Computers in Industry 127 (2021) 103399

Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.elsevier.com/locate/compind

A double-layer attention based adversarial network for partial transfer learning in machinery fault diagnosis

Yafei Deng[a,b], Delin Huang[c], Shichang Du[a,b,*], Guilong Li[a,b], Chen Zhao[a,b], Jun Lv[d]

[a] *State Key Lab of Mechanical System and Vibration, School of Mechanical Engineering, Shanghai Jiao Tong University, No. 800 Dongchuan Road, Shanghai, 200240, China*
[b] *Department of Industrial Engineering and Management, School of Mechanical Engineering, Shanghai Jiao Tong University, No. 800 Dongchuan Road, Shanghai, 200240, China*
[c] *College of Mechanical Engineering, Donghua University, Shanghai, 201620, China*
[d] *Faculty of Economics and Management, Shanghai, 200241, China*

ARTICLE INFO

Article history:
Received 16 August 2020
Received in revised form 3 January 2021
Accepted 13 January 2021

Keywords:
Partial transfer learning
Mechanical fault diagnosis
Generative adversarial network
Domain adaptation

ABSTRACT

Recently, the deep transfer learning approaches have been widely developed for mechanical fault diagnosis issue, which could identify the health state of unlabeled data in the target domain with the help of knowledge learned from labeled data in the source domain. The tremendous success of these methods is generally based on the assumption that the label spaces across different domains are identical. However, the partial transfer scenario is more common for industrial applications, where the label spaces are not identical. This partial transfer scenario arises a more difficult problem that it is hard to know where to transfer since the shared label spaces are unavailable. To tackle this challenging problem, a double-layer attention based adversarial network (DA-GAN) is proposed in this paper. The proposed method sheds a new angle to deal with the question where to transfer by constructing two attention matrices for domains and samples. These attention matrices could guide the model to know which parts of data should be concentrated or ignored before conducting domain adaptation. Experimental results on both transfer in the identical machine (TIM) and transfer on different machines (TDM) suggest that the DA-GAN model shows great superiority on mechanical partial transfer problem.

1. Introduction

With the increase of complexity of manufacturing systems, the machine fault diagnosis serves an important role to guarantee the stability of industrial production. With the rapid development and integration of sensor techniques for modern industry, huge amount of monitoring data could be collected in engineering scenarios (Lei et al., 2018). The data-driven approach gradually shows its superiority on machine fault diagnosis, which is mainly regard to two aspects: (a) developing advanced signal processing methods to extract representative features, such as wavelet analysis (Liang et al., 2019), empirical mode decomposition (EMD) (Flandrin et al., 2004), singular value decomposition (SVD) (Liu, 2020), and (b) applying machine learning methods to seek the hidden relationship between the collected data and the health states of machines, such as artificial neural network (ANN), support vector machine (SVM) and recent deep learning approaches (Márquez et al., 2020; Li et al., 2019a, b). Among these research, deep learning method has gain great popularity regarding to its capacity of multi-layer feature learning from mechanical big data.

The key of applying deep learning models for mechanical fault diagnosis is sufficient labeled training data. However, it is unpractical to collect sufficient labeled fault data in real engineering scenarios, which can be mainly attributed to two reasons: First, the degradation of machines is usually a time-consuming process, which takes much cost to obtain sufficient data and further label them. Second, some machines may not be allowed to run to failure because the unexpected fault could lead to the break down or even catastrophic accidents (Guo et al., 2018). The aforementioned problems limit the successful development deep learning diagnostic model in real industrial fields. On the other hand, sufficient mechanical fault label data can be collected in the laboratory platform with specific working conditions. In this background, one promising idea is to promote the generalization ability of current diagnosis model, which could transfer the diagnostic knowledge

[*] Corresponding author at: State Key Lab of Mechanical System and Vibration, School of Mechanical Engineering, Shanghai Jiao Tong University, No. 800 Dongchuan Road, Shanghai, 200240, China.
E-mail address: lovbin@sjtu.edu.cn (S. Du).

https://doi.org/10.1016/j.compind.2021.103399
0166-3615/© 2021 Elsevier B.V. All rights reserved.

from the source labeled data to the target unlabeled data. For this issue, the transfer learning methods have been investigated to expand deep diagnosis model from academic research to engineering scenarios.

Generally, the transfer scenarios could be divided into two categories (Lei et al., 2020): transfer diagnosis knowledge in the identical machine (TIM) and transfer diagnosis knowledge across different machines (TDM). Aimed at these two different transfer scenarios, many deep transfer-learning based models have been proposed, which can be classified into three categories according to the transfer techniques: fine-tune approaches (Zhang et al., 2017; Cao et al., 2018a; Shao et al., 2018), feature-based approaches (Wen et al., 2017; Li et al., 2018a; Yang et al., 2019) and adversarial-based approaches (Li et al., 2018b; Zhang et al., 2020):

Despite the successful development of deep transfer learning approaches in mechanical diagnosis field, the existing approaches mainly deal with the problem of how to transfer without considering the problem of where to transfer. They carry out the transfer model based on the assumption that the source domain and target domain have the same label space. However, a more general case for real engineering applications is that the label spaces between two domains are different, which can be referred to $Y^t \subset Y^s$. This scenario can be defined as partial transfer problem, which was initially proposed in the image processing fields (Cao et al., 2018b). The partial transferring scenario would produce a more difficult challenge, in which we even do not know which part of the source domain label space Y^s is shared with the target domain label space Y^t. Moreover, the outlier source domain labeled data $Y^s \backslash Y^t$ will lead to negative transfer effect to the overall transfer performance. It is essential to select the effective part from the source domain to determine where to transfer for the target domain. For the partial transfer problem in mechanical diagnosis field, only a few researchers made exploratory work based on the adversarial-based approaches (Li et al., 2020a; Cao et al., 2018c). However, there are mainly two limitations in the existing methods:

1) The current researchers mainly focus on transferring the mechanical diagnosis knowledge in one identical machine (Lei et al., 2020), but the partial transfer problem across different machines have not been studied comprehensively.
2) To address the problem of where to transfer in the partial transfer problem, the common approach is to assign different weights for the corresponding domain discriminators. However, the effects of samples from different domains are neglected. The transferability would be severely degraded if the irrelative samples from different domains are fed into the network indiscriminately (Cao et al., 2018b).

Aiming at the above limitations, a novel double layer attention-based generative adversarial network (DA-GAN) is proposed in this paper to expand the diagnosis model for more general engineering applications. The proposed DA-GAN network consists of three modules: a feature generator, a source classifier G and a double layer attention-based discriminator D. The generator F automatically extract deep features f from both domains and the classifier G could accurately recognize the different fault types in the trained source domain. The proposed double layer attention-based discriminator D deal with the problem of where to transfer in aspects of both selecting effective domain and samples. Based on these three modules, the proposed DA-GAN approach is expected to address the partial transfer learning problem for both TIM and TDM scenarios. The main contributions are summarized as follows:

1) Different from existing deep transfer diagnosis models where either the label spaces or the mechanical components to be transferred across two domains are assumed to be same, the problem in which both the label space and mechanical components are different has been investigated in this paper. A novel deep-transfer learning model called as double layer attention-based generative adversarial network (DA-GAN) is proposed to address the partial transfer learning problem across different machines. This exploration contributes one of the first attempts to deal with this practical problem for expanding academic research to engineering applications.
2) A novel double layer attention mechanism is designed in the proposed DA-GAN model to better solve the problems of where to transfer and how to transfer. The proposed double layer attention mechanism enables the whole adversarial network to know which discriminators should be concentrated or be neglected for partial domain adaptation, as well as to decide which part of the source domain data should be shared for the target domain during each discriminator's training process.
3) Comparative experimental studies based on three different bearing datasets are investigated to evaluate the proposed method comprehensively, in which totally 42 transfer tasks across different working conditions, different machines and different types of fault characteristic are all considered.

The remainder of this paper is organized as follows: Section 2 introduces the theoretical background of transfer learning and adversarial strategy for transfer learning; Section 3 details the proposed method; Section 4 demonstrates the experiment results and discussions; Section 5 draws the conclusions.

2. Theoretical background

2.1. Background of transfer learning

The transfer learning aims at sharing reusable information across different scenarios, in which the domain and task are two basic concepts. The domain is denoted as a pair of $\mathcal{D} = \left\{X, P(X)\right\}$, including the sampled data $X = \left\{\boldsymbol{x}_i\right\}_{i=1}^{N}$ and its marginal distribution $P(X)$. The task $\mathcal{T} = \left\{Y, P\left(Y\middle|X\right)\right\}$ consists of the label space $Y = \left\{y_i\right\}_{i=1}^{N}$ and the objective prediction function $f(\cdot) = P\left(Y\middle|X\right)$.

For the mechanical fault diagnosis issue, transfer learning is applied for promoting the generality of the diagnosis model to cover the divergence of working conditions, the variation within component family type as well as the difference between machine types. The domain and task are detailed to describe the transfer learning problem in mechanical diagnosis as follow:

1) The source domain serves as the one which could provide diagnosis knowledge to other diagnosis tasks (Pan and Yang, 2009). The source domain is denoted as: $\mathcal{D}^s = \left\{X^s, P_s(X)\right\}$, where the dataset X^s contains sufficient labeled samples and follows a marginal distribution $P_s(X)$. The source task is denoted as $\mathcal{T}^s = \left\{Y^s, P\left(Y^s\middle|X^s\right)\right\}$, where the label space $Y_s = \left\{1, 2, \ldots, k\right\}$ contains different k kinds of health state and the diagnosis model could be obtained as $f_s(\cdot) = P\left(Y^s\middle|X^s\right)$, which could be learned from the labeled dataset $\left\{\boldsymbol{x}_i^s, \boldsymbol{y}_i^s\right\}_{i=1}^{N_s}$.
2) The target domain serves as the one where the diagnosis knowledge could be reused. The target domain is denoted as: $\mathcal{D}^t = \left\{X^t, P_t(X)\right\}$. If the labels in the target domain could be obtained as $\left\{\boldsymbol{x}_i^t, \boldsymbol{y}_i^t\right\}_{i=1}^{Nt}$, the transferring task from the source domain to the target domain could be attributed to inductive transfer problem. On the other hand, if there are only unlabeled samples $X^t = \left\{\boldsymbol{x}_i^t\right\}_{i=1}^{N_t}$ in the target domain, where the latent diagnosis model $f_t(\cdot) = P\left(Y^t\middle|X^t\right)$ is unavailable. In this case, the trans-

ferring task could be regard as a transductive transfer problem (Li et al., 2020b).

3) In order to guarantee the effective transfer performance from the source domain to the target domain. The label space of the source domain is expected to cover or at least equal to that of the target domain, i.e., $Y^t \subseteq Y^s \subseteq Y$ (Cao et al., 2018c). An intuitive explanation is that only when the source domain contains similar failure modes as the target domain, can it transfer reusable diagnostic knowledge to the target task.

2.2. Generative adversarial strategy for transfer learning

The GAN (generative adversarial network) based transfer learning approach develops an adversarial strategy to combine the domain adaptation and feature learning in one training process. A simple GAN model consists of two modules: a generator $G_f(\cdot)$ and a discriminator $G_d(\cdot)$. The generated feature vectors $\boldsymbol{f} = G_f\left(\boldsymbol{x}, \theta_f\right)$ is usually obtained by a multi-layer mapping function, such as SAE and CNN, where $\boldsymbol{x}$ is a series of raw sampling points and θ_f be defined as $P_f = G_d\left(\boldsymbol{f}, \theta_d\right)$, where P_f is the probability that $\boldsymbol{f}$ comes from the source domain rather than the target domain and θ_d is the discriminator parameters.

During the training process, the discriminator adjusts its parameters θ_d to maximize the probability the P_f, thus the generated feature $\boldsymbol{f}_t$ from the target domain can be distinguished from the generated $\boldsymbol{f}_s$ from the source domain. In contrast, the generator is designed to minimize the P_f to confuse the discriminator by generating fake samples with the similar distribution to the source domain feature. As the minimax two-player game continues, the GAN model is optimized to capture domain-invariant features, which can be formulated as:

$$\underset{G_f}{min}\ \underset{G_d}{max}\quad E_{\boldsymbol{x}^s \in X^s}\left[logG_d\left(\boldsymbol{f}_s\right)\right] + E_{\boldsymbol{x}^t \in X^t}\left[\log\left(1 - G_d\left(\boldsymbol{f}_t\right)\right)\right] \tag{1}$$

3. Proposed method

3.1. Problem formulation

In this paper, a partial transfer learning problem is studied for fault diagnosis of identical machines and different machines, where the learned diagnosis knowledge from the source domain is expected to be transferred to the target domain. Generally, this study is carried out under the following assumptions:

1) The fault diagnosis problems from two domains are different, specifically the health state labels from the source domain are not identical as the target domain, $Y^t \neq Y^s$.
2) For the source domain, there are sufficient labeled data, $\left\{\boldsymbol{x}_i^s, \boldsymbol{y}_i^s\right\}_{i=1}^{N_s}$ for supervising learning, but there are only unlabeled data $\left\{\boldsymbol{x}_i^t\right\}_{i=1}^{N_t}$ in the target domain, which can be attributed to the transductive transfer problem.
3) Since the fault diagnosis transfer across different machines is also investigated, the data attributes from source and target domains could be totally different, such as the difference across sample length and the variance in the sample distribution.

