Mechanical Systems and Signal Processing 205 (2023) 110846

Contents lists available at ScienceDirect

Mechanical Systems and Signal Processing

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

IDSN: A one-stage interpretable and differentiable STFT domain adaptation network for traction motor of high-speed trains cross-machine diagnosis

Chao He, Hongmei Shi *, Jianbo Li

School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, Beijing 100044, China
Collaborative Innovation Center of Railway Traffic Safety, Beijing 100044, China

ARTICLE INFO

Communicated by O. Fink

Keywords:
Differentiable signal processing
Interpretable AI
Cross-machine diagnosis
Transfer learning
High-speed trains

ABSTRACT

A surge of transfer fault diagnosis techniques has been proposed to guarantee the safe operation of traction motor systems. However, existing efforts highly depend on the availability of fault data in source domain, which is rare in practice due to the regular maintenance. Fortunately, self-customized testbeds provide an opportunity to easily obtain fault data, assuming that the simulated data can be utilized to monitor the real-world traction motor systems via the cross-machine diagnosis method. Besides, current deep learning-based cross-machine fault diagnosis methods suffer from the poor physical interpretability and the troublesome hype-parameter selection. To tackle aforementioned issues, a one-stage Interpretable and Differentiable STFT cross-machine dual-driven adaptation Network (IDSN) is proposed. In IDSN, a new paradigm termed interpretable differentiable STFT layer is devised, where a derivable coefficient is introduced to adjust pivotal parameters of STFT such as window length by the gradient descent. Prominently, it is a plug-and-play module, which can be embedded into the arbitrary typical network without conflict. Besides, a novel adaptive trade-off coefficient is developed to tackle the weight matching of the domain discrepancy metric. Finally, to ensure the reliability and effectiveness of cross-machine diagnosis, a concise yet valid smoothed joint maximum mean discrepancy is proposed, which simultaneously promotes intra-class compactness and inter-class separability. The results of experiments confirm that the proposed IDSN outperforms the state of the art.

1. Introduction

High-speed trains view the traction motor systems as the heart from which they derive sustaining power. As operation time increases, in comparison to standard mechanical components, the aging and wear of traction motor components will inevitably lead to irreversible features or parameter deviations, referred to as "faults" or "failures", which can occur in traction transformers, gearboxes, motor bearings, and so forth. Hence, it is of substantial utility to monitor the health of the traction motor systems [1]. As one of the three maintenance technologies of high-speed trains, Prognostics and Health Management (PHM) has become the primary technology for stable, safe and economic operation of traction motor systems. In practice, however, it often encounters some intractable problems: Non-Gaussianity, Nonlinearity, Dynamic operation, Big data, Heterogeneity, and Realtime capability. Artificial intelligence has considerable application potential with the capacities to adapt to ever-changing scenarios and continuously upgrade performance over time [2].

* Corresponding author at: School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, Beijing 100044, China.
E-mail addresses: chaohe@bjtu.edu.cn (C. He), hmshi@bjtu.edu.cn (H. Shi), jbli@bjtu.edu.cn (J. Li).

https://doi.org/10.1016/j.ymssp.2023.110846
Received 27 March 2023; Received in revised form 4 August 2023; Accepted 3 October 2023

Until now, the predecessors have carried out a substantial of extensive and thorough research on PHM for traction motor systems. Dynamical modeling-based [3], signal analysis-based [4–7], heuristic algorithms-based [8], shallow machine learning-based [9], and deep learning-based methods have been covered. Additionally, temperature signal-based [10,11] and acoustic emission signal-based [12] analysis methods have sprung up.

Industrial big data has become a reality thanks to breakthroughs in sensors and the accessibility of powerful computational resources [13,14]. Schemes based on signal processing and mechanism analysis entail extensive human expertise. However, deep learning-based methods with powerful end-to-end (one-stage) capacity are constantly emerging for fault diagnosis in traction motor systems of high-speed trains [15–20]. For the drive system of high-speed trains, Cheng et al. developed a collaborative deep learning fault identification method [21]. As a recent follow-up, a federal transfer neural network on the basis of variational autoencoder was put forward [22].

Focusing on PHM of high-speed trains exploiting transfer learning techniques, transfer learning is a continuous trend in fault diagnosis domain, and a significant research effort has been devoted [23,24]. Fine-tuning-based, domain discrepancy-based, domain adversarial-based and domain generalization-based approaches have been preliminarily investigated. Based on the thoughts of Fine-tuning, Bai et al. [25] converted vibration signals into spectral Markov transition field frequency spectrum images and identified wheelset-axlebox faults using AlexNet pre-trained by ImageNet data set. Inspired by the domain discrepancy-based insights, Qin et al. [26] proposed a stepwise adaptive convolutional neural network for the cross-running speed fault diagnosis of high-speed train bogies. Derived from the domain adversarial techniques, aiming at the train wheels of Lanxin Railway, Chen et al. [27] employed semi-supervised domain adversarial to align the marginal distribution of raw signals to achieve fault identification between different operating conditions, such as from high mean elevation to low mean elevation. For imbalanced fault diagnosis of train bearings, Ren et al. [28] then introduced the attention mechanism into the domain adversarial. Besides, Zhang et al. [29] introduced federated learning to protect the confidential train bogie fault data. Yu et al. [30] proposed conditional adversarial domain adaptation with discrimination embedding for cross-speed locomotive fault diagnosis. As for the direction of domain generalization, Hu et al. [31] applied continuous wavelet transform (CWT) and local maximum mean discrepancy (LMMD) to monitor the running state of brake pads in high-speed trains.

Despite the inspiring results of the prior gains, they are limited by two major issues that require attention:

(1) In light of the application scenarios of fault diagnosis for traction motor systems, innumerable efforts, including existing partial domain adaptation and domain generalization methodologies, resort to a multitude of annotation data gleaned from different working conditions but the same mechanical equipment, and exploit data in the source domain. However, the expensive and painstaking labeling cost precludes the usage of traction motor data in the source domain. Nonetheless, there are public, self-collected data available, which provides flexible solutions to tackle bearing diagnosis from simulated domain to the actually damaged traction motor domain. The cross-machine task poses particularly challenging because data collected from two different machines exhibits a more significant distribution discrepancy in contrast to cross-speed, cross-load tasks. In summary, no research concerns about cross different machines yet in the field of PHM for high-speed trains to date.

(2) In view of an algorithmic standpoint, the aforesaid works are purely data-driven models that, on the one hand, overlook the active role of differentiable signal processing-based interpretable “knowledge” in domain adaptation. On the other hand, these models are overly susceptible to the quantity and quality of the “data” set, leading to the results reported by different data sets vary greatly [32]. To break these limitations, dual-driven approaches have reached new heights because of the progressive transparency, interpretability, and analytic capabilities. In this regard, this work investigates the knowledge and data dual-driven cross-machine transfer network in traction motor bearings of high-speed trains, where minimal human expertise is in demand. In essence, the knowledge, namely differential STFT pertains to the new pre-processing paradigm, while data refers to the data-driven DAN.

In light of the data scarcity with fault labels in traction motor bearing data of high-speed trains, conventional supervised learning-based and same-machine transfer learning-based schemes cannot achieve the desired results. Cross-machine transfer learning is an alternative and optimal solution to solve this problem, which can transfer from labeled dataset to unlabeled traction motor bearing dataset between different machines. To address this challenge, this work proposes a knowledge and data dual-driven model integrating differentiable interpretable STFT “knowledge” and a smaller amount of “data”, which does not resort to the usage of traction motor fault data as the source domain.

This work postulates that prior knowledge—signal processing, can bolster the similarity between data sets from different machines, thereby decreasing the adverse effects of distribution shift, preventing negative transfer, which also can alleviate the heavy reliance on the quality and quantity of data, domain metric loss. The crux of this study is an exploration of whether integrating effective data pre-processing namely differential signal processing and cross-machine domain adaptation can facilitate cross-machine transfer diagnostic tasks. Holistic evaluation on five data sets indicates differentiable STFT is compatible with domain adaptation network, which can yield competitive outcomes in this domain.

Regrettably, differentiable signal processing as an extremely promising paradigm has yet to be prevalent in the field of fault diagnosis. This research thus aims to fill this gap. For one thing, physical-informed, interpretable, and differentiable STFT layer allows the interpretable input into DAN, which combines the interpretability of signal processing with the expressiveness of deep learning. For another thing, in non-differentiable signal processing algorithms, parameter selection strategies are task-specific or data-specific, which may not be suitable for arbitrary task, even using the same data set. Thus, there is no authoritative and robust hyper-parameter selection strategy that can be adapted to all data sets or transfer diagnostic tasks. In contrast, aiming at arbitrary data set and its corresponding fault diagnosis tasks, differentiable signal processing integrates the decisive hyper-parameters as part

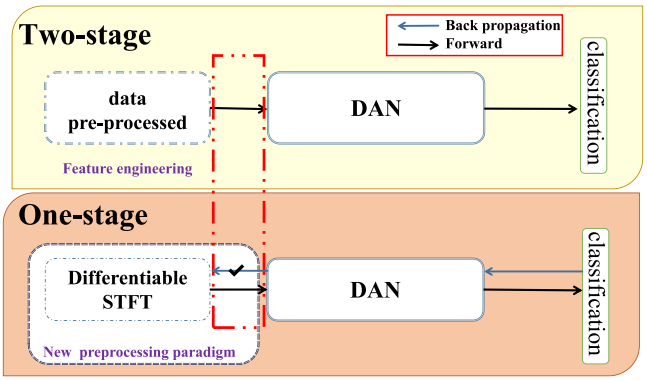


Fig. 1. One-stage vs. two-stage strategy.

of the parameters of neural networks, allowing them to be systematically embedded within neural networks to adaptively match the most appropriate hyper-parameters by back propagation without heavy reliance on human expertise. As such, continued research on differentiable signal processing has the potential to drive immense significance in fault diagnosis.

As shown in Fig. 1, IDSN is markedly distinct from current dual-driven research methods on account of the incorporation of differentiable STFT layers. Prior investigations generally adopt the two-stage strategy, where the raw signals must be pre-processed before being fed into the domain adaptation network. In recent years, some state-of-the-art methods prefer to analyzing raw signals by signal processing algorithms to capture the frequencies of faults, followed by an operation that uses them as weights of neural networks. These methodologies embed the "priori knowledge" into neural networks and force the neural networks to learn information related to the frequency of faults. However, the two-stage strategy resorts to precisely handcrafted hyper-parameters [33–36], because non-differentiable signal processing algorithms cannot be involved in the training process of networks.

The bulk of methods solely consider single domain information and intend to ignore time-frequency information. STFT, as one of the staple signal processing algorithms, has a broad range of applications [37]. In fault diagnosis, the two-stage strategy domain adaptation method based on none-differentiable STFT is generally targeted for simple transfer tasks such as cross-speed, cross-load [38,39], rather than cross-machine tasks as in this paper. In addition, it is evident that they will fail in the face of the defined scenarios as they rely on traction motor data utilized in the source domain. In contrast, IDSN has an end-to-end one-stage strategy due to the differentiable STFT, allowing for direct diagnostic output from the raw signal input.

In particular, IDSN is not a mere mechanical patchwork, but a differentiable STFT layer can be embedded in arbitrary deep structure to enable the input interpretable, the critical hyper-parameter of which is determined by back propagation algorithm. To the best of our knowledge, the ideology of interpretable differentiable signal processing algorithms has not been documented for transfer fault diagnosis across different machines [40–43], which is a striking distinguish between this study and majority cross-machine algorithms.

The aforementioned analysis elucidates the primary distinctions between the proposed IDSN and the existing STFT-based or dual-driven or cross-machine domain adaptation algorithms. Under the context, an interpretable STFT-based knowledge and data dual-driven cross-machine network is proposed to implement traction motor fault diagnosis from simulated to real-world domains.

In realm of domain adaptation, loss functions such as maximum mean discrepancy (MMD) [44], local maximum mean discrepancy (LMMD) [45], and joint maximum mean discrepancy (JMMD) [46] leave inter- and intra-class discriminative features out of consideration [47], for which smoothed joint maximum mean discrepancy (S-JMMD) is proposed. As illustrated in Fig. 2, due to the discrepancies between the distributions of the source and target domain, once the domain adaptation loss without smoothing strategy [48] (Fig. 2a), the features of some targeted samples will fall on the edge of the decision boundary, causing the classifier to misclassify them. On the contrary, by label smoothing (Fig. 2b), the distance between different classes in the source domain and the target domain can be enlarged [49], so it will contribute to increasing the decision margin between different faults. The misclassified samples can be thus close to the correct category. Noteworthy, a novel, concise yet valid perspective is concluded that label smoothing is compatible with domain adaptation in this paper.

The main four-fold contributions of the research are as follows:

(1) In light of the scenario on the scarcity of annotation traction motor fault data set, a novel alternative considering the prior knowledge of signal processing is proposed herein, that is, one-stage domain adaptation networks embedded in differentiable signal processing, which does not resort to the usage of traction motor fault data as the source domain.

(2) In contrast to existing cross-machine methodologies without interpretability that tend to devise intricate domain-shift statistical metrics or network structures, ISDN is proposed to integrate STFT into DAN for the first time, in which the hyper-parameters of STFT are part of the parameters of ISDN. They are dynamically updated using back-propagation algorithm. Meanwhile, ISDN is a dual-driven model, which draws on the knowledge-driven of differentiable STFT and the data-driven of neural networks. In particular, the differentiable STFT is embedded into ISDN as priori knowledge, enabling ISDN with intrinsic interpretability.

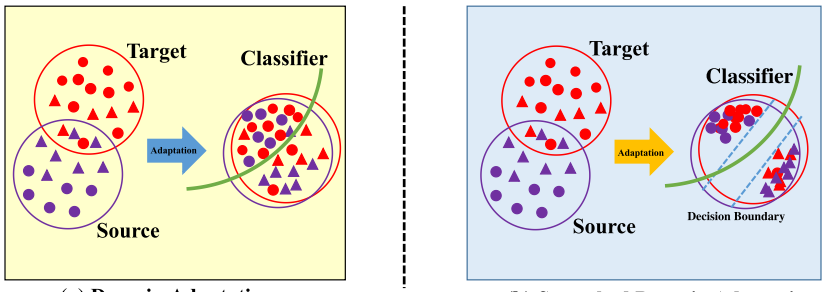


Fig. 2. Domain adaptation and smoothed domain adaptation.

