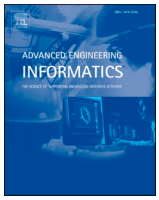
Advanced Engineering Informatics 65 (2025) 103243

Contents lists available at ScienceDirect

Advanced Engineering Informatics

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



Full length article

A transfer learning method: Universal domain adaptation with noisy samples for bearing fault diagnosis

Yi Sun, Hongliang Song, Liang Guo *, Hongli Gao, Ao Cao

Southwest Jiaotong University, China

ARTICLE INFO

Keywords:
Universal domain adaptation with noisy samples
Fault diagnosis
Transfering learning
Multibranch learning

ABSTRACT

Under the influence of frequent start-stop driving and rail launching during the service of urban rail vehicles, the source domain samples contain a large number of noise labels and noise samples. Moreover, the feature distribution and sample categories of the target domain and source domain are different because the urban rail vehicles are affected by the fluctuation of passenger flow and long-term service. This paper summarizes this real task in rail transportation as universal domain adaptation with noisy samples (UDANS). A novel multibranch convolutional neural network is proposed to solve the above problem. By optimizing the divergence of the two classifier outputs, the following objectives can be achieved: detecting noisy source samples, finding private classes in the target domain, and aligning the distribution of the source domain and the target domain. Finally, the results of the wheelset bearing dataset show that the method has advantages in rail transportation fault diagnosis.

1. Introduction

In recent years, the computing algorithm has developed rapidly, and the computing ability of the equipment has been continuously improved. The research on fault diagnosis in rail transportation based on deep learning has also made great progress [1,2]. Different from traditional fault diagnosis methods, deep learning-based fault diagnosis methods can automatically learn features from raw data [3]. Industrial applications of deep learning models have proliferated, such as ResNet-50 for turbine bearing monitoring, which achieves greater than 95 % accuracy with sub-second inference delays [4]. Therefore, such models are gaining increasing interest from academia and industry due to their high diagnostic accuracy, fast response times, and ease of training and deployment [5,6].

The vibration signals of wheelset bearings in urban rail vehicles hold a wealth of latent fault information under various operational conditions. These signals offer the potential for assessing the health and diagnosing faults in the wheelset bearings of the vehicles [2,7]. Fig. 1 shows the rotational speed and vibration signals of wheelset bearings. However, there are some difficulties and challenges in diagnosing the vibration signals of wheelset bearings, mainly in the following two aspects:

Noise samples. The features of wheelset bearing faults are only obvious when the vehicle is running stably, but the vehicle inevitably starts and stops frequently between platforms according to the actual situation. When the vehicle accelerates or decelerates, the signal amplitude changes dramatically due to the speed change, which masks the features of wheelset bearing faults [8]. Therefore, the signals during acceleration or deceleration phases are not helpful for identifying wheelset bearing faults, but rather interfere with the model's learning and classification [9,10]. These noise-induced disruptions impede the efficacy of feature extraction and classification, resulting in diminished accuracy and robustness of the model [11]. Therefore, how to effectively identify and eliminate noise samples, or reduce their impact on the model, is an urgent problem to be solved.

Domain adaptation. The dynamics of urban rail vehicles, influenced by passenger flow variations and service duration, lead to differences in the feature distribution and sample categories between the training (source domain) and testing (target domain) datasets [12]. This will cause domain drift phenomenon, which makes the model trained on fixed dataset perform poorly in actual use. Domain drift can be divided into two types: marginal distribution drift and conditional distribution drift [13]. The former pertains to differences in the distribution of sample features across the two domains, such as variations in vibration signals generated by vehicles with varying passenger loads. The latter involves differences in the relationship between sample labels and features across the two domains, like variations in the types of bearing

* Corresponding author.
E-mail address: guoliang@swjtu.edu.cn (L. Guo).

https://doi.org/10.1016/j.aei.2025.103243
Received 20 November 2024; Received in revised form 1 February 2025; Accepted 3 March 2025
Available online 22 March 2025

Nomenclature

$\{\boldsymbol{x}^s, y^s\}$	the labeled sample in source domain
$\{\boldsymbol{x}^t\}$	the unlabeled sample in target domain
$\boldsymbol{y}^s$	the unlabeled sample's one-hot matrix
D_S, D_t	the source and target domain dataset
P_s, P_t	the marginal distribution discrepancy of D_S and D_t
C_S, C_t	the fault categories of the source and target domain
C	the common categories shared by source and target domain
C_1, C_2	the two classifiers
E	the feature extractor
$\boldsymbol{p}^1(\boldsymbol{y}\|\boldsymbol{x}), \boldsymbol{p}^2(\boldsymbol{y}\|\boldsymbol{x}), \boldsymbol{p}^1, \boldsymbol{p}^2$	the softmax classification probabilities of C_1 and C_2
$p_k^1(\boldsymbol{y}\|\boldsymbol{x}_i), p_k^2(\boldsymbol{y}\|\boldsymbol{x}_i)$,	the predicted probabilities of the classifier for the sample $\boldsymbol{x}_i$
n_s, n_t	the number of a mini-batch from D_S and D_t
D_t'	a common sample set is composed of $\boldsymbol{x}_i^{t'}$
$D_{KL}(\cdot)$	the KL divergence formula
$H(\cdot)$, $\mathrm{L}_{ent}(\cdot)$	the entropy formula
$\mathrm{L}_{SKL}(\cdot)$	the SKL divergence loss
$\mathrm{L}_{crs}()$	the divergence loss
δ	the threshold
m	the margin of threshold
λ	the regularization parameter
$\mathrm{L}_{sup}(\cdot)$	the cross-entropy formula
$\widetilde{\mathrm{L}}_{JD}(\cdot)$	the distance formula
α	the scaling parameter of training sample scale
$\mathrm{L}(\cdot)$	the loss formula
$\boldsymbol{x}_i^{t'}$	a common category sample

faults found in different vehicles [14]. Therefore, how to effectively narrow or eliminate the gap between two domains, so that the model can adapt to the new domain's feature distribution and sample categories, is a challenging problem.

The problem of noise samples in urban rail vehicle can be addressed by two types of methods proposed by researchers [15]. The first type is to use data preprocessing techniques, including feature extraction, data cleaning, data transformation, data normalization, etc., which can improve data quality and suppress noise effects [16]. Although these techniques are simple and versatile, they may introduce biases, distortions, or data loss, and rely heavily on prior knowledge of data distribution or structure, leading to limitations. The second type is to use robust classification techniques, such as artificial neural networks (ANN) [17], deep belief networks (DBN) [18], recurrent neural networks (RNN) [19]. These techniques can learn from data without explicit rules or equations, handle various types of noise, such as label noise or feature noise, and offer robust and flexible solutions for different applications. However, they may experience overfitting or underfitting, require large datasets and computational resources, and lack transparency or interpretability. The choice of method depends on a profound understanding of problem complexity, dataset characteristics, and expected outcomes, demanding significant insight and experience from researchers.

Urban rail vehicle fault diagnosis exemplifies a classic domain adaptation problem. In current research, attempts have been made to apply domain adaptation methods to address fault diagnosis challenges in urban rail vehicles. Marginal distribution shift, denoted as domain adaptation (DA), is a process of transferring knowledge from a source domain to a target domain that lacks labeled samples. Qian et al. introduced a model to enhance joint distribution adaptability [20]. Their approach employs sparse filtering (SF) and principal component analysis (PCA) to reduce dimensionality of resampled data, coupled with Maximum Mean Discrepancy (MMD) for aligning feature distributions between domains. In a parallel vein of research, Ainapure et al. optimized PCA as a convolutional neural network [21]. Furthermore, Xiao et al. introduced the Marginal Stacking Denoising Auto-encoder (mSDA) into the realm of fault diagnosis [22].

Conditional distribution shift, denoted as universal domain adaptation (UDA), pertains to scenarios where source and target domains possess asymmetric fault categories. Zhao et al. integrated an attention mechanism to assign category-specific weights, enhancing dynamic adjustment and self-adaptive capabilities of their model [23]. Wang et al. proposed a triplet loss guided adversarial domain adaptation method [24]. Their approach utilizes Wasserstein distance to align

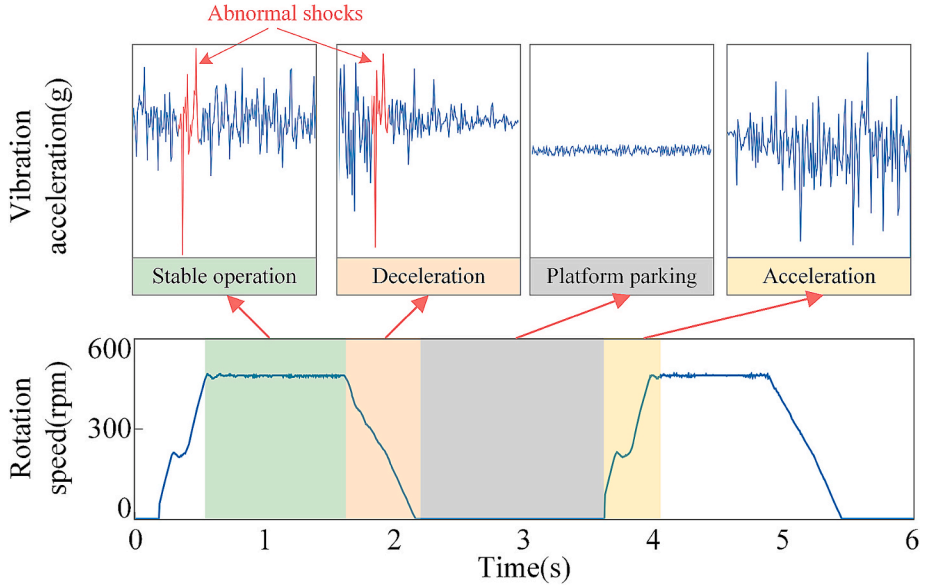


Fig. 1. Monitoring signals of rotational speed and vibration acceleration of axle box bearings of urban rail vehicles.

domain-level distributions and triplet loss for category-level distribution alignment. Notably, Zhang et al. have made innovative strides by amalgamating diverse loss functions [25]. Their model integrates classification loss, center-based discriminative loss, and correlation alignment loss, facilitating intra-class coherence and inter-class separability in feature extraction. Additionally, Zhang et al. employed $\ell 2$-norm, and Li et al. introduced manifold alignment, introducing novel strategies for addressing UDA challenges [26,27]. To conclude, domain adaptation holds promise for addressing urban rail vehicle fault diagnosis problems, surmounting challenges posed by disparate data distributions and enhancing model generalization and robustness. However, current domain adaptation methods still exhibit notable limitations. Formulas employed to quantify the similarity between source and target domains lack the requisite precision or stability, thereby inducing fluctuations or declines in the model's performance.