Since the target domain label space is assumed to be the subset of the source domain label space, $Y^t \subseteq Y^s$, the outlier data from source domain $\complement_{Y^s} Y^t = \left\{y \,\middle|\, y \in Y^s,\ y \notin Y^t\right\}$ will lead to the unnecessary negative transfer. Correspondingly the larger the outlier label space $\complement_{Y^s} Y^t$ compared to the Y^t, the worse the transferability across different domains will be. To combat the negative transfer effect caused by $\complement_{Y^s} Y^t$ and determine which part should be shared from the source domain, the DA-GAN model is proposed to achieve the partial mechanical diagnosis transfer task under the given assumptions, and the framework of DA-GAN is illustrated in Fig. 1.

Since the direct feature learning from raw noisy signal generally leads to low network training efficiency, data-preprocessing techniques including resampling, fast Fourier transformation (FFT) and frequency spectrum rescaling are applied to transform the raw mechanical vibration data. It should be noticed that data from source domain and target domain may have different sampling frequency, which would lead to the misalignment of frequency features and further degrade the model transferability. Therefore, the spectrum rescaling is applied to the source and target spectrums after fast Fourier transformation, which facilitates the frequency feature alignment through sharing the same frequency range then the aligned spectrum images will be fed into the separable CNN network and deep layer feature generator is trained to extract domain-invariant features. The extracted feature f_s from the source domain are used to train the label classifier G_y to guarantee the fault diagnosis functions. At the same time, a two-stage attention-based discriminator is conducted to deal with the problems of where to transfer and how to transfer gradually. The details of the feature generator and attention-based discriminator will be introduced in section 3.2 and 3.3.

3.2. A separable convolutional neural network

In this section, the feature generator G_f and the source label classifier G_y are built based on a separable convolutional neural network (S-CNN). As a powerful tool to achieve deep feature extraction and classification, a variety of CNN based models have been proposed recently, and the model performance is optimized by changing the networks connecting architecture as well as introducing the heterogeneous convolution kernels. However, these stacked deep models are usually designed for huge-scale classification tasks, such as face recognition with millions of samples. Considering the characteristics of mechanical monitoring data, we introduce the S-CNN model with less parameters to replace the traditional CNN model.

The traditional convolutional neural network learns features from all three dimensions of the input images, which include the width, the height and the channels. Therefore, the kernel in CNN is expected to characterize the spatial relationships and the cross-channel correlations synchronously. The process of CNN is demonstrated in Fig. 2 (a), and the general formulation of convolutional kernel is given as:

$$\mathrm{Conv}(W, x)_{(i,j)} = \sum\nolimits_{m,n,k}^{M,N,K} W_{(m,n,k)} \cdot x_{(i+m,j+n,k)} \tag{2}$$

where W is the parameters of the kernel weight matrix to be trained. The input is denoted as x and the (i, j) denotes the coordination of the output feature. m, n, k indicate the width, the height and the channel number of the convolutional kernel respectively.

Different from the traditional CNN model, the introduced separable convolution neural network characterizes the spatial relationships and the cross-channel correlations independently, including two simpler steps: a depthwise convolution and a pointwise convolution, which is demonstrated in Fig. 2 (b).

In the first step, the convolutional kernel is applied to extract each channel information of the input data. After the operation of depthwise convolution, the number of the channels does not change. The depthwise convolution is expressed as:

$$\mathrm{DW} - \mathrm{Conv}(W, x)_{(i,j)} = \sum\nolimits_{m,n}^{M,N} W_{(m,n)} \cdot x_{(i+m,j+n)} \tag{3}$$

In the second step, a pointwise operation with a 1 × 1 convolutional kernel is developed to concatenate the outputs from the depthwise convolution, which is described in Eq. (4). The point-

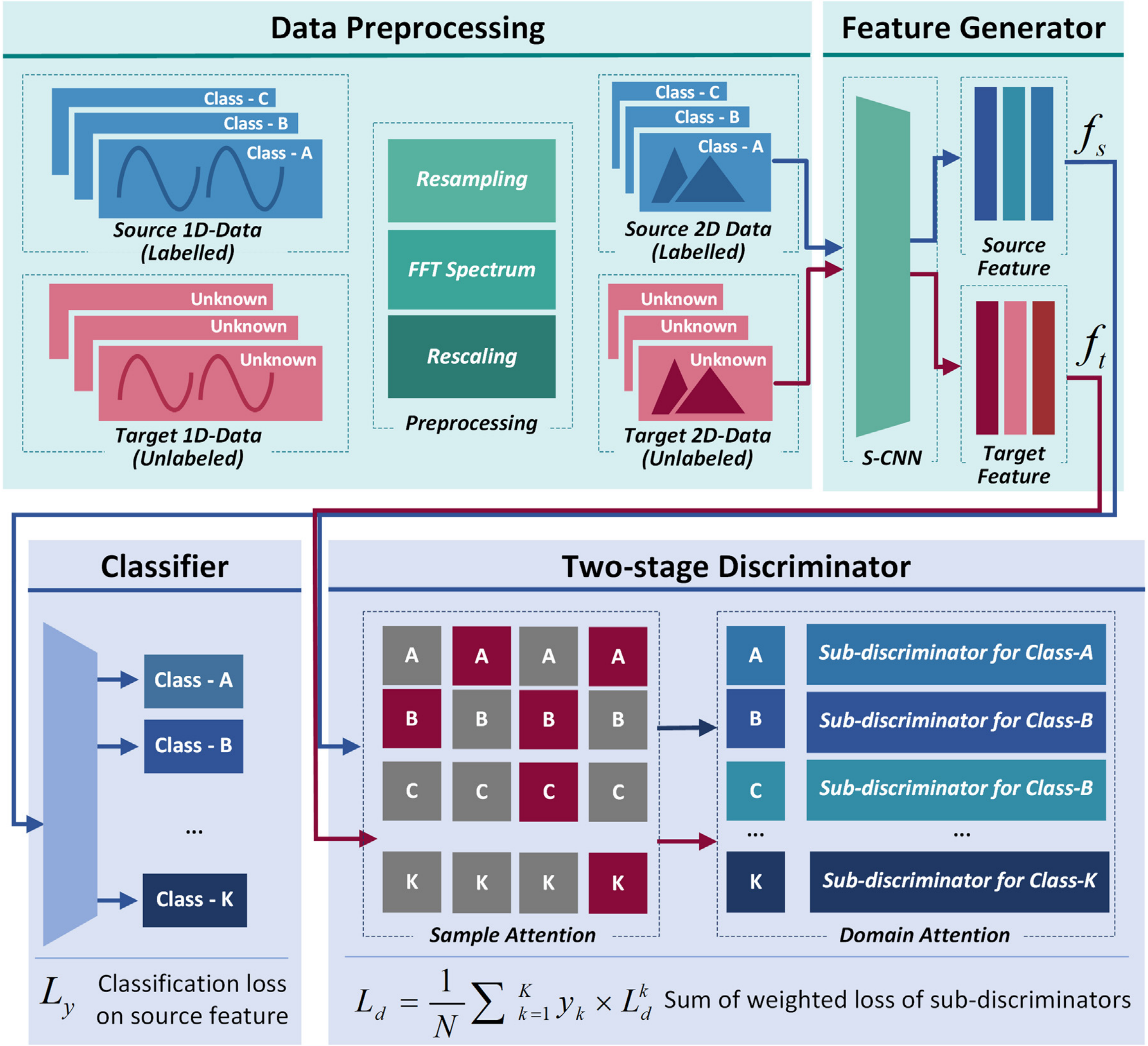

Data Preprocessing
Class - C
Class - B
Class - A
Source 1D-Data (Labelled)
Unknown
Unknown
Unknown
Target 1D-Data (Unlabeled)
Resampling
FFT Spectrum
Rescaling
Preprocessing
Class - C
Class - B
Class - A
Source 2D Data (Labelled)
Unknown
Unknown
Unknown
Target 2D-Data (Unlabeled)
Feature Generator
S-CNN
Source Feature
Target Feature
f_s
f_t
Classifier
Class - A
Class - B
...
Class - K
L_y Classification loss on source feature
Two-stage Discriminator
A A A A
B B B B
C C C C
K K K K
Sample Attention
A
B
C
...
K
Sub-discriminator for Class-A
Sub-discriminator for Class-B
Sub-discriminator for Class-B
...
Sub-discriminator for Class-K
Domain Attention
$L_d = \frac{1}{N}\sum_{k=1}^{K} y_k \times L_d^k$ Sum of weighted loss of sub-discriminators

Fig. 1. Framework of proposed DA-GAN model.

wise convolution is applied to extract spatial information, which would not change the spatial size but change the channel size.

$$\mathrm{PW}-\mathrm{Conv}(W,x)_{(i,j)} = \sum\nolimits_{k}^{K} W_k \cdot x_{(i,j)} \tag{4}$$

Combined with the Eq. (3) and Eq. (4), the overall expression of the separable convolution can be expressed as:

$$\mathrm{S}-\mathrm{Conv}(W_D, W_P, x)_{(i,j)} = \mathrm{PW}-\mathrm{Conv}(W,x)_{(i,j)} \left[W_P, \mathrm{DW}-\mathrm{Conv}(W,x)_{(i,j)}\right] \tag{5}$$

To quantitively evaluate the difference between the CNN and the separable CNN, the model parameters are calculated. W, H, C indicate the width, the height and the channel number of the convolutional kernel respectively, and the kernel number is denoted as K. The ratio of total required parameters is given in Eq. (6). It can be seen that the separable CNN could effectively reduce the model complexity since the entire feature extraction is divided into two simpler steps independently.

$$\frac{P_{S-CNN}}{P_{CNN}} = \frac{W \times H \times K + C \times K}{W \times H \times K \times C} = \frac{1}{W \times H} + \frac{1}{C} \tag{6}$$

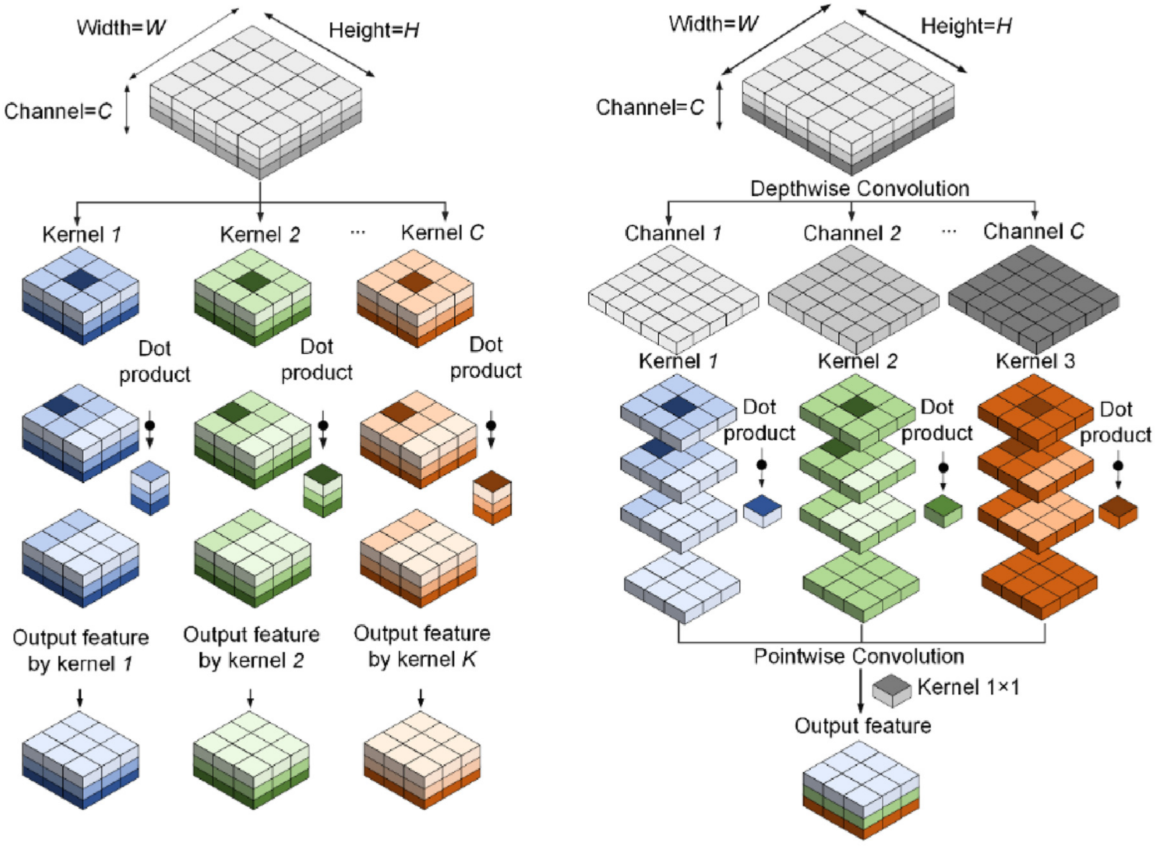
The architecture of proposed feature generator is shown in Fig. 3, which mainly consists of the separable convolutional layer, pooling layer and residual connection layer. By replacing the CNN with separable CNN, the calculation complexity is greatly reduced without sacrificing prediction accuracy. The pooling layer can quickly decrease the dimension of extracted features, which could reduce the layers needed in the model and introduce some nonlinear changes. The residual connection is designed to avoid the gradient degradation during training and information loss, which could promote the feature extraction from different levels.

The detailed parameters of proposed feature generator and subdomain discriminator are given in Tables 1a 1b. After the operation of global average pooling P4, a fully connected layer is conducted to flatten the outputs and to map them into features f, which can be formulated as:

$$f = \sigma\left[\left(w_f\right)^T x + b_f\right] \tag{7}$$

where w_f indicates the weight matrix and b_f is the corresponding bias vector, and x is the input vector from the above pooling layer.

After building the feature generator G_f, the source label classifier could be established subsequently. The classifier G_y is a simple



(a) Traditional CNN

(b) Depthwise Separable CNN

Fig. 2. Architecture of CNN and separable CNN.