(3) ISDN outperforms the prior two-stage method (signal pre-processing + DAN) in diminishing the distribution shift of source- and target-domain, making a simpler domain alignment process.
(4) To enhance the effect of domain adaptation, a novel statistical metric weight matching coefficient and smoothed JMMD loss are respectively, which contribute to the high performance in all experiments.

Section 2 is mostly concerned about prior works. In Section 3, IDSN is discussed in detail. Sections 4 and 5 includes several experiments and analyses to demonstrate the superior performance of IDSN. Section 6 is the interpretability analysis. This work form a conclusion in Section 7.

2. Preliminary knowledge

2.1. Practical scenario-based problem definition

In the source domain, the task cannot exploit traction motor data due to the expensive and laborious labeling cost, however there are public, self-collected data available, which provides flexible solutions to tackle from simulated to the actually damaged traction motor bearing diagnosis. The cross-machine task is particularly challenging because of more significant distribution discrepancy of data collected from two different devices in contrast to cross-speed, cross-load tasks.

Differentiable STFT is utilized to extract spectrograms that are suitable for transferring from a strongly supervised source domain $\mathscr{D}_s : \mathscr{X}^s = \{(x_i^s, y_i^s) | i = 1, 2, \ldots, n_s\}$ to an unsupervised target domain $\mathscr{D}_t : \mathscr{X}^t = \{x_j^t | j = 1, 2, \ldots, n_t\}$. Source domain is gleaned from one machine, where samples follow the distribution of P_s, and the label corresponds to $Y^s \in \mathbb{R}^R$. Corresponding, target domain is gleaned from another machine, where samples follow the distribution of P_t.

The purpose of joint distribution alignment (JDA) is to discover an optimization space $\mathscr{T}$ and minimize $\mathscr{T}$, which can be viewed as a combination of the marginal distribution and the conditional distribution.

The marginal MMD statistic of the marginal probability density (MDA) between different domains can be defined as:

$$\mathrm{MMD}^2_{\mathscr{H}}(P_s, P_t) = \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \phi(x_i) - \frac{1}{n_t} \sum_{j=1}^{n_t} \phi(x_j) \right\|^2_{\mathscr{H}} \tag{1}$$

where $\|\|^2_{\mathscr{H}}$ represents the reproducing kernel Hilbert space (RKHS), and $\phi(\cdot)$ is a kernel function map.

The conditional MMD statistic of the conditional probability density (CDA) between different domains can be defined as:

$$\mathrm{MMD}^2_{\mathscr{H}}(Q_s^{(c)}, Q_t^{(c)}) = \left\| \frac{1}{n_s^{(c)}} \sum_{x_i \in D_s^{(c)}} \phi(x_i) - \frac{1}{n_t^{(c)}} \sum_{x_j \in D_t^{(c)}} \phi(x_j) \right\|^2_{\mathscr{H}} \tag{2}$$

It is assumed that x_i^s and l_i^s denote FC layer and ground-truth labels of source domain, respectively, while t_i^s and l_i^s denote FC layer and pseudo-labels of target domain. Considering MDA and CDA, JMMD can be obtained:

$$\begin{aligned} \hat{L}_{\mathrm{JMMD}}(P, Q) = & \frac{1}{n_s^2} \sum_{i,j=1}^{n_s} \phi_1(x_i^s, x_j^s) \cdot \phi_2(l_i^s, l_j^s) \\ & - \frac{2}{n_s n_t} \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} \phi_1(x_i^s, x_j^t) \cdot \phi_2(l_i^s, l_j^t) \\ & + \frac{1}{n_t^2} \sum_{i,j=1}^{n_t} \phi_1(x_i^t, x_j^t) \cdot \phi_2(l_i^t, l_j^t) \end{aligned} \tag{3}$$

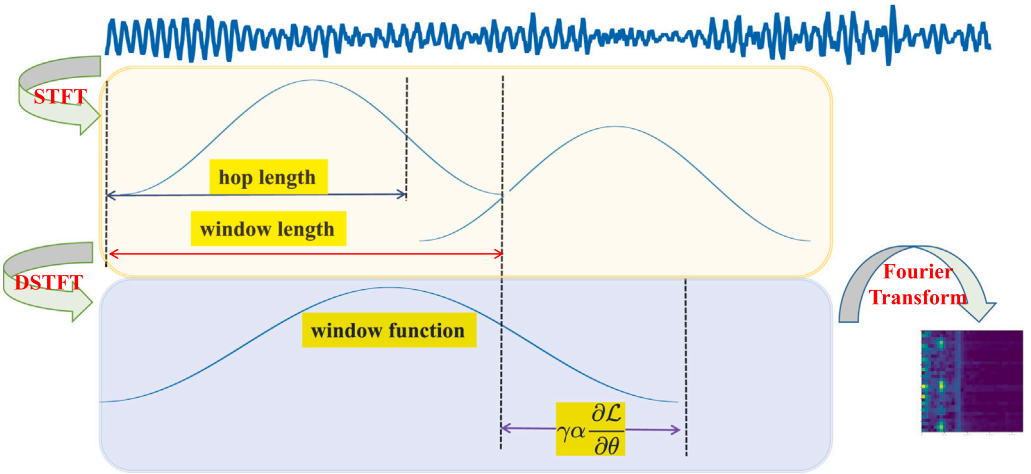


Fig. 3. Differentiable STFT layer.

2.2. Short-time Fourier transform

STFT is a staple time-frequency analysis algorithm for non-stationary signals. The 1-D signal is transformed into the 2-D feature matrix in the time-frequency domain, the idea of which is to intercept signals utilizing a window function with the fixed length and perform the Fourier transform on them. After translating on the time horizon using the window function, a growing number of spectrograms are obtained.

$$\mathrm{STFT}(t,f)=\int_{-\infty}^{+\infty}x(s)g(s-t)\cdot e^{-ifts}ds \tag{4}$$

where $x(s)$ is the signals to be processed, $g(s-t)$ is the window function, both s and t are time, and f is frequency.

3. The proposed dual-driven transfer model

3.1. Differentiable STFT layer

As shown in Fig. 3, the length of the window function directly determines the temporal resolution and frequency resolution of the spectrogram. Due to the game between time and frequency, a proper trade-off will provide the feature representation that is conducive to cross-machine domain adaptation. The link between window length and resolution can be expressed as [39]:

$$T_r=\left[\frac{N_x-N_o}{n-N_o}\right],F_r=\frac{n}{2}+\frac{2}{3+(-1)^{n-1}} \tag{5}$$

where N_x is the length of signals, N_o is the hop length, and n is the length of the window function $W_{t,\theta}$, which is adaptively adjusted by the gradient descent method. $[\cdot]$ denotes rounding to the nearest whole number. $\theta(\theta>0)$ is the derivable coefficient.

Note that, the vanilla STFT is non-differentiable on account of the non-derivable cyclic sliding operation performed by the window function in raw signals, thus presenting a limitation for usage in neural networks. In this regard, making the window length differentiable, this paper cleverly introduces a derivable coefficient θ that acts as a bridge between the coefficient and the window length. Specifically, a linear relationship between the two parameters is established, as shown in Eq. (7), allowing the derivable coefficient to participate in the back-propagation process, which in turn indirectly optimizing the window length and hop length by BP. As a result, the differentiable STFT makes an arbitrary loss function optimize the length of the window function, which is defined:

$$F_W[t,f]=\sum_{n=1}^{N-1}x[t+n]W_{t,\theta}[n]e^{-j\frac{2\pi}{N}fn} \tag{6}$$

where t is the sample position, $W_{t,\theta}[n]$ is the window function centered at the tth sample, $F_W[t,f]$ is denoted as STFT of t and f, j is the unit imaginary number.

During the optimization, this work takes the same window function and fixed hop/window ratio γ, where

$$n=\gamma\theta \tag{7}$$

$$N_o=\delta n \tag{8}$$

where δ is a hop coefficient.

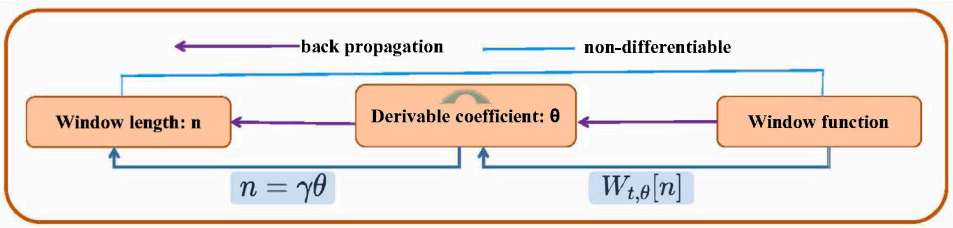


Fig. 4. How to make STFT differentiable.

Note that the window function is also bounded by the window size n and the derivable factor θ. As an illustration, let us consider the Hann window.

$$\begin{aligned} W_{t,\theta}^{Hann}[n] &= 0.5 - 0.5\cos(\tfrac{2\pi m}{n-1}), (0 \le m \le n-1) \\ &= 0.5 - 0.5\cos(\tfrac{2\pi m}{\gamma\theta-1}), (0 \le m \le \gamma\theta - 1) \end{aligned} \tag{9}$$

The window function is not derivable owing to its dependence on the window length, but a clever usage of Eq. (7) makes the window function derivable. In summary, as shown in Fig. 4, STFT becomes differentiable STFT, which can be regarded as an interpretable pre-processing paradigm that can be integrated into arbitrary network structure without conflict.

According to Eqs. (7), (8), and parameter update rule of neural networks, the update rule for θ, n, N_o is shown in Eq. (10).

$$\begin{aligned} &\theta \leftarrow \theta - \alpha\frac{\partial \mathscr{L}}{\partial \theta} \\ &\Rightarrow \gamma\theta \leftarrow \gamma\theta - \gamma\alpha\frac{\partial \mathscr{L}}{\partial \theta} \\ &\Rightarrow n \leftarrow n - \gamma\alpha\frac{\partial \mathscr{L}}{\partial \theta} \\ &\Rightarrow \delta n \leftarrow \delta n - \delta\gamma\alpha\frac{\partial \mathscr{L}}{\partial \theta} \\ &\Rightarrow N_o \leftarrow N_o - \delta\gamma\alpha\frac{\partial \mathscr{L}}{\partial \theta} \end{aligned} \tag{10}$$

where α is the learning rate exclusive to the derivable factor θ. The updating relationship between the derivable coefficient θ and window length and hop length is established through Eq. (10), which is crucial for the temporal and frequency resolution of the STFT spectrogram, obtaining a more reasonable resolution mapping.

In this research, the performance of several well-known window functions will be examined. Gaussian window, Bartlett window, Blackman window, Hamming window, Hann window and Kaiser window, to name a few, will be considered.

The salient property of the differentiable coefficient θ is that it is optimized by gradient back propagation and the chain rule, which in turn the window size n and the hop size N_o are optimized. Analytically, the differentiable θ can be expressed as follows:

$$\frac{\partial F_W[t,f]}{\partial \theta} = \sum_{n=0}^{N-1} e^{-j\frac{2\pi}{N}fn} x[t+n]\frac{\partial W_{t,\theta}[n]}{\partial \theta} \tag{11}$$

Assuming:

$$F_W[t,f]'_\theta = \frac{\partial F_W[t,f]}{\partial \theta} \tag{12}$$

θ is ultimately determined by optimizing the overall loss function $\mathscr{L}$, taking into account both the classification loss $\mathscr{L}_{cls}$ and the domain adaptation loss $\mathscr{L}_{DA}$. The final optimization process can be expressed as:

$$\frac{\partial \mathscr{L}}{\partial \theta} = \sum_{n=0}^{N-1} \frac{\partial \mathscr{L}}{\partial F_W[t,f]}\frac{\partial F_W[t,f]}{\partial \theta} = \sum_{n=0}^{N-1} \frac{\partial \mathscr{L}}{\partial F_W[t,f]} F_W[t,f]'_\theta \tag{13}$$

For the window length and hop size, according to Eqs. (7) and (8):

$$\frac{\partial n}{\partial \theta} = \gamma, \frac{\partial N_o}{\partial \theta} = \frac{\partial N_o}{\partial n}\frac{\partial n}{\partial \theta} = \gamma\delta. \tag{14}$$

Finally, according to Eq. (5), the time and frequency resolution of the spectrogram can be determined by θ:

$$\frac{\partial T_r}{\partial \theta} = \left[\frac{-N_x}{\gamma\theta^2(1-\delta)}\right], \frac{\partial F_r}{\partial \theta} = \frac{\gamma}{2} \tag{15}$$

Overall, differentiable STFT with adaptively adjusted window length simply demands for the authentic value of learnable parameter θ to impel the time/frequency resolution trade-off, which is high-performance.

3.2. Smoothed Joint Maximum Mean Discrepancy(S-JMMD)

In light of certain erroneously annotated samples in the source domain, label smoothing (LS) is utilized to prevent DAN from placing an excessive amount of faith the labels of the samples in the source domain. Prior research has revealed the potential of label smoothing in advancing knowledge distillation [50], while this work have investigated the bond of LS and domain adaptation (DA), indicating that label smoothing is compatible with domain adaptation.

The logits of the source domain with label smoothing are flattening, which serves to prevent DAN from overfitting the training samples of the source domain, thus enhancing the extrapolation capacity for the unseen targeted samples, which creates a favorable initialization for domain adaptation and works in tandem with the alignment loss to improve prediction accuracy. Therefore, this work deem that label smoothing that erases the correlation information in the intra-class is an advantage to distinguishing similarity fault features in domain adaptation. For label smoothing, the phenomenon that weakens the intra-class correlation information, can enforce the clustering centers of similar but different classes away from each other, i.e., in the process of domain adaptation, expand the distance between these the clustering centers and make the same fault cluster more closely, thus forcing these samples of different faults in the target domain to stay away from each other.

