalling) were involved in target domain. In such case, successful diagnosis not only depends on the quality of knowledge transfer model, but also relies on the links of fault types between source and target domains. For example, the outer race fault and roller fault are from the same type of bearing, and the gear root crack and tooth surface spalling are from the same type of gear. Kim [180] also proposed a selective parameter freezing (SPF) method for fault diagnosis of rolling element bearings. In the tests, four fault types (normal, ball fault, inner raceway fault, and outer raceway fault) exist in source domain, but other five types (normal, early spalling, early flaking, severe spalling, and severe flaking) exist in target domain. Transfer process was successfully implemented by adjusting the sensitivity to distinguish available or unavailable parameters of the pre-trained network. The key to success in the similar tasks transfer model is that the corresponding fault types in two domains are similar enough.

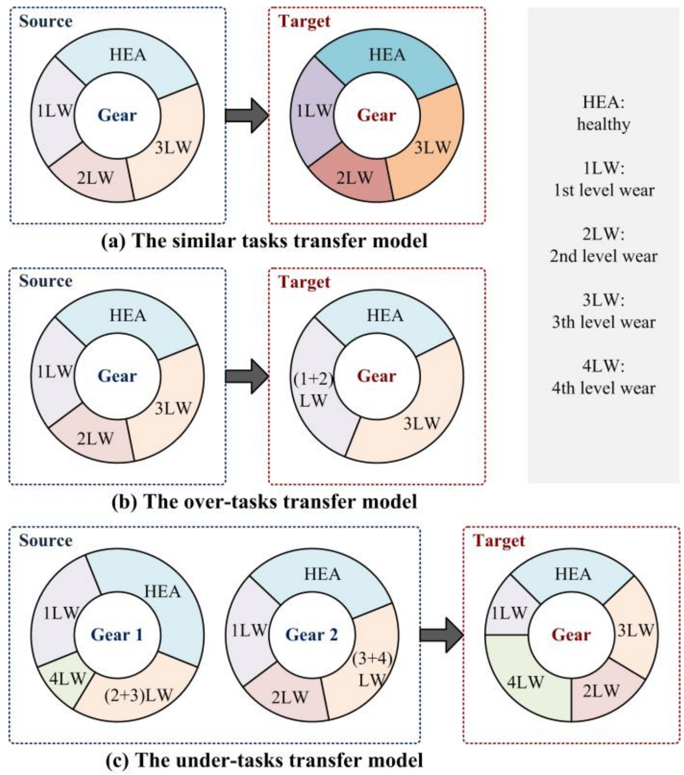

Fig. 9. The gear fault diagnosis examples of three main transfer styles.

3) The over-tasks transfer model. In this model, the number of tasks in source domain is more than those in target domain. Most of existing transfer algorithms is appropriate for this situation, such as deep CNNs model [130], TJM model [148], deep distance metric learning model [149], and TICNN model [131], etc. However, an extra decision algorithm requires to be added, aiming to allocate redundant fault types to existing fault types after transferring. Some basic classifiers are always considered in this stage because the differences between $\mathcal{D}_s$ and $\mathcal{D}_t$ have been weakened after domain adaptation, such as K nearest neighbor (KNN) [181], non-convex sparse regularization [182], and adaptive neuro fuzzy classifier [183] etc. In other words, the over-tasks transfer model will cost more time than basic transfer learning due to the redundant fault types in source domain.

4) The under-tasks transfer model. In this model, although the tasks in each source domain are less than those in the target domain, more than one source domain is selected, thus including all tasks. To solve the under-tasks transfer problem, two general steps are included: Firstly, the basic domain adaptation algorithms are designed to transfer the diagnostic model from each source domain to the target domain. And each domain adaptation may not be the same, which depends on the signal property of each source domain. Secondly, diagnostic results by these domain adaptation algorithms are fused using related measures, such as multi-agent decision fusion [184], class-specific Bayesian fusion [185], Dempster-Shafer evidence theory [186], etc. For example, Wang [187] proposed a hierarchical deep domain adaptation (HDDA) approach to transfer the classifier under one load to identify faults under another load, where the effective information for diagnosis was fused by layer-wisely capturing representative features. Wang's model achieved good under-tasks transfer performance in the fault diagnosis of power plant thermal system. In addition, this type of transfer strategy relies on basic transfer models, and then adds extra steps to adapt the change of fault types. However, the under-tasks transfer model will not be considered if optional source domains are sufficient.

E. Case Studies

1) Descriptions of datasets. To systematically investigate the performance of TL-based RMFD, datasets from four systems were analyzed in this study:

- CWRU bearing dataset. Dataset was collected from a platform provided by CWRU, and included signals from fan end (FE) and drive end (DE). Testing bearings contain four categories: normal condition (NC), inner race fault (IF), outer race fault (OF) and ball fault (BF). Rotating speeds were set from 1730 to1797RPM, and loads were set from 0 to 3HP.
- DDS gear dataset. Dataset was collected from a drivetrain dynamics simulator provided by Yan's group [129], and included signals from six channels in reduction gearbox (RG) and planetary gearbox (PG). Testing gears contain five categories: normal condition (NC), root crack fault (RCF), chipped tooth fault (CTF), miss tooth fault (MTF), surface wear fault (SWF). Rotating speeds were set from 20 to50Hz. Loads were set from 0 to 12.8N·m.
- Qianpeng gear dataset. Dataset was collected from a platform provided by Xi'an Jiaotong University, and included signals from four channels in drive end (DE) and follower end (FE). Testing gears contain three categories: normal condition (NC), root crack fault (RCF), miss tooth fault (MTF). Rotating speeds were set from 30 to 50Hz, and loads were set as 0N·m.
- Induction motor dataset. Dataset was collected from a platform provided by Sun's group [188], and included signals from motor end (ME) and bearing end (BE). Testing motors contain six categories: normal condition (NC), broken rotor bar (BRB), stator winding fault (SWDF), build-in bowed rotor (BBR), build-in rotor unbalance (BRU), build-in fault bearing (BFB). Rotating speed conditions were set from 5 to50Hz, and loads were set as 0N·m.

2) Experimental setup. In this study, Table II lists selected source and target domains with different working conditions, different locations, different machines as well as different fault types. Meanwhile, seven diagnostic models are compared as follows:

- SVM: Support vector machine [40];
- LSSVM: Least square support vector machine [44];
- TCA: Transfer component analysis [76];
- JDA: Joint domain adaptation [77];
- DAM: Domain adaptation machine [49];
- DSM: Domain selection machine [50];
- DTL: Deep transfer learning [132,189,190];

TABLE II
RELATED EXPERIMENT CONDITIONS

No.	Dataset	Channel	Rotating speed	Load
A	CWRU bearing dataset (NC, IF, OF,BF)	FE	1730 RPM	3HP
B	CWRU bearing dataset (NC, IF, OF,BF)	DE	1730 RPM	3HP
C	CWRU bearing dataset (NC, IF, OF,BF)	DE	1797 RPM	0HP
D	DDS gear dataset (NC, RCF, CTF, MTF, SWF)	RG	30Hz	0N·m
E	DDS gear dataset (NC, RCF, CTF, MTF, SWF)	PG	30Hz	0N·m
F	DDS gear dataset (NC, RCF, CTF, MTF, SWF)	PG	30Hz	12.8N·m
G	DDS gear dataset (NC, RCF, CTF)	RG, PG	30Hz	0N·m
H	DDS gear dataset (NC, MTF, SWF)	RG, PG	30Hz	0N·m
I	Qianpeng gear dataset (NC, RCF, MTF)	FE	30Hz	0N·m
J	Qianpeng gear dataset (NC, RCF, MTF)	DE	30Hz	0N·m
K	Qianpeng gear dataset (NC, RCF, MTF)	DE	50Hz	0N·m
L	Induction motor dataset (NC, BRB, SWDF, BBR, BRU, BFB)	ME	30Hz	0N·m
M	Induction motor dataset (NC, BRB, SWDF, BBR, BRU, BFB)	ME	50Hz	0N·m
N	Induction motor dataset (NC, BRB, SWDF, BBR, BRU, BFB)	BE	30Hz	0N·m

In SVM algorithm, the number of training samples was set as 50×N. In other transfer algorithms, the number of samples in target domain and source domain was set as 10×N and 40×N respectively. Finally, the machine fault diagnostic results will be assessed using the root mean square error (RMSE) index, by

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widetilde{y_i} - y_i)^2} \tag{24}$$

where N represents the number of fault types; $\widetilde{y_i}$ and y_i represent the predicted and real fault types.

