Research Article

Spacecraft Intelligent Fault Diagnosis under Variable Working Conditions via Wasserstein Distance-Based Deep Adversarial Transfer Learning

Gang Xiang^{1,2} and Kun Tian²

¹School of Automation and Electrical Engineering, Beihang University, Beijing 100191, China

²Beijing Aerospace Automatic Control Institute, Beijing 100854, China

Correspondence should be addressed to Gang Xiang; xianggang@buaa.edu.cn

Received 31 May 2021; Revised 25 August 2021; Accepted 30 August 2021; Published 13 October 2021

Academic Editor: Giovanni Palmerini

Copyright © 2021 Xiang Gang and Tian Kun. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In recent years, deep learning methods which promote the accuracy and efficiency of fault diagnosis task without any extra requirement of artificial feature extraction have elicited the attention of researchers in the field of manufacturing industry as well as aerospace. However, the problems that data in source and target domains usually have different probability distributions because of different working conditions and there are insufficient labeled or even unlabeled data in target domain significantly deteriorate the performance and generalization of deep fault diagnosis models. To address these problems, we propose a novel Wasserstein Generative Adversarial Network with Gradient Penalty- (WGAN-GP-) based deep adversarial transfer learning (WDATL) model in this study, which exploits a domain critic to learn domain invariant feature representations by minimizing the Wasserstein distance between the source and target feature distributions through adversarial training. Moreover, an improved one-dimensional convolutional neural network- (CNN-) based feature extractor which utilizes exponential linear units (ELU) as activation functions and wide kernels is designed to automatically extract the latent features of raw time-series input data. Then, the fault model classifier trained in one working condition (source domain) with sufficient labeled samples could be generalized to diagnose data in other working conditions (target domain) with insufficient labeled samples. Experiments on two open datasets demonstrate that our proposed WDATL model outperforms most of the state-of-the-art approaches on transfer diagnosis tasks under diverse working circumstances.

1. Introduction

Fault diagnosis utilizing the data acquired by monitoring the equipment is a technology which aims to detect the potential anomalies as early as possible and identify the root causes of failures and help make maintenance decision to avoid catastrophic breakdown. Therefore, fault diagnosis is urgently required and of great importance for spacecraft [1, 2].

In the past few decades, some advanced signal processing methods [3] and machine learning algorithms [4, 5] were employed for spacecraft fault diagnosis. Machine learning usually plays an essential role in exploring the mapping relationships between features extracted by the signal processing techniques and the health states of spacecraft [6, 7]. How-

ever, these methods usually require abundant experience and sufficient expert knowledge [8], making them time-consuming and labor-intensive. With the rapid increase of test data and fast development in data-driven technology, deep learning models, such as Deep Belief Network (DBN) [9, 10], Sparse Auto-Encoder (SAE) [11, 12], convolutional neural network (CNN) [8, 13–15], and Recurrent Neural Network (RNN) [16, 17], can extract effective features and show superior learning capability in fault diagnosis. These deep models are becoming research hotspots in the fault diagnosis of industrial equipment and spacecraft [8, 18, 19]. However, these algorithms must be subject to a basic hypothesis that the training and test data should be sampled from identical distribution. Unfortunately, this hypothesis

does not always hold in most aerospace application scenarios. For example, in the area of spacecraft fault diagnosis, training dataset and test dataset are often collected in different working conditions, which results in shift of data distribution. Therefore, the diagnosis model learned in the training dataset (usually contains sufficient labeled samples) could not be generalized well to the test dataset (usually contains few or even none labeled samples).

In order to tackle these challenges, transfer learning (TL) approaches [20] aiming to leverage knowledge obtained in source domain (with adequate labeled samples under one working condition) to train a model that could generalize to a target domain (with insufficient labeled or unlabeled data under a novel but related working condition) have been introduced into fault diagnosis of spacecraft task and show superior performance in cross-domain fault diagnosis scenarios [21–23]. TL can be roughly classified into four categories: instance-based TL [24], parameter-based TL [25], relation-based TL [26], and feature-based TL [27, 28]. Among them, the feature-based TL is the most investigated method because of the ability of correcting cross-domain discrepancy. The Transfer Component Analysis (TCA) [29] is a typical feature-based TL method, which adopts Maximum Mean Discrepancy (MMD) [30–32] in a low-dimensional feature space as distance metric to reduce distribution discrepancy.