Table 1a
Parameters of proposed feature generator.

Layer	Symbol	Operator	Parameter
1	Input	Input data	3 × 64 × 64
2	SC1	Separable convolution2D	Channel number:128, kernel size:3
3	LR1	LeakyReLU	alpha=0.3
4	D1	Dropout	$p = 0.25$
5	SC2	Separable convolution2D	Channel number:64, kernel size:3
6	LR2	LeakyReLU	alpha=0.3
7	D2	Dropout	$p = 0.25$
8	SC3	Separable convolution2D	Channel number:32, kernel size:3
9	LR3	LeakyReLU	alpha=0.3
10	D3	Dropout	$p = 0.25$
11	R1	Residual connection	/
12	FC	Fully connection	Dense number:128

softmax regression based on the output from the fully connection layer, which can be expressed as:

$$y = \frac{1}{\sum_{i=1}^{K} e^{\left[\left(w_y^i\right)^T f + b_y^i\right]}} \begin{bmatrix} e^{\left[\left(w_y^1\right)^T f + b_y^1\right]} \\ e^{\left[\left(w_y^2\right)^T f + b_y^2\right]} \\ \vdots \\ e^{\left[\left(w_y^K\right)^T f + b_y^K\right]} \end{bmatrix} \tag{8}$$

where w_y^i is the weight matrix connecting to the ith output neuron, b_y^i is the corresponding bias vector, and K denotes the different kinds of fault modes or health conditions.

3.3. Double layer Attention-based domain discriminator

The double layer attention-based domain discriminator mainly contains two parts: sub-domain discriminators and double layer attention matrices. Firstly, the original domain discriminator G_d is split into a series of sub-domain discriminators G_d^k, and each of them is responsible for matching the source and target domain data corresponded to the source domain label $\left\{y^k \middle| y^k \in Y^s\right\}$. Compared with the traditional domain discriminator, which conducts

Table 1b
Parameters of proposed sub-domain discriminator.

Layer	Symbol	Operator	Parameter
1	Input	Input data	128 × 1
2	C1	Convolution1D	Channel number:32, kernel size:3
3	LR1	LeakyReLU	alpha=0.3
4	D1	Dropout	$p = 0.25$
5	C2	Convolution1D	Channel number:16, kernel size:3
6	LR2	LeakyReLU	alpha=0.3
7	D2	Dropout	$p = 0.25$
8	FC1	Fully connection	Dense number:512
9	LR3	LeakyReLU	alpha=0.3
10	FC2	Fully connection	Dense number:128
12	LR4	LeakyReLU	alpha=0.3
13	FC3	Fully connection	Dense number:2
14	S1	Softmax	/

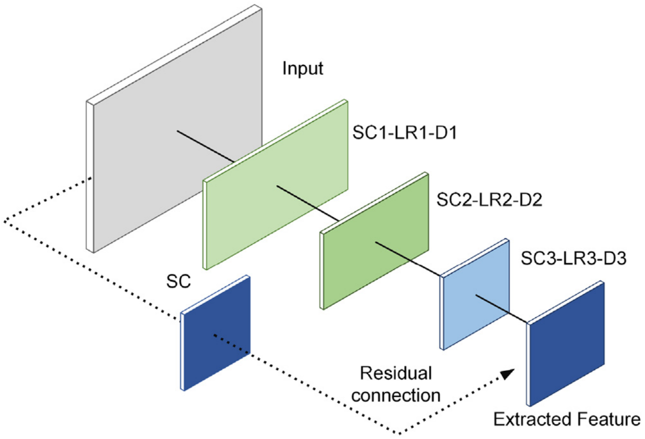


Fig. 3. Architecture of the proposed feature generator.

the domain adaptation considering the whole distribution of X^s and X^t, the sub-domain discriminators could achieve better flexibility when the label spaces across two domains are different. The negative transfer effect caused by unbalanced label space $\complement_{Y^s} Y^t$ could be suppressed if the sub-domain discriminators corresponded to the outlier label space could be correctly identified. Therefore, the double layer attention mechanism is employed subsequently as the transfer indicators for sub-domain discriminators. The double layer attention mechanism is constructed based on two matrices defined as domain attention matrix M_d and sample attention matrix M_s, which are introduced as follows:

$$M_d = [y_k]\,, k = 1, 2, \cdots, C_s$$
$$M_s = \begin{bmatrix} s_1^i \\ \vdots \\ s_k^i \end{bmatrix}, i = 1, 2, \cdots, N_{s+t} \tag{9}$$

where the number of label space in the source domain is denoted as C_s, and the number of total source and target domain samples is denoted as N_{s+t}.

The first layer is designed to determine which sub-domain discriminators should be activated for the current transfer task. Since the label space of target domain is unknown during the training process, it is hard to know which label spaces should be shared across the source domain and the target domain. Correspondingly, the domain attention matrix M_d is designed to assign different weights y_k to each sub-domain discriminator. More attention is expected on those discriminators sharing the same label space, as well as less attention is laid on the discriminators responsible for the outlier label space.

It should be noticed that only assigning domain attentions y_k to discriminators could not guarantee the whole transfer performance of the diagnosis model. Because there would be no guidance for these weighted sub-domain discriminators to decide which part of the samples should be exploited as training data for each domain adaptation process. If samples from different domains are fed into these discriminators indiscriminately for each sub-domain adaptation, there would lead to a problem that the sub-domain discriminator would learn the wrong pattern according to the outlier source sample although this sub-discriminator belongs to the shared label spaces. Therefore, the second layer M_s is designed to generate attentions s_k^i of each sample for sub-domain discriminators. The sample attention matrix is expected to ensure each data point could be only aligned to one or several most relevant classes with high attention value s_k^i, while the irrelevant classes with low s_k^i would be filtered out.

The comparisons across transfer without attention mechanism, transfer with only domain attention mechanism and double-layer mechanism is demonstrated in Fig. 4. It could be seen that the key step of successful implement of partial transfer based on double layer attention mechanism is to select reasonable metrics to construct y_k and s_i^k, which could judge whether the unknown target domain data share the same label space with the source domain. As mentioned above, the index MMD has been applied in many feature-based transfer approaches attributed to its superior capacity of characterizing distributions similarity. Therefore, in this paper the index MMD is exploited as the metrics to construct the double layer attention matrices to assign different weights. The detailed calculation of *MMD* is described as follows:

$$MMD_{\mathcal{H}}(X, Y) := \sup_{\Phi \,\in\, \mathcal{H}} \left\{ E_{X\sim p}[\Phi(x)] - E_{Y\sim p}[\Phi(y)] \right\} \tag{10}$$

Where $sup\{\cdot\}$ is the supremum of the input aggregate, $\mathcal{H}$ indicates the reproduced kernel Hilbert space (RKHS), and $\Phi(x)$ is a nonlinear mapping function from the original space to RKHS. The nonlinear mapping function $\Phi(x)$ in RKHS is assumed to be rich enough to obtain an appropriate mode which could maximize the distance between the source data and the target data. If the value of *MMD* is small, it can be concluded that these two samples follow similar distributions, which indicates that the target data may share the same fault type as the sub-domain data.

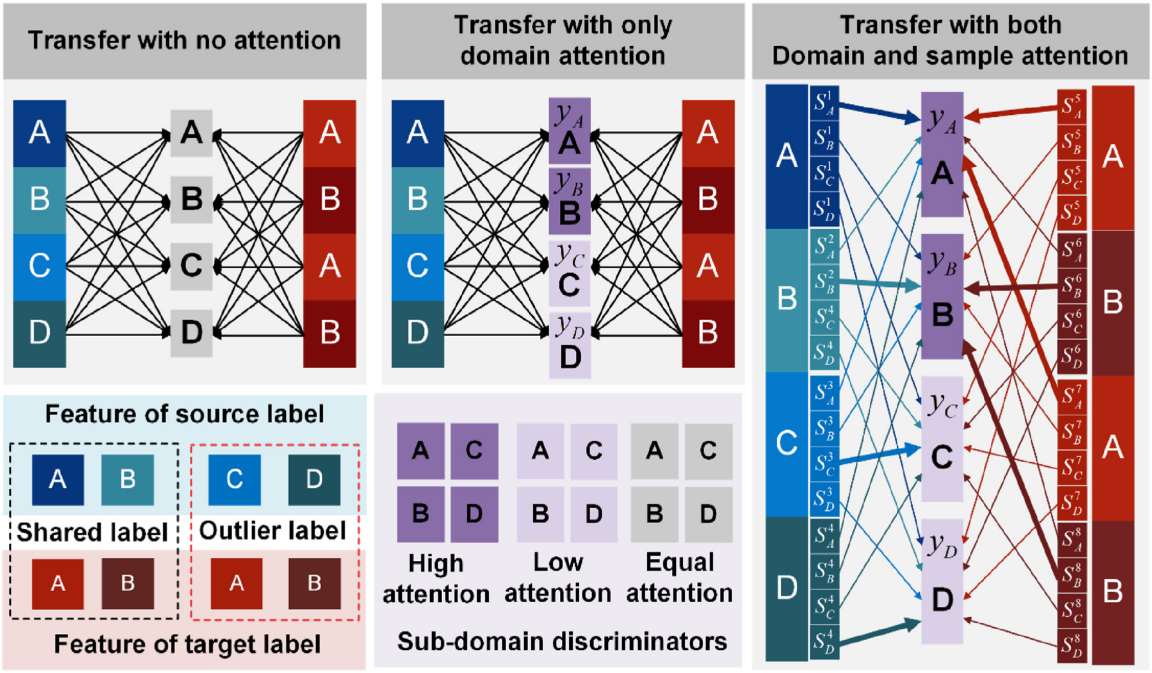

Transfer with no attention
Transfer with only domain attention
Transfer with both Domain and sample attention
A
B
C
D
y_A
y_B
y_C
y_D
S_A^1 S_B^1 S_C^1 S_D^1
S_A^2 S_B^2 S_C^2 S_D^2
S_A^3 S_B^3 S_C^3 S_D^3
S_A^4 S_B^4 S_C^4 S_D^4
S_A^5 S_B^5 S_C^5 S_D^5
S_A^6 S_B^6 S_C^6 S_D^6
S_A^7 S_B^7 S_C^7 S_D^7
S_A^8 S_B^8 S_C^8 S_D^8
Feature of source label
Shared label
Outlier label
Feature of target label
High attention
Low attention
Equal attention
Sub-domain discriminators

Fig. 4. Domain adaptation based on discriminators with different attention mechanisms.

Based on the MMD metrics, the first layer attention $M_d = [y_k]$ can be formulated as:

$$d_i^k = \begin{cases} d_i^{sk} : mean \left\{ \sum_{j=1}^{s_k} MMD_{\mathcal{H}} \left[G_f\left(x_j^{s_k}\right), G_f\left(x_i^s\right) \right] \right\} \\ d_i^{tk} : mean \left\{ \sum_{j=1}^{s_k} MMD_{\mathcal{H}} \left[G_f\left(x_j^{s_k}\right), G_f\left(x_i^t\right) \right] \right\} \end{cases}$$
$$y_k = \frac{1/mean\left(\sum_{i=1}^{N_t} d_i^{tk}\right)}{\sum_{k=1}^{C_s}\left[1/mean\left(\sum_{i=1}^{N_t} d_i^{tk}\right)\right]} \tag{11}$$

where the feature generator based on the separable CNN is denoted as $G_f(\cdot)$. $x_j^{s_k}$ refers to the jth source domain data of kth fault type, and total sample number of source domain in kth class is s_k. The distance d_i^k is calculated to measure the similarity between the distribution of ith target domain data (source domain data) and the whole distribution of source domain data from the kth label space. Correspondingly, the larger distance d_i^k is, the lower probability that the data x_i^t belongs to the label space k will be. After obtaining d_i^k, the domain attention y_k could be derived subsequently by evaluating the similarity with the label space k from whole target domain. Note that the probability y_i^k represents a negative correlation with the MMD distance metrics, the reciprocal operation is applied in Eq. (11).

Similarly, the second layer attention could be obtained as follows:

$$s_i^k = \begin{cases} s_i^{sk} : \dfrac{1/d_i^{sk}}{\sum_{k=1}^{C_s} 1/d_i^{sk}} \\ s_i^{tk} : \dfrac{1/d_i^{tk}}{\sum_{k=1}^{C_s} 1/d_i^{tk}} \end{cases} \tag{12}$$

After the calculation of attention matrices M_d and M_s, the traditional optimization function described in Eq. (1) is transformed to a double-layer weighted loss function as follows:

$$\underset{G_f}{min}\ \underset{G_d}{max}\ \sum_{k=1}^{C_s}\left[\sum_{i=1}^{N_s}\left[log\left[y_k \cdot G_d^k\left(s_i^{sk} \times f_i^s\right)\right]\right] + \sum_{i=1}^{N_t}\left[\log\left[1 - y_k \cdot G_d^k\left(s_i^{tk} \times f_i^t\right)\right]\right]\right]$$

Compared with the single domain discriminator in GAN, the proposed weighted loss function enables attention-based adaptation where the target data is only focused on those relevant sub-domain discriminators according to the probabilities y_i^k and s_i^k. The proposed double layer attention-based domain discriminator has three main advantages:

1) The proposed multiple sub-domain discriminators provide a soft and flexible transfer mechanism compared with the hard assignment of all source and target data to only one discriminator. The formulation of multiple discriminators with different parameters θ_d^k could enhance the learning performance in each sub-domain.
2) The designed domain attention layer enables the model to know which label spaces are shared and which sub-domain discriminators should be emphasized. The sub-domain discriminators with different weights could suppress the negative effect from outlier label space, which provides guidance for the model to know where to transfer.
3) The designed sample attention layer further explores the problem of where to transfer in the aspect of better using the source and target domain samples. The target data with unknown labels is only aligned to one or several most relevant fault classes with high value of attentions s_i^k, which could avoid introducing the redundant information and promote the positive transfer performance for each sub-domain discriminator.

