Artificial Intelligence Review
https://doi.org/10.1007/s10462-022-10230-4

A survey of transfer learning for machinery diagnostics and prognostics

Siya Yao[1,2] · Qi Kang[1,2] · MengChu Zhou[3] · Muhyaddin J. Rawa[4,5] · Abdullah Abusorrah[4,5]

Abstract
In industrial manufacturing systems, failures of machines caused by faults in their key components greatly influence operational safety and system reliability. Many data-driven methods have been developed for machinery diagnostics and prognostics. However, there lacks sufficient labeled data to train a high-performance data-driven model. Moreover, machinery datasets are usually collected from different operation conditions and mechanical components, leading to poor model generalization. To address these concerns, cross-domain transfer learning methods are applied to enhance the feasibility and accuracy of data-driven methods for machinery diagnostics and prognostics. This paper presents a comprehensive survey about how recent studies apply diverse transfer learning methods into machinery tasks including diagnostics and prognostics. Three types of commonly-used transfer methods, i.e., model and parameter transfer, feature matching and adversarial adaptation, are systematically summarized and elaborated on their main ideas, typical models and corresponding representative studies on machinery diagnostics and prognostics. In addition, ten widely-used open-source machinery datasets are presented. Based on recent research progress, this survey expounds emerging challenges and future research directions of transfer learning for industrial applications. This survey presents a systematic review of recent research with clear explanations as well as in-depth insights, thereby helping readers better understand transfer learning for machinery diagnostics and prognostics.

Keywords Transfer learning · Fault diagnosis · Remaining useful life prediction · Domain adaptation · Manufacturing automation

1 Introduction

A manufacturing field is inseparable from the extensive use of various types of machines. In industrial machines, such as induction motors and turbines, some key machinery components such as gears and bearings, tend to have various failures due to their harsh working

✉ Qi Kang
qkang@tongji.edu.cn

✉ MengChu Zhou
zhou@njit.edu

Extended author information available on the last page of the article

Published online: 16 August 2022

environment and long operation time. Such failures may affect the performance of the whole machine, causing huge economic losses or even casualties (Xia et al. 2018; Lei et al. 2018; Li et al. 2020; Chen et al. 2021). To improve the safety and reliability when a machine runs, various advanced signal processing techniques and data-driven methods are developed for machinery diagnosis and prognostics. Machinery diagnosis aims to detect machine faults in time and accurately identify their fault types, which is considered as a classification task by machine learning algorithms. Machinery prognostics refers to the prediction of the remaining useful life (RUL) of a machine or predicts some potential faults in advance, which helps reduce maintenance costs (Zhong et al. 2019; Jiao et al. 2021). Such prognostics is mainly regarded as a regression task in machine learning.

Intelligent data-driven methods are proved to be effective and have many advantages (Zhao et al. 2019; Li et al. 2021; Zhao et al. 2018). Compared with traditional physical model-based methods, they do not require specialized operational mechanisms and high-level expertise to build models and are not limited by corresponding background knowledge (Liu et al. 2022; Ding et al. 2021; Wang et al. 2020). Therefore, in recent years, many studies have been using data-driven learning methods for diagnosis and prognostics (Zhong et al. 2019), and most of them have achieved outstanding results. However, like other machine learning algorithms, the high performance of a data-driven model requires the following two prerequisites: (1) there is sufficient labeled data, namely, historical data is well labeled in terms of whether it is normal or not and what faults have occurred if not; and (2) training and testing data follow the identical distribution, which is an assumption widely used in the application of machine learning algorithms.

However, in a real-world scenario, it is difficult to satisfy such prerequisites. Generally speaking, it is very hard to obtain sufficient labeled data (Ko and Kim 2019; Deng et al. 2021). On the one hand, machines are not allowed to run with faults because severe faults can cause significant losses. Hence, labeled data, especially failure data, are very scarce. On the other hand, some machines have long life cycles, and it is time-consuming and even impossible to obtain complete data that represent their full life cycles. Borrowing data from other relevant datasets to train the model is thus a possible solution. But the data obtained by different machines come from different operating conditions and different working environments, which do not necessarily satisfy the assumption of independent and identical distribution. The model learned directly in this way has poor generalization and sometimes fails completely.

The above problems greatly limit the practical application of machine learning methods for intelligent diagnosis and prognostics. Transfer learning (TL) emerges fortunately and can break such limitations. As an emerging artificial intelligence technology, TL methods have been successfully applied in diverse fields, such as medical image diagnosis, bioinformatics analysis and transportation applications. TL solves the tough problem of training models without labels by generalizing knowledge from related domains. Its main goal is to learn domain-invariant features and reduce the differences among cross-domain features such that models trained on labeled source data can maintain their good performance on target domain data. Since there is distribution divergence between source and target domains, most TL methods focus on matching feature representations to aligning their data distributions. Among existing research, feature-based domain adaption methods are most commonly seen. They make source and target features closer by learning domain-invariant representations. Depending on whether neural networks are used or not, they can be divided into shallow and deep ones. The former maps source and target data into a shared subspace to reduce their distribution divergence, which is simple and efficient to train but has limited accuracy. The latter extracts invariant features by convolutional neural

networks (CNN) which are excellent and powerful when they are used to mine complex structure and learning feature representation. This well suits the TL necessities of feature extraction and distribution matching among different domains. Therefore, deep learning methods have been widely used.

A large number of TL methods have been applied to machinery diagnosis and prognostics. This work provides a comprehensive review of them and aims to make the following contributions:

(1) It summarizes the research progress of TL-based machinery diagnosis and prognosis tasks and provides researchers with a clear and complete understanding of this field. It systematically organizes TL methods and sort out three types of methods: model and parameter transfer, feature matching, and adversarial adaptation-based ones. For each type of TL methods, we conclude its representative algorithms and summarize their advantages, disadvantages, development and evolution.
(2) It puts forward a general framework of cross-domain transfer methods. Existing methods can be viewed as instantiations of such framework with different choices: whether a feature extracting method is shallow or deep; whether source and target feature extractors are shared or not; whether a feature matching term to reduce domain divergence is used or not; and whether an adversarial discriminator is used or not. To the best of our knowledge, it is the first time that various transfer methods are organized in the same framework to show their similarities and strengths, which should inspire researcher in the development of new and novel methods.
(3) It analyzes and compares machinery TL applications, based on which valuable conclusions and useful suggestions are drawn. The machinery application is divided into three transfer settings, which provides a quick start for beginners to grasp the basics of machinery transfer tasks. Ten widely-used datasets in machinery diagnostics and prognostics are described with detailed discussions as well as resource links. The results of different methods on machinery classification and regression tasks (i.e., fault diagnosis and remaining useful life prediction) are summarized, which offers intuitive comparisons in terms of the advantage and performance of various types of TL methods.
(4) Based on comprehensive analyses, it points out the shortcomings of current methods and presents suggestions to extend their applicability. It also indicates future prospects for TL-based machinery applications, and summarizes emerging challenges and future directions.

Firstly, we introduce the basis of TL methods and different settings of transfer tasks in machinery applications. Then, we present a detailed literature review of recent applications that use cross-domain transfer methods for machinery diagnosis and prognostics. Afterward, some popular open-source machinery datasets are described. Last but not least, we present the challenges and future research directions of developing and applying TL methods to machinery diagnosis and prognostics.

2 Background

2.1 Concepts and definitions

In transfer learning, we need to clarify which one is “source” and which one is “target” before conducting any transfer. A dataset with labels functions as the “source” denoted as

Table 1 Symbols and descriptions

Symbols	Description
D_S , D_T	Source/target domain
$\boldsymbol{X_S}$, $\boldsymbol{X_T}$	Source/target data
$\boldsymbol{Y_S}$, $\boldsymbol{Y_T}$	Source/target label
n_1 , n_2	Number of source/target samples
n_c	Number of classes
$\boldsymbol{W}$	Classifier coefficient vector
$\boldsymbol{K}$	Kernel matrix
$\boldsymbol{H}$	Centering matrix
$\boldsymbol{L_M}$	Marginal MMD matrix
$\boldsymbol{L_C}$	Conditional MMD matrix
λ	Coefficient of classifier regularization
α	Coefficient of domain divergence
μ	Coefficient of conditional distribution matching
ϵ	Coefficient of domain discriminator
θ	Network parameters (weights and biases)
η	Learning rate

$D_S = (\boldsymbol{X_S}, \boldsymbol{Y_S})$ where $\boldsymbol{X_S} \in R^{d \times n_1}$ is a d-dimensional feature space, and $\boldsymbol{Y_S} \in R^{n_1}$ is the label and n_1 is the total number of source samples. The dataset without labels is the "target" denoted as $D_T = (\boldsymbol{X_T})$ where $\boldsymbol{X_T} \in R^{d \times n_2}$ and n_2 is the total number of target samples. Some frequently used notations and their descriptions are presented in Table 1.

Definition 1 (*Domain*). A "domain" $D = \{\boldsymbol{X}, P(x)\}$ consists of two main parts: feature data $\boldsymbol{X}$ and its corresponding marginal probability distribution $P(x)$ that falls into [0,1].

Definition 2 (*Task*). A "task" $T = \{\boldsymbol{Y}, f(x)\}$ consists of two parts: labels $\boldsymbol{Y}$ and their corresponding function $f(x)$ that predicts them.

According to above definitions, given a domain D, a task T depicts the relationship between labels and features, and $f(x)$ is the conditional probability distribution $Q(y \mid x)$, which predicts the probability of label y given a sample x.

Given a labeled source domain $D_S = (\boldsymbol{X_S}, \boldsymbol{Y_S})$ and an unlabeled target domain $D_T = (\boldsymbol{X_T})$, let $P(\boldsymbol{X_S})$ and $P(\boldsymbol{X_T})$ be the marginal probability distributions of a source and target, and $Q_S(\boldsymbol{Y_S} \mid \boldsymbol{X_S})$ and $Q_T(\boldsymbol{Y_T} \mid \boldsymbol{X_T})$ be the conditional probability distributions of a source and target respectively. TL assumes that $P(\boldsymbol{X_S}) \neq P(\boldsymbol{X_T})$ and $Q_S(\boldsymbol{Y_S} \mid \boldsymbol{X_S}) \neq Q_T(\boldsymbol{Y_T} \mid \boldsymbol{X_T})$. In plain English, the feature space of source and target domains are different, and so are their prediction functions. This indicates that directly using a source model to predict the labels of a target dataset does not work, especially when there is a large difference between their data distributions. Such TL setting is the most commonly encountered in practice. Therefore, TL methods aim to learn features of D_S and D_T and draw their distributions closer to each other in learned space where their divergence is greatly reduced. Once the domain-invariant features are learned, the classifier trained based on such features of D_S can also perform well on D_T with more accurate predictions than directly re-using a source model.

2.2 Different transfer settings

According to the definitions of TL in machinery diagnosis and prognostics, Table 2 summarizes related studies into three different transfer settings.

(1) Transfer between different working conditions of the same machine. Sensor data under different operating situations and working environments usually have changing features, thus their intrinsic data distributions are different. In other words, data collected under varying working conditions belong to different domains. Hence, learning knowledge from other working conditions can be regarded as a type of cross-domain transfer.
(2) Transfer between machine components. Intuitively, different machine components are installed in different places (Shen et al. 2020), and each of them has its unique characteristic. It is obvious that data of various components are from different domains, so it is also a kind of cross-domain transfer. Compared with the transfer between different working conditions of the same machine, the knowledge transfer from one component to another is more difficult.
(3) Transfer from simulation to real-world. Many researchers build their own experimental platforms in laboratories and artificially simulate various machinery faults. Since the laboratory simulation cannot fully take the environmental noise or disturbance into consideration, it is hard to completely simulate complex working environments of real-world scenarios. The real-world industrial machines are often operated under some harsh settings, which is very hard, even impossible, to be fully simulated. Therefore, there is conspicuous distribution divergence between laboratory simulations and real-world machines, thus be regarded as the toughest transfer.

3 Cross-domain transfer methods

We summarize the cross-domain transfer methods for machinery diagnostics and prognostics into 3 classes: model and parameter transfer-based, feature matching-based, and adversarial adaptation-based ones. Specifically, their general architectures are depicted in Fig. 1. To conclude these methods, we draw a framework in Fig. 2. Existing methods can be viewed as instantiations of such framework with different choices. We have: (1) whether the method used for extracting features from raw data input is shallow or deep; (2) whether the feature extractors for source and target inputs are shared or separated; (3) whether the feature matching term is added to reduce domain divergence; and (4) whether the domain discriminator trained with adversarial objectives is used. Next, we present the basic principles and representative work of each type. A comprehensive overview is presented in Table 3.

3.1 Model and parameter transfer-based methods

Model and parameter transfer-based methods for machinery diagnostics and prognostics are the most commonly used ones. More importantly, they are also served as the basis of other transfer methods. Model transfer is a basic operation in TL. It takes the advantage of pre-trained models and directly transfers their parameters. With model transfer methods,

Table 2 Summary of three transfer seetings

Transfer setting	Fault classification	RUL prediction
Transfer between different working conditions of the same machine	Chen et al. (2021), Deng et al. (2021), Shao et al. (2020), Zhang et al. (2017), Shao et al. (2018), Zhao et al. (2020), Zhang et al. (2020), Zhao et al. (2021), Qian et al. (2021), Ma et al. (2020), Zhang et al. (2020), Wang et al. (2016), Shen et al. (2015), Chen et al. (2021), Azamfar et al. (2020), Si et al. (2021), Wang et al. (2020), Han et al. (2020), Shen et al. (2021), Li et al. (2018), Jin et al. (2021), Zhu et al. (2019), Lu et al. (2021), Li et al. (2020), Lu et al. (2016), Wen et al. (2017), Wang et al. (2019), Han et al. (2019), Michau and Fink (2021), Huang et al. (2021), Jiao et al. (2020), Li et al. (2021), Li et al. (2020), Li et al. (2020), Li et al. (2021), Zhao et al. (2021), Li et al. (2020), Qian et al. (2019), Cheng et al. (2020), Jiao et al. (2020), Yang B et al. (2021), Wen et al. (2017), Qin et al. (2021), Chai et al. (2021)	Ding et al. (2021), Huang et al. (2021), Mao et al. (2019),Wang et al. (2021), Yu et al. (2019), Zhang et al. (2021),da Costa et al. (2020), Zhu et al. (2020), Miao and Yu (2021), Ragab et al. (2020), Ragab et al. (2020), Ding et al. (2021)
Transfer between different machine components	Liu et al. (2022), Shen et al. (2020), Zhao et al. (2020), Chen et al. (2019), Li et al. (2020),Zhiyi et al. (2020), Dong et al. (2022), Liao et al. (2021), Kim et al. (2021), Li et al. (2021),Han et al. (2021), Guo et al. (2018)	Huang et al. (2021), Sun et al. (2018),Xia et al. (2021), Sun et al. (2018)
Transfer from simulation to real-world	Zhiyi et al. (2020), Li et al. (2019), Wu et al. (2020), Yang et al. (2019), Wu et al. (2020)	/

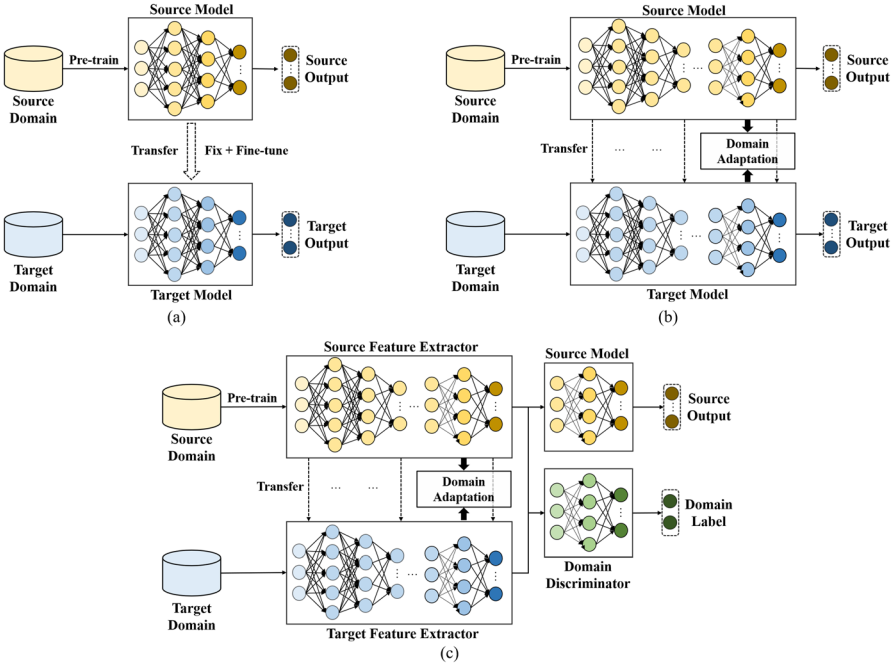


Fig. 1 General architecture of **a** Model and parameter transfer-based methods, **b** Feature matching-based methods (only deep methods are presented), and **c** Adversarial adaptation-based machinery diagnostics and prognostics. Source and target models (or feature extractors) can be shared or separated. For the convenience of reading, we draw two separated models with different color to represent their models

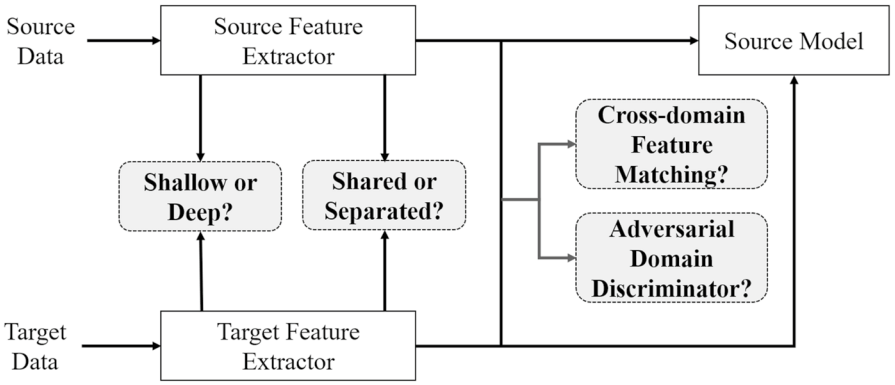


Fig. 2 General framework of cross-domain transfer methods

there is no need to train a model from scratch, thus reducing the training burden. Although both shallow models and deep neural networks models can be transferred, it is more common to transfer neural networks because their structures are more suitable for such transfer. Deep neural networks are regarded as a powerful feature extractor with multiple hidden layers which can directly learn hierarchical features from raw data. Following Yosinski

Table 3 Overview of reviewed transfer methods and related machinery applications

Transfer methods	Typical transfer models	Model architecture				Related machinery applications[1]
		Shallow or deep method	Shared or separated feature extractor	Cross- domain feature matching	Adversarial domain discri- minator	
Model and Parameter Transfer	Pre-train and Fine-tune[2]	Deep	Shared, Separated	None	None	Shao et al. (2020), Zhang et al. (2017), Shao et al. (2018), Zhao et al. (2020),Zhang et al. (2020), Zhao et al. (2021), Huang et al. (2021), Chen et al. (2019), Li et al. (2020), Zhiyi et al. (2020), Dong et al. (2022), Sun et al. (2018), Li et al. (2019), Hasan et al. (2019)
Feature Matching	TCA (Pan et al. 2010)	Shallow	Shared	Marginal MMD	None	Qian et al. (2021), Ma et al. (2020), Zhang et al. (2020), Wang et al. (2016), Shen et al. (2015), Chen et al. (2021), Mao et al. (2019), Wang et al. (2021), Zhang et al. (2020)
	JDA (Long et al. 2013), TJM (Long et al. 2014)	Shallow	Shared	Marginal and Conditional MMD	None	
	CORAL (Sun et al. 2017)	Shallow	Shared	Second- order Correlation	None	
	DDC (Tzeng et al. 2014)	Deep	Separated	Marginal MMD	None	Azamfar et al. (2020), Si et al. (2021), Wang et al. (2020), Han et al. (2020), Shen et al. (2021), Li et al. (2018), Zhu et al. (2019), Yu et al. (2019), Liao et al. (2021), Kim et al. (2021), Wu et al. (2020), Deebak and Al-Turjman (2021)
	DAN (Long et al. 2015)	Deep	Shared	MK-MMD	None	Jin et al. (2021), Lu et al. (2021), Li et al. (2020), Lu et al. (2016), Wen et al. (2017), Wang et al. (2019), Zhang et al. (2021), Li et al. (2021), Yang et al. (2019), Ding et al. (2021)

Table 3 (continued)

Transfer methods	Typical transfer models	Model architecture				Related machinery applications[1]
		Shallow or deep method	Shared or separated feature extractor	Cross- domain feature matching	Adversarial domain discri- minator	
Adversarial Adaptation	DANN (Ganin et al. 2016)	Deep	Shared	None	Gradient Reversal Layer, Minimax Loss	Han et al. (2019), Michau and Fink (2021), Huang et al. (2021), Jiao et al. (2020), Li et al. (2021), Li et al. (2020), Li et al. (2020), da Costa et al. (2020), Zhu et al. (2020), Miao and Yu (2021), Han et al. (2021), Guo et al. (2018), Xia et al. (2021)
	ADDA (Tzeng et al. 2017)	Deep	Separated	None[3]	GAN Loss	Li et al. (2021), Zhao et al. (2021), Li et al. (2020), Ragab et al. (2020), Ragab et al. (2020), Wu et al. (2020)

[1]The listed related machinery applications in each row are built on corresponding transfer methods, and most of them also extend original model architectures.