Inspired by generative adversarial network (GAN), more and more researchers have found that learning domain-invariant features using GAN is a novel approach in the field of TL [21, 33–35]. Through the adversarial training, domain-invariant features could be learned when the networks reach a state known as Nash equilibrium [36]. When training the GAN, there may be some difficulties to be solved, such as the instability of learning and mode collapse [37]. Arjovsky et al. [38] proposed Wasserstein GAN (WGAN), which introduced a more sensible Wasserstein distance to replace Jensen-Shannon (JS) divergence and Kullback-Leibler (KL) divergence as the loss function in GAN. To alleviate the gradient vanishing or exploding problem when training the WGAN, Gulrajani et al. [39] proposed a new WGAN with gradient penalty (WGAN-GP) term for domain alignment. Miyato et al. [40] proposed the Spectrally Normalized GAN (SNGAN), which found the maximum singular value of the weight matrix in the discriminator. The motivation of this paper is to explore how the Wasserstein-based adversarial learning method performs in the field of spacecraft fault diagnosis.

In this article, we propose a new WGAN-GP-based deep adversarial transfer learning (WDATL) model to tackle the problems of spacecraft fault diagnosis under various working conditions. The WDATL model contains four components: source domain feature extractor, target domain feature extractor, fault model classifier, and domain critic. The source and target domain feature extractors are deep convolutional neural networks which learn feature representations from the raw one-dimensional (1-D) input signals in source and target domain, respectively. The domain critic minimizes the discrepancy of the source and target feature distributions using WGAN-GP. The fault model classifier

is designed to map the features with corresponding fault labels by optimizing the loss function. Through the above adversarial learning process, the domain-invariant and fault model discriminative feature representations could be learned to help promote the diagnosis performance in the target domain. The main contributions of this paper can be summarized as follows:

- (1) We propose a novel WGAN-GP-based deep adversarial transfer learning (WDATL) model, which could learn domain-invariant features and promote diagnosis performance under multiple working conditions. To the best of our knowledge, it is the first attempt in the field of spacecraft fault diagnosis, which prospectively overcomes the obstacles when applying deep learning methods to real engineering scenarios
- (2) We propose an improved 1-D convolutional neural network, which builds an end-to-end model to extract features automatically without requiring experienced experts and a high skill set. We design wide kernels in the top convolutional layer to suppress high-frequency noise and promote robustness of our model. Moreover, we adopt exponential linear units (ELU) instead of commonly used rectified linear units (ReLU) as activation function to speed up learning
- (3) The t-SNE is used to visualize the feature representations learned by the feature extractor for our proposed method and other compared models, helping to validate the superior performance of the WDATL
- (4) Experimental results on two public datasets demonstrate that our proposed WDATL model performs well on diverse working circumstances and achieves significantly better results than most of the baseline approaches

The rest of this work is organized as follows. Preliminaries including transfer learning and generative adversarial network are introduced in Section 2. Section 3 details the proposed fault diagnosis methods under various working circumstances. Extensive experiments are conducted, and results are discussed in Section 4. Finally, conclusions are drawn in Section 5.

2. Preliminaries

2.1. Transfer Learning. TL could utilize knowledge from one or multiple related datasets called source domains to promote the model's diagnosis accuracy and performance in the current dataset called target domain.

There are two basic concepts in TL, domain and task [20]. A domain is denoted as $\mathcal{D} = \{\mathcal{X}, \mathbb{P}(X)\}$, which includes two components: a feature space $\mathcal{X}$ and a marginal probability distribution $\mathbb{P}(X)$. $X = \{x_1, x_2, x_3, \dots, x_n\} \in \mathcal{X}$ is a n -dimensional vector. There exists source domain $\mathcal{D}^s$ and target domain $\mathcal{D}^t$. If $X^s \neq X^t$ and (or) $\mathbb{P}(X^s) \neq \mathbb{P}(X^t)$,

then we consider that two domains $\mathcal{D}^s$ and $\mathcal{D}^t$ are with different distribution, i.e., $\mathcal{D}^s \neq \mathcal{D}^t$. A task $\mathcal{T} = \{Y, \mathbb{P}(Y|X)\}$ is also composed of two components: a label space Y and a prediction function which is equal to the conditional probability distribution $\mathbb{P}(Y|X)$.