For a given backbone $\mathbf{D}$ and samples x_i^s, the logits of $\mathbf{D}$ are z_i^s, and after SoftMax, the category k^s of x_i^s satisfies $k^s \in [0, K]$: $p(k^s|x_i^s) = \frac{\exp(z_i)}{\sum_{i=1}^{K}\exp(z_i)}$. In general, for the ground-truth labels, the cross-entropy loss of the source domain $H^s(p,q)$ is calculated as:

$$H^s(p,q) = -\sum_{k^s=1}^{K} q(k^s|x_i^s)\log(p(k^s|x_i^s)) \tag{16}$$

where $q(k^s|x_i^s)$ is the ground-truth distribution of labels in source domain.

Considering label smoothing, k is the corresponding one-hot label, and $k^s \in \{0,1\}^K$, where element k_c^s is 1 for ground-truth class and 0 for others. For x_i^s with label y_i^s, consider the uniform label distribution $u(k^s) = \frac{1}{K}$ and the smoothing factor ε. $q\left(k^s|x_i^s\right) = \delta_{k^s,y^s}$ is superseded by:

$$\begin{aligned} q'(k^s|x_i^s) &= (1-\varepsilon)\delta_{k^s,y^s} + \varepsilon u(k^s) \\ &= (1-\varepsilon)\delta_{k^s,y^s} + \frac{\varepsilon}{K} \end{aligned} \tag{17}$$

Eq. (17) is substituted into Eq. (16) to derive the novel smoothing cross-entropy of the source domain:

$$\begin{aligned} \mathscr{L}_{cls} &= H^{s\prime}(p,q) \\ &= -\sum_{k^s=1}^{K} q'(k^s|x_i^s)\log(p(k^s|x_i^s)) \\ &= -\sum_{k^s=1}^{K}[(1-\varepsilon)\delta_{k^s,y^s} + \tfrac{\varepsilon}{K}]\log(p(k^s|x_i^s)) \end{aligned} \tag{18}$$

In process of back propagation, the gradient is calculated as:

$$\nabla\mathscr{L}_{cls} = \nabla H^s(p,q) = p(k^s|x_i^s) - (1-\varepsilon)\delta_{k^s,y^s} - \frac{\varepsilon}{K} \tag{19}$$

Assuming that the probability of change for label is fixed, the smoothed label can be expressed as:

$$z_i^{s\prime} = y_{smoothing}^s = (1-\varepsilon)y_{one-hot}^s + \frac{\varepsilon}{K} \tag{20}$$

The marginal probability distributions of the source and target domains are lowered by the smoothed cross-entropy loss and minimizing MDA.

In this work, the smoothed labels of the source domain and the pseudo-labels of the target domain are aligned by CDA, thus verifying the compatibility of label smoothing with domain adaptation. Using Eqs. (20) and (2), the conditional distribution alignment loss can be viewed as:

$$\begin{aligned} \mathscr{L}_{\mathrm{CDA}} &= \mathrm{MMD}_{\mathscr{H}}^2(Q_s^{(c)}, Q_t^{(c)}) = \left\| \frac{1}{n_s^{(c)}} \sum_{x_i \in D_s^{(c)}} \phi(z_i^{s\prime}) - \frac{1}{n_t^{(c)}} \sum_{x_j \in D_t^{(c)}} \phi(z_j^t) \right\|_{\mathscr{H}}^2 \\ &= \left\| \frac{1}{n_s^{(c)}} \sum_{x_i \in D_s^{(c)}} \phi[(1-\varepsilon)y_{one-hot}^s + \tfrac{\varepsilon}{K}] - \frac{1}{n_t^{(c)}} \sum_{x_j \in D_t^{(c)}} \phi(z_j^t) \right\|_{\mathscr{H}}^2 \end{aligned} \tag{21}$$

Eqs. (1) and (21) are synthetically considered, and S-JMMD is written as:

$$\mathscr{L}_{\mathrm{S-JMMD}} = \left\| \frac{1}{n_s} \sum_{i,j=1}^{n_s} \phi_1\left(x_i^s, x_j^s\right)\phi_2\left(z_i^{s\prime}, z_j^{s\prime}\right) - \frac{1}{n_t} \sum_{i,j=1}^{n_s} \phi_1\left(x_i^t, x_j^t\right)\phi_2\left(z_i^t, z_j^t\right) \right\|_{\mathscr{H}}^2 \tag{22}$$

Under the strong supervision of hard labels, the maximum position of one-hot is infinitely close to 1 due to the distribution of SoftMax activation function. To a certain extent, the optimization will reach the saturation zone at low efficiency, while the smoothing label can ensure that the optimization process always stays in the middle region of the highest optimization efficiency and escape the saturation zone.

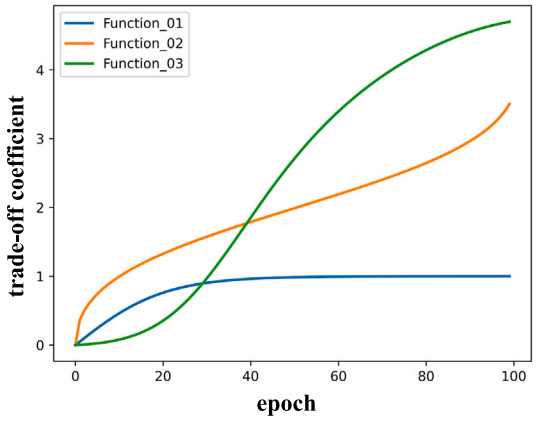


Fig. 5. Different trade-off coefficients.

The main difference between S-JMMD and JMMD is that the pseudo labels using S-JMMD are deducted by smooth label-guided network, while JMMD is inferred by hard labels.

In the training process of JMMD, the pseudo-label of the target domain matches the ground-truth label of the source domain, while S-JMMD is to make the smoothed label of the source domain match the SoftMax distribution of the target domain. Intuitively, the SoftMax distribution inherently contains certain valuable knowledge that it may reveal that the sample is most likely to be an inner ring fault, unlikely to be an outer ring fault, but never the normal state. In summary, in S-JMMD, the smoothing-guided soft label of the target domain statistically analyzes the valid information of the data set, which retains the inter-class correlation information and eliminates some invalid and redundant information.

Finally, the total loss function of IDSN is:

$$\mathscr{L} = \mathscr{L}_{cls} + \lambda \mathscr{L}_{\text{S-JMMD}} \tag{23}$$

where λ is the trade-off factor.

3.3. The devised trade-off coefficient

As shown in Fig. 5, when IDSN learn domain invariant features in the later stage, the existing adaptive trade-off coefficient adjust slowly, leading in a sluggish updating of parameters (**Function_01**). So it is time-consuming to learn the domain invariant features, and even generates negative transfer [51]. Luo et al. [52] proposes a novel coefficient, but the fast increment of the coefficient in the early period cannot completely fit the source domain, which can hinder the domain-invariant feature learning in the later period of training process and thereby impair the extrapolation ability (**Function_02**). For this reason, a novel coefficient strategy (**Function_03**) is proposed to allow IDSN to better learn domain-invariant time-frequency features that favor transfer diagnostic tasks. where the slow increment of the coefficients in the early training process can focus on fitting the source domain, and in the later stage, it focuses primarily on the domain-invariant feature learning with fast increment of the coefficient, at the maintenance of generality and significant performance.

As illustrated in Fig. 5, the corresponding function representations of **Function_01** and **Function_02** are:

$$\lambda_{Function\_01} = \frac{2}{\exp(\frac{-10epoch}{\max\_epoch})} - 1 \tag{24}$$

$$\lambda_{Function\_02} = \frac{-4}{\sqrt{\frac{epoch}{\max\_epoch}} + 1} + 4 \tag{25}$$

The proposed novel trade-off factor $\lambda^* = \lambda_{Function\_03}$ is shown in Eq. (26).

$$\lambda^* = \lambda_{Function\_03} = \frac{epoch * \cos(\frac{epoch}{4\max\_epoch}\pi)}{15 * [1 + \exp(\frac{-10epoch}{\max\_epoch} + 3)]} \tag{26}$$

The trade-off coefficient grow monotonically in Fig. 5. Compared with the generic settings, λ^* update slowly in the early stage of optimization, and the smoothed cross-entropy loss $\mathscr{L}_{cls}$ occupies the prominent role, which is conducive for IDSN to learning the features of the source domain and circumventing the underfitting phenomenon. As epoch increases, the proportion of $\mathscr{L}_{\text{S-JMMD}}$ in $\mathscr{L}$ evolves larger and larger, and λ^* increases from fast to slowly (the slope changes from considerably to tiny), thus pushing IDSN to learn the domain-invariant features.

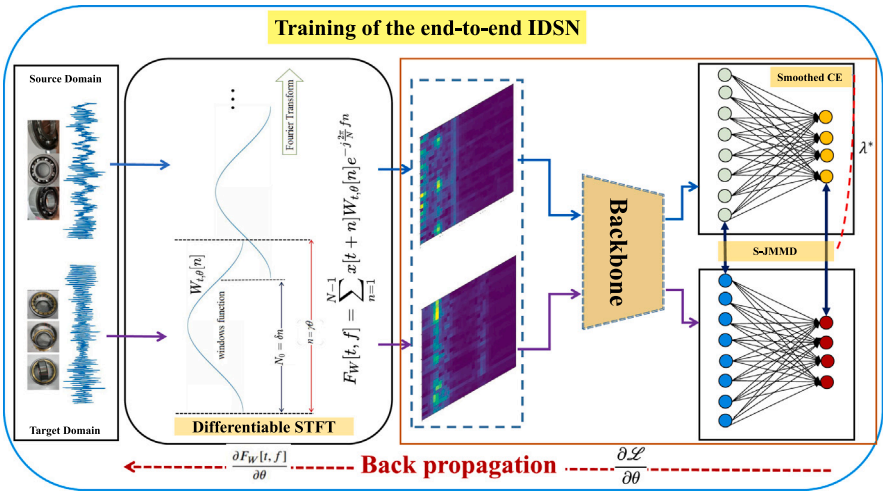


Fig. 6. Training of the one-stage IDSN model.

The proposed S-JMMD matches the smoothed labels with pseudo-labels to increase the intra-class aggregation and enlarge the inter-class distance while aligning the joint distribution. Finally, domain confusion achieved.

The gradients $\mathscr{L}$ corresponding to the training parameters $\mathscr{F}$ of IDSN are:

$$\nabla\mathscr{L} = \nabla\mathscr{L}_{cls} + \lambda^*\nabla\mathscr{L}_{\mathrm{S-JMMD}} \tag{27}$$

where

$$\nabla\mathscr{L}_{\mathrm{S-JMMD}} = \begin{cases} \frac{2}{n_s}[\frac{1}{n_s}\sum_{i=1}^{n_s}\sum_{j=1}^{n_s}\phi_1(x_i^s, x_j^s)\phi_2(z_i^{s\prime}, z_j^{s\prime}) - \frac{1}{n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_s}\phi_1(x_i^t, x_j^t)\phi_2(z_i^t, z_j^t)], & \mathscr{F} \in \mathscr{X}^s \\ \frac{2}{n_t}[\frac{1}{n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_s}\phi_1(x_i^t, x_j^t)\phi_2(z_i^t, z_j^t) - \frac{1}{n_s}\sum_{i=1}^{n_s}\sum_{j=1}^{n_s}\phi_1(x_i^s, x_j^s)\phi_2(z_i^{s\prime}, z_j^{s\prime})], & \mathscr{F} \in \mathscr{X}^t \end{cases} \tag{28}$$

Adam is used to update $\mathscr{F}$.

$$\mathscr{F} \leftarrow \mathscr{F} - [\alpha\frac{\partial\mathscr{L}}{\partial\theta} + \beta\lambda^*(\nabla\mathscr{L}_{cls} + \nabla\mathscr{L}_{\mathrm{S-JMMD}})] \tag{29}$$

where α and β denote the learning rate. In order to search for θ over a larger range, α will take a larger value.

3.4. Pipeline of IDSN for cross-machine transfer diagnosis

Fig. 7 exhibits the workflow of the proposed knowledge and data dual-drive cross-machine transfer learning framework, with analogue data as the source domain and actual damaged faults to the traction motor bearings of high-speed trains as the target domain. From the diagram of IDSN (Fig. 6), IDSN with S-JMMD loss implements the deep binding and systematic connection between differentiable STFT and DAN. Specifically, the STFT with differentiable window length processes the raw time-domain signals to produce the spectrograms, and they are fed to DAN as input, and both the training parameters of DAN and the hyper-parameter of the STFT are update by means of a back-propagation algorithm, which is a one-stage strategy. Thus with adjustable trade-off coefficients, smoothed cross-entropy loss, and S-JMMD loss, IDSN finally is trained to accomplish the cross-machine diagnosis tasks. The detailed training procedure is displayed in Algorithm 1.

Post-training, the training model parameters are saved. Then load the model parameters and input the testing set from target domain into the saved model to get the predicted labels, and the predicted labels are compared against the ground-truth labels to get the test accuracy.

4. Application to traction motor bearings of high-speed trains

4.1. The description of data sets

To validate the effectiveness of IDSN, IDSN is evaluated against two public data sets, self-gleaned bearing and gearbox data sets, and an actual damaged bearing data set for high-speed train traction motors as shown in Table 1. The experiments are divided into two major tasks: fault identification with the same distribution and transfer diagnosis across machines. The profiles of the data sets are displayed below:

Algorithm 1 The training process of IDSN.