[2]For model and parameter transfer method, there is no typical model, but it follows the transfer paradigm of pre-train and fine-tune. Almost all deep methods follow it.

[3]In the adversarial adaptation transfer method, no feature matching is used in basic models, i.e., DANN and ADDA, but many of their extensions integrate feature matching in their frameworks to reduce cross-domain feature divergence

et al. (2014), lower layers learn abstract and general features such as edges and curves, which can be applied to common image recognition tasks; and higher layers learn different task-specific and discriminative features which suit different application fields. In the process of model training, the deep architecture is able to automatically select and extract features, and it is trained to learn discriminative features based on the training data such that they are useful for accurate prediction in subsequent classification. However, training such a large and high-performance network from scratch usually requires sufficient labeled training data, many computational resources and considerable time.

Model transfer is applied to overcome the training difficulties. Rather than training and optimizing all networks weights from random initialization, model transfer directly takes the weights that have been trained in another application as the initial weights. Note that, in order to ensure an effective and successful transfer and avoid negative transfer, it is more appropriate to use the knowledge obtained from a different but related application. After transferring the parameters, according to whether the transferred weights are fixed or not, there are two ways to accomplish the model training in terms of the current target task: (1) loading the weights of a pre-trained network as the initial setting and proceeding to update and optimize them based on the target data; and (2) freezing the weights in the lower layer and only updating those in several higher layers when training. Such a process of updating and optimizing the weights in higher layers is called fine-tuning. The latter is less time-consuming because it essentially reduces the number of parameters to be trained, thus it becomes a more widely-used approach. As mentioned before, lower layers extract some common features and higher layers learn task-specific ones. How many lower layers should be fixed and how many higher layers should be fine-tuned depend in part on how different the source and target datasets are from each other. For similar datasets, only fine-tuning the fully connected layers can achieve satisfactory transfer performance. For datasets with considerable differences, more convolutional layers need to be updated. Apart from network weights that need to be optimized via model training, the hyperparameters (such as learning rates, dropout rates, and regularization coefficients) are also required to be tuned.

Model and parameter transfer have been applied to the research on machinery diagnostics and prognostics. Shao et al. (2020) present a modified transfer CNN to diagnose faults in a rotor-bearing system under varying working conditions. Their method belongs to the first kind of model and parameter transfer, i.e., transfer all parameters and then adjust all weights in all layers. As shown in Fig. 3, parameters (weights and bias) of a source CNN are transferred to the target model. Their basic CNN architecture is LeNet-5 that is a classical and concise network including an input layer, two convolutional layers, two pooling layers, and a fully connected layer. Note that, since LeNet-5 has only five layers, updating all weights is not too burdensome. But for larger and deeper CNN with hundreds of layers, training all layers is time-consuming and undesirable. In order to enhance the performance of a basic CNN, stochastic pooling and leaky rectified linear units are developed to form the modified CNN in Shao et al. (2020). As presented in Fig. 3, infrared thermal images are collected for both source and target domains and then input to the modified CNN. They are used for characterizing the health condition of a rotor-bearing system. The procedures of model and parameter transfer in Shao et al. (2020) are quite representative with the below steps:

(1) The thermal images of a rotor-bearing system under different operating conditions are collected, and then converted into grayscale images, and then divided into source and

target domains according to their different working conditions. The source domain has enough labeled samples, but the target one has a small number of labeled samples.
(2) Combine the techniques of stochastic pooling and leaky rectified linear unit to form a modified CNN.
(3) Use sufficient samples from the source domain to train a modified source CNN by minimizing cross-entropy loss between predictions and true labels.
(4) Initialize a target modified CNN with the same structure and hyperparameters as the source model, and directly transfer all the pre-trained weights and biases from the source modified CNN to the target one.
(5) Train the target modified CNN using small samples in the target domain, and optimize and update its weights and biases.
(6) Use the remaining samples in the target domain to test the diagnostic performance of the trained target model.

Many methods are based on a similar transfer paradigm. When only a small number of target data samples are available, we prepare a target model with the same architecture and parameters as a source model, and then update the target model based on training data in the target domain. Chen et al. (2019) present a slightly different transfer scheme for intelligent diagnosis. Their work assumes that the number of categories (i.e., labels) of target tasks are not always equal to those of source tasks. Thus, exactly copying their architectures may not work. To address such issue, they propose to modify the number of softmax output nodes during a transfer stage such that it corresponds to the number of labels in target tasks. Besides, one-dimensional raw vibration signals are used as the input of CNN. Zhang et al. (2017) and Li et al. (2020) indicate that such one-dimensional vibration data is collected by sensors and can be very long. Thus, they adopt a sliding frame to extract samples by small steps, as shown in Fig. 4. All data in a time frame is regarded as a row of the input data. The length of the vibration signal is fixed, and the step size and frame size should be set appropriately so as to generate more effective samples. They also affect the size of the input layer. For example, both a large frame size and a small step size result in a large input size. Since the source and target data are collected under different working conditions, they may have different signal lengths and sample sizes. Therefore, the structure of target models in Zhang et al. (2017) and Li et al. (2020) are altered according to the dimensionalities of target data and labels.

Shao et al. (2018) use a deep transfer framework to accelerate the training of deep neural networks as well as achieve accurate machine fault diagnosis. Figure 5 shows the general pipeline of the proposed framework. They first use a Wavelet transformation to convert the original one-dimensional sensor data into images which present time-frequency distributions. A VGG-16 network pretrained on the ImageNet dataset is used as the source model for transfer. To fit the network architecture of VGG-16, a time window consisting of 1024 data points is converted to a 224×224 time-frequency image. Since the source dataset, i.e., ImageNet dataset that consists of natural images, is considered to be quite different from the studied target dataset that is composed of time-frequency images, three highest-level blocks of a target network are fine-tuned based on labeled training data while the weights of other lower layers are fixed. Three mechanical datasets including induction motors, gearboxes, and bearings with sizes of 5000, 6000 and 9000 time-series samples, are used to verify the effectiveness of the proposed pipeline.

Apart from AlexNet and VGG architectures, the model and parameter transfer method can be applied to an auto-encoder (AE). For rolling bearing fault diagnosis, Li et al. (2019)

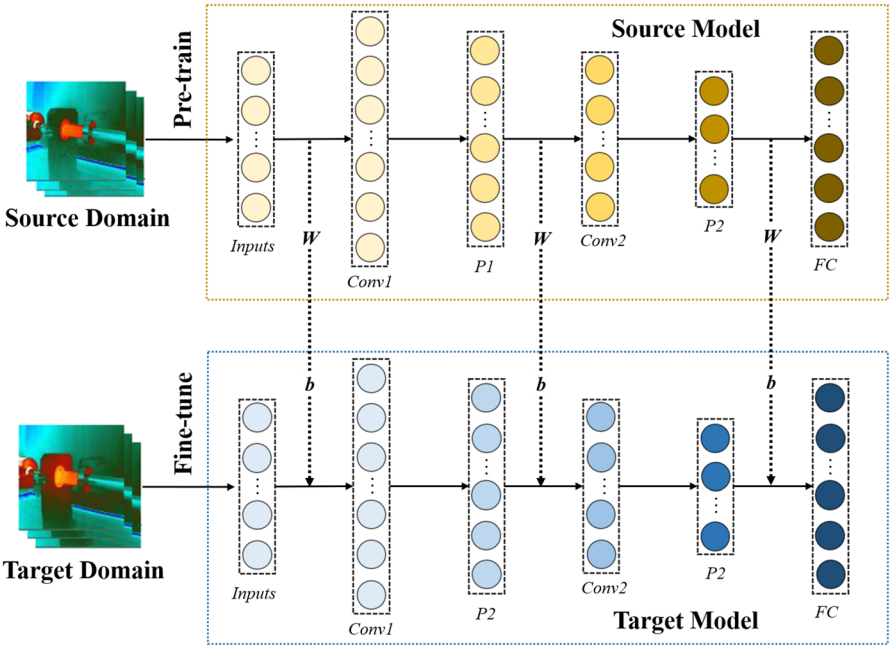


Fig. 3 General framework of CNN transfer (Shao et al. 2020)

construct a deep transfer nonnegativity-constraint sparse AE to tackle the transfer diagnosis problem with limited labeled target data. They first build a base deep AE by stacking multiple nonnegativity-constraint sparse AEs, then train the base model with source data. The model architecture and parameters (including weights and biases) of the source deep AE are transferred to initialize the target model which is then fine-tuned by limited labeled target data. The deep transfer framework is verified by a motor bearing source dataset and a real-world bearing target dataset collected from railway locomotive. The work Zhiyi et al. (2020) applies a similar transfer method in Li et al. (2019), except that it enhances autoencoder by replacing the standard activation function with a scaled exponential linear unit and adding a nonnegative constraint term into its loss function. Sun et al. (2018) select a sparse AE to construct a deep transfer network for RUL prediction. A case study on the cutting tool is performed to validate its effectiveness. Following the idea of model and parameter transfer, in Sun et al. (2018), they directly transfer and then fine-tune weights and bias learned in pre-trained source sparse AE networks.

More related machinery applications based on model and parameter transfer can be found in Zhao et al. (2020), Zhang et al. (2020), Zhao et al. (2021), Huang et al. (2021), Dong et al. (2022) and Hasan et al. (2019).

To conclude, the model and parameter transfer methods accomplish the transfer by using pre-trained models that are already well-trained using other large datasets. A target model can share the network structure, model parameters and model hyperparameters of a pre-trained source model. The majority of this type of research deploys such method in deep neural networks where the lower-level weights of the target model are obtained from a pre-trained model, and the higher-level weights are fine-tuned to fit the specific fault diagnosis or RUL prediction task. By this means, model and parameter transfer can offer reasonable and suitable initialization to a target model and decrease the number of parameters

that need to be updated in a target model, which greatly improves the target model's training process. It is noted that model and parameter transfer-based machinery diagnostics and prognostics methods require the target data to be partially labeled at least because model fine-tune needs such labeled data. Moreover, when the distributions of source and target data are quite different, only fine-tuning is not enough to achieve satisfactory results, thus demanding considerations of feature alignment to be discussed next.

3.2 Feature matching-based methods

Feature matching-based transfer methods aim to reduce the distribution difference between source and target features through feature transformations. With feature matching, the knowledge of source domain can be transferred to target model. Some methods transform the source features so as to match them with the target features, and vice versa. Some methods transform both source and target features into a shared feature space where their distributions are drawn closer and become similar to each other. According to whether deep neural networks are used or not, feature matching-based methods can be divided into shallow methods and deep ones.

3.2.1 Shallow methods

To better present different shallow feature matching approaches, we introduce some classical transfer methods in two classes: (1) distribution adaptation that draws source and target features closer by aligning their statistical features (e.g., align marginal or conditional distributions or both of them); and (2) subspace feature learning which projects original

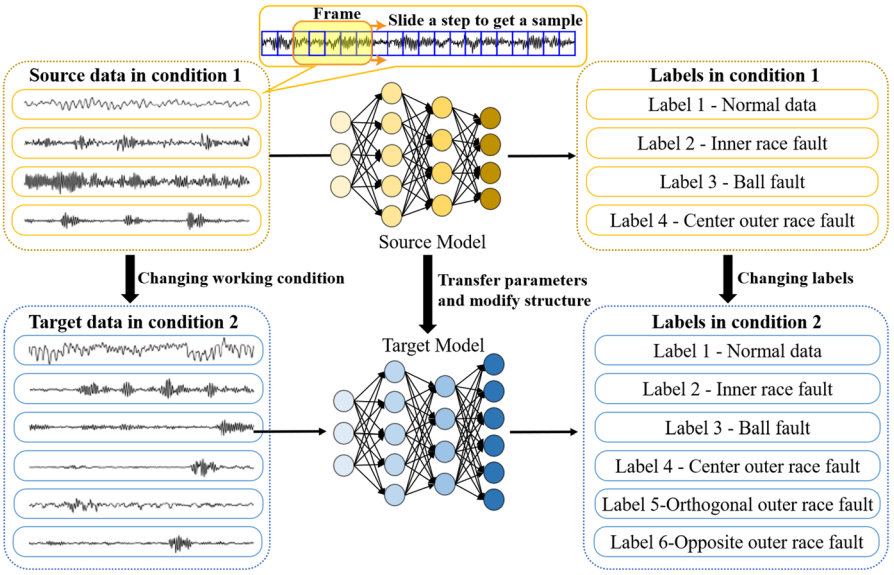


Fig. 4 General framework of fault diagnosis models (Zhang et al. 2017)

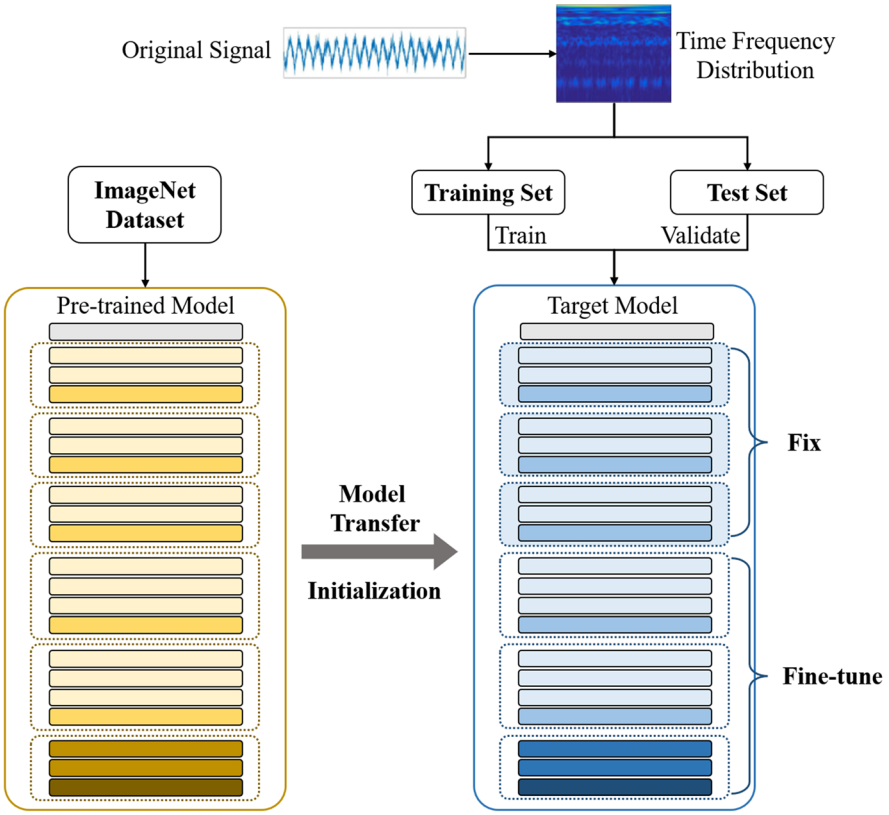


Fig. 5 General pipeline (Shao et al. 2018) based on fine-tuning pre-trained models

source and target features into a shared subspace based on matching statistical features or manifold learning.

(1) Distribution adaptation

Distribution adaptation of source and target features is a widely used TL method. As shown in Fig. 6 , the basic motivation behind this approach is that since the data probability distributions in source and target domains are different, then the most straightforward way is to bring these different data distributions closer by some transformations. According to the nature of data distribution, these methods can be classified into marginal, conditional, and joint distribution adaptation.

Marginal distribution adaptation aims to reduce the difference between the marginal probability distribution of source data and that of target one, i.e., minimizing the distance between $P(\mathbf{X_S})$ and $P(\mathbf{X_T})$:

$$minD(D_S, D_T) = minD(P(\boldsymbol{X_S}), P(\boldsymbol{X_T})) \tag{1}$$

where $D(\cdot, \cdot)$ measures the distance. More specifically, $D(D_S, D_T)$ denotes the distance between source and target data, and $D(P(\boldsymbol{X_S}), P(\boldsymbol{X_T})$ means the distance between source and target marginal probability distributions.

TCA (Transfer Component Analysis) (Pan et al. 2010) is a representative method that applies marginal distribution adaptation. It takes the Maximum Mean Discrepancy (MMD) (Gretton et al. 2006) as its distance metric. MMD is the most frequently used metric in TL, which measures the distance between two distributions in the Reproducing Kernel Hilbert Space (RKHS) (Borgwardt et al. 2006). It is a nonparametric approach to estimate distance and is also a kernel learning method. MMD between $\mathbf{X_S}$ and $\mathbf{X_T}$ is:

$$MMD(\boldsymbol{X_S}, \boldsymbol{X_T}) = \| \frac{1}{n_1} \sum_{i=1}^{n_1} \phi(\boldsymbol{X_S}) - \frac{1}{n_2} \sum_{j=1}^{n_2} \phi(\boldsymbol{X_T}) \|_H^2 \tag{2}$$

where kernel mapping $\phi : x \to \phi(x)$ maps data into RKHS where the inner product calculation in (2) can be shifted to the form of kernel functions. Namely, MMD can be obtained by directly calculating the kernel functions. Moreover, TCA assumes that the feature mapping function ϕ can achieve $P(\phi(\boldsymbol{X_S})) \approx P(\phi(\boldsymbol{X_T}))$, namely, the marginal distribution of source and target are similar after mapping. Therefore, in TCA, the objective of minimizing the domain difference is written as

$$\begin{aligned} minD(D_S, D_T) &= min \| \frac{1}{n_1} \sum_{i=1}^{n_1} \phi(\boldsymbol{X_S}) - \frac{1}{n_2} \sum_{j=1}^{n_2} \phi(\boldsymbol{X_T}) \|_H^2 \\ &= min \| \frac{1}{n_1} \sum_{i=1}^{n_1} \boldsymbol{W}^T k_i - \frac{1}{n_2} \sum_{j=1}^{n_2} \boldsymbol{W}^T k_{j+n_1} \|_H^2 \\ &= tr(\boldsymbol{W^T K L_M K W}) \end{aligned} \tag{3}$$

where $tr(\cdot)$ is the trace operation, k_i is the i-th row vector of matrix $\boldsymbol{K} = \phi(\boldsymbol{X})^T \phi(\boldsymbol{X}) \in R^{(n_1+n_2)\times(n_1+n_2)}$ which denotes the kernel matrix, $\boldsymbol{W}$ is a transformation matrix, and $\boldsymbol{L_M}$ is the MMD matrix whose entry at (i, j) is calculated as:

$$\boldsymbol{L_{M(ij)}} = \begin{cases} \frac{1}{n_1^2} & x_i, x_j \in X_S, \\ \frac{1}{n_2^2} & x_i, x_j \in X_T, \\ -\frac{1}{n_1 n_2} & otherwise. \end{cases} \tag{4}$$

Adding a regularization term $tr(\boldsymbol{W^T W})$ and a constraint $\boldsymbol{W^T KHKW = I}$ ($\boldsymbol{I}$ is an identity matrix) that maintains the variance of mapped data, TCA is formulated as:

$$\begin{aligned} &min\ tr(\boldsymbol{W^T K L_M K W}) + \lambda tr(\boldsymbol{W^T T W}) \\ &s.t.\ \boldsymbol{W^T TKHKW = I} \end{aligned} \tag{5}$$

where $\boldsymbol{H} = \boldsymbol{I}_{(n_1+n_2)} - (1/n_1 + n_2)\mathbf{1}\mathbf{1}^T$ denotes the centering matrix, $\mathbf{1}$ is a $(n_1 + n_2)$-dimension column vector whose all entries are 1, and λ is a trade-off coefficient. Solving (5) with the Lagrange duality theory, the solution of $\mathbf{W}$ in TCA is the p ($p \le n_1 + n_2 - 1$) largest eigenvectors of $(\boldsymbol{K L_M K + \mu I})^{-1} \boldsymbol{KHK}$. With the optimal $\mathbf{W}$, the original source data and target data are transformed to be similar, i.e., the transformed features are matched. Therefore, a well-trained source model based on such source features can achieve good results on transformed target features.