3) Experimental results. Table III gives the RMSE values under four knowledge transfer strategies with different models.

Some conclusions can be obtained:

Firstly, by comparing SVM with other transfer models, the RMSE value of the former are much greater than those of the latter. It proves the effectiveness of knowledge transfer on rotary machine fault diagnosis field. During varying working conditions, DSM is superior to others because the difference on time frequency plane can be adjusted well by this kind of domain adaptation machine. During varying locations, TCA has the highest diagnostic performance, thus illustrating that the mapping feature spaces are close between different locations. During varying machines and fault types, DTL model can obtain the best transfer results, which shows that the main transfer components in these two cases are the abstract features.

TABLE III
THE RMSE VALUES IN TRANSFER TESTS

Model	Transfer between multiple working conditions			Transfer between multiple locations		
	B→C	E→F	L→M	A→B	E→D	N→L
SVM	28.11	25.78	28.03	36.76	40.77	38.04
LSSVM	17.03	15.94	17.66	19.78	23.65	20.32
TCA	16.44	14.78	19.20	**7.09**	**10.57**	**9.42**
JDA	11.99	10.05	13.51	12.66	15.87	13.23
DAM	11.30	10.95	12.04	10.75	18.56	14.07
DSM	**7.03**	**6.79**	**9.11**	9.53	12.54	11.01
DTL	10.7	9.16	13.06	8.96	10.87	10.55
Model	Transfer between multiple machines			Transfer between multiple fault types		
	D→I	E→I	I→D	D→G	{G,H}→D	
SVM	47.12	50.77	45.25	24.30	30.21	
LSSVM	20.56	24.63	18.62	17.33	21.46	
TCA	13.20	14.30	13.28	10.94	17.22	
JDA	15.94	16.30	10.83	15.30	19.36	
DAM	14.00	18.29	16.24	10.03	15.28	
DSM	14.17	15.63	12.02	9.40	15.00	
DTL	**10.74**	**12.11**	**9.96**	**7.12**	**10.38**	

Secondly, by observing the RMSE values when transfer between multiple working conditions, the indexes are ordered by: RMSE(E→F)<RMSE(B→C)<RMSE(L→M), which proves that signals measured from motor are easier to be affected by external conditions than those measured from gears and bearings. On the contrary, signals from gears are the easiest to be affected by sensor locations, especially in the complex mechanical system like DDS.

Thirdly, by observing the RMSE values when transfer between multiple machines, the indexes are ordered by: RMSE(D→I) > RMSE(I→D), which means that it is easier to transfer from simple Qianpeng device to the complex DDS device than transfer from complex system to simple system. Therefore, these tests offer an idea for source machine selection in this case. Meanwhile, by observing the RMSE values when transfer between multiple fault types, the diagnostic performance of over-tasks transfer model (RMSE=7.12) is superior to the under-tasks transfer model (RMSE=10.38).

Through these case studies, along with other experiments in [191-195], transfer learning has been proved a useful tool for rotatory machine fault diagnosis when insufficient data is not available in the target domain.

IV. RESEARCH TRENDS OF KNOWLEDGE TRANSFER IN ROTARY MACHINE FAULT DIAGNOSIS

A. *The Applicability of TL in RMFD*

Suitable knowledge transfer strategies require to be taken into consideration when facing with different RMFD tasks. Referring to [24,169,196-198], we recommend several rules for selecting appropriate transfer strategy.

- ❖ If part of features in source domain is affected by external conditions, the feature-based knowledge transfer strategy is recommended. If part of samples in source domain are affected by external conditions, the instance-based knowledge transfer strategy is recommended;
- ❖ Models like DANN and TCA are helpful to weaken the difference of MPD. Models like LSSVM and TrAdaboost are helpful to weaken the difference of CPD. Models like DAM and DSM are helpful to weaken the difference of both [4];
- ❖ Among available datasets, the priority is sorted by: labeled target domain samples > labeled source domain samples > unlabeled target domain samples > unlabeled source domain samples.

Although rules above are helpful to select a suitable transfer strategy, the design of knowledge transfer model relies on specific RMFD task. In [198], the transfer compact coding for hyper plane classifier (TCCHC) was proposed to assess the transfer ability between historical and current bearing vibration signals using the minimum description length (MDL) theory [199,200]. The historical data with the smallest code length were selected to help learning of current data. By combining TCCHC with exponential semi-deterministic extended Kalman filter (EKF) [201], Shen succeeded in tracking the bearing degradation curves. Tzeng [202] proposed an adversarial discriminative domain adaptation (ADDA) method, which added two domain discriminator modules to maximize loss function [203,204] between source and target domains, as shown in Fig. 10. By the feedback of domain discriminator, the details of selected transfer strategy can be understood, including: generative or discriminative model, tied or untied weights, and adversarial objectives. Despite these studies, how to select a suitable transfer strategy is still challenging due to the uncertainty of signals measured on real mechanical system, thus needing further investigation in the future.

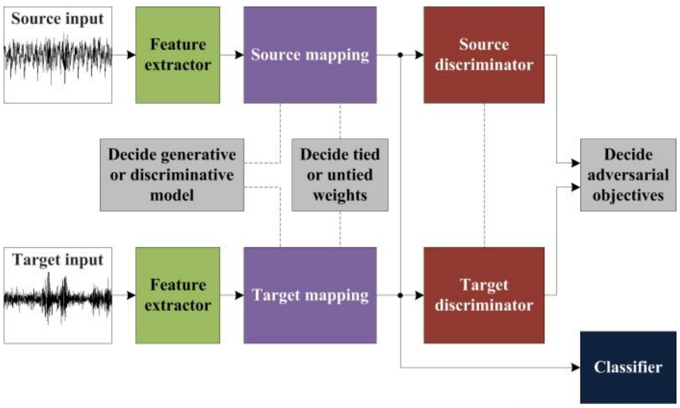

Fig. 10. The adversarial discriminative domain adaptation framework [202].

B. The Effectiveness of TL in RMFD

1) Negative transfer study. Negative transfer [37,205,206] may lead to the failure of RMFD model, which means that the diagnostic performance of transfer learning is even worse than traditional methods. The following strategies can be carried out to avoid negative transfer:

- Excluding the non-useful information from the measured signals. For example, the DTN with JDA framework [133] can exclude the upper layer information and achieved better diagnostic results;
- Evaluating the task similarity between signals, and selecting suitable source domains. For example, the working condition scan method, the signal separation method, and the factor analytical method can be applied.
- Improving classical domain mapping algorithms to achieve similar feature distributions between measured signals. For example, joint domain adaptation (JDA) [77] can be improved to the adaptation regularization based transfer learning (ARTL) [88].

2) Sample quantity study. The sample quantity has direct influence on final diagnostic performance, which always improves with the increase of number of samples in target domain samples. In Cao's [130] experiments, 2% to 80% target samples were included in training dataset and more samples indicated higher diagnostic accuracy. However, too many target samples may bring negative effects if their quality is poor because target samples are always considered as standard in most transfer models. For example, they may be also poor in the time-varying mechanical system [65], so two basic rules are recommended to determine the sample quantity:

- The length of signals in target domain depends on the rate of working conditions. The faster the working condition changes, the less the target samples are selected.
- The signals of at least one running cycle are selected each time for full fault information.

Table IV lists fault diagnostic accuracy under the case of B→C in Table III. It can be seen that the classification accuracy has up to 27.2% improvement if knowledge transfer models are used when the number of samples in target domain is less (n_t=20), but only 2% improvement when the number of samples is large in target domain (n_t=120), which indicates that transfer learning is more suitable for the situation of insufficient data samples in target domain. It is believed that further study is needed on investigating the effect of the sample quantity for performance improvement when the TL-based RMFD technology is applied.

