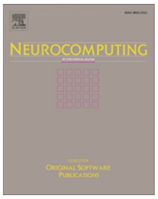
Neurocomputing 548 (2023) 126391

Contents lists available at ScienceDirect

Neurocomputing

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



Combining the theoretical bound and deep adversarial network for machinery open-set diagnosis transfer

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

[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] Faculty of Economics and Management, East China Normal University, 200241 Shanghai, China
[d] College of Intelligent Manufacturing and Control Engineering, Shanghai Polytechnic University, China

ARTICLE INFO

Article history:
Received 16 September 2021
Revised 12 January 2023
Accepted 26 May 2023
Available online 3 June 2023
Communicated by Zidong Wang

Keywords:
Intelligent fault diagnosis
Deep transfer learning
Open-set domain adaptation
Adversarial learning

ABSTRACT

Recently, deep transfer learning-based intelligent machine diagnosis has been well investigated, and the source and the target domain are commonly assumed to share the same fault categories, which can be called as the closed-set diagnosis transfer (CSDT). However, this assumption is hard to cover real engineering scenarios because some unknown new fault may occur unexpectedly due to the uncertainty and complexity of machinery components, which is called as the open-set diagnosis transfer (OSDT). To solve this challenging but more realistic problem, a Theory-guided Progressive Transfer Learning Network (TPTLN) is proposed in this paper. First, the upper bound of transfer learning model under open-set setting is thoroughly analyzed, which provides a theoretical insight to guide the model optimization. Second, a two-stage module is designed to carry out distracting unknown target samples and attracting known samples through progressive learning, which could effectively promote inter-class separability and intra-class compactness. The performance of proposed TPTLN is evaluated in two OSDT cases, where the diagnosis knowledge is transferred across bearings and gearbox running under different working conditions. Comparative results show that the proposed method achieves better robustness and diagnostic performance under different degrees of domain shift and openness variance.

The source codes and links to the data can be found in the following GitHub repository: https://github.com/phoenixdyf/Theory-guided-Progressive-Transfer-LearningNetwork.

1. Introduction

Recent development of deep learning (DL) approaches has greatly improved the performance of mechanical fault diagnosis tasks [1,2], and the substantial prerequisite of the diagnosis accuracy and stability boost is usually based on two assumptions: 1.) large amounts of labeled data are available, and 2) the training dataset and testing dataset follow the same distribution [3]. Considering the practical applications of DL approaches in many industrial scenarios, however, it is time-consuming and labor-intensive to collect sufficient labeled data, especially the labeled fault data, because the machines are always in normal condition with scheduled maintenance. Moreover, similar machines often work in different regimes, due to the specific task demands and working environments, which leads to the distribution discrepancy between the testing data and the training data. As a result, the well performed DL diagnosis models under laboratory settings would degenerate greatly when encountering real-world situations.

The transfer learning (TL) has been demonstrated as a powerful tool for helping deep learning to bridge the gap between performance in the laboratory and in the real world because it allows the knowledge obtained in one or more tasks to be reused to another [4]. In the scenarios of mechanical fault diagnosis, different working conditions and machine components can be regarded as different domains, and the different fault types can be regarded as different tasks. TL aims at transferring the diagnostic knowledge from the source domain (where the diagnostic models can be fully trained with the sufficient labeled data) to the target domain (where the models are difficult to be trained due to the insufficient labeled data and distribution discrepancy). Compared with the DL approaches, the TL approaches have two significant advantages:

1) **Relaxing the labeling setting:** Transfer learning has continuously relaxed the labeling constraints for the target tasks. It started with a supervised setting, where the testing data in

[*] 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.neucom.2023.126391
0925-2312/

the target domain are fully labeled, followed by the semi-supervised setting, where only part of labeled data in the target domain are available, and finally converged at unsupervised setting, where no labeled data could be used in the target domain. The unsupervised transfer learning approaches are expected to deal with the mechanical fault diagnosis problem where the labeled data are hard to collect.

2) **Covering the domain shift:** Transfer learning bridges the gap that there exist different probability distributions between the source domain (data for training) and the target domain (data for testing). It means that a new but related target task could be addressed well through transferring learned knowledge from the source domain. Transfer learning shows great potential to extend the applications of existed DL diagnosis model for mechanical fault diagnostic because it enables the model to cover the difference of working conditions and even the variance within machine component family type.

Generally, the TL-based fault diagnosis approaches could be classified into three types: instance-based approaches [5,6], model-based approaches [7,8] and feature-based approaches [9–13]. Concretely, the goal of instance-based approaches is to promote the diagnostic knowledge transfer through singling out positive instances from the source domain and merging the source data into the target data with the instance weighting strategies. Model-based approaches explore which part of the source pretrained model could facilitate the target domain parameters learning and fine tune the rest components with target domain data. Feature-based approaches construct the mapping function to convert the raw data from source and target domain into a common latent space and extract the domain-invariant features, which mainly includes three strategies: reducing the domain shift with discrepancy-based metrics, generating the domain confusion with adversarial-based mechanism and improving the domain invariance representation with reconstruction-based models.

One of the main assumptions of most machinery diagnosis transfer learning methods is based on the closed-set condition, which indicates that training and testing data should cover the same machine health state. However, this closed-set diagnosis transfer (CSDT) assumption is difficult to satisfy in real engineering fields. Due to the huge economic cost and human labor of data collection, it is often hard to collect sufficient labeled training data with various machine health states, resulting in limited source domain categories. In the testing phase, new unknown fault modes absent in the source domain would occur, and the diagnosis model would wrongly classify the emerging unknown fault type as known fault type. Therefore, a more challenging and practical scenario called as open-set diagnosis transfer (OSDT) is proposed [14]. The OSDT not only transfers the diagnosis knowledge from source domain, but also detects the unknown faults absent in the source domain to expand its diagnosis knowledge in target domain.

Fig. 1 shows the diagram of CSDT and OSDT, the main difference is that OSDT aims at transferring diagnosis knowledge from class-scarce source domain to class-rich target domain and constructing a decision boundary of known and unknown samples. Therefore, the OSDT problems could be solved from two aspects: 1) Learning a decision boundary between known and unknown target samples to achieve the inter-class level separability and 2) Matching the distribution of source samples and target samples in the shared label space to achieve intra-class level compactness.

However, some pending issues limit the development of OSDT solutions:

1) **Uncertain decision boundary:** It is difficult to formulate a certain decision bound to recognize the unknown and known samples without information about the degree of openness. Using a too tight decision bound would ignore some unknown samples under a large degree of openness while using too loose decision bound would include some known samples under a small degree of openness.
2) **Interactive negative transfer:** The two objectives influence each other during the model training negatively. In detail, the known and unknown target samples would become confused under large domain shifts, making it harder to learn an accurate decision boundary. Subsequently, these misclassified samples, belonging to shared (private) space but recognized as private (shared), would adversely influence the distribution matching, which further mislead the decision boundary learning.

To overcome the aforementioned challenges, a novel principle-guided deep adversarial network is proposed to detect the emerging faults and reduce the domain discrepancy. The open-domain detection loss and the invariant learning loss are optimized through minimizing the theoretically derived error bound, which could achieve inter-class level separability and intra-class level compactness with theoretical guarantee. Moreover, a two-stage progressively learning strategy is designed to suppress the interactive negative transfer. Finally, the adversarial-based mechanism is employed to endow the model with the capacity to learn discriminative representations from imbalance data, which could perform well generalization under different degrees of openness.

The overview of proposed Theory-guided Progressive Transfer Learning Network (TPTLN) is illustrated in Fig. 2. As shown in Fig. 2, in the distract stage, a flexible decision bound is adaptively obtained to accurately discriminate target unknown samples, in which the source risk and open set risk would be optimized iteratively to facilitate outlier data (inter-class) separability; and in the attract stage, the domain-invariant features across the shared label space of source and target domains will be learned to promote intra-class compactness.

(a) An example of intial state under open set setting, in which the source domain and target domain share several known fault classes (triangles and squares marked in the figure), and the target domain also contains the unknown fault classes not included in the source domain(circles in the figure). (b) The situation after the distract stage: Firstly the classification boundary of source known faults is built to enable the TPTLN to achieve diagnosis. Secondly, a coarse-to-fine decision boundary is constructed to distract the target samples of known classes and unknown classes. (c) The situation after the attract stage. An adversarial distribution alignment strategy is conducted to attract the samples of the source domain and target domain in the shared space. (d) The situation after progressive learning between distract and attract stages iteratively until to convergence, in which the target samples of unknown classes would be pushed away from shared space, and target samples of the known classes would be close to their source domain corresponding classes.

The main contributions are summarized as follows:

1) A more realistic diagnosis transfer scenario for rotating machines called as OSDT is explored, in which the target domain not only contains shared label space with source domain but also contains private label space (emerging unknown fault types) not included in the source domain.
2) A progressive learning structure TPTLN is designed to execute the target samples dividing (distract stage) and domain distribution aligning (attract stage) separately. The proposed uncertainty calibration, adaptive openness estimation, and weighted distribution modules are attached to the two-stage learning process, which could better accommodate under larger domain shifts and effectively avoid the problem of interactive negative transfer within one model.

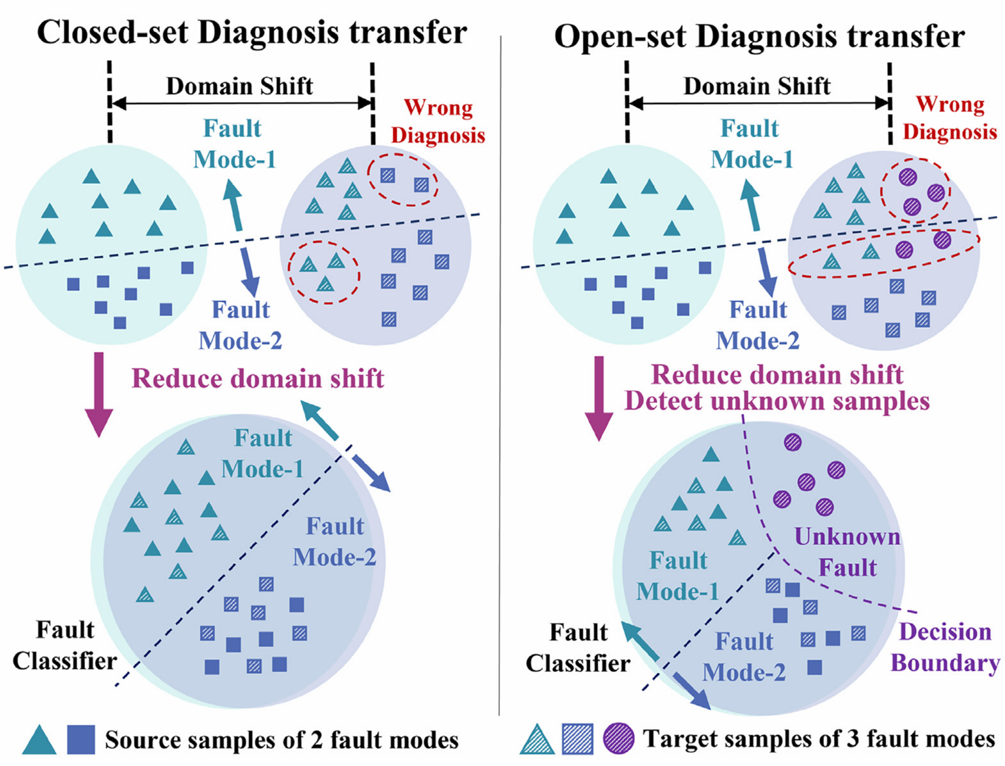


Fig. 1. Comparison diagram of CSDT and OSDT.

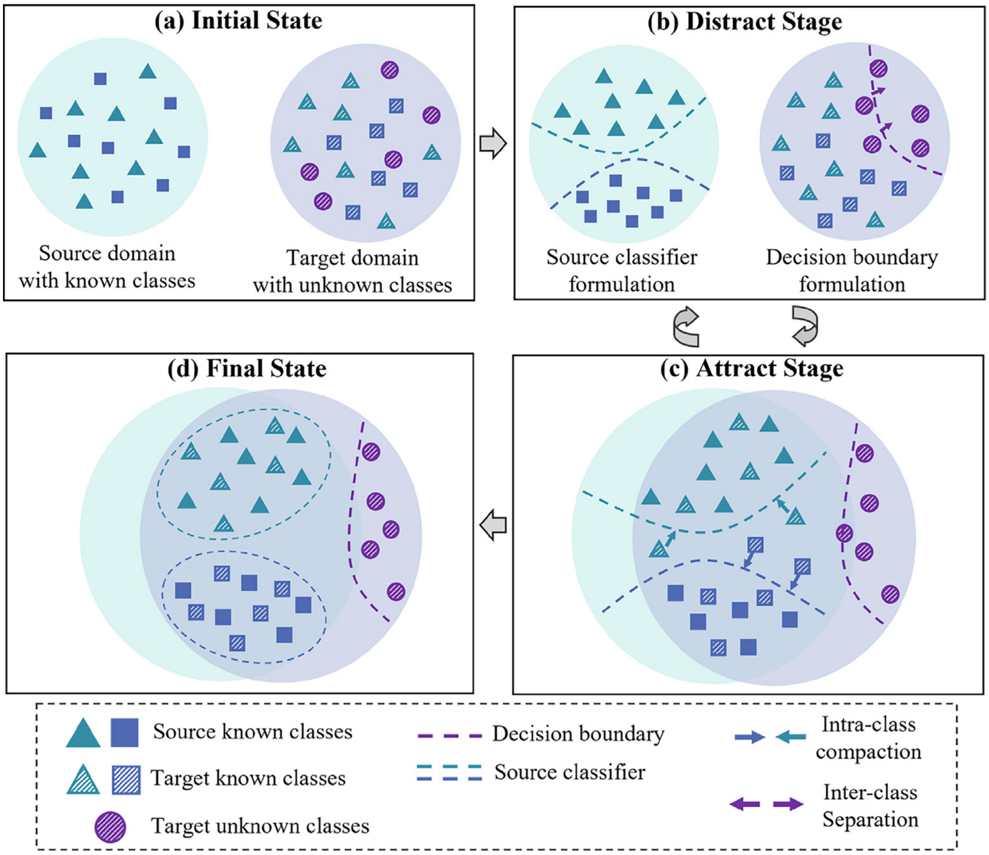


Fig. 2. An overview of proposed TPTLN approach for OSDT problem.

3) The proposed TPTLN is designed with the guidance of theoretical bound analysis. Minimizing the source risk aims to obtain a source classifier, optimizing the distribution discrepancy is to learn domain invariant features, and the open set risk is controlled with proposed calibrated similarity and domain consensus score, providing an adaptive decision boundary to detect the target unknown samples, which accommodates different degrees of target data openness.
4) Various open set diagnosis transfer tasks under different degrees of domain shifts and openness variance are designed to verify the effectiveness of proposed model. Totally seven representative models are selected as baseline and several visualization methods are explored to illustrate the model performance on improving diagnosis accuracy and enhancing transfer robustness.

The rest of this paper is organized as follows. In section 2, related works are briefly discussed. In section 3, the preliminaries are briefly described. In section 4, the method is introduced in detail. In section 5, the experimental cases and comparative results are analyzed. In section 6, the conclusions and future work are presented.