The above studies show that ztransfer learning model has achieved remarkable achievements in the field of mechanical fault diagnosis. However, most existing transfer learning methods are limited by the specific conditions and scenarios they can handle, and fail to cope with the complex and diverse vibration signals of urban rail vehicles in service [28]. Therefore, a more challenging task in rail transportation needs to be solved, which is called universal domain adaptation with noise samples (UDANS). The essence of UDANS lies in utilizing unsupervised techniques to detect faults in rail vehicles across varying passenger load scenarios. The difficulty of UDANS lies in the fact that both the source domain and the target domain contain noisy samples, and the target domain has different label categories from the source domain. It has the following properties, with data features shown in Fig. 2.

A convolutional neural network with multiple classifiers is proposed, which can effectively solve the UDANS problem and outperform the state-of-the-art methods in urban rail transportation. The flow of the proposed model is shown in Fig. 3. The main novelties of this paper are as follows:

- A multi-branch convolutional neural network (CNN) architecture was developed to simultaneously detect noisy samples and classify distinct fault types. Each classifier branch processes input features independently, allowing the model to robustly filter out mislabeled or irrelevant samples in both source and target domains. This design significantly enhances overall classification accuracy and resilience against noise interference.
- Incorporating adversarial learning principles enables the proposed model to achieve macro-scale alignment of edge distributions between the source and target domains. This allows the model to learn a common feature representation for both domains, thereby improving classification accuracy in the target domain.
- A new loss function, called joint divergence, is proposed to address the asymmetric fault category issue in UDA. The design of joint divergence aims to address the asymmetry in fault categories between the source and target domains, thereby enhancing the model's generalization capability.

The subsequent sections of this paper are structured as follows. Section II, a survey of related work is carried out. The details of the architecture proposed in this paper are described in Section III. Section IV, the experimental set-up is described, and the performance of our architecture is analyzed. Section V presents the conclusions.

2. Preliminary

2.1. Problem formulation

$\{\boldsymbol{x}^s, y^s\}$ represents a labeled sample in the source domain, while $\{\boldsymbol{x}^t\}$ represents an unlabeled sample in the target domain. The one-hot matrix expanded by y^s is $\boldsymbol{y}^s$. n_s and n_t is denoted as the number of a mini-batch from the source and target domain. Let $D_S = \left\{\left(\boldsymbol{x}_i^s, y_i^s\right)\right\}_{i=1}^{n_s}$ denotes the source domain dataset, and $D_t = \left\{\left(\boldsymbol{x}_i^t\right)\right\}_{i=1}^{n_t}$ denotes the target domain dataset. P_s and P_t represent the marginal distribution discrepancy of D_S and D_t, respectively. The fault categories of the source and target domain are denoted as C_S and C_t, and $C = C_s \cap C_t$ to represent the common categories shared by both domains. For UDANS in mechanical fault diagnosis, this paper needs to train an accurate diagnostic model across the source and target domains. The model can learn domain-invariant features under the following conditions.

- D_S and D_t are collected in different equipment, that is, $P_s \neq P_t$.
- The samples in D_S with high noise, which means that a certain percentage of $\boldsymbol{x}^s$ cannot be classified correctly.
- Some fault categories only appear in the source domain, denoted as $C \subset C_s$. $\overline{C_s} = C_s \backslash C$ is used to represent the private category of the source domain. Some fault categories only appear in the source domain, denoted as $C \subset C_t$. $\overline{C_t} = C_t \backslash C$ is used to represent the private category of the source domain.

2.2. Multiple classifier combination

This paper tackles the UDANS problem from the perspective of multiple classifier combination. Inspired by the fusion algorithm, models with different parameters and structures are trained using the same dataset. For each sample, different models output different feature distributions. Therefore, this paper constructs a model with one feature extractor, denoted as E, and two classifiers, denoted as C_1 and C_2. When sample $\boldsymbol{x}$ is input into the model, $\boldsymbol{p}^1(\boldsymbol{y}|\boldsymbol{x})$ and $\boldsymbol{p}^2(\boldsymbol{y}|\boldsymbol{x})$ are the softmax classification probabilities of C_1 and C_2, respectively. The predicted probabilities of the classifier for the sample $\boldsymbol{x}_i$ are $p_k^1(\boldsymbol{y}|\boldsymbol{x}_i)$ and $p_k^2(\boldsymbol{y}|\boldsymbol{x}_i)$, k is the number of categories in the corresponding domain of sample $\boldsymbol{x}_i$.

A model with two classifiers is proposed in the article, which has a

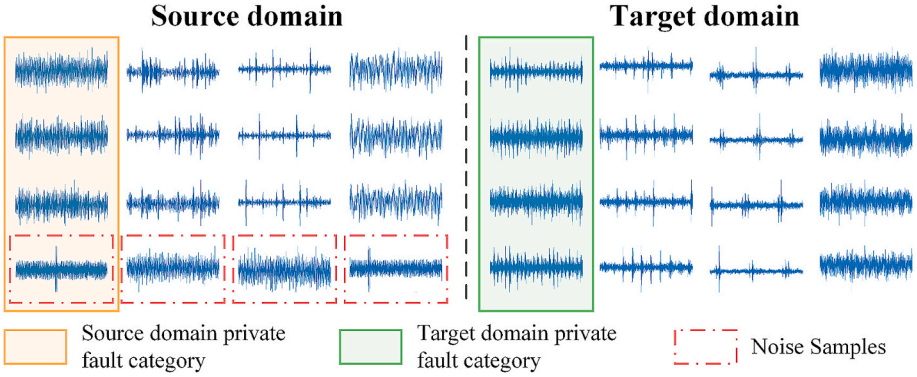


Fig. 2. Setting of UDANS.

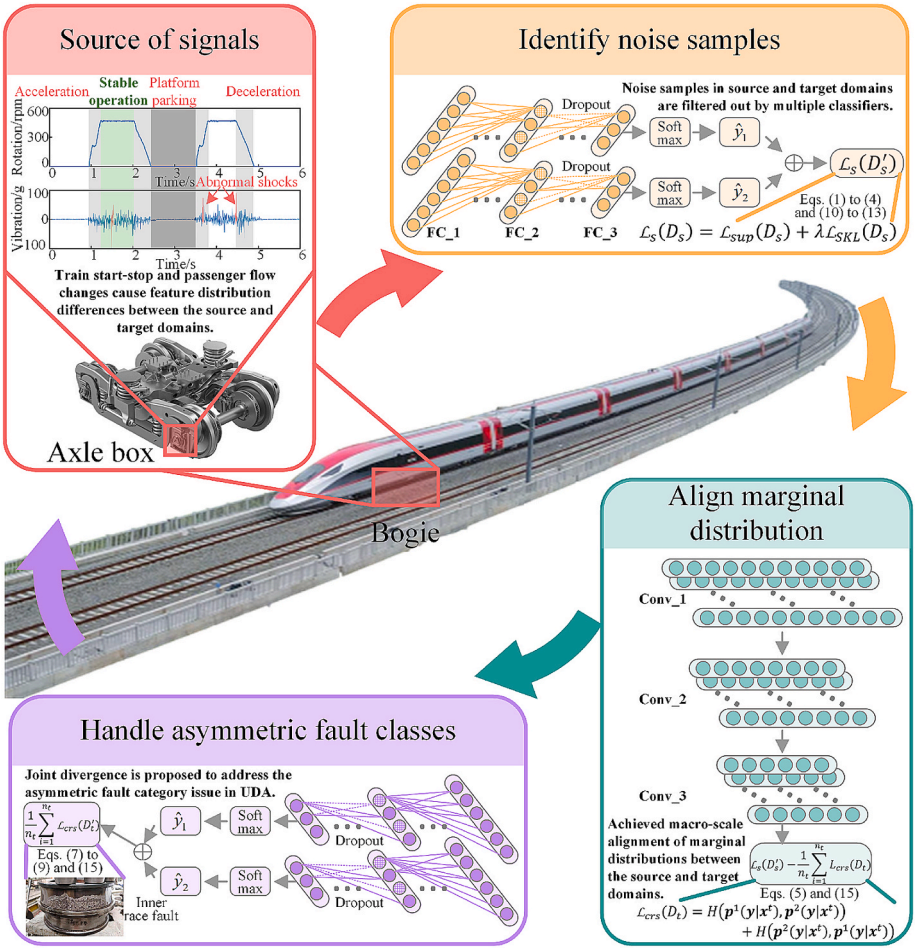


Fig. 3. The flow of the proposed model.

better diagnostic effect on the source domain samples with noise. The two classifiers have different initialization parameters and they are trained independently. The specific reasons are as follows, if the sample $\boldsymbol{x}_i$ is clean, the outputs of C_1 and C_2 are close to the correct labels, and $\boldsymbol{p}^1(\boldsymbol{y}|\boldsymbol{x})$ and $\boldsymbol{p}^2(\boldsymbol{y}|\boldsymbol{x})$ are similar. Conversely, if the sample $\boldsymbol{x}_i$ with noise, the outputs of C_1 and C_2 are far from to the correct labels, and $\boldsymbol{p}^1(\boldsymbol{y}|\boldsymbol{x})$ and $\boldsymbol{p}^2(\boldsymbol{y}|\boldsymbol{x})$ have a significant difference. The model calculates the divergence between two classifiers for each sample in each mini-batch, and selects the sample with smaller divergence to update the gradient.