Given a source domain $\mathcal{D}^s$ and learning task $\mathcal{T}^s$, a target domain $\mathcal{D}^t$ and learning task $\mathcal{T}^t$, TL aims to promote the performance of target predictive function in $\mathcal{D}^t$ through leveraging the knowledge in $\mathcal{D}^s$ and $\mathcal{T}^s$, where $\mathcal{D}^s \neq \mathcal{D}^t$ or $\mathcal{T}^s \neq \mathcal{T}^t$. Usually, there are sufficient labeled samples $\{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ in source domain to train a diagnosis model, in which N_s is the size of dataset and y_i^s is the corresponding label of x_i^s . While in the target domain, samples are not labeled, and our aim is to construct a deep learning model which could diagnose or classify the unlabeled data through learning domain-invariant features.

2.2. Generative Adversarial Network. GAN proposed by Goodfellow et al. [41] in 2014 consists of two independent networks: a generator network $G(z; \theta)$ and a discriminator network $D(x; \phi)$ as shown in Figure 1, where θ and ϕ denote the parameters in generator and discriminator, respectively. The generator aims to learn the distribution of the generator $p_\theta(x)$ from a randomly generated variable z , while the discriminator tries to differentiate between the real data distribution $p_r(x)$ and the fake data distribution $p_\theta(x)$ generated by generator network. When GAN reaches a stage at which the discriminator is unable to distinguish between real and fake data distribution, the stage is called Nash equilibrium.

In summary, the optimizing object of GAN is a min-max problem, and the loss function $\mathcal{L}$ can be formulated as

$$\begin{aligned} \mathcal{L} &= \mathbb{E}_{x \sim p_r(x)} \left[\log \frac{p_r(x)}{p_r(x) + p_\theta(x)} \right] + \mathbb{E}_{x \sim p_\theta(x)} \left[\log \frac{p_\theta(x)}{p_r(x) + p_\theta(x)} \right] \\ &= 2\text{JS}(p_r, p_\theta) - 2 \log 2, \end{aligned} \quad (1)$$

where $\mathbb{E}_{x \sim p_r(x)}(*)$ calculates the expectation and the variable x follows the distribution of $p_r(x)$. $\text{JS}(\cdot)$ is the Jensen-Shannon divergence which measures the difference between $p_r(x)$ and $p_\theta(x)$. When there are no overlaps between $p_r(x)$ and $p_\theta(x)$, the gradients of loss function $\mathcal{L}$ will become zero. Therefore, the Jensen-Shannon divergence is not suitable and reasonable for the loss function of GAN, which may cause instability of training and mode collapse [37]. In order to address this problem, Wasserstein distance was first used in [38] to evaluate the difference between margin distributes p_r and p_θ instead of Jensen-Shannon divergence. The Wasserstein distance definition reads as

$$W^1(p_r, p_\theta) = \inf_{\gamma \sim \Gamma(p_r, p_\theta)} \mathbb{E}_{(x,y) \sim \gamma} [\|x - y\|], \quad (2)$$

where γ denotes the joint probability distribution and $\Gamma(p_r, p_\theta)$ is the set of all joint distributions $\gamma(p_r, p_\theta)$. However, Equation (2) is extremely intractable to calculate,

FIGURE 1: The schematic diagram of GAN.

and it can be rewritten according to the Kantorovich-Rubinstein duality.

$$W^1(p_r, p_\theta) = \sup_{\|f\| \leq 1} \left(\mathbb{E}_{x \sim p_r} [f(x)] - \mathbb{E}_{x \sim p_\theta} [f(x)] \right), \quad (3)$$

where $f: \mathbb{R}^d \rightarrow \mathbb{R}$ is a 1-Lipschitz function subjected to

$$\|f\|_L \triangleq \sup_{x \neq y} \frac{|f(x) - f(y)|}{|x - y|} \leq 1. \quad (4)$$

Eventually, calculating the Wasserstein distance between p_r and p_θ could be converted into exploring the maximum expectation difference of a continuous function subjected to 1-Lipschitz under distribution p_r and p_θ .