2. Related works

2.1. Closed set diagnosis transfer

Benefiting from the superiority of transfer learning, various CSDT methods have been proposed, and the related research are presented as follows:

1) Instance-based approaches are typically based on instance selection or instance weighting strategies. Song et al. proposed a retraining strategy-based domain adaption network (DAN-R), in which the pseudo-labels were generated to annotate the unlabeled instances in the target domain and the model was retrained with both training instances and pseudo-labeled testing instances [5]. Zheng et al. proposed an instance-based discriminative loss for rolling bearing diagnosis under a more practical and challenging scenario, in which only normal samples could be available in the dataset of the target machine [6].
2) Model-based approaches aim at sharing the neural network structures and parameters across different domains. Zhang et al. proposed a parameter transfer model based on CNN to diagnose motor bearings under different working conditions [7]. Shao et al. proposed a pre-trained VGG-16 network to extract lower-level features and successfully transferred the fault diagnosis knowledge on both motor bearings and gearboxes [8].
3) Feature-based approaches endow the diagnosis model the ability to transfer knowledge by learning domain invariant features. In the aspect of discrepancy-based metrics, Yang et al. proposed a multi-layer MMD to optimize the CNN-based diagnosis model, which transferred diagnosis knowledge from laboratory-used bearings to the locomotive bearings [9]. Jia et al. proposed a diverse feature aggregation module to enhance feature extraction capability across large domain gaps and the joint maximum mean discrepancy was designed to diminish the distribution discrepancy [10]. In the aspect of adversarial-based mechanism, Li et al. employed the adversarial strategy to deal with the machine fault diagnosis issue with imbalanced data [11]. Deng et al. proposed a double-layer adversarial domain adaptation network combined with attention mechanism to achieve diagnosis knowledge transfer across different machines [12]. In the aspect of reconstruction-based models, Wen et al. designed a deep transfer learning (DTL) to extract deep features with an auto-encode model and employed maximum mean discrepancy (MMD) as the metric to reduce domain discrepancy [13].

The CSDT approaches including transfer diagnosis knowledge across different working conditions, different sensor locations, and different types of components focus on addressing the issue of distribution matching, and they offer scope for extending the diagnosis transfer application under open set scenario at the same time.

2.2. Open set diagnosis transfer

The open set diagnosis transfer aims at detecting the emerging unknown fault classes from the target domain while correctly transferring diagnosis knowledge for known fault classes from the source domain to the target domain. In recent years, several approaches have been proposed to address the open set domain adaptation issues in computer vision fields, such as OSBP [15], STA [16], and CMU [17]. There are also some researchers making exploratory works for industrial applications. Li et al. first proposed a deep adversarial transfer learning network (DATLN), which is motivated by the OSBP approach and has been successfully applied into the rotating machinery emerging fault detection [18]. The DATLN formulates the decision boundary through an adversarial binary classifier, in which the unknown fault would be recognized when the classification probability of corresponding samples exceeds the fixed threshold (usually set as 0.5). Zhang et al. develop an instance-level weighted adversarial network to solve the OSDT issue [19]. The instance-level weights of target-domain samples are proposed to describe the similarities with the source classes, which are derived from the domain discriminator scores, and the entropy minimization is applied to enhance the target outlier detection. Zhu et al. proposed an adversarial network with multiple auxiliary classifiers (ANAMC), which develops multiple classifiers to calibrate the threshold set in the OSBP approach [20]. These auxiliary classifiers could better utilize the domain knowledge with representative weights, which assist to formulate an accurate decision bound more flexibly instead of forcing the whole target data under a certain fixed threshold to one category.

Reviewing recent literatures, the existing OSDT models lack essential theoretical analysis, thus omitting potential solutions for improvement and leading to a biased solution [21]. In this paper, the theoretical bound analysis is combined into OSDT model design to complete the blank, which endows the OSDT model with the capacity to detect the target outlier data newly occur with an unsupervised manner. Moreover, the progressive adversarial learning strategy is designed to suppress the interactive negative transfer under large degrees of openness and domain shifts.

3. Preliminaries

The definitions of the OSDT problem and some important concepts are introduced in this section. The notations used in this paper are summarized in Table 1.

3.1. OSDT problem definition

Assume that the labeled source data $\mathscr{D}_s = \left\{\left(x_{s_i}, y_{s_i}\right)\right\}_{i=1}^{n_s}$ $\mathbb{P}^s$ and unlabeled target data $\mathscr{D}_t = \left\{x_{t_j}\right\}_{j=1}^{n_t}$ ${\mathbb{Q}^t}_X$ are given, where $\mathbb{P}^s$ is the joint probability distribution of the source domain, ${\mathbb{Q}^t}_X$ is the mar-

Table 1
Notations and Their Descriptions.

Notation	Description	Notation	Description
$\mathscr{D}_s$	Source domain	C_s	Number of known classes
$\mathscr{D}_t$	Target domain	$\boldsymbol{G}$	Feature extractor
x_{s_i}	The i-th sample of source domain	$\boldsymbol{C}$	Multi-categories classifier
y_{s_i}	The i-th label of source domain	$\mathscr{H}$	Hypothesis space, set of classifiers $\boldsymbol{C}$
$x_{t_j,}$	The j-th sample of target domain	$R_s(\cdot), R_t(\cdot)$	Source, target risk
n_s, n_t	Source, target domain samples number	$\mathscr{L}(\cdot)$	Symmetric loss function
$\mathbb{P}^s$	Source domain joint distribution	π_i^s	The source i-th class-prior probability
$\mathbb{Q}^t$	Target domain joint distribution	π_i^t	The target i-th class-prior probability
${\mathbb{Q}^t}_X$	Target domain marginal distribution	$R_{s,i}(\boldsymbol{C}), R_{t,i}(\boldsymbol{C})$	Source, target partial risk of i-th class
$\mathscr{X}_s, \mathscr{X}_t$	Source, target sample space	$d_{\mathscr{H}}^{\ell}$	Discrepancy distance
$\mathscr{Y}^s, \mathscr{Y}^t$	Source, target label space	λ	Shared error
$\boldsymbol{y}_k$	Label vector of k-th class	$\Delta_{\boldsymbol{o}}$	Open set risk

ginal distribution of the target domain, with n_s and n_t indicating the number of source and target samples respectively. The goal of OSDT is to train an optimal target domain classifier $\boldsymbol{C}: \mathscr{X}_t \to \mathscr{Y}_t$ with drawing samples from both domains, which should meet the following requirements:

1) $\boldsymbol{C}$ should classify the known target samples into the corresponding known classes.
2) $\boldsymbol{C}$ should discriminatexzdqq4eles.

According to the definition of problem OSDT, the target domain classifier $\boldsymbol{C}: \mathscr{X}_t \to \mathscr{Y}_t$ only needs to detect unknown target data and classify other target data. It is not necessary to classify unknown target data, and all unknown target data are recognized as "unknown fault type". Correspondingly, the source label space and target label space can be defined as $\mathscr{Y}^s = \{\boldsymbol{y}_k\}_{k=1}^{C_s}$ and $\mathscr{Y}^t = \{\boldsymbol{y}_k\}_{k=1}^{C_s+1}$ respectively, where the label $\boldsymbol{y}_k$ denotes the k-th class and $\boldsymbol{y}_{C_s+1}$ denotes the unknown fault class.

3.2. Theoretical upper bound

Given the hypothesis space $\mathscr{H}$ with a mild condition that constant function $C_s + 1 \in \mathscr{H}$, for $\forall \boldsymbol{C} \in \mathscr{H}$, the expected risk on target samples $R_t(\boldsymbol{C})$ can be bounded as:

$$\frac{R_t(\boldsymbol{C})}{1-\pi_{C_s+1}^t} \le R_s(\boldsymbol{C}) + d_H^l\left(Q_{X|Y\le C_s}^t, P_X^s\right) + \lambda + \frac{R_{t,C_s+1}(\boldsymbol{C})}{1-\pi_{C_s+1}^t} - R_{s,C_s+1}(\boldsymbol{C}) \tag{1}$$

where λ is a constant called as the shared error [22].

As shown in equation (1), the upper error of classifier $\forall \boldsymbol{C} \in \mathscr{H}$ on the target domain is bounded by four terms. Correspondingly, the optimization of the diagnosis model for the OSDT problem could be carried out under the guidance of the following four aspects.

1) **Source risk** $R_s(\boldsymbol{C})$: The source risk represents the classification loss on the known data. According to the assumption of OSDT, the source domain does not include any unknown samples, thus the source risk could be bounded easily by only minimizing the classification error on the source labeled data.
2) **Discrepancy distance** $d_H^l\left(Q_{X|Y\le C_s}^t, P_X^s\right)$: The discrepancy distance represents the divergence across different domains at the feature level. In order to learn task-sensitive but domain-insensitive features, the generative adversarial mechanism is conducted to encourage domain confusion. The optimization objective to train the generator $\boldsymbol{G}$ (to generate domain invariant features of diverse domains) and the discriminator $\boldsymbol{D}$ (to enhance the discriminative architectures) is expressed as:

$$\min{}_G \max{}_D E_{x-D_s}[\log D(G(x))] + E_{x-D_t}[\log(1-D(G(x)))] \tag{2}$$

where $x - \mathscr{D}_s$ and $x - \mathscr{D}_t$ are respectively the probability density function of the source and target domain samples [23].

3) **Shared error** λ: The constant $\lambda = \min\limits_{C\in H} \frac{R_t^*(C)}{1-\pi_{C_s+1}^t} + R_s(C)$, where $R_t^*(\boldsymbol{C})$ indicates the partial risk of classifier $\boldsymbol{C}$ on known target samples. λ tends to be large when the conditional shift is encountered, where the class-wise conditional distributions are not aligned even with marginal distribution aligned [24].
4) **Open set risk** $\Delta_{\boldsymbol{o}} = \frac{R_{t,C_s+1}(\boldsymbol{C})}{1-\pi_{C_s+1}^t} - R_{s,C_s+1}(\boldsymbol{C})$: The open set risk $\Delta_{\boldsymbol{o}}$ is designed to estimate the classification loss on the unknown data [24]. It can be seen that $\Delta_{\boldsymbol{o}}$ consists of two parts: the positive term $R_{t,C_s+1}(\boldsymbol{C})$ and the negative term $R_{s,C_s+1}(\boldsymbol{C})$. The larger value of positive part $R_{t,C_s+1}(\boldsymbol{C})$ indicates that more target samples are recognized as unknown and the negative part $R_{s,C_s+1}(\boldsymbol{C})$ prevents the source samples from being recognized as unknown. Since the source samples are labeled, the negative part $R_{s,C_s+1}(\boldsymbol{C})$ could be eliminated, and the value of $\Delta_{\boldsymbol{o}}$ is contributed by two parts: the openness $\pi_{C_s+1}^t$ (the proportion of unknown samples in the target domain) and the partial risk on unknown target samples $R_{t,C_s+1}(\boldsymbol{C})$.

4. Methodology

4.1. Overview of TPTLN

The proposed TPTLN is illustrated in Fig. 3, which mainly includes three modules: feature extraction, distract stage, and attract stage. Firstly, in the feature extraction stage, the frequency spectrums of mechanical vibration signals are fed into the feature generator $\boldsymbol{G}$ to extract the hierarchical features f_s and f_t. Secondly, in the distract stage, the theoretically derived source risk term and the open set risk term are estimated through uncertainty calibration and domain consensus learning strategy. Subsequently, in the attract stage, the adversarial based weighting mechanism is proposed to minimize the discrepancy distance term of shared label space by assigning larger weights to target known classes and alleviate the negative transfer from the private space by assigning fewer weights to the target unknown classes. Finally, the progressively learning strategy between the distract stage and attract stage could enable the TPTLN to learn a tighter theoretical bound for detecting the unknown fault classes in the target

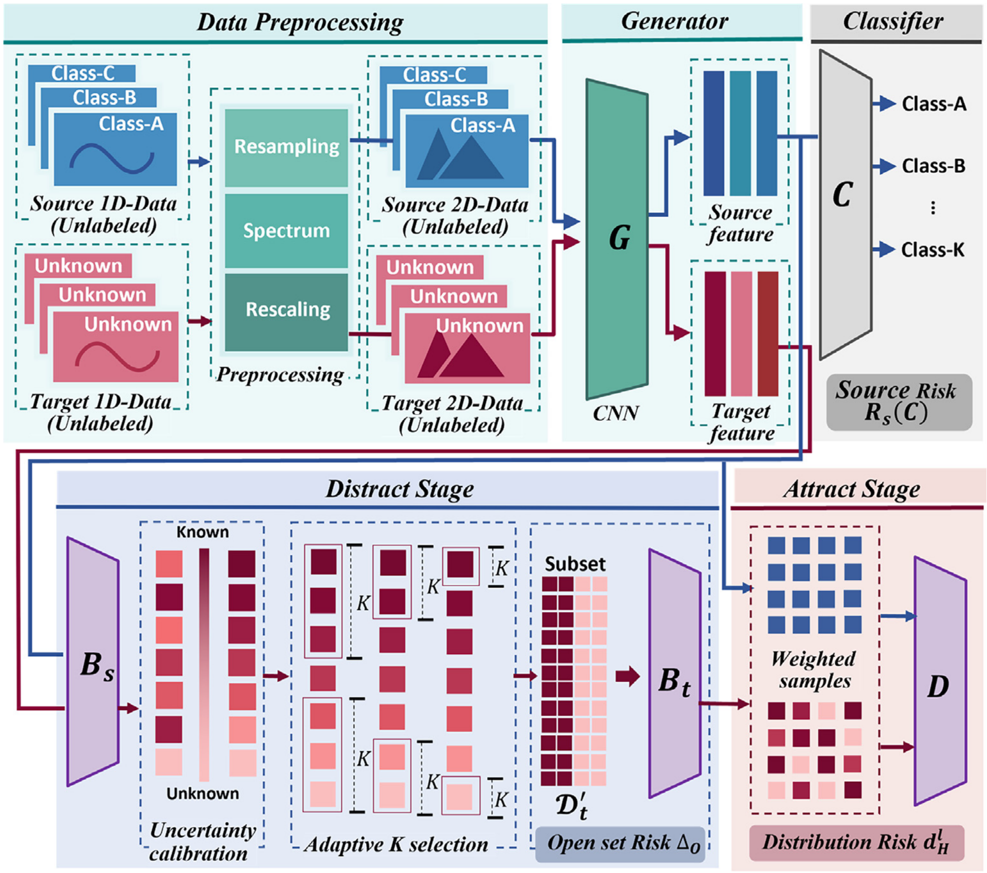


Fig. 3. The detailed architecture of proposed TPTLN.

domain and to build a more robust classifier for diagnosing the known fault classes cross domains.

4.2. Distract stage

In this section, detailed descriptions of the distract stage in the progressive learning framework are given. As the name suggests, the distract stage mainly focuses on distracting those unknown fault classes in the unlabeled target domains from known classes, which could alleviate the interactive negative transfer caused by wrongly matching the distributions of source known samples and whole target samples. To carry out the distract stage theoretically, the aforementioned source risk $R_s(\mathbf{C})$ and open-set risk $\Delta_{\mathbf{o}}$ are employed to guide the model optimization.

4.2.1. Source risk $\mathbf{R_s}(\mathbf{C})$ optimization

The source risk ensures that the cross-domain classifier could accurately classify the known fault categories of the source domain. To achieve this, a multi-categories classifier $\mathbf{C}$ is defined to calculate the source risk, the source risk $R_s(\mathbf{C})$ is given as follows:

$$R_s(\mathbf{C}) = L_{cls} = \frac{1}{n^s}\sum_{\mathbf{x}_i\in\mathscr{D}_s} L_{ce}\left(\mathbf{C}^{1:|C_s|}(\mathbf{G}(\mathbf{x}_i)),\mathbf{y}_i^s\right) \tag{3}$$

where the loss function is cross-entropy loss, which is commonly used for multi-categories classification function, $\mathbf{C}$ is a general classifier for $C_s + 1$ categories, i.e., the $|C_s|$ known classes in the source domain plus the additional unknown class in the target domain. $\mathbf{G}(\mathbf{x}_i)|\mathbf{x}_i \in \mathscr{D}_s$ denotes the extracted features from the source domain and $\mathbf{C}^{1:|C_s|}(\mathbf{G}(\mathbf{x}_i))$ indicates the probabilities of each sample to the corresponding $|C_s|$ known classes.$\mathbf{y}_i^s$ is the ground truth label information of each source sample.