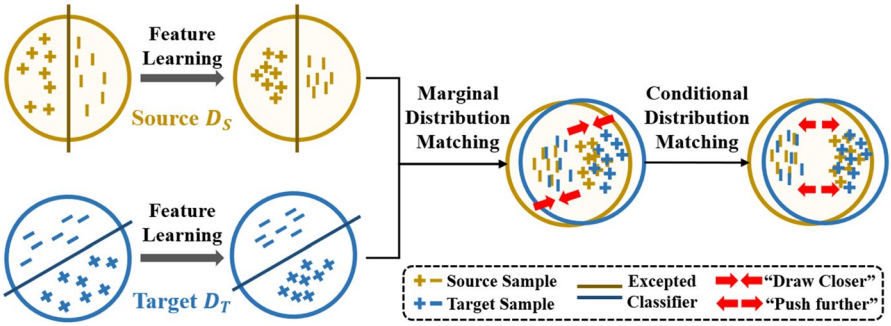


Fig. 6 Motivation of distribution matching

Conditional distribution adaptation aims to reduce the difference between the conditional probability distributions of source and target data, i.e., minimizing the distance between source conditional distribution $Q_S(\boldsymbol{Y_S} \mid \phi(\boldsymbol{X_S}))$ and target one $Q_T(\boldsymbol{Y_T} \mid \phi(\boldsymbol{X_T}))$. They are hard to be directly computed, so they are approximated by $Q_S(\boldsymbol{X_S} \mid \phi(\boldsymbol{Y_S}))$ and $Q_T(\boldsymbol{X_T} \mid \phi(\mathbf{Y_T}))$. Note that, there are few studies that apply conditional adaptation alone (Saito et al. 2017). Most studies utilize both marginal and conditional adaptations with a coefficient μ, so as to achieve robust transfer performance, i.e.,

$$\begin{aligned} minD(D_S, D_T) = min[&D(P(\phi(\mathbf{X_S})), P(\phi(\mathbf{X_T}))) \\ &+ \mu D(Q_S(\mathbf{X_S} \mid \phi(\mathbf{Y_S})), Q_T(\mathbf{X_T} \mid \phi(\mathbf{Y_T})))] \end{aligned} \tag{6}$$

If target labels are unavailable, $Q_T(\mathbf{X_T} \mid \phi(\mathbf{Y_T}))$ cannot be directly calculated. To deal with this, many studies apply pseudo target labels to complete the calculation in (6). In most work (Kang et al. 2020), the pseudo labels of target data are predicted by the source model or unsupervised clustering methods. e.g., K-Nearest Neighbor Classifier. Joint Distribution Adaptation (JDA) (Long et al. 2013) and Transfer Joint Matching (TJM) (Long et al. 2014) are typical transfer methods that consider both marginal and conditional probability distributions. They take MMD as the distance metric, i.e.,

$$\begin{aligned} &D(Q_S(\mathbf{X_S} \mid \boldsymbol{\phi}(\mathbf{Y_S})), Q_T(\mathbf{X_T} \mid \phi(\mathbf{Y_T}))) \\ &= \| \frac{1}{n_1^{(c)}} \sum_{i \in \mathbf{X_S}^{(c)}} \mathbf{W}^T k_i - \frac{1}{n_2^{(c)}} \sum_{j \in \mathbf{X_T}^{(c)}} \mathbf{W}^T k_j \|_H^2 \\ &= \sum_{c=1}^{C} tr(\mathbf{W^T K L_c K W}) \\ &= tr(\mathbf{W^T K \widehat{L_c} K W}) \end{aligned} \tag{7}$$

where $\mathbf{X_S^{(c)}} = \{x_i : x_i \in \mathbf{X_S} \cap y(x_i) = c\}$ is a set of source data with label c, i.e., their true label $y(x_i) = c$, and $n_1^{(c)}$ is the total number of data samples in such set. Similarly, $\mathbf{X_T^{(c)}} = \{x_j : x_j \in \mathbf{X_T} \cap y_p(x_j) = c\}$ represents a set of target data whose pseudo labels $y_p(x_j)$ are c, $n_2^{(c)}$ is the total number of target samples that belong to label c, and $\widehat{\boldsymbol{L}_c} = \sum_{c=1}^{N_c} \mathbf{L_c}$. The entry at (i, j) of conditional MMD matrix $\mathbf{L_c}$ is:

$$\mathbf{L}_{\mathbf{c}(ij)} = \begin{cases} \frac{1}{n_1^{(c)} n_1^{(c)}} & x_i, x_j \in \mathbf{X_S}^{(c)}, \\ \frac{1}{n_2^{(c)} n_2^{(c)}} & x_i, x_j \in \mathbf{X_T}^{(c)}, \\ -\frac{1}{n_1^{(c)} n_2^{(c)}} & \begin{cases} x_i \in \mathbf{X_S}^{(c)}, x_j \in \mathbf{X_T}^{(c)} \\ x_j \in \mathbf{X_S}^{(c)}, x_i \in \mathbf{X_S}^{(c)} \end{cases}, \\ 0 & otherwise. \end{cases} \tag{8}$$

Integrating (3) and (8) into (6), the objective of distribution adaptation can be reformulated as

$$minD(D_S, D_T) = tr(\mathbf{W^T K}(\mathbf{L_M} + \boldsymbol{\mu}\widehat{\mathbf{L_C}})\mathbf{KW}) \tag{9}$$

Following the similar solving steps, JDA can get its optimal transformation matrix that makes $P(\phi(\mathbf{X_S})) \approx P(\phi(\mathbf{X_T}))$ and $Q_S(\mathbf{Y_S} \mid \phi(\mathbf{X_S})) \approx Q_T(\mathbf{Y_T} \mid \phi(\mathbf{X_S}))$. In JDA, the trade-off coefficient μ is set to be 1. Balanced Distribution Adaptation (Liu et al. 2021) and Dynamic Distribution Adaptation (Wang et al. 2020) extend this method by exploring the best trade-off coefficient μ between two distributions and propose an approach to estimate and update it dynamically.

(2) Subspace feature learning

Subspace feature learning methods assume that the source and target data samples are similar in the transformed subspace. They firstly compute a domain-specific subspace for the source data and another one for the target data independently. Then, they project source and target data into intermediate subspaces. Next, subspace features of source and target mapped data are learned such that they are closer to each other. There are two ways to learn subspace features: statistical distribution alignment and manifold one. The former projects both source and target data to a commonly shared subspace by aligning their statistical distributions. The latter builds a (potentially large) set of intermediate representations along the shortest geodesic path that connects source and target subspaces.

Statistical distribution alignment-based methods focus on aligning the statistical features by transformations in a subspace. The aligned data can be learned by traditional machine learning methods. Subspace Alignment (SA) (Fernando et al. 2013) is a typical method to learn features in a statistical alignment way. In SA, a subspace is composed of p eigenvectors obtained by a PCA (Principal Component Analysis). First, the source data $\mathbf{X_S}$ and target data $\mathbf{X_T}$ are represented by their respective subspaces whose basis vectors $\mathbf{B_S}$ and $\mathbf{B_T}$ are derived from PCA. Namely, data in their respective subspaces can be formulated as $\mathbf{X_S^P} = \mathbf{X_S B_S}$ and $\mathbf{X_T^P} = \mathbf{X_T B_T}$. SA learns a mapping function $\mathbf{M}$ that aligns the source subspace basis vectors with the target ones by finding:

$$\mathbf{M}^* = \arg\min_{\mathbf{M}} \|\mathbf{B_S M} - \mathbf{B_T}\|_F^2 \tag{10}$$

where $\|\cdot\|_F^2$ is the Frobenius norm. Since $\mathbf{B_S}$ and $\mathbf{B_T}$ are generated from the first p eigenvectors using PCA, they tend to be intrinsically regularized. Hence, there is no need to add a regularization term in (10). Thus, the closed-form solution of (10) can be obtained as $\mathbf{M}^* = \mathbf{B_S^T B_T}$. The new basis vectors of a source coordinate system are $\mathbf{B_S B_S^T B_T}$, which is called "the target-aligned source coordinate system".

SA is simple to implement with an efficient computational process, and is a representative method for subspace learning. Based on SA, Sun et al. have proposed SDA (Subspace Distribution Alignment) (Sun and Saenko 2015) by adding probability distribution

adaptation into SA. According to Sun and Saenko (2015), SA does not address the problem of distribution alignment. Assume that subspace transformation $\mathbf{M}$ can fully align source and target subspace bases. Yet the subspace distributions of $\mathbf{X_S^P}$ and $\mathbf{X_T^P}$ can be different, thereby leading to degraded performance of source-trained models since such distribution divergence affects the decision boundary. Hence in SDA, apart from a subspace transformation matrix, a probability distribution adaptation transformation is added such that the source and target distributions in the subspaces are aligned as well, thereby promoting TL performance.

Different from SA and SDA, which only perform first-order feature alignment between source and target domains, Sun et al. have proposed CORAL (CORrelation Alignment) (Sun et al. 2017) to perform second-order feature alignment between these two domains. Assume that $\mathbf{C_S}$ and $\mathbf{C_T}$ are the covariance matrices of source and target domains respectively. It learns a second-order feature transformation $\mathbf{A}$ to minimize the cross-domain feature distance, i.e.,

$$\mathbf{A}^* = \arg\min_{\mathbf{A}} \|\mathbf{A}^\mathbf{T}\mathbf{C_S}\mathbf{A} - \mathbf{C_T}\|_F^2 \tag{11}$$

CORAL is very simple and efficient. It is applied to the neural networks and results in DeepCORAL (Sun and Saenko 2016). It calculates a CORAL metric to measure the loss of cross-domain distribution divergence in a neural network.

Many methods embed manifold learning in the process of subspace learning for different domains. Manifold alignment aims to match source and target data from two different manifolds so that they are able to be mapped into a shared common space. As mentioned before, manifold alignment projects data into a (potentially large) set of intermediate subspaces along the shortest geodesic path that links the source and target subspaces on the Grassmann manifold. The approach of Sampling Geodesic Flow (SGF) (Gopalan et al. 2011) is classic, and then Geodesic Flow Kernel (GFK) (Gong et al. 2012) is SGF's extension and achieves enhanced performance. Hence, the latter is the representative work in the field of manifold alignment.

SGF is motivated by incremental learning and regards the generative subspaces of source and target domains as two points on the Grassmann manifold. Then, there is a geodesic path that connects source and target subspaces. SGF samples several points along this geodesic path to create intermediate subspaces, and then transform a source to a target in order of these sampled subspaces. Intuitively, these subspaces offer a meaningful description to model domain shift. Hence, a transfer is achieved by these intermediate representations between source and target domains. Concretely, SGF has the following steps: i) build a geodesic flow path bridging source and target domains on the Grassmannian manifold; ii) sample a certain number of subspaces along this geodesic path; iii) map original source and target feature vectors onto these subspaces and integrate them to form new feature vectors; and iv) construct discriminative models based on new features of source and target data.

The work (Gong et al. 2012) points out several limitations of SGF, e.g., it is not clear how many subspaces should be sampled to ensure a successful transfer in SGF. To address this problem, it proposes GFK to model domain shift by integrating an infinite number of subspaces, which describes incremental changes between the source and target domains in a more elaborated way. GFK has four key steps: i) determine the best dimensionality of the subspaces to embed domains; ii) construct the geodesic flow that parameterizes the smooth changes from the source domain to the target one; iii) calculate their geodesic flow kernel; and iv) use the kernel to build a discriminative model with labeled data.

Many methods combine subspace feature learning and distribution adaptation so as to improve their TL performance. For example, Domain Invariant Projection (DIP) (Baktashmotlagh et al. 2013) integrates the marginal distribution alignment and manifold subspace learning, and Manifold Embedded Distribution Alignment (MEDA) (Wang et al. 2018) performs dynamic distribution alignment in the Grassmann manifold and integrate a graph Laplacian regularization term in its objective.

As for the machinery applications based on feature matching transfer, Mao et al. (2019) utilize TCA (Pan et al. 2010) to transfer features in the task of rolling bearings RUL prediction. Qian et al. (2021) use MEDA (Wang et al. 2018) for rotating machine fault diagnosis under variable working conditions. Ma et al. (2020) present a diagnosis framework by extending TCA to weighted TCA (WTCA). To improve the separability of class labels, WTCA adds the objective function of linear discriminate analysis into the original TCA objective, which minimizes the within-class distance as well as maximizes the distance between different classes. It also considers the conditional distribution alignment in its objective, which is similar to TJM (Long et al. 2014). WTCA is validated by five transfer diagnosis tasks with varying experimental positions, fault severity levels, fault types, working conditions and experiment setups. To tackle the problem of class imbalance, Zhang et al. (2020) enhance TJM by integrating a distance metric: maximum variance discrepancy (MVD), and their method is verified by two case studies of bearing fault diagnosis. Note that, MMD matching reduces the domain distribution divergence concerning the first-order statistics, while MVD matching achieves such goal from the perspective of the second-order statistics. Zhang et al. (2020) present a GFK-based domain adaptation method for fault diagnosis and validate it with real-world gears and bearings datasets.

In addition to the aforementioned transfer methods, many other approaches integrate time-frequency analysis and matrix factorization into the transfer framework for machinery diagnosis or prognostics. Wang et al. (2021) propose a joint dictionary matrix factorization method to handle transferable regression tasks, i.e., RUL predictions of bearings under varying operating conditions. They use joint-domain projection dictionaries to construct a shared latent space for source and target data. Wang et al. (2016) present a transfer factor analysis (TFA) method for gearbox diagnosis under various operating conditions. TFA is built based on factor analysis that transforms data into a low-dimensional latent space where the key properties of original data are preserved. It also takes domain difference into consideration. Thus, TFA can capture the pivot features of the original source and target domain data, and can be used to reduce their domain divergence in the learned latent space. In TFA, the knowledge transfer between source and target domains is achieved by a shared factor loading matrix. TFA utilizes two different noise terms to represent domain differences, which are described by two different diagonal covariance matrixes. During the process of dimension reduction, the learned features can minimize domain differences as well as preserving data properties. Such features acquired by TFA can be fed into a machine learning model for classification. In Wang et al. (2016), experiments based on support vector machines are performed to validate that TFA can be effectively applied to gearbox diagnosis under different operating conditions. Shen et al. (2015) present a bearing fault diagnosis strategy by transferring knowledge from selective auxiliary data (also source data) to their target model. Single Value Decomposition (SVD) (Sun et al. 2021) is used for feature extraction, and the TrAdaboost transfer learning algorithm is applied to improve the classification accuracy. In Shen et al. (2015), in order to avoid negative transfer, each source sample is evaluated for transferability and possibility according to how similar it is to target samples. A similarity criterion is determined by the vector angle cosine of SVD features between target data and auxiliary source data. Experiments using datasets from a

bearing test system verify its effectiveness. Chen et al. (2021) also use the TrAdaboost transfer learning algorithm in their framework for wind turbine fault diagnosis.

3.2.2 Deep methods

With the wide popularity of deep learning methods, more and more researchers are using deep neural networks for TL. Compared to traditional shallow TL methods (TCA, GFK, etc.), deep ones greatly improve learning performance on different tasks with more accurate results. Moreover, because they learn directly from raw data, they have two additional advantages over shallow ones: automating the extraction of more discriminative features, and meeting the end-to-end needs of real-world applications.

In the previous section, we have answered why deep networks are transferrable and introduced the simplest form of deep network transfer: model and parameter transfer followed by finetuning. In this section, feature matching techniques used in deep networks are explored. Note that, deep adversarial networks, as a promising and commonly-used method to perform the transfer, are introduced separately in the next section. We describe the basic ideas and core methods for matching features in general deep networks. It is worth noting that due to the vast development of research efforts in deep TL, it is impossible to cover all the latest methods. But their basic principles are similar. Therefore, we introduce some basic but representative methods.

Model and parameter transfer with fine-tune serves as the basis in deep transfer, which can save training time and also improve accuracy. But when source and target data samples are of different distributions, feature matching methods, such as distribution adaptation and subspace feature learning depicted in shallow methods, should be incorporated into deep transfer networks. Thereby, many deep learning methods develop adaptation layers to achieve cross-domain adaptation. Such adaptation layers can draw the data distributions of source and target domain closer to each other, so as to improve the transferability of networks. From the above analysis, it is concluded that self-adaptive deep networks mainly concentrate on two parts. The first part answers which layer can be self-adaptive. It determines the transferability of the networks. The second one answers what kind of adaptation methods should be used. This determines the generalization ability of transfer networks.

The most important issue in deep networks is to define network loss. Most methods adopt the following loss definition:

$$l = l_S(D_S) + \alpha l_A(D_S, D_T) \tag{12}$$

where l denotes the total network loss, $l_S(D_S)$ is the source model loss on the labeled source data D_S (which is the same as that in general deep networks), $l_A(D_S, D_T)$ represents the adaptation loss between D_S and D_T, and coefficient α is used to trade off two network losses. The second part, i.e., adaptation loss, is unique to TL and not present in traditional deep networks. Its expression is the same as the distribution difference between source and target domains as discussed about shallow methods.

Domain adaptive neural network (DaNN) (Ghifary et al. 2014) is an early work that integrates the distribution alignment into network training. But its structure is quite simple. The base networks in DaNN consist of only two layers: a feature extracting layer and a classification one. DaNN proposes to add an MMD adaptation layer after the feature extracting layer, which is used to calculate the distance between source and target domains. Such distance is used as the adaptation loss in (12). However, due to the too-simple network architecture, DaNN's capability to learn good representations is limited. Thus, it cannot

effectively solve the problem of domain adaptation. Therefore, most subsequent researchers expand it by using deeper architectures, such as AlexNet (Krizhevsky et al. 2012), VGG (Simonyan and Zisserman 2014) and ResNet (He et al. 2016), or replacing MMD with multi-kernel MMD (MK-MMD).

Deep Domain Confusion (DDC) (Tzeng et al. 2014) follows the above ideas and employs pre-trained AlexNet to achieve deep transfer. As shown in Fig. 7, DDC fixes the first seven layers of AlexNet and takes features of the eighth layer (the layer before the classifier) to measure cross-domain distance.

Its distance metric adopts the widely used MMD criterion. DDC can be expressed as:

$$\min_{\Theta} l = l_S(\theta(\mathbf{X_S}), \mathbf{Y_S}) + \alpha l_{MMD}(D_S, D_T) \tag{13}$$

where Θ denotes all the network parameters (weights and bias), $l_S(\theta(\mathbf{X_S}), \mathbf{Y_S})$ is the loss between the predicted labels $\theta(\mathbf{X_S})$ and true labels $\mathbf{Y_S}$, $l_{MMD}(D_S, D_T)$ computes the cross-domain distance based on MMD. Noting that DDC only aligns marginal distributions, its extension (Tzeng et al. 2015) jointly aligns conditional and marginal distributions in CNN architecture for domain and task transfer.

Deep Adaptation Networks (DAN) (Long et al. 2015) extend DDC in two aspects. First, different from DDC that adds only one adaptive layer, DAN adds three adaptation layers (three fully connected layers before the classifier layer) and freezes the first five layers in AlexNet as shown in Fig. 8. Secondly, MK-MMD, which is considered to have better representation ability than MMD, is adopted in DAN to replace plain MMD in DDC. The parameter learning of MK-MMD is integrated into the training of the deep network, which does not increase the extra network training time. DAN can achieve better classification performance than DDC on multiple tasks. It is written as

$$\min_{\Theta} l = l_S(\theta(\mathbf{X_S}), \mathbf{Y_S}) + \alpha \sum_{l=l_1}^{l_2} l_{MMD}(D_S^l, D_T^l) \tag{14}$$

where l_1 and l_2 denote the start and end layers of adaption, and l_1= 6 and l_2= 8 in DAN. Joint Adaptation Network (JAN) (Long et al. 2017) extends DAN by adding conditional distribution alignment and proposes joint MMD.

Using the structure of DDC (Tzeng et al. 2014), Azamfar et al. (2020) present a deep domain adaptation methodology for ball screw fault diagnosis. Deebak and Al-Turjmann (2021) also take a DDC architecture for fault diagnosis, but their base network is a stacked sparse auto-encoder. Si et al. (2021) add CORAL (Sun et al. 2017) matching into DDC's framework and use ResNet as a feature extractor in their fault diagnosis task. Replacing marginal MMD in DDC with conditional MMD and extracting multi-scale features by multiple different convolution kernels, Wang et al. (2020) propose a multi-scale deep transfer method. Han et al. (2020) present a deep transfer network (DTN) for bearing fault diagnosis. As shown in Fig. 9, DTN embeds the idea of joint distribution adaptation (JDA) (Long et al. 2013) to ensure accurate distribution matching. It is built on the basis of DDC and extends it by replacing marginal MMD matching with joint MMD matching, i.e., both marginal and conditional probability distributions are employed to reduce the domain divergence across domains. Three fault datasets, i.e., wind turbine, bearing and gearbox datasets, covering different operating conditions, fault severities and fault types, are used for validation. Based on such datasets, ten transfer tasks are used to evaluate the practicability and applicability of DTN. Similarly, Wu et al. (2020) integrate JDA in their deep transfer network, but they take the long-short term memory (LSTM) as the feature

extractor network. Shen et al. (2021) and Liao et al. (2021) also utilize JDA for distribution alignment for bearing fault diagnosis, but they dynamically adjust the weight of conditional MMD matching. Yu et al. (2019) also integrate LSTM in their transfer framework, but they conduct RUL prediction (instead of extracting features) by LSTM due to its excellent ability to handle time-series data.