TABLE IV
FAULT DIAGNOSTIC ACCURACIES UNDER THE CASE OF B→C

Model	n_t: n_s				
	20:120	30:120	50:120	80:120	120:120
SVM	0.642 (20:0)	0.781 (30:0)	0.899 (50:0)	0.927 (80:0)	0.939 (120:0)
LSSVM	0.894	0.901	0.923	0.944	0.945
TCA	0.902	0.913	0.925	0.952	0.959
JDA	0.887	0.887	0.901	0.932	0.938
DAM	0.860	0.894	0.925	0.945	0.952
DSM	0.914	0.954	0.975	0.976	0.976
DTL	0.885	0.911	0.937	0.968	0.969

3) Transfer parameter study. Some transfer parameters have been designed to improve the diagnostic accuracy. For example, in Cheng's WD-DTL [132] and Wang's improved deep learning network [207], the max pooling layer was applied to extracting the maximum feature values. As a result, the CNN features within the small window are similar over disjoint regions for fault diagnosis. In Li's deep distance metric learning [150], 2000 epochs were first run to initialize the parameters of the model. Then, the objective function was used for another 2000 training epochs. Consequently, the extracted features were expected to be domain-invariant and robust against noise, that serves the final fault classification. Other parameters have also been designed to improve the diagnostic speed. For example, Tong [162] reduced the impact of discrepancies from both the marginal and conditional distributions simultaneously by resorting the pseudo labels of test data on diagnosis. In Cao's deep convolutional neural network [130], small datasets for gear fault diagnosis were utilized to speed up the implementation of transfer learning. Stochastic gradient descent (SGD) method [208] was also used for network minimization during knowledge transfer-based RMFD.

TABLE V
DETAILED CONFIGURATION OF VGG-16 ARCHITECTURE

Layer	Block	Type	Field size - number of channels	Output
1	Block1	Convolution	3×3-64	224×224×64
2		Convolution	3×3-64	224×224×64
		MaxPooling	2×2	112×112×64
3	Block2	Convolution	3×3-128	112×112×128
4		Convolution	3×3-128	112×112×128
		MaxPooling	2×2	56×56×128
5	Block3	Convolution	3×3-256	56×56×256
6		Convolution	3×3-256	56×56×256
7		Convolution	3×3-256	56×56×256
		MaxPooling	2×2	28×28×256
8	Block4	Convolution	3×3-512	28×28×512
9		Convolution	3×3-512	28×28×512
10		Convolution	3×3-512	28×28×512
		MaxPooling	2×2	14×14×512
11	Block5	Convolution	3×3-512	14×14×512
12		Convolution	3×3-512	14×14×512
13		Convolution	3×3-512	14×14×512
		MaxPooling	2×2	7×7×1024
14	Block6	Fully-connected	1×1×4096	4096
15		Fully-connected	1×1×4096	4096
16		Fully-connected	1×1×4	4

TABLE VI
FAULT DIAGNOSTIC RESULTS UNDER THE CASE OF A→B

Locked layers	Accuracy	Training time (s)
Only Block1	0.960	1664
Block1 ~ Block2	0.957	751
Block1 ~ Block3	0.956	339
Block1 ~ Block4	0.923	153
Block1 ~ Block5	0.890	69

To illustrate the influence of transfer parameters on diagnostic performance, a VGG-16 network in Table V is adopted for the case of A→B in Table III, which contains 6 blocks. Then, the diagnostic results by locking different layers are listed in Table VI. It can be found that the diagnostic accuracy and time consumption both increase with the growth of locked layers. Therefore, the effectiveness of transfer learning is affected by structure of the VGG-16 network.

C. The Flexibility of TL in RMFD

1) Automatic transfer study. As one of the main automated transfer strategies, DTL can adaptively transfer after giving network structure, thus being widely used in rotary machine fault diagnosis, such as VGG-16 [209], LeNet-5 [210], DBN [166] as well as auto-encoder [211] networks. However, some necessary parameters still need to be fine-tuned to match target domain using known knowledge in these models, e.g., the weights of neurons and the locked layers. So, some automatic fine-tuning algorithms have been developed to match different domains without priori knowledge (e.g., rotating speed), such as between-layer scale adjustment, and deep random forest [212]. Self-taught clustering [213] is another automatic transfer strategy, which aims to minimize the reformulated objective function, as

$$\mathcal{J} = I(\mathcal{X}_t;\mathcal{Z}) - I(\widetilde{\mathcal{X}_t};\tilde{\mathcal{Z}}) + \lambda\left[I(\mathcal{X}_s;\mathcal{Z}) - I(\widetilde{\mathcal{X}_s};\tilde{\mathcal{Z}})\right] \quad (25)$$

where target data $\mathcal{X}_t$ and source domain $\mathcal{X}_s$ share the same features clustering $\tilde{\mathcal{Z}}$ on the feature set $\mathcal{Z}$; I denotes the mutual information [214]; $I(\mathcal{X}_t;\mathcal{Z})$- $I(\widetilde{\mathcal{X}_t};\tilde{\mathcal{Z}})$ is the clustering term on the $\mathcal{X}_t$; $I(\mathcal{X}_s;\mathcal{Z})$- $I(\widetilde{\mathcal{X}_s};\tilde{\mathcal{Z}})$ is clustering term on the $\mathcal{X}_s$; λ represents the compensation parameter. Besides, the self-taught clustering offers a valid idea for automatic TL because it is a kind of unsupervised knowledge transfer classifier, like TFusion [215], unsupervised deep learning [216,217]. However, it is hard to make the transfer model automatic compared with traditional models because the parameters in the latter are basically stable and can be pre-trained. So, automatic transfer provides a new direction for future research.

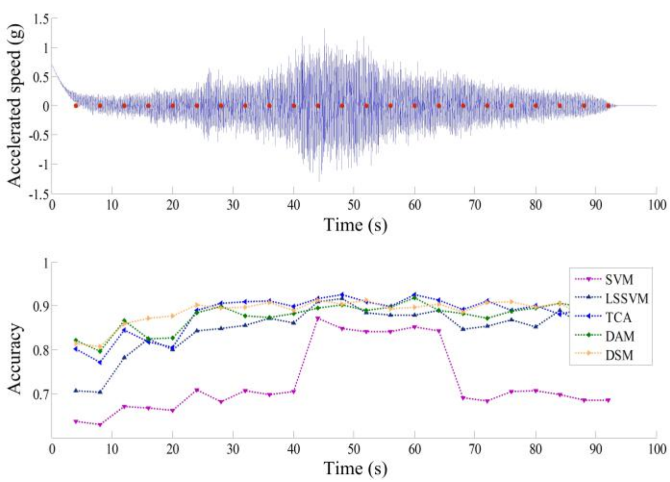

Fig. 11. The diagnostic results of continuous varying rotating speed.

2) Real-time transfer study. In recent years, some fast transfer algorithms have been developed to improve the computational efficiency by 30%~70% in different diagnostic datasets. Xu [218] proposed an online fault diagnosis method based on transfer convolutional neural networks (TCNN), which is made up of an online CNN based on LeNet-5 and several offline CNNs with a shallow structure. Experimental study proved that TCNN method significantly improved the real-time performance and successfully addressed the issue of achieving the desired diagnostic accuracy within limited training time. Except for fast transfer algorithms, the domain updating is another strategy to speed up transfer process by leaving out old training samples. A domain updating framework for real-time transfer is recommended, which includes three major steps: Firstly, all low-quality samples in source domain are discarded and all high-quality samples in target domain are added to source domain. Secondly, the earliest samples in source domain are discarded if too many samples are added into source domain. Finally, all new samples are put into target domain. To verify real-time capability of domain updating framework, the induction motor dataset is adopted as an example in this study, and Fig. 11 illustrates the vibration signal and diagnostic results under continuous varying rotating speed condition (0→60→0Hz). Unlike SVM, the high diagnostic accuracy of transfer algorithms can be achieved all the time, even though the speed changes rapidly.

V. Conclusions

In this study, we have provided an overview on utilizing transfer learning as a powerful tool for the purpose of fault diagnosis in rotary machines. Most of RMFD studies related to TL in recent years have been reviewed based on the following four categories: transfer between multiple working conditions, transfer between multiple locations, transfer between multiple machines, and transfer between multiple fault types. They possess certain research values under certain circumstances. Although the interest of this research is growing, from the theoretical background of knowledge transfer to applicability of TL in RMFD, from the effectiveness of TL to the flexibility of TL, it should be noted that there still exist many challenges when using knowledge transfer for rotary machine fault diagnosis. For example, there is almost no discussion about the link between effective transfer spaces and real vibration signals. Also there are a few studies on the RMFD application of relevance-based knowledge transfer. However, with the transfer learning technology becoming more and more mature and new theoretical contributions being made, it is believed that transfer learning will be one of the most appealing techniques, like classical machine learning, that make contributions to the filed of RMFD.

References

[1] H. C. Shin *et al.*, “Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning,” *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[2] M. Qiu *et al.*, “Transfer Learning for Context-Aware Question Matching in Information-seeking Conversations in E-commerce,” 2019, pp. 208–213.

[3] J. Park, R. J. Javier, T. Moon, and Y. Kim, “Micro-doppler based classification of human aquatic activities via transfer learning of convolutional neural networks,” *Sensors (Switzerland)*, vol. 16, no. 12, 2016.