4.2.2. Open-set risk $\Delta_{\mathbf{o}}$ optimization

The open-set risk indicates the model's capability of dividing the unknown target classes and known target classes, and it should be noticed that the openness $\pi_{C_s+1}^t$ should be first evaluated accurately. If the openness is too large, indicating most percentage data in the target domain is unknown fault type $(\pi_{C_s+1}^t \to 1)$, this term would dominate the open set risk $\Delta_{\mathbf{o}}$ and contribute most to the whole upper bound. On the other hand, if $\pi_{C_s+1}^t = 0$, which means there has no unknown target samples, the whole upper error bound would be same as CSDT problem. In the previous studies [21], the unknown parameter openness $\pi_{C_s+1}^t$ in $\Delta_{\mathbf{o}}$ is evaluated tentatively, such as aggressively setting a large value to $\pi_{C_s+1}^t$, which could suffer fluctuation under different degrees of domain shifts. To estimate the openness accurately, a coarse-to-fine discriminator is employed in this paper, which is employed to learn an adaptive openness boundary between known and unknown target samples. The designed discriminator includes a source discriminator $\mathbf{B_s}$ for coarse separation and a target discriminator $\mathbf{B_t}$ for fine separation.

4.2.2.1. Source discriminator $\mathbf{B_s}$ based on uncertainty calibration. A multi-binary discriminator $\mathbf{B_s}$ is designed to achieve a coarse separation only with source labeled samples. The loss of proposed source-only discriminator $\mathbf{B_s}$ can be optimized as follows:

$$L_s = \sum_{c=1}^{|C_s|}\frac{1}{n^s}\sum_{\mathbf{x}_i\in\mathscr{D}_s} L_{bce}\left(\mathbf{B_s}(\mathbf{G}(\mathbf{x}_i)),\mathbf{I}(\mathbf{y}_i^s,c)\right) \tag{4}$$

where L_{bce} is the binary cross-entropy loss, and there are totally $|C_s|$ source-only discriminators in $\boldsymbol{B_s}$. The labelling function is defined as $\boldsymbol{I}(\boldsymbol{y}_i^s, c) = 1 if \boldsymbol{y}_i^s = c$ and $\boldsymbol{I}(\boldsymbol{y}_i^s, c) = 0$ otherwise.

For each target feature $\boldsymbol{G}(\boldsymbol{x}_i) \in \mathscr{D}_t$ fed into $\boldsymbol{B_s}$, the output probability p_c from each binary discriminator could be seen as the prediction confidence according to the respective source class $|C_s|$, and the source-only discriminator would output a probability vector $\left[p_1, p_2, \cdots p_{|C_s|}\right]$. Since the target data of known classes prone to have higher probabilities in one of the shared spaces than target data of unknown classes, the maximum probability in the vector $\left[p_1, p_2, \cdots p_{|C_s|}\right]$ is always used as the similarity s_j^t between the j_{th} target sample and known class in previous studies [16]:

$$s_j^t = max\left(\boldsymbol{B_s}\left(\boldsymbol{G}\left(\boldsymbol{x}_{t_j}\right)\right)|c = 1, 2, \cdots, |C_s|\right) \tag{5}$$

The prediction confidence could perform well for the similarity estimation for target data from the known label space because these data would share similar structures which have occurred in the source domain. However, for target data from the unknown label space, it is hard to learn the structure-specific features which have never occurred in the source domain and lead to weak discriminability and biased similarity estimation.

In order to overcome this problem, a self-supervised uncertainty calibration technique called as entropy analysis is combined into similarity evaluation. The entropy is used to measure the smoothness of the class distribution, and a larger entropy indicates the class distribution has higher uncertainty.

Thus, the calibrated similarity with entropy analysis can be expressed as:

$$s_j^t = \frac{1}{2} \times \left(\underbrace{\max\left(B_s\left(G\left(x_{t_j}\right)\right)|c = 1, 2, \ldots, |C_s|\right)}_{confidence\ \ term} + 1 - \underbrace{\left(\sum_{j=1}^{|C^s|} - p_j^t \log\left(p_j^t\right)\right) / \log\left(|C^s|\right)}_{entropy\ \ term} \right) \tag{6}$$

From equation (6), the calibrated similarity is composed of confidence and entropy, which are complementary to discriminate different degrees of uncertainty clearly and provide more accurate estimation, the detailed explanation is given as follows:

1) If the target data are from the known classes, the confidence term would be larger because the target feature may have similar structures that occurred in the source domain and the entropy term would be smaller because the predictions tend to be sharp without uncertainty. Thus, the entropy term could increase the similarity between target data and their corresponding known categories.
2) If the target data are from the unknown classes, it is hard to obtain accurate similarity only from the confidence term. The target unknown data may have structure-specific features which have never been captured before, therefore the similarity from prediction confidence would be uncertain. The entropy term would be larger when the prediction distributions are uncertain, which would decrease the similarity between the target data and all known categories.

4.2.2.2. Target discriminator $\boldsymbol{B_t}$ based on the consensus score. The target discriminator $\boldsymbol{B_t}$ is expected to tune the coarse openness boundary learned by $\boldsymbol{B_s}$ to a more-fine boundary, aims at further separating the unknown and known target samples. To achieve this, the similarity s_j^t of all target domain samples are ranked in descending order, and top-K samples with higher similarity and bottom-K samples with lower similarity would be chosen to build a subset $\mathscr{D}_t'$. It should be noticed an unsuitable K will cause the unexpected model degeneration: too small K would cause some unknown samples to be ignored by the subset $\mathscr{D}_t'$, and too large K would allow some irrelevant samples to be included by the subset $\mathscr{D}_t'$. Therefore, the value of K should be increased or decreased to expand or contract the discriminative bound of the subset according to the degree of openness variance and batch size.

In this paper, an adaptive K initialization approach based on domain consensus score (DCS) is introduced. The domain consensus score aims at drawing the cross-domain knowledge to facilitate the shared class clustering and private class discovery [25]. The definition of domain consensus score is illustrated in Fig. 4. Given a pair of matched clusters $\{v_i^s\}_{i=1}^c$ and $\{v_i^t\}_{i=1}^n$ with corresponding centers μ_c^s and μ_l^t. From the source view, the similarity with all target cluster centers $\{\mu_1^t, \cdots, \mu_L^t\}$ is calculated as:

$$r_{i,l}^s = Sim\left(v_i^s, \mu_l^t\right), l = \{1, \cdots, L\} \tag{7}$$

where $Sim(\cdot)$ denotes the cosine similarity. Then the source consensus score could be formulated as the proportion of samples reach consensus:

$$S_{(c,l)}^s = \frac{\sum_{i=1}^m 1\left\{\arg\max_l\left(r_{i,l}^s\right) = l\right\}}{m} \tag{8}$$

where $1\left\{\arg\max_l\left(r_{i,l}^s\right) = l\right\}$ is indicated to judge whether v_i^s holds corresponding cluster index (l) across domains. Analogously, the similarity with all source cluster centers could be obtained as equation (8) from the target view. And the target consensus score could be formulated as:

$$S_{(c,l)}^t = \frac{\sum_{i=1}^n 1\left\{\arg\max_c\left(r_{i,c}^t\right) = c\right\}}{n} \tag{9}$$

Then the consensus score of this matched pair could be written as: $S_{(c,l)} = \frac{S_{(c,l)}^s + S_{(c,l)}^t}{2}$.

Since the number of underlying target classes L is unknown, multiple target clusters with different L would be used during the domain consensus score calculation. And the instantiation of L with the highest score is chosen for initializing K, which is given as:

$$K = \frac{\sum_{i=1}^C N_i^t \times S_{(c,i)}}{N^t} \times batchsize \tag{10}$$

where N_i^t is the number of target samples clustered by the paired cluster v_i^t, and $S_{(c,i)}$ indicates the domain consensus score of the corresponding matched pair, and N^t is the number of all target domain samples.

Compared with the previous research using threshold-based method [15,17] or fixing the value of K [16] for detection of unknown target samples, the proposed adaptive K selection approach provides a more robust way to train the known and unknown target samples separation under different levels of domain shifts, since the discriminative bound of the subset could be adjusted flexibly.

Based on the filtered target subset $\mathscr{D}_t'$, the target domain discriminator $\boldsymbol{B_t}$ can be optimized as follows:

$$L_t = \frac{1}{n_t'} \sum_{\boldsymbol{x}_j \in \mathscr{D}_t'} L_{bce}\left(\boldsymbol{B_t}(\boldsymbol{G}(\boldsymbol{x}_i)), d_j\right) \tag{11}$$

where L_{bce} is a binary cross-entropy loss, and d_j indicates whether the target samples of the filtered subset $\mathscr{D}_t'$ is known $(d_j = 0)$ or unknown $(d_j = 1)$.

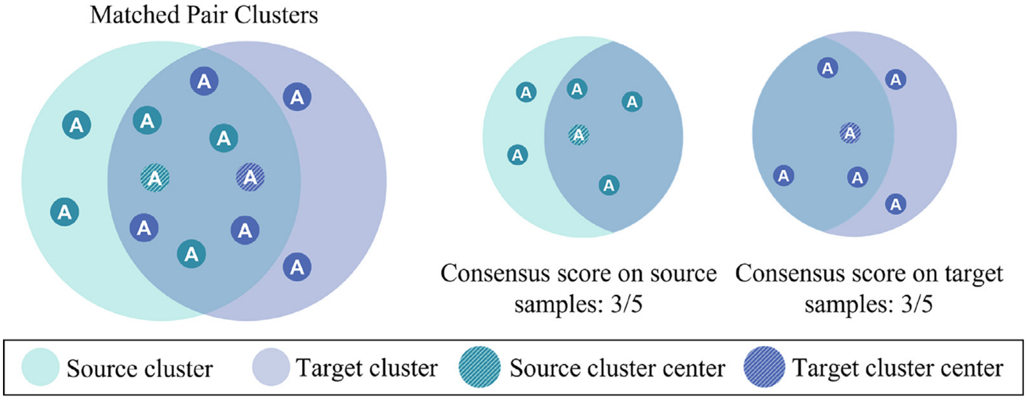


Fig. 4. Illustration of domain consensus score.

With the introduced $\boldsymbol{B_s}$ (providing the similarity of each target sample for the coarse separation) and $\boldsymbol{B_t}$ (providing the probability of each target sample being known or unknown fault classes for the fine separation), an adaptive decision boundary of target samples could be obtained gradually, and the unknown parameter openness $\pi^t_{C_s+1}$ of open-set risk could be estimated as follows:

$$\pi^t_{C_s+1} = \frac{1}{n^t}\sum_{\boldsymbol{x}_j\in\mathscr{D}_t} 1 - \boldsymbol{B_t}\left(\boldsymbol{G}\left(\boldsymbol{x}_j\right)\right) \tag{12}$$

After estimating the openness $\pi^t_{C_s+1}$ by the coarse-to-fine discriminator, the open-set risk $\Delta_{\boldsymbol{o}}$ could be obtained and the optimization function is given as follows:

$$\Delta_{\boldsymbol{o}} = \frac{\pi^t_{C_s+1}}{1-\pi^t_{C_s+1}}\sum_{\boldsymbol{x}_j\in\mathscr{D}_t} L_{mse}\left(\boldsymbol{C}^{|C_s|+1}\left(\boldsymbol{G}\left(\boldsymbol{x}_j\right)\right), \boldsymbol{y}_{|C_s|+1}\right) \tag{13}$$

4.3. Attract stage

In the attract stage, the distribution discrepancy between the source domain and target domain should be reduced to achieve the learning domain-invariant features. It should be noticed that only the shared space of the source domain and target domain need to be aligned. Therefore, a weighted distribution risk optimization strategy is proposed to promote the alignment of source classes and target known classes, and to suppress the unexpected alignment of source classes and target unknown classes.

4.3.1. Weighted distribution risk optimization

In order to assign different weights to each target sample, the output of target discriminator $\boldsymbol{B_t}$ is used as a soft instance-level weight $w_j = \boldsymbol{B_t}\left(\boldsymbol{G}\left(\boldsymbol{x}_j\right)\right), \boldsymbol{x}_j \in \mathscr{D}_t$. Based on the definition of target known/unknown discriminator $\boldsymbol{B_t}$, a larger w_j indicates that the sample has higher probabilities of being from the shared space (known classes) and should be paid more attention during the distribution risk minimization, while a smaller w_j implies that the sample has a higher probability of being from the private space (unknown classes) and should be suppressed during the distribution risk minimization. Correspondingly, the target sample weights $w_{j}{}_{j=1}^{n_t}$ is exploited for the adversarial domain adaptation process, and the weighted distribution risk optimization function can be obtained as follows:

$$L_d = \frac{1}{n^s}\sum_{\boldsymbol{x}_i\in\mathscr{D}_s} L_{bce}(\boldsymbol{D}(\boldsymbol{G}(\boldsymbol{x}_i)), d_i) - \frac{1}{\sum_{\boldsymbol{x}_j\in\mathscr{D}_t} w_j} \times \sum_{\boldsymbol{x}_j\in\mathscr{D}_t} (1 - w_j L_{bce}(\boldsymbol{D}(\boldsymbol{G}(\boldsymbol{x}_j)), d_j)) \tag{14}$$

where the domain adversarial loss L_d aims at minimizing over $\boldsymbol{D}$ and maximizing over $\boldsymbol{G}$. The domain discriminator $\boldsymbol{D}$ is trained to identify whether the input features are from the source domain or target domain, while the feature generator $\boldsymbol{G}$ is expected to confuse the discriminator $\boldsymbol{D}$ through extracting the domain-invariant features. Due to the weighting mechanism, target samples with higher probabilities of being known classes would dominate the adversarial process compared with unknown target samples, which means that the distribution alignment will be concentrated on target known samples and source samples iteratively during the minimax game.

4.4. Training procedure

The training procedure of proposed TPTLN is divided into two steps, which include the distract stage of detecting the unknown fault classes from target domain and attract stage of aligning the target known data with source data. The two separate stages are alternated progressively to minimize the theoretical bound of OSDT issue and to effectively suppress the interactive negative transfer problem.

Step 1. Distract stage training

In the first step, the feature generator $\boldsymbol{G}$ and classifier $\boldsymbol{C}$ are trained to accurately classify the source fault classes. Subsequently, a source-only discriminator $\boldsymbol{B_s}$ and a target-only discriminator $\boldsymbol{B_t}$ is trained to build a coarse-to-fine decision boundary to discriminate the known and unknown fault classes in the target domain. Denote by $\theta_f, \theta_y, \theta_t$ and $\theta^c_s\big|^{|C_s|}_{c=1}$ the parameters of the feature generator $\boldsymbol{G}$, the classifier $\boldsymbol{C}$, the target-only discriminator $\boldsymbol{B_t}$ and the source-only discriminator $\boldsymbol{B_s}$, and the optimal parameters $\widehat{\theta}_f, \widehat{\theta}_y, \widehat{\theta}_t$ and $\widehat{\theta}^c_s\Big|^{|C_s|}_{c=1}$ can be obtained as follows:

$$\left(\widehat{\theta}_f, \widehat{\theta}_y, \widehat{\theta}_t, \widehat{\theta}^c_s\Big|^{|C_s|}_{c=1}\right) = \underset{\theta_f,\theta_y,\theta_t,\theta^c_s|^{|C_s|}_{c=1}}{\operatorname{argmin}}\ L_{cls} + \alpha_s L_s + \alpha_t L_t + \alpha_o \Delta_{\boldsymbol{o}} \tag{15}$$

Step 2. Attract stage training

In the second step, the feature generator $\boldsymbol{G}$ and domain discriminator $\boldsymbol{D}$ are trained in an adversarial way to conduct distribution alignment, in which only the target known fault classes would be attracted with source classes. Denote by θ_d the parameters of the domain discriminator $\boldsymbol{D}$, the optimal parameters of $\widehat{\theta}_f, \widehat{\theta}_y, \widehat{\theta}_d$ can be given as:

$$\left(\widehat{\theta}_y, \widehat{\theta}_d\right) = \underset{\theta_y, \theta_d}{\operatorname{argmin}} L_{cls} + \alpha_d L_d \tag{16}$$

$$\left(\widehat{\theta}_f\right) = \underset{\theta_f}{\operatorname{argmin}} L_{cls} - \alpha_d L_d \tag{17}$$

With the proposed TPTLN model, the theoretical error upper bound of OSDT issue can be well optimized and the pending problem of negative transfer caused by the unknown fault classes could be effectively suppressed. The detailed algorithm of TPTLN is given in Table 2, it should be noticed that the distract stage and attract stage could benefit each other through the proposed progressively training strategy between two steps. Step 1 performs a coarse-to-fine way to discriminate unknown target data and estimate the openness adaptively, which could suppress the unexpected alignment of unknown classes and better facilitate the alignment of known classes in step 2. Step 2 performs an adversarial way to conduct domain alignment, which helps building the decision boundary of unknown data in step 1 more accurately.