Input: $\mathscr{D}_s$: source domain; $\mathscr{D}_t$: target domain; batch size; α: learning rate of θ; β: learning rate; **D**: model; λ^*: the trade-off coefficient; ε: the smoothing factor;

Output: Parameters of IDSN: $\mathscr{F}$.

for D not converged **do**
  for $epoch = 0$ to max_epoch **do**
    A batch samples (x_i^s, y_i^s) from $\mathscr{D}_s$
    A batch samples x_j^t from $\mathscr{D}_t$
    Extract domain invariant features using **D**
    Optimize θ by Eq. (13)
    Output pseudo labels of x_j^t: z_i^t
    Calculate smoothed labels x_i^s by Eq. (20)
    Calculate smoothed cross entropy loss $\mathscr{L}_{cls}$ by Eq. (18)
    Calculate S-JMMD loss $\mathscr{L}_{\text{S-JMMD}}$ by Eq. (22)
    Calculate and update λ^* by Eq. (26)
    Optimize the parameters of IDSN: $\mathscr{F}$ by Eq. (29)
  end for
end for

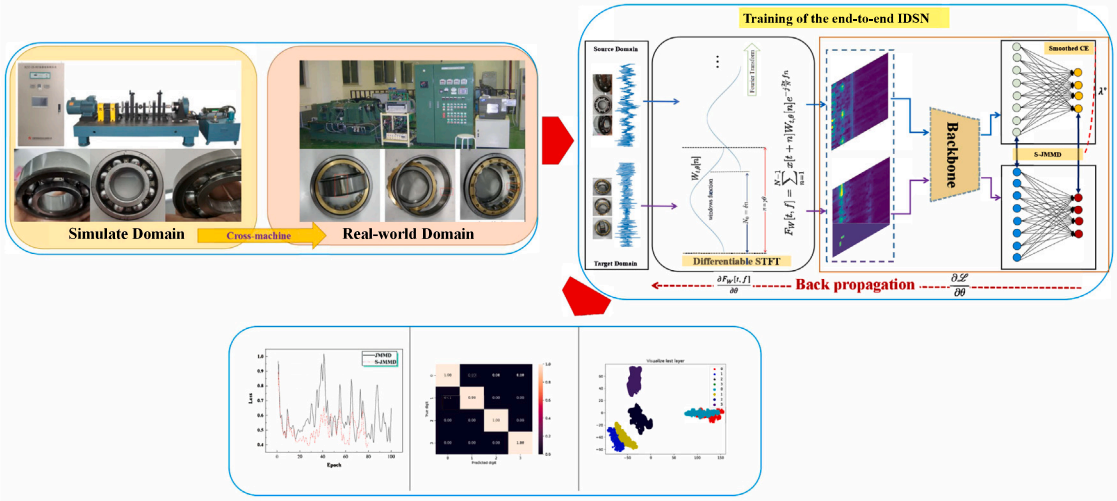
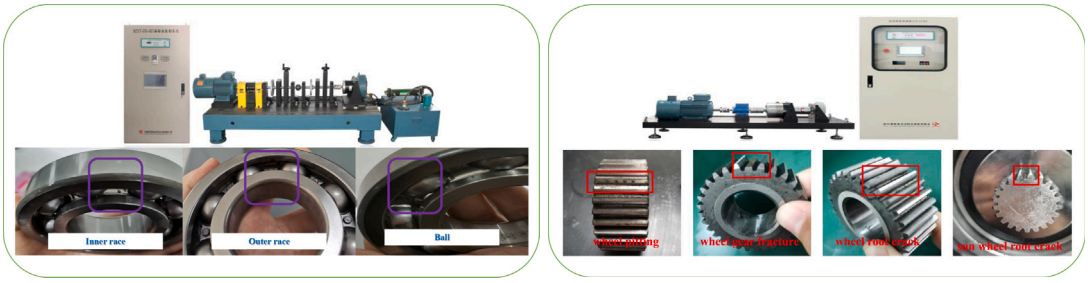


Fig. 7. The overall pipeline of IDSN.

Table 1
The detailed description of data sets.

Name	Data source	Speed	Sampling frequency	Health status
A_0	JNU	600 r/min	50 kHz	N,IF,OF,BF
A_1		800 r/min		
A_2		1000 r/min		
B	CWRU	1797 r/min	48 kHz	N,IF,OF,BF
C_0	DDB	900 r/min	10 kHz	N,IF,OF,BF
C_1		900 r/min	20 kHz	
C_2	DDP	1800 r/min	10 kHz	H,PP,PB,PC,SC
C_3		2500 r/min		
D_0	NTN	165 km/h	100 kHz	N,IF,OF,BF
D_1		200 km/h		
D_2		250 km/h		
D_3		300 km/h		
D_4		350 km/h		



(a) Rotor-gear comprehensive fault experiment platform.

(b) Gearbox fault experiment platform.

Fig. 8. Fault experiment platform.

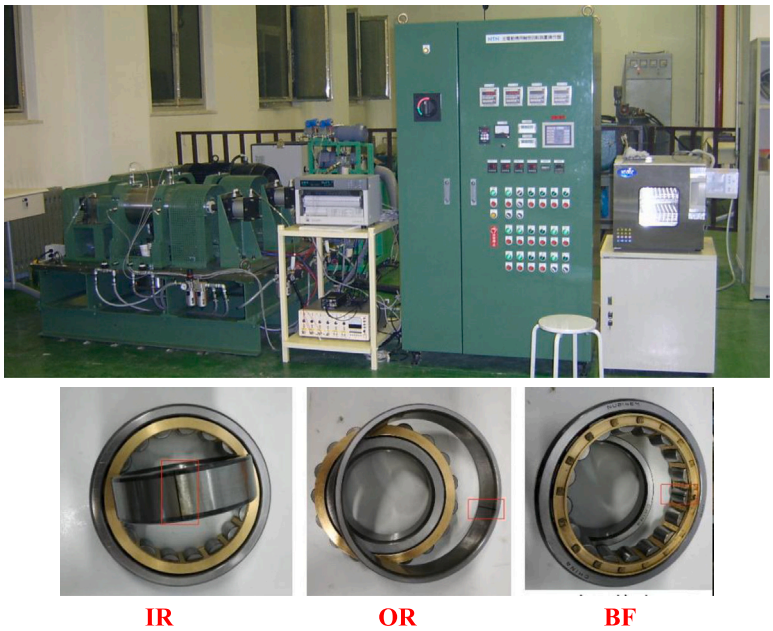


Fig. 9. High-speed train traction motor bearing platform.

(1) CWRU [53]: It is widely utilized as a benchmark data set in fault diagnosis. Raw signals from four states at 48 kHz, 0 hp are used: normal (NC), inner ring fault (IR), outer ring fault (OR), and ball fault(BF).

(2) JNU [54]: Jiangnan University data set has a greater difficulty identifying faults in preliminary experimental analysis, the sampling frequency of which is 50 kHz and are gleaned from four states at 600 r/min, 800 r/min, 1000 r/min: NC, IF, OF, BF.

(3) DDB [38]: The double-span dual-rotor fault test bench, depicted in Fig. 8(a), consists of a control cabinet, variable frequency motor, drive motor, loading device, single-span rolling bearing rotor system, sensors. Using a deep groove ball bearing called NSK-6308, the vibration acceleration signals are gleaned by the triaxial ICP piezoelectric sensor placed on the bearing housing with the sampling frequency 10, 20 kHz, which contains NC, IF, OF, BF.

(4) DDP [55]: The planetary gearbox fault test bench is shown in Fig. 8(b). Four types of gearbox failures are simulated at different locations, including planetary wheel pitting (PP), planetary wheel gear fracture (PB), planetary wheel root crack (PC), and sun wheel root crack (SC), and additionally including healthy planetary gearbox (H).

(5) NTN: The data set obtained from the NTN traction motor bearing experimental bench of Beijing Jiaotong University. As shown in Fig. 9, it is a special equipment for high-speed train traction motor bearing experiments with an advanced international level. The vibration acceleration sensor is fixed on the outer side of the bearing and the sampling frequency is set to 100 kHz. The bearing model used for the experiment is HRB NU214 EM 32214H, with the same dimensions of the actual bearing. Therefore, four types of mechanical states, including three types of failures and healthy bearings: IF, OF, BF, NC. The bearing rotation speeds are equivalent to the speeds of high-speed trains: 165 km/h, 200 km/h, 250 km/h, 300 km/h and 350 km/h.

Table 2
The list of parameter settings.

Settings	Batch size	γ	δ	Backbone	Window function	N_0	θ	max_epoches	β	seed
Values	16	6	0.5	ResNet-18	Gaussian	$0.5n$	$n/6$	100	0.001	3407 [56]

4.2. Task description

Sliding window sampling is used to divide the raw signals and expand the samples with a length of 1024, including overlapping points between two neighboring samples.

In tasks of the fault identification from the same distribution, the numbers of the training set, validation set, and test set are 280, 70, 150 for each class.

In cross-machine transfer diagnosis tasks, the number of the source domain and the target domain is 20 for each category. So there are 20K samples in the source domain and the target domain, respectively, where the test set is from the target domain with a total of 480K samples (K is the number of classes).

Also, according to Ref. [35], these transfer tasks also belong to the small sample diagnosis tasks.

Multiple cross-machine transfer diagnostic tasks are constructed, $\mathrm{A}_1 \rightarrow \mathrm{D}_1$ denoting A_1 as the supervised source domain and D_1 as the unsupervised target domain. Also, the models are consistently implemented on PyTorch 1.12.0 with the GPU of NVIDIA Tesla V100 32 GB. The experiments are repeated five times to take the average, using the unified random seed with fair comparison.

4.3. Parameter setting

The values of parameters play an essential part in deep learning. The generic parameter settings are shown in Table 2, which are not explicitly discussed. In this work, some parameters are pivotal and will be researched specifically.

(1) $n \in [16, 32, 64, 128, 256, 512, 1024]$: The initial value of the window length, generally set by artificial experience, determines the size that IDSN will ultimately seek. It is set to 128.
(2) $\alpha \in [0.9, 1.0, 1.2, 2.0]$: To make θ have a wider search range, use a larger learning rate, generally set to 1.2.
(3) $\varepsilon = 0.1 \sim 0.55$: The smoothing coefficient is to determine the degree of trust of the model for the training samples, and it is untrue that the larger the value is, the better it is for learning discriminative features. Generally, the parameter ε is set to 0.1.

In the experimental setup, this work do not utilize heavy training, validation, and testing samples, where $\mathrm{A}, \mathrm{C}_0, \mathrm{C}_1$ and D are bearing data sets and $\mathrm{C}_2 \sim \mathrm{C}_3$ are gearbox data sets.

4.4. Case 1: Fault identification from the same distribution

As shown in Fig. 10, the average accuracy of ResNet-18 with adaptive window length is 99.65% on the seven data sets, while 98.47% for fixed window length. It is amply illustrated the vital value of differentiable data pre-processing prior to input to the neural networks.

A reasonable window length of STFT can make the time resolution and frequency resolution trade-off, highlighting helpful information conducive to fault identification, thus exhibiting promising performance. Compared with manually setting the window length, through the gradient descent algorithm, optimizing it to select an optimal and task-driven time/frequency trade-off is more beneficial for the interoperability of the new data preprocessing paradigm and neural networks. Although the tasks of fault identification with the same distribution is not the crux of this study, it lays a solid foundation for the subsequent cross-machine transfer diagnosis.

4.5. Case 2: Fault identification from the different distribution

4.5.1. Two-stage vs. one-stage

Fig. 11 showcases the results of cross-dataset transfer utilizing the one-stage or two-stage training strategies on the public data sets CWRU and JNU. According to Ref. [57] for the parameter settings of pre-processing, both Wavelet-DAN and STFT-DAN belong to two-stage strategies, i.e., raw signals processed by wavelet transform or STFT before feeding them to the DAN, where the hyper-parameters are constant and artificially specified. In contrast, IDSN is a one-stage strategy that employs the back-propagation algorithm to determine the window length and hop length of STFT. By adaptively selecting a suitable parameter, IDSN can better cope with the concrete transfer task while purifying source- and target-domain distribution shift. The results demonstrate that IDSN is able to intelligently select the most appropriate window length and hop length for different tasks and data sets to construct spectrograms to assist in achieving high-performance cross-machine diagnosis. The proposed method outperforms the two-stage strategy with an average score of over 96% in six tasks. The performance of STFT-DAN and IDSN is comparable in the $B \rightarrow A_0$ and $B \rightarrow A_1$ due to the high similarity between the two data sets, indicating that employing a two-stage strategy can achieve a single-stage similar performance on the similar data distribution shift, but IDSN also outperforms STFT-DAN.

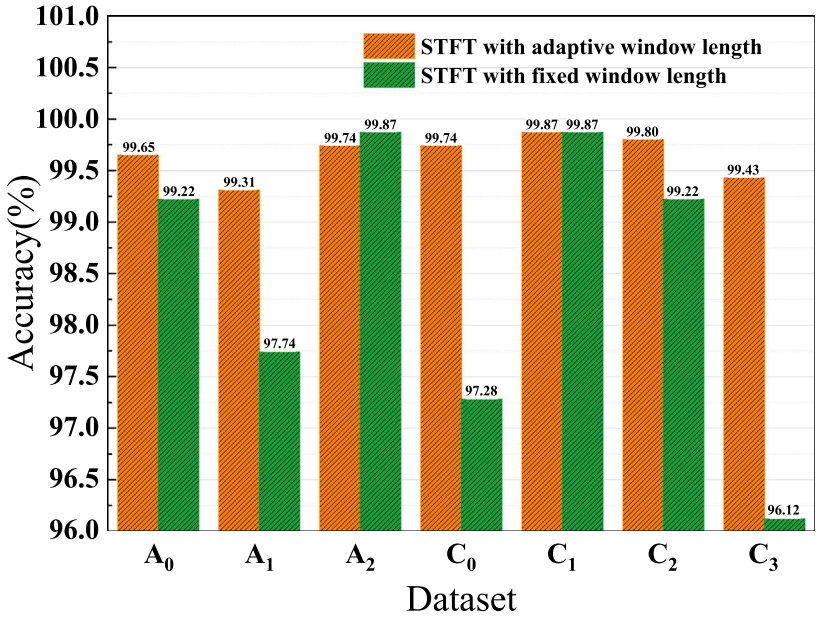


Fig. 10. Performance with adaptive/fixed window length of different data sets.