[4] K. Weiss, T. M. Khoshgoftaar, and D. D. Wang, “A survey of transfer learning,” *J. Big Data*, vol. 3, no. 1, 2016.

[5] B. Zoph, D. Yuret, J. May, and K. Knight, “Transfer Learning for Low-Resource Neural Machine Translation,” 2016, pp. 1568–1575.

[6] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, and G. Zhang, "Transfer learning using computational intelligence: A survey," *Knowledge-Based Syst.*, vol. 80, pp. 14–23, 2015.
[7] Q. Wang, G. Michau, and O. Fink, "Domain adaptive transfer learning for fault diagnosis," 2019 Prognostics and System Health Management Conference (PHM-Paris), 2019, pp. 279–285.
[8] S. Lu, Q. He, H. Zhang, and F. Kong, "Rotating machine fault diagnosis through enhanced stochastic resonance by full-wave signal construction," *Mech. Syst. Signal Process.*, vol. 85, pp. 82–97, 2017.
[9] W. Li, S. Zhang, and S. Rakheja, "Feature Denoising and Nearest-Farthest Distance Preserving Projection for Machine Fault Diagnosis," *IEEE Trans. Ind. Informatics*, vol. 12, no. 1, pp. 393–404, 2016.
[10] Y. Tian, J. Ma, C. Lu, and Z. Wang, "Rolling bearing fault diagnosis under variable conditions using LMD-SVD and extreme learning machine," *Mech. Mach. Theory*, vol. 90, pp. 175–186, 2015.
[11] R. Liu, B. Yang, E. Zio, and X. Chen, "Artificial intelligence for fault diagnosis of rotating machinery: A review," *Mechanical Systems and Signal Processing*, vol. 108. pp. 33–47, 2018.
[12] D. Yang, Y. Liu, S. Li, X. Li, and L. Ma, "Gear fault diagnosis based on support vector machine optimized by artificial bee colony algorithm," *Mech. Mach. Theory*, vol. 90, pp. 219–229, 2015.
[13] M. Xia, T. Li, L. Xu, L. Liu, and C. W. De Silva, "Fault Diagnosis for Rotating Machinery Using Multiple Sensors and Convolutional Neural Networks," *IEEE/ASME Trans. Mechatronics*, vol. 23, no. 1, pp. 101–110, 2018.
[14] L. Yang, L. Jing, and M. K. Ng, "Robust and non-negative collective matrix factorization for text-to-image transfer learning," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4701–4714, 2015.
[15] Y. Zhu *et al.*, "Heterogeneous transfer learning for image classification," in *Proceedings of the National Conference on Artificial Intelligence*, 2011.
[16] N. Zou, Y. Zhu, J. Zhu, M. Baydogan, W. Wang, and J. Li, "A Transfer Learning Approach for Predictive Modeling of Degenerate Biological Systems," *Technometrics*, vol. 57, no. 3, pp. 362–373, 2015.
[17] W. Zhang *et al.*, "Deep Model Based Transfer and Multi-Task Learning for Biological Image Analysis," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, 2015, pp. 1475–1484.
[18] L. F. Villa, A. Reñones, J. R. Perán, and L. J. De Miguel, "Statistical fault diagnosis based on vibration analysis for gear test-bench under non-stationary conditions of speed and load," in *Mechanical Systems and Signal Processing*, 2012, vol. 29, pp. 436–446.
[19] S. M. A. Cruz, "An active-reactive power method for the diagnosis of rotor faults in three-phase induction motors operating under time-varying load conditions," *IEEE Trans. Energy Convers.*, vol. 27, no. 1, pp. 71–84, 2012.
[20] T. Ince, S. Kiranyaz, L. Eren, M. Askar, and M. Gabbouj, "Real-Time Motor Fault Detection by 1-D Convolutional Neural Networks," *IEEE Trans. Ind. Electron.*, vol. 63, no. 11, pp. 7067–7075, 2016.
[21] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10. pp. 1345–1359, 2010.
[22] M. E. Taylor and P. Stone, "Transfer learning for reinforcement learning domains: A survey," *J. Mach. Learn. Res.*, vol. 10, pp. 1633–1685, 2009.
[23] HL. Zheng, R.X. Wang, Y.T. Yang, J.C. Yin, Y.B. Li, Y.Q. Li, and M.Q. Xu, "Cross-domain fault diagnosis using knowledge transfer strategy: a review," *IEEE Access*, vol. 7, pp. 115368–115377, 2019.
[24] B. Yang, Y. Lei, F. Jia, and S. Xing, "An intelligent fault diagnosis approach based on transfer learning from laboratory bearings to locomotive bearings," *Mech. Syst. Signal Process.*, vol. 122, pp. 692–706, 2019.
[25] 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.
[26] 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, 2012.
[27] X. Glorot, A. Bordes, and Y. Bengio, "Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach," *Proc. 28th Int. Conf. Mach. Learn.*, no. 1, pp. 513–520, 2011.
[28] B. Gong, K. Grauman, and F. Sha, "Geodesic flow kernel and landmarks: kernel methods for unsupervised domain adaptation," in *Advances in Computer Vision and Pattern Recognition*, no. 9783319583464, 2017, pp. 59–79.
[29] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2066–2073.
[30] F. Li, S. J. Pan, O. Jin, Q. Yang, and X. Zhu, "Cross-domain co-extraction of sentiment and topic lexicons," in *50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference*, 2012, vol. 1.
[31] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1717–1724.
[32] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, "Cross-domain sentiment classification via spectral feature alignment," in *Proceedings of the 19th international conference on World wide web - WWW '10*, 2010, p. 751.
[33] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Networks*, vol. 22, no. 2, pp. 199–210, 2011.
[34] H. Zhao, S. Zhang, G. Wu, J. P. Costeira, J. M. F. Moura, and G. J. Gordon, "Multiple Source Domain Adaptation with Adversarial Training of Neural Networks," *Adv. Neural Inf. Process. Syst. 31*, 2017.
[35] H. Daumé, "Frustratingly Easy Domain Adaptation," *Coling*, 2009.
[36] T. Tommasi, F. Orabona, and B. Caputo, "Safety in numbers: Learning categories from few examples with multi model knowledge transfer," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3081–3088.
[37] P. Huang, G. Wang, and S. Qin, "Boosting for transfer learning from multiple data sources," *Pattern Recognit. Lett.*, vol. 33, no. 5, pp. 568–579, 2012.
[38] A. Rozantsev, M. Salzmann, and P. Fua, "Residual Parameter Transfer for Deep Domain Adaptation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4339–4348.
[39] Y. Zhou, T. M. Hospedales, and N. Fenton, "When and where to transfer for Bayesian network parameter learning," *Expert Syst. Appl.*, vol. 55, pp. 361–373, 2016.
[40] S. Mahadevan and S. L. Shah, "Fault detection and diagnosis in process data using one-class support vector machines," *J. Process Control*, vol. 19, no. 10, pp. 1627–1639, 2009.
[41] Z. Liu, H. Cao, X. Chen, Z. He, and Z. Shen, "Multi-fault classification based on wavelet SVM with PSO algorithm to analyze vibration signals from rolling element bearings," *Neurocomputing*, vol. 99, pp. 399–410, 2013.
[42] J. Zheng, H. Pan, and J. Cheng, "Rolling bearing fault detection and diagnosis based on composite multiscale fuzzy entropy and ensemble support vector machines," *Mech. Syst. Signal Process.*, vol. 85, pp. 746–759, 2017.
[43] T. Tommasi, F. Orabona, and B. Caputo, "Learning categories from few examples with multi model knowledge transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 928–941, 2014.
[44] W. Deng, R. Yao, H. Zhao, X. Yang, and G. Li, "A novel intelligent diagnosis method using optimal LS-SVM with improved PSO algorithm," *Soft Comput.*, vol. 23, no. 7, pp. 2445–2462, 2019.
[45] F. Tan, M. Yin, L. Wang, and G. Yin, "Spindle thermal error robust modeling using LASSO and LS-SVM," *Int. J. Adv. Manuf. Technol.*, vol. 94, no. 5–8, pp. 2861–2874, 2018.
[46] 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, 2010.
[47] S. Zhou, W. Li, C.R. Cox, et al. "Domain Independent SVM for Transfer Learning in Brain Decoding", 2019.
[48] B. B. Damodaran, N. Courty, and S. Lefevre, "Sparse Hilbert Schmidt Independence Criterion and Surrogate-Kernel-Based Feature Selection for Hyperspectral Image Classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 4, pp. 2385–2398, 2017.
[49] L. Duan, D. Xu, and I. W. H. Tsang, "Domain adaptation from multiple sources: A domain-dependent regularization approach," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 23, no. 3, pp. 504–518, 2012.
[50] L. Duan, D. Xu, and S. F. Chang, "Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1338–1345.
[51] A. Krizhevsky and G. Hinton, "Using Very Deep Autoencoders for Content-Based Image Retrieval," *Proc. Eur. Symp. Artif. Neural Networks*, pp. 1–7, 2011.
[52] M. Simon and E. Rodner, "Neural activation constellations: Unsupervised part model discovery with convolutional networks," in *Proceedings of the*

IEEE International Conference on Computer Vision, 2015, vol. 2015 International Conference on Computer Vision, ICCV 2015, pp. 1143–1151.