5. Experimental case study

5.1. Compared methods and evaluation metrics

Totally seven representative deep transfer learning approaches are selected as the comparative methods and introduced as follows:

1) **FTNN (Feature-based Transfer Neural Network)** [9] aims at extracting the transferable features through multi-layer domain adaptation and pseudo-label learning. The maximum mean discrepancy (MMD) across different layers of the CNN network is minimized to align the distributions in source and target domains.
2) **DCTLN (Deep Convolutional Transfer Learning Network)** [11] employs an adversarial way to transfer diagnosis knowledge across different machines. The domain adversarial loss and the MMD loss are combined to minimize the distribution discrepancy between the source domain and target domain together.
3) **OSBP (Open Set Domain Adaptation by Backpropagation)** [15] aims at constructing a decision boundary to detect the unknown target samples. A classifier is trained to make a boundary between the source and the target samples whereas a generator is trained to make target samples far from the boundary.
4) **STA (Separate to Adapt)** [16] employs a two-stage network structure to solve the open-set transfer learning problem, in which the first stage is trained to discriminate the target unknown data by exploring the source data, and the second stage is trained to adapt the distributions in the scope of known classes.
5) **CMU (Calibrated Multiple Uncertainties)** [17] proposes a novel transferability measure to detect the outlier data, which is estimated by a mixture of uncertainty quantities in complementation: entropy, confidence, and consistency, defined on conditional probabilities calibrated by a multi-classifier ensemble model.
6) **DATLN (Deep Adversarial Transfer Learning Network)** [18] first discusses the open set setting for transferring diagnosis knowledge in industrial scenarios. An adversarial classifier is designed to align the samples (with the known classes) in both source and target domains and to detect samples with the unknown classes.
7) **IW-OSDA (Instance-Level Weighted Open-set Domain Adaptation)** [19] combines an outlier classifier into the adversarial-based network to enhance the unknown fault samples detection for OSDT problem. The instance weight obtained from the domain discrepancy is developed to describe the similarity of target samples with the source classes.

Table 2
Details of the training procedure.

```
Algorithm: Training procedure of proposed method

Input: source samples ${x_i^s, y_i^s}_{i=1}^{n^s}$, target samples ${x_i^s}_{i=1}^{n^t}$
Parameter: learning rate $\gamma$, batch size $m$, the number of iterations $T$,
    network parameters $\theta_f, \theta_y, \theta_t, \theta_s^c|_{c=1}^{|C_s|}, \theta_d$.
Output: predicted target label $\widehat{y}_t$.
1: Initialize $\theta_f, \theta_y, \theta_t, \theta_s^c|_{c=1}^{|C_s|}, \theta_d, \alpha_s, \alpha_t, \alpha_o, \alpha_d$
Initialization stage
2: Sample source minibatch $\left\{\left(x_{i_1}^s, y_{i_1}^s\right), \cdots, \left(x_{i_m}^s, y_{i_m}^s\right)\right\}$
3: Sample target minibatch $\left\{x_{i_1}^t, \cdots, x_{i_m}^t\right\}$.
4: Initialize $L_{cls}, L_s$ according to Eqs. (5), (6).
5: Initialize the Top-K value according to Eqs. (7)–(12)
Distract stage
6: $t = 0$
7: while $t < T$ do
8: Calculate $L_t$ according to Eqs. (13)
9: Calculate $\Delta_o$ according to Eqs. (14) and (15)
10: Update parameter: $\widehat{\theta}_f, \widehat{\theta}_y, \widehat{\theta}_t$ and $\widehat{\theta}_s^c\Big|_{c=1}^{|C_s|}$
$\left(\widehat{\theta}_f, \widehat{\theta}_y, \widehat{\theta}_t, \widehat{\theta}_s^c\Big|_{c=1}^{|C_s|}\right) = \underset{\theta_f, \theta_y, \theta_t, \theta_s^c|_{c=1}^{|C_s|}}{\operatorname{arg}min} L_{cls} + \alpha_s L_s + \alpha_t L_t + \alpha_o \Delta_o$
Attract stage
11: Calculate $L_d$ according to Eqs. (16) by leveraging weighted target samples.
12: Update parameter:$\theta_f, \theta_y, \theta_d$
$\left(\widehat{\theta}_y, \widehat{\theta}_d\right) = \underset{\theta_y, \theta_d}{\operatorname{argmin}} L_{cls} + \alpha_d L_d$
$\left(\widehat{\theta}_f\right) = \underset{\theta_f}{\operatorname{argmin}} L_{cls} - \alpha_d L_d$
13:$t = t + 1$
14: end while
```

To comprehensively evaluate the proposed TPTLN and other baseline methods, two widely used evaluation metrics, normalized accuracy for all classes (OS) and normalized accuracy for the known classes only (OS*), are given as follows:

$$\begin{aligned} OS &= \frac{1}{C_s + 1} \sum_{c=1}^{C_s+1} \frac{\left|x : x \in D_t^c \wedge C(G(x)) = c\right|}{\left|x : x \in D_t^c\right|} OS^* \\ &= \frac{1}{C_s} \sum_{c=1}^{C_s} \frac{\left|x : x \in D_t^c \wedge C(G(x)) = c\right|}{\left|x : x \in D_t^c\right|} \end{aligned} \tag{18}$$

where $\mathscr{D}_t^c$ denotes target samples belonging to the c-th fault class, and $\boldsymbol{C}(\boldsymbol{G}(x)) = c$ indicates that classifier $\boldsymbol{C}$ correctly assign the sample x to the corresponding category.

5.2. Dataset description

5.2.1. PU bearing dataset

The rolling bearing dataset is acquired from the Paderborn University which consists of bearing artificial faults and real damages [26]. Vibration signals of the bearing housing were collected by a piezoelectric accelerometer with a sampling frequency as 64 kHz. By changing the rotational speed of the drive system, the radial force onto the test bearing, and the load torque on the drive train, different working conditions could be performed. As shown in Table 3, eight different bearings with real damages caused by the accelerated lifetime tests are selected as the PU dataset.

The frequency spectrums of Paderborn bearing vibration signal are selected as the model input, which is illustrated in Fig. 5. It can be found that the fault characteristics of different health conditions may have shifts and divergence across different working conditions. Thus, the first challenge for the diagnosis model is to transfer the domain-invariant knowledge across different working conditions. Furthermore, the source domain and target domain have different label spaces in the OSDT scenario, which means that some emerging fault categories would occur during the testing stage. Therefore, the second challenge is how to discriminate these unknown data in the target domain. For example, if the source domain does not have IRF samples, the challenge is how to avoid misclassifying target data of IRF as the known faults, since the source known samples may share similar frequency structures as the unknown IRF samples.

5.2.2. PHM 2009 gearbox dataset

The planet gearbox dataset is acquired from PHM 2009 data challenge, which contains 3 shafts, 4 gears, and 6 bearings [27]. Two sets of gears including spur gears and helical gears with different fault types are tested. The dataset is comprised of 2 channels of accelerometer signals and 1 channel of tachometer signal acquired by corresponding sensors. Signals were collected for each health condition with a sampling frequency of 66.67 kHz and an acquisition time of 4 s. Totally 6 different health conditions of the planet gearbox are selected, and the detailed description is given in Table 4.

5.3. OSDT task and implementation details

5.3.1. Descriptions of OSDT tasks

In this study, seven different open set diagnosis transfer tasks based on bearing and gearbox datasets are investigated. Specifically, the detailed settings of bearing OSDT tasks and gearbox OSDT tasks are shown in Table 5 and Table 6 respectively.

Based on the selected PU bearing dataset, four OSDT tasks are designed, which are given in Table 5. The openness gap between the source domain and the target domain is gradually larger, which

Table 3
Paderborn university accelerating life test bearing dataset specification.

Health Label	Element	Specification	Type	Working conditions
1-OSF	Outer Ring	Fatigue pitting	S	Speed: 900 rpm/1500 rpm
2-OSP	Outer Ring	Plastic deformation	S	Load torque: 0.7 Nm/0.1 Nm
3-ORF	Outer Ring	Fatigue pitting	R	Radial force: 1000 N/400 N
4-ISF	Inner Ring	Fatigue pitting	S	
5-IRF	Inner Ring	Fatigue pitting	R	
6-IORF	Inner & Outer Ring	Fatigue pitting	R	
7-IORP	Inner & Outer Ring	Plastic deformation	R	
8-H	Health	/	/	

O: Outer ring fault; I: Inner ring fault; IO: Inner ring & Outer ring fault.
S: Single fault; R:Repetitve fault; F: Fatigue pitting; P: Plastic deformation.

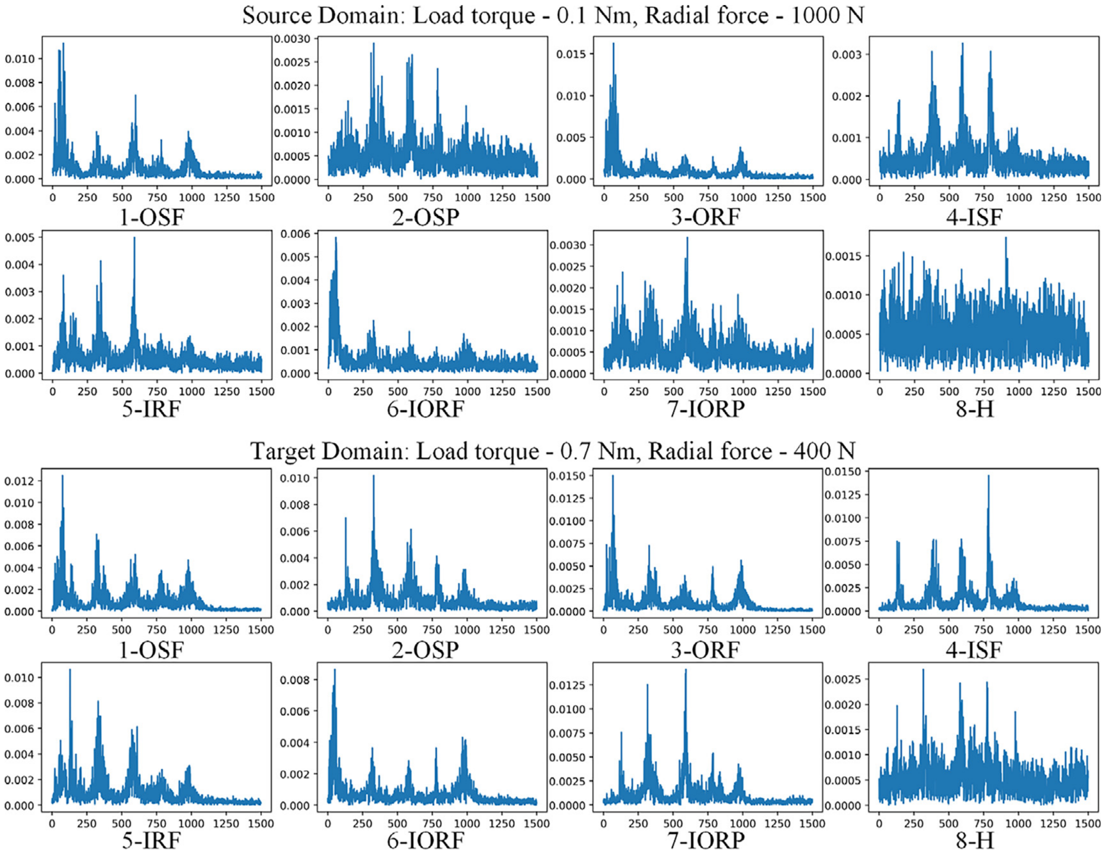


Fig. 5. The frequency spectrums of Paderborn bearing data across different domains.

Table 4
2009 PHM gearbox dataset specification.

Health Label	Element	Fault specification	Working condition
1-N	/	Health	Shaft speed: 30/35/40/45/50 Hz
2-G	Helical	Gear chipped	Load: Low/High
3-BF&SI	Helical	Bearing combination & ball fault, shaft imbalance	
4-G&E	Spur	Gear chipped & eccentric	
5-BI&SK	Spur	Bearing inner fault, shaft keyway sheared	
6-BF&BO&SI	Spur	Bearing ball & outer fault, shaft imbalance	

Table 5
Detailed descriptions of designed bearing OSDT tasks.

Task No.	Source label	Target label	Sample number	Openness ($\mathbb{O}$)
T1	1,2,3,4,6,7,8	1,2,3,4, 5*,6,7,8	Source:700, Target:800	0.125
T2	1,3,4,6,7,8	1,2*,3,4,5*,6,7,8	Source:600, Target:800	0.25
T3	1,2,3,4,5	1,2,3,4,5,6*,7*,8*	Source:500, Target:800	0.375
T4	1,4,6,8	1,4, 5*,6,8	Source:400, Target:800	0.5

Source domain working condition: Load torque: 0.1 Nm, Radial force: 1000 N.
Target domain working condition: Load torque: 0.7 Nm, Radial force: 400 N.

Table 6
Detailed descriptions of designed gearbox OSDT tasks.

Task No.	Source label	Target label	Sample number	Openness($\mathbb{O}$)
T1	1,3,4	1,3,4,6*	Source:400, Target:400	0.25
T2	1,3,4	1,3,4, 5*, 6*	Source:500, Target:500	0.4
T3	1,3,4	1,2*,3, 4,5*,6*	Source:600, Target:600	0.5

Source domain working condition: Load torque: Low, Shaft speed: 40 Hz.
Target domain working condition: Load torque: High, Radial force: 45 Hz.

is expected to evaluate the robustness and accurateness of all methods under different degrees of openness comprehensively.

Based on the selected gearbox dataset, three open-set diagnosis transfer learning tasks are designed, which is given in Table 6. The openness gap between the source domain and the target domain are gradually larger. Moreover, the source samples and target samples are imbalanced across different health states, and the detailed samples distributions of the source domain and target domain among three tasks are illustrated as Fig. 6. The gearbox OSDT tasks not only evaluate the transferability of compared models under

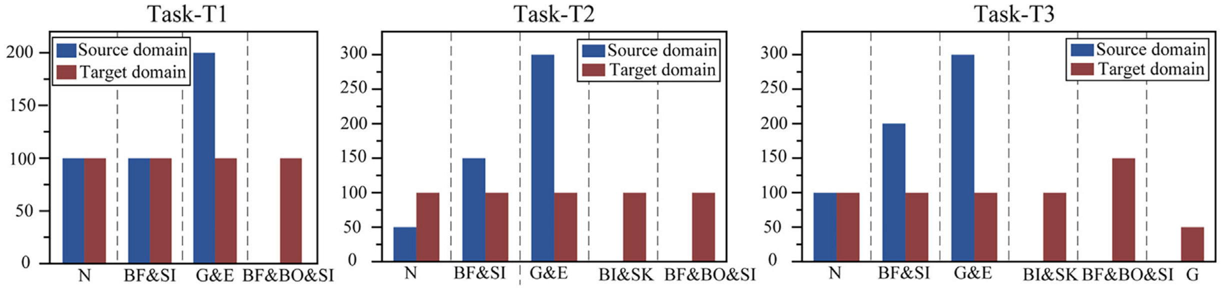


Fig. 6. Detailed sample distribution across different health states.