3.4. Overall objective function and training strategy

3.4.1. Optimization objective

The proposed DA-GAN model consists of two optimization objects:

1) Minimization of the fault identification error L_c on the source domain data.
2) Minimization of domain adaptation loss L_d with respect to the feature generator G_f and maximization of L_d with respect to the double layer attention-based discriminator G_d;

Object 1: To achieve the effective diagnosis transferability, the extracted domain-share feature is expected to discriminate different mechanical fault types. Since the target label is not available during training, the source domain dataset $\mathcal{D}^s = \left\{ \boldsymbol{x}_i^s, \boldsymbol{y}_i^s \right\}_{i=1}^{N}$ is developed to minimize the fault classification error. The classification loss could be formulated as:

$$L_c = \frac{1}{n_s} \sum_{f_i \in D_s} L_y \left[G_y \left(f_i, y_i \right) \right] \tag{14}$$

where G_y is the fault classifier proposed in section 3.1, f_i is the generated features with proposed separable CNN network, and L_y denotes the cross-entropy loss.

Object 2: The domain adaptation module is accomplished through a minimax adversarial strategy. During the competitive training process, the discriminator G_d is trained to distinguish features from source and target domains by maximize the domain adaptation loss L_d, and at the same time, the generator G_f is expected to capture the domain-invariant features by minimize the domain adaptation loss L_d. The optimization function of domain adaptation can be rewritten as:

$$L_d = \frac{1}{N_s} \sum_{k=1}^{C_s} \times \left[\sum_{f_i^s \in D_s} y_k \times L_d^k \left(G_d^k \left(s_i^{sk} \times f_i^s, \varepsilon_i^k \right) \right) \right] + \frac{1}{N_t} \sum_{k=1}^{C_s} \times \left[\sum_{f_i^t \in D_t} y_k \times L_d^k \left(G_d^k \left(s_i^{tk} \times f_i^t, \varepsilon_i^k \right) \right) \right] \tag{15}$$
$$L_d^k \left[G_d^k \left(f_i, d_i \right) \right] = d_i log \frac{1}{G_d^k \left(f_i \right)} + \left(1 - d_i \right) \times log \frac{1}{1 - G_d^k \left(f_i \right)}$$

where ε_i^k indicates the binary variable of k th sub-domain discriminator G_d^k, f_i^s and f_i^t are extracted features from source domain and target domain respectively.

Combining all these optimization functions, the overall object can be expressed as:

$$L \left(\theta_f, \theta_y, \theta_d^k |_{k=1}^{K} \right) = L_c \left(\theta_f, \theta_y \right) - \alpha \times L_d \left(\theta_f, \theta_d \right) \tag{16}$$

where α is the hyperparameter which trade-off these objectives in the unified optimization problem.

3.4.2. Training strategy

Once the overall optimization function is built, the stochastic gradient descent (SGD) algorithm could be applied to train the proposed method, in which the parameters $\left(\theta_f, \theta_y, \theta_d \right)$ can be trained as follows:

$$\begin{aligned} \left(\hat{\theta}_f, \hat{\theta}_y \right) &= \underset{\theta_f, \theta_y}{argmin} L \left(\theta_f, \theta_y, \hat{\theta}_d^{C_s} |_{k=1}^{C_s} \right) \\ \left(\hat{\theta}_d^1, \ldots, \hat{\theta}_d^{C_s} \right) &= \underset{\theta_d^k |_{k=1}^{C_s}}{arg\,max} L \left(\hat{\theta}_f, \hat{\theta}_y, \theta_d^k |_{k=1}^{C_s} \right) \end{aligned} \tag{17}$$

It should be noticed that the DA-GAN network could not possess an explicit loss function for model training since the parameters $\left(\theta_f, \theta_d^k \right)$ are updated in the opposite direction during the adversarial stage. To achieve the flexible implement of SGD algorithm, this paper update the gradient of the generator and discriminator iteratively, in which the parameter θ_d^k will be frozen during the training process of parameter θ_f. Through this circuitous training strategy, the parameters θ_f, θ_y, θ_d can be updated with the standard backpropagation algorithm, which can be expressed as:

$$\begin{aligned} \theta_f &\longleftarrow \theta_f - \mu \left(\frac{\partial L_c}{\partial \theta_f} - \alpha \frac{\partial \overline{L_d}}{\partial \theta_f} \right) \\ \theta_y &\longleftarrow \theta_y - \mu \left(\frac{\partial L_c}{\partial \theta_y} \right) \\ \theta_d^k &\longleftarrow \theta_d^k - \mu \left(\frac{\partial L_d}{\partial \theta_d^k} \right) \end{aligned} \tag{18}$$

where μ represents the learning rate taken by the SGD algorithm during training progresses.

4. Experimental study

4.1. Dataset description

To validate the performance of DA-GAN on partial transfer for both TIM and TDM scenarios, two rolling bearing datasets are exploited in this section to build three partial transfer mechanical diagnosis experiments.

1) Dataset A: The CWRU bearing dataset (Bearings Data Center)

The CWRU bearing dataset is commonly used for mechanical fault diagnosis, in which the vibration data were measured from the motor bearings. There were totally ten kinds of health states in the monitoring data, which were generally separated as: (1) healthy (H), (2) inner race fault (IF), (3) outer race fault (OF) and (4) ball fault (BF). These three faults are further classified according to the fault size as 0.07 in., 0.14 in. and 0.21 in., respectively. The vibration signal was sampled with 12.8 kHz and each sample contained 96,678 data points.

2) Dataset B: The Paderborn University bearing dataset (Bearing DataCenter)

In the Paderborn University bearing dataset, the ball bearings fault could be divided into artificial faults and real damages caused by the accelerated lifetime test (ADT). In the artificial fault data, two kinds of common fault type inner race fault and outer race fault were recorded. But for the real damages fault data, it also contained some unusual failures, such as failure caused by plastic deformation and combined damage modes. The collected bearing fault signals consist of motor current, vibration, dynamic loading and temperature. The vibration signal is sampled with 64 kHz and each sample contained 25,600 data points.

3) Dataset C: The XJTU-SY bearing dataset (XJTU-SY Bearing Datasets)

Different from datasets A and B collecting data with obvious fault characteristics, XJTU-SY bearing dataset collects the full life cycle bearing information with different degradation modes as inner fault, outer fault and combined fault. Therefore, the dataset C not only contains similar obvious fault characteristics data as datasets A and B, but also collects early-stage life cycle data with weak fault characteristics. The vibration signal is sampled with 25.6 kHz and each sample contained 32,768 points.

The three datasets details are summarized in Table 2, and three experimental cases are designed to comprehensively evaluate the proposed method under different degrees of domain shifts.

4.2. Compared approaches

In this paper, different strategies including no transfer, feature-based transfer and adversarial-based transfer are implemented on the partial transfer learning problem for comparison study. To evaluate the proposed method fairly and comprehensively, all the following approaches would share the similar network (same structure for feature generator) and hyperparameters with DA-GAN model.

(1) Baseline

First, a baseline method is applied for comparison to show the transferring performance without transfer. The baseline model has no special designed structures for domain adaptation and partial transfer learning. The feature extractor and classifier are trained with the labeled source domain data and would directly predict the unlabeled target domain data.

(2) Feature-based model: Domain adaptation based on MMD

In the feature-based domain adaption method, the metric MMD is employed as the optimization item, training the whole network to extract domain-invariant features and to achieve better transferability. In this section, the popular methods in the existing studies multi-kernel MMD (MK-MMD) (Che et al., 2020) and multi-layer MMD (ML-MMD) (Yang et al., 2019) are applied.

(3) Adversarial-based model: Domain adaptation based on GAN

In the adversarial-based model, two kinds of GAN model, deep adversarial CNN (DACNN) (Han et al., 2019) and selective adversarial network (SAN) (Cao et al., 2018b), are employed for comparisons. In the DACNN model, the sub-domain discriminators and double-attention layer are removed, the domain adaptation is implemented by one discriminator. In the SAN model, the multiple sub-domain discriminators are also constructed but only the domain attention is calculated based on pseudo-label from the target domain.

4.3. Brief introduction of designed experimental cases

In this sub-section, a brief introduction of designed transfer scenario cases is provided. The comparisons of three experimental cases are illustrated in Fig. 5. It can be seen that the challenges among three partial transfer learning scenarios are increased progressively. In Case I, the model evaluation is focused on transferring diagnosis knowledge under different working conditions, in which the target domain data have different label space compared with source domain data. In Case II, the model evaluation is focused on the partial transferring performance across different machines, also known as TDM issues which have not been investigated comprehensively. In Case III, the partial transfer task further concerns the effect of different fault characteristics, in which the model trained with obvious fault data is expected to transfer diagnosis knowledge on weak fault characteristics data.

4.4. Case study I: partial transfer problem for TIM scenario

4.4.1. Experiment description

In this sub-section, the partial transfer learning problem is studied for TIM scenario. Different fault diagnosis knowledge transfer tasks are designed to make comprehensive comparisons between proposed DA-GAN and other deep transfer approaches. The detailed information of the concerned transfer tasks is given in Table 3, which are randomly selected from dataset A and B.

For the transfer task TIM_A, 2000 source-domain and target domain samples with 12,800 sample length at each health state are employed for model training. For the transfer task TIM_B, the sample number is 1000 and the sample length turns to 64,000 to match the sampling frequency. It should be noticed that the target domain data are unlabeled, therefore samples are shuffled to guarantee that the labeled source domain samples and the unlabeled target domain samples from the same label space would not be aligned before domain adaptation. Afterwards, totally 2000 samples from the target domain are tested, and the predicted result of each task is averaged 10 trials to reduce the randomness. The training iterations of each model is set as 500 to guarantee a convergent result, the learning rate is set as 0.01, and the batch size is set as 100. For DA-GAN, the trade-off parameter α is set as 0.5. Both mean value and standard deviations of predicted results are provided to reduce the effect of randomness.

4.4.2. Experimental results and performance comparisons

The TIM fault diagnosis results for dataset A are presented in Table 4. It can be observed that all the compared deep transfer approaches could obtain excellent transfer performance in the non-partial transfer problem (Task C_{A-1}). However, when dealing with the partial transfer problems, where the target samples have large biased label space compared with the source domain, the transfer performance of these approaches are degraded and even the negative transfer occur, such as C_{A-6}, C_{A-8} and C_{A-9}. The confusion matrices of testing accuracy on these partial transfer tasks are illustrated in Fig. 5. Based on the confusion matrix, it can be found that the testing accuracies of all the feature-based approaches (MK-MMD & ML-MMD) and adversarial-based approaches (GAN & SAN) reduce greatly when only the limited label spaces exist in the testing data. Especially in the extreme case C_{A-9}, where one health state, inner race fault with 0.14 in. fault size, is included in the target domain data, the SAN model could only reach 44.9 % accuracy on testing data, which is even much lower than the accuracy of baseline model without transfer.

To further investigate the severe degradation of transfer performance on the SAN approach, the domain attention matrices of 10 transfer tasks constructed by SAN and DA-GAN are illustrate in Fig. 6. The true domain attention is given based on the latent target domain labels, and it can be seen the domain attention for task-9 should be concentrated on label-3. However, the SAN attention on task-9 has been mainly scattered to label-2, label-3 and label-4, in which samples from irrelevant label spaces produce negative effect on domain adaptation process. Since the domain attention in the SAN approach is built on the pseudo label trained from source data, the large bias between two domains could lead to a wrong pseudo label distributions and interference the transferability correspondingly.

On the other hand, the proposed DA-GAN model could construct the domain attention correctly compared with SAN. Benefit from better representation of domain similarity based on MMD-metrics instead of depending on pseudo labels, DA-GAN model shows great superiority on partial transfer learning problems, especially in the cases where the source domain and target domain has large-biased label space. It should be noticed in the engineering machinery diagnosis scenario, the unlabeled testing data for one machine usually

Table 2
The detailed information of three experimental datasets.

Name		Dataset Specification		
Dataset A: CWRU dataset	Bearing type	Ball bearing, SKF 6205−2RS JEM		
	Working condition	Load : 0Hp / 1Hp / 2Hp / 3Hp	Speed: 1720 rpm to 1797 rpm	
	Health state label & Health state specification	1		Health
		2		Inner race fault with 0.07 in. fault size
		3		Inner race fault with 0.14 in. fault size
		4		Inner race fault with 0.21 in. fault size
		5		Ball fault with 0.07 in. fault size
		6		Ball fault with 0.14 in. fault size
		7		Ball fault with 0.21 in. fault size
		8		Outer race fault with 0.07 in. fault size
		9		Outer race fault with 0.14 in. fault size
		10		Outer race fault with 0.21 in. fault size
	Fault mode	Artificial damage		
	Data type	Fault data sampled with 12,800 Hz		
Dataset B: Paderborn dataset	Bearing type	Ball bearing, SKF 6203		
	Working condition	Load :400 N / 1000N	Speed: 900 rpm/1500 rpm	Torque: 0.1 Nm / 0.7 Nm
	Health state label & Health state specification	1a		Inner race fault (Artificial damage)
		2a		Health
		3a		Outer race fault (Artificial damage)
		1b		Inner race fault (Fatigue pitting)
		2b		Health
		3b		Outer race fault (Fatigue pitting)
		4b		Inner race fault + Outer race fault(Fatigue pitting)
		5b		Inner race fault + Outer race fault (Plastic deformations)
	Fault mode	Artificial damage & Real damage by accelerated life test		
	Data type	Fault data sampled with 64,000 Hz		
Dataset C: XJTU dataset	Bearing type	Rolling bearing, LDK UER204		
	Working condition	Load : 10 kN / 11 kN / 12 kN	Speed: 2100 rpm/2250 rpm / 2400 rpm	
	Health state label & Health state specification	1c		Inner race fault
		2c		Outer race fault
		3c		Inner race fault + Outer race fault
		4c		Health
	Fault mode			Real damage by accelerated life test
	Data type			Full life cycle running data sampled with 25,600 Hz

Table 3
Transfer tasks for experimental case I.