[53] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*, 2016.

[54] D. Han, Q. Liu, and W. Fan, "A new image classification method using CNN transfer learning and web data augmentation," *Expert Syst. Appl.*, vol. 95, pp. 43–56, 2018.

[55] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep Learning for Emotion Recognition on Small Datasets using Transfer Learning," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI '15*, 2015, pp. 443–449.

[56] F. Doshi-Velez and G. Konidaris, "Hidden parameter markov decision processes: A semiparametric regression approach for discovering latent task parametrizations," in *IJCAI International Joint Conference on Artificial Intelligence*, 2016, vol. 2016-January, pp. 1432–1440.

[57] T. Killian, S. Daulton, G. Konidaris, et al. "Robust and Efficient Transfer Learning with Hidden-Parameter Markov Decision Processes", 2017.

[58] S. Mei, "Multi-kernel transfer learning based on Chou's PseAAC formulation for protein submitochondria localization," *J. Theor. Biol.*, vol. 293, pp. 121–130, 2012.

[59] A. Argyriou, A. Maurer, and M. Pontil, "An algorithm for transfer learning in a heterogeneous environment," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2008, vol. 5211 LNAI, no. PART 1, pp. 71–85.

[60] H. Chang, J. Han, C. Zhong, A. M. Snijders, and J. H. Mao, "Unsupervised Transfer Learning via Multi-Scale Convolutional Sparse Coding for Biomedical Applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1182–1194, 2018.

[61] W. Fischer, S. S. Moudgalya, J. D. Cohn, N. T. T. Nguyen, and G. T. Kenyon, "Sparse coding of pathology slides compared to transfer learning with deep neural networks," *BMC Bioinformatics*, vol. 19, 2018.

[62] X. Huang, Y. Rao, H. Xie, T.-L. Wong, and F. L. Wang, "Cross-domain sentiment classification via topic-related tradaboost," in *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, 2017.

[63] K. Li and J. C. Principe, "Transfer Learning in Adaptive Filters: The Nearest Instance Centroid-Estimation Kernel Least-Mean-Square Algorithm," *IEEE Trans. Signal Process.*, vol. 65, no. 24, pp. 6520–6535, 2017.

[64] B. Liu, Y. Xiao, and Z. Hao, "A Selective Multiple Instance Transfer Learning Method for Text Categorization Problems," *Knowledge-Based Syst.*, vol. 141, pp. 178–187, 2018.

[65] F. Shen, C. Chen, R. Yan, and R. X. Gao, "Bearing fault diagnosis based on SVD feature extraction and transfer learning classification," in *Proceedings of 2015 Prognostics and System Health Management Conference, PHM 2015*, 2016.

[66] R. Chattopadhyay, Q. Sun, W. Fan, I. Davidson, S. Panchanathan, and J. Ye, "Multisource domain adaptation and its application to early detection of fatigue," *ACM Trans. Knowl. Discov. Data*, vol. 6, no. 4, pp. 1–26, 2012.

[67] Y. Cheng, G. Cao, X. Wang, and J. Pan, "Weighted multi-source TrAdaBoost," *Chinese J. Electron.*, 2013.

[68] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997.

[69] D. Kotzias, M. Denil, N. de Freitas, and P. Smyth, "From Group to Individual Labels Using Deep Features," 2015, pp. 597–606.

[70] J. Xie, M. Kiefel, M. T. Sun, and A. Geiger, "Semantic Instance Annotation of Street Scenes by 3D to 2D Label Transfer," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016-December, pp. 3688–3697.

[71] D. Haase, E. Rodner, and J. Denzler, "Instance-weighted transfer learning of active appearance models," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1426–1433.

[72] X. Huang and S. Belongie, "Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, vol. 2017-October, pp. 1510–1519.

[73] B. Babenko, "Multiple instance learning: algorithms and applications," *View Artic. PubMed/NCBI Google Sch.*, pp. 1–19, 2008.

[74] B. Tan, E. Zhong, E. W. Xiang, and Q. Yang, "Multi-transfer: Transfer learning with multiple views and multiple sources," *Stat. Anal. Data Min.*, vol. 7, no. 4, pp. 282–293, 2014.

[75] H. Ren, W.Y. Liu, M.C. Shan, and X. Wang, "A new wind turbine health condition monitoring method based on VMD-MPE and feature-based transfer learning", *Measurement*, vol.148, 2019.

[76] G. Matasci, M. Volpi, M. Kanevski, L. Bruzzone, and D. Tuia, "Semisupervised Transfer Component Analysis for Domain Adaptation in Remote Sensing Image Classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3550–3564, 2015.

[77] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2200–2207.

[78] I. O. Tolstikhin, B. K. Sriperumbudur, and P. B. Schölkopf, "Minimax Estimation of Maximum Mean Discrepancy with Radial Kernels," *Iclr*, no. Nips, pp. 1930–1938, 2016.

[79] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo, "Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, vol. 2017-January, pp. 945–954.

[80] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, 2006.

[81] S. Pang, S. Leung, I. Ben Nachum, Q. Feng, and S. Li, "Direct automated quantitative measurement of spine via cascade amplifier regression network," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, vol. 11071 LNCS, pp. 940–948.

[82] Y. Zhang, Z. Qiu, T. Yao, D. Liu, and T. Mei, "Fully Convolutional Adaptation Networks for Semantic Segmentation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6810–6818.

[83] H. Nam and B. Han, "Learning Multi-domain Convolutional Neural Networks for Visual Tracking," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016-December, pp. 4293–4302.

[84] M. Long, J. Wang, Y. Cao, J. Sun, and P. S. Yu, "Deep learning of transferable representation for scalable domain adaptation," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 8, pp. 2027–2040, 2016.

[85] A. Gretton and L. Györfi, "Consistent Nonparamtric Tests of Independence," *J. Mach. Learn. Res.*, vol. 11 (20, pp. 1391–1423, 2010.

[86] C. Kandaswamy, L. M. Silva, L. A. Alexandre, R. Sousa, J. M. Santos, and J. M. De Sá, "Improving transfer learning accuracy by reusing stacked denoising autoencoders," in *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, 2014, vol. 2014-January, no. January, pp. 1380–1387.

[87] M. Rajesh and Manikanthan, "Annoyed realm outlook taxonomy using twin transfer learning," *Int. J. Pure Appl. Math.*, vol. 116, no. 21 Special Issue, pp. 549–557, 2017.

[88] 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, 2014.

[89] H. V. Nguyen, H. T. Ho, V. M. Patel, and R. Chellappa, "DASH-N: Joint Hierarchical Domain Adaptation and Feature Learning," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5479–5491, 2015.

[90] B. Kulis, K. Saenko, and T. Darrell, "What you saw is not what you get: Domain adaptation using asymmetric kernel transforms," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1785–1792.

[91] M. Kandemir, "Asymmetric transfer learning with deep Gaussian processes," in *32nd International Conference on Machine Learning, ICML 2015*, 2015, vol. 1.

[92] M. Harel and S. Mannor, "Learning from Multiple Outlooks," *Icml*, 2011.

[93] P. Rebentrost, A. Steffens, I. Marvian, and S. Lloyd, "Quantum singular-value decomposition of nonsparse low-rank matrices," *Phys. Rev. A*, vol. 97, no. 1, 2018.

[94] H. Li, Y. Kluger, and M. Tygert, "Randomized algorithms for distributed computation of principal component analysis and singular value decomposition," *Adv. Comput. Math.*, vol. 44, no. 5, pp. 1651–1672, 2018.

[95] C. Wang and S. Mahadevan, "Heterogeneous domain adaptation using manifold alignment," in *IJCAI International Joint Conference on Artificial Intelligence*, 2011, pp. 1541–1546.