3. Proposed Method

3.1. The Overall Structure of WDATL. Inspired by GAN, many researchers have introduced the adversarial learning strategy into the field of fault diagnosis [36]. We detail the newly proposed WDATL approach in this section to solve the difficulties of spacecraft fault diagnosis under diverse working circumstances. The WDATL model is composed of four parts: source domain feature extractor, target domain feature extractor, fault model classifier, and domain critic. The overall structure of WDATL is shown in Figure 2.

Unlike the generator aiming to generate fake data in Figure 1, the generator called source and target feature extractors in WDATL is implemented to extract feature characteristics from source input data and target input data, respectively. The source and target feature extractors are with the same network architecture sharing the identical parameters. The domain critic is designed to decrease the discrepancy of feature distributions between source and target data extracted by the feature extractors, and the fault model classifier is implemented to predict the corresponding labels. Through adversarial training, the domain-invariant and fault model discriminative characteristics could be learned, so the fault model classifier could not only classify the labeled data in source domain but also be generalized and applied to diagnose the unlabeled data in target domain with a satisfying accuracy.

3.2. Feature Extraction and Pretraining

3.2.1. Improved CNN-Based Feature Extraction. In the TL-based fault diagnosis under various working conditions, we assume that there are sufficient labeled data denoted as

FIGURE 2: The overall structure of WDATL.

$\{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ available in source domain to train a deep neural network in a supervised way. Due to the excellent feature extraction and powerful fitting abilities of CNN, we introduce one dimensional (1-D) CNN to the spacecraft fault diagnosis and propose an improved 1-D CNN-based feature extractor, which is an end-to-end deep model working directly on raw time-serial data and helping automatically learns fault characteristics without any requirement of artificial feature extraction. The source and target feature extractors in our proposed WDATL model have the same network architecture and share the identical parameters.

The feature extractor in the proposed WDATL contains five 1-D convolutional layer-pooling layer pairs that come with 1-D Batch Normalization (BN) layers to tackle the issue of internal covariate shifting and dropout layer in the last layer to avoid overfitting. Moreover, there is a flatten layer behind the fifth convolutional layer-pooling layer pair, and the number of neurons in the flatten layer is equal to the size of output features. The improved 1-D CNN-based feature extractor is illustrated in Figure 3.

The convolutional layer in WDATL convolves the raw input data with 1-D filter kernels and is then followed by an activation function to generate output features. Usually, there are multiple kernels in each convolutional layer, which aims to learn comprehensive feature characteristics. The mathematical formula of the convolutional layer is as follows:

$$x^l = f(W^l x^{l-1} + b^l), \quad (5)$$

where W^l represents kernels of the l th layer. x^{l-1} and x^l denote input and output of the l th layer, respectively. b^l is the bias term, and the f is the activation function, such as sigmoid, tanh, and ReLU, which enables the layer to acquire the ability of nonlinearity.

The pooling operation after a convolutional layer aims to select and reduce the dimensions of features, making the features more robust after convolution. Max pooling and average pooling are two most commonly used pooling operations. In our WDATL, we use max-pooling in the former fourth convolutional layers and adaptive max-pooling in the last convolutional layer, which takes the maximum value in the input features.

Specifically, we modify and improve the CNN-based feature extractor mainly in two aspects.

Firstly, we adopt large 64×1 kernels instead of small kernels in the initial convolutional layer. In real industrial application, equipment, especially spacecraft, often work in various conditions and harsh environments, resulting in much noise in data. The larger kernel could extract more comprehensive fault information because large size of kernel has better characterization capability and is less sensible to noise [42] compared with small kernel. Then, small kernels are followed for formulating deeper network architecture and finer feature learning.