Li et al. (2018) extend DDC in a different way. They present a deep distance metric learning for fault diagnosis. To improve the generalization ability of a transfer model, apart from domain adaptation by matching marginal MMD as in DDC, they propose a representation clustering method that minimizes the within-class distance of features and maximize the between-class distance of features. Hence, data from the same class can be mapped closer, while data in different classes are well-separated. The deep distance metric learning method in Li et al. (2018) has three steps: (1) the raw machinery vibration signal data are input to a deep learning network which functions as a feature extractor; (2) distance metric learning is deployed to handle the extracted high-level features. It consists of representation clustering and domain adaptation, and both of them utilize the top fully-connected layer as data representations; and (3) a classifier is used for final fault classification. Kim et al. (2021) present a similar semantic clustering-based method for fault diagnosis of rotating machinery, but their

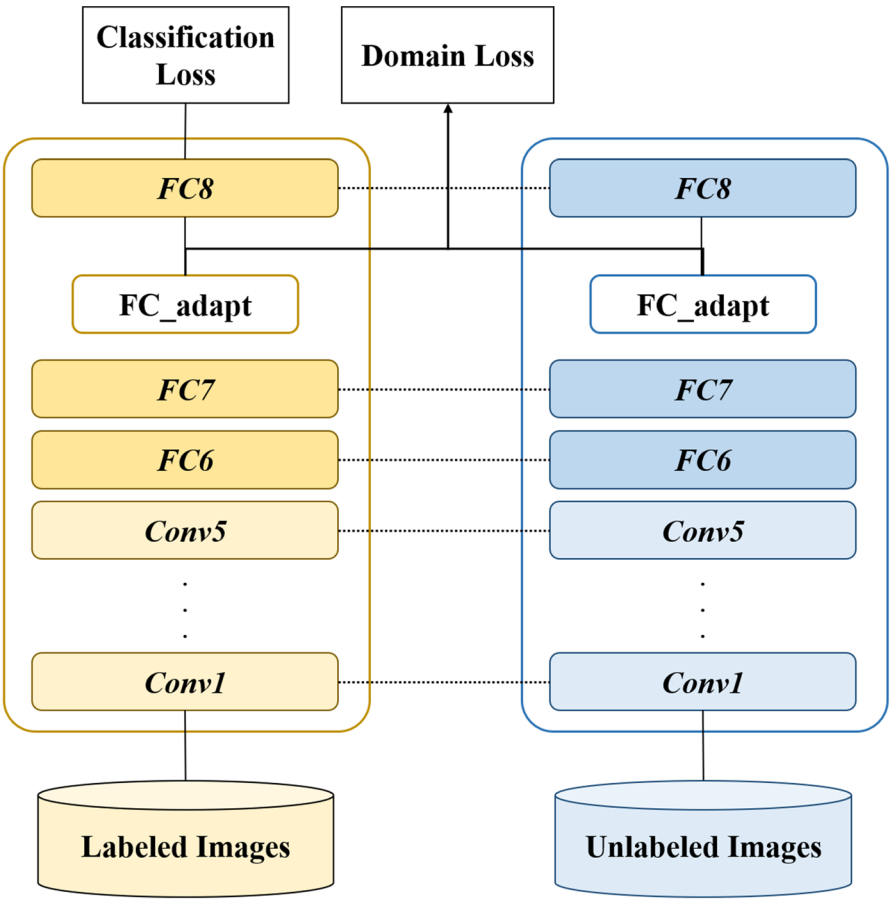


Fig. 7 General framework of DDC (Tzeng et al. 2014)

additional semantic clustering loss is computed at multiple feature levels, i.e., outputs from every pooling layer in a source feature extractor are used for calculating such clustering loss.

Based on the architecture of DAN (Long et al. 2015), Jin et al. (2021) achieve domain adaptation bearing fault diagnosis under different working conditions. Zhu et al. (2019) build their transfer network on DDC and calculate domain loss by a linear combination of multiple Gaussian kernels. Ding et al. (2021) match MK-MMD between source and target features extracted by auto-encoders, and realize the cross-domain transfer for RUL prediction, while Lu et al. (2021) match marginal and conditional MMD for deep feature adaptation in fault diagnosis. By using DAN, Li et al. (2021) utilize DAN architecture and present an ensemble of twelve transfer networks with 12 different kernel MMD (such as linear kernel, Gaussian kernel, polynomial kernel, and exponential kernel) such that diverse transferable feature representations can be learned.

A feature-based transfer neural network (FTNN) (Yang et al. 2019) is built on DAN but matches the marginal MMD in multiple layers (including two convolutional layers and two full-connected layers) as shown in Fig. 10. It utilizes two separated networks for source and target. Their network architectures are shared such that the multi-layer domain adaptation can be employed to reduce the cross-domain distribution divergence. There are three loss functions to be jointly minimized in FTNN: (1) source network loss between predicted labels for source samples and their true labels, (2) target network loss between predicted labels for target samples and their pseudo labels, and (3) multi-layer MMD between extracted features in source and target networks. FTNN is designed to transfer the diagnosis knowledge of laboratory bearings simulating both normal and faulty conditions to real-case machines. Two fault diagnosis cases, i.e., transfer from laboratory motor bearings to real-world locomotive bearings and from laboratory gearbox bearings to real-world locomotive bearings, are used to verify the effectiveness of the FTNN model.

Especially, to deal with RUL prediction, Zhang et al. (2021) present a data alignment scheme to ensure that source and target data follow similar degradation trace in the learned subspace, thus facilitating prognostics knowledge transfer. The proposed data alignment scheme includes healthy state alignment, degradation direction alignment, degradation level regularization and degradation fusion. Features extracted by high-level layers are

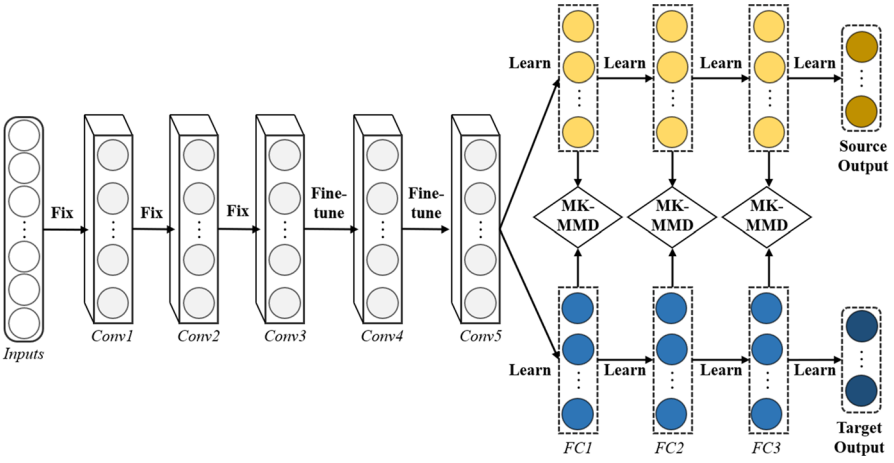


Fig. 8 General framework of DAN (Long et al. 2015)

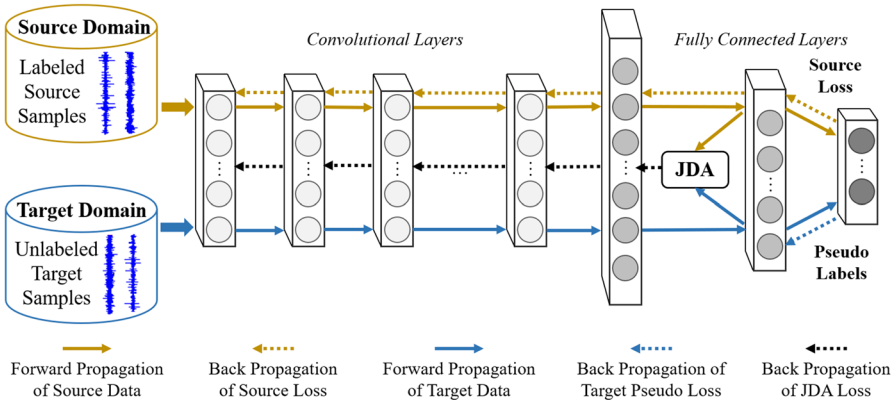


Fig. 9 General framework of DTN (Han et al. 2020)

input to the proposed data alignment module. Other model architectures are similar to DAN.

Apart from popular CNN, an auto-encoder and its variants are widely used in deep transfer methods for machinery diagnosis. Studies (Li et al. 2020; Lu et al. 2016) employ an auto-encoder as a feature extractor. In fault diagnosis tasks, they achieve domain adaptation by applying MMD to minimize the domain discrepancy when transferring a source model. The difference is that one auto-encoder is taken as a basic single-layer representation model in Lu et al. (2016), while a sparse auto-encoder is used in Wen et al. (2017). They both stack multiple layers to develop a deep network. Similarly, by stacking multiple layers of a denoising autoencoder (DAE) to extract features in various levels, the work (Wang et al. 2019) presents a hierarchical deep domain adaptation (HDDA) approach for fault diagnosis of a thermal system under varying operating conditions. A fault classifier trained by labeled source data collected under one working load is transferred to classify unlabeled target data acquired from another different load. In HDDA, both marginal and conditional CORAL (Sun et al. 2017) distances are used to minimize the distribution differences.

To conclude the feature matching-based transfer methods and their machinery applications, researchers achieve the transfer by reducing the feature distribution difference across domains via feature transformations. Both shallow and deep methods can be used for feature extraction and domain adaptation with the goal of drawing source and target features closer. Techniques established in shallow methods, such as distribution adaptation and subspace learning, are integrated into deep transfer networks. Since deep transfer methods usually gain higher performance than shallow ones, most existing research tends to implement deep ones in their applications of machinery fault diagnosis or RUL prediction.

3.3 Adversarial adaptation-based methods

In recent years, Generative Adversarial Networks (GAN) (Liu et al. 2021; Han et al. 2020; Yang N et al., 2021), as the representative of adversarial learning, have attracted the attention of many researchers. Various kinds of GAN-based variants have emerged. They have greatly improved the learning performance compared to traditional deep neural networks. Therefore, GAN-based TL is also a popular research topic.

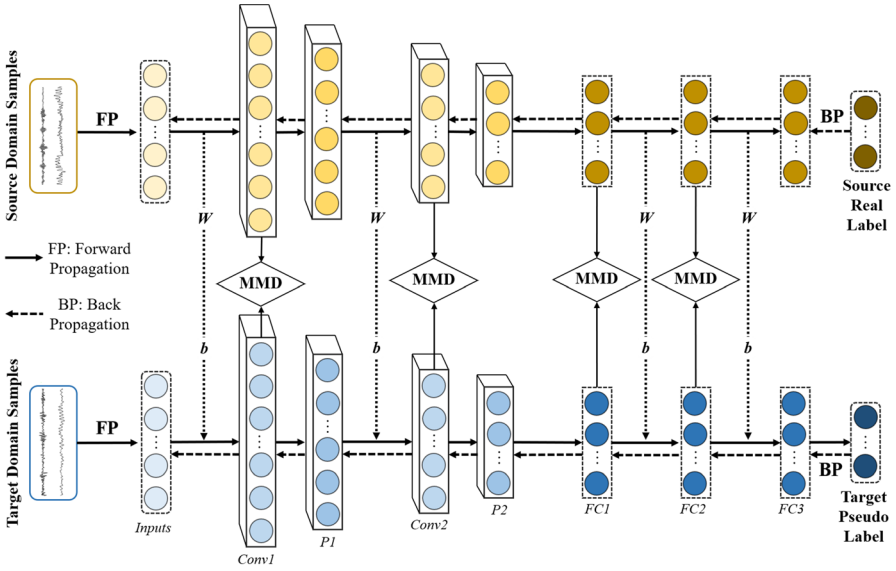


Fig. 10 General framework of FTNN (Yang et al. 2019)

GAN is inspired by the idea of a two-player zero-sum game in game theory. It consists of two parts: a generative network and a discriminative one. The former, known as a generator, aims to generate fake samples that are as real as possible. The latter, called a discriminator, determines whether the samples are real or generated by the generator. The generator tries to confuse the discriminator, but the discriminator tries to not be fooled by the generator. As a consequence, their training objectives are adversarial. GAN is completed by playing the generators and discriminators against each other.

The original GAN is mostly used for generating samples, which seems to have no relation with TL. But its adversarial structure is exactly suitable for transfer. In GAN, the generator tries to produce samples that are similar to the real samples, while in TL the source and target features are also supposed to be similar. In adversarial adaption-based transfer, there is no need to generate samples, and thus we take features of a source or target domain as the output of a generator, i.e., "generated samples". Then the discriminator is used to tell the difference between source and target features. Therefore, the generator in adversarial adaption tasks does not function as a sample generator, but it is used for extracting features. Its goal is to learn the features of one domain such that the discriminator cannot distinguish them from features of the other domain. In this way, the original generator can also be called a feature extractor. Similar to deep transfer networks, the loss of an adversarial adaptation network consists of two components: the loss of network training loss l_S and domain discriminative loss l_d, i.e.,

$$l = l_S(D_S) + \epsilon l_d(D_S, D_T) \tag{15}$$

A Domain-Adversarial Neural Network (DANN) (Ganin et al. 2016) is the first one to applying an adversarial mechanism into the transfer network. It promotes features in two ways. First, the learned features are discriminative for the primary learning task on a source domain. Secondly, they are indiscriminate in terms of the domain shift, i.e., the features are domain-invariant and transferable. As shown in Fig. 11, DANN consists of three parts: a deep feature extractor $G_f(\cdot;\theta_f)$ with parameter θ_f, a deep label predictor $G_y(\cdot;\theta_y)$

with θ_y and a domain classifier $G_d(\cdot;\theta_d)$ with θ_d. The feature extractor G_f and label predictor G_y form a standard feed-forward path. Along this path, the labeled source data can be input and used for training. A standard network training for G_f and G_y is achieved by minimizing the label prediction loss and back-propagating corresponding gradients. As for domain adaptation, DANN uses a gradient reversal layer (GRL) to connect the domain classifier and feature extractor. As its name implies, GRL multiplies back-propagated gradients by a negative constant. As a consequence, the domain classifier has an adversarial impact on G_f compared to that in the training path of G_f and G_y. The domain classifier G_d is designed to distinguish source and target features by minimizing its loss. With GRL, the feature extractor network maximizes such loss such that G_d cannot tell the difference between source and target domains.

To conclude, (1) when training G_f and G_y, both of them are optimized to minimize the prediction error on source data, which can learn discriminative features that can be accurately classified; (2) when training G_f and G_d, G_d is to minimize the discrimination loss but G_f is to maximize such loss, which ensures the features of two domains are similar such that domain-invariant features are learned. In DANN, for sample x_i, the network training loss L_s^i and domain discriminative loss L_d^i are formulated as:

$$L_s^i(\theta_f, \theta_y) = L_s(G_y(G_f(x_i;\theta_f);\theta_y), y_i) \tag{16}$$

$$\begin{aligned} L_d^i(\theta_f, \theta_d) &= L_d(G_d(G_f(x_i;\theta_f);\theta_d), d_i) \\ &= d_i \log \frac{1}{G_d(G_f(x_i;\theta_f);\theta_d)} + (1-d_i)\log \frac{1}{1-G_d(G_f(x_i;\theta_f);\theta_d)} \end{aligned} \tag{17}$$

where d_i is the binary domain indicator for the i-th sample. It is 0 if x_i comes from a source domain and 1 if it is from a target domain.

The overall objective in DANN can be written as:

$$L(\theta_f, \theta_y, \theta_d) = \frac{1}{n_1}\sum_{i=1}^{n_1} L_s^i(\theta_f, \theta_y) - \epsilon\left(\frac{1}{n_1}\sum_{i=1}^{n_1} L_d^i(\theta_f, \theta_d) + \frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2} L_d^i(\theta_f, \theta_d)\right) \tag{18}$$

where ϵ is a coefficient to balance L_s and L_d. As analyzed previously, the optimal parameters satisfy

$$(\theta_f^*, \theta_y^*) = \arg\min_{\theta_f, \theta_y} L(\theta_f, \theta_y, \theta_d) \tag{19}$$

$$\theta_d^* = \arg\max_{\theta_d} L(\theta_f^*, \theta_y^*, \theta_d) \tag{20}$$

Following the gradient descent rule, parameters in (19) and (20) can be updated as:

$$\theta_f \leftarrow \theta_f - \eta\left(\frac{\partial L_s^i}{\partial \theta_f} - \epsilon\frac{\partial L_d^i}{\partial \theta_f}\right) \tag{21}$$

$$\theta_y \leftarrow \theta_y - \eta\frac{\partial L_s^i}{\partial \theta_y} \tag{22}$$

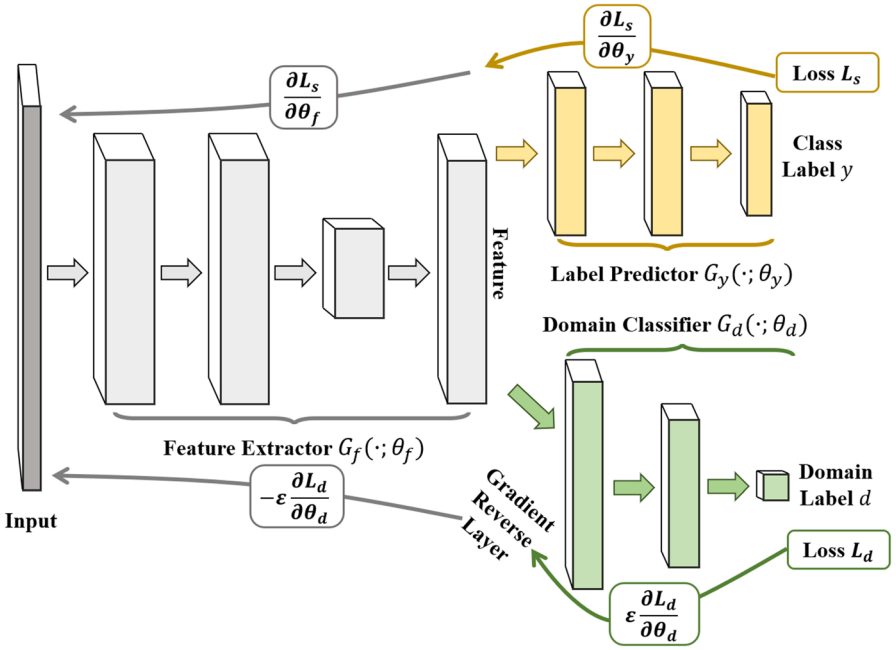


Fig. 11 General framework of DANN (Ganin et al. 2016)

$$\theta_d \leftarrow \theta_d - \eta\epsilon\frac{\partial L_d^i}{\partial \theta_d} \tag{23}$$

where η is the learning rate.

Different from DANN where source and target data share the same feature extractor, Tzeng et al. propose Adversarial Discriminative Domain Adaptation (ADDA) (Tzeng et al. 2017) that uses separated feature extractors networks M_S and M_T for source and target data, respectively. ADDA is a representative method that directly uses the architecture of GAN. According to Tzeng et al. (2017), the optimization in DAAN corresponds to the minimax objective in GAN, but such optimization setting tends to face the problem of vanished gradients because its discriminator converges quickly at the early stage of training. Therefore, instead of utilizing such objective function, the work (Tzeng et al. 2017) trains its generator via a standard loss function with inverted labels.