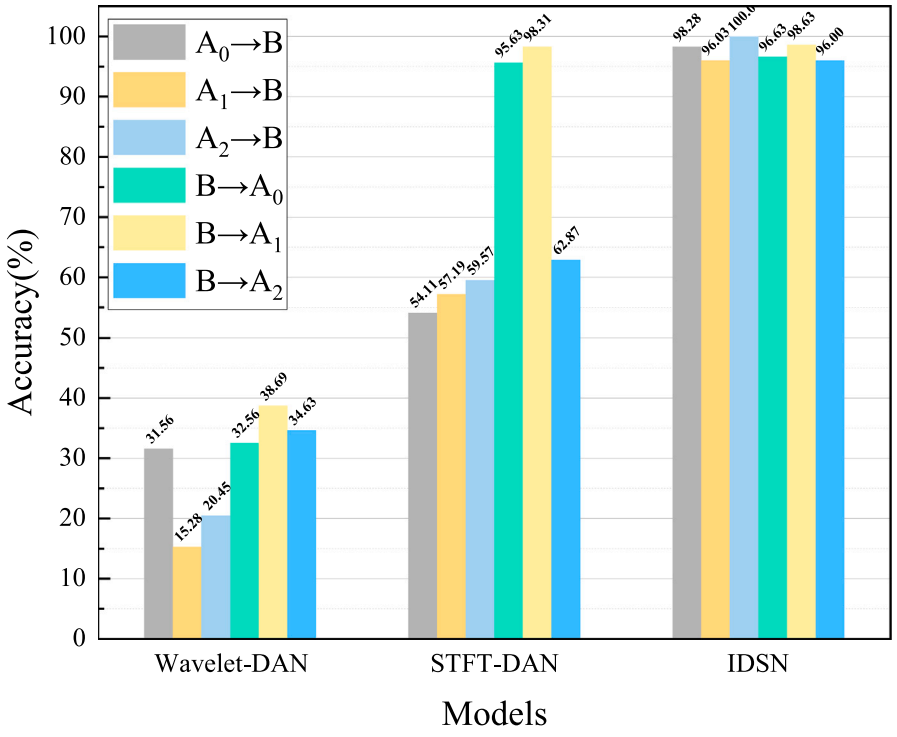


Fig. 11. One-stage vs. two-stage.

4.5.2. The plug-and-play differentiable STFT

Table 3 illustrates that the proposed differentiable STFT can be seamlessly merged into various backbone networks without conflict. To demonstrate this, several well-known models DenseNet-121, ResNet-18, GhostNet, MobileNetV3, EfficientNetV2, WDCNN, DRSN-CW and MIXCNN are conducted on $A_0 \rightarrow D_3$, w/o indicating that models without differentiable STFT layer. The results show that adding the differentiable STFT layer can significantly improve the cross-machine transfer performance of the model,

Table 3
Comparing different backbones.

Models	Accuracy	Accuracy(w/o)	Window Length	Improvement
GhostNet	99.95%	50.00%	108	↑49.95%
MobileNetV3	99.23%	97.50%	84	↑1.73%
EfficientNetV2	99.74%	22.81%	126	↑76.93%
DenseNet-121	96.43%	93.98%	132	↑2.45%
ResNet-18	98.37%	47.50%	222	↑80.87%
WDCNN-2D [58]	99.59%	96.63%	228	↑2.96%
DRSN-CW-2D [59]	98.78%	66.38%	222	↑32.40%
MIXCNN-2D [60]	99.11%	93.75%	78	↑5.36%

Table 4
Performance comparison with other methods on D_0 (%).

Task	$A_0 \rightarrow D_0$	$A_1 \rightarrow D_0$	$A_2 \rightarrow D_0$	$B \rightarrow D_0$	$C_0 \rightarrow D_0$	$C_1 \rightarrow D_0$	Average
BN-CNN [61]	25.00	25.00	47.16	39.17	28.17	49.89	35.81 ± 10.25
TCA [62]	17.50	19.50	29.00	19.00	09.50	26.00	20.08 ± 6.25
DAN [63]	99.89	**100.00**	99.78	47.78	74.28	61.67	80.57 ± 20.78
DDC [64]	99.67	**100.00**	99.83	48.00	99.78	99.28	91.09 ± 19.27
DSAN [52]	30.06	25.00	62.61	49.78	46.17	47.33	43.49 ± 12.58
BBDA [65]	25.00	62.83	49.44	32.33	35.73	25.00	38.30 ± 13.63
CMMD [66]	41.17	32.61	66.83	41.94	99.78	99.94	63.71 ± 27.61
DDNTL [47]	46.35	51.51	30.00	45.10	70.83	35.36	39.24 ± 13.64
CK-MMD [67]	42.13	42.94	27.20	65.97	26.56	30.61	46.53 ± 13.00
IDSN	99.83	**100.00**	**100.00**	**100.00**	**100.00**	99.89	**99.95 ± 0.07**

with a maximum improvement of more than 80% on certain challenging tasks. The above results demonstrate that the differentiable STFT layer can be used as a plug-and-play module to improve the cross-machine performance of the baseline.

4.5.3. The performance analysis of IDSN on rolling bearings

Compare the superior performance of IDSN with the following state-of-the-art methods: BN-CNN [61], TCA [62], DAN(MMD) [63], DDC(MK-MMD) [64], DSAN [52], BBDA [65], CMMD [66], DDNTL [47], CK-MMD [67]. Six cross-machine transfer tasks are constructed, achieving health monitoring from common bearings to damaged traction motor bearings with 165 km/h as the target domain. To ensure fairness, only the loss function of domain adaptation is altered, and other conditions are set the same.

Performance improvements over conventional domain alignment methods are experimentally shown in Table 4. The proposed IDSN has an average accuracy of 99.95% across the six tasks, with the accuracy of 99.83%, 100.00%, 100.00%, 100.00%, 100.00%, 100.00%, 99.89%, and 99.95%, respectively. In addition, the other counterparts are susceptible to the data sets, where it can reach rather high accuracy in some data sets as the source domain, but the transfer learning will fail if it changes to other data sets. Ref. [67] refers to cross-machine transfer from simulated data to experimental data. The Ref. [47] belongs to transfer between different machines. As indicated by the results, CK-MMD and DDNTL are more sensitive to data sets and have different transfer accuracies for different data sets, however, the average diagnostic accuracy is lower, primarily because the methods rely solely on statistical metric regularization, while the proposed IDSN reduces the domain shift between two domains by means of differentiable STFT, and achieves a high transfer effect without complex statistical metrics.

A purely data-driven approach demands a significant amount of data to train the model and is reliant on the quality of available data. Additionally, some of the underlying approaches exhibit great differences on different tasks. For some transfer tasks, the source- and target-domain is noisy or contains little fault information, so some baselines are less accurate. In contrast to these methods, IDSN takes a differentiable short-time Fourier transform from a signal processing perspective to process the data in the source- and target-domain, diminishing the domain shift and thus making the tasks easier and thus improving the accuracy. Differentiable signal processing is a different idea from the previous methods.

Overall, the knowledge and data dual-driven approach can decrease the dependence on the quality and quantity of source and target domain samples. Markedly, IDSN can identify more than 99.80% of the tested samples, which only entails 80 source domain samples and 80 target domain samples.

As shown in Fig. 12, the performance of deep JAN (DJAN, JMMD) [46] and IDSN are tested under conditions, where the number of samples available for training in the source and target domains is 20K and for testing in target domain is 480K. The proposed IDSN is also higher performance and more stable. Apparently, DJAN, purely data-driven model, is overly susceptible to the quantity and quality of data sets, and the results reported by different data sets vary greatly. The performance of IDSN is more reliable, performance of which is mainly affected by α. The best result can be recorded by one of the choices from $[1.0, 1.2, 2.0]$, with acceptable parameter adjustment costs.

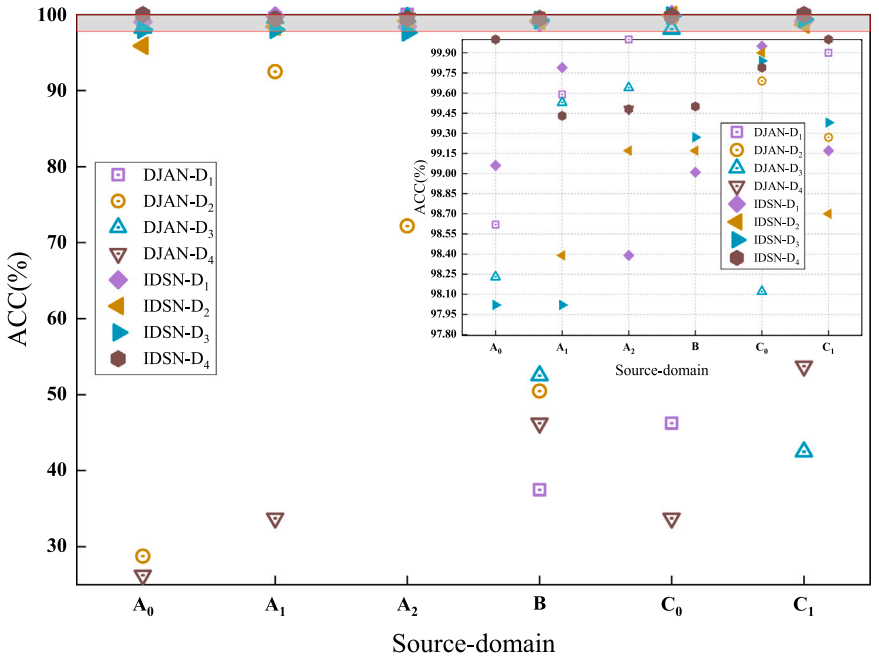


Fig. 12. Performance with DJAN/IDSN on more extreme conditions.

Table 5
Performance of different window functions.

		Hamming	Hann	Gaussian	Blackman	Bartleet	Kaiser
$C_0 \rightarrow D_3$	accuracy (%)	99.83	99.78	99.56	**99.94**	99.67	99.83
	Time (s)	38.33	22.71	25.47	**15.28**	33.08	27.69
$C_1 \rightarrow D_3$	Accuracy (%)	99.67	**100.00**	99.11	**100.00**	**100.00**	99.72
	Time (s)	17.03	17.58	18.36	**15.43**	34.29	30.16

The study of STFT has emerged multiple elegant window functions, the expressions of which can be referred to PyTorch[1] and Scipy.[2] The accuracy of the test set and training time are obtained on C_0 and C_1 as the source domain and D_3 as the target domain, as shown in Table 5. It can be observed that the elected blackman window function shines in terms of accuracy and training time. However, it does not mean that blackman window function is omnipotent. Different transfer tasks can affect the type of window function. The accuracy of Gaussian window function is the lowest, so the results reported in this paper are only the minimum of the proposed IDSN.

4.5.4. *The performance analysis of IDSN on planetary gearbox*

The abovesaid studies are based on bearings, as one of the mechanical components. In addition to these, the gearbox fault diagnosis for different speeds is explored using C_2, C_3 since the actual high-speed train gearbox fault data has not been gleaned. From Table 6, it can be seen that the proposed IDSN achieves the highest accuracy of 99.89% and 99.61% in two tasks. It is demonstrated that IDSN is practical for cross-machine bearing transfer tasks and also applicable to cross-running speed gearbox transfer tasks.

4.5.5. *Variations in window length*

As depicted in Fig. 13, the source domain data set A_0 and the target domain D are utilized to demonstrate the variations process of window length. Specifically, the window length is adaptively adjusted by the backpropagation algorithm via continuous tuning of the variable θ. This dynamic optimization process aims to determine an optimal window length in the context of domain adaptation.

[1] https://pytorch.org/docs/stable/torch.html#spectral-ops.
[2] https://docs.scipy.org/doc/scipy/reference/signal.windows.html?highlight=windows#module-scipy.signal.windows.

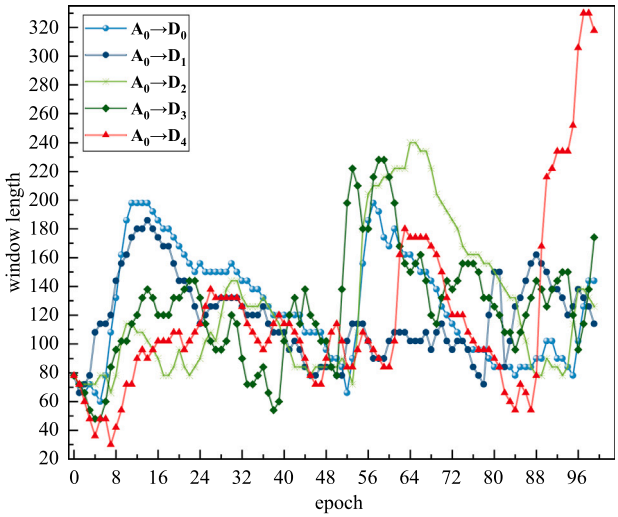


Fig. 13. Variations in window length (%).

Table 6
Experimental performance based on gearboxes at different speeds.

Task	$C_2 \rightarrow C_3$	$C_3 \rightarrow C_2$	Average
BN-CNN	70.72	69.83	70.28 ± 0.45
TCA	29.50	27.50	28.50 ± 1.00
DAN	99.22	98.72	98.97 ± 0.25
DDC	99.61	65.44	97.53 ± 17.09
DSAN	70.39	97.89	84.14 ± 13.75
BBDA	99.50	96.56	98.03 ± 1.47
CMMD	99.17	97.72	98.45 ± 0.73
IDSN	**99.89**	**99.61**	**99.75 ± 0.14**

Table 7
Experimental results (%) are based on $C_0, C_1 \rightarrow D$.