TIM for dataset A: CWRU			TIM for dataset B: Paderborn bearing dataset		
Task	Transfer scenario	Target Classes	Task	Transfer scenario	Target Classes
C_{A-1}	1797 rpm→1730 rpm	Non-Partial	C_{B-1}	1000 N →400 N	Non-Partial
C_{A-2}	1797 rpm→1730 rpm	1,3,5,7,9	C_{B-2}	1000 N →400 N	1a,3a
C_{A-3}	1797 rpm→1730 rpm	246,810	C_{B-3}	1000 N →400 N	2a,3a
C_{A-4}	1797 rpm→1730 rpm	1,2,3,8	C_{B-4}	1000 N →400 N	3a
C_{A-5}	1797 rpm→1730 rpm	2,5,8	C_{B-5}	1000 N →400 N	Non-Partial
C_{A-6}	1797 rpm→1730 rpm	3,7,9	C_{B-6}	1000 N →400 N	1b,3b,5b
C_{A-7}	1797 rpm→1730 rpm	1,2	C_{B-7}	1000 N →400 N	2b,4b,5b
C_{A-8}	1797 rpm→1730 rpm	3,5	C_{B-8}	1000 N →400 N	1b,5b
C_{A-9}	1797 rpm→1730 rpm	3	C_{B-9}	1000 N →400 N	3b,4b
C_{A-10}	1797 rpm→1730 rpm	2	C_{B-10}	1000 N →400 N	3b

Table 4
Means and standard deviations of the testing accuracies on TIM for dataset A.

Task Name	Baseline	MK-MMD	ML-MMD	GAN	SAN	DA-GAN
C_{A-1}	96.4(±1.5)	98.5(±1.1)	95.5(±0.3)	99.3(±0.2)	97.2(±0.6)	**99.7(±0.1)**
C_{A-2}	93.7(±1.7)	95.4(±1.3)	96.2(±0.2)	97.4(±0.4)	91.5(±0.7)	**99.6(±0.2)**
C_{A-3}	99.6(±0.1)	98.2(±0.9)	98.2(±0.2)	99.6(±0.1)	98.4(±0.1)	**99.9(±0.1)**
C_{A-4}	94.6(±0.1)	96.7(±1.5)	93.6(±0.4)	98.0(±0.4)	86.2(±1.1)	**99.4(±0.1)**
C_{A-5}	99.3(±0.1)	99.8(±0.2)	99.8(±0.1)	99.4(±0.2)	99.9(±0.1)	**99.9(±0.1)**
C_{A-6}	88.1(±0.3)	67.5(±2.8)	74.5(±0.7)	89.4(±0.6)	83.6(±0.5)	**96.2(±0.4)**
C_{A-7}	99.8(±0.1)	99.8(±0.1)	99.8(±0.1)	99.8(±0.1)	99.9(±0.1)	**97.3(±0.4)**
C_{A-8}	81.9(±2.9)	73.4(±3.0)	90.6(±0.4)	94.8(±0.6)	74.3(±1.2)	**99.8(±0.1)**
C_{A-9}	66.9(±2.9)	78.1(±3.1)	78.8(±0.7)	83.7(±1.0)	44.9(±1.5)	**91.4(±1.3)**
C_{A-10}	99.7(±0.1)	99.9(±0.2)	99.7(±0.2)	99.8(±0.1)	99.9(±0.1)	**99.6(±0.1)**
Average	92.0	90.7	92.6	96.1	87.6	**98.3**

has limited categories of faults or only one certain health state compared with multi-labeled training data. Therefore, the proposed DA-GAN is well suited for this type of partial transfer problem.

The TIM fault diagnosis results for dataset B are presented in Table 5. Similar as Table 4, all the deep transfer approaches could reach high prediction accuracy in non-partial scenario. Therefore, the main comparison is focused on the partial transfer problems. It can be observed that the transferability of these compared approaches has been degraded dramatically for task C_{B-9} and C_{B-10}. One main reason is that the outlier label space would lead to negative transfer when conducting the domain adaptation process. To compare the negative effect of outlier source data within different approaches, in Fig. 7 the confusion matrices of testing accuracy on partial transfer tasks C_{B-9} and C_{B-10} are given.

It can be observed that in the task of C_{B-9}, the ML-MMD, GAN and SAN approaches have categorized target data of label-3 and label-4 into one class, which greatly affect the transferability and even produce negative transfer. In the task of C_{B-10}, the negative effect on the transferability from outlier source data is even more obvious, the ML-MMD approach incorrectly classifies more than 80 % of the target data from label-3 into label-4. A possible reason causing negative transfer is that these transfer learning models may learn the common feature mode from label-3 and label-4 to adapt source and target domains but fail to capture the discriminative features between them. Since the label-4 called as inner fault and outer fault includes the identical fault type as in the label-3 called as outer fault, it could easily lead to the negative transfer when conducting domain adaptation. Especially when there is only one type of data in the target domain but there is another type source domain data including identical fault modes as the target domain data, this outlier data would affect the domain-shared feature learning process and lead to unexpected negative transfer correspondingly.

While in the proposed DA-GAN model, this negative transfer effect could be well suppressed by applying the double layer attention mechanism. The domain and sample attention matrices of transfer tasks C_{B-9} and C_{B-10} are showed in Fig. 8. For task C_{B-9}, the actual label spaces in the target domain are label-3 (Outer fault) and label-4 (Inner fault and outer fault). It could be observed that these two kinds of labels have been lied on more weights compared with other labels in the domain attention matrix. What's more, in the sample attention matrix, the weight of each sample has been arranged correctly according to its latent fault class. For instance, the target samples belonged to label-4 have more weights on the fourth row of the attention matrix, which means that the sub-domain of discriminator label-4 would pay more attention to training with these samples compared with other sub-domain discriminators. For task C_{B-10} where the target data has only one fault class of label-3, the domain attention matrix and sample attention matrix both successfully distribute more weights on label-3 to conduct domain adaptation process.

According to the above comparative experimental results for partial transfer tasks C_B, it could be further concluded that DA-GAN model could effectively suppress the negative transfer effect cause by the outlier data. Even in the case where the source outlier data share the identical feature mode as the target data (e.g. outlier label "inner fault and outer fault" in the source domain and label "outer fault" in the target domain), the designed double layer attention mechanism could still guide the sub-domain discriminators to know where to transfer and to learn more discriminative features for positive transfer.

4.5. Case study II: Partial transfer problem for TDM scenario with similar fault characteristic

4.5.1. Experiment description

In this sub-section, the partial transfer learning problem is explored for TDM scenario with similar fault characteristic. Two types of TDM transfer tasks are designed to evaluate the transfer performance comparatively, which are listed in Table 6. In the first type of TDM transfer tasks, the source data and target data have the same fault mode called as artificial fault (electric discharge machining, EDM), but the data from two domains are collected from different machines with different sensors. In the second type of TDM transfer tasks, the fault mode, sensor location and testing machine of these two domains are all different, which could further explore the transferability of these approaches under large domain variance. The detailed parameters used for training the TDM model is given in Table 7.

4.5.2. Experimental results and performance comparisons

The TDM fault diagnosis knowledge transfer performance of different approaches is given in Table 8, similar as case I, each transfer task is averaged 10 trials to reduce randomness and to provide mean value and standard deviation of the testing accuracies.

From the comparative results shown in Table 8, it could be observed that the feature-based methods (MK-MMD and ML-MMD) and adversarial-based methods (GAN and SAN) could promote positive transfer compared with baseline method for TDM scenario, but the transferability is much lower than these model's

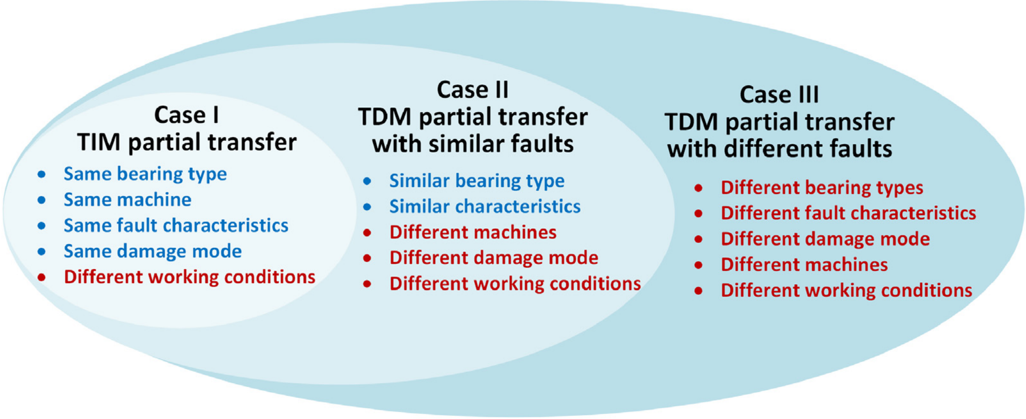
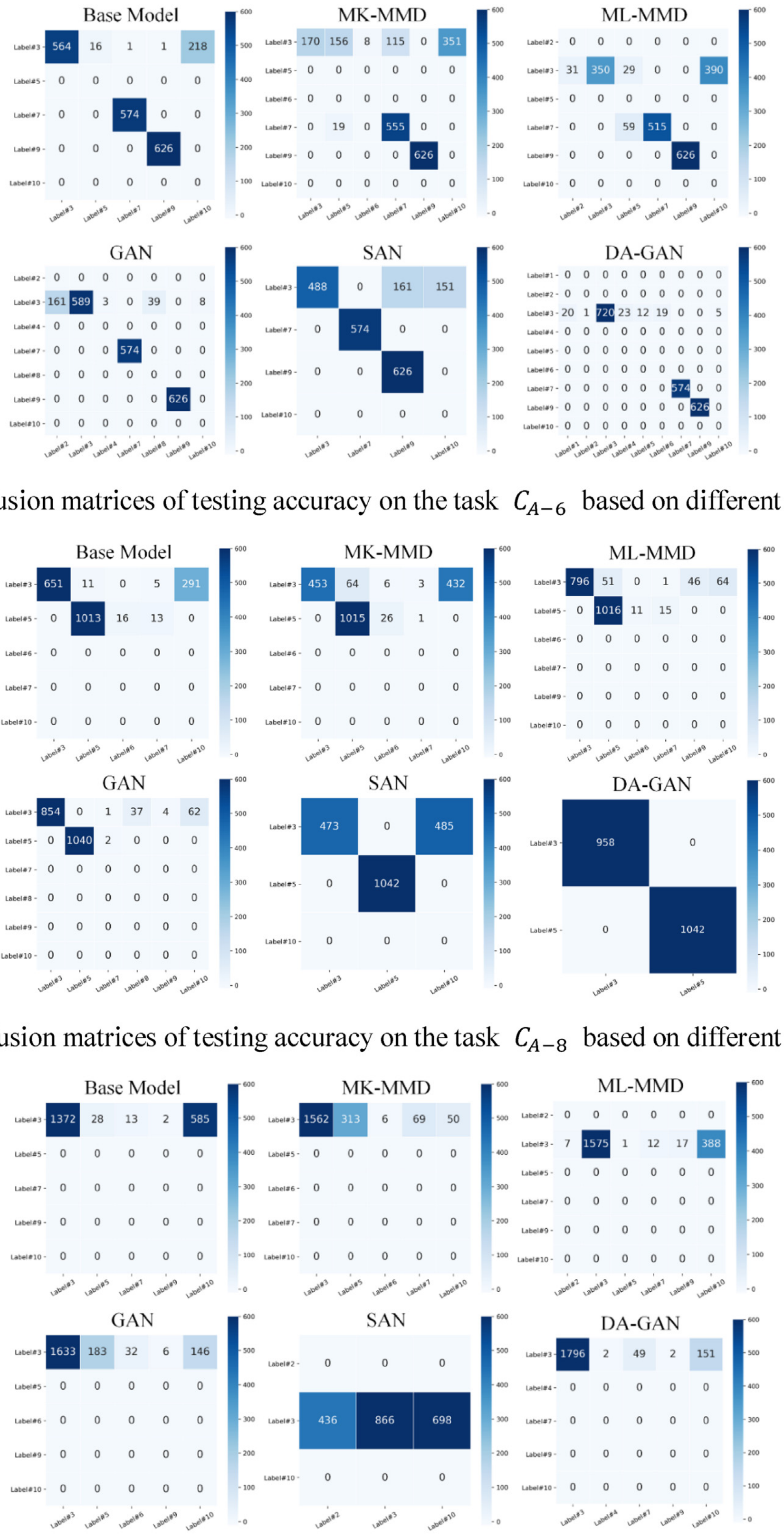


Fig. 5. The comparisons of designed experimental cases.
Confusion matrices of testing accuracy on the task C_{A-6} based on different approaches.
Confusion matrices of testing accuracy on the task C_{A-8} based on different approaches.
Confusion matrices of testing accuracy on the task C_{A-9} based on different approaches.



a Confusion matrices of testing accuracy on the task C_{A-6} based on different approaches

b Confusion matrices of testing accuracy on the task C_{A-8} based on different approaches

c Confusion matrices of testing accuracy on the task C_{A-9} based on different approaches

Fig. 5. (*Continued*)

performance on TIM scenario. More seriously, in some partial transfer tasks as C_{A-B-4}, C_{A-B-5} and C_{B-A-4}, the feature-based methods and SAN model even fail to transfer diagnosis knowledge. While the proposed DA-GAN model provides significant improvement on the positive transfer among these TDM tasks, which achieves the average testing accuracy of 87.9 %.