As shown in Fig. 12, first, as a source CNN, M_S is pre-trained based on samples of labeled source images, and a source classifier C is well-trained. During the pre-training, the loss of source classifier $l_S(X_S, Y_S)$ is optimized over M_S and C based on the labeled source data (X_S, Y_S), i.e.,

$$\min_{M_S,C} l_S(\mathbf{X_S}, \mathbf{Y_S}) = \min_{M_S,C} -E_{(x_s,y_s)\sim(\mathbf{X_S},\mathbf{Y_S})} \sum_{k=1}^{K} \mathbf{1}_{[k==y_s]} \log C(M_S(x_s)) \tag{24}$$

Next, as a target CNN, M_T is trained to confuse discriminator D such that it cannot reliably predict domain labels of source and target features, i.e., target data are mapped

to a shared feature space where they are similar to source features. Note that, the network parameters of M_S are fixed during adversarial training, which mimics the original GAN setting where the real distribution remains fixed, and the generated distribution is updated to match it. As mentioned previously, the objective of D is to correctly identify the binary domain labels of source features $M_S(x_s)$ and targets features $M_T(x_t)$. Thus the domain discrimination loss $l_D(X_S, X_T, M_S, M_T)$ is expressed as:

$$\begin{aligned} &\min_D l_D(\mathbf{X_S}, \mathbf{X_T}, M_S, M_T) \\ &= \min_D -E_{x_s \sim \mathbf{X_S}}[\log D(M_S(x_s))] - E_{x_t \sim \mathbf{X_T}}[\log(1 - D(M_T(x_s)))] \end{aligned} \tag{25}$$

The generator, i.e., target CNN M_T, is trained by the standard loss function with inverted labels, i.e.,

$$\min_{M_T} l_T(\mathbf{X_S}, \mathbf{X_T}, D) = \min_{M_T} -E_{x_T \sim \mathbf{X_T}}[\log D(M_T(x_T))] \tag{26}$$

Note that, (26) is in the same form as the first term in (25) except that it uses target data while (25) applies source data. This indicates that target features are trained to close to source features. Hence, the adversarial adaptation is successfully performed. Last, the target image is input into the learned target CNN M_T and classified by the source classifier C.

DANN and ADDA represent two general frameworks, and many existing approaches can be considered as their extensions. Based on them, Generative Adversarial Distribution Matching (Kang et al. 2020) and Dynamic Adversarial Adaptation Networks (Yu et al. 2019) integrate well-established techniques in distribution alignment and/or subspace learning. Moreover, some studies are motivated by the advanced development in GAN, then borrow their ideas to refine the deep adversarial transfer network by adding more components (Liu and Tuzel 2016), e.g., feature extractor, label predictor and domain classifier and designing more effective architectures, e.g., such as residual connections (Cai et al. 2019).

Han et al. (2019) propose a deep adversarial CNN for intelligent diagnosis of mechanical faults. Its adversarial learning framework shares the same architecture as DANN, and mechanical signals of source and target datasets are regarded as inputs. Such application is verified on a wind turbine fault dataset and a gearbox fault one. Da Costa et al. (2020) predict RUL by utilizing the architecture of DANN with LSTM as a feature extractor. Michau and Fink (2021) adopt DANN and introduce a novel multi-dimensional scaling loss for unsupervised anomaly detection tasks. Han et al. (2021) extend DANN to the scenario of sparse target data by adding multiple domain adversarial discriminators, such that it can realize convincing feature distribution alignment. Guo et al. (2018) also consider the distribution alignment. But different from Han et al. (2021), they integrate the techniques in the feature matching method and present a deep convolutional transfer learning network (DCTLN) for machine fault diagnosis. As shown in Fig. 13. It has two modules: condition recognition (blue part) and domain adaptation (orange part). The former is constructed by a feature extractor that utilizes CNN to learn features from the input of one-dimensional sensor data, and a health condition classifier that recognizes different health states of machines. The latter learns domain-invariant features along two directions: a domain classifier to maximize domain recognition errors, and a distribution matching term that takes MMD as distribution discrepancy metric and minimizes the probability distribution distance between extracted source and target features. Note that DCTLN constructs its framework based on DANN, and extends it by adding distribution matching into DANN. In DCTLN, three bearing

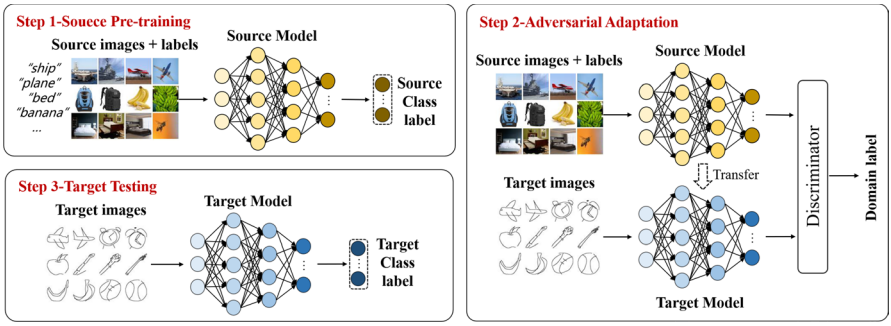


Fig. 12 General framework of ADDA (Tzeng et al. 2017)

datasets collected from different machines (motor bearing, shaft support bearing and railway locomotive bearing) are used to construct six transfer fault diagnosis experiments. Since these machines are operated under different speeds and loads, their data distributions are with obvious divergence. Experimental results with regards to the classification of bearing health conditions indicate that accuracies of DCTLN are about 32.1% higher than traditional methods without TL.

A structure similar to DCTLN is utilized in Zhu et al. (2020) for RUL prediction and in Huang et al. (2021) and Jiao et al. (2020) for fault diagnosis. There are some differences: Huang et al. (2021) takes the dense convolutional neural network (Dense-net, Huang et al. 2017) for extracting features and replaces MMD metric with CORAL (Sun et al. 2017) metric, and Jiao et al. (2020) takes residual network as a feature extractor and replaces MMD metric with joint MMD. Besides, Miao and Yu (2021) present a similar adversarial framework for RUL prediction, but they take the selective convolutional recurrent neural networks as a feature extractor to learn both temporal and spatial features from raw vibration signal. Motivated by DAN (Long et al. 2015), they utilize MK-MMD to measure the cross-domain distribution discrepancy. Li et al. (2021) extend DCTLN by adding data structure alignment. In their domain adversarial framework for fault diagnosis, apart from fault classifier and domain discriminator, they construct graph generation layers to model the relationship of structure characteristics, and then MMD matching is used to minimize the structure difference between such instance graphs from different domains. Xia et al. (2021) use the same framework as DCTLN, but integrate fault information and ensemble multiple LSTMs to capture different degradation patterns resulting from varying faults, thus enhancing RUL prediction performance.

Weighted Adversarial Transfer Network (WATN) (Li et al. 2020) pays attention to a more real-world transfer problem in machinery fault diagnosis. Most existing methods assume that source and target domains share identical label space. But in practice, there is a situation where small target data only cover a subset of classes in relatively larger source data. To address this partial transfer issue, WATN introduces a weighting learning strategy by adopting an additional pair of G/D (i.e., Generator/Discriminator) modules, i.e., the pink parts in Fig. 14. Since source data include more categories that are not shown in target data, WTAN filters out irrelevant source data such that they do not impact transfer performance. A weighting strategy is designed to assess the transferability of samples in source domain, which assigns them different weights according to their contributions to additional G/D modules. Concretely, WTAN utilizes an additional pair of a source classifier

and domain discriminator to quantify sample weights. If the additional discriminator gives a large probability to a source sample, it means that this sample can be easily recognized as source data, which indicates that such sample is far away from the target domain and may belong to the outlier classes. On the contrary, a small probability value means that the source sample is closer to the target domain and is more likely drawn from the shared label space. Thus, for a source sample, if its output probability by additional G/D modules is small, a larger weight is assigned, and vice versa. Except that WATN adds auxiliary networks to identify and filter out some irrelevant samples in the source domain, its basic network architecture is similar to DANN, i.e., it applies the minimax adversarial training to learn both class-discriminative and domain-invariant features. Two sets of experiments based on a rolling element bearing dataset and a gearbox dataset are carried out, and different diagnosis tasks under partial transfer setting are presented to verify the effectiveness of WATN. Experimental results show that it can achieve promising performance by matching source and target distributions in their shared label space. Similarly, Li et al. (2020) present class-weighted adversarial networks for partial TL in machinery cross-domain fault diagnostics. Its network architecture is the same as DANN, and the probability outputs of its domain discriminator are used to determine of the corresponding weights.

Based on ADDA, Li et al. (2021) present knowledge mapping-based adversarial domain adaptation for fault diagnosis. Zhao et al. (2021) propose a joint distribution adaptation network in ADDA's adversarial learning framework, and an improved joint MMD (which directly calculates the joint probability by using Bayesian theorem rather than approximation) is used to match cross-domain features for bearing fault diagnosis. Ragab et al. (2020) indicate that in transfer tasks of RUL prediction, conventional domain adaptation approaches focus on learning domain invariant features, but fail to consider target specific information, which leads to uncomplete feature learning and limited transfer performance. To address this problem, they extend ADDA by presenting a Contrastive Adversarial Domain Adaptation (CADA) method for cross-domain RUL prediction under various working conditions. Their framework extends ADDA with a contrastive loss. As presented in Fig. 15, CADA adds an InfoNCE module in ADDA, such that it can preserve

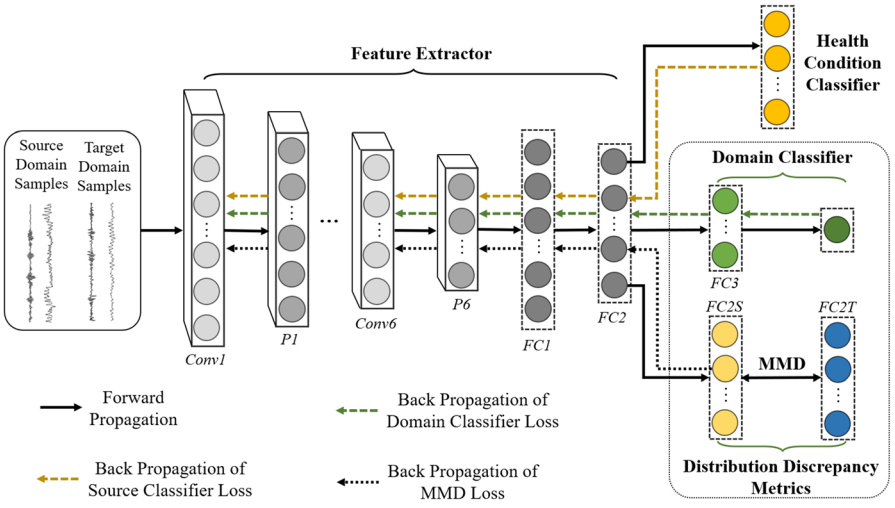


Fig. 13 General framework of DCTLN (Guo et al. 2018)

target-specific information when learning domain-invariant features. The contrastive loss in CADA takes InfoNCE (Henaff 2020) as a metric to measure the mutual information between the extracted features of the target domain and original data. By maximizing InfoNCE, task-specific information is preserved. Aero-engines datasets are used to construct 12 cross-domain scenarios, and experiments of predicting their RUL are conducted to validate the performance of CADA. Experimental results show that when compared to other state-of-the-art methods, CADA achieves over 21% and 38% improvements in terms of root mean square error (RMSE) in a regression task, and score metric that imposes penalties for late RUL predictions (A late RUL prediction refers to a situation where a machine has already failed before the predicted time, which usually causes large economic loss).

Besides, Ragab et al. (2020) estimate RUL based on ADDA and utilize multi-layer bidirectional LSTM as a feature encoder. Wu et al. (2020) also take LSTM as a feature extractor, and conduct the adversarial strategy by applying maximum classifier discrepancy (MCD) (Saito et al. 2018), so as to conduct rolling bearing fault diagnosis under few labeled data. Similarly, Li et al. (2020) take advantage of MCD to present an adversarial multi-classifier optimization method for machinery fault diagnostics.

In closing, adversarial adaptation-based transfer is achieved by learning domain-invariant features during the adversarial training of a generator and discriminator. In adversarial adaptation transfer, the generator is not used for generating samples but functions as a feature extractor; and the discriminator aims to distinguish the difference between extracted source and target features. Due to their satisfactory transfer performance, deep adversarial adaptation methods are widely used for machinery applications (especially for fault diagnosis). Many studies extend the general adversarial framework by integrating feature matching techniques such as distribution alignment and advanced development in GAN.

3.4 Discussion

In Tables 4 and 5, we summarize the model performance of existing methods. Table 4 presents the classification accuracy results of different transfer fault diagnosis methods on

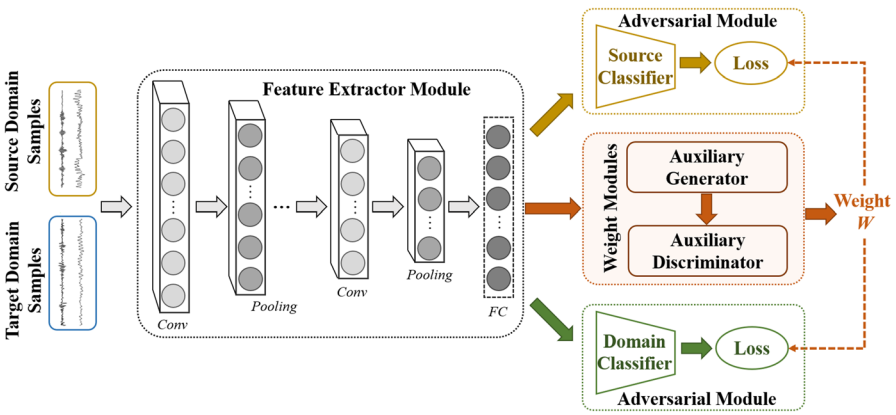


Fig. 14 General framework of WATN (Li et al. 2020)

CWRU Dataset. Table 5 shows RMSE results of different transfer fault diagnosis methods on C-MAPSS Dataset.

From the review of the research about cross-domain transfer for machinery diagnostics and prognostics, it is clear that most of researchers prefer to deploy a deep transfer method for their applications instead of using a shallow one. The wider range of applications of deep TL methods is probably because they have an unbeatable advantage over traditional shallow ones in terms of accuracy. It should be noted that the transfer technique developed in shallow methods, such as distribution alignment and subspace feature learning with manifold, are integrated into the deep transfer models to improve their performance. Moreover, the training of deep networks requires sufficient data, expensive GPU resources and a long time, which may be unsuitable for machinery tasks with a small dataset and limited computation resources. For example, although the idea of edge computing (Silva et al. 2021; Yuan and Zhou 2020) is a recent research hotspot, engineers face some difficulties to deploy a deep transfer model for machines in the edge.

Moreover, based on the survey of existing machinery applications, it is worth noting that fault diagnosis tasks draw more attention than RUL prediction tasks. Two reasons can account for it: (1) limitation of available datasets. Many open-source datasets are collected for fault classification, but there are fewer RUL datasets because it is hard and much more time-consuming to collect the complete data of machines' full life cycles, and (2) limitation of deep models. Deep network architectures are fit for classification tasks and can achieve high accuracy, but for regression tasks, their training process tends to be unstable and it is not easy to obtain satisfactory prediction results. Most TL models are originally designed for classification tasks and thus not always effective for regression tasks (Ding et al. 2021). Nevertheless, the RUL prediction tasks, which contribute to maintenance planning and avoidance of large economic loss, are quite important for real-world industrial machines.

Among the previously discussed TL or domain adaptation methods, MMD is the most widely-used metric for measuring the distribution distance between a source and target. Apart from MMD, there are many other distance metrics. For example, Maximum Density Divergence (Li et al. 2020), Kullback-Leibler divergence (Qian et al.

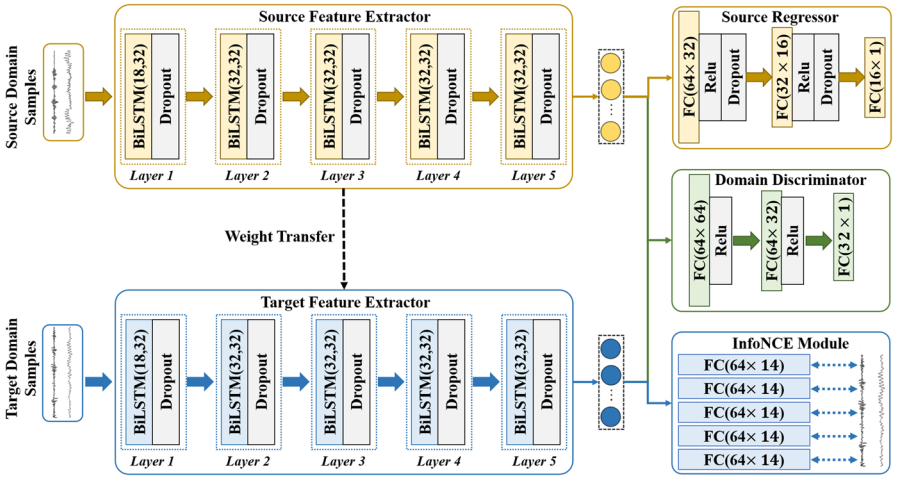


Fig. 15 General framework of CADA (Ragab et al. 2020)

2019; Sun et al. 2018) and Wasserstein distance (Cheng et al. 2020; Jiao et al. 2020; Yang B et al. 2021; Shen et al. 2018 and Chen et al. 2018) can be used for cross-domain distance metrics. But such advanced development in TL has yet been widely applied to machinery tasks of diagnostics and prognostics.

4 Open-source machinery datasets

Most machinery datasets are based on the sensor data (such as speed, vibration and load) and other environmental data (e.g., temperature and humidity). Monitoring systems record and store such data every day. Therefore, most datasets are constituted by time-series data that covers sensor records in a certain period. It is noted that image datasets are also used in the machinery area. But there are no open-source image datasets, to our best knowledge. Some research (Ko and Kim 2019; Shao et al. 2018; Si et al. 2021; Wen et al. 2017; Hasan et al. 2019; Saufi et al. 2020; Oh et al. 2017) produces their own image datasets by converting original sensor data into time-frequency images, or two-dimensional images with variable and time axes, or acoustic spectral images. Besides, some work (Shao et al. 2020; Janssens et al. 2018; Janssens et al. 2017; Nasiri et al. 2019) uses an infrared camera to take thermal images and then trains deep learning models to process these images. However, according to our best knowledge, no such image-based machinery datasets are online available. Therefore, in this paper, we introduce ten widely-used time-series datasets in machinery diagnostics and prognostics, including CWRU (2003), Gearbox Dataset (2018) by Southeast University, Paderborn University Dataset (2016), PHM09 Gearbox Dataset (2009), DIRG (Daga et al. 2019), MFPT (2013), FEMTO (2012) (PRONOSTIA Dataset), IMS (Qiu et al. 2006), C-MAPSS (Saxena et al. 2008) and N-CMAPSS (Chao et al. 2021). As shown in Table 6, most of them are used for the classification tasks of fault diagnostics. FEMTO (PRONOSTIA), IMS, C-MAPSS and N-CMAPSS datasets are for the regression task of RUL prediction. More prognostic datasets can be found in the Prognostics Data Repository (http://ti.arc.nasa.gov/project/prognostic-data-repository) which is collected by NASA and includes various kinds of run-to-failure data, such as battery usage data (Bole et al. 2014) and capacitor degradation data (Renwick et al. 2015).