[96] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, "Comparative Study of Attribute Selection Using Gain Ratio and Correlation Based Feature Selection," *Int. J. Inf. Technol. Knowl. Manag.*, vol. 2, no. 2, pp. 271–277, 2010.

[97] S. Bahassine, A. Madani, M. Al-Sarem, and M. Kissi, "Feature selection using an improved Chi-square for Arabic text classification," *Journal of King Saud University - Computer and Information Sciences*, 2018.

[98] P. Gupta and T. Dallas, "Feature selection and activity recognition system using a single triaxial accelerometer," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 6, pp. 1780–1786, 2014.

[99] Z. T. Liu, M. Wu, W. H. Cao, J. W. Mao, J. P. Xu, and G. Z. Tan, "Speech emotion recognition based on feature selection and extreme learning machine decision tree," *Neurocomputing*, vol. 273, pp. 271–280, 2018.

[100] J. B. and C. C. Guang Qiu, Bing Liu., "Expanding Domain Sentiment Lexicon throughDouble Popagation.pdf," in *IJCAI International Joint Conference on Artificial Intelligence*, 2009.

[101] N. Jakob and I. Gurevych, "Extracting opinion targets in a single-and cross-domain setting with conditional random fields," *Proc. 2010 Conf. Empir. Methods Nat. Lang. Process.*, no. October, pp. 1035–1045, 2010.

[102] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Web-scale training for face identification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, vol. 07-12-June-2015, pp. 2746–2754.

[103] A. Nibali, Z. He, and D. Wollersheim, "Pulmonary nodule classification with deep residual networks," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 12, no. 10, pp. 1799–1808, 2017.

[104] L. Zhang, W. Zuo, and D. Zhang, "LSDT: Latent Sparse Domain Transfer Learning for Visual Adaptation," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1177–1191, 2016.

[105] L. Mihalkova, T. N. Huynh, and R. J. Mooney, "Mapping and Revising Markov Logic Networks for Transfer Learning," *Proc. Twenty-Second Conf. Artif. Intell.*, pp. 608–614, 2007.

[106] L. Torrey and J. Shavlik, "Policy transfer via Markov logic networks," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2010, vol. 5989 LNAI, pp. 234–248.

[107] A. Yenter and A. Verma, "Deep CNN-LSTM with combined kernels from multiple branches for IMDb review sentiment analysis," in *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON 2017*, 2018, vol. 2018-January, pp. 540–546.

[108] H. Poon, P. Domingos, and M. Sumner, "A General Method for Reducing the Complexity of Relational Inference And its Application to MCMC," *Proc. 23rd Natl. Conf. Artif. Intell.*, 2008.

[109] M. Labani, P. Moradi, F. Ahmadizar, and M. Jalili, "A novel multivariate filter method for feature selection in text classification problems," *Eng. Appl. Artif. Intell.*, vol. 70, pp. 25–37, 2018.

[110] J. Davis and P. Domingos, "Deep Transfer: A Markov Logic Approach," *AI Mag.*, vol. 32, no. 1, p. 51, 2011.

[111] V. Sharma and A. Parey, "A Review of Gear Fault Diagnosis Using Various Condition Indicators," in *Procedia Engineering*, 2016, vol. 144, pp. 253–263.

[112] M. Cerrada, R. V. Sánchez, F. Pacheco, D. Cabrera, G. Zurita, and C. Li, "Hierarchical feature selection based on relative dependency for gear fault diagnosis," *Appl. Intell.*, vol. 44, no. 3, pp. 687–703, 2016.

[113] G. Cheng, X. Chen, H. Li, P. Li, and H. Liu, "Study on planetary gear fault diagnosis based on entropy feature fusion of ensemble empirical mode decomposition," *Meas. J. Int. Meas. Confed.*, vol. 91, pp. 140–154, 2016.

[114] Z. Li, Y. Jiang, C. Hu, and Z. Peng, "Recent progress on decoupling diagnosis of hybrid failures in gear transmission systems using vibration sensor signal: A review," *Measurement: Journal of the International Measurement Confederation*, vol. 90. pp. 4–19, 2016.

[115] J. Ben Ali, N. Fnaiech, L. Saidi, B. Chebel-Morello, and F. Fnaiech, "Application of empirical mode decomposition and artificial neural network for automatic bearing fault diagnosis based on vibration signals," *Appl. Acoust.*, vol. 89, pp. 16–27, 2015.

[116] X. Zhang, Y. Liang, J. Zhou, and Y. Zang, "A novel bearing fault diagnosis model integrated permutation entropy, ensemble empirical mode decomposition and optimized SVM," *Meas. J. Int. Meas. Confed.*, vol. 69, pp. 164–179, 2015.

[117] J. Chen, J. Pan, Z. Li, Y. Zi, and X. Chen, "Generator bearing fault diagnosis for wind turbine via empirical wavelet transform using measured vibration signals," *Renew. Energy*, vol. 89, pp. 80–92, 2016.

[118] J. Wang, J. Zhang, C. Chen, F. Tian, and L. Cui, "Basic pursuit of an adaptive impulse dictionary for bearing fault diagnosis," in *Proceedings - 2014 International Conference on Mechatronics and Control, ICMC 2014*, 2015, pp. 2425–2430.

[119] F.-J. Yu, F.-X. Zhou, and B.-K. Yan, "Bearing initial fault feature extraction via sparse representation based on dictionary learning," *Zhendong yu Chongji/Journal Vib. Shock*, vol. 35, no. 6, 2016.

[120] L.-L. Cui, J. Wang, N. Wu, and L.-X. Gao, "Bearing fault diagnosis based on self-adaptive impulse dictionary matching pursuit," *Zhendong yu Chongji/Journal Vib. Shock*, vol. 33, no. 11, 2014.

[121] W. Sun, R. Zhao, R. Yan, S. Shao, and X. Chen, "Convolutional Discriminative Feature Learning for Induction Motor Fault Diagnosis," *IEEE Trans. Ind. Informatics*, vol. 13, no. 3, pp. 1350–1359, 2017.

[122] E. Germen, M. Başaran, and M. Fidan, "Sound based induction motor fault diagnosis using Kohonen self-organizing map," *Mech. Syst. Signal Process.*, vol. 46, no. 1, pp. 45–58, 2014.

[123] K. M. Siddiqui, K. Sahay, and V. K. Giri, "Health Monitoring and Fault Diagnosis in Induction Motor- A Review," *Int. J. Adv. Res. Electr. Electron. Instrum. Eng.*, 2014.

[124] H. Liu, X. Wang, and C. Lu, "Rolling bearing fault diagnosis under variable conditions using hilbert-huang transform and singular value decomposition," *Math. Probl. Eng.*, vol. 2014, 2014.

[125] Y. Wang, G. Xu, L. Liang, and K. Jiang, "Detection of weak transient signals based on wavelet packet transform and manifold learning for rolling element bearing fault diagnosis," *Mech. Syst. Signal Process.*, vol. 54, pp. 259–276, 2015.

[126] J. Burriel-Valencia, R. Puche-Panadero, J. Martinez-Roman, A. Sapena-Bano, and M. Pineda-Sanchez, "Short-Frequency Fourier Transform for Fault Diagnosis of Induction Machines Working in Transient Regime," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 3, pp. 432–440, 2017.

[127] S. Wang, X. Chen, I. W. Selesnick, Y. Guo, C. Tong, and X. Zhang, "Matching synchrosqueezing transform: A useful tool for characterizing signals with fast varying instantaneous frequency and application to machine fault diagnosis," *Mech. Syst. Signal Process.*, vol. 100, pp. 242–288, 2018.

[128] T. Mulumba, A. Afshari, K. Yan, W. Shen, and L. K. Norford, "Robust model-based fault diagnosis for air handling units," *Energy Build.*, vol. 86, pp. 698–707, 2015.

[129] S. Shao, S. McAleer, R. Yan, and P. Baldi, "Highly Accurate Machine Fault Diagnosis Using Deep Transfer Learning," *IEEE Trans. Ind. Informatics*, vol. 15, no. 4, pp. 2446–2455, 2019.

[130] 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, 2018.

[131] 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, 2018.

[132] C. Cheng, B. Zhou, G. Ma, et al. "Wasserstein Distance based Deep Adversarial Transfer Learning for Intelligent Fault Diagnosis", 2019.

[133] T. Han, C. Liu, W. Yang, et al. "Deep Transfer Network with Joint Distribution Adaptation: A New Intelligent Fault Diagnosis Framework for Industry Application", 2018.