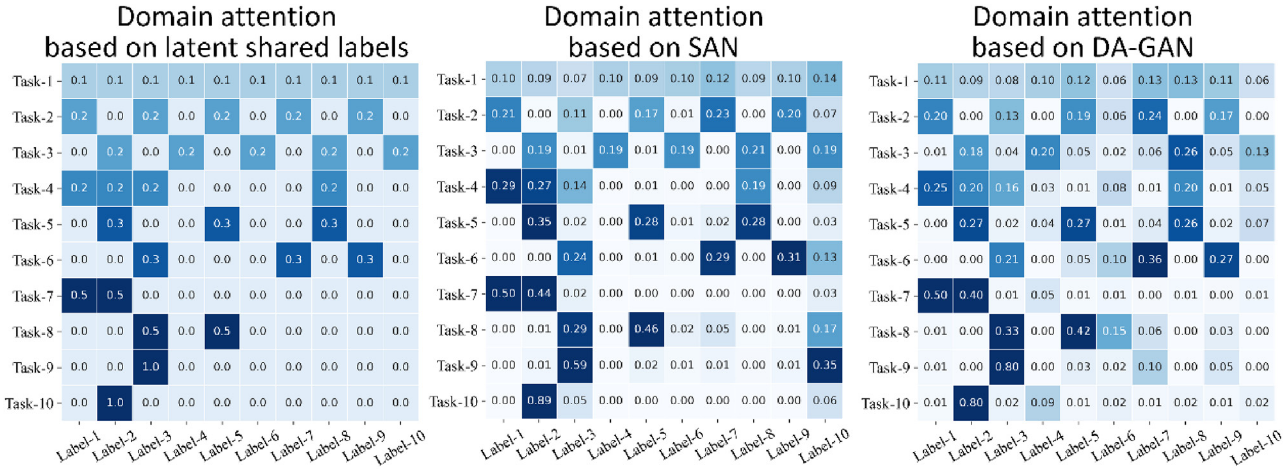


Fig. 6. Domain attention matrices on transfer tasks for dataset A based on SAN and DA-GAN.

Table 5
Means and standard deviations of the testing accuracies on TIM for dataset B.

Task Name	Baseline	MK-MMD	ML-MMD	GAN	SAN	DA-GAN
C_{B-1}	89.9(±1.3)	97.8(±0.6)	95.7(±0.7)	98.8(±0.1)	98.9(±0.4)	**99.9(±0.1)**
C_{B-2}	81.0(±3.2)	99.4(±0.3)	97.7(±0.5)	89.5(±1.7)	86.1(±1.4)	**99.9(±0.1)**
C_{B-3}	98.8(±0.1)	99.3(±0.4)	98.6(±0.4)	93.3(±1.4)	95.9(±1.1)	**99.5(±0.1)**
C_{B-4}	99.3(±0.4)	97.7(±0.6)	99.9(±0.1)	94.9(±1.0)	99.9(±0.1)	**98.4(±0.4)**
C_{B-5}	92.5(±1.1)	81.7(±1.7)	99.7(±0.2)	99.9(±0.1)	87.6(±3.8)	**99.9(±0.1)**
C_{B-6}	84.8(±1.9)	76.5(±1.6)	78.5(±1.4)	87.9(±3.1)	91.4(±2.2)	**99.6(±0.1)**
C_{B-7}	97.7(±1.2)	99.7(±0.2)	99.8(±0.1)	89.6(±2.8)	99.5(±0.6)	**99.8(±0.1)**
C_{B-8}	97.0(±1.6)	99.8(±0.2)	99.7(±0.1)	78.2(±2.9)	87.2(±2.4)	**99.8(±0.1)**
C_{B-9}	77.8(±2.0)	74.5(±1.7)	58.3(±1.5)	59.0(±3.2)	40.5(±3.5)	**98.8(±0.2)**
C_{B-10}	62.8(±2.8)	83.9(±1.6)	14.5(±1.0)	84.1(±2.7)	51.5(±5.8)	**99.7(±0.1)**
Average	88.1	91.0	84.2	87.5	83.8	**99.5**

Table 6
Transfer tasks for experimental case II.

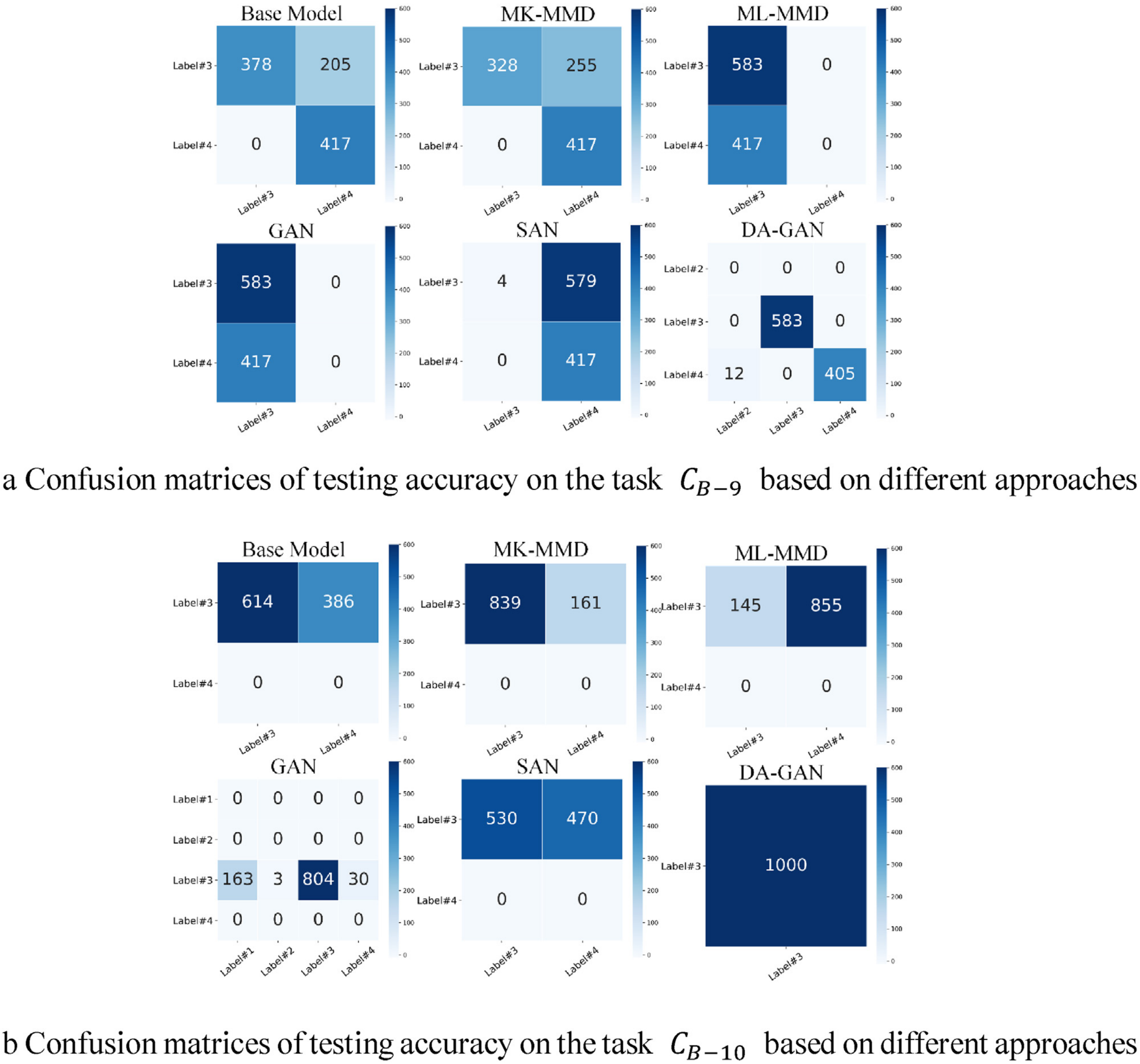
TDM for dataset A and dataset B		
Task	Transfer scenario	Target Classes
C_{A-B-1}	CWRU artificial damage→Paderborn artificial damage	1a,2a,3a
C_{A-B-2}	CWRU artificial damage→Paderborn artificial damage	1a,2a
C_{A-B-3}	CWRU artificial damage→Paderborn artificial damage	1a,3a
C_{A-B-4}	CWRU artificial damage→Paderborn artificial damage	2a
C_{A-B-5}	CWRU artificial damage→Paderborn artificial damage	3a
C_{B-A-1}	Paderborn natural degradation→CWRU artificial damage	1,2,3
C_{B-A-2}	Paderborn natural degradation→CWRU artificial damage	1,2
C_{B-A-3}	Paderborn natural degradation→CWRU artificial damage	1,3
C_{B-A-4}	Paderborn natural degradation→CWRU artificial damage	2
C_{B-A-5}	Paderborn natural degradation→CWRU artificial damage	3

Table 7
Parameters used for models in experimental case II.

Parameter	Value	Parameter	Value
Source domain sample number	2000	Training iterations	500
Target domain sample number	2000	Batch size	100
Source domain sample length	1280	Learning rate	0.001
Target domain sample length	6400	Tuning parameters α	0.5

To further investigate the effectiveness of proposed model, the confusion matrix of each model on task C_{A-B-1} in Fig. 9. From the comparative results shown in Fig. 9, all the transfer learning approaches could promote the positive transfer on the label#1, but the transferability on the label-2 and label-3 is far poor. These unexpected negative transfer results could be mainly attributed to two reasons: incorrect adaptation and outlier data interference:

1) Incorrect adaptation: Since the transfer task belongs to TDM scenario, there has a large distribution shift across the source and target domains. It is possible that the source domain data and target domain data belong to the different classes, but still share some common feature distributions. In such case, the transfer learning model tends to learn incorrect domain-invariant features and fail for diagnosis knowledge transferring.



a Confusion matrices of testing accuracy on the task C_{B-9} based on different approaches

b Confusion matrices of testing accuracy on the task C_{B-10} based on different approaches

Fig. 7. Confusion matrices of testing accuracy on the task C_{B-9} based on different approaches.
Confusion matrices of testing accuracy on the task C_{B-10} based on different approaches.

Table 8
Means and standard deviations of the testing accuracies on case II.

Task Name	Baseline	MK-MMD	ML-MMD	GAN	SAN	DA-GAN
C_{A-B-1}	17.7(±2.7)	51.0(±3.3)	60.0(±0.7)	64.6(±2.8)	49.5(±3.2)	**98.1(±1.1)**
C_{A-B-2}	17.8(±3.5)	49.1(±2.9)	28.9(±0.7)	50.4(±2.9)	50.2(±2.8)	**83.9(±3.3)**
C_{A-B-3}	17.5(±2.0)	49.3(±3.3)	49.8(±0.7)	49.6(±3.1)	49.5(±0.3)	**90.4(±2.3)**
C_{A-B-4}	/	7.9(±2.7)	11.9(±0.5)	96.4(±1.8)	/	**98.4(±1.1)**
C_{A-B-5}	/	34.6(±3.2)	70.8(±0.7)	99.8(±0.1)	/	**99.4(±0.4)**
C_{B-A-1}	16.4(±2.7)	40.4(±3.9)	44.5(±0.7)	41.7(±3.4)	50.3(±3.9)	**81.3(±3.0)**
C_{B-A-2}	15.2(±2.8)	43.3(±2.7)	17.3(±0.6)	49.8(±2.5)	72.1(±3.0)	**88.0(±1.5)**
C_{B-A-3}	27.9(±3.3)	36.2(±3.8)	37.8(±0.7)	18.7(±2.8)	61.6(±2.2)	**61.0(±1.2)**
C_{B-A-4}	/	/	/	43.5(±3.4)	/	**88.7(±2.1)**
C_{B-A-5}	27.1(±3.0)	27.8 (±2.6)	86.5(±0.5)	63.6(±2.7)	63.3(±3.7)	**90.4(±1.5)**
Average	13.9	33.9	40.7	57.8	39.7	**87.9**

2) Outlier data interference: The outlier data in the source domain may also contain similar feature mode as the target data, which would lead to the extra confusion for domain adaptation process. The transferability would be severely reduced under the negative effect of outlier data.