4.1 CWRU dataset

CWRU dataset (2003) is a bearing dataset collected by Case Western Reserve University. Many types of data are included in this dataset, but here we detailed the most widely used data which are for carrying out transfer experiment(the same for other datasets). CWRU is composed of data from both normal and faulty bearings with single-point drive end defects or fan end defects. Three positions in bearing, i.e., the inner raceway, rolling element (i.e., ball) and outer raceway, are with different faults. There are three kinds of single-point faults with fault diameters of 0.007, 0.014, and 0.021 inches. Faulted bearings are reinstalled into the test motor with loads of 0 to 3 horsepower (motor speeds of 1797 to 1730 rpm), as in Table 7. Vibration signals data (as shown in Fig. 16) are collected and then processed in a Matlab environment. Therefore,

Table 4 Accuracy (%) results of different transfer fault diagnosis methods on CWRU

Transfer task	Model and parameter transfer	Feature matching-based transfer[1]							Adversarial Adaptation -based Transfer		
	Zhao et al. (2020)	DFADA (Qian et al. 2021)	DAGSZ (Zhang et al. 2020)	MACNN (Jin et al. 2021)	DTLCNN (Zhu et al. 2019)	MDIAN (Wang et al. 2020)	Li et al. 2020)	Kim et al. (2021)	Jiao et al. (2020)	Zhao et al. (2021)	Huang et al. (2021)
L0-L1	99.20	100	99.40	99.53	94.17	99.60	98.80	99.30	99.20	99.16	99.98
L0-L2	99.17	99.95	99.00	99.96	95.00	99.30	98.20	96.20	99.37	99.24	99.99
L0-L3	99.33	100	99.60	99.17	92.50	99.10	99.50	91.90	99.37	98.89	99.97
L1-L0	98.80	100	100	99.32	95.83	99.70	98.8	99.50	99.01	99.11	99.98
L1-L2	99.57	100	99.10	99.96	93.33	99.65	99.60	99.80	99.92	99.75	100
L1-L3	99.63	100	98.90	99.01	91.67	99.80	99.40	96.80	99.31	99.42	99.99
L2-L0	98.93	99.94	99.8	98.93	95.00	97.60	98.20	99.10	99.13	99.28	98.92
L2-L1	99.27	100	99.70	98.41	95.00	99.45	98.30	99.70	99.40	99.05	99.42
L2-L3	99.43	100	100	99.41	99.17	99.45	99.30	99.30	99.40	99.38	100
L3-L0	98.60	100	99.30	98.84	97.50	97.45	98.40	96.90	98.84	99.05	99.31
L3-L1	99.20	100	99.30	98.58	95.83	98.60	98.90	98.90	99.24	99.20	99.44
L3-L2	99.67	100	99.80	99.97	92.50	99.50	99.30	99.90	99.61	99.31	100
Average	99.23	99.99	99.49	99.26	94.79	99.10	98.89	98.11	99.32	99.24	99.75

[1] DFADA and DAGSZ are shallow feature matching methods, and the others listed in this category are deep ones

Table 5 RMSE results of different transfer fault diagnosis methods on C-MAPSS

Transfer task	Feature matching-based transfer[1]					Adversarial adaptation-based transfer		
	TCA-NN	TCA-DNN	CORAL-NN	CORAL-DNN	Zhang et al. (2021)	LSTM -DANN (da Costa et al. 2020)	Ragab et al. (2020)	CADA (Ragab et al. 2020)
FD001-FD002	94.10	90.00	99.20	77.50	43.40	46.40	19.87	19.52
FD001-FD003	120.00	116.10	60.00	69.60	24.50	37.30	39.74	39.58
FD001-FD004	20.10	113.80	107.70	84.60	45.10	43.50	31.78	31.23
FD002-FD001	94.70	85.60	77.90	80.90	34.60	31.20	14.33	13.88
FD002-FD003	107.40	111.50	60.90	79.80	43.20	32.20	32.60	33.53
FD002-FD004	93.50	94.40	37.50	43.60	33.10	27.70	34.35	33.71
FD003-FD001	98.70	90.50	26.50	26.50	18.30	30.60	19.97	19.54
FD003-FD002	90.50	80.80	113.20	75.60	48.80	43.10	23.47	19.33
FD003-FD004	78.90	102.60	113.90	77.20	52.50	49.70	26.33	20.61
FD004-FD001	98.50	85.60	119.10	94.00	35.30	25.40	37.89	20.10
FD004-FD002	75.30	80.80	37.30	30.90	24.60	26.90	28.77	18.50
FD004-FD003	77.20	102.90	68.10	68.60	38.70	23.60	14.13	14.49
Average	87.41	96.22	76.78	67.40	36.84	34.80	26.94	23.67

[1]The results of feature matching-based transfer methods, i.e., TCA and CORAL, are cited from da Costa et al. (2020). TCA-NN method use TCA feature representations to train a shallow NN for RUL prediction, while TCA-DNN trains a deep NN. CORAL-NN and CORAL-DNN use the same combination

all data files in CWRU dataset are in *.mat format. Each file includes fan vibration data, drive end vibration data, motor rotational speed, and time information.

4.2 Gearbox dataset

Gearbox Dataset (2018) is from Southeast University, China, and can be found at mlmechanics.ics.uci.edu. This dataset is divided into two sub-datasets: bearing data and gear data, which are both acquired on Drivetrain Dynamics Simulator (DDS). Figure 17 shows the experimental setup of the test rig. All the data are the original vibration signals acquired by sensors. For the bearing dataset, besides the normal healthy working state, there are four types of faults, i.e., ball fault, inner ring fault, outer ring fault and a combination fault with both inner ring and outer ring. Gear data also contain four faults: chipped tooth, missing tooth, root fault and surface fault. Both are collected under two kinds of working conditions. For the names of data files, their suffix "20-0" or "30-2" represents the "rotating speed-load" working configuration. All files in this dataset are in the format of *.csv. Within each file, for each row of record, there are eight variables that represent different vibration signals.

4.3 Paderborn university dataset

Paderborn University Dataset (2016) provides the data of rolling bearings. A general view of the test rig is presented in Fig. 18a. There are 26 faulty bearing states and 6 healthy states. Among the faulty states, both 12 artificially damaged states and 14 real damages exist at the inner and outer ring of the ball bearing. As shown in Fig. 18b, artificial faults are introduced by manually drilling, electric discharge and engraving. Besides, accelerated lifetime tests are performed to acquire the real bearing damages. For the experiment, the test rig is operated under four operating conditions with three different main operation parameters, i.e., the rotational speed of the drive system, radial force onto the test bearing and the load torque in the drive train. For each setting, 20 measurements in every 4 seconds are recorded. All data files are restored via Matlab and saved as *.mat files.

4.4 PHM09 gearbox dataset

PHM09 Gearbox Dataset (2009) comes from the IEEE PHM (Prognostics and Health Management) 2009 Challenge Competition. It is a representative of generic industrial gearbox data with gear faults (crack in gear, cracked/broken tooth, and excessive wear or clearance), bearing faults (bearing race defect, and excessive bearing clearance) and shaft faults (rotor imbalance, shaft misalignment, and mechanical looseness). Figure 19a shows the test rig and three examples of gears (Left to right: normal, missing tooth, chipped tooth). Two kinds of gear are used, i.e., a spur gear and a spiral cut (helical) gear. Moreover, data are collected at five different shaft speeds (30, 35, 40, 45 and 50 Hz) under high and low loads. In addition, data from different repeated runs are presented in this dataset. There are a total of 560 runs. For each run, data are restored in a *.csv file that has three columns (input voltage, output voltage and tachometer).

Table 6 Summary of open-source machinery datasets

Task	Dataset	Source	Data format	Download source
Fault Classification	CWRU (2003)	Case Western Reserve University	*.mat	https://engineering.case.edu/bearingdatacenter/download-data-file
	Gearbox Dataset (2018)	Southeast University	*.csv	http://mlmechanics.ics.uci.edu/data/mechanical/
	Paderborn University Dataset (2016)	Paderborn University	*.mat	https://mb.uni-paderborn.de/kat/forschung/datacenter/bearing-data-center/data-sets-and-download
	DIRG (Daga et al. 2019)	Dynamic and Identification Research Group	*.mat	https://zenodo.org/record/ 3559553#.YhCA1-gzY2w
	PHM09 Gearbox (2009)	IEEE PHM 2009 Challenge Competition	*.csv	https://c3.ndc.nasa.gov/dashlink/resources/997/
	MFPT (2013)	Society for Machinery Failure Prevention Technology	*.mat	https://www.mfpt.org/fault-data-sets/
RUL Prediction	FEMTO (PRONOSTIA Dataset) (2012)	Franche-Comté Electronics Mechanics Thermal Science and Optics-Sciences and Technologies, IEEE PHM 2012 Challenge	*.csv	https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/#femto
	IMS (Qiu et al. 2006)	Intelligent Maintenance System	*.txt	https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/#bearing
	C-MAPSS (Saxena et al. 2008), N-CMAPSS Arias (Chao et al. 2021)	NASA Commercial Modular Aero-Propulsion System Simulation	*.txt files in C-MAPSS, and *.h5 files in N-CMAPSS	C-MAPSS can be found at https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/#turbofan, and N-CMAPSS is at https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/#turbofan-2

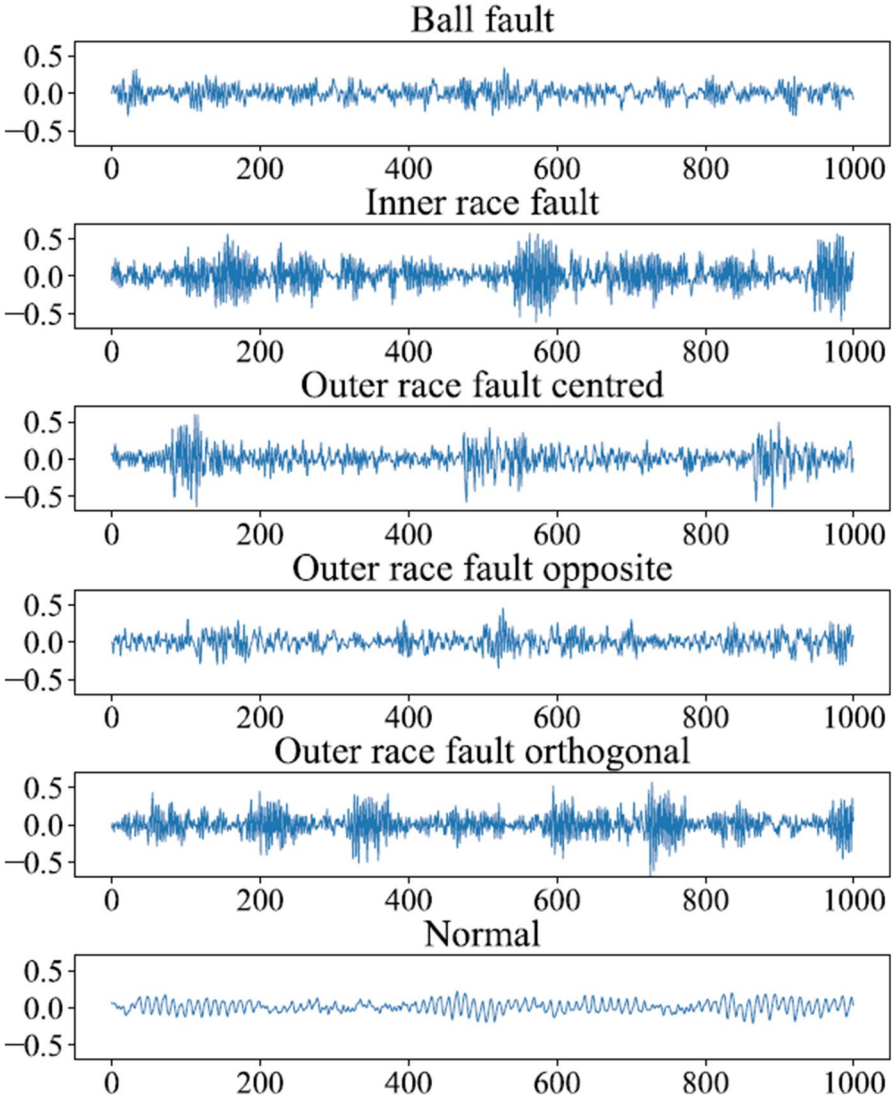


Fig. 16 Examples of vibration signals (time v.s. vibration amplitude) under different faults in CWRU (2003)

Table 7 Overview of CWRU dataset

Operating condition	Speed (rpm)	Load (hp)	Number of health conditions	Number of total samples
L0	1797	0	10	10000
L1	1772	1	10	10000
L2	1750	2	10	10000
L3	1730	3	10	10000

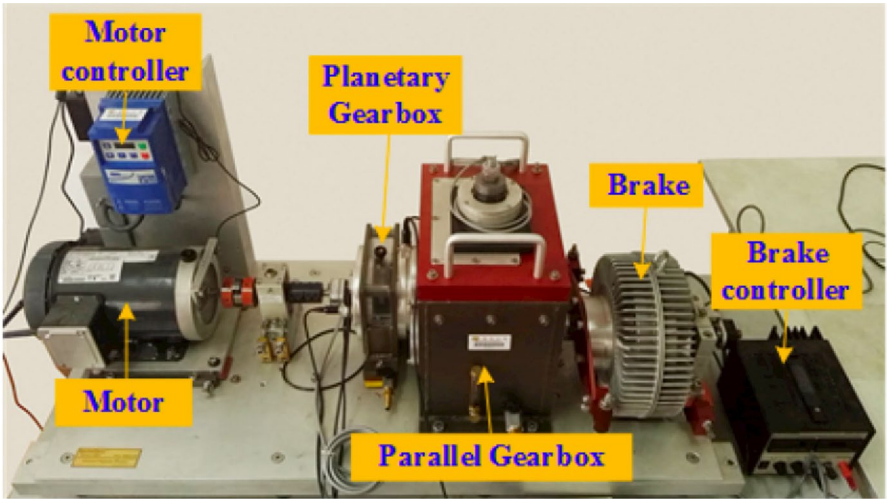
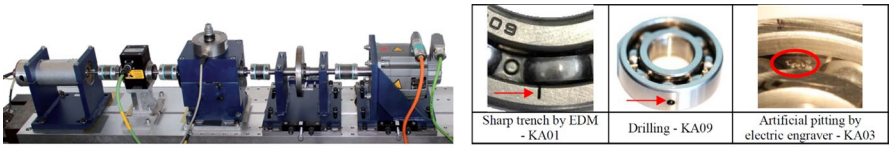


Fig. 17 General view of the test rig in Gearbox dataset (2018)



(a) General view of the test rig **(b)** Three types of artificial damages

Fig. 18 Examples in Paderborn University dataset (2016)

4.5 DIRG dataset

The DIRG dataset Daga et al. (2019) is acquired on the rolling bearing test rig of the Dynamic and Identification Research Group (DIRG). Figure 20a presents its general view. There are two different experiments. The first one reports the vibration data of bearings operating with different damage and different speed and load combinations (speed: 0, 100, 200, 300, 400, 500Hz; load: 0, 1000, 1400, 1800N). It can be used for the classification task of fault diagnostics. The second one presents the vibration data of a single damaged bearing tested at constant speed and load for a long period of time (about 330 hours). It monitors the damage evolution, which can be used for the regression task of RUL prediction. Apart from the healthy state, three types of faults are produced to cause a conical indentation on the inner ring or on a single roller. The faults are of different severities, and the diameter of the resulting circular indentation is 150, 250 and 450 μm, respectively. Six channels of vibration signals are recorded. They correspond to the outputs of accelerometers placed in axial and radial points in x, y, and z directions. Data are recorded in Matlab *.mat files.

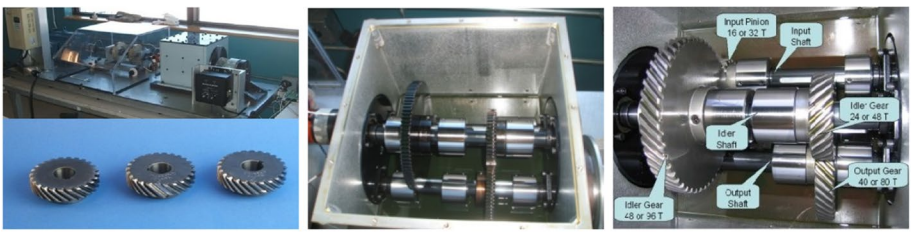
4.6 MFPT dataset

The MFPT dataset (2013) is from the Society for Machinery Failure Prevention Technology (MFPT). It records vibration data of rolling bearing. Besides normal healthy conditions, two faulty conditions (outer race fault and inner race fault under various loads) are introduced in a bearing test rig. Three healthy datasets are provided and serve as the baseline condition (270 lbs of load, input shaft rate of 25 Hz, sample rate of 97656 sps for 6 seconds). Under the same condition, three datasets with outer race faults are provided. Besides, under a sample rate of 48,828 sps for 3 seconds, seven more outer race fault datasets are provided with seven different loads (25, 50, 100, 150, 200, 250 and 300 lbs), and seven inner race fault datasets are provided with seven different loads (0, 50, 100, 150, 200, 250 and 300 lbs). All data are stored in Matlab *.mat files. Additionally, three real-world examples are also presented in MFPT dataset: an intermediate shaft bearing from a wind turbine, a planet bearing fault, and an oil pump shaft bearing from a wind turbine. Figure 20b shows an example of the vibration signal under the inner race fault in this dataset.

4.7 FEMTO dataset

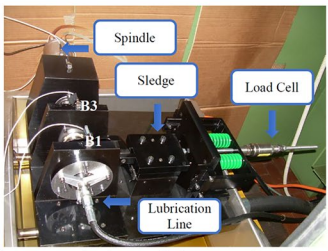
FEMTO Dataset (2012) includes vibration data of bearings. It is from FEMTO-ST Institute (Franche-Comté Electronics Mechanics Thermal Science and Optics-Sciences and Technologies). This dataset is used for estimation of the RUL of bearings and served as the data in the IEEE PHM 2012 Challenge Competition. Its data are acquired from experiments carried on a laboratory experimental platform called PRONOSTIA (Nectoux et al. 2012). So, it is also known as PRONOSTIA Dataset. The test rig is shown in Fig. 20c.

This dataset has two sub-datasets: training one with 6 bearings and test one with 11 bearings. The former is acquired from run-to-failure experiments, and the latter is set to include truncated experimental data. Note that, the specific faulty type is not declared in each dataset. Consequently, most methods (Sutrisno et al. 2012; Mosallam et al. 2013; Li and Wang 2013; Sloukia et al. 2013) use this dataset for estimating the remaining useful life of ball bearings. It data are collected under three different operation conditions as summarized in Table 8. Vibration and temperature signals are measured to monitor the health of the test bearings. Every record is stored as a *.csv file. Each bearing data filecontains 268 to 3269 records. As mentioned in Sutrisno et al. (2012), there are multiple challenges in analyzing such data, i.e., limited training samples, no information about failure modes, no fixed failure threshold, and a wide range of failure times.

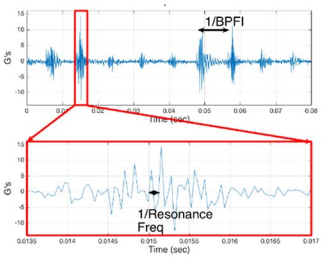


(a) Overview of the test rig (top); Gears (bottom) **(b)** The tested gearbox **(c)** Details of the gearbox

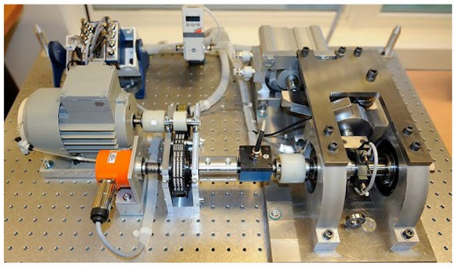
Fig. 19 Examples in PHM09 dataset (2009)



(a) General view of the test rig in DIRG



(b) An example under inner race fault in MFPT



(c) General view of PRONOSTIA platform

Fig. 20 Examples in DIRG and MFPT datasets, and an overview of PRONOSTIA platform

4.8 IMS dataset

IMS (Intelligent Maintenance System) Dataset (Qiu et al. 2006) is used for the RUL prediction of bearings. It has three datasets under test-to-failure experiments that run at 2000 rpm rotation speed with a radial load of 6000 lbs. There are eight channels of vibration data in Dataset 1 and four channels in Datasets 2 and 3. Each dataset describes a test-to-failure experiment and includes multiple individual files (*.txt files in the ASCII format) that are 1-second vibration signal snapshots recorded at specific intervals. There are 2156, 984 and 4448 files in Datasets 1-3, respectively. At the end of the test-to-failure experiment, there is an inner race defect in bearing 3 and roller element defect in bearing 4 in Dataset 1, an outer race failure in bearing 1 in Dataset 2, and outer race failure in bearing 3 in Dataset 3. As mentioned in Zhang et al. (2020), IMS Dataset is particularly appropriate for the RUL prediction of rolling bearing because it contains the complete records of vibration signals from the beginning to the failure of tested bearings and its records have time stamps.

4.9 C-MAPSS and N-CMAPSS datasets

C-MAPSS (Saxena et al. 2008) and N-CMAPSS (Arias Chao et al. 2021) Datasets originate from NASA Commercial Modular Aero-Propulsion System Simulation, and they include a great number of run-to-failure data of aircraft turbofan engine. C-MAPSS (Saxena et al. 2008) was released in 2008, and N-CMAPSS (Arias Chao et al. 2021) was online available in 2021. C-MAPSS (Saxena et al. 2008) has four sub-datasets collected under different operation conditions (represented by altitude, flight mach number and throttle resolver angle). Different faults occur in these datasets, leading to varying life spans. Each record shows the degradation process of an engine, which starts from

varying wears and runs in a healthy state for a while and then degrades to the end of record time. 21 sensor data are recorded. All files are provided in the *.txt format.

Different from C-MAPSS that only records data from standard cruise conditions, N-CMAPSS (Arias Chao et al. 2021) considers a more complete flight cycle that includes climbing, cruise and descend flight conditions in different flight routes. It also provides additional information: flight class (1, 2, or 3) determined by flight length and the number of flights, which is used for representing operating history, thus allowing a degradation model to be built with more fidelity. Additionally, a binary health state label (faulty or healthy) is included in the dataset. There are 8 sub-datasets from varying flight classes and fault modes. More sensor data are provided, i.e., 14-dimension measured signals, 14-dimension virtual sensors data and 10-dimension engine health parameters (efficiency and flow for five rotating sub-components). All data are stored in *.h5 files.