Table 7a
Architecture of the common modules in the compared models.

Layers	Parameter size	Output size	Activation
Feature Generator *G*			
Input	/	−1 × 3 × 64 × 64	/
Convolutional_1	Channel:128, kernel size:3, padding = 0	−1 × 128 × 62 × 62	ReLU
Convolutional_2	Channel:64, kernel size:3, padding = 0	−1 × 64 × 60 × 60	ReLU
Convolutional_3	Channel:32, kernel size:3, padding = 0	−1 × 32 × 58 × 58	ReLU
Convolutional_4	Channel:32, kernel size:3, padding = 0	−1 × 32 × 56 × 56	ReLU
Convolutional_5	Channel:16, kernel size:3, padding = 0	−1 × 16 × 54 × 54	ReLU
Linear	Dense number: 1024	−1 × 1024	
Classifier *C*			
Linear	Dense number: K + 1	−1×(K + 1)	Softmax
Discriminator *D*			
Linear_1	Dense number: 1024,Batch Normalization	−1 × 1024	LeakyReLU
Linear_2	Dense number: 1024,Batch Normalization	−1 × 1024	LeakyReLU
Linear_3	Dense number: 1	−1 × 1	Sigmoid

different degrees of openness but also explore the robustness of compared models under imbalanced training and testing data.

5.3.2. Implementation details

For fair comparisons, the architectures of all the methods, including the feature generator, the domain discriminator, and the classifier are implemented with the same architecture parameters as the proposed approaches, which are shown in Table 7a. Besides, the parameters of the proposed coarse-to-fine discriminator are presented in Table 7b. The detailed information of model hyperparameters, such as the learning rate, the optimizers and the number of training iterations are initialed as Table 8, which would be adjusted to produce the optimal results during the training process.

5.4. Experimental results and performance analysis

5.4.1. Experimental results

The compared results on the above OSDT tasks are given in Table 9a and Table 9b, including the normalized accuracy for all classes (OS) and normalized accuracy for the known classes only (OS^{*}). Each task is averaged 10 trials to reduce randomness and to provide the mean value and standard deviation of the testing accuracies. It can be seen the proposed TPTLN method generally outperforms other methods in all concerned tasks. It should be noticed that some methods perform better than TPTLN when only considering the known classes accuracy (OS^{*}) but perform much worse on all classes. For example, the CMU model performs better than TPTLN on the bearing taskT3 without considering the unknown class and reaches high accuracy as 97.9%. However, when considering the emerging fault data from the target domain, the misclassification of unknown faults would greatly degenerate the model transferability, leading to the lower accuracy as 85.7%. The DATLN model also suffers this problem, which performs well on classifying the known data but performs poorly on all the classes.

To further compare the model performance under different degrees of openness, the classification results of all methods for the bearing task T3 and gearbox task T3 are given in Fig. 7. As shown in Fig. 7, the FTNN and DCTLN models both suffer the negative transfer caused by the emerging fault class. The poor diagnostic performance of these standard transfer learning models could be attributed to the global domain alignment strategy without considering the effect of outlier data from $\boldsymbol{y}_{C_s+1}$ in the target domain, and this non-discriminative domain alignment strategy would lead to two problems: 1) wrongly recognizing unknown fault data as known fault class (marked with blue dashed lines) and 2) learning biased features because of matching the unknown data with source data (marked with purple dashed lines). These two problems limit the diagnostic performance of FTNN and DCTLN on all concerned OSDT scenarios, and as the openness increases, the transferability decreases more significantly.

Different from the FTNN and DCTLN matching the whole target domain with the source domain, the open-set transfer learning models achieve significant improvements by extracting shared features across domains for fault diagnosis and recognizing the unknown class to avoid the negative transfer. However, the problem of discriminability fluctuation is still not well addressed under different degrees of openness. It could be observed that some models tend to be over-discriminative. For example, in the bearing task T3, STA, OSBP, and IW-OSDA models misclassify the known samples as the unknown fault category (marked with the red dashed lines). While some models prone to be insufficiently discriminative. For example, in the gearbox task T3, OSBP, DATLN, and STA miss some unknown samples and recognize them as the known fault categories. These issues can be attributed to the fact that the above-compared methods could not adjust the discriminative bound according to the degree of openness in the target domain, leading to over-discrimination or under-discrimination of outlier samples. Correspondingly, the proposed model could well address the pending issue by the coarse-to-fine discriminator module, estimating the underlying openness and adjusts the decision bound adaptively. Noticeable performance increases can be seen between the TPLTN and other OSDT models concerning both known and unknown classes.

Table 7b
Architecture of the common modules in the compared models.

Layers	Parameter size	Output size	Activation
***K*-Coarse discriminator *B*$_s$**			
Input	/	-1×1024	/
Linear_1	Dense number: 256,Batch Normalization	-1×256	LeakyReLU
Linear_2	Dense number: 256,Batch Normalization	-1×256	LeakyReLU
Linear_3	Dense number: 1	-1×1	Sigmoid
Fine discriminator *B*$_t$			
Input	/	-1×1024	/
Linear	Dense number: 2	-1×2	Softmax

Table 8
Training parameters in the compared models.

Item	Detailed parameter
$\alpha_s, \alpha_t, \alpha_o$	0.25
α_d	0.5
Optimizer of the coarse-to-fine discriminator	Adam $(lr = 1 \times 10^{-4}, weight\_decay = 5 \times 10^{-4})$
Optimizer of the domain discriminator	Adam $(lr = 1 \times 10^{-4}, weight\_decay = 5 \times 10^{-4})$
Optimizer of the source classifier	Adam $(lr = 1 \times 10^{-4}, weight\_decay = 5 \times 10^{-4})$
Optimizer of the feature generator	Adam $(lr = 1 \times 10^{-4}, weight\_decay = 5 \times 10^{-4})$
Training iterations	500
Batch size	40

Table 9a
Accuracy (%) of each method on the bearing OSDT tasks.

Method	Task T1		Task T2		Task T3		Task T4	
	OS	***OS*** *	***OS***	***OS*** *	***OS***	***OS*** *	***OS***	***OS*** *
FTNN	79.8 ± 4.9	89.1 ± 3.0	67.4 ± 3.8	88.9 ± 2.3	52.0 ± 5.7	82.6 ± 3.6	37.0 ± 3.8	76.2 ± 4.7
DCTLN	71.1 ± 3.9	82.1 ± 4.3	70.2 ± 5.3	92.2 ± 3.6	56.6 ± 5.0	87.1 ± 3.2	38.5 ± 5.1	75.3 ± 6.1
OSBP	69.3 ± 4.5	77.8 ± 4.0	79.2 ± 3.4	91.4 ± 4.4	53.1 ± 5.5	79.3 ± 4.8	40.3 ± 7.1	78.2 ± 5.4
STA	63.5 ± 4.9	73.0 ± 3.6	63.6 ± 3.5	58.2 ± 4.1	63.1 ± 4.2	76.9 ± 5.0	45.1 ± 6.3	91.1 ± 4.8
DATLN	82.0 ± 3.7	94.1 ± 1.5	63.0 ± 4.3	86.7 ± 3.9	72.8 ± 4.9	98.2 ± 1.0	51.9 ± 3.7	100
CMU	97.5 ± 1.4	95.1 ± 1.3	95.5 ± 1.4	93.9 ± 1.75	85.7 ± 2.1	97.9 ± 1.9	86.9 ± 2.6	73.2 ± 5.2
IWOSDA	88.2 ± 1.8	86.4 ± 2.2	97.5 ± 0.9	96.7 ± 1.2	88.9 ± 2.0	82.2 ± 3.4	100.0	100.0
TPTLN	98.5 ± 1.0	98.2 ± 1.2	99.9 ± 0.3	99.8 ± 0.3	97.9 ± 0.3	95.7 ± 1.0	100.0	100.0

Table 9b
Accuracy (%) of each method on the gearbox OSDT tasks.

Method	Task T1		Task T2		Task T3	
	OS	***OS****	***OS***	***OS****	***OS***	***OS****
FTNN	79.8 ± 4.9	89.1 ± 3.0	67.4 ± 3.8	88.9 ± 2.3	52.0 ± 5.7	82.6 ± 3.6
DCTLN	71.1 ± 3.9	82.1 ± 4.3	70.2 ± 5.3	92.2 ± 3.6	56.6 ± 5.0	87.1 ± 3.2
OSBP	69.3 ± 4.5	77.8 ± 4.0	79.2 ± 3.4	91.4 ± 4.4	53.1 ± 5.5	79.3 ± 4.8
STA	63.5 ± 4.9	73.0 ± 3.6	63.6 ± 3.5	58.2 ± 4.1	63.1 ± 4.2	76.9 ± 5.0
DATLN	82.0 ± 3.7	94.1 ± 1.5	63.0 ± 4.3	86.7 ± 3.9	72.8 ± 4.9	98.2 ± 1.0
CMU	80.8 ± 1.6	74.3 ± 1.9	77.1 ± 2.0	61.9 ± 3.4	80.1 ± 2.1	61.7 ± 3.3
IWOSDA	93.4 ± 1.4	91.3 ± 1.9	91.0 ± 1.4	86.9 ± 2.2	86.0 ± 2.0	85.6 ± 3.9
TPTLN	94.0 ± 2.0	92.0 ± 2.6	94.20 ± 1.4	90.3 ± 2.3	92.3 ± 1.6	93.3 ± 2.1

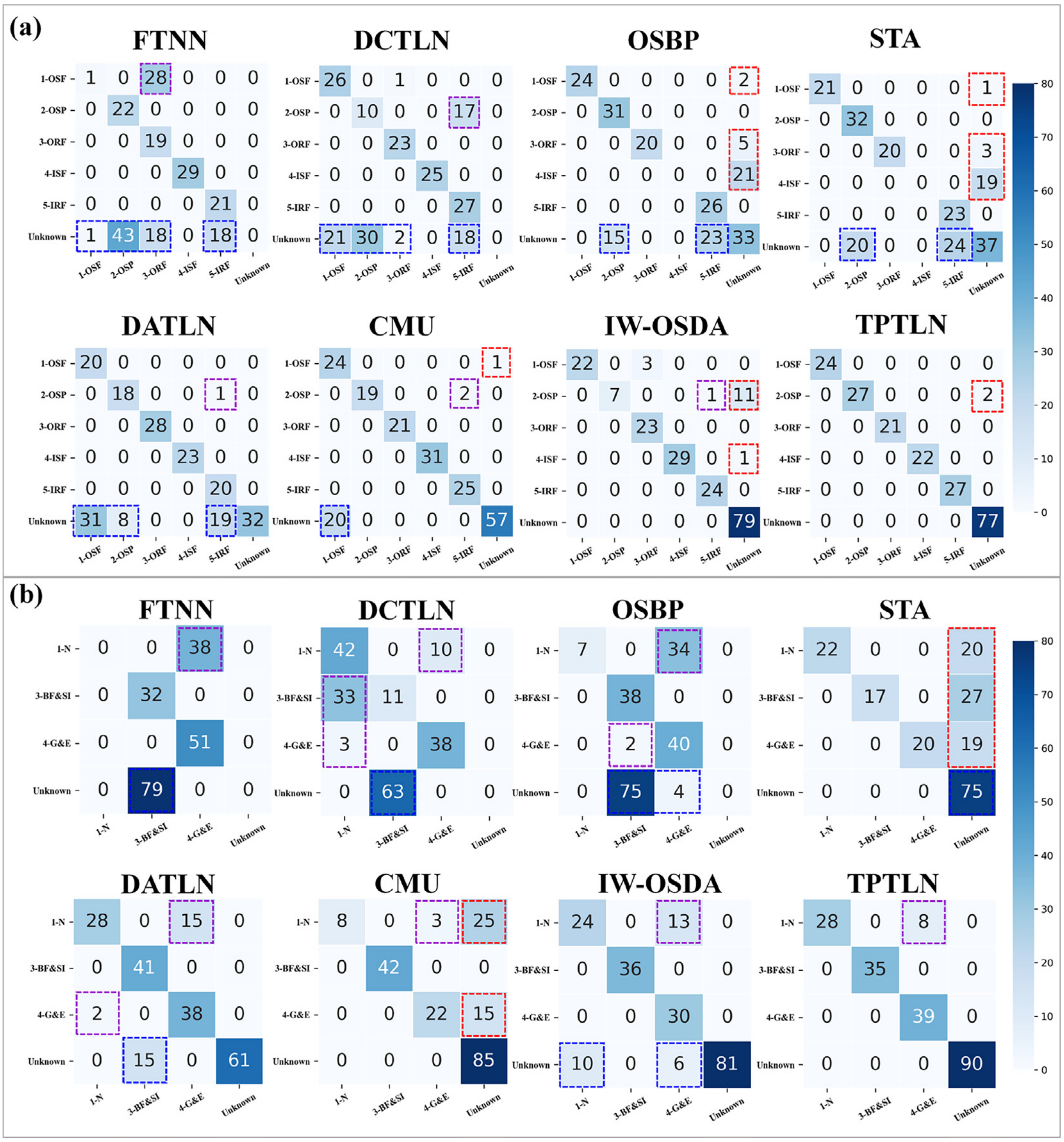


Fig. 7. Classification results of all methods on the designed OSDT tasks: (a) confusion matrices on bearing task T3 and (b) confusion matrices on gearbox task T3.

5.4.2. Performance investigation

To evaluate the performance of proposed method intuitively, the learned instance-level weights of the target domain samples in all OSDT tasks are investigated. Since the STA, CMU, and IW-OSDA models discriminate the outlier samples by assigning different weights, the results of these methods are given for comparison, and visualized results are shown in Fig. 8, where the weights of target outlier instances are marked in red.