To visually understand how these factors influence the performance of transfer models, the FFT spectrum distributions of the source data and target data for task C_{A-B-1} are given in Fig. 10. It can be seen that the source data from label-1 has some common distribution model as the target data from label-2. It is also

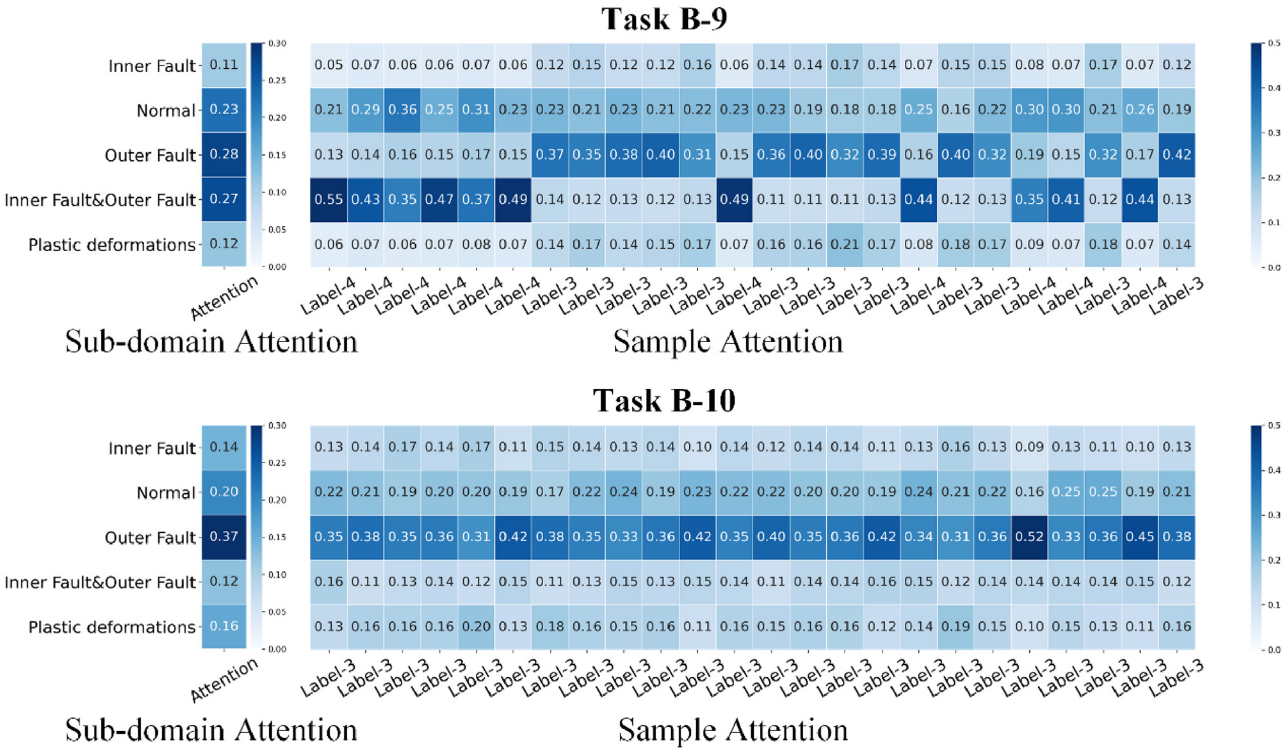


Fig. 8. Domain and sample attention matrices constructed by DA-GAN for C_{B-9} and C_{B-10}.

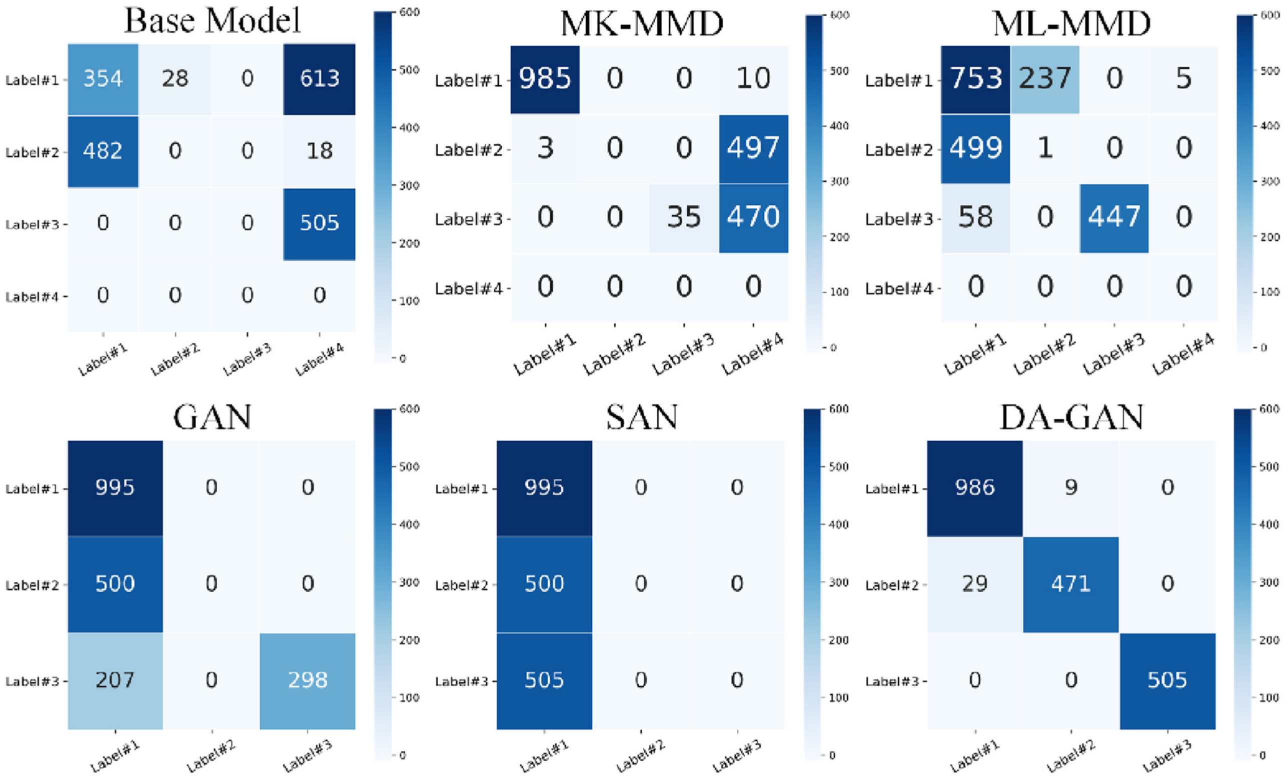


Fig. 9. Confusion matrices of testing accuracy on C_{A-B-1} based on different approaches.

obvious that the outlier data from label-4 has some similar distributions as the target data from label-2 and label-3. This could well explain the transferability degradation of compared models in Fig. 9. For instance, the ML-MMD method has recognized target data from label-2 as label-1, which could be attributed to the incorrect adaptation. The MK-MMD method categorizes the target data from label-2 and label-3 into label#4, which is mainly influenced by the effect of outlier data.

In the proposed DA-GAN model, these two unexpected problems could be well suppressed by applying double layer attention

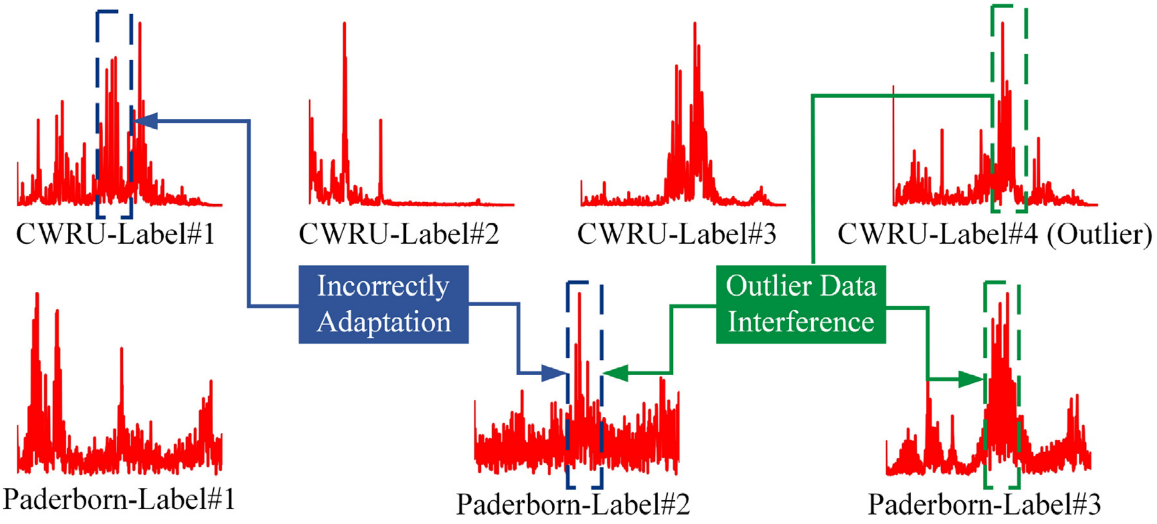


Fig. 10. FFT spectrum distributions of source domain data and target domain data.

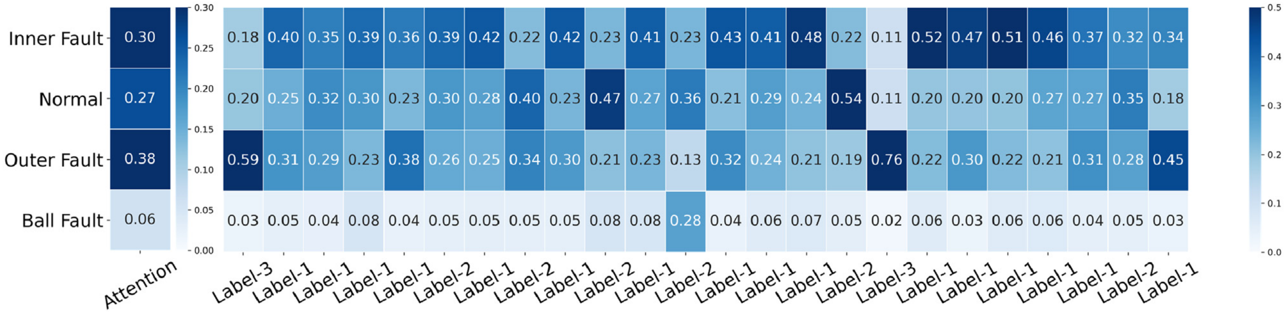


Fig. 11. Domain and sample attention matrices constructed by DA-GAN for C_{A-B-1}.

mechanism, which is illustrated in Fig. 11. It can be seen that the domain attention matrix could guide the model to activate correct sub-domain discriminators label-1, label-2 and label-3 for subsequent adaptation process. This could alleviate the interference from outlier data label space. Further the sample attention matrix is constructed to utilize data discriminately for each sub-domain discriminators. It can be observed that each sample would be assigned more probability on its corresponding label, this could promote the correct domain alignment across different domain data and greatly reduce the negative effect caused by wrong adaptation.

4.6. Case study III: Partial transfer problem for TDM scenario with different fault characteristics

4.6.1. Experiment description

In this sub-section, a more challenged partial transfer scenario, where the data from source domain and target domain are not only from different machines, but also have different fault characteristics, such as obvious bearing fault through artificial damage and weak bearing fault from the early-stage & middle-stage of nature degradation.

Totally 12 transfer tasks with different fault characteristics are designed in case III, including different types of fault characteristic transfer problems: transferring diagnosis knowledge from artificial damage fault data to early-stage & middle stage fault data (Task C_{A-C-1} to C_{A-C-6}) and transferring diagnosis knowledge from nature degraded fault data to early-stage & middle stage fault data (Task C_{B-C-1} to C_{B-C-6}). The detailed transfer tasks specification and parameters setting are listed in Tables 9 and 10respectively.

4.6.2. Experimental results and performance comparisons

The comparative results of transferring diagnosis knowledge across different machines with different degrees of fault characteristics is given in Table 11. Each transfer task is averaged 10 trials to reduce the randomness and to provide the mean value and standard deviation of the testing accuracies.

From Table 11 it can be seen that the proposed DA-GAN could effectively improve the model accuracy compared with other approaches in the partial transfer scenarios with different fault characteristics. For example, in the transfer task C_{A-C-6}, where the labelled training data are from CWRU dataset with obvious artificial fault characteristics and testing unlabeled data are from XJTU dataset with early fault characteristics, the proposed method (DA-GAN with 89.4 % accuracy) shows great superiority in promoting positive transfer compared with other GAN-based method (GAN with 71.4 % accuracy & SAN with 75.2 % accuracy). What's more, the feature-based model (MK-MMD with 39.6 % accuracy) even leads to unexpected negative transfer, which degenerate the diagnosis model trained from source data solely without transfer.

The classification results on task C_{B-C-2} based on different approaches are given in Fig. 12. It can be seen that the feature-based transfer models (MK-MMD and ML-MMD) classify all the inner fault and outer fault data as normal data wrongly, which lead to unexpected negative transfer and degenerate the base model. Compared with feature-based models, adversarial-based models (GAN and SAN) could effectively transfer diagnosis knowledge about classifying fault and health data, however, they could not discriminate the difference between inner fault and outer fault data since all fault data are classified as inner fault. The proposed DA-GAN model could effectively solve the issue through the double-layer atten-

Table 9
Transfer tasks for experimental case III.

TDM from dataset A to dataset C, TDM from dataset B to dataset C		
Task	Transfer scenario	Target Classes
C_{A-C-1}	CWRU artificial damage→XJTU-SY middle-stage fault	1c,2c,3c
C_{A-C-2}	CWRU artificial damage→XJTU-SY middle -stage fault	1c,2c
C_{A-C-3}	CWRU artificial damage→XJTU-SY middle -stage fault	1c,3c
C_{A-C-4}	CWRU artificial damage→XJTU-SY early-stage fault	1c,2c,3c
C_{A-C-5}	CWRU artificial damage→XJTU-SY early-stage fault	1c,2c
C_{A-C-6}	CWRU artificial damage→XJTU-SY early-stage fault	1c,3c
C_{B-C-1}	Paderborn natural degradation→XJTU-SY middle-stage fault	1c,2c,3c,4c
C_{B-C-2}	Paderborn natural degradation→XJTU-SY middle-stage fault	1c,3c,4c
C_{B-C-3}	Paderborn natural degradation→XJTU-SY middle-stage fault	1c,3c
C_{B-C-4}	Paderborn natural degradation→XJTU-SY early-stage fault	1c,2c,3c,4c
C_{B-C-5}	Paderborn natural degradation→XJTU-SY early-stage fault	1c,3c,4c
C_{B-C-6}	Paderborn natural degradation→XJTU-SY early-stage fault	1c,3c

Table 10
Parameters used for models in experimental case III.