5 Challenges and future research

Figure 21 presents the overlook of TL machinery applications, including emerging challenges in yellow boxes and future directions in green boxes.

5.1 Emerging challenges

5.1.1 Generalized knowledge transfer

Most of the existing transfer diagnostics approaches are performed among strongly similar domains, e.g., different operating states of the same machine. A more general knowledge transfer, i.e., learning cross-domain knowledge from data in very different domains, remains to be explored. For example, how to transfer between different machine components (e.g., bearings vs. gearboxes) is yet to be investigated. It is considered as a difficult generalized transfer task. Moreover, many failure data are not available in practice, and hence the existing open-source datasets are obtained from simulation experiments. Laboratory simulations do not fully simulate complex scenarios such as working environments and disturbances in real industrial applications, and thus there are large differences in their data distribution. Hence, how to effectively transfer the laboratory simulation to the real-world application remains open.

Table 8 Overview of FEMTO dataset

Operating condition	Speed (rpm)	Load (N)	Number of training datasets	Number of test datasets
C1	1800	4000	2	5
C2	1650	4200	2	5
C3	1500	5000	2	1

5.1.2 Few-shot incremental data

In fault detection tasks, data that records machinery failure in industrial scenarios are usually in a smaller size (i.e., few-shot) when compared to data recording normal operation states. Thus, there is an issue of data imbalance when building diagnostics models (Zhang et al. 2020; Wang et al. 2021; Zhang et al. 2020; Liu et al. 2019; Wang et al. 2016; Han et al. 2022). Moreover, as the machine runs hour by hour and day by day, its operation data is continuously accumulated and updated. Some previously never-seen failures may occur in such an incremental data stream, which means such new failure data are inevitably scarce. In addition, without human supervision, such failure data are unknown to algorithms and models. Namely, labeled data are very limited in such an incremental data stream, leading to insufficient discriminative information. Therefore, there is no sufficient new data to train a satisfactory TL model. In sum, it is very difficult to perform knowledge transfer with few-shot incremental data. This calls for more studies.

5.1.3 Label-noise

The existing methods usually assume that data labels are completely correct. However, in practical applications, the mechanical data are not always labeled correctly due to the influence of various factors, e.g., working environment noise, human mistakes and digitalization/instrument errors. Manual checking and re-calibration of these labels are costly and sometimes it is impossible. Wrong label information brings wrong gradient information, which inevitably degrades the performance of a diagnosis and prediction model. More importantly, label-noise tends to cause negative transfer. So, it is of critical importance to deal with such issue of noisy labels, such that effective and robust transfer can be performed to assist the task of machinery diagnostics and prognostics.

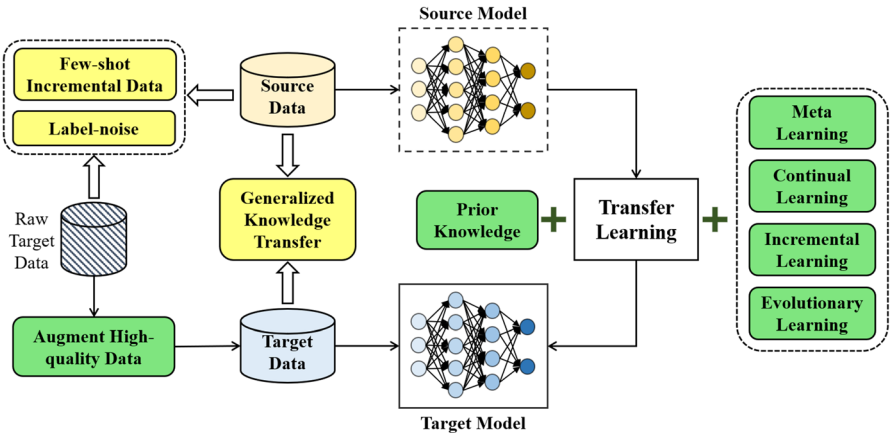


Fig. 21 Overlook of transfer learning for machinery applications

5.2 Future directions

5.2.1 Combining promising techniques

The field of machine learning has been developing rapidly in recent years, and researchers are refining related promising theories step by step. Many emerging algorithms and models, such as meta learning, continual learning, incremental learning, and evolutionary learning, can be applied to important fields, e.g., computer vision and natural language processing. However, at present, their applications in traditional industries are very limited. If we further combine them when applying TL in tasks of mechanical fault diagnosis and RUL prediction, advanced approaches can be obtained. These technologies can empower the model with self-learning ability. With meta learning, the model can automatically learn from which machine data to transfer knowledge; With continual learning, a diagnostic/prediction model learns continuously and achieves online dynamic monitoring; With incremental learning, a model learns how to deal with the constantly updated incremental sensor data in industrial sites; With evolutionary learning (Deng et al. 2022), it can automatically search a network architecture and achieve better optimization results. The application of these emerging technologies can greatly enhance the self-adaptive capability of TL models, which enables cost-effective industrial applications (Shi et al. 2022).

5.2.2 Supporting complex applications

Many of the current investigations focus on the classical standard settings of transfer tasks, i.e., homogeneous, feature-aligned, and closed-set. As for labels in the target domain, TL methods have been well elaborated to handle unsupervised learning problems. But in many machinery diagnosis and prediction experiments, their target mechanical datasets still use fully or partially labeled data. Obviously, this is a relatively simple task setting. However, the situation may be more complex in real scenarios. For example, feature dimensions are different when transferring (Qin et al. 2021), and labels of fault types are not exactly consistent between source and target domains (Li et al. 2020; Chai et al. 2021). The data in complex application cases do not conform to the above task assumptions, thus limiting the application of some transfer methods in diagnosis and prediction. Current research on such transfer tasks is scarce and needs in-depth exploration.

5.2.3 Leveraging prior knowledge

Existing fault diagnosis and prediction approaches for mechanical components usually borrow ideas from well-established TL methods that have been developed in other fields. Nevertheless, how to design specific transfer methods for mechanical diagnosis that integrate its task characteristics when transferring, remains a question to be answered. Prior knowledge about fault failure information and expert insights in related fields can be used to instruct data preprocessing, such as extracting representative features based on some professional understanding such that they well characterize different types of faults. Beyond that, how to take advantage of prior knowledge and expert insights in the process of transferring across domains and how to maximize the useful information embedding in the prior knowledge needs further research.

5.2.4 Augmenting high-quality data

Mechanical data often face the situation of sparse data and scarce labels, which is very unfavorable for the training and learning of transfer models. Employing data augmentation techniques to generate more available samples is a feasible solution to such issue. Data generation methods based on GAN and auto-encoder have achieved desired results in the image processing field, which inspires researchers to apply them in the industrial field. They can be applied to generate sufficient data needed for transfer model training. Especially, when failure data is scarce and with noisy labels as discussed above, generative models can be used to augment useful data and produce more high-quality training samples, so as to ensure the model performance.

6 Conclusion

This survey comprehensively explains different cross-domain transfer methods applied in machinery diagnostics and prognostics, which are summarized into three categories: model and parameter transfer, feature matching (including shallow and deep methods), and adversarial adaptation. Main ideas, typical algorithms and models of representative methods in each category are introduced in detail. More importantly, their application to machinery diagnosis and prognostics are presented in the context of recent related investigations, e.g., deep transfer methods are a research hotspot in machinery diagnostics and prognostics, and diverse transfer architectures are developed for different application tasks. Although many studies place more emphasis on the classification task of machinery fault diagnostics, the regression task of remaining useful life prediction (Chen et al 2021; Jiao et al 2021; Lin et al 2021) is of equal importance in the industrial field. Besides, ten widely-used open-source machinery datasets are briefly introduced in this survey. At last, according to the previous discussions, emerging challenges of applying transfer learning in machinery diagnostics/prognostics and potential directions are given.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant 51775385 and Grant 61703279, in part by the Strategy Research Project of Artificial Intelligence Algorithms of Ministry of Education of China, in part by the Shanghai Industrial Collaborative Science and Technology Innovation Project (2021-cyxt2-kj10), in part by the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0100) and the Fundamental Research Funds for the Central Universities, and in part by the Ministry of Education and King Abdulaziz University (KAU)/Deanship of Scientific Research (DSR), Jeddah, Saudi Arabia via Institutional Fund Projects under grant no. (IFPRP: 693-135-1442). We are also grateful for the efforts from our colleagues in Sino-German Center of Intelligent Systems, Tongji University.

References

Arias Chao M, Kulkarni C, Goebel K, Fink O (2021) Aircraft engine run-to-failure dataset under real flight conditions for prognostics and diagnostics. Data 6(1):5

Azamfar M, Li X, Lee J (2020) Intelligent ball screw fault diagnosis using a deep domain adaptation methodology. Mech Mach Theory 151:103932

Baktashmotlagh M, Harandi MT, Lovell BC, Salzmann M (2013) Unsupervised domain adaptation by domain invariant projection. In: Proceedings of the IEEE International conference on computer vision, pp 769–776

Bole B, Kulkarni CS, Daigle M (2014) Adaptation of an electrochemistry-based li-ion battery model to account for deterioration observed under randomized use. Inc., Moffett Field United States, Technical report, SGT
Borgwardt KM, Gretton A, Rasch MJ, Kriegel H-P, Schölkopf B, Smola AJ (2006) Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics 22(14):49–57
Cai G, Wang Y, He L, Zhou M (2019) Unsupervised domain adaptation with adversarial residual transform networks. IEEE Trans Neural Netw Learn Syst 31(8):3073–3086
Case Western Reserve University Bearing Data Center, CWRU Dataset. https://csegroups.case.edu/bearingdatacenter
Chai Z, Zhao C, Huang B (2021) Multisource-refined transfer network for industrial fault diagnosis under domain and category inconsistencies. IEEE Trans Cybernet
Chen Z, Gryllias K, Li W (2019) Intelligent fault diagnosis for rotary machinery using transferable convolutional neural network. IEEE Trans Ind Inform 16(1):339–349
Chen W, Qiu Y, Feng Y, Li Y, Kusiak A (2021) Diagnosis of wind turbine faults with transfer learning algorithms. Renew Energy 163:2053–2067
Chen C, Shen F, Xu J, Yan R (2021) Model parameter transfer for gear fault diagnosis under varying working conditions. Chin J Mech Eng 34(1):1–13
Chen Q, Liu Y, Wang Z, Wassell I, Chetty K (2018) Re-weighted adversarial adaptation network for unsupervised domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7976–7985
Chen C, Lu N, Jiang B, Wang C (2021) A Risk-Averse Remaining Useful Life Estimation for Predictive Maintenance. IEEE/CAA J Autom Sin 8(2):412–422. https://doi.org/10.1109/JAS.2021.1003835
Cheng C, Zhou B, Ma G, Wu D, Yuan Y (2020) Wasserstein distance based deep adversarial transfer learning for intelligent fault diagnosis with unlabeled or insufficient labeled data. Neurocomputing 409:35–45
da Costa PRDO, Akçay A, Zhang Y, Kaymak U (2020) Remaining useful lifetime prediction via deep domain adaptation. Reliab Eng System Saf 195:106682
Daga AP, Fasana A, Marchesiello S, Garibaldi L (2019) The politecnico di torino rolling bearing test rig: Description and analysis of open access data. Mech Syst Signal Process 120:252–273
Deebak B, Al-Turjman F (2021) Digital-twin assisted: fault diagnosis using deep transfer learning for machining tool condition. Int J Intell Syst
Deng M, Deng A, Zhu J, Shi Y, Liu Y (2021) Intelligent fault diagnosis of rotating components in the absence of fault data: a transfer-based approach. Measurement 173:108601
Deng Q, Kang Q, Zhang L, Zhou M, An J (2022) Objective Space-based Population Generation to Accelerate Evolutionary Algorithms for Large-scale Many-objective Optimization. IEEE Trans Evol Comput 1–1:9762228. https://doi.org/10.1109/TEVC.2022.3166815
Ding Y, Ding P, Jia M (2021) A novel remaining useful life prediction method of rolling bearings based on deep transfer auto-encoder. IEEE Trans Instrum Measure 70:1–12
Ding Y, Jia M, Cao Y (2021) Remaining useful life estimation under multiple operating conditions via deep subdomain adaptation. IEEE Trans Instrum Measure 70:1–11
Ding Y, Jia M, Miao Q, Huang P (2021) Remaining useful life estimation using deep metric transfer learning for kernel regression. Reliab Eng Syst Saf 212:107583
Dong Y, Li Y, Zheng H, Wang R, Xu M (2022) A new dynamic model and transfer learning based intelligent fault diagnosis framework for rolling element bearings race faults: Solving the small sample problem. ISA Trans 121:327–348
FEMTO-ST Institute, FEMTO Dataset. https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/#femto
Fernando B, Habrard A, Sebban M, Tuytelaars T (2013) Unsupervised visual domain adaptation using subspace alignment. In: Proceedings of the IEEE international conference on computer vision, pp 2960–2967
Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, Marchand M, Lempitsky V (2016) Domain-adversarial training of neural networks. J Mach Learn Res 17(1):2030–2096
Ghifary M, Kleijn WB, Zhang M (2014) Domain adaptive neural networks for object recognition. In: Pacific Rim International Conference on Artificial Intelligence, pp 898–904. Springer
Gong B, Shi Y, Sha F, Grauman K (2012) Geodesic flow kernel for unsupervised domain adaptation. In: 2012 IEEE Conference on computer vision and pattern recognition, pp. 2066–2073. IEEE
Gopalan R, Li R, Chellappa R (2011) Domain adaptation for object recognition: an unsupervised approach. In: 2011 International conference on computer vision, pp. 999–1006. IEEE
Gretton A, Borgwardt K, Rasch M, Schölkopf B, Smola A (2006) A kernel method for the two-sample-problem. Adv Neural Inform Process Syst 19:513–520

Guo L, Lei Y, Xing S, Yan T, Li N (2018) Deep convolutional transfer learning network: a new method for intelligent fault diagnosis of machines with unlabeled data. IEEE Trans Ind Electron 66(9):7316–7325
Han T, Liu C, Wu R, Jiang D (2021) Deep transfer learning with limited data for machinery fault diagnosis. Appl Soft Comput 103:107150
Han T, Liu C, Yang W, Jiang D (2019) A novel adversarial learning framework in deep convolutional neural network for intelligent diagnosis of mechanical faults. Knowl-Based Syst 165:474–487
Han T, Liu C, Yang W, Jiang D (2020) Deep transfer network with joint distribution adaptation: a new intelligent fault diagnosis framework for industry application. ISA Trans 97:269–281
Han H, Ma W, Zhou M, Guo Q, Abusorrah A (2020) A novel semi-supervised learning approach to pedestrian reidentification. IEEE Internet Things J 8(4):3042–3052
Han S, Zhu K, Zhou M, Liu X (2022) Evolutionary weighted broad learning and its application to fault diagnosis in self-organizing cellular networks. IEEE transactions on cybernetics, 1–13. https://doi.org/10.1109/TCYB.2021.3126711
Hasan MJ, Islam MM, Kim J-M (2019) Acoustic spectral imaging and transfer learning for reliable bearing fault diagnosis under variable speed conditions. Measurement 138:620–631
He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
Henaff, O.: Data-efficient image recognition with contrastive predictive coding. In: International Conference on Machine Learning, pp. 4182–4192 (2020). PMLR
Huang Z, Lei Z, Wen G, Huang X, Zhou H, Yan R, Chen X (2021) A multi-source dense adaptation adversarial network for fault diagnosis of machinery. IEEE Trans Ind Electron 69:6298–6307
Huang G, Zhang Y, Ou J (2021) Transfer remaining useful life estimation of bearing using depth-wise separable convolution recurrent network. Measurement 176:109090
Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017): Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708
Janssens O, Loccufier M, Van Hoecke S (2018) Thermal imaging and vibration-based multisensor fault detection for rotating machinery. IEEE Trans Ind Electron 15(1):434–444
Janssens O, Van de Walle R, Loccufier M, Van Hoecke S (2017) Deep learning for infrared thermal image based machine health monitoring. IEEE/ASME Transa MechD 23(1):151–159
Jiao J, Lin J, Zhao M, Liang K (2020) Double-level adversarial domain adaptation network for intelligent fault diagnosis. Knowl-Based Syst 205:106236
Jiao R, Peng K, Dong J (2021) Remaining useful life prediction for a roller in a hot strip mill based on deep recurrent neural networks. IEEE/CAA J Autom Sin 8(7):1345–1354
Jiao J, Zhao M, Lin J, Liang K (2020) Residual joint adaptation adversarial network for intelligent transfer fault diagnosis. Mech Syst Signal Process 145:106962
Jin T, Yan C, Chen C, Yang Z, Tian H, Guo J (2021) New domain adaptation method in shallow and deep layers of the CNN for bearing fault diagnosis under different working conditions. Int J Adv Manuf Technol 12:1–12
Kang Q, Yao S, Zhou M, Zhang K, Abusorrah A (2020) Enhanced subspace distribution matching for fast visual domain adaptation. IEEE Trans Comput Soc Syst 7(4):1047–1057
Kang Q, Yao S, Zhou M, Zhang K, Abusorrah A (2020) Effective visual domain adaptation via generative adversarial distribution matching. IEEE Trans Neural Netw Learn Syst 32(9):3919–3929
Kim M, Ko JU, Lee J, Youn BD, Jung JH, Sun KH (2021) A domain adaptation with semantic clustering (dasc) method for fault diagnosis of rotating machinery. ISA transactions
Ko T, Kim H (2019) Fault classification in high-dimensional complex processes using semi-supervised deep convolutional generative models. IEEE Trans Ind Inform 16(4):2868–2877
Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inform Process Syst 25:1097–1105
Lei Y, Li N, Guo L, Li N, Yan T, Lin J (2018) Machinery health prognostics: a systematic review from data acquisition to rul prediction. Mech Syst Signal Process 104:799–834
Li J, Chen E, Ding Z, Zhu L, Lu K, Shen HT (2020) Maximum density divergence for domain adaptation. IEEE Trans Pattern Anal Mach Intell 43(11):3918–3930
Li W, Chen Z, He G (2020) A novel weighted adversarial transfer network for partial domain fault diagnosis of machinery. IEEE Trans Ind Inform 17(3):1753–1762
Li X, Hu Y, Li M, Zheng J (2020) Fault diagnostics between different type of components: A transfer learning approach. Appl Soft Comput 86:105950
Li H, Hu G, Li J, Zhou M (2021) Intelligent fault diagnosis for large-scale rotating machines using binarized deep neural networks and random forests. IEEE Trans Autom Sci Eng 5:1–11. https://doi.org/10.1109/TASE.2020.3048056