From Fig. 8 it can be observed that in general, large weights would be obtained for the samples sharing the same health states with the source domain data, and small weights should be assigned to the target outlier samples. However, there are still have some incorrect weights assignment in the compared methods, leading to the decrease of discriminability. For example, concerning the bearing OSDT task T3, the target outlier instances with label 6 should have low weight values but obtain a high level of weight

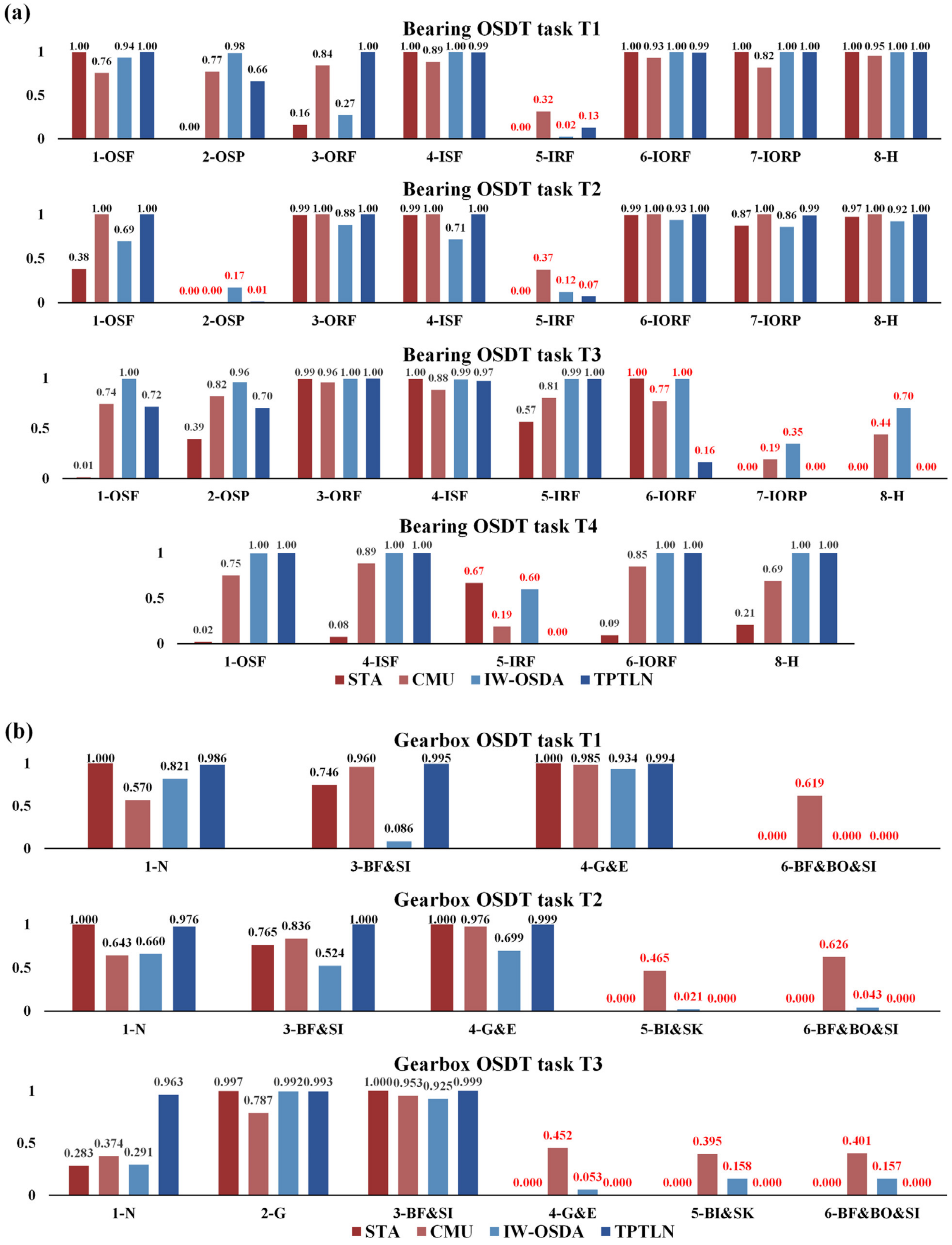


Fig. 8. Mean values of weights of the target instances in each class: (a) the results on the bearing OSDT tasks and (b) the results on the gearbox OSDT tasks.

values, which would be paid more attention during the domain adaptation with source the known samples and lead to the biased feature alignment. The results of bearing task T4 also follow similar patterns. On the other hand, with respect to the gearbox OSDT task T3, the target known instances with label 1 should obtain high weights to be aligned with the source domain data but are assigned with a low level of weight values, which would be paid less attention in the domain adaptation and lead to the insufficient feature alignment. The reasons for the performance degradation of compared methods are discussed as follows:

1) STA method employs a two-step discriminator structure to find the outlier sample with extreme similarity. Since STA only focuses on the samples with the significant divergence, treating the outlier samples as one general class. It is arguably sub-optimal since the target outlier intrinsic structures could not be fully exploited and those outlier samples with less diverged features would be missed.
2) CMU method constructs the target instance weights by calculating the statistical metrics from confidence, entropy, and consistency. When the outlier samples share similar intrinsic structures as the known classes, the weights of outlier samples are difficult to be further decreased. For the gearbox tasks, although CMU assigns a high level of weights to the target known samples, the target outlier samples also achieve a middle level of weights, which prevents fine discrimination.
3) IW-OSDA method employs the instance-level weight to detect the target outlier samples, and it chooses a fixed proportion of target samples with low similarity (set as 10%) for further discrimination. However, those outlier samples outside the selected proportion would be always unable to be chosen for further discrimination, resulting in the incorrect classification of certain categories.

Compared with the above approaches, the proposed method solves the above problems through learning an adaptive decision bound from the coarse-to-fine discriminator. The coarse discriminator would estimate the underlying outlier samples proportion and adjust the decision bound accordingly, and the fine discriminator would compress the weights of outlier samples to zero for further suppressing the negative effect on the shared classes alignment. The results of all tasks intuitively validate that the proposed method could accommodate different degrees of openness shift better and assign the corresponding level of weights to the target samples.

5.4.3. Feature visualization

To compare the model performance and reflect the advantages of proposed method, the t-distributed stochastic neighbor embedding (t-SNE) algorithm [28] is used to visualize the extracted features from the generator ***G***. Take the gearbox task T1 for example. The visualized features of t-SNE from different models are shown in Fig. 9.

As shown in Fig. 9, the FTNN and DCTLN could not discriminate the emerging fault data from the target domain, which classify the target unknown samples (marked as a purple cross) into the known category (marked as a blue circle). This is because the FTNN and DCTLN only have domain distribution module but lack target outlier data detection module. Moreover, the significant overlapping between different health states caused by the large-biased target domain could be observed in the FTNN and DCTLN module, which leads to misclassification and unexpected negative transfer. Compared with FTNN and DCTLN, the OSBP designs an outlier data discrimination module to recognize the target emerging fault data as an unknown category. However, the unknown data recognition depends on a fixed threshold (set as 0.5 empirically [15]), which lacks flexibility and robustness under different degrees of openness. It could be seen that there has multiple partial overlapping across different health states, leading to a great performance reduction on classification consequently.

The STA, DATLN, CMU, and IW-OSDA perform better compared with the above three approaches. It can be observed data from the known categories could be well separated, but there is still have noticeable overlapping between the target outlier data and the known data, leading to less effective cross-domain diagnostic performance.

In the proposed TPTLN model, the distributions of different domains are drawn closer to each other, and distributions of shared classes are drawn farther to each other, which indicates the high transferability of the known classes. Furthermore, the outlier samples are fully pushed away from the known classes, which facilitates the unknown target samples detection and suppresses the negative transfer on the shared space. Compared with other models, the TPTLN model could learn promising cross-domain features to achieve accurate fault diagnosis under the large degrees of

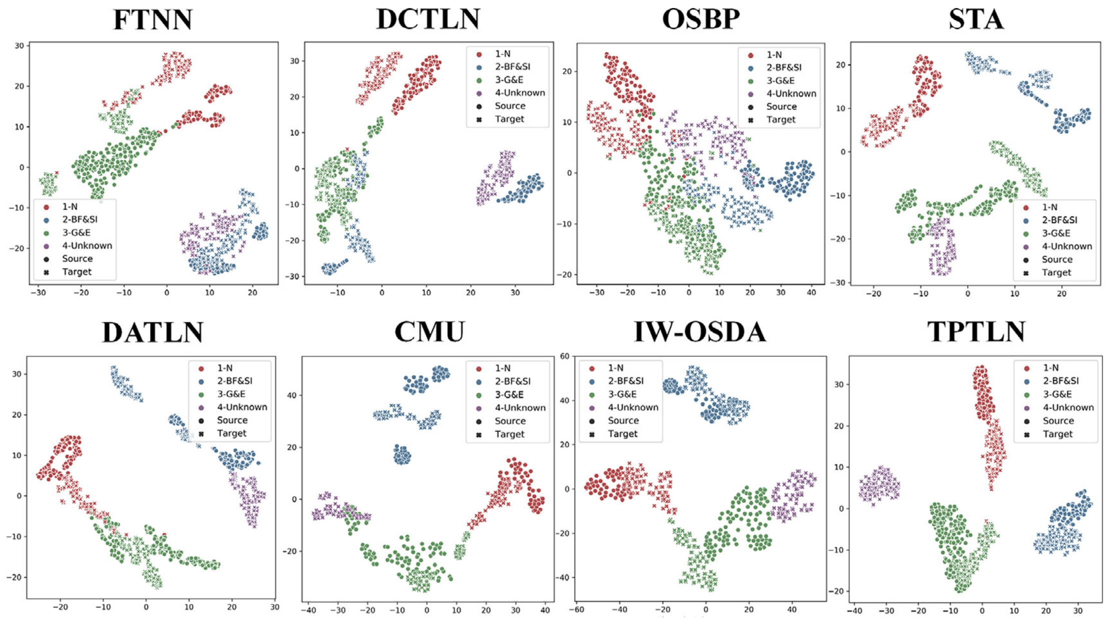


Fig. 9. The visualization of the domain-invariant features on gearbox task T1.

openness fluctuation, which considers both the intra-class compactness and the inter-class separability.

5.5. Ablation study

5.5.1. Component analysis

In this section, the ablation studies are conducted to investigate the effect of model components guided by the theoretical bound analysis on the final performance. Concretely, three specially designed modules are evaluated, including a coarse discriminator $\boldsymbol{B_s}$ based on the calibrated similarity, an adaptive fine discriminator $\boldsymbol{B_t}$ based on the domain consensus score and an adversarial discriminator $\boldsymbol{D}$ based on weighted distribution. The comparative results of their variants are given in Table 10.

As shown in Table 10, the TPTLN model with three modules shows superiority on all OSDT tasks of rolling bearings and gearbox compared with other variants. Detailed analyses for above results and corresponding discussions are listed as follows:

1) Calibrated similarity module.

The variants V_1 and V_2 both employ only one index (entropy or confidence) to describe the similarity of the unknown target samples with the source known samples. To visualize the difference on detecting the unknown samples by these variants, the estimated similarities of target samples on rolling bearing task T4 are shown in Fig. 10. It can be seen both V_1 and V_2 have their disadvantages. Concretely, the entropy-based method (V_1) assigns high weights to all the unknown target samples to guarantee the discriminability, but it assigns excessive weights to some target known samples causing them to be misclassified as an unknown category (the part enclosed by the red dotted line). On the other hand, the confidence-based method (V_2) avoids the over-discriminating by assigning low weights to all the known samples, but it misses some unknown samples and assigns insufficient weights causing them to be misclassified as the known category (the part enclosed by the red dotted line).

These situations could be attributed that the entropy-based method could easily detect the target unknown samples based on high uncertainty from their structure-specific features, but it is insensitive to uncertain known samples, leading to the incorrect recognition for part of known categories. The confidence-based method is opposite to entropy, which could recognize those known samples with high certainty from their structure-similar features as the source domain, but it exhibits low discriminability for uncertainty and is prone to miss part of unknown samples. Based on the above analyses, the V_5 is designed to overcome the disadvantage of each variant. From Fig. 10 it can be found that the proposed calibrated similarity measurement method could distinguish the known samples and unknown samples accurately,

Table 10
Comparative results of TPTLN with the variants.

Methods		V_1	V_2	V_3	V_4	V_5
Confidence similarity in $\boldsymbol{B_s}$			√	√	√	√
Entropy similarity in $\boldsymbol{B_s}$		√		√	√	√
Adaptive K selection in $\boldsymbol{B_t}$		√	√		√	√
Weighted matching in $\boldsymbol{D}$		√	√	√		√
OSDT for the rolling bearings						
Accuracy for all classes OS (%)	T1	87.45 ± 1.92	89.60 ± 1.45	94.95 ± 1.21	97.7 ± 0.93	**98.45 ± 1.01**
	T2	96.45 ± 1.47	94.30 ± 1.60	90.05 ± 1.98	89.7 ± 1.65	**99.90 ± 0.30**
	T3	77.55 ± 1.67	87.25 ± 1.94	73.65 ± 2.21	85.80 ± 2.16	**97.9 ± 0.34**
	T4	90.15 ± 2.01	100.00	93.30 ± 2.05	100.00	**100.00**
Accuracy for known classes OS* (%)	T1	85.41 ± 1.44	88.32 ± 2.46	94.25 ± 1.30	96.94 ± 1.31	**98.22 ± 1.18**
	T2	94.28 ± 1.58	91.38 ± 1.83	86.93 ± 2.12	87.71 ± 2.43	**99.86 ± 0.28**
	T3	75.11 ± 3.11	78.87 ± 2.92	76.69 ± 3.74	37.10 ± 2.74	**95.69 ± 1.01**
	T4	77.99 ± 3.06	100.00	86.14 ± 4.37	100.00	**100.00**
OSDT for the gearbox						
Accuracy for all classes OS (%)	T1	80.20 ± 1.47	86.90 ± 1.67	94.00 ± 0.62	91.80 ± 1.49	**94.00 ± 1.99**
	T2	82.05 ± 2.63	91.80 ± 1.43	92.50 ± 2.00	65.80 ± 2.38	**94.20 ± 1.38**
	T3	87.90 ± 1.99	87.80 ± 1.08	83.50 ± 1.91	37.30 ± 2.02	**92.25 ± 1.69**
Accuracy for known classes OS* (%)	T1	73.80 ± 3.40	82.28 ± 2.48	92.03 ± 1.41	89.81 ± 1.83	**91.96 ± 2.65**
	T2	69.15 ± 5.13	88.99 ± 2.13	87.29 ± 2.47	47.46 ± 1.68	**90.28 ± 2.31**
	T3	87.0 ± 3.51	79.35 ± 2.87	88.56 ± 2.28	75.62 ± 3.26	**93.18 ± 2.19**

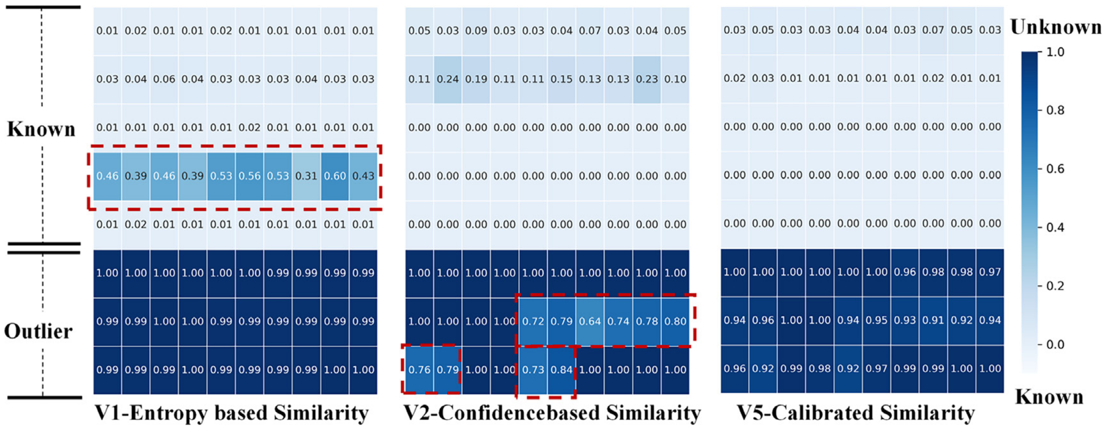


Fig. 10. The estimated similarities based on different indexes.