2.3. Symmetric Kullback-Leibler (KL) Divergence

The symmetric KL(SKL) is used to solve the DA problem. KL divergence, also known as relative entropy, is used to measure the difference between two probability distributions. Because it satisfies non-negativity, the KL divergence can also describe the distance between one softmax classification probabilities $\boldsymbol{p}^1$ and another softmax classification probabilities $\boldsymbol{p}^2$. For the source domain, $D_{KL}(\boldsymbol{p}^1\|\boldsymbol{p}^2)$ and $D_{KL}(\boldsymbol{p}^2\|\boldsymbol{p}^1)$ are shown in the following equations:

$$D_{KL}(\boldsymbol{p}^1\|\boldsymbol{p}^2) = \sum_{k=1}^{C_s} p_k^1(\boldsymbol{y}|\boldsymbol{x}_i^s) log \frac{p_k^1(\boldsymbol{y}|\boldsymbol{x}_i^s)}{p_k^2(\boldsymbol{y}|\boldsymbol{x}_i^s)}$$

$$= \sum_{k=1}^{C_s} p_k^1(\boldsymbol{y}|\boldsymbol{x}_i^s) log p_k^1(\boldsymbol{y}|\boldsymbol{x}_i^s) - \sum_{k=1}^{C_s} p_k^1(\boldsymbol{y}|\boldsymbol{x}_i^s) log p_k^2(\boldsymbol{y}|\boldsymbol{x}_i^s)$$

$$= -H(\boldsymbol{p}^1(\boldsymbol{y}|\boldsymbol{x}^s)) + H(\boldsymbol{p}^1(\boldsymbol{y}|\boldsymbol{x}^s), \boldsymbol{p}^2(\boldsymbol{y}|\boldsymbol{x}^s)) \quad (1)$$

$$D_{KL}(\boldsymbol{p}^2\|\boldsymbol{p}^1) = \sum_{k=1}^{C_s} p_k^2(\boldsymbol{y}|\boldsymbol{x}_i^s) log \frac{p_k^2(\boldsymbol{y}|\boldsymbol{x}_i^s)}{p_k^1(\boldsymbol{y}|\boldsymbol{x}_i^s)} \quad (2)$$

where $H(\boldsymbol{p}^1(\boldsymbol{y}|\boldsymbol{x}^s))$ is the entropy of $\boldsymbol{p}^1(\boldsymbol{y}|\boldsymbol{x}^s)$, and the corresponding $H(\boldsymbol{p}^1(\boldsymbol{y}|\boldsymbol{x}^s), \boldsymbol{p}^2(\boldsymbol{y}|\boldsymbol{x}^s))$ is the cross-entropy of $\boldsymbol{p}^1(\boldsymbol{y}|\boldsymbol{x}^s)$ and $\boldsymbol{p}^2(\boldsymbol{y}|\boldsymbol{x}^s)$.

From the above equations, it can be seen that the calculation process of KL divergence contains logarithmic operations, so $D_{KL}(\boldsymbol{p}^1\|\boldsymbol{p}^2)$ and $D_{KL}(\boldsymbol{p}^2\|\boldsymbol{p}^1)$ are asymmetric, that is, $D_{KL}(\boldsymbol{p}^1\|\boldsymbol{p}^2) \neq D_{KL}(\boldsymbol{p}^2\|\boldsymbol{p}^1)$. This paper uses SKL divergence, because it accurately reflects the difference between two probability distributions. SKL divergence can be rewritten as:

$$\mathrm{L}_{SKL}(D_s) = \frac{1}{n_s}\sum_{i=1}^{n_s} D_{KL}(\boldsymbol{p}^1\|\boldsymbol{p}^2) + \frac{1}{n_s}\sum_{i=1}^{n_s} D_{KL}(\boldsymbol{p}^2\|\boldsymbol{p}^1) \quad (3)$$

$$\mathrm{L}_{SKL}(D_t) = \frac{1}{n_t}\sum_{i=1}^{n_t} L_{crs}(D_t) - \frac{1}{n_t}\sum_{i=1}^{n_t} L_{ent}(D_t) \quad (4)$$

where, $\mathrm{L}_{crs}(D_t)$ shows the divergence of the two classifiers' outputs and $\mathrm{L}_{ent}(D_t)$ shows the entropy of each classifier output. They are expressed as follows:

$$\mathrm{L}_{crs}(D_t) = H(\boldsymbol{p}^1(\boldsymbol{y}|\boldsymbol{x}^t), \boldsymbol{p}^2(\boldsymbol{y}|\boldsymbol{x}^t)) + H(\boldsymbol{p}^2(\boldsymbol{y}|\boldsymbol{x}^t), \boldsymbol{p}^1(\boldsymbol{y}|\boldsymbol{x}^t)) \quad (5)$$

$$\mathrm{L}_{ent}(D_t) = H(\boldsymbol{p}^1(\boldsymbol{y}|\boldsymbol{x}^t)) + H(\boldsymbol{p}^2(\boldsymbol{y}|\boldsymbol{x}^t)) \quad (6)$$

2.4. Joint Divergence

Joint divergence is proposed to address the UDA problem. In Eq. (5), SKL divergence is calculated by subtracting entropy from divergence. However, the entropy of private category samples in the target domain is uncertain, causing their SKL divergence to be smaller than that of common category samples. Consequently, the private category samples are misidentified as common samples with wrong prediction labels. The above errors will be transmitted and amplified during the training of the model. To address this problem, joint divergence is proposed, which is defined by the following equation:

$$\mathrm{L}_{JD}(D_t) = \frac{1}{n_t}\sum_{i=1}^{n_t}\widetilde{\mathrm{L}}_{crs}(D_t) + \frac{1}{n_t}\sum_{i=1}^{n_t}\widetilde{\mathrm{L}}_{ent}(D_t) \tag{7}$$

$$\widetilde{\mathrm{L}}_{crs}(D_t) = \begin{cases} |\mathrm{L}_{crs}(\boldsymbol{x}^t) - \delta \,|if|L_{crs}(\boldsymbol{x}^t) - \delta\,| > m \\ 0if|L_{crs}(\boldsymbol{x}^t) - \delta\,| \le m \end{cases} \tag{8}$$

$$\widetilde{\mathrm{L}}_{ent}(D_t) = \begin{cases} |\mathrm{L}_{ent}(\boldsymbol{x}^t) - \delta \,|if|L_{ent}(\boldsymbol{x}^t) - \delta\,| > m \\ 0if|L_{ent}(\boldsymbol{x}^t) - \delta\,| \le m \end{cases} \tag{9}$$

where a threshold δ and a margin m are introduced to distinguish between private and common categories. The threshold $\delta = logC_s$, because $H(\boldsymbol{p}^1) + H(\boldsymbol{p}^2) = 2logC_s$ in the training process. The margin $m = 1$.

3. Proposed method

In this section, the training process of the proposed model is described in detail, as shown in Fig. 4. The architecture of the model includes one feature extractor and two classifiers, which are Resnet-18 and three fully connected layers. Four steps are included in the training process of the model.

3.1. Step 1

The main purpose of step 1 is to enable the model to filter clean samples from the source domain and extract their features. The classifier and feature extractor are trained simultaneously. According to [29], the model cannot accurately classify noise samples because their fault features are contaminated. Noise samples typically have large losses, while clean samples tend to have small losses. Joint loss, defined by Eq. (10), is used to train the model, incorporating SKL divergence and cross-entropy, with λ as a regularization parameter.

$$\mathrm{L}_s(D_s) = \mathrm{L}_{sup}(D_s) + \lambda \mathrm{L}_{SKL}(D_s) \tag{10}$$

The SKL divergence measures the difference between the outputs of samples from two classifiers. For clean samples in the source domain, the SKL divergence between the outputs of the two classifiers is small. In contrast, for noise samples, the SKL divergence is large. Consequently, the SKL divergence effectively identifies clean samples in the source domain. And, by minimizing the SKL divergence, the consistency of the two classifiers can be improved.

The cross-entropy is used to measure the difference between the output of a sample and its correct label. The ability of feature extractor to extract fault features of samples can be improved by minimizing cross-entropy. The cross-entropy can be defined as:

$$\mathrm{L}_{sup}(D_s) = -\frac{1}{n_s}\sum_{i=1}^{n_s}\sum_{k=1}^{C_s} y_i^s log p_k^1(\boldsymbol{y}|\boldsymbol{x}_i^s) - \frac{1}{n_s}\sum_{i=1}^{n_s}\sum_{k=1}^{C_s} y_i^s log p_k^2(\boldsymbol{y}|\boldsymbol{x}_i^s) \tag{11}$$

To select clean samples, Eq. (10) is used to measure the loss of samples one by one in each batch-size. And the samples with smaller losses are selected to train the model. The details of selecting samples are shown in Eq. (12).

$$D_s' = arg\,min_{D':|D'|\ge\alpha|D'|}\mathrm{L}_{sup}(D_s) \tag{12}$$

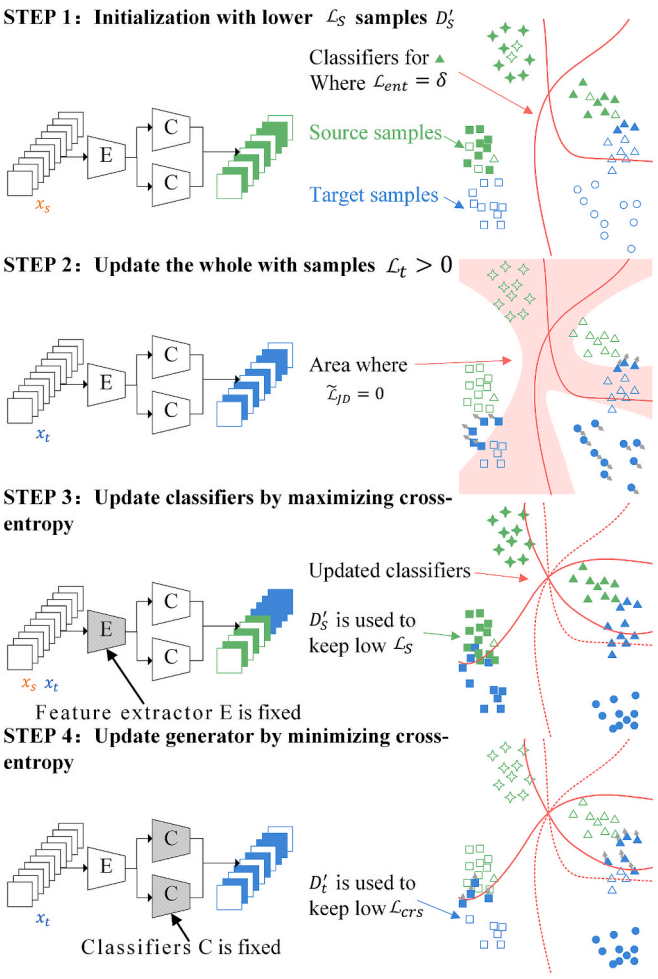


Fig. 4. Training steps in the proposed method. There are four classes in both the source and target domains, with two classes common to both. The white samples are not used for training, while the filled ones are used.