Li X, Jia X-D, Zhang W, Ma H, Luo Z, Li X (2020) Intelligent cross-machine fault diagnosis approach with deep auto-encoder and domain adaptation. Neurocomputing 383:235–247
Li X, Jiang H, Wang R, Niu M (2021) Rolling bearing fault diagnosis using optimal ensemble deep transfer network. Knowl-Based Syst 213:106695
Li X, Jiang H, Zhao K, Wang R (2019) A deep transfer nonnegativity-constraint sparse autoencoder for rolling bearing fault diagnosis with few labeled data. IEEE Access 7:91216–91224
Li X, Li X, Ma H (2020) Deep representation clustering-based fault diagnosis method with unsupervised data applied to rotating machinery. Mech Syst Signal Process 143:106825
Li Q, Shen C, Chen L, Zhu Z (2021) Knowledge mapping-based adversarial domain adaptation: a novel fault diagnosis method with high generalizability under variable working conditions. Mech Syst Signal Process 147:107095
Li X, Zhang W, Ding Q (2018) A robust intelligent fault diagnosis method for rolling element bearings based on deep distance metric learning. Neurocomputing 310:77–95
Li X, Zhang W, Ma H, Luo Z, Li X (2020) Partial transfer learning in machinery cross-domain fault diagnostics using class-weighted adversarial networks. Neural Netw 129:313–322
Li X, Zhang W, Ma H, Luo Z, Li X (2020) Deep learning-based adversarial multi-classifier optimization for cross-domain machinery fault diagnostics. J Manuf Syst 55:334–347
Li T, Zhao Z, Sun C, Yan R, Chen X (2021) Domain adversarial graph convolutional network for fault diagnosis under variable working conditions. IEEE Trans Instrument Measure 70:1–10
Li H, Wang Y (2013) Rolling bearing reliability estimation based on logistic regression model. In: 2013 International conference on quality, reliability, risk, maintenance, and safety engineering (QR2MSE), pp. 1730–1733. IEEE
Liao Y, Huang R, Li J, Chen Z, Li W (2021) Dynamic distribution adaptation based transfer network for cross domain bearing fault diagnosis. Chin J Mech Eng 34(1):1–10
Lin J, Lin Z, Liao G, Yin H (2021) A Novel Product Remaining Useful Life Prediction Approach Considering Fault Effects. IEEE/CAA J Autom Sin 8(11):1762–1773. https://doi.org/10.1109/JAS.2021.1004168
Liu K, Ye Z, Guo H, Cao D, Chen L, Wang FY (2021) FISS GAN: A Generative Adversarial Network for Foggy Image Semantic Segmentation. IEEE/CAA J Autom Sin 8(8):1428–1439
Liu M, Li X, Chakrabarty K, Gu X (2022) Knowledge transfer in board-level functional fault diagnosis enabled by domain adaptation. IEEE Trans Comput-Aided Des Integr Circuits Syst 41(3):762–775
Liu M-Y, Tuzel O (2016) Coupled generative adversarial networks. Adv Neural Inform Process Syst 29:469–477
Liu H, Zhou M, Liu Q (2019) An embedded feature selection method for imbalanced data classification. IEEE/CAA J Autom Sin 6(3):703–715
Long M, Zhu H, Wang J, Jordan MI (2017) Deep transfer learning with joint adaptation networks. In: International Conference on Machine Learning, pp. 2208–2217. PMLR
Long M, Wang J, Ding G, Sun J, Yu PS (2013) Transfer feature learning with joint distribution adaptation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2200–2207
Long M, Wang J, Ding G, Sun J, Yu PS (2014) Transfer joint matching for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1410–1417 (2014)
Long M, Cao Y, Wang J, Jordan M (2015) Learning transferable features with deep adaptation networks. In: International conference on machine learning, pp 97–105. PMLR
Lu W, Liang B, Cheng Y, Meng D, Yang J, Zhang T (2016) Deep model based domain adaptation for fault diagnosis. IEEE Trans Ind Electron 64(3):2296–2305
Lu N, Xiao H, Sun Y, Han M, Wang Y (2021) A new method for intelligent fault diagnosis of machines based on unsupervised domain adaptation. Neurocomputing 427:96–109
Ma P, Zhang H, Fan W, Wang C (2020) A diagnosis framework based on domain adaptation for bearing fault diagnosis across diverse domains. ISA Trans 99:465–478
Mao W, He J, Zuo MJ (2019) Predicting remaining useful life of rolling bearings based on deep feature representation and transfer learning. IEEE Trans Instrum Measure 69(4):1594–1608
Miao M, Yu J (2021) A deep domain adaptative network for remaining useful life prediction of machines under different working conditions and fault modes. IEEE Trans Instrum Measure 70:1–14
Michau G, Fink O (2021) Unsupervised transfer learning for anomaly detection: application to complementary operating condition transfer. Knowl-Based Syst 216:106816
Mosallam A, Medjaher K, Zerhouni N (2013) Nonparametric time series modelling for industrial prognostics and health management. The Int J Adv Manuf Technol 69(5–8):1685–1699
NASA Ames Prognostics Data Repository. http://ti.arc.nasa.gov/project/prognostic-data-repository

Nasiri A, Taheri-Garavand A, Omid M, Carlomagno GM (2019) Intelligent fault diagnosis of cooling radiator based on deep learning analysis of infrared thermal images. Appl Thermal Engi 163:114410
Nectoux P, Gouriveau R, Medjaher K, Ramasso E, Chebel-Morello, B., Zerhouni, N., Varnier, C.: PRONOSTIA: An experimental platform for bearings accelerated degradation tests. In: IEEE International conference on prognostics and health management, PHM'12., pp. 1–8 (2012). IEEE Catalog Number: CPF12PHM-CDR
Oh H, Jung JH, Jeon BC, Youn BD (2017) Scalable and unsupervised feature engineering using vibration-imaging and deep learning for rotor system diagnosis. IEEE Trans Ind Electron 65(4):3539–3549
PHM Society, PHM09 Gearbox Datasets. https://phmsociety.org/public-data-sets/
Paderborn University, Paderborn University Dataset. https://mb.uni-paderborn.de/kat/datacenter
Pan SJ, Tsang IW, Kwok JT, Yang Q (2010) Domain adaptation via transfer component analysis. IEEE Trans Neural Netw 22(2):199–210
Qian W, Li S, Jiang X (2019) Deep transfer network for rotating machine fault analysis. Pattern Recognit 96:106993
Qian W, Li S, Yao T, Xu K (2021) Discriminative feature-based adaptive distribution alignment (dfada) for rotating machine fault diagnosis under variable working conditions. Appl Soft Comput 99:106886
Qin A-S, Mao H-L, Hu Q (2021) Cross-domain fault diagnosis of rolling bearing using similar features-based transfer approach. Measurement 172:108900
Qiu H, Lee J, Lin J, Yu G (2006) Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics. J Sound Vibr 289(4–5):1066–1090
Ragab M, Chen Z, Wu M, Foo CS, Kwoh CK, Yan R, Li X (2020) Contrastive adversarial domain adaptation for machine remaining useful life prediction. IEEE Trans Ind Inform 17(8):5239–5249
Ragab M, Chen Z, Wu M, Kwoh CK, Li X (2020) Adversarial transfer learning for machine remaining useful life prediction. In: 2020 IEEE International Conference on Prognostics and Health Management (ICPHM), pp. 1–7. IEEE
Renwick J, Kulkarni CS, Celaya JR (2015) Analysis of electrolytic capacitor degradation under electrical overstress for prognostic studies. In: Proceedings of the Annual Conference of the Prognostics and Health Management Society, vol. 6 (2015)
Saito K, Watanabe K, Ushiku Y, Harada T (2018) Maximum classifier discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3723–3732
Saito K, Ushiku Y, Harada T (2017) Asymmetric tri-training for unsupervised domain adaptation. In: International Conference on Machine Learning, pp. 2988–2997. PMLR
Saufi SR, Ahmad ZAB, Leong MS, Lim MH (2020) Gearbox fault diagnosis using a deep learning model with limited data sample. IEEE Trans Ind Inform 16(10):6263–6271
Saxena A, Goebel K, Simon D, Eklund N *008(Damage propagation modeling for aircraft engine run-to-failure simulation. In: 2008 International conference on prognostics and health management, pp 1–9 (2008). IEEE
Shao S, McAleer S, Yan R, Baldi P (2018) Highly accurate machine fault diagnosis using deep transfer learning. IEEE Trans Ind Inform 15(4):2446–2455
Shao H, Xia M, Han G, Zhang Y, Wan J (2020) Intelligent fault diagnosis of rotor-bearing system under varying working conditions with modified transfer convolutional neural network and thermal images. IEEE Trans Ind Inform 17(5):3488–3496
Shen F, Langari R, Yan R (2020) Transfer between multiple machine plants: a modified fast self-organizing feature map and two-order selective ensemble based fault diagnosis strategy. Measurement 151:107155
Shen C, Wang X, Wang D, Li Y, Zhu J, Gong M (2021) Dynamic joint distribution alignment network for bearing fault diagnosis under variable working conditions. IEEE Trans Instrum Meas 70:1–13
Shen J, Qu Y, Zhang W, Yu Y (2018) Wasserstein distance guided representation learning for domain adaptation. In: Thirty-Second AAAI conference on artificial intelligence
Shen F, Chen C, Yan R, Gao RX (2015) Bearing fault diagnosis based on svd feature extraction and transfer learning classification. In: 2015 Prognostics and System Health Management Conference (PHM), pp. 1–6. IEEE
Shi X, Kang Q, An J, Zhou M (2021) Novel L1 Regularized Extreme Learning Machine for Soft-Sensing of an Industrial Process. IEEE Trans Industr Inform 18(2):1009–1017. https://doi.org/10.1109/TII.2021.3065377
Si J, Shi H, Chen J, Zheng C (2021) Unsupervised deep transfer learning with moment matching: a new intelligent fault diagnosis approach for bearings. Measurement 172:108827
Silva L, Magaia N, Sousa B, Kobusińska A, Casimiro A, Mavromoustakis CX, Mastorakis G, De Albuquerque VHC (2021) Computing paradigms in emerging vehicular environments: a review. IEEE/CAA J Autom Sin 8(3):491–511

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
Sloukia F, El Aroussi M, Medromi H, Wahbi M (2013) Bearings prognostic using mixture of gaussians hidden markov model and support vector machine. In: 2013 ACS international conference on computer systems and applications (AICCSA), pp. 1–4. IEEE
Society For Machinery Failure Prevention Technology, MFPT Dataset. https://www.mfpt.org/fault-data-sets/
Southeast University, Gearbox Dataset. http://mlmechanics.ics.uci.edu/
Sun C, Ma M, Zhao Z, Tian S, Yan R, Chen X (2018) Deep transfer learning based on sparse autoencoder for remaining useful life prediction of tool in manufacturing. IEEE Trans Ind Inform 15(4):2416–2425
Sun B, Saenko K (2015) Subspace distribution alignment for unsupervised domain adaptation. BMVC 4:1–24
Sun B, Saenko K (2016) Deep coral: correlation alignment for deep domain adaptation. In: European conference on computer vision, pp 443–450. Springer
Sun B, Feng J, Saenko K (2017) Correlation alignment for unsupervised domain adaptation. In: Domain adaptation in computer vision applications, pp 153–171. Springer, Cham
Sun C, Yin H, Li Y, Chai Y (2021) A Novel Rolling Bearing Vibration Impulsive Signals Detection Approach Based on Dictionary Learning. in IEEE/CAA J Autom Sin 8(6): 1188–1198
Sutrisno E, Oh H, Vasan ASS, Pecht M (2012) Estimation of remaining useful life of ball bearings using data driven methodologies. In: 2012 IEEE Conference on Prognostics and Health Management, pp. 1–7 (2012). IEEE
Tzeng E, Hoffman J, Darrell T, Saenko K (2015) Simultaneous deep transfer across domains and tasks. In: Proceedings of the IEEE international conference on computer vision, pp 4068–4076
Tzeng E, Hoffman J, Zhang N, Saenko K, Darrell T (2014) Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474
Tzeng E, Hoffman J, Saenko K, Darrell T (2017) Adversarial discriminative domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7167–7176
Wang J, Chen Y, Feng W, Yu H, Huang M, Yang Q (2020) Transfer learning with dynamic distribution adaptation. ACM Trans Intell Syst Technol (TIST) 11(1):1–25
Wang X, He H, Li L (2019) A hierarchical deep domain adaptation approach for fault diagnosis of power plant thermal system. IEEE Trans Ind Inform 15(9):5139–5148
Wang B, Lei Y, Yan T, Li N, Guo L (2020) Recurrent convolutional neural network: a new framework for remaining useful life prediction of machinery. Neurocomputing 379:117–129
Wang X, Shen C, Xia M, Wang D, Zhu J, Zhu Z (2020) Multi-scale deep intra-class transfer learning for bearing fault diagnosis. Reliab Eng Syst Saf 202:107050
Wang X, Wanga T, Ming A, Zhang W, Li A, Chu F (2021) Cross-operating-condition degradation knowledge learning for remaining useful life estimation of bearings. IEEE Trans Instrum Measure 70:1–11
Wang C, Xin C, Xu Z (2021) A novel deep metric learning model for imbalanced fault diagnosis and toward open-set classification. Knowl-Based Syst 220:1106925
Wang F, Xu T, Tang T, Zhou M, Wang H (2016) Bilevel feature extraction-based text mining for fault diagnosis of railway systems. IEEE Trans Intell Trans Syst 18(1):49–58
Wang J, Chen Y, Hao S, Feng W, Shen Z (2017) Balanced distribution adaptation for transfer learning. In: 2017 IEEE international conference on data mining (ICDM), pp. 1129–1134. IEEE
Wang J, Feng W, Chen Y, Yu H, Huang M, Yu PS (2018) Visual domain adaptation with manifold embedded distribution alignment. In: Proceedings of the 26th ACM international conference on multimedia, pp 402–410
Wang J, Xie J, Zhang L, Duan L (2016) A factor analysis based transfer learning method for gearbox diagnosis under various operating conditions. In: 2016 International Symposium on Flexible Automation (ISFA), pp. 81–86. IEEE
Wen L, Gao L, Li X (2017) A new deep transfer learning based on sparse auto-encoder for fault diagnosis. IEEE Trans Syst Man Cybernet 49(1):136–144
Wen L, Li X, Gao L, Zhang Y (2017) A new convolutional neural network-based data-driven fault diagnosis method. IEEE Trans Ind Electron 65(7):5990–5998
Wu Z, Jiang H, Lu T, Zhao K (2020) A deep transfer maximum classifier discrepancy method for rolling bearing fault diagnosis under few labeled data. Knowl-Based Syst 196:105814
Wu Z, Jiang H, Zhao K, Li X (2020) An adaptive deep transfer learning method for bearing fault diagnosis. Measurement 151:107227
Xia P, Huang Y, Li P, Liu C, Shi L (2021) Fault knowledge transfer assisted ensemble method for remaining useful life prediction. IEEE Trans Ind Inform 18(3):1758–1769
Xia M, Li T, Shu T, Wan J, De Silva CW, Wang Z (2018) A two-stage approach for the remaining useful life prediction of bearings using deep neural networks. IEEE Trans Ind Inform 15(6):3703–3711

Yang B, Lee C-G, Lei Y, Li N, Lu N (2021) Deep partial transfer learning network: a method to selectively transfer diagnostic knowledge across related machines. Mech Syst Signal Process 156:107618
Yang B, Lei Y, Jia F, Xing S (2019) An intelligent fault diagnosis approach based on transfer learning from laboratory bearings to locomotive bearings. Mech Syst Signal Process 122:692–706
Yang N, Zheng Z, Zhou M, Guo X, Qi L, Wang T (2021) A Domain-Guided Noise-Optimization-Based Inversion Method for Facial Image Manipulation. IEEE Trans. on Image Processing 30:6198–6211
Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural networks? arXiv preprint arXiv:1411.1792
Yu C, Wang J, Chen Y, Huang M (2019) Transfer learning with dynamic adversarial adaptation network. In: 2019 IEEE international conference on data mining (ICDM), pp 778–786. IEEE
Yu S, Wu Z, Zhu X, Pecht M (2019) A domain adaptive convolutional lstm model for prognostic remaining useful life estimation under variant conditions. In: 2019 Prognostics and System Health Management Conference (PHM-Paris), pp. 130–137. IEEE
Yuan H, Zhou M (2020) Profit-maximized collaborative computation offloading and resource allocation in distributed cloud and edge computing systems. IEEE Trans Autom Scid Eng 18(3):1277–1287
Zhang Z, Chen H, Li S, An Z (2020) Unsupervised domain adaptation via enhanced transfer joint matching for bearing fault diagnosis. Measurement 165:108071
Zhang Z, Chen H, Li S, An Z, Wang J (2020) A novel geodesic flow kernel based domain adaptation approach for intelligent fault diagnosis under varying working condition. Neurocomputing 376:54–64
Zhang L, Guo L, Gao H, Dong D, Fu G, Hong X (2020) Instance-based ensemble deep transfer learning network: A new intelligent degradation recognition method and its application on ball screw. Mech Syst Signal Process 140:106681
Zhang W, Li X, Jia X-D, Ma H, Luo Z, Li X (2020) Machinery fault diagnosis with imbalanced data using deep generative adversarial networks. Measurement 152:107377
Zhang W, Li X, Ma H, Luo Z, Li X (2021) Transfer learning using deep representation regularization in remaining useful life prediction across operating conditions. Reliab Eng Syst Saf 211:1075560
Zhang R, Tao H, Wu L, Guan Y (2017) Transfer learning with neural networks for bearing fault diagnosis in changing working conditions. IEEE Access 5:14347–14357
Zhang S, Zhang S, Wang B, Habetler TG (2020) Deep learning algorithms for bearing fault diagnostics-a comprehensive review. IEEE Access 8:29857–29881
Zhao K, Jiang H, Li X, Wang R (2021) Ensemble adaptive convolutional neural networks with parameter transfer for rotating machinery fault diagnosis. Int J Mach Learn Cybernet 12(5):1483–1499
Zhao K, Jiang H, Wang K, Pei Z (2021) Joint distribution adaptation network with adversarial learning for rolling bearing fault diagnosis. Knowl-Based Syst 222:106974
Zhao M, Jiao J, Lin J (2018) A data-driven monitoring scheme for rotating machinery via self-comparison approach. IEEE Trans Ind Inform 15(4):2435–2445
Zhao R, Yan R, Chen Z, Mao K, Wang P, Gao RX (2019) Deep learning and its applications to machine health monitoring. Mech Syst Signal Process 115:213–237
Zhao B, Zhang X, Zhan Z, Pang S (2020) Deep multi-scale convolutional transfer learning network: a novel method for intelligent fault diagnosis of rolling bearings under variable working conditions and domains. Neurocomputing 407:24–38
He Z, Shao H, Jing L, Cheng J, Yang Y (2020) Transfer fault diagnosis of bearing installed in different machines using enhanced deep auto-encoder. Measurement 152:107393
Zhong K, Han M, Han B (2019) Data-driven based fault prognosis for industrial systems: a concise overview. IEEE/CAA J Autom Sin 7(2):330–345
Zhu J, Chen N, Shen C (2019) A new deep transfer learning method for bearing fault diagnosis under different working conditions. IEEE Sens J 20(15):8394–8402
Zhu J, Chen N, Shen C (2020) A new data-driven transferable remaining useful life prediction approach for bearing under different working conditions. Mech Syst Signal Process 139:106602

Siya Yao received her B.S. degree in Automation, from Donghua University, Shanghai, China in 2017. She is currently pursuing a Ph.D. degree in Control Science and Engineering with the Department of Control Science and Engineering, Tongji University, Shanghai, China. From 2019 to 2021, she worked as a joint Ph.D. Student with the Department of ECE, New Jersey Institute of Technology, Newark, NJ, USA. She

received a scholarship from the China Scholarship Council. Her research interests are in transfer learning and anomaly detection.

Qi Kang received his B.S. degree in Automatic Control, M.S. and Ph.D. degrees in Control Theory and Control Engineering, from Tongji University, Shanghai, China, in 2002, 2005, and 2009, respectively. From 2007 to 2008, he was a Research Associate with University of Illinois, Chicago, IL, USA. From 2014 to 2015, he was a visiting scholar with New Jersey Institute of Technology, NJ, USA. He is currently a Professor with the Department of Control Science and Engineering, Tongji University, Shanghai, China, and a Professor with the Shanghai Institute of Intelligent Science and Technology, Tongji University, Shanghai, China. His research interests are in swarm intelligence, evolutionary computation, machine learning, intelligent control and engineering optimization in transportation, energy and industrial systems.

MengChu Zhou received his B.S. degree in Control Engineering from Nanjing University of Science and Technology, Nanjing, China in 1983, M.S. degree in Automatic Control from Beijing Institute of Technology, Beijing, China in 1986, and Ph. D. degree in Computer and Systems Engineering from Rensselaer Polytechnic Institute, Troy, NY in 1990. He joined New Jersey Institute of Technology, Newark, NJ in 1990, and is now a Distinguished Professor of Electrical and Computer Engineering. His research interests are in Petri nets, intelligent automation, Internet of Things, big data, web services, and intelligent transportation. He has over 1000 publications including 12 books, 700+ journal papers (over 600 in IEEE transactions), 31 patents and 30 book-chapters. He is a Fellow of IEEE, IFAC, AAAS, CAA, and NAI.

Muhyaddin J. Rawa received a PhD in electrical and electronic engineering from the University of Nottingham in 2014. He has more than 7 years experience in Saudi Electricity Company and currently he is an associate professor with the Department of Electrical and Computer Engineering at King Abdulaziz University, where he is the deputy director of Center of Research Excellence in Renewable Energy and Power Systems. He is actively involved in industrial consultancy for major corporations in power systems projects. His research interest includes power quality, renewable energy and smart grids. He is a member of IEEE.

Abdullah Abusorrah is a Professor in the Department of Electrical and Computer Engineering at King Abdulaziz University. He is the head of the Center for Renewable Energy and Power Systems at King Abdulaziz University. His field of interest includes renewable energy, smart grid and system analysis. He received his PhD degree in Electrical Engineering from the University of Nottingham, UK, in 2007. He is a senior member of IEEE.

Authors and Affiliations

Siya Yao[1,2] · Qi Kang[1,2] · MengChu Zhou[3] · Muhyaddin J. Rawa[4,5] · Abdullah Abusorrah[4,5]

Siya Yao
yaosiya@tongji.edu.cn

Muhyaddin J. Rawa
mrawa@kau.edu.sa

Abdullah Abusorrah
aabusorrah@kau.edu.sa

1 Department of Control Science and Engineering, Tongji University, Shanghai 201804, China

2 Shanghai Institute of Intelligent Science and Technology, Tongji University, Shanghai 200092, China

3 Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ 07102, USA

4 Department of Electrical and Computer Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia

5 Center of Research Excellence in Renewable Energy and Power Systems, King Abdulaziz University, Jeddah 21589, Saudi Arabia