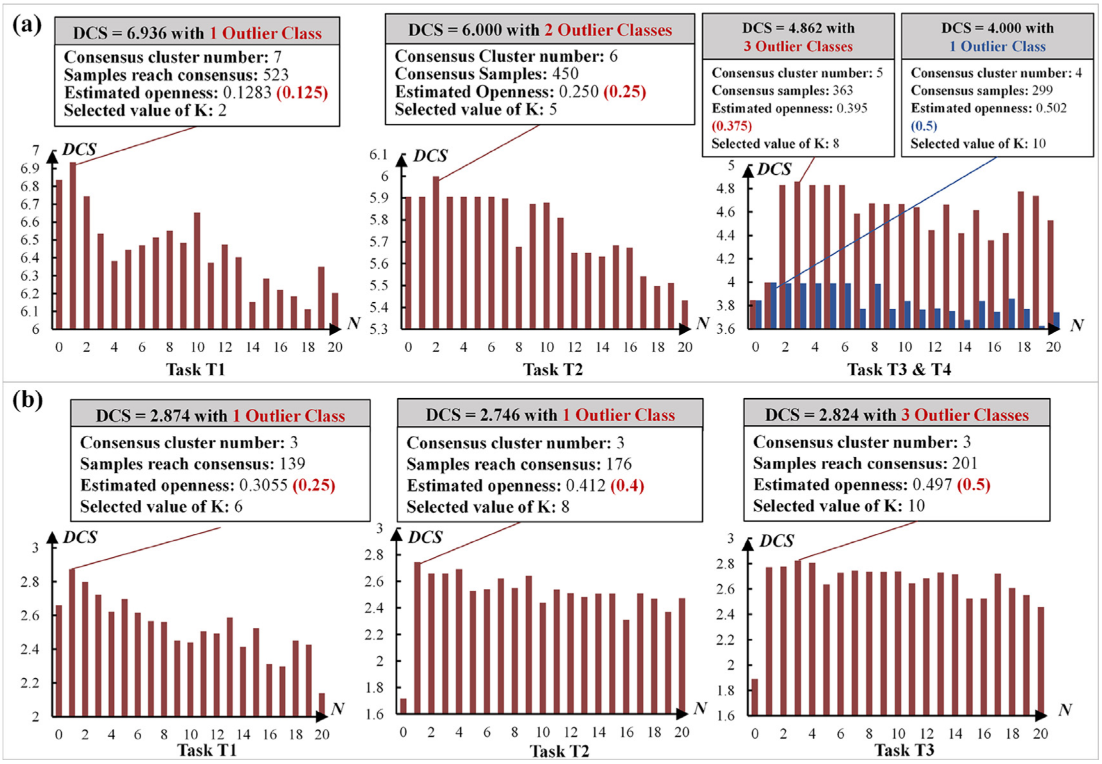


Fig. 11. The openness estimation and selection of K values based on domain consensus score: (a) results on bearing OSDT tasks and (b) results on gearbox OSDT tasks.

because it exploits complementary characteristics of both entropy and confidence, building a more robust discriminator to cover all types of predictions.

2) Adaptive K selection module.

The main difference between V_3 and V_5 is the choice of K value. In V_3, the subset $\mathscr{D}_t'$ only consists of one sample with the highest similarity (denoted as unknown) and one sample with the lowest similarity (denoted as known). In V_5, the subset $\mathscr{D}_t'$ is built through an adaptive K selection method based on the proposed domain consensus score (DCS), which would choose top-K samples with high similarities (bottom-K samples with low similarities) according to the degree of openness. The details of adaptive K selection method on bearing tasks and gearbox tasks are shown in Fig. 11.

From Fig. 11 the proposed adaptive K selection method firstly assumes that there are different numbers of outlier classes, calculates the domain consensus score of each candidate class, and determines the candidate with the highest score as the number of true outlier classes. Subsequently, the openness of each task is estimated by accumulating samples reaching consensus under the current outlier setting. From the result, it can be found the

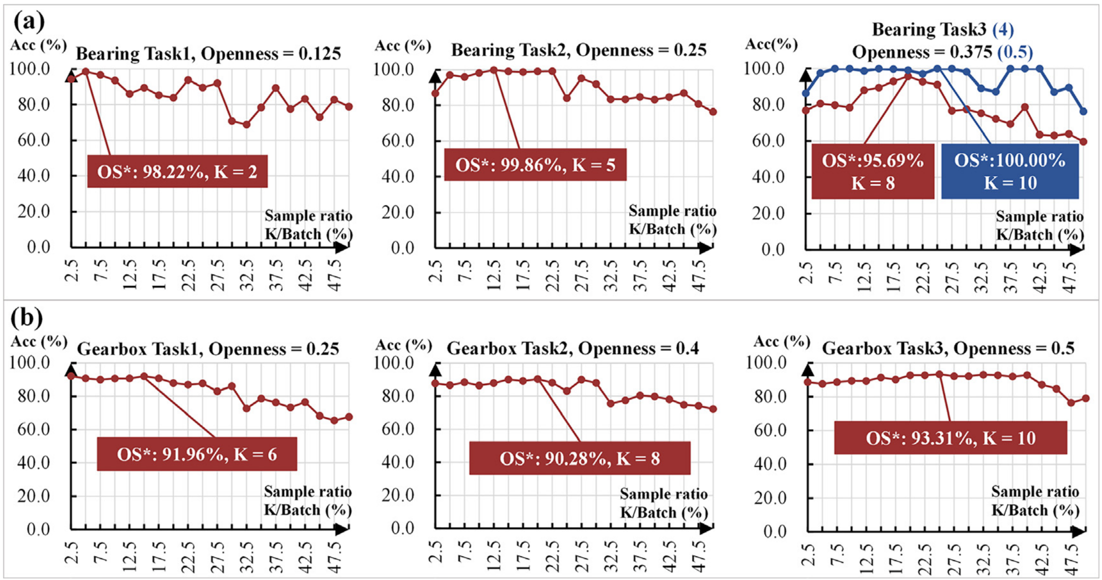


Fig. 12. The OS* of OSDT tasks with different values of K: (a) results on bearing OSDT tasks and (b) results on gearbox OSDT tasks.

divergence between the calculated openness and the true underlying openness (marked with red color) of each task is slightly small, which proves the effectiveness of the proposed method on estimating the outlier samples proportion. Finally, the subset size K is determined according to the calculated openness and batch size, and the classification results of all OSDT tasks based on different K are shown in Fig. 12. It can be found that the normalized accuracy for the known classes (OS*) on each task reaches optimum with the calculated K based on DCS, which proves that the proposed adaptive K selection approach could adjust the subset bound under different degrees of openness.

3) Weighted distribution matching module.

The difference between V_4 and V_5 is that whether assigning different weights on target samples during the domain-invariant feature extraction. From Table 10 it can be found the performance of V_4 would suffer dramatic degeneration on gearbox OSDT task T2 and T3, and the classification results of these tasks are visualized in Fig. 13. In task T2 it can be found the model not only misclassified the known samples but also recognize the known samples (3-BF&SI) as the unknown fault type. On the other hand, in task T3 there has a great negative transfer caused by the wrong recognition of all unknown samples as the known fault type. This unexpected negative transfer can be attributed to the indiscriminately matching all target samples with the source data, and the outlier samples would be aligned with the source known samples and influence the model discriminability interactively, leading to incorrect classification of the unknown (known) samples as the known (unknown) categories.

Based on the above discussions, the proposed three modules could effectively improve the model performance from different aspects. The calibrated similarity module ensures the discriminator accurately detecting the outlier data by employing the comple-

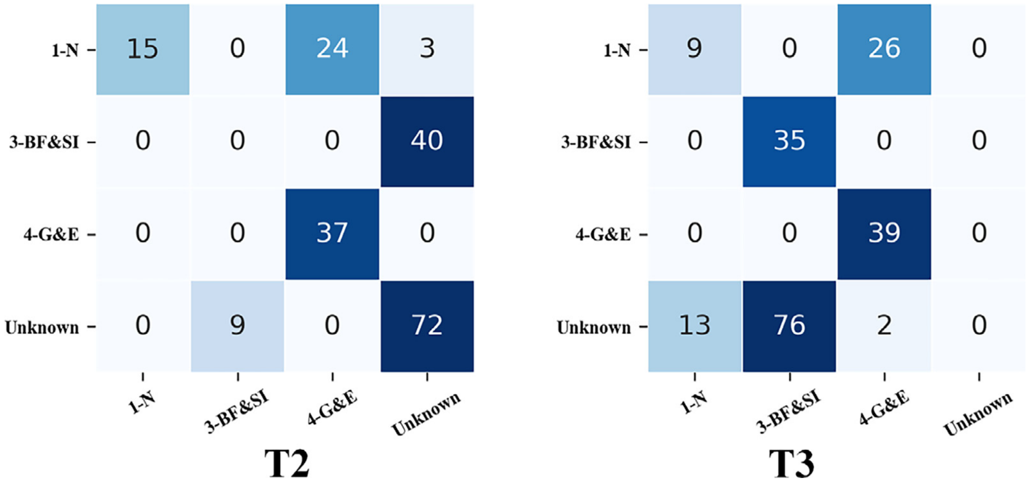


Fig. 13. The classification result of V_4 on gearbox tasks.

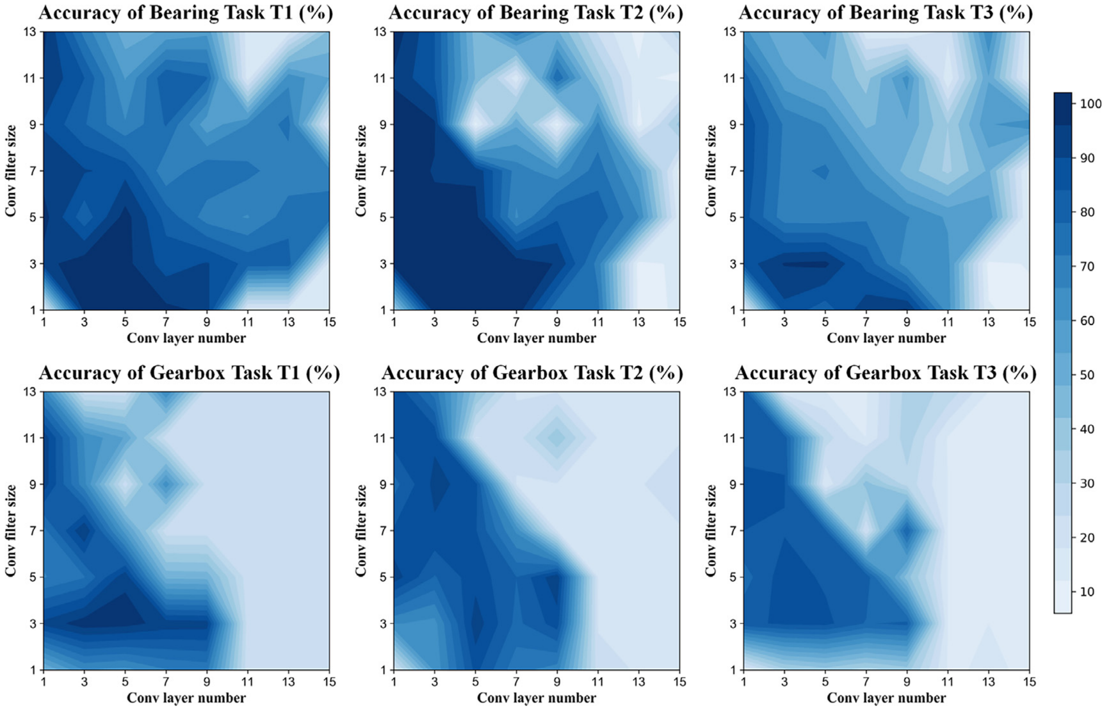


Fig. 14. The sensitivity analysis to the CNN network configurations.

mentary characteristics from entropy and confidence. The adaptive K selection module provides the discriminator a flexible subset based on DCS, which could accommodate different degrees of openness scenarios during the distract stage. The weighted distribution matching module avoids the unexpected negative transfer by assigning low weights to outlier data during the attract stage. The experimental results show that the model attached with all modules could achieve the best performance, and the ablation of a certain module will cause worse performance under all tasks.

5.5.2. Hyper-parameters sensitivity

In this section, the detailed ablation studies of key parameters in proposed model are carried out to evaluate the effect on diagnosis accuracy and transfer robustness. Concretely, the sensitivity analysis of key parameters is investigated from three aspects: network architectures in representation learning process, weight coefficients of objective function in optimizing process and hyper-parameters in training process.

1) Detailed network architectures.

In this subsection, configurations of feature learning backbones are analyzed to explore the effect on the final performance. According to the practical guide to CNN configuration [29], the filter region size and the number of feature layers should be investigated on proposed model. Specifically, the filter size and the layer number are set in range of [1 to 13], and ten experiments are performed for each parameter combinations to reduce randomness. The averaged results are shown in Fig. 14.

From the results it can be observed that the diagnosis accuracy decreases as the filter size and the number of layers increase from the default values. Especially, there has significant decrease when the parameters exceed specific value (e.g., when the number of convolutional layers exceeds 11 or the filter size exceeds 7 among the OSDT bearing tasks). This performance degeneration could be attributed to two aspects: 1) excessive irrelevant information from neighboring receptive fields being fed into the extracted features and 2) overfitting and gradient vanishing caused by too many convolutional layers. On the other hand, the model performance would decrease when the extracted features lose interactions with neighboring pixels because of too small filter size or when the learned representation combinations become weak caused by the insufficient number of convolutional layers. Therefore, the default values of network configuration are set as 3 for the filter size and 5 for the number of layers to obtain stable classification results according to the input scale and parameters analysis.

2) Optimization weight coefficients.

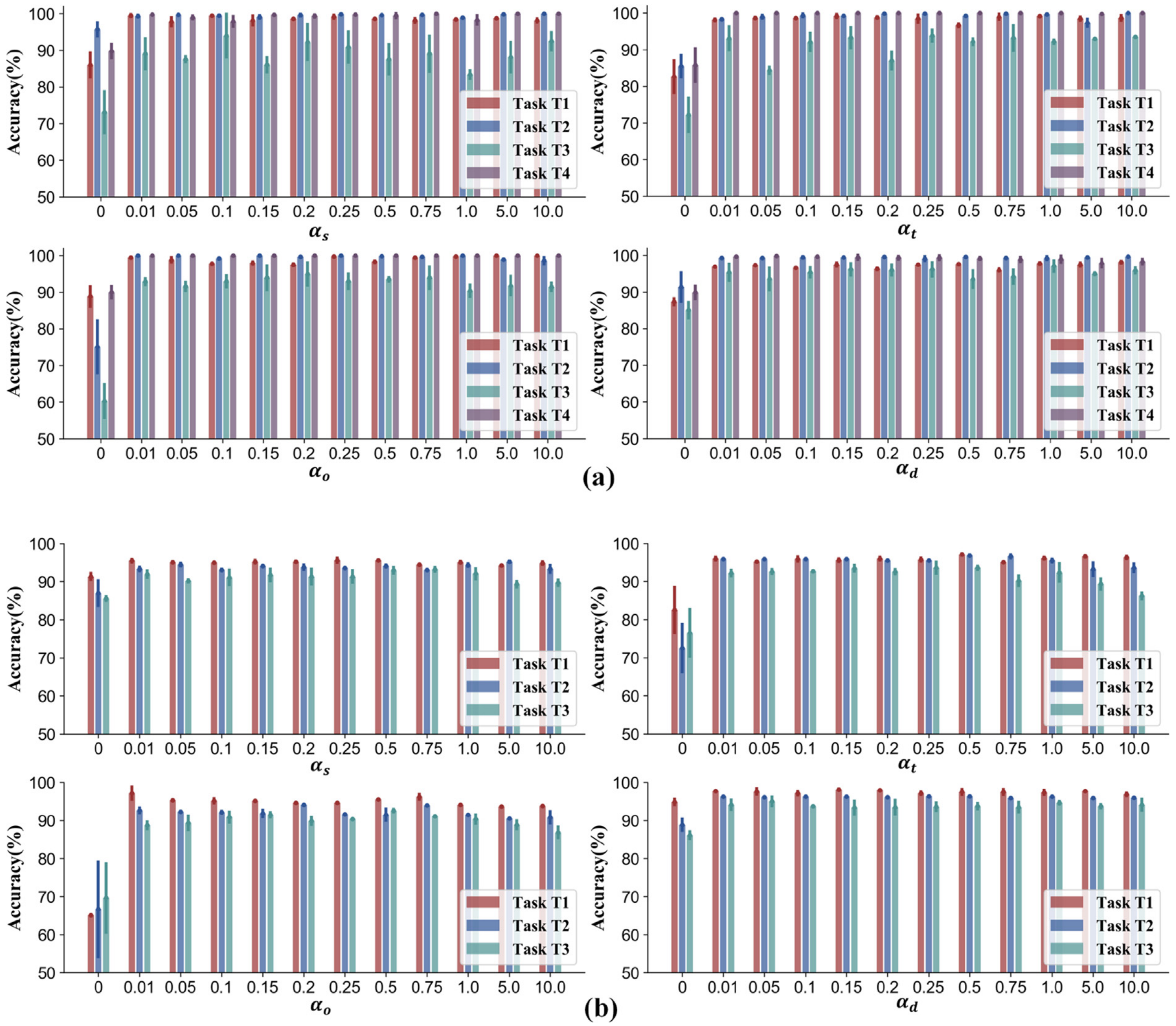


Fig. 15. The sensitivity analysis to the weight coefficients $\alpha_s, \alpha_t, \alpha_o$ and α_d(a) results on the bearing tasks (b) results on the gearbox tasks.