Parameter	Value	Parameter	Value
Source domain sample number	2000	Training iterations	500
Target domain sample number	2000	Batch size	100
Source domain sample length	1280 & 6400	Learning rate	0.001
Target domain sample length	2560	Tuning parameters α	0.5

Table 11
Means and standard deviations of the testing accuracies on case III.

Task Name	Baseline	MK-MMD	ML-MMD	GAN	SAN	DA-GAN
C_{A-C-1}	52.0(±7.8)	50.8(±4.4)	65.2(±3.5)	83.8(±6.2)	79.0(±3.9)	**93.6(±3.2)**
C_{A-C-2}	34.2(±6.7)	55.4(±3.5)	70.0(±2.9)	81.6(±4.3)	93.2(±2.7)	**99.0(±0.6)**
C_{A-C-3}	50.4(±6.7)	68.0(±3.4)	98.4(±0.7)	80.2(±3.8)	84.2(±4.4)	**93.4(±1.8)**
C_{A-C-4}	42.2(±6.4)	51.0(±3.9)	60.0(±3.6)	61.8(±6.2)	67.6(±6.6)	**77.2(±5.4)**
C_{A-C-5}	28.6(±6.7)	37.2(±3.4)	64.4(±2.9)	68.0(±6.4)	84.4(±2.1)	**94.0(±2.7)**
C_{A-C-6}	59.4(±9.0)	39.6(±3.3)	75.0(±2.4)	71.4(±8.6)	75.2(±6.9)	**89.4(±5.1)**
C_{B-C-1}	28.4(±4.8)	51.0(±4.1)	47.8(±3.3)	74.2(±4.4)	71.2(±5.7)	**90.0(±3.7)**
C_{B-C-2}	66.0(±5.2)	52.0(±4.7)	52.0(±3.2)	72.2(±4.7)	70.0(±4.2)	**94.0(±3.2)**
C_{B-C-3}	/	32.6(±3.5)	/	66.6(±5.9)	57.6(±4.4)	**89.0(±2.9)**
C_{B-C-4}	33.8(±5.5)	40.6(±3.6)	25.4(±2.9)	72.6(±4.5)	65.6(±6.9)	**76.6(±5.4)**
C_{B-C-5}	68.8(±6.0)	50.0(±4.3)	48.8(±3.2)	79.4(±6.5)	98.0(±1.8)	**98.8(±1.0)**
C_{B-C-6}	29.6(±6.1)	37.8(±3.9)	48.6(±3.2)	71.4(±6.8)	51.4(±5.8)	**86.2(±6.1)**
Average	41.1	47.2	54.6	73.6	74.8	**90.1**

tion mechanism. The attention matrices based on DA-GAN for task C_{B-C-2} is illustrated as Fig. 13. It can be seen that all samples will be guided to conduct domain adaptation adaptively according to the double layer attention matrices. Especially for testing data with weak fault characteristics (such as inner fault and outer fault during early-stage degradation), the sample weight matrix could promote the positive transfer by assigning discriminative weights to these confusing data, which could effectively improve the transferability in the scenario where source domain and target domain have different fault characteristics.

5. Conclusions

In this paper, a novel adversarial-based approach is proposed to address the partial transfer problem for mechanical fault diagnosis. The designed double layer attention mechanism could promote the positive transfer and alleviate the negative effect of irrelevant source data. The first domain attention is applied to decide which label space in the target domain should be shared for the current transfer task, and the second sample attention is implemented to know which samples should be focused on for each sub-domain discriminator.

We believe that the proposed DA-GAN model could shed a new angle for solving mechanical fault diagnosis transfer learning problems, which enables the network to know where to transfer instead of conducting domain adaptation indiscriminately. Three experimental case studies have been investigated and the comparative results validate that the proposed method shows great superiority on promoting positive transfer under different degrees of domain shifts, such as different working conditions, different machines and different fault characteristics. Especially in the extreme cases where the target label space has large-biased data compared with source domain, the proposed method could effectively alleviate the negative transfer caused by the outlier source data. Consider that above extreme cases, in which only one fault occurs in a certain machine, are more common in practice, the proposed DA-GAN could promote the extension of diagnosis model from academic research to engineering scenarios.

Another possible case may also occur in practice, where some additional faults not belonging to the source domain classes happen in the testing machine. This problem called as open set domain adaptation will be explored in the future research.

CRediT authorship contribution statement

Yafei Deng: Conceptualization, Methodology, Software, Writing - original draft. **Delin Huang:** Investigation, Validation. **Shichang**

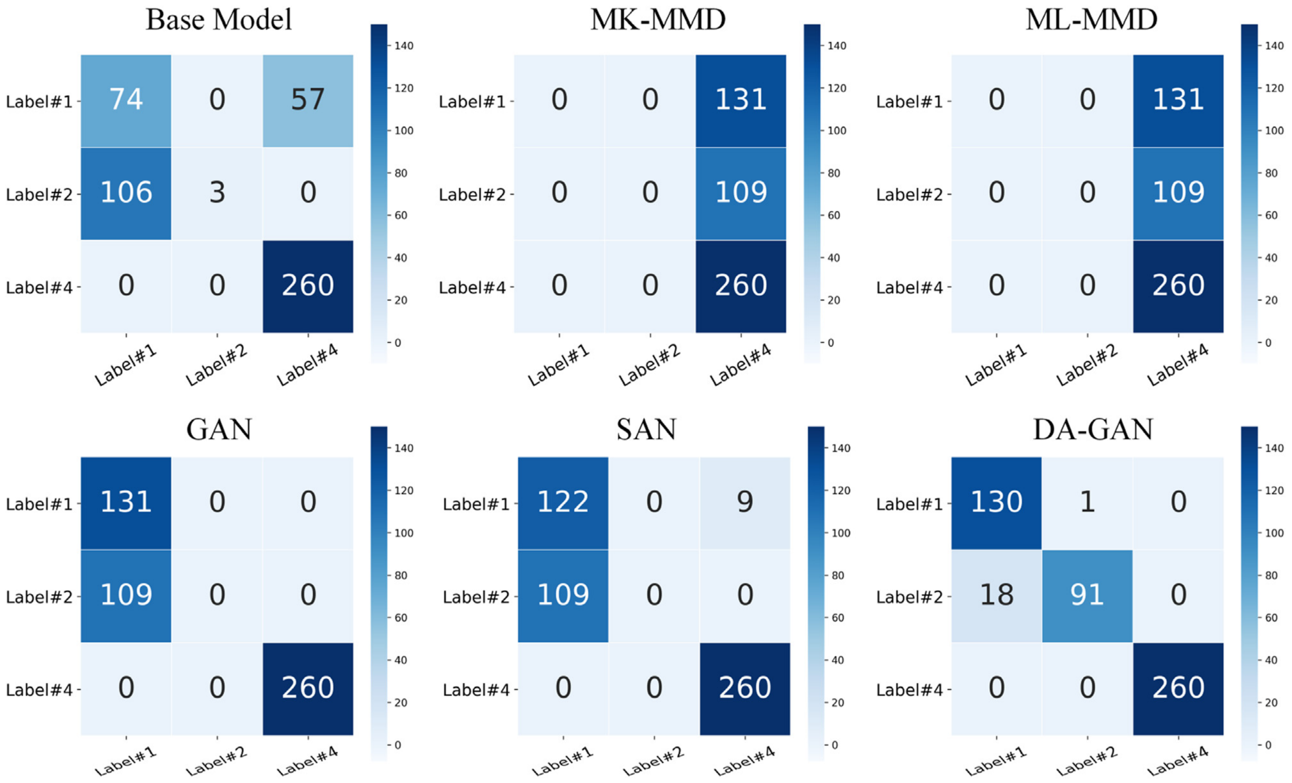


Fig. 12. Confusion matrices of testing accuracy on C_{B-C-2} based on different approaches.

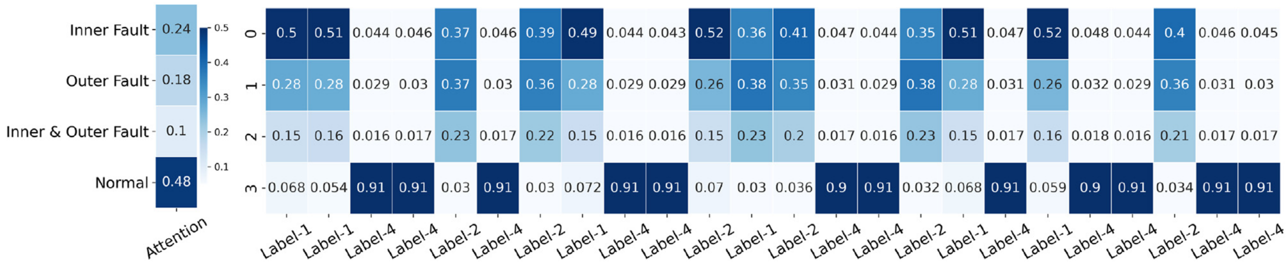


Fig. 13. Domain and sample attention matrices constructed by DA-GAN for C_{B-C-2}.

Du: Supervision. **Guilong Li:** Software. **Chen Zhao:** Visualization. **Jun Lv:** Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was funded by National Natural Science Foundation of China (Grant No. 51775343) and Shanghai Pujiang Program (Grant No. 18PJC031).

References

Bearing DataCenter, Paderborn University. [Online]. Available: https://mb.uni-paderborn.de/kat/forschung/datacenter/bearing-datacenter.

Bearings Data Center, Seeded Fault Test Data, Case Western Reserve University. [Online]. Available: http://csegroups.case.edu/bearingdatacenter/pages/download-data-file.

Cao, Pei, Zhang, Shengli, Tang, Jiong, 2018a. Preprocessing-free gear fault diagnosis using small datasets with deep convolutional neural network-based transfer learning. IEEE Access 6, 26241–26253.

Cao, Zhangjie, et al., 2018b. Partial transfer learning with selective adversarial networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Cao, Zhangjie, et al., 2018c. Partial adversarial domain adaptation. Proceedings of the European Conference on Computer Vision (ECCV).

Che, Changchang, et al., 2020. Domain adaptive deep belief network for rolling bearing fault diagnosis. Comput. Ind. Eng., 106427.

Flandrin, Patrick, Rilling, Gabriel, Goncalves, Paulo, 2004. Empirical mode decomposition as a filter bank. IEEE Signal Process. Lett. 11.2, 112–114.

Guo, Liang, et al., 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, Te, et al., 2019. A novel adversarial learning framework in deep convolutional neural network for intelligent diagnosis of mechanical faults. Knowledge Based Syst. 165, 474–487.

Lei, Yaguo, et al., 2018. Machinery health prognostics: a systematic review from data acquisition to RUL prediction. Mech. Syst. Signal Process. 104, 799–834.

Lei, Yaguo, et al., 2020. Applications of machine learning to machine fault diagnosis: a review and roadmap. Mech. Syst. Signal Process. 138, 106587.

Li, Xiang, Zhang, Wei, Ding, Qian, 2018a. A robust intelligent fault diagnosis method for rolling element bearings based on deep distance metric learning. Neurocomputing 310, 77–95.

Li, Xiang, Zhang, Wei, Ding, Qian, 2018b. Cross-domain fault diagnosis of rolling element bearings using deep generative neural networks. IEEE Trans. Ind. Electron. 66.7, 5525–5534.

Li, Xin, et al., 2019a. A novel deep stacking least squares support vector machine for rolling bearing fault diagnosis. Comput. Ind. 110, 36–47.

Li, Zhe, Wang, Yi, Wang, Kesheng, 2019b. A deep learning driven method for fault classification and degradation assessment in mechanical equipment. Comput. Ind. 104, 1–10.

Li, Xiang, et al., 2020a. Partial transfer learning in machinery cross-domain fault diagnostics using class-weighted adversarial networks. Neural Netw.

Li, Xudong, et al., 2020b. Fault diagnostics between different type of components: a transfer learning approach. Appl. Soft Comput. 86, 105950.

Liang, Pengfei, et al., 2019. Compound fault diagnosis of gearboxes via multi-label convolutional neural network and wavelet transform. Comput. Ind. 113, 103132.

Liu, Tao, 2020. A bearing fault diagnosis method based on enhanced singular value decomposition. IEEE Trans. Industr. Inform.

Márquez, Adolfo Crespo, Crespo Del Castillo, Adolfo, Fernández, Juan F.G.ómez, 2020. Integrating artificial intelligent techniques and continuous time simulation modelling. Practical predictive analytics for energy efficiency and failure detection. Comput. Ind. 115, 103164.

Pan, Sinno Jialin, Yang, Qiang, 2009. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22.10, 1345–1359.

Shao, Siyu, et al., 2018. Highly accurate machine fault diagnosis using deep transfer learning. IEEE Trans. Industr. Inform. 15.4, 2446–2455.

Wen, Long, Gao, Liang, Li, Xinyu, 2017. A new deep transfer learning based on sparse auto-encoder for fault diagnosis. IEEE Trans. Syst. Man Cybern. Syst. 49.1, 136–144.

XJTU-SY Bearing Datasets. [Online]. Available: https://biaowang.tech/xjtu-sy-bearing-datasets/.

Yang, Bin, et al., 2019. An intelligent fault diagnosis approach based on transfer learning from laboratory bearings to locomotive bearings. Mech. Syst. Signal Process. 122, 692–706.

Zhang, Ran, et al., 2017. Transfer learning with neural networks for bearing fault diagnosis in changing working conditions. IEEE Access 5, 14347–14357.

Zhang, Wei, et al., 2020. Machinery fault diagnosis with imbalanced data using deep generative adversarial networks. Measurement 152, 107377.