		D_0	D_1	D_2	D_3	D_4	Average
C_0	JMMD	99.49	47.89	99.78	74.33	99.78	84.24 ± 20.66
	S-JMMD	100.00	99.61	100.00	99.78	99.28	**99.73 ± 0.27**
	NS-JMMD	99.44	98.89	99.67	99.50	99.89	99.48 ± 0.33
C_1	JMMD	100.00	99.83	98.28	99.94	99.78	99.57 ± 0.64
	S-JMMD	99.89	99.72	99.61	99.67	99.94	**99.77 ± 0.13**
	NS-JMMD	100.00	84.50	100.00	97.28	99.61	96.28 ± 5.98

4.5.6. The ablation experiment of IDSN

The DDB bearing data set is used as the source domain, and NTN is the target domain. NS-JMMD indicates that DAN is without the adaptive window length and the S-JMMD loss function. The results are shown in Table 7. S-JMMD has the best stability, wherein the tasks of $C_0 \rightarrow D_1$, $C_0 \rightarrow D_3$, and the cross-machine diagnosis task of JMMD without label smoothing will fail. With a fixed window length, the accuracy of NS-JMMD is only 84.50% in the $C_1 \rightarrow D_1$. On the whole, faced with different source and target domains, IDSN with label smoothing and differentiable STFT can relief the sensitivity to different data sets, avoid negative transfer, and maintain maximum accuracy. In addition, the average accuracy of 99.73% and 99.77% are attained in the cross-machine tasks for $C_0 \rightarrow D$ and $C_1 \rightarrow D$, respectively.

4.5.7. The parameter analysis of IDSN

The effects of the important parameters are studied for $C_0 \rightarrow D_3$ as an illustration.

As shown in Fig. 14, the smoothing factor determines the confidence level for the source domain sample labels. As the smoothing factor changes, the accuracy of the model fluctuates. Combining Table 7 and Fig. 14, this work can clearly see the necessity of adopting the smoothed labels to align pseudo-labels. However, a larger smoothing factor may make the network fail to converge, so the results of only $\varepsilon = 0.1 \sim 0.55$ is reported. By setting different ε, the inter-class distance is enlarged while the intra-class dispersion is minimized.

It takes somewhat level of expertise to determine valid initial values for the window length. In accordance with Ref. [39] and the sample length of 1024, some initial values are set as shown in Table 8. The IDSN hits 99.89% when the initial values of window

Table 8
Performance of different window length.

Before	16	32	64	128	256	512	1024
Accuracy	99.83%	**99.89%**	97.44%	99.11%	99.33%	99.17%	**99.89%**
After	42	114	84	96	96	168	618

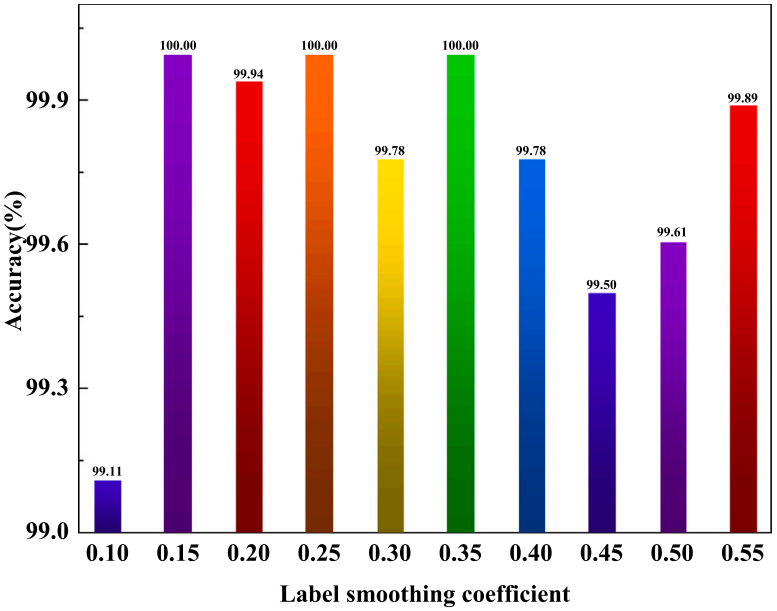


Fig. 14. The influence of the smoothing coefficient.

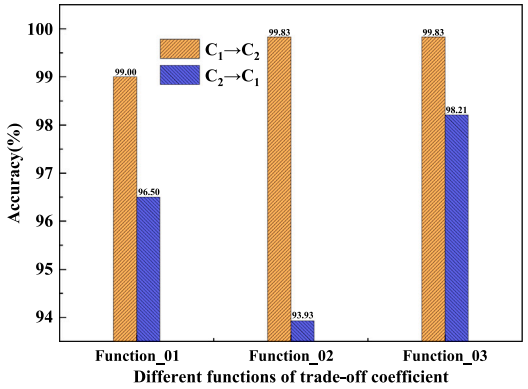


Fig. 15. The performance of different functions of trade-off coefficient.

length are 32 and 1024. Furthermore, the final Gaussian window lengths are tuned from 32 to 114 or from 1024 to 618. Naturally, regarding the results of the state-of-the-art window length methods [68] as the initial value may deliver even superior performance.

As shown in Fig. 15, the transfer performance between different sampling frequencies of the same mechanical device is contrasted. It can be noticed from Table 1 that there are differences in sampling frequencies between different mechanical devices, so this work maintain that it can be viewed as another form of the cross-machine task. The proposed **Function_03** achieves optimal performance in both tasks. In particular, it reaches 98.21% in the $C_1 \rightarrow C_0$, which improves by 1.71% and 4.28% compared to its competitors, respectively. For **Function_01**, the reason why the optimal alignment is not achieved is because less weight is endowed to S-JMMD loss in the later stage. Also, for **Function_02**, the model cannot adequately learn the source domain knowledge due to the faster increase of weights λ in the early stage, which affects the domain adaptation stage and produces a severe negative transfer phenomenon.

To further illustrate the utility of label smoothing, t-SNE [69] of the learned features is displayed in Fig. 16. Features without label smoothing are nearer between distinct classes in the same domain, which thus weakens the generalization capacity of the model

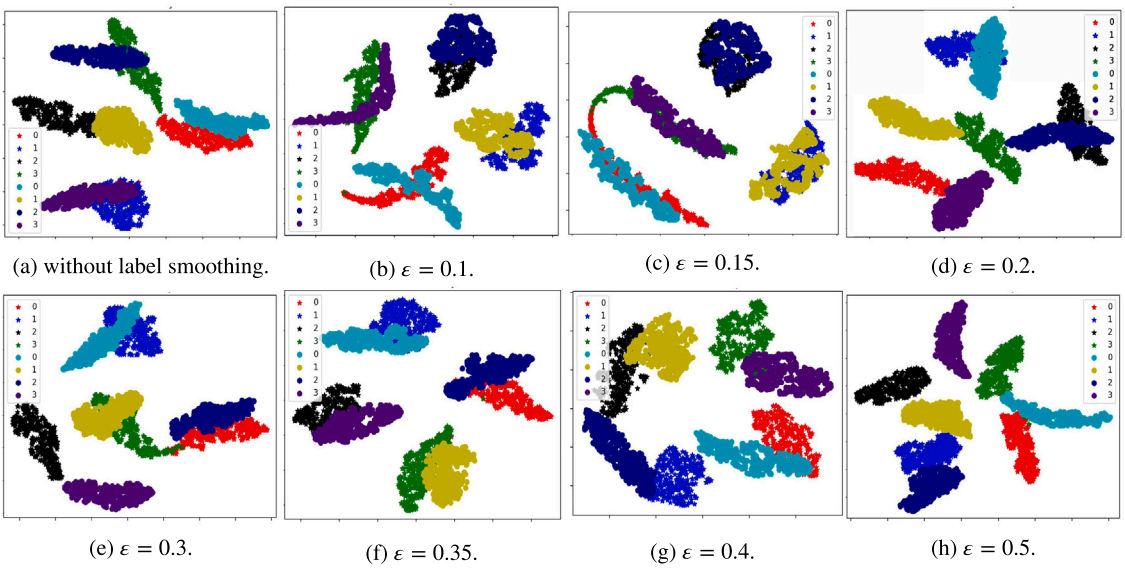


Fig. 16. The t-SNE mappings of learned features obtained by different ε.

Table 9
The impact of window function learning rate on the proposed IDSN.

α	0.001	0.1	0.9	1.0	1.2	2.0
Accuracy (%)	56.69	83.44	94.81	98.70	98.80	99.33
Window length	126	84	108	54	96	90

(Fig. 16(a)). In contrast, a appropriate smoothing coefficient (Figs. 16(b), 16(g)) widens the distance between different classes in the same domain. Furthermore, it can be observed that unreasonable smoothing coefficients can confuse the alignment (Fig. 16(f)). A case in point is $\varepsilon = 0.35$. Although separating different types in the same domain is straightforward, IR in the source domain is closer to NC of the target domain. Nevertheless, from another viewpoint, although the mappings are poorly interpretable, it does not seriously undermine the recognition performance of the target domain, which is an intriguing finding, and this work speculate that this may be the property of T-SNE, independent of the extracted discriminative features. In short, label smoothing is compatible with domain adaptation but entails a reasonable smoothing factor.

The learning rate of DAN is 0.001. As shown in Table 9, it can be found that the best accuracy is not produced when the learning rate of differentiable short-time Fourier transform is comparable to that of DAN. This is due to the limited search range of the derivable coefficient, which restricts the capability to search for an appropriate derivable coefficient and thus hindering the calculation of a reasonable window size. However, when a larger learning rate is used, the search range can be extended, thereby facilitating the identification of an appropriate window function. The difference of learning rate further highlights the dissimilarities between differentiable STFT and DAN. The reason why a larger learning rate can be adopted is that the differentiable STFT layer follows the computational rules of STFT and has physical interpretability, it is capable of adopting larger learning rates differing from those of DAN, such as $1.0, 1.2, 2.0$, and so forth.

5. Further experimental study

In addition, the cross-speed transfer diagnosis tasks for the bearings of the traction motor systems are further analyzed, as shown in Fig. 17. Among these, the minimum accuracy surpasses 98.50%, demonstrating the utility of IDSN for cross-running speed challenges.

To further assure the reproducibility of the experimental results, the performance of the cross-machine transfer between two shared data sets is specifically documented. As shown in Table 10, the cross-machine transfer tasks of data sets between Jiangnan University (Dataset A) and Case Western Reserve University (Dataset B) achieve over 96%, which outperforms other traditional models.

The algorithm from Refs. [70,71] is implemented in the same way. As shown in Table 11, for multiple tasks based on public data sets, IDSN outperforms Refs. [70,71] in generalization for multi-task transfer recognition, achieving 97.60% average accuracy and improving by 5.39%, with shorter training time.

Different implementations of the differentiable STFT also exist [70,71]. Dissimilarly, ISDN introduces the derivative coefficient θ to guide the window length, window function, support set size, hop size, and trade-off between time-frequency resolution, where

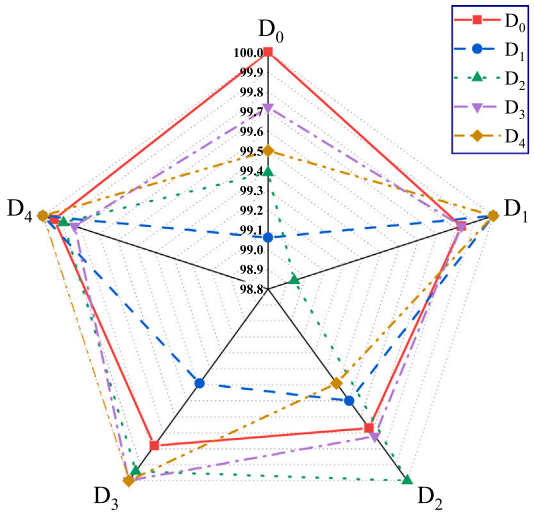


Fig. 17. The cross-speed transfer diagnosis tasks.

Table 10
The cross-machine results based on published data sets (%).

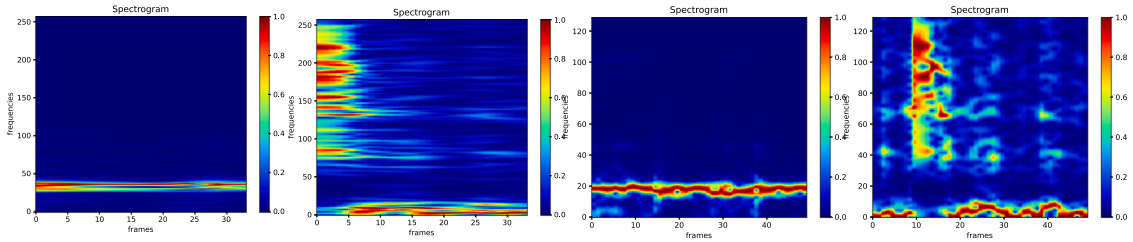
Task	$A_1 \rightarrow B$	$A_2 \rightarrow B$	$A_3 \rightarrow B$	$B \rightarrow A_1$	$B \rightarrow A_2$	$B \rightarrow A_3$
DAN	68.75	57.91	60.18	70.62	75.78	64.43
DDC	41.98	75.38	72.54	89.17	74.32	60.78
DSAN	49.57	74.47	24.92	49.64	49.11	66.35
BBDA	43.75	49.87	37.50	57.40	81.82	73.85
CMMD	43.04	51.93	25.00	60.26	83.28	60.83
IDSN	98.28	96.03	100.00	96.63	98.63	96.00

Table 11
The time and accuracy of ISDN and Refs. [70,71].