[134] H. Liu, C. Liu, and Y. Huang, "Adaptive feature extraction using sparse coding for machinery fault diagnosis," *Mech. Syst. Signal Process.*, vol. 25, no. 2, pp. 558–574, 2011.

[135] A. Widodo and B. S. Yang, "Application of nonlinear feature extraction and support vector machines for fault diagnosis of induction motors," *Expert Syst. Appl.*, vol. 33, no. 1, pp. 241–250, 2007.

[136] B. Cao, N. Nan Liu, and Q. Yang, "Transfer Learning for Collective Link Prediction in Multiple Heterogenous Domains.," in *Icml*, 2010, pp. 159–166.

[137] A. Van Opbroek, M. A. Ikram, M. W. Vernooij, and M. De Bruijne, "A transfer-learning approach to image segmentation across scanners by maximizing distribution similarity," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, vol. 8184 LNCS, pp. 49–56.

[138] P. Hamel, M. E. P. Davies, K. Yoshii, and M. Goto, "Transfer Learning in MIR : Sharing Learned Latent Representations for Music Audio Classification and Similarity," *Proc. ISMIR*, no. Ismir, pp. 9–14, 2013.

[139] I. Grosse, P. Bernaola-Galván, P. Carpena, R. Román-Roldán, J. Oliver, and H. E. Stanley, "Analysis of symbolic sequences using the Jensen-Shannon divergence," *Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top.*, vol. 65, no. 4, p. 16, 2002.

[140] Y. Yuan, M. Chao, and Y. C. Lo, "Automatic Skin Lesion Segmentation Using Deep Fully Convolutional Networks with Jaccard Distance," *IEEE Trans. Med. Imaging*, vol. 36, no. 9, pp. 1876–1886, 2017.

[141] S. Zhang and J. Tang, “Integrating angle-frequency domain synchronous averaging technique with feature extraction for gear fault diagnosis,” *Mech. Syst. Signal Process.*, vol. 99, pp. 711–729, 2018.
[142] B. Sun and K. Saenko, “Deep CORAL: Correlation alignment for deep domain adaptation,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, vol. 9915 LNCS, pp. 443–450.
[143] M. Wang and W. Deng, “Deep visual domain adaptation: A survey,” *Neurocomputing*, vol. 312, pp. 135–153, 2018.
[144] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, “Deep hashing network for unsupervised domain adaptation,” in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, vol. 2017-January, pp. 5385–5394.
[145] C.X. Gao, T Wu, and Z.Y. Fu, "Advanced Rolling Bearing Fault Diagnosis Using Ensemble Empirical Mode Decomposition, Principal Component Analysis and Probabilistic Neural Network", 2018.
[146] Y. Wang, J. Xiang, R. Markert, and M. Liang, “Spectral kurtosis for fault detection, diagnosis and prognostics of rotating machines: A review with applications,” *Mechanical Systems and Signal Processing*, vol. 66–67. pp. 679–698, 2016.
[147] A. Nguyen, J. Yosinski, and J. Clune, “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, vol. 07-12-June-2015, pp. 427–436.
[148] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, “Transfer joint matching for unsupervised domain adaptation,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1410–1417.
[149] J. Wang, W. Feng, Y. Chen, H. Yu, M. Huang, and P. S. Yu, “Visual Domain Adaptation with Manifold Embedded Distribution Alignment,” in *2018 ACM Multimedia Conference on Multimedia Conference - MM ’18*, 2018, pp. 402–410.
[150] X. Li, W. Zhang, and Q. Ding, “A robust intelligent fault diagnosis method for rolling element bearings based on deep distance metric learning,” *Neurocomputing*, vol. 310, pp. 77–95, 2018.
[151] 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. Electr. Eng. Technol.*, vol. 14, no. 1, pp. 407–419, 2019.
[152] L. Wen, L. Gao, Y. Dong, and Z. Zhu, “A negative correlation ensemble transfer learning method for fault diagnosis based on convolutional neural network,” *Math. Biosci. Eng.*, vol. 16, no. 5, pp. 3311–3330, 2019.
[153] F. Jia, Y. Lei, L. Guo, J. Lin, and S. Xing, “A neural network constructed by deep learning technique and its application to intelligent fault diagnosis of machines,” *Neurocomputing*, vol. 272, pp. 619–628, 2018.
[154] F. Jia, Y. Lei, N. Lu, and S. Xing, “Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization,” *Mech. Syst. Signal Process.*, vol. 110, pp. 349–367, 2018.
[155] C. Yi, Y. Lv, and Z. Dang, “A fault diagnosis scheme for rolling bearing based on particle swarm optimization in variational mode decomposition,” *Shock Vib.*, vol. 2016, 2016.
[156] B. Koçer and A. Arslan, “Genetic transfer learning,” *Expert Syst. Appl.*, vol. 37, no. 10, pp. 6997–7002, 2010.
[157] B. Pardo, D. Little, and D. Gergle, “Building a personalized audio equalizer interface with transfer learning and active learning,” 2012, p. 13.
[158] R. C. De Amorim and C. Hennig, “Recovering the number of clusters in data sets with noise features using feature rescaling factors,” *Inf. Sci. (Ny).*, vol. 324, pp. 126–145, 2015.
[159] R. Attux, C. Cavalcante, J. Romano, and R. Suyama, “Blind Source Separation,” in *Unsupervised Signal Processing*, 2011, pp. 181–225.
[160] C. Yi, Y. Lv, H. Xiao, G. You, and Z. Dang, “Research on the Blind Source Separation Method Based on Regenerated Phase-Shifted Sinusoid-Assisted EMD and Its Application in Diagnosing Rolling-Bearing Faults,” *Appl. Sci.*, vol. 7, no. 4, p. 414, 2017.
[161] T. Benkedjouh, N. Zerhouni, and S. Rechak, “Tool wear condition monitoring based on continuous wavelet transform and blind source separation,” *Int. J. Adv. Manuf. Technol.*, vol. 97, no. 9–12, pp. 3311–3323, 2018.
[162] 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, 2018.
[163] 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 *2016 IEEE International Conference on Prognostics and Health Management, ICPHM 2016*, 2016.
[164] J. Wang, J. Xie, L. Zhang, and L. Duan, “A factor analysis based transfer learning method for gearbox diagnosis under various operating conditions,” in *International Symposium on Flexible Automation, ISFA 2016*, 2016, pp. 81–86.
[165] L. Wen, L. Gao, and X. Li, “A New Deep Transfer Learning Based on Sparse Auto-Encoder for Fault Diagnosis,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2017.
[166] Z. Chen and W. Li, “Multisensor feature fusion for bearing fault diagnosis using sparse autoencoder and deep belief network,” *IEEE Trans. Instrum. Meas.*, vol. 66, no. 7, pp. 1693–1702, 2017.
[167] 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.
[168] B. Zhang, W. Li, Z. Tong, et al. "Bearing fault diagnosis under varying working condition based on domain adaptation", 2017.
[169] 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 (Switzerland)*, vol. 17, no. 2, 2017.
[170] 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, 2019.
[171] X. Tang, J. Wang, J. Lu, G. Liu, and J. Chen, “Improving Bearing Fault Diagnosis Using Maximum Information Coefficient Based Feature Selection,” *Appl. Sci.*, vol. 8, no. 11, p. 2143, 2018.
[172] J. Ben Ali, L. Saidi, A. Mouelhi, B. Chebel-Morello, and F. Fnaiech, “Linear feature selection and classification using PNN and SFAM neural networks for a nearly online diagnosis of bearing naturally progressing degradations,” *Eng. Appl. Artif. Intell.*, vol. 42, pp. 67–81, 2015.
[173] C. Shen, F. Liu, D. Wang, A. Zhang, F. Kong, and P. W. Tse, “A Doppler transient model based on the laplace wavelet and spectrum correlation assessment for locomotive bearing fault diagnosis,” *Sensors (Switzerland)*, vol. 13, no. 11, pp. 15726–15746, 2013.
[174] M. C. Su and H. Te Chang, “Fast self-organizing feature map algorithm,” *IEEE Trans. Neural Networks*, vol. 11, no. 3, pp. 721–733, 2000.
[175] M. Abdel-Basset, M. Mohamed, F. Smarandache, and V. Chang, “Neutrosophic association rule mining algorithm for big data analysis,” *Symmetry (Basel).*, vol. 10, no. 4, 2018.
[176] A. M. Palacios, J. L. Palacios, L. Sánchez, and J. Alcalá-Fdez, “Genetic learning of the membership functions for mining fuzzy association rules from low quality data,” *Inf. Sci. (Ny).*, vol. 295, pp. 358–378, 2015.
[177] W. Shi, A. Zhang, and G. I. Webb, “Mining significant crisp-fuzzy spatial association rules,” *Int. J. Geogr. Inf. Sci.*, vol. 32, no. 6, pp. 1247–1270, 2018.
[178] M. Abdel-Basset and M. Mohamed, “The role of single valued neutrosophic sets and rough sets in smart city: Imperfect and incomplete information systems,” *Meas. J. Int. Meas. Confed.*, vol. 124, pp. 47–55, 2018.
[179] J. Ye, “A multicriteria decision-making method using aggregation operators for simplified neutrosophic sets,” *J. Intell. Fuzzy Syst.*, vol. 26, no. 5, pp. 2459–2466, 2014.
[180] 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, 2019.
[181] D. H. Pandya, S. H. Upadhyay, and S. P. Harsha, “Fault diagnosis of rolling element bearing with intrinsic mode function of acoustic emission data using APF-KNN,” *Expert Syst. Appl.*, vol. 40, no. 10, pp. 4137–4145, 2013.
[182] S. Wang, I. Selesnick, G. Cai, Y. Feng, X. Sui, and X. Chen, “Nonconvex Sparse Regularization and Convex Optimization for Bearing Fault Diagnosis,” *IEEE Trans. Ind. Electron.*, vol. 65, no. 9, pp. 7332–7342, 2018.
[183] R. Tiwari, V. K. Gupta, and P. K. Kankar, “Bearing fault diagnosis based on multi-scale permutation entropy and adaptive neuro fuzzy classifier,” *JVC/Journal Vib. Control*, vol. 21, no. 3, pp. 461–467, 2015.
[184] G. Niu, T. Han, B. S. Yang, and A. C. C. Tan, “Multi-agent decision fusion for motor fault diagnosis,” *Mech. Syst. Signal Process.*, vol. 21, no. 3, pp. 1285–1299, 2007.
[185] K. Choi *et al.*, “Novel classifier fusion approaches for fault diagnosis in automotive systems,” *IEEE Trans. Instrum. Meas.*, vol. 58, no. 3, pp. 602–611, 2009.