Secondly, we utilize exponential linear unit (ELU) instead of the commonly used ReLU as activation function. A drawback of ReLU lies in that it has zero gradients whenever the input is negative, which may lead to neural units being inactivated. This will greatly restrict the learning capability of CNN [43]. To overcome the disadvantage of ReLU, we propose to use ELU [44] in our WDATL model. The ELU is as

$$\text{ELU}(x) = \begin{cases} \alpha(e^x - 1) & \text{for } x \leq 0, \\ x & \text{for } x > 0, \end{cases} \quad (6)$$

where $\alpha \geq 0$. In contrast to ReLU, the ELU does not only produce a permanent active state for all inputs but also saturates to a negative value when the input is a small negative value, which decreases the forward propagated variation, making it more robust to noise. Moreover, ELU could accelerate training of deep neural networks and bring about higher classification accuracies.

3.2.2. Pretraining in Source Domain. The fault model classifier consists of 2 fully connected forward neural layers, which is designed to directly map the feature representations learned by the feature extractor to their corresponding fault states of spacecraft. A softmax activation function is added to the last layer to predict the labels of input representations.

The first step in our WDATL is to pretrain the source feature extractor and fault model classifier by employing the labeled data x^s in source domain, and the corresponding parameters are updated by minimizing the cross-entropy loss $l_c(x^s, y^s)$ between the predicted labels $\hat{y}_i^s$ and ground true labels y_i^s using the Adam [45] algorithm.

$$l_c(x^s, y^s) = -\frac{1}{N_s} \sum_{i=1}^{N_s} \sum_{k=1}^K l(y_i^s = k) \log \hat{y}_i^s, \quad (7)$$

FIGURE 3: Architecture of CNN-based feature extractor.

FIGURE 4: The framework of proposed WDATL.

where $l(y_i^s = k)$ is the indicator function, and when $y_i^s = k$, it returns 1 while $y_i^s \neq k$, it returns 0. K represents the number of fault models. The purposes of pretraining are not only to initialize the target domain feature extractor by sharing the same parameters with source domain feature extractor but also to set a reference model during adversarial training to accelerate training process.

3.3. Adversarial Learning of the WDATL. The schematic of the WDATL model proposed in this article is illustrated in Figure 4. The essential aim of adversarial learning in WDATL is a min-max game among the feature extractors, domain critic, and fault model classifier. More specially, the domain critic tries to decrease the distribution discrepancy between the source and target features by using the

Wasserstein distance, while the feature extractor is trained to extract target features which could maximize the output of domain critic and the fault model classifier is trained to classify the source features with minimal cross-entropy loss. Thereby, the fault model classifier trained in this adversarial manner could be generalized to diagnose the target domain data (under different working conditions) with unsupervised learning.

Given two minibatch of samples $\{x_i^s\}_{i=1}^n$ and $\{x_i^t\}_{i=1}^n$ collected from source domain x^s and target domain x^t , in which n represents the size of minibatch and $n < N_s, n < N_t$. The source features $r^s = f_{\theta_e}(x^s)$ and the target features $r^t = f_{\theta_e}(x^t)$, where θ_e denotes the parameters of feature extractor. Let p^s and p^t be the margin distribution of source and target features, respectively.

FIGURE 5: The flowchart of the proposed WDATL.

The domain critic f_{θ_d} consists of two fully connected neural layers, whose output is the scores corresponding to their inputs. Assume that there exists a set of parameters θ_d , which makes f_{θ_d} satisfy Lipschitz constraint, according to Equation (3), the Wasserstein distance l_{wd} between p^s and p^t could be computed by

$$\begin{aligned}
 l_{wd} &= \max \left(\mathbb{E}_{x \sim p^s} [f_{\theta_d}(x)] - \mathbb{E}_{x \sim p^t} [f_{\theta_d}(x)] \right) \\
 &= \max \left(\frac{1}{N_s} \sum_{x^s \in X^s} f_{\theta_d}(f_{\theta_e}(x^s)) - \frac{1}{N_t} \sum_{x^t \in X^t} f_{\theta_d}(f_{\theta_e}(x^t)) \right). \quad (8)
 \end{aligned}$$

In order to minimize the distribution discrepancy between the source and target features, the domain critic attempts to output a big score for source feature while a small score for target features.