where α is a scaling parameter that can be used to adjust the proportion of training samples selected in each batch. The optimization objective of step 1 is as follows:

$$\min_{E,C_1,C_2} \mathrm{L}_s(D_s') \tag{13}$$

3.2. Step 2

The main purpose of step 2 is that the model can select private category samples from the target domain. Accordingly, the classifier and the feature extractor of the model are trained simultaneously. The model uses $\widetilde{\mathrm{L}}_{JD}(D_t)$ to measure the distance between the two classifier outputs for each target domain sample. And the model can identify the private category samples because of the threshold δ and the margin m. Further, in order to increase the $\widetilde{\mathrm{L}}_{JD}(D_t)$ of private category samples and decrease the $\widetilde{\mathrm{L}}_{JD}(D_t)$ of common category samples. During the training process of the model, a negative $\widetilde{\mathrm{L}}_{JD}(D_t)$ distance is chosen as follows:

$$\mathrm{L}_t(D_t) = \widetilde{\mathrm{L}}_{JD}(D_t) \tag{14}$$

$\mathrm{L}_s(D_t)$ is used to expand the difference between the outputs of the two classifiers for the private category samples. This makes the feature distributions of the private category samples and common category samples gradually move away from each other. In summary, the optimization objective is expressed as:

$$\min_{E,C_1,C_2} \mathrm{L}_t(D_t) \tag{15}$$

In addition, $\boldsymbol{x}_i^{t'}$ is defined as a common category sample, which is $\boldsymbol{x}_i^{t'}$: $\boldsymbol{x}_i^{t'} \in D_t$, $\mathrm{L}_{crs}(\boldsymbol{x}^t) < \delta - m$. In each batch, a common sample set D_t' is composed of $\boldsymbol{x}_i^{t'}$.

3.3. Step 3

The main purpose of step 3 is that the model can align the edge distribution of the source domain and the target domain, that is, the domain adaptation on the macro-scale. Therefore, only the classifier of the model is trained. The classifier of the model can be considered as a sample discriminator. And the divergence of classifiers is increased during training, which makes the model unable to detect whether the sample is from the source or target domain. In this step, both source and target domain samples are used to train the model. By integration $\mathrm{L}_s(D_s')$ with cross-entropy of target domain samples, the overall objective function can be rewritten as:

$$\min_{C_1,C_2} \mathrm{L}_s(D_s') - \frac{1}{n_t}\sum_{i=1}^{n_t} L_{crs}(D_t) \tag{16}$$

3.4. Step 4

Finally, the main purpose is to make the model align the conditional distribution of the common samples in the source domain and the target domain, that is, the domain adaptation on the micro-scale. Therefore, the classifier of the model is fixed and the feature extractor is trained to reduce the divergence between the common samples D_t'. In the network training, the optimization objective is to minimize $\mathrm{L}_{crs}(D_t')$ which is defined as:

$$\min_{E} \frac{1}{n_t}\sum_{i=1}^{n_t} \mathrm{L}_{crs}(D_t') \tag{17}$$

Moreover, Eq. (16) is repeated n times to better align the conditional distribution of common class samples.

4. Experiment

4.1. Experimental Setup

4.1.1. Induction Bearing Dataset

The wheelset bearings of high-speed trains are mostly in normal condition when they are in service. As a result, normal data is easy to obtain, however, obtaining real fault data is a challenge. In this paper, the fault data is obtained by establishing a wheelset bearing test platform. Although the experimental platform is different from the real high-speed train in monitoring variable speed and variable load conditions, it can generate the most real and closest fault data when the high-speed train is running in a short time.

The main components of the test platform test bench include driving motor, wheelset and its connecting shaft, two wheelset bearings and their axle boxes, transmission system, control system, radial and axial loading system. As shown in Fig. 5. The acceleration sensor is placed on the top surface of the right wheelset bearing axle box, and the sampling frequency is $20kHz$. To simulate a train traveling at $120km/h$, the rotational speed of the drive motor is set to $655r/min$. To simulate the fluctuation of passenger flow, the radial loads on the wheelset bearings were $56kN$, $146kN$, $236kN$. The fault category is set to inner race fault, rolling element fault, cage fault, outer race fault. Table 1 shows the detailed descriptions of the wheelset bearing domain adaptation dataset.

The actual speed of trains can vary due to factors such as track conditions, weather, and other external influences. To evaluate the performance of models under varying rotational speeds, a dataset encompassing three representative speed conditions has been established. Specifically, train speeds of 90, 120, and 150 km/h were selected, corresponding to bearing rotational speeds of 589, 786, and 983 rpm, respectively. To realistically simulate operating conditions under passenger loads, a radial load of 236 kN was applied to two sets of bearings. This dataset provides a comprehensive basis for analyzing model robustness and classification accuracy under practical operating conditions, as presented in Table 2.

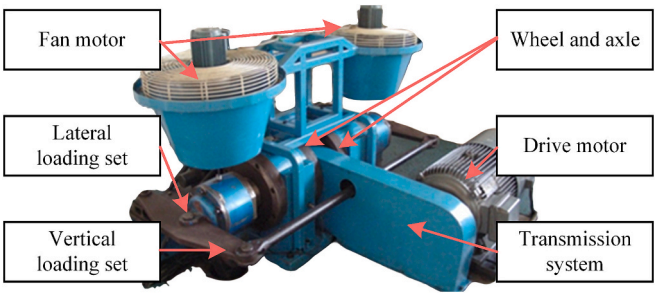


Fig. 5. Wheel bearing testbed.

Table 1
Description of the Wheelset Bearing Domain Adaptation Dataset.

Task	Source (radial load)	Target (radial load)	Proportion of noise samples	Category Settings	Problem
T1	$56kN$	$146kN$	$\rho = 0.15$	$C/\overline{C_s}/\overline{C_t} = 3/1/1$	UDANS
T2	$56kN$	$146kN$	$\rho = 0.3$	$C/\overline{C_s}/\overline{C_t} = 3/1/1$	UDANS
T3	$56kN$	$236kN$	$\rho = 0.15$	$C/\overline{C_s}/\overline{C_t} = 3/1/1$	UDANS
T4	$56kN$	$236kN$	$\rho = 0.3$	$C/\overline{C_s}/\overline{C_t} = 3/1/1$	UDANS
T5	$146kN$	$56kN$	$\rho = 0.15$	$C/\overline{C_s}/\overline{C_t} = 3/1/1$	UDANS
T6	$146kN$	$56kN$	$\rho = 0.3$	$C/\overline{C_s}/\overline{C_t} = 3/1/1$	UDANS
T7	$146kN$	$236kN$	$\rho = 0.15$	$C/\overline{C_s}/\overline{C_t} = 4/0/0$	WSDA
T8	$146kN$	$236kN$	$\rho = 0.3$	$C/\overline{C_s}/\overline{C_t} = 4/0/0$	WSDA
T9	$236kN$	$56kN$	$\rho = 0.15$	$C/\overline{C_s}/\overline{C_t} = 3/1/1$	UDANS
T10	$236kN$	$56kN$	$\rho = 0.3$	$C/\overline{C_s}/\overline{C_t} = 3/1/1$	UDANS
T11	$236kN$	$146kN$	$\rho = 0.15$	$C/\overline{C_s}/\overline{C_t} = 4/0/0$	WSDA
T12	$236kN$	$146kN$	$\rho = 0.3$	$C/\overline{C_s}/\overline{C_t} = 4/0/0$	WSDA

Table 2
Description of Wheelset Bearing Variable-speed Dataset.

Task	Source (Rotational speed)	Target (Rotational speed)	Proportion of noise samples	Category Settings	Problem
R1	589 r/min	786/983 r/min	$\rho = 0.15$	$C/\overline{C_s}/\overline{C_t} = 4/0/0$	WSDA
R2	589 r/min	786/983 r/min	$\rho = 0.3$	$C/\overline{C_s}/\overline{C_t} = 4/0/0$	WSDA
R3	786 r/min	589/983 r/min	$\rho = 0.15$	$C/\overline{C_s}/\overline{C_t} = 3/1/1$	UDANS
R4	786 r/min	589/983 r/min	$\rho = 0.3$	$C/\overline{C_s}/\overline{C_t} = 3/1/1$	UDANS
R5	983 r/min	589/786 r/min	$\rho = 0.15$	$C/\overline{C_s}/\overline{C_t} = 3/1/1$	UDANS
R6	983 r/min	589/786 r/min	$\rho = 0.3$	$C/\overline{C_s}/\overline{C_t} = 3/1/1$	UDANS

4.1.2. Compared methods

In this section, various classic methods in transfer learning are compared with the method proposed in this paper. In order to ensure the accuracy of the results, the structure and experimental setup of classic methods have been fine-tuned to make them similar to the method proposed in this paper, as shown in Table 3.

Baseline: The baseline model is unmodified Resnet18 [30].

Domain adaptation (DA): Transfer component analysis (TCA) is a classic shallow DA method [31]. TCA extracts the features of the source and target domain data and maps them into a new Hilbert space. Then, the distance between domain distributions is shortened by maximum mean discrepancy. Domain adversarial neural network (DANN), the most popular DA method [32]. DANN performs well under significant domain drift and demonstrates robust results in DA scenarios.

Weakly-supervised domain adaptation (WSDA): The domain adaptive deep belief network (DA-DBN) uses multi-kernel maximum mean discrepancy distance to measure domain differences [33]. The anti-noise neural network framework (AAnNet) is based on convolutional neural network architecture [34]. AAnNet predicts faults under variable load conditions with high robustness to noise.