[186] K. H. Hui, M. H. Lim, M. S. Leong, and S. M. Al-Obaidi, "Dempster-Shafer evidence theory for multi-bearing faults diagnosis," *Eng. Appl. Artif. Intell.*, vol. 57, pp. 160–170, 2017.

[187] X. Wang, H. He, and L. Li, "A Hierarchical Deep Domain Adaptation Approach for Fault Diagnosis of Power Plant Thermal System," *IEEE Trans. Ind. Informatics*, pp. 1–1, 2019.

[188] C. Sun, M. Ma, Z. Zhao, and X. Chen, "Sparse Deep Stacking Network for Fault Diagnosis of Motor," *IEEE Trans. Ind. Informatics*, vol. 14, no. 7, pp. 3261–3270, 2018.

[189] J. Li, X. Li, D. He, and Y. Qu, "A domain adaptation model for early gear pitting fault diagnosis based on deep transfer learning network," *Proc. Inst. Mech. Eng. Part O J. Risk Reliab.*, p. 1748006X1986777, 2019.

[190] Z. He, H. Shao, X. Zhang, J. Cheng, and Y. Yang, "Improved deep transfer auto-encoder for fault diagnosis of gearbox under variable working conditions with small training samples," *IEEE Access*, vol. 7, pp. 115368–115377, 2019.

[191] T. Han, C. Liu, W. Yang, and D. Jiang, "Learning transferable features in deep convolutional neural networks for diagnosing unseen machine conditions," *ISA Trans.*, 2019.

[192] 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, 2019.

[193] 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.*, pp. 1–1, 2019.

[194] X. Li, W. Zhang, Q. Ding, and J. Q. Sun, "Multi-Layer domain adaptation method for rolling bearing fault diagnosis," *Signal Processing*, vol. 157, pp. 180–197, 2019.

[195] 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, 2019.

[196] Y. Lei, N. Li, L. Guo, N. Li, T. Yan, and J. Lin, "Machinery health prognostics: A systematic review from data acquisition to RUL prediction," *Mechanical Systems and Signal Processing*, vol. 104. pp. 799–834, 2018.

[197] G. Csurka, "A comprehensive survey on domain adaptation for visual applications," in *Advances in Computer Vision and Pattern Recognition*, no. 9783319583464, 2017, pp. 1–35.

[198] F. Shen, J. Xu, C. Sun, X. Chen, and R. Yan, "Transfer between multiple working conditions: A new TCCHC-based exponential semi-deterministic extended Kalman filter for bearing remaining useful life prediction," *Meas. J. Int. Meas. Confed.*, vol. 142, pp. 148–162, 2019.

[199] B. Hu, T. Rakthanmanon, Y. Hao, S. Evans, S. Lonardi, and E. Keogh, "Using the minimum description length to discover the intrinsic cardinality and dimensionality of time series," *Data Min. Knowl. Discov.*, vol. 29, no. 2, pp. 358–399, 2014.

[200] Y. Jaw and G. Sohn, "Wind adaptive modeling of transmission lines using minimum description length," *ISPRS J. Photogramm. Remote Sens.*, vol. 125, pp. 193–206, 2017.

[201] M. J. Carr and W. Wang, "An approximate algorithm for prognostic modelling using condition monitoring information," *Eur. J. Oper. Res.*, vol. 211, no. 1, pp. 90–96, 2011.

[202] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, vol. 2017-January, pp. 2962–2971.

[203] X. Guo, L. Chen, and C. Shen, "Hierarchical adaptive deep convolution neural network and its application to bearing fault diagnosis," *Meas. J. Int. Meas. Confed.*, vol. 93, pp. 490–502, 2016.

[204] H. Shao, H. Jiang, H. Zhao, and F. Wang, "A novel deep autoencoder feature learning method for rotating machinery fault diagnosis," *Mech. Syst. Signal Process.*, vol. 95, pp. 187–204, 2017.

[205] W. Dai, O. Jin, G.-R. Xue, Q. Yang, and Y. Yu, "EigenTransfer: a unified framework for transfer learning," in *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, 2009.

[206] Y. Yao and G. Doretto, "Boosting for transfer learning with multiple sources," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1855–1862.

[207]H.Q. Wang, S. Li, L.Y Song, L.L. Cui and P.X. Wang. "An Enhanced Intelligent Diagnosis Method Based on Multi-Sensor Image Fusion via Improved Deep Learning Network," *IEEE Trans. Instrum. Meas.*, 2019.

[208] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, vol. 2017-January, pp. 7130–7138.

[209] Z. Liu, L. Wang, C. Li, and Z. Han, "A High-Precision Loose Strands Diagnosis Approach for Isoelectric Line in High-Speed Railway," *IEEE Trans. Ind. Informatics*, vol. 14, no. 3, pp. 1067–1077, 2018.

[210] 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, 2018.

[211] W. Sun, S. Shao, R. Zhao, R. Yan, X. Zhang, and X. Chen, "A sparse auto-encoder-based deep neural network approach for induction motor faults classification," *Meas. J. Int. Meas. Confed.*, vol. 89, pp. 171–178, 2016.

[212] C. Li, R. V. Sanchez, G. Zurita, M. Cerrada, D. Cabrera, and R. E. Vásquez, "Gearbox fault diagnosis based on deep random forest fusion of acoustic and vibratory signals," *Mech. Syst. Signal Process.*, vol. 76–77, pp. 283–293, 2016.

[213] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Self-taught clustering," 2008, pp. 200–207.

[214] M. Koch-Janusz and Z. Ringel, "Mutual information, neural networks and the renormalization group," *Nat. Phys.*, vol. 14, no. 6, pp. 578–582, 2018.

[215] J. Lv, W. Chen, Q. Li, and C. Yang, "Unsupervised Cross-Dataset Person Re-identification by Transfer Learning of Spatial-Temporal Patterns," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7948–7956.

[216] Y. Bengio, "Deep Learning of Representations for Unsupervised and Transfer Learning," in *JMLR: Workshop and Conference Proceedings*, 2011, vol. 7, pp. 1–20.

[217] G. Mesnil *et al.*, "Unsupervised and Transfer Learning Challenge: a Deep Learning approach," *JMLR W& CP Proc. Unsupervised Transf. Learn. Chall. Work.*, 2012.

[218] G. Xu, M. Liu, Z. Jiang, W. Shen, and C. Huang, "Online Fault Diagnosis Method Based on Transfer Convolutional Neural Networks," *IEEE Transactions on Instrumention and Measurement*,2019.