In this subsection, the effects of different weight coefficients in the objective optimization functions ($\alpha_s, \alpha_t, \alpha_o$ in the distract stage and α_d in the attract stage) on the final performance are investigated. Results of classification accuracy on all classes are illustrated in Fig. 15 with a wide range of $\alpha_s, \alpha_t, \alpha_o, \alpha_d$ (from 0 to 10) and ten experiments of each task are conducted to obtain the mean value (marked with dot) and standard deviation (marked with line).

From the results it can be found there has no significant performance fluctuations within the range of $\alpha_s, \alpha_t, \alpha_o$, α_d (0.05 to 1) and decreases slightly only when one of weight groups is too large (set as 10). It can be also observed dramatic performance degeneration without the corresponding weight (set as 0), which justifies the necessity of each designed module in the objective functions. In summary, the model is insensitive to $\alpha_s, \alpha_t, \alpha_o, \alpha_d$ as the design of progressive learning, and weights could be set in the range of [0.1 to 0.75] with similar scale according to the analysis results.

3) Training hyper-parameters.

In this subsection, sensitivity analysis of the initial learning rate and the batch size in the training process is conducted, and interactions between these hyper-parameters and diagnosis accuracy are demonstrated in Fig. 16. The initial learning rate and batch size are selected in a wide range of [2.5×10^{-5} to 1×10^{-3}] and [10 to 120] respectively, ten experiments of each task are conducted to obtain averaged accuracy.

From the results it could be observed that the model suffers performance decrease when learning rate is lower than 2.5×10^{-5} or higher than 5×10^{-4}, this fluctuation could be attributed to the fact that 1) the model gets stuck in an undesirable local minimum with too a small learning rate and 2) the model jumps over the global minimum with a too large learning rate and leads to divergence. On the other hand, when learning rate is initialized in an appropriate range as [5×10^{-5} to 2.5×10^{-4}], the model fluctuates only in a narrow range under the change of batch size, which is consistent across all tasks and proves the insensitivity of proposed model to batch size. Additionally, the averaged convergence time of the model training process with different hyperparameters combinations are illustrated in Fig. 17. The proposed model is constructed and conducted under the Windows 10 system, hardware platform with an AMD 5900X CPU and one NVIDIA GeForce RTX 3080 GPU. From Fig. 17 it can be seen that large batch size or small learning rate would slow down the model learning process, which could be attributed to excessive computational burden.

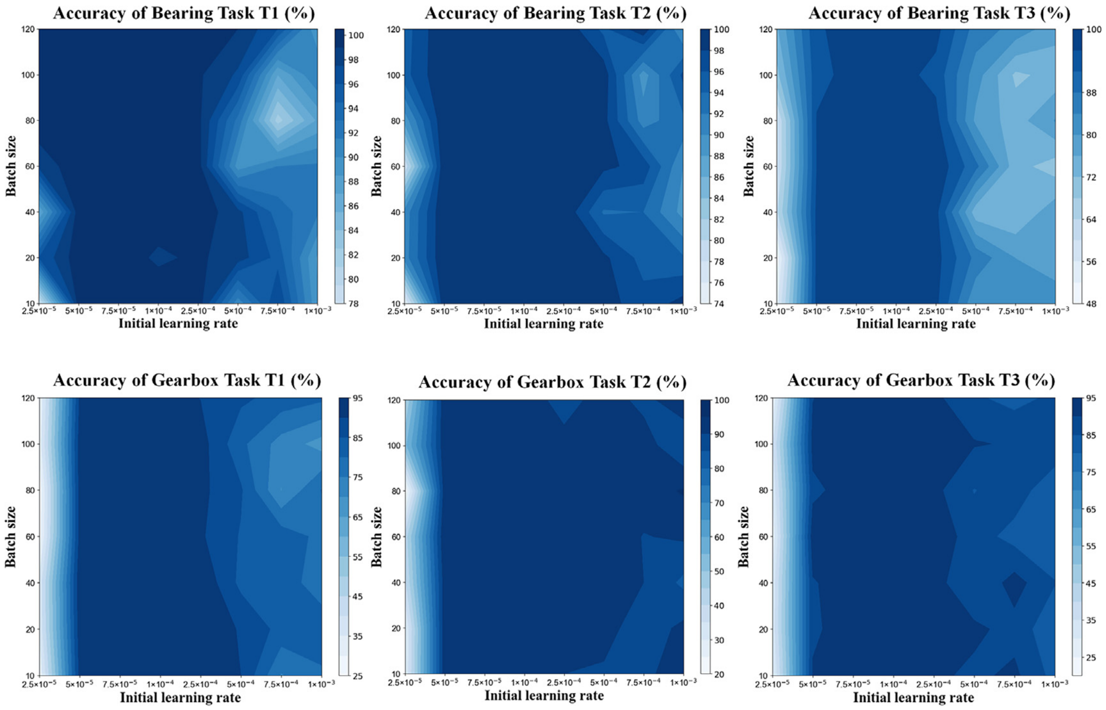


Fig. 16. The sensitivity analysis to the training hyper-parameters.

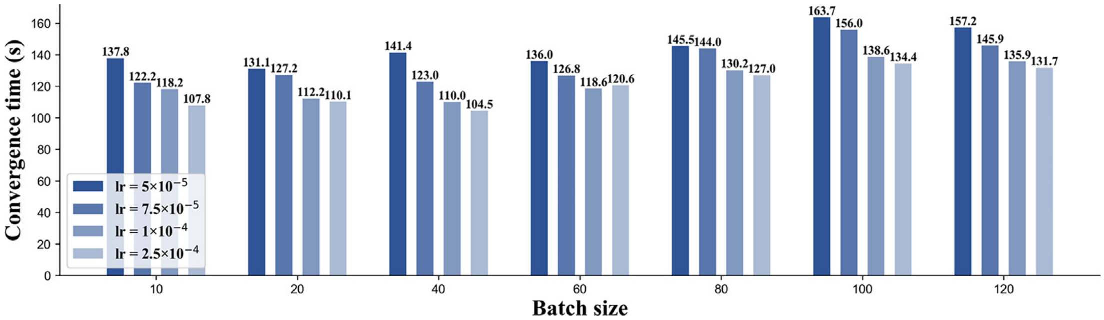


Fig. 17. The averaged convergence time of proposed model on all OSDT tasks.

According to above sensitivity analysis of training hyper-parameters, the learning rate is initialized as 1×10^{-4} and the batch size is selected as 40 to achieve stable performance considering both diagnosis accuracy and computation efficiency.

6. Conclusions

In this paper, a theory-guided transfer learning model named as TPTLN is proposed to tackle the open set diagnosis transfer problem (OSDT), in which the target domain has the unknown fault category. The TPTLN model tackles the OSDT issue through the distract stage and attract stage, in which the uncertainty calibration, adaptive openness estimation, and weighted distribution modules are designed for better accommodating different openness shifts. In the distract stage, the target unknown data are pushed away from the known classes to avoid participating domain alignment process, in which a robust discriminative boundary for the outlier data could be learned through the complementary similarities and domain consensus score. In the attract stage, data from the shared label space between the source domain and target domain will be aligned through an adversarial learning strategy to conduct diagnosis knowledge transfer. Furthermore, the theoretical upper bound of each stage is analyzed and combined into the optimization process, which could facilitate the inter-class separability for the distract stage and intra-class compactness for the attract stage. Various OSDT tasks based on the bearing and gearbox are designed to evaluate the proposed method, the experimental results demonstrate the TPTLN shows superior performance to other representative methods in the scope of diagnosis accuracy and transferring robustness.

In the future, a more challenging problem called as universal domain adaptation (UDA) for the mechanical diagnosis will be explored. In the UDA problem, both the source domain and target domain contain the unknown fault data, which could be seen as a more general setting for OSDT problem.

CRediT authorship contribution statement

Yafei Deng: Conceptualization, Methodology, Software, Writing – original draft. **Jun Lv:** Writing – review & editing. **Delin Huang:** Investigation, Validation. **Shichang Du:** Supervision.

Data availability

All data in this study are openly available online and the links have been included in the reference.

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.

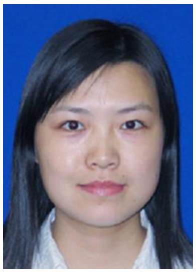
Acknowledgements

This work was funded by National Natural Science Foundation of China (Grant No. 52275499), the National Key Research and Development Program of China (No. 2022YFF0605700), and Shanghai Sailing Program (NO. 21YF1400500).

References

[1] C. Li, S. Zhang, Y. Qin, E. Estupinan, A systematic review of deep transfer learning for machinery fault diagnosis, Neurocomputing 407 (2020) 121–135.
[2] C. Cheng, B. Zhou, G. Ma, D. Wu, Y. Yuan, Wasserstein distance based deep adversarial transfer learning for intelligent fault diagnosis with unlabeled or insufficient labeled data, Neurocomputing 409 (2020) 35–45.
[3] X. Li, Y. Hu, M. Li, J. Zheng, Fault diagnostics between different type of components: A transfer learning approach, Appl. Soft Comput. 86 (2020) 105950.
[4] W. Li, R. Huang, J. Li, Y. Liao, Z. Chen, G. He, R. Yan, K. Gryllias, A perspective survey on deep transfer learning for fault diagnosis in industrial scenarios: Theories, applications and challenges, Mech. Syst. Sig. Process. 167 (2022) 108487.
[5] Y. Song, Y. Li, L. Jia, M. Qiu, Retraining strategy-based domain adaption network for intelligent fault diagnosis, IEEE Trans. Ind. Inf. 16 (9) (2019) 6163–6171.
[6] H. Zheng, Y. Yang, J. Yin, Y. Li, R. Wang, M. Xu, Deep domain generalization combining a priori diagnosis knowledge toward cross-domain fault diagnosis of rolling bearing, IEEE Trans. Instrum. Meas. 70 (2020) 1–11.
[7] R. Zhang, H. Tao, L. Wu, Y. Guan, Transfer learning with neural networks for bearing fault diagnosis in changing working conditions, IEEE Access 5 (2017) 14347–14357.
[8] S. Shao, S. McAleer, R. Yan, P. Baldi, Highly accurate machine fault diagnosis using deep transfer learning, IEEE Trans. Ind. Inf. 15 (4) (2018) 2446–2455.
[9] B. Yang, Y. Lei, F. Jia, S. Xing, An intelligent fault diagnosis approach based on transfer learning from laboratory bearings to locomotive bearings, Mech. Syst. Sig. Process. 122 (2019) 692–706.
[10] S. Jia, Y. Deng, J. Lv, S. Du, Z. Xie, Joint distribution adaptation with diverse feature aggregation: A new transfer learning framework for bearing diagnosis across different machines, Measurement 187 (2022) 110332.
[11] X. Li, W. Zhang, Q. Ding, Cross-domain fault diagnosis of rolling element bearings using deep generative neural networks, IEEE Trans. Ind. Electron. 66 (7) (2018) 5525–5534.
[12] Y. Deng, D. Huang, S. Du, G. Li, C. Zhao, J. Lv, A double-layer attention based adversarial network for partial transfer learning in machinery fault diagnosis, Comput. Ind. 127 (2021) 103399.
[13] L. Wen, L. Gao, X. Li, A new deep transfer learning based on sparse auto-encoder for fault diagnosis, IEEE Trans. Systems, Man, Cybern.: Syst. 49 (1) (2017) 136–144.
[14] Panareda Busto, P., & Gall, J. (2017). Open set domain adaptation. In Proceedings of the IEEE international conference on computer vision (pp. 754-763).
[15] Saito, K., Yamamoto, S., Ushiku, Y., & Harada, T. (2018). Open set domain adaptation by backpropagation. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 153-168).
[16] Liu, H., Cao, Z., Long, M., Wang, J., & Yang, Q. (2019). Separate to adapt: Open set domain adaptation via progressive separation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2927-2936).
[17] Fu, B., Cao, Z., Long, M., & Wang, J. (2020, August). Learning to detect open classes for universal domain adaptation. In European Conference on Computer Vision (pp. 567-583). Springer, Cham.
[18] J. Li, R. Huang, G. He, S. Wang, G. Li, W. Li, A deep adversarial transfer learning network for machinery emerging fault detection, IEEE Sens. J. 20 (15) (2020) 8413–8422.
[19] W. Zhang, X. Li, H. Ma, Z. Luo, X.u. Li, Open set domain adaptation in machinery fault diagnostics using instance-level weighted adversarial learning, IEEE Trans. Ind. Inf. 17 (11) (2021) 7445–7455.
[20] J. Zhu, C. Huang, C. Shen, Y. Shen, Cross-domain open set machinery fault diagnosis based on adversarial network with multiple auxiliary classifiers, IEEE Trans. Ind. Inf. (2021).
[21] Luo, Y., Wang, Z., Huang, Z., & Baktashmotlagh, M. (2020, November). Progressive graph learning for open-set domain adaptation. In International Conference on Machine Learning (pp. 6468-6478). PMLR.
[22] Zhang, Y., Liu, T., Long, M., & Jordan, M. (2019, May). Bridging theory and algorithm for domain adaptation. In International Conference on Machine Learning (pp. 7404-7413). PMLR.
[23] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.
[24] L. Zhong, Z. Fang, F. Liu, B. Yuan, G. Zhang, J. Lu, Bridging the theoretical bound and deep algorithms for open set domain adaptation, IEEE Trans. Neural Networks Learn. Syst. (2021).
[25] Li, G., Kang, G., Zhu, Y., Wei, Y., & Yang, Y. (2021). Domain Consensus Clustering for Universal Domain Adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9757-9766).
[26] Bearing DataCenter, Paderborn University. [Online]. Available: https://mb.uni-paderborn.de/kat/forschung/datacenter/bearing-datacenter.
[27] PHM Data Challenge 2009, PHM (Prognostics and Health Management) society [Online]. Available: https://www.phmsociety.org/competition/PHM/09.
[28] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008).
[29] Y. Zhang, B. Wallace, A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820, 2015.

Yafei Deng received B.S degree in Mechanical Engineering from Harbin Institute of Technology, Harbin, China in 2017. He is currently pursuing Ph.D. degree with State Key Laboratory of Mechanical System and Vibration, School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, China. His current research interests include prognostic and diagnostic for the key components in CNC machines using transfer learning approaches and hybrid models.



Jun Lv received Ph.D. degree in International Business Administration from Shanghai University of Finance and Economics, Shanghai, China, in 2008. Currently, she is an Associate Professor of the Faculty of Economics and Management at East China Normal University. Her major research interests are sustainability and operations research and reliability engineering.

Delin Huang received B.S. degree in Industrial Engineering and Management from Northeastern University, Qinhuangdao, China in 2013, and Ph.D. degree in Industrial Engineering and Management from Shanghai Jiao Tong University, Shanghai, China, in 2019. He is a Lecturer with Donghua University. His current research interests include manufacturing process monitoring and control, industrial measurement and computer vision.

Shichang Du received B.S. and M.S.E. degrees in Mechanical Engineering from the Hefei University of Technology, Hefei, China, in 2000 and 2003, respectively, and Ph.D. degree in Industrial Engineering and Management from Shanghai Jiao Tong University, Shanghai, China, in 2008. He was a Visiting Scholar with the University of Michigan. He is a Professor with Shanghai Jiao Tong University. His current research interests include quality and reliability engineering, quality control with analysis of error flow, and monitoring and diagnosis of manufacturing process.