Universal domain adaptation (UDA): UDA with hybrid weighted deep adversarial learning is a hybrid selection adaptive algorithm with source class-wise and target instance-wise weighting mechanism [25]. The latest research in UDA includes a deep learning-based open set DA method [35].

4.1.3. Implementation Details

When a high-speed vehicle passes through a turnout or an inter-track weld, the vibration signal changes abruptly and the fault feature is masked by high background noise. The above samples are called noisy samples, and the proportion ρ of noise samples was chosen from {0.15, 0.3}.

During the service of high-speed trains, new types of faults may occur in their bearings. That is, in order to satisfy the UDA experimental setup, the fault categories of the dataset are asymmetric in Table 1.

4.1.4. Domain Adaptation Analysis

The classification accuracy under domain adaptation dataset are shown in Fig. 6. The model proposed in this paper outperforms other models in most migration tasks, which proves its superiority in UDANS problem. It is not difficult to find that the baseline model and the DA models are difficult to solve the UDANS task, and the UDA models have better diagnostic effect than the WSDA models.

From the task setting, it can be found that the proposed model has better stability when the noise sample ratio ρ increases from 0.15 to 0.3. The experimental tasks TI and T2 only differ in the proportion of noisy samples, as they can be considered as the same group. The standard deviation was used to assess the effect of the proportion of noisy samples on each model. First, the classification accuracy difference was calculated for each of the six groups of experiments, e.g. |T1-T2|. Subsequently, the standard deviation was calculated for the classification accuracy difference. The standard deviation of the proposed model is the lowest among all models, 0.0115. The standard deviations of AAnNet and DA-DBN models were 0.0363 and 0.0405, respectively, and ranked

Table 3
Task Settings for Each Model.

Model	Noisy samples	DA	UDA
Baseline [30]	×	×	×
TCA [31]	×	√	×
DANN [32]	×	√	×
DA-DBN [33]	√	√	×
AAnNet [34]	√	√	×
UDA-1 [25]	×	√	√
UDA-2 [35]	×	√	√
Propssed	√	√	√

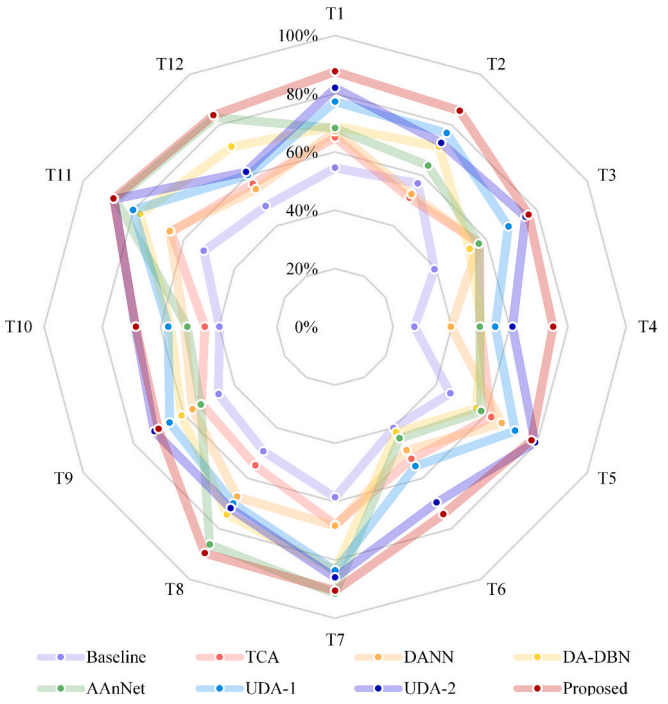


Fig. 6. Classification accuracy of each model under domain adaptation dataset.

third and fourth in all models. It can be seen that the model based on WSDA task has good stability when the source domain data contain noise. Moreover, for WSDA tasks (T7, T8, T11 and T12), the classification accuracies of the above two models are significantly higher than other tasks. In particular, the AAnNet model achieved a classification accuracy of 89.2 % at T7, which was the highest of all models.

Both the UDA models and the proposed model have good classification accuracy when the source domain samples contain a small number of noisy samples, $\rho = 0.15$. Nevertheless, when the source domain samples contain a large number of noise samples, $\rho = 0.3$, the classification accuracy of the UDA model decreases sharply. To be specific, in task T5, the UDA-2 model achieved slightly higher classification accuracy than the proposed model in this paper. However, with the increase in the proportion of noise samples, the classification accuracy of the UDA-2 model drops sharply and is much lower than the proposed model.

4.1.5. Variable-speed Analysis

To systematically investigate the influence of operational speed variations (589–983 rpm, equivalent to 90–150 km/h) on bearing fault signatures, vibration data were acquired from four critical bearing components: inner race, outer race, cage, and rolling elements, as detailed in Table 2. Time–frequency analysis revealed distinct variations in fault characteristics among these components, as shown in Figs. 7-10. For inner race faults, the frequency and intensity of fault impacts increased with speed in the time domain, whereas their amplitudes remained stable in the frequency domain. Outer race faults exhibited increased impact frequency and intensity in the time domain with higher speeds, alongside notable amplitude and frequency shifts in the frequency domain, making these faults more discernible under high-speed conditions. Cage faults maintained stable amplitude and frequency characteristics in the frequency domain despite elevated fault impact frequency in the time domain. Conversely, rolling element faults were strongly affected by speed in both domains, shifting toward higher frequencies and exhibiting greater amplitudes at high speeds. At lower speeds, 90 km/h, these signals were more susceptible to noise interference, potentially concealing underlying fault features.

The classification accuracy under variable-speed datasets is shown in

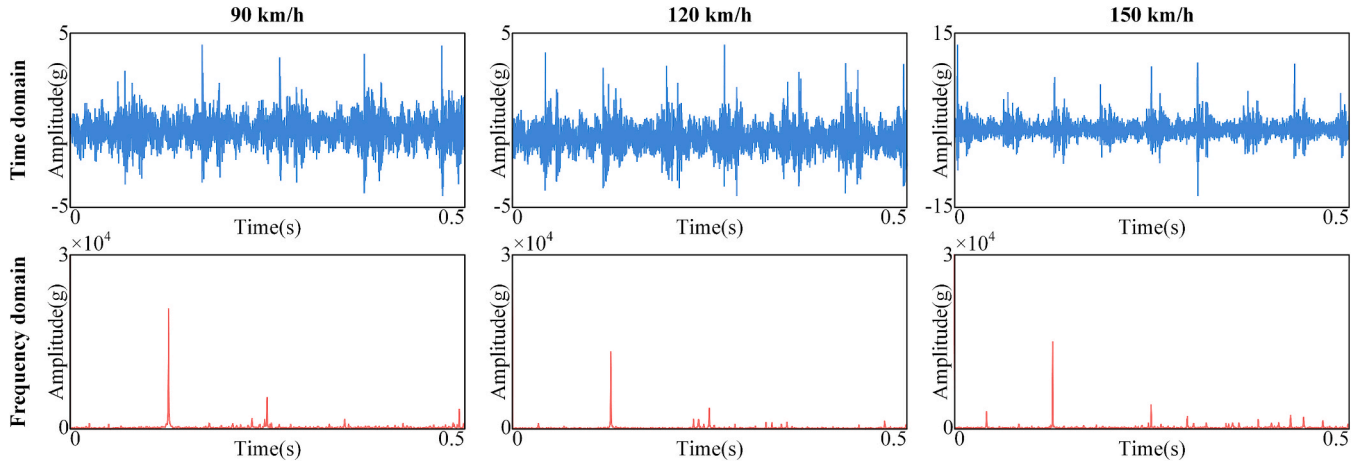


Fig. 7. Inner race fault time and frequency domain signal.

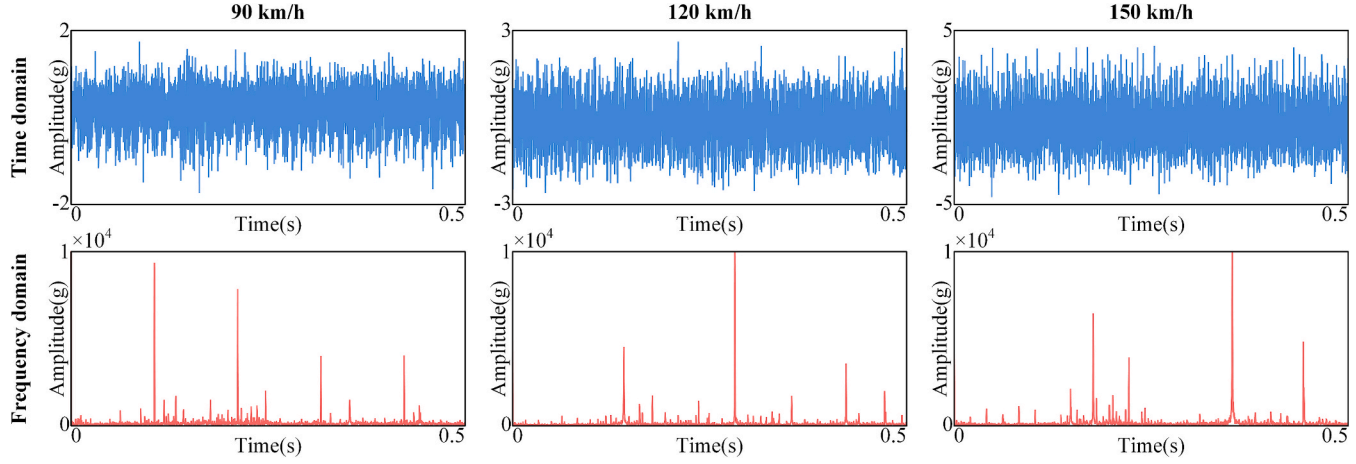


Fig. 8. Outer race fault time and frequency domain signal.