Task		$A_1 \rightarrow B$	$A_2 \rightarrow B$	$A_3 \rightarrow B$	$B \rightarrow A_1$	$B \rightarrow A_2$	$B \rightarrow A_3$	Average
Refs. [70,71]	Acc (%)	90.61	85.47	97.94	88.21	98.17	92.86	92.21 ± 4.70
	Time (s)	72.21	69.22	75.57	203.00	193.35	191.74	
ISDN	Acc (%)	98.28	96.03	100.00	96.63	98.63	96.00	97.60 ± 1.49
	Time (s)	16.01	16.67	28.09	18.53	23.21	33.79	

relationships between θ and the aforementioned variables is established. Meanwhile, Refs. [70,71] predefine window length and sliding steps, thus fixing sampled window numbers and adjusting the window function by window length. Since the number of sampling windows cannot be adaptively adjusted, it lacks the flexibility. However, the proposed differentiable STFT in this paper can optimize both sample numbers and window functions. To validate the proposed theory from IDSN, the correlation between θ and the time-frequency resolution is derived. Moreover, θ is guided by the transfer metric loss, which is more suitable for cross-machine fault diagnosis tasks. In addition to this, the number of parameters in Refs. [70,71] are related to the size of the support set, and the number of parameters in ISDN proposed is only one. The smaller number of parameters makes it easier to optimize. Assuming that an initial window length is 64, the learning rate of derivable coefficient is 2.0, enabling a large search range. In contrast, Refs. [70,71] directly tune 965 window parameters, hindering training and increasing time. The learning rate of 2.0 causes vanished gradients, while 0.001 only permits small tuning ranges.

In all six tasks, IDSN consistently achieves an accuracy rate exceeding 96%. On the contrary, Refs. [70,71] showcase an accuracy rate lower than 90% in the $A_2 \rightarrow B$ and $B \rightarrow A_1$ tasks. These results suggest that Refs. [70,71] exhibit a lack of generalization capability in specific tasks, thereby demonstrating suboptimal performance and limited adaptability across different cross-machine tasks. Moreover, it is noteworthy that IDSN possesses a single parameter, while Refs. [70,71] require 965 parameters. The inclusion of a significantly larger number of parameters in Refs. [70,71] introduce greater complexity in optimization, leading to a prolonged training time compared to IDSN. Consequently, the proposed technique advantages include higher transfer learning accuracy, fewer parameters, shorter training time, wider optimization window length ranges, and the flexible STFT form.



(a) source domain. (b) target domain. (c) source domain(w/o). (d) target domain(w/o).

Fig. 18. The spectrograms on source- or target-domain with/without DSTFT.

6. The interpretability analysis of DSTFT

The traditional data-driven deep learning model is black box. However, ISDN is a knowledge and data dual-driven model, which draws on the knowledge-driven of differentiable STFT and the data-driven of neural networks. In particular, the differentiable STFT is embedded into ISDN as priori knowledge, enabling ISDN with intrinsic interpretability.

Intrinsic interpretability based on physical prior knowledge and post-hoc explanations based on activation mapping, are two forms of interpretable fault diagnosis [72]. Intrinsic physical interpretability is the structure design that follows some specific physical knowledge, such as the differentiable STFT proposed in this paper. The post-hoc interpretability lies in the networks which learn the fault frequency bands of signals, such as convolutional kernel visualization and Grad-CAM. Differentiable STFT belongs to the former, and it follows the solid theoretical foundation of STFT. Based on the theory of STFT, the trade-off between time and frequency resolution is crucial, so the derivable coefficient θ is introduced to adjust the time-frequency resolution adaptively, and the relationship between θ and the time-frequency resolution is analyzed in detail.

IDSN enables the coordination of the differentiable STFT layer and DAN. The paradigm design strictly follows the theory of SFTF theory, so it is intrinsic interpretability. On the other hand, the differentiable STFT paradigm has only one parameter, whose learning rate is typically 1.0 about 1000 times that of DAN, indicating that it is different from the black box domain adaptation model. In summary, this scheme does not destroy the theoretical interpretability of STFT itself, producing reliable results. There are structures with interpretability, such as SincNet and WavekernelNet, which combine more with the nature of convolution, replacing the convolution kernel with some wavelet function for extracting interpretable features, not the same as the proposed interpretable differentiable STFT.

In terms of post-hoc interpretability, as shown in Fig. 18, we draw the spectrogram obtained with (Figs. 18(a), 18(b)) and without (Figs. 18(c), 18(d)) differentiable STFT, respectively. The spectrograms represent the outer ring faults of CWRU and JNU data sets, respectively. The proposed differentiable STFT, with the aid of the derivable coefficient, concentrates the energy of the spectrogram and reduce noises, and the difference between the two domains reduces from 0.96 to 0.04, according to the $\mathscr{A}$-distance: $\mathscr{A}\left(\mathscr{D}_s, \mathscr{D}_t\right) = 2\left(1 - 2err\left(h\right)\right)$, which implies an increase in the transferability between the two domains.

7. Conclusion

Due to the scarcity of fault labels in traction motor bearing data of high-speed trains, conventional supervised learning-based and same-machine transfer learning-based schemes will fail. However, drawing on the idea of cross-machine transfer, it is an alternative and optimal solution to transfer public, self-collected data to traction motor bearing data without annotation. To address this challenge, this work proposes a dual-driven model integrating differentiable interpretable short-time Fourier transform "knowledge" and a smaller amount of "data". This is the first record of differentiable signal processing in domain adaptation fault diagnosis. What we propose and five-fold contributions are as follows:

(1) There are two existing main ideas to enable effective cross-machine transfer fault diagnosis. One is trying to find the better method to establish the domain discrepancy statistical metric, and the other is devoted to design the elaborate network structure. Different from the aforesaid two ways, a novel alternative considering the prior knowledge of signal processing is proposed herein, that is, one-stage domain adaptation networks embedded in differentiable signal processing.

(2) A one-stage cross-machine transfer diagnosis method, termed ISDN, is proposed to integrate STFT into DAN for the first time, in which the hyper-parameters of STFT are part of the parameters of ISDN. They are dynamically updated using back-propagation algorithm, which tackles the challenge of time-consuming and laborious optimization process owing to the non-differentiable hyper-parameters of STFT, and ISDN outperforms the prior two-stage method (signal pre-processing + DAN) in diminishing the distribution shift of source- and target-domain, making a simpler domain alignment process.

(3) A new paradigm for implementing differentiable STFT is put forward, which a variable is introduced as the derivative coefficient. By deducting the relationships between the derivative coefficient and pivotal parameters such as window length, window function, and hop length, these parameters can adaptively update using just the derivative coefficient that is treated as one hyperparameter of ISDN. Eventually, a trade-off between time-frequency resolution is realized.

(4) Smoothed joint maximum mean discrepancy (S-JMMD) is proposed. For conditional distribution alignment, the smoothed labels of source domain are utilized to align the pseudo-labels of target domain, improving the generalizability of the source model while further enhancing domain confusion. Additionally, a new adaptive weighting factor is also developed to address the weight matching of S-JMMD.
(5) ISDN is a dual-driven model, which draws on the knowledge-driven of differentiable STFT and the data-driven of neural networks. In particular, the differentiable STFT is embedded into ISDN as priori knowledge, enabling ISDN with intrinsic interpretability.

Several nontrivial issues that naturally arise demand deeper examinations:

(1) Although a linear relationship between the window function and the derivative coefficient has been established in this work, further research into more intricate relationships, such as exponential and trigonometric ones, may produce even more promising outcomes.
(2) While the current research focuses on the fixed window length, investigating the utilization of variable window lengths may extend its application to more complex bearing transfer fault diagnosis scenarios on speed fluctuation, which thus move beyond purely data-driven domain adaptation methodology.
(3) Continuous wavelet transform and Synchrosqueezing transform, as well-known signal processing algorithms, are worth studying to make them become the differentiable form.
(4) The current research is limited to closed set, and the cross-machine study regarding partial set and open set is still rare and difficult in fault diagnosis, which is also a worthy breakthrough point. The knowledge and data dual-driven model may provide an optimal solution.
(5) Differentiable signal processing requires a guide, such as setting an approximate value for the window size.

CRediT authorship contribution statement

Chao He: Writing – original draft, Software, Validation, Visualization, Methodology, Investigation, Proof reading. **Hongmei Shi:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Jianbo Li:** Investigation, Writing – review & editing, Data curation, Resources.

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.

Data availability

The data that has been used is confidential.

Acknowledgments

The authors are grateful for the supports of the National Natural Science Foundation of China (No. 52272429).

References

[1] Y. Sun, J. Wang, X. Wang, Fault diagnosis of mechanical equipment in high energy consumption industries in China: A review, Mech. Syst. Signal Process. 186 (2023) 109833, http://dx.doi.org/10.1016/j.ymssp.2022.109833.
[2] S. Xu, C. Chen, Z. Lin, X. Zhang, J. Dai, L. Liu, Review and prospect of maintenance technology for traction system of high-speed train, Transp. Saf. Secur. 3 (3) (2021) tdab017, http://dx.doi.org/10.1093/tse/tdab017.
[3] Z. Wang, G. Mei, Q. Xiong, Z. Yin, W. Zhang, Motor car–track spatial coupled dynamics model of a high-speed train with traction transmission systems, Mech. Mach. Theory 137 (2019) 386–403, http://dx.doi.org/10.1016/j.mechmachtheory.2019.03.032.
[4] Y. Zou, Y. Zhang, H. Mao, Fault diagnosis on the bearing of traction motor in high-speed trains based on deep learning, Alex. Eng. J. 60 (1) (2021) 1209–1219, http://dx.doi.org/10.1016/j.aej.2020.10.044.
[5] C. Cheng, W. Wang, H. Chen, B. Zhang, J. Shao, W. Teng, Enhanced fault diagnosis using broad learning for traction systems in high-speed trains, IEEE Trans. Power Electron. 36 (7) (2021) 7461–7469, http://dx.doi.org/10.1109/TPEL.2020.3043741.
[6] H. Wang, Z. Liu, D. Peng, M.J. Zuo, Interpretable convolutional neural network with multilayer wavelet for Noise-Robust Machinery fault diagnosis, Mech. Syst. Signal Process. 195 (2023) 110314, http://dx.doi.org/10.1016/j.ymssp.2023.110314.
[7] Z. Wang, J. Yang, Y. Guo, Unknown fault feature extraction of rolling bearings under variable speed conditions based on statistical complexity measures, Mech. Syst. Signal Process. 172 (2022) 108964, http://dx.doi.org/10.1016/j.ymssp.2022.108964.
[8] Z. Mao, M. Xia, B. Jiang, D. Xu, P. Shi, Incipient fault diagnosis for high-speed train traction systems via stacked generalization, IEEE Trans. Cybern. 52 (8) (2022) 7624–7633, http://dx.doi.org/10.1109/TCYB.2020.3034929.
[9] Y. Li, Exploring real-time fault detection of high-speed train traction motor based on machine learning and wavelet analysis, Neural Comput. Appl. 34 (12) (2022) 9301–9314, http://dx.doi.org/10.1007/s00521-021-06284-0.
[10] H. Dong, H. Ma, Z. Wang, J. Man, L. Jia, Y. Qin, An online health monitoring framework for traction motors in high-speed trains using temperature signals, IEEE Trans. Ind. Inform. 19 (2) (2023) 1389–1400, http://dx.doi.org/10.1109/TII.2022.3200357.
[11] J. Man, H. Dong, X. Yang, Z. Meng, L. Jia, Y. Qin, G. Xin, GCG: Graph Convolutional network and gated recurrent unit method for high-speed train axle temperature forecasting, Mech. Syst. Signal Process. 163 (2022) 108102, http://dx.doi.org/10.1016/j.ymssp.2021.108102.

[12] Q. Hao, Y. Shen, Y. Wang, J. Liu, An adaptive extraction method for rail crack acoustic emission signal under strong wheel-rail rolling noise of high-speed railway, Mech. Syst. Signal Process. 154 (2021) 107546, http://dx.doi.org/10.1016/j.ymssp.2020.107546.
[13] G. Yan, J. Chen, Y. Bai, C. Yu, C. Yu, A survey on fault diagnosis approaches for rolling bearings of railway vehicles, Processes 10 (4) (2022) 724, http://dx.doi.org/10.3390/pr10040724.
[14] C. Han, W. Lu, H. Wang, L. Song, L. Cui, Multistate fault diagnosis strategy for bearings based on an improved convolutional sparse coding with priori periodic filter group, Mech. Syst. Signal Process. 188 (2023) 109995, http://dx.doi.org/10.1016/j.ymssp.2022.109995.
[15] S. Liu, J. Chen, S. He, Z. Shi, Z. Zhou, Few-shot learning under domain shift: Attentional contrastive calibrated transformer of time series for fault diagnosis under sharp speed variation, Mech. Syst. Signal Process. 189 (2023) 110071, http://dx.doi.org/10.1016/j.ymssp.2022.110071.
[16] J. Wang, J. Yang, Y. Wang, Y. Bai, T. Zhang, D. Yao, Ensemble decision approach with dislocated time–frequency representation and pre-trained CNN for fault diagnosis of railway vehicle gearboxes under variable conditions, Int. J. Rail Transp. 10 (5) (2022) 655–673, http://dx.doi.org/10.1080/23248378.2021.2000897.
[17] Z. Chen, W. Chen, X. Fan, T. Peng, C. Yang, JITL-MBN: A real-time causality representation learning for sensor fault diagnosis of traction drive system in high-speed trains, IEEE Trans. Neural Netw. Learn. Syst. (2022) 1–12, http://dx.doi.org/10.1109/TNNLS.2022.3173337.
[18] J. Xu, H. Ke, Z. Chen, X. Fan, T. Peng, C. Yang, Over-smoothing relief graph convolutional network-based fault diagnosis method with application to the rectifier of high-speed trains, IEEE Trans. Ind. Inform. 19 (1) (2023) 771–779, http://dx.doi.org/10.1109/TII.2022.3167522.
[19] H. Dong, F. Chen, Z. Wang, L. Jia, Y. Qin, J. Man, An adaptive multisensor fault diagnosis method for high-speed train traction converters, IEEE Trans. Power Electron. 36 (6) (2021) 6288–6302, http://dx.doi.org/10.1109/TPEL.2020.3034190.
[20] H. Wang, T. Men, Y.-F. Li, Transformer for high-speed train wheel wear prediction with multiplex local–global temporal fusion, IEEE Trans. Instrum. Meas. 71 (2022) 1–12, http://dx.doi.org/10.1109/TIM.2022.3154827.
[21] C. Cheng, W. Wang, G. Ran, H. Chen, Data-driven designs of fault identification via collaborative deep learning for traction systems in high-speed trains, IEEE Trans. Transp. Electrif. 8 (2) (2022) 1748–1757, http://dx.doi.org/10.1109/TTE.2021.3129824.
[22] C. Cheng, X. Li, P. Xie, X. Yang, Transfer learning-aided fault detection for traction drive systems of high-speed trains, IEEE Trans. Artif. Intell. (2022) 1, http://dx.doi.org/10.1109/TAI.2022.3177387.
[23] F. Lu, Q. Tong, Z. Feng, Q. Wan, Unbalanced bearing fault diagnosis under various speeds based on spectrum alignment and deep transfer convolution neural network, IEEE Trans. Ind. Inform. (2022) 1–12, http://dx.doi.org/10.1109/TII.2022.3217541.
[24] 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. Signal Process. 167 (2022) 108487, http://dx.doi.org/10.1016/j.ymssp.2021.108487.
[25] Y. Bai, J. Yang, J. Wang, Y. Zhao, Q. Li, Image representation of vibration signals and its application in intelligent compound fault diagnosis in railway vehicle wheelset-axlebox assemblies, Mech. Syst. Signal Process. 152 (2021) 107421, http://dx.doi.org/10.1016/j.ymssp.2020.107421.
[26] N. Qin, B. Wu, D. Huang, Y. Zhang, Stepwise adaptive convolutional network for fault diagnosis of high-speed train bogie under variant running speeds, IEEE Trans. Ind. Inform. 18 (12) (2022) 8389–8398, http://dx.doi.org/10.1109/TII.2022.3152540.
[27] S.-X. Chen, L. Zhou, Y.-Q. Ni, Wheel condition assessment of high-speed trains under various operational conditions using semi-supervised adversarial domain adaptation, Mech. Syst. Signal Process. 170 (2022) 108853, http://dx.doi.org/10.1016/j.ymssp.2022.108853.
[28] H. Ren, J. Wang, J. Dai, Z. Zhu, J. Liu, Dynamic balanced domain-adversarial networks for cross-domain fault diagnosis of train bearings, IEEE Trans. Instrum. Meas. 71 (2022) 1–12, http://dx.doi.org/10.1109/TIM.2022.3179468.
[29] W. Zhang, X. Li, Federated transfer learning for intelligent fault diagnostics using deep adversarial networks with data privacy, IEEE/ASME Trans. Mechatron. 27 (1) (2022) 430–439, http://dx.doi.org/10.1109/TMECH.2021.3065522.
[30] X. Yu, Z. Zhao, X. Zhang, C. Sun, B. Gong, R. Yan, X. Chen, Conditional adversarial domain adaptation with discrimination embedding for locomotive fault diagnosis, IEEE Trans. Instrum. Meas. 70 (2021) 1–12, http://dx.doi.org/10.1109/TIM.2020.3031198.
[31] R. Hu, M. Zhang, X. Meng, Z. Kang, Deep subdomain generalisation network for health monitoring of high-speed train brake pads, Eng. Appl. Artif. Intell. 113 (2022) 104896, http://dx.doi.org/10.1016/j.engappai.2022.104896.
[32] V. Monga, Y. Li, Y.C. Eldar, Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing, IEEE Signal Process. Mag. 38 (2) (2021) 18–44, http://dx.doi.org/10.1109/MSP.2020.3016905.
[33] J. Wang, Y. Li, R.X. Gao, F. Zhang, Hybrid physics-based and data-driven models for smart manufacturing: Modelling, simulation, and explainability, J. Manuf. Syst. 63 (2022) 381–391, http://dx.doi.org/10.1016/j.jmsy.2022.04.004.
[34] T. Yin, N. Lu, G. Guo, Y. Lei, S. Wang, X. Guan, Knowledge and data dual-driven transfer network for industrial robot fault diagnosis, Mech. Syst. Signal Process. 182 (2023) 109597, http://dx.doi.org/10.1016/j.ymssp.2022.109597.
[35] T. Zhang, J. Chen, S. He, Z. Zhou, Prior knowledge-augmented self-supervised feature learning for few-shot intelligent fault diagnosis of machines, IEEE Trans. Ind. Electron. 69 (10) (2022) 10573–10584, http://dx.doi.org/10.1109/TIE.2022.3140403.
[36] J. Yu, S. Wang, L. Wang, Y. Sun, Gearbox fault diagnosis based on a fusion model of virtual physical model and data-driven method, Mech. Syst. Signal Process. 188 (2023) 109980, http://dx.doi.org/10.1016/j.ymssp.2022.109980.
[37] A. Jablonski, K. Dziedziech, Intelligent spectrogram–A tool for analysis of complex non-stationary signals, Mech. Syst. Signal Process. 167 (2022) 108554, http://dx.doi.org/10.1016/j.ymssp.2021.108554.
[38] J. Si, H. Shi, J. Chen, C. Zheng, Unsupervised deep transfer learning with moment matching: A new intelligent fault diagnosis approach for bearings, Measurement 172 (2021) 108827, http://dx.doi.org/10.1016/j.measurement.2020.108827.
[39] H. Tao, P. Wang, Y. Chen, V. Stojanovic, H. Yang, An unsupervised fault diagnosis method for rolling bearing using STFT and generative neural networks, J. Franklin Inst. 357 (11) (2020) 7286–7307, http://dx.doi.org/10.1016/j.jfranklin.2020.04.024.
[40] B. Yang, Y. Lei, X. Li, C. Roberts, Deep targeted transfer learning along designable adaptation trajectory for fault diagnosis across different machines, IEEE Trans. Ind. Electron. 70 (9) (2023) 9463–9473, http://dx.doi.org/10.1109/TIE.2022.3212415.
[41] R. Zhu, W. Peng, D. Wang, C.-G. Huang, Bayesian transfer learning with active querying for intelligent cross-machine fault prognosis under limited data, Mech. Syst. Signal Process. 183 (2023) 109628, http://dx.doi.org/10.1016/j.ymssp.2022.109628.
[42] Y. Shi, A. Deng, X. Ding, S. Zhang, S. Xu, J. Li, Multisource domain factorization network for cross-domain fault diagnosis of rotating machinery: An unsupervised multisource domain adaptation method, Mech. Syst. Signal Process. 164 (2022) 108219, http://dx.doi.org/10.1016/j.ymssp.2021.108219.
[43] C. He, H. Shi, J. Si, J. Li, Physics-informed interpretable wavelet weight initialization and balanced dynamic adaptive threshold for intelligent fault diagnosis of rolling bearings, J. Manuf. Syst. 70 (2023) 579–592, http://dx.doi.org/10.1016/j.jmsy.2023.08.014.
[44] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, T. Darrell, Deep domain confusion: Maximizing for domain invariance, 2014, http://dx.doi.org/10.48550/arXiv.1412.3474.
[45] Y. Zhu, F. Zhuang, J. Wang, G. Ke, J. Chen, J. Bian, H. Xiong, Q. He, Deep subdomain adaptation network for image classification, IEEE Trans. Neural Netw. Learn. Syst. 32 (4) (2021) 1713–1722, http://dx.doi.org/10.1109/TNNLS.2020.2988928.
[46] M. Long, H. Zhu, J. Wang, M.I. Jordan, Deep transfer learning with joint adaptation networks, in: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, Vol. 70, 2017, pp. 2208–2217, http://proceedings.mlr.press/v70/long17a.html.
[47] Q. Qian, Y. Qin, J. Luo, Y. Wang, F. Wu, Deep discriminative transfer learning network for cross-machine fault diagnosis, Mech. Syst. Signal Process. 186 (2023) 109884, http://dx.doi.org/10.1016/j.ymssp.2022.109884.

[48] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, 2016, pp. 2818–2826, http://dx.doi.org/10.1109/CVPR.2016.308.
[49] R. Müller, S. Kornblith, G.E. Hinton, When does label smoothing help? in: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, 2019, pp. 4696–4705, https://proceedings.neurips.cc/paper/2019/hash/f1748d6b0fd9d439f71450117eba2725-Abstract.html.
[50] Z. Shen, Z. Liu, D. Xu, Z. Chen, K. Cheng, M. Savvides, Is label smoothing truly incompatible with knowledge distillation: An empirical study, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, 2021, https://openreview.net/forum?id=POBuuGVrGaZ.
[51] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, P. Yu, Generalizing to unseen domains: A survey on domain generalization, IEEE Trans. Knowl. Data Eng. (2022) 1, http://dx.doi.org/10.1109/TKDE.2022.3178128.
[52] J. Luo, H. Shao, H. Cao, X. Chen, B. Cai, B. Liu, Modified DSAN for unsupervised cross-domain fault diagnosis of bearing under speed fluctuation, J. Manuf. Syst. 65 (2022) 180–191, http://dx.doi.org/10.1016/j.jmsy.2022.09.004.
[53] Case western reserve university (CWRU) bearing data center, 2019, https://engineering.case.edu/bearingdatacenter/download-data-file.
[54] K. Li, School of mechanical engineering, Jiangnan university, 2020, http://mad-net.org:8765/explore.html?t=0.5831516555847212.
[55] J. Si, H. Shi, T. Han, J. Chen, C. Zheng, Learn generalized features via multi-source domain adaptation: Intelligent diagnosis under variable/constant machine conditions, IEEE Sens. J. 22 (1) (2022) 510–519, http://dx.doi.org/10.1109/JSEN.2021.3126864.
[56] D. Picard, Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision, 2021, http://dx.doi.org/10.48550/arXiv.2109.08203, abs/2109.08203.
[57] Z. Zhao, T. Li, J. Wu, C. Sun, S. Wang, R. Yan, X. Chen, Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study, ISA Trans. 107 (2020) 224–255, http://dx.doi.org/10.1016/j.isatra.2020.08.010.
[58] W. Zhang, G. Peng, C. Li, Y. Chen, Z. Zhang, A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals, Sensors 17 (2) (2017) 425, http://dx.doi.org/10.3390/s17020425.
[59] M. Zhao, S. Zhong, X. Fu, B. Tang, M. Pecht, Deep residual shrinkage networks for fault diagnosis, IEEE Trans. Ind. Inform. 16 (7) (2020) 4681–4690, http://dx.doi.org/10.1109/TII.2019.2943898.
[60] Z. Zhao, Y. Jiao, A fault diagnosis method for rotating machinery based on CNN with mixed information, IEEE Trans. Ind. Inform. (2022) 1–11, http://dx.doi.org/10.1109/TII.2022.3224979.
[61] Z. Zhao, Q. Zhang, X. Yu, C. Sun, S. Wang, R. Yan, X. Chen, Applications of unsupervised deep transfer learning to intelligent fault diagnosis: A survey and comparative study, IEEE Trans. Instrum. Meas. 70 (2021) 1–28, http://dx.doi.org/10.1109/TIM.2021.3116309.
[62] R. van de Sand, S. Corasaniti, J. Reiff-Stephan, Data-driven fault diagnosis for heterogeneous chillers using domain adaptation techniques, Control Eng. Pract. 112 (2021) 104815, http://dx.doi.org/10.1016/j.conengprac.2021.104815.
[63] N. Lu, H. Hu, T. Yin, Y. Lei, S. Wang, Transfer relation network for fault diagnosis of rotating machinery with small data, IEEE Trans. Cybern. 52 (11) (2022) 11927–11941, http://dx.doi.org/10.1109/TCYB.2021.3085476.
[64] S. Schwendemann, Z. Amjad, A. Sikora, Bearing fault diagnosis with intermediate domain based layered maximum mean discrepancy: a new transfer learning approach, Eng. Appl. Artif. Intell. 105 (2021) 104415, http://dx.doi.org/10.1016/j.engappai.2021.104415.
[65] K. Zhao, H. Jiang, Z. Wu, T. Lu, A novel transfer learning fault diagnosis method based on Manifold Embedded Distribution Alignment with a little labeled data, J. Intell. Manuf. 33 (1) (2022) 151–165, http://dx.doi.org/10.1007/s10845-020-01657-z.
[66] X. Wang, C. Shen, M. Xia, D. Wang, J. Zhu, Z. Zhu, Multi-scale deep intra-class transfer learning for bearing fault diagnosis, Reliab. Eng. Syst. Saf. 202 (2020) 107050, http://dx.doi.org/10.1016/j.ress.2020.107050.
[67] Y. Xiao, H. Shao, S. Han, Z. Huo, J. Wan, Novel joint transfer network for unsupervised bearing fault diagnosis from simulation domain to experimental domain, IEEE/ASME Trans. Mechatron. 27 (6) (2022) 5254–5263, http://dx.doi.org/10.1109/TMECH.2022.3177174.
[68] J. Zhong, Y. Huang, Time-frequency representation based on an adaptive short-time Fourier transform, IEEE Trans. Signal Process. 58 (10) (2010) 5118–5128, http://dx.doi.org/10.1109/TSP.2010.2053028.
[69] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (86) (2008) 2579–2605, http://jmlr.org/papers/v9/vandermaaten08a.html.
[70] M. Leiber, A. Barrau, Y. Marnissi, D. Abboud, A differentiable short-time Fourier transform with respect to the window length, in: 30th European Signal Processing Conference, EUSIPCO 2022, Belgrade, Serbia, August 29 - Sept. 2, 2022, 2022, pp. 1392–1396, http://dx.doi.org/10.23919/EUSIPCO55093.2022.9909963.
[71] M. Leiber, Y. Marnissi, A. Barrau, M.E. Badaoui, Differentiable adaptive short-time Fourier transform with respect to the window length, in: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5, http://dx.doi.org/10.1109/ICASSP49357.2023.10095245.
[72] S. Vollert, M. Atzmueller, A. Theissler, Interpretable Machine Learning: A brief survey from the predictive maintenance perspective, in: 2021 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), 2021, pp. 01–08, http://dx.doi.org/10.1109/ETFA45728.2021.9613467.