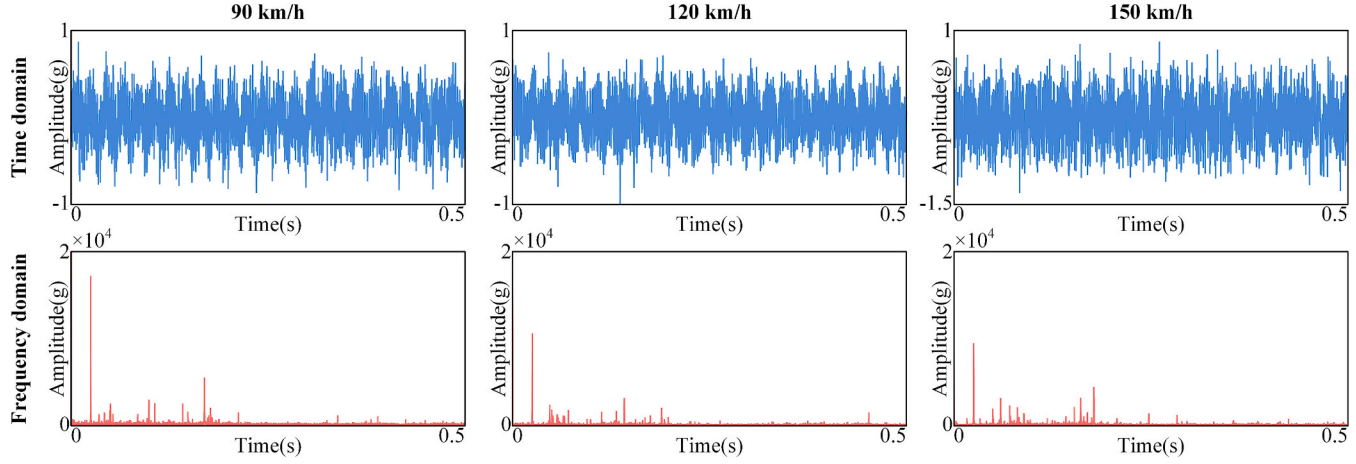


Fig. 9. Cage fault time and frequency domain signal.

Fig. 11. The proposed model demonstrated consistently high classification accuracy across all tasks, showcasing robust adaptability and its ability to effectively leverage data from both source and target domains to learn shared features for fault diagnosis. Notably, it exhibited strong performance across various rotational speed scenarios. In R1, where fault types in the source and target domains are identical and the proportion of noisy samples is low, all models achieved improved classification accuracy. AAnNet outperformed the proposed model in this scenario due to its dynamic sampling strategy, which adjusts model parameters based on data complexity. In contrast, the fixed structure of the proposed model limited its flexibility. In R2, the proposed model's advantages became evident. Both UDA-1 and UDA-2 experienced sharp

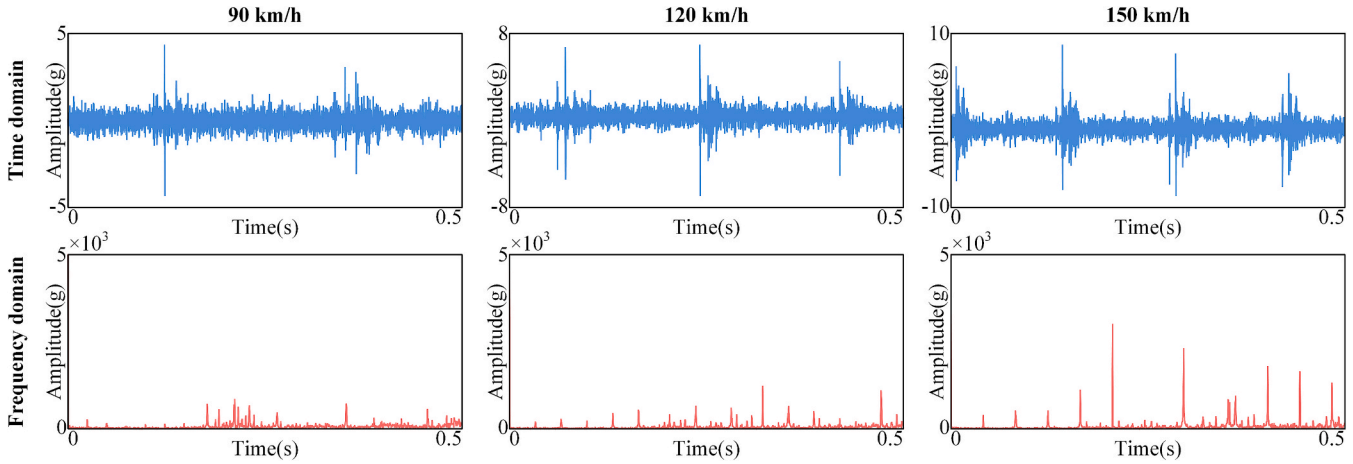


Fig. 10. Rolling element fault time and frequency domain signal.

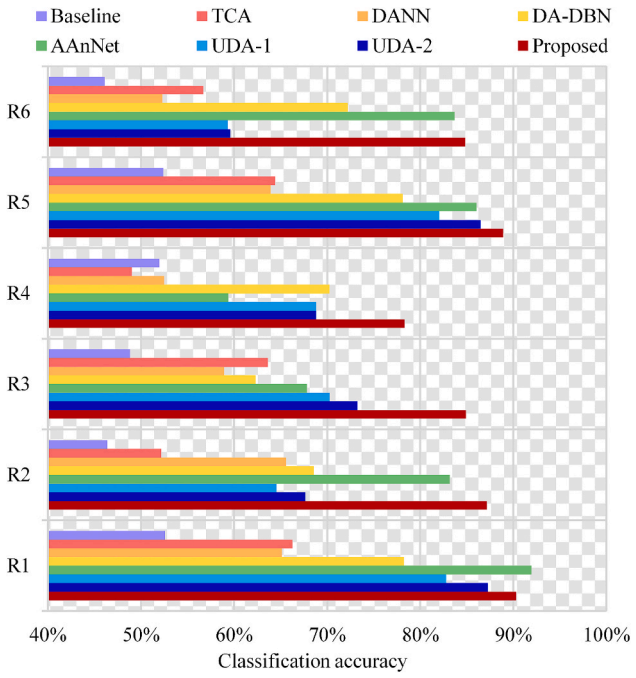


Fig. 11. Classification accuracy of each model under variable-speed dataset.

accuracy drops of 18.2 % and 19.6 %, respectively, due to their sensitivity to noise and issues related to overfitting or underfitting. By employing multiple classifier combinations, the proposed model effectively filtered useful samples and enhanced source-target domain alignment, ensuring robust performance. In R3, where there was a disparity in fault types between the source and target domains, DA-DBN and AAnNet faced challenges, leading to reduced classification accuracy and highlighting their limited robustness. The proposed model addressed this challenge through the implementation of Joint Divergence, significantly enhancing its generalization capability.

4.1.6. Timeliness Analysis

When evaluating the practical applicability of models in rail transportation, the focus extends beyond validating the accuracy of axle box bearing recognition to include timeliness indicators such as number of model parameters and multiply–add divide operations (MADDs). Experiments were conducted on the Jetson TX2 platform, repeated 20 times to ensure reliability, with the minimum values selected for analysis to minimize the impact of external task interference. The results are shown in Fig. 12, where circle sizes represent the number of model parameters.

4.1.7. Hyperparameter Analysis

In deep learning, batch-size and learning rate are critical hyperparameters that significantly influence model training and performance. Using the T1 sub-dataset from Table 1 as a case study, this study analyzes the impact of learning rate and batch-size on the proposed model classification accuracy, with results shown in Fig. 13.

Small batch-sizes enable models to leverage finer-grained sample distributions during each update. However, the limited sample count per batch increases gradient noise, leading to unstable optimization. For example, at a learning rate of 1e-1, the classification accuracy is relatively low, achieving 75.6 % for batch-size 16 and 80.3 % for batch-size 32. As the learning rate decreases, accuracy improves but fails to match the performance of larger batch-sizes. Small batch-sizes introduce significant gradient fluctuations, particularly under large learning rates, often leading to unstable optimization and suboptimal convergence.

Large batch-sizes significantly reduce gradient noise, stabilizing optimization dynamics. At a learning rate of 1e-5, batch-size 256 achieves the highest accuracy of 87.8 %. Medium batch-sizes balance training stability and computational efficiency, demonstrating better generalization. At a learning rate of 1e-3, batch-size 128 achieves an accuracy of 85.6 %, only 1.9 % below the highest accuracy attained by larger batches. This configuration effectively balances gradient

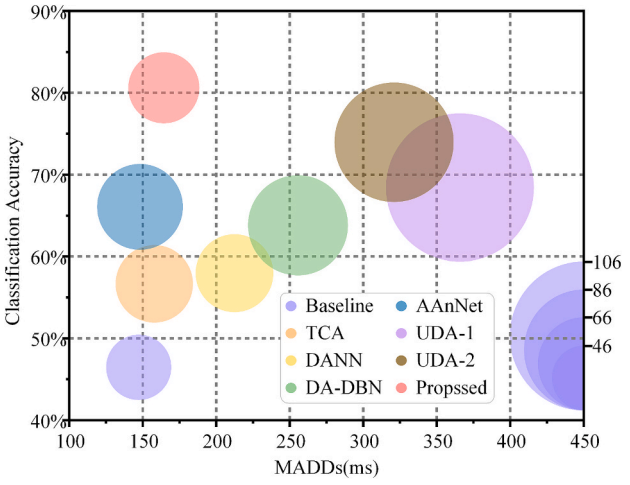


Fig. 12. The comparison between CNN-MC and the current mainstream intelligent models.

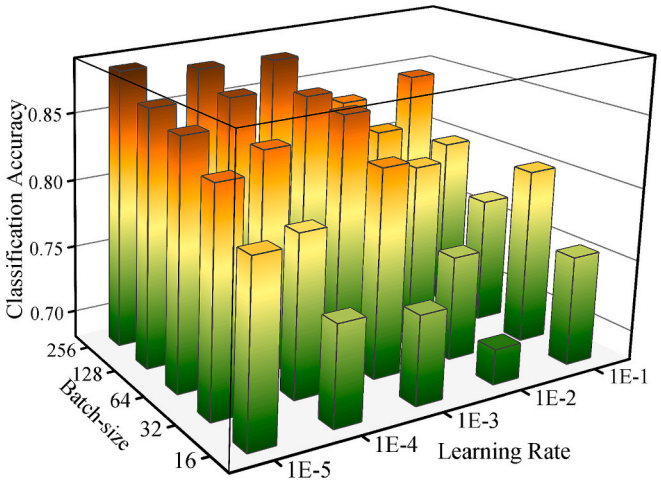


Fig. 13. The impact of learning rate and batch-size on the proposed model classification accuracy.

estimation precision and noise reduction, exhibiting strong stability during training.

4.2. Performance Analysis

4.2.1. Varying Size of $\overline{C_s}$ and .$\overline{C_t}$

To test the performance of the method in this paper under different $\overline{C_s}$ and $\overline{C_t}$, T2-supplement is proposed in this section. In T2 and T2-supplement, the proportion of noise sample is fixed at 0.3, and the type of source domain sample is changed, as shown in Table 4.

Each task is repeated 10 times, as shown in Fig. 14. The category of T2-supplement was set to $C/\overline{C_s}/\overline{C_t} = 4/0/0$, which can be considered as a WSDA task. In the T2-supplement, the proposed model is close to the DA-DBN model in terms of classification accuracy and has a smaller interquartile distance, which shows that it has better stability. In T2, the classification accuracy of the proposed model is 83.8 %, which is 5.6 % lower than that of the T2-supplement, showing good robustness. However, the classification accuracy of the DA-DBN model was 68.7 %, a decrease of 16.8 % compared to the T2-supplement. More interestingly, as can be seen from the box plots, the quartile distances and standard deviations of the DA-DBN model increase significantly. In summary, it can be seen that in the WSDA task, the proposed model has close accuracy and higher stability than the mainstream model; in the UDANS domain, the proposed model has a great advantage over the other models.

4.2.2. Varying noise level

In this section, the influence of proportion of noise samples on the model performance is tested by changing ρ. The values of ρ range from 0 to 0.3 with an interval of 0.05. Data of the source domain and target domain is obtained from T1, and the results are shown in Fig. 15.

It is noteworthy that when $\rho = 0$ (no noise samples in the source domain), the performance of the proposed model is close to that of the state-of-the-art UDA model. However, when $\rho = 0.3$, the proposed model still has a classification accuracy of 83.8 %, compared to 74.7 % and 71.2 % for UDA-1 and UDA-2 respectively. The above phenomenon proves that the proposed model is not susceptible to noise samples, and can extract fault feature information.

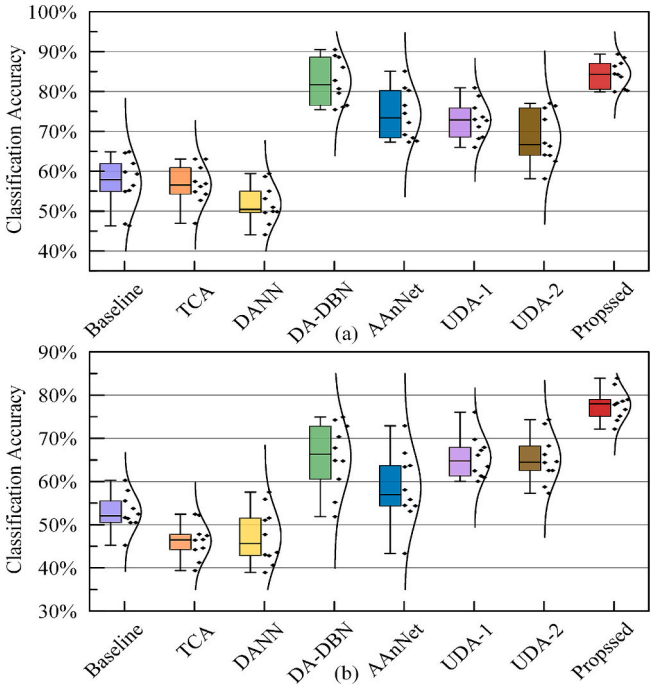


Fig. 14. Performance of different size of $\overline{C_s}$ and $\overline{C_t}$. (a) T2-supplement. (b) T2. In plot, the black line in the middle represents the median, the solid black points on the right show the classification accuracy for each times, and the black curve on the right represents the distribution of results.

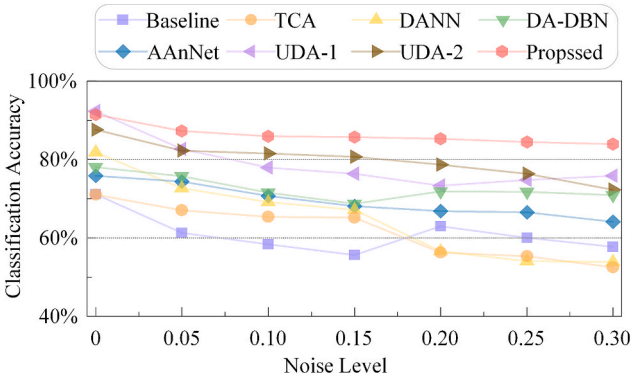


Fig. 15. The effect of noise levels on transfer models.

4.2.3. Visualizationss

Both source domain and target domain data are from T2, with 150 samples. To reduce the complexity of visualization, source domain and target domain only select three fault types. The same fault types in source domain and target domain are inner race fault and rolling

Table 4
Description of T2 and T2-supplement.

Task	Source Domain Categories	Target Domain Categories	Category Settings
T2	inner race fault, rolling element fault, cage fault, outer race fault	health condition, inner race fault, rolling element fault and outer race fault	$C/\overline{C_s}/\overline{C_t} = 3/1/1$
T2-supplement	health condition, inner race fault, rolling element fault and outer race fault	health condition, inner race fault, rolling element fault and outer race fault	$C/\overline{C_s}/\overline{C_t} = 4/0/0$

element fault. The sample type of private source domain is outer race fault, and the sample type of private target domain is cage fault. Notably, the distribution of these two private sample types is significantly distinct from each other. The learning decision boundaries of the proposed method and its variants are visualized in Fig. 16 as follows.

Without using the small loss selection The method of small loss selection in Eq. (9) is not used for training, which means that the objectives are as follows:

$$\min_{E,C_1,C_2} \mathrm{L}_s(D_s) \tag{18}$$

Without joint divergence The method of joint divergence in Eq. (7) is not used in training, and KL divergence is used instead, that is, the following formula is used instead of Eq. (7):

$$\mathrm{L}_{JD}(D_t) = \frac{1}{n_t}\sum_{i=1}^{n_t}\widetilde{\mathrm{L}}_{crs}(D_t) - \frac{1}{n_t}\sum_{i=1}^{n_t}\widetilde{\mathrm{L}}_{ent}(D_t) \tag{19}$$

Without separating the divergence between classifiers The method of small loss selection in Eq. (8) and Eq. (9) are not used in training, namely, the following formulas are used instead of Eq. (8) and Eq. (9):

$$\widetilde{\mathrm{L}}_{crs}(D_t) = 0 \tag{20}$$

$$\widetilde{\mathrm{L}}_{ent}(D_t) = 0 \tag{21}$$

As shown in Fig. 16(a), due to the noise samples in the source domain, a large area is marked as the target private class with large L_{crs}. At this time, the model cannot accurately extract the features of the source domain samples. For Fig. 16(b) and Fig. 16(c), although they are not affected by noise samples, they cannot well distinguish private samples in the target domain. Because their loss functions are not suitable for processing private category samples in the target domain. The proposed method, shown in Fig. 16(d), achieves the best performance compared to the other variants. In the proposed method, the classifier is not affected by the noise samples in the source domain. In addition, when there is a large divergence between the private samples and the common samples in the target domain, the proposed method attempts to increase the divergence of these private samples and reduce the divergence of the common samples.

5. Conclusion

This study proposed a novel transfer learning model that integrates a multi-classifier architecture, adversarial learning mechanisms, and a joint divergence loss to tackle the UDANS challenge in rail transportation. The proposed model integrates multiple classifiers, adversarial learning principles, and a joint divergence loss function to overcome the limitations of existing transfer learning methods. Experimental evaluations demonstrated that the proposed model substantially enhances fault diagnosis accuracy and robustness compared to classical and state-of-the-art transfer learning methods. Such advancements have immediate implications for urban rail systems, where frequent speed changes and varying passenger loads often lead to complex, noisy sensor data.

Despite these contributions, the proposed method exhibits limitations that merit future investigation. Chief among them is the high computational overhead, particularly when operating on high-dimensional datasets collected under diverse operating conditions. Additional opportunities include extending the method to other mechanical systems beyond rail transportation. By pursuing these directions, the adaptability and trustworthiness of the proposed approach may be further enhanced, thereby promoting safer and more reliable fault detection in a broader range of industrial applications.

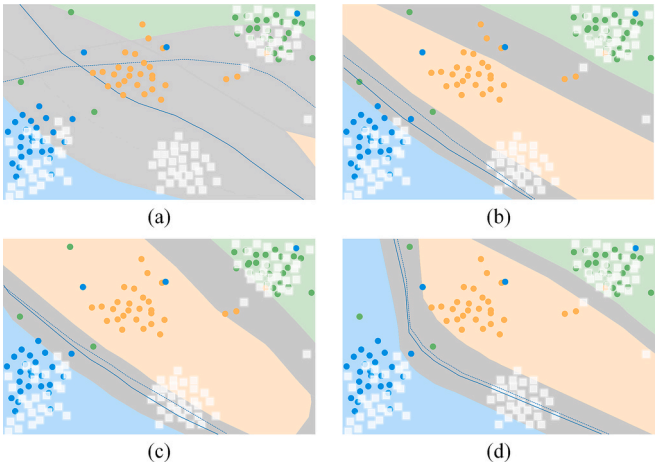


Fig. 16. Visualization of the learning decision boundaries. The proportion of noise samples in the source domain is set to $\rho = 0.15$. The white block represents the target domain sample, with the target domain private category located at the bottom right of Fig. 16. The blue lines and red dashed lines indicate the decision boundaries of the two classifiers. The light blue, light orange, and light green areas show where both classifiers classify samples as blue, orange, and green, respectively. The gray region represents the private class of the target domain, showing a significant difference.

CRediT authorship contribution statement

Yi Sun: Writing – review & editing, Writing – original draft, Visualization, Supervision, Conceptualization. **Hongliang Song:** Validation, Software, Project administration. **Liang Guo:** Resources, Funding acquisition, Formal analysis. **Hongli Gao:** Methodology, Investigation. **Ao Cao:** Resources, Investigation.

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.

Acknowledgment

This research was supported by the National Key R&D Program of China under Grant 2022YFB3402100, the National Natural Science Foundation of China under Grant 52275134, the Foundation of Sichuan under Grant 2023NSFSC1960 and the Sichuan Science and Technology Major Project under Grant 2022ZDZX0044.

Data availability

The authors do not have permission to share data.

References

[1] D.Z. Zhao, W.B. Cai, L.L. Cui, Adaptive thresholding and coordinate attention-based tree-inspired network for aero-engine bearing health monitoring under strong noise, Adv Eng Inform 61 (2024).
[2] Y. Qin, H.Y. Liu, Y.F. Mao, Faulty rolling bearing digital twin model and its application in fault diagnosis with imbalanced samples, Adv Eng Inform 61 (2024).
[3] J.Y. Wu, Y.L. Li, L.M. Jia, G.P. An, Y.F. Li, J. Antoni, G. Xin, Semi-supervised fault diagnosis of wheelset bearings in high-speed trains using autocorrelation and improved flow Gaussian mixture model, Eng Appl Artif Intel 132 (2024).
[4] G.H. Qin, K. Zhang, X.W. Lai, Q. Zheng, G.F. Ding, M.H. Zhao, Y.H. Zhang, An adaptive symmetric loss in dynamic wide-kernel resnet for rotating machinery fault diagnosis under noisy labels, Ieee T Instrum Meas 73 (2024).
[5] Y.P. Gao, D.F. Chang, C.H. Chen, Z.Y. Xu, Design of digital twin applications in automated storage yard scheduling, Adv Eng Inform 51 (2022).
[6] F. Li, C.H. Lee, C.H. Chen, L. Khoo, Hybrid data-driven vigilance model in traffic control center using eye-tracking data and context data, Adv Eng Inform 42 (2019).

[7] B. Pang, Q.H. Liu, Z.L. Xu, Z.D. Sun, Z.Y. Hao, Z.Q. Song, Fault vibration model driven fault-aware domain generalization framework for bearing fault diagnosis, Adv Eng Inform 62 (2024).
[8] K. Xu, X.G. Kong, Q.B. Wang, B. Han, L.Q. Sun, Intelligent fault diagnosis of bearings under small samples: a mechanism-data fusion approach, Eng Appl Artif Intel 126 (2023).
[9] X.J. Yao, X.C. Lu, Q.S. Jiang, Y.H. Shen, F.Y. Xu, Q.X. Zhu, SSPENet: Semi-supervised prototype enhancement network for rolling bearing fault diagnosis under limited labeled samples, Adv Eng Inform 61 (2024).
[10] Q.N. Zhu, H.Q. Liu, C.Y. Bao, J.M. Zhu, X.Y. Mao, S.P. He, F.Y. Peng, Decoupled interpretable robust domain generalization networks: A fault diagnosis approach across bearings, working conditions, and artificial-to-real scenarios, Adv Eng Inform 61 (2024).
[11] M.Y. Ye, X.A. Yan, N. Chen, Y. Liu, A robust multi-scale learning network with quasi-hyperbolic momentum-based Adam optimizer for bearing intelligent fault diagnosis under sample imbalance scenarios and strong noise environment, Struct Health Monit (2023).
[12] W. Cheng, X. Liu, J. Xing, X.F. Chen, B.Q. Ding, R.Y. Zhang, K.N. Zhou, Q. Huang, AFARN: Domain adaptation for intelligent cross-domain bearing fault diagnosis in nuclear circulating water pump, Ieee T Ind Inform 19 (2023) 3229–3239.
[13] P.F. Liang, J.Y. Tian, S.Y. Wang, X.M. Yuan, Multi-source information joint transfer diagnosis for rolling bearing with unknown faults via wavelet transform and an improved domain adaptation network, Reliab Eng Syst Safe 242 (2024).
[14] S.Z. Xie, P. Xia, H.Q. Zhang, Domain adaptation with domain specific information and feature disentanglement for bearing fault diagnosis, Meas Sci Technol 35 (2024).
[15] X.Y. Zheng, P.X. Yang, K. Yan, Y.Z. He, Q.J. Yu, M.Y. Li, Rolling bearing fault diagnosis based on multiple wavelet coefficient dimensionality reduction and improved residual network, Eng Appl Artif Intel 133 (2024).
[16] Y. Sun, H.L. Song, H.L. Gao, J. Li, S. Yin, Interpretable tool wear monitoring: Architecture with large-scale CNN and adaptive EMD, J Manuf Syst 78 (2025) 294–307.
[17] X.D. Sun, L. Chen, Z.B. Yang, H.Q. Zhu, Speed-sensorless vector control of a bearingless induction motor with artificial neural network inverse speed observer, Ieee-Asme T Mech 18 (2013) 1357–1366.
[18] Y. Zhang, J.C. Ji, Intelligent fault diagnosis of a reciprocating compressor using mode isolation convolutional deep belief networks, Ieee-Asme T Mech 26 (2021) 1668–1677.
[19] W.X. Zhang, Z.T. Ling, S. Heinrich, X.L. Ding, Y.G. Feng, Walking speed learning and generalization using Seq2Seq gated and adaptive continuous-time recurrent neural network (S2S-GACTRNN) for a hip exoskeleton, Ieee-Asme T Mech 28 (2023) 2375–2386.
[20] W.W. Qian, S.M. Li, P.X. Yi, K.C. Zhang, A novel transfer learning method for robust fault diagnosis of rotating machines under variable working conditions, Measurement 138 (2019) 514–525.
[21] A. Ainapure, X. Li, J. Singh, Q.B. Yang, J. Lee, Enhancing intelligent cross-domain fault diagnosis performance on rotating machines with noisy health labels, Procedia Manuf 48 (2020) 940–946.
[22] D.Y. Xiao, C.J. Qin, H.G. Yu, Y.X. Huang, C.L. Liu, J.W. Zhang, Unsupervised machine fault diagnosis for noisy domain adaptation using marginal denoising autoencoder based on acoustic signals, Measurement 176 (2021).
[23] B. Zhao, X.M. Zhang, Z.H. Zhan, Q.Q. Wu, Deep multi-scale adversarial network with attention: a novel domain adaptation method for intelligent fault diagnosis*, J Manuf Syst 59 (2021) 565–576.
[24] X.D. Wang, F. Liu, Triplet loss guided adversarial domain adaptation for bearing fault diagnosis, Sensors-Basel 20 (2020).
[25] W. Zhang, X. Li, H. Ma, Z. Luo, X. Li, Universal domain adaptation in fault diagnostics with hybrid weighted deep adversarial learning, Ieee T Ind Inform 17 (2021) 7957–7967.
[26] Z.W. Zhang, H.H. Chen, S.M. Li, Z.H. An, Unsupervised domain adaptation via enhanced transfer joint matching for bearing fault diagnosis, Measurement 165 (2020).
[27] Y. Li, L. Cheng, Y.X. Peng, Z.J. Wen, S.H. Ying, Manifold alignment and distribution adaptation for unsupervised domain adaptation, Ieee Int Con Multi (2019) 688–693.
[28] Y. Sun, S.C. Li, H.L. Gao, X.Q. Zhang, J.Z. Lv, W.X. Liu, Y.C. Wu, Transfer learning: a new aerodynamic force identification network based on adaptive EMD and soft thresholding in hypersonic wind tunnel, Chinese J Aeronaut 36 (2023) 351–365.
[29] Q. Yu, A. Hashimoto, Y. Ushiku, Divergence optimization for noisy universal domain adaptation, in: 2021 Ieee/cvf Conference on Computer Vision and Pattern Recognition, Cvpr 2021, 2021, pp. 2515–2524.
[30] K.M. He, X.Y. Zhang, S.Q. Ren, J. Sun, Deep residual learning for image recognition, Proc Cvpr Ieee (2016) 770–778.
[31] S.J. Pan, I.W. Tsang, J.T. Kwok, Q.A. Yang, Domain adaptation via transfer component analysis, Ieee T Neural Networ 22 (2011) 199–210.
[32] Y. Ganin, V. Lempitsky, Unsupervised domain adaptation by backpropagation, Pr Mach Learn Res 37 (2015) 1180–1189.
[33] C.C. Che, H.W. Wang, X.M. Ni, Q. Fu, Domain adaptive deep belief network for rolling bearing fault diagnosis, Comput Ind Eng 143 (2020).
[34] G.Q. Jin, T.Y. Zhu, M.W. Akram, Y. Jin, C.G. Zhu, An adaptive anti-noise neural network for bearing fault diagnosis under noise and varying load conditions, IEEE Access 8 (2020) 74793–74807.
[35] W. Zhang, X. Li, H. Ma, Z. Luo, X. Li, Open-set domain adaptation in machinery fault diagnostics using instance-level weighted adversarial learning, Ieee T Ind Inform 17 (2021) 7445–7455.

YiSun received the M.S. degree in mechanical engineering from Southwest Jiaotong University, Sichuan, P. R. China, in 2020, where he is currently pursuing the Ph.D. degree in mechatronic engineering. His research interests focus on the universal domain adaptation and health index construction.

Hongliang Song received the B.S. degree in mechanical engineering from Southwest Jiaotong University, Sichuan, China, in 2020. His research interests include machinery condition monitoring and structural load identification.

Liang Guo (Member, IEEE) received Ph.D. degrees in mechanical engineering from Southwest Jiaotong University, Sichuan, China in 2016. He is currently an Associate Professor at Southwest Jiaotong University. His research interests focus on machinery intelligent fault diagnostics.

Hongli Gao received the Ph. D degree in mechanical engineering from Southwest Jiaotong University, Sichuan, P. R. China in 2005. He is currently a Professor at Southwest Jiaotong University. His research interests focus on the reliability analysis of complex electromechanical equipment.

Ao Cao received the B.S. degree in mechanical engineering from Southwest Jiaotong University, Sichuan, China, in 2020, where he is currently pursuing the Ph.D. degree in mechatronic engineering. His research interests focus on the mechanical system dynamics.