In literature [38], Arjovsky et al. adopted a weight-clipping approach to drive the parameters θ_d inside a limit space to satisfy Lipschitz constraint; however, this will cause gradient vanishing or exploding. Gulrajani et al. [39] proposed a new gradient penalty (WGAN-GP) term added to the l_{wd} as a regularization, so the final optimization object of domain critic is

$$l_d = \max (l_{wd} - \rho l_{\text{grad}}), \quad (9)$$

where ρ denotes balancing coefficient, and $l_{\text{grad}} = (\|\nabla_h f(h)\|_2 - 1)^2$ represents the gradient penalty term.

As for the target feature extractor, the aim is to maximize the output of domain critic and the loss function is computed by

$$l_e = \max \frac{1}{N_t} \sum_{x^t \in X^t} f_{\theta_d}(f_{\theta_e}(x^t)). \quad (10)$$

Moreover, the fault model classifier should be trained to minimize the cross-entropy loss in Equation (7). Through adversarial learning among Equations (7), (9), and (10), domain-invariant and fault model discriminative features could be learned. The flowchart of the proposed WDATL is illustrated in Figure 5. Therefore, the classifier trained in source domain could be generalized to classify data in target data. In this way, the unlabeled target domain is expected to benefit from knowledge in the source domain with sufficient supervised information.

3.4. Epilog. The training strategy of our WDATL approach is summarized in Algorithm 1, which can be divided into 5 steps as follows:

- (1) The feature extractor and fault model classifier are first pretrained in source domain with sufficient labeled data, and the parameters of θ_e and θ_c can be updated using Adam
- (2) Then, source and target data are fed into the pre-trained feature extractor to get source features and target features, and domain critic tries to decrease

Require 1: source dataset X^s , the size of mini-batch n , the pre-training epochs Num1, learning rate of feature extractor α_1 , learning rate of fault model classifier α_2 .

Require 2: target dataset X^t , the adversarial learning epochs Num2, learning rate of domain critic α_3 , domain critic training steps n_d .

Pre-training:

- 1: Initialize the parameters of θ_e and θ_c .
- 2: for $i = 1, 2, 3, \dots, \text{Num1}$ do.
- 3: Sample mini-batch $\{(x_i^s, y_i^s)\}_{i=1}^n$ from X^s
- 4: $\theta_e \leftarrow \theta_e - \alpha_1 \nabla_{\theta_e} l_c(x^s, y^s)$
- 5: $\theta_c \leftarrow \theta_c - \alpha_2 \nabla_{\theta_c} l_c(x^s, y^s)$
- 6: end for
- 7: the target feature extractor shares the parameters with source feature extractor.

Adversarial training:

- 8: Initialize the parameters of θ_d .
- 9: for $i = 1, 2, 3, \dots, \text{Num2}$ do
- 10: Sample mini-batch $\{(x_i^s, y_i^s)\}_{i=1}^n$ from X^s
- 11: Sample mini-batch $\{x_i^t\}_{i=1}^n$ from X^t
- 12: for $i = 1, 2, 3, \dots, n_d$ do
- 13: $r^s \leftarrow f_{\theta_e}(x^s), r^t \leftarrow f_{\theta_e}(x^t)$
- 14: $l_{\text{grad}} = (\|\nabla_h f(h)\|_2 - 1)^2$
- 15: $\theta_d \leftarrow \theta_d - \alpha_3 \nabla_{\theta_d} (-l_d)$
- 16: end for
- 17: $\theta_e \leftarrow \theta_e - \alpha_1 \nabla_{\theta_e} (l_c + \lambda l_e)$
- 18: $\theta_c \leftarrow \theta_c - \alpha_2 \nabla_{\theta_c} l_c(x^s, y^s)$
- 19: end for

ALGORITHM 1: Training strategy and procedure of WDATL.

the distribution discrepancy between the source and target features according to Equation (9)

- (3) The parameters of θ_e are updated according to Equation (10) while θ_c are updated according to Equation (7)
- (4) Repeat step (2) and step (3) until the WDATL model reaches Nash equilibrium, then store all the parameters
- (5) Finally, the trained feature extractor and fault model can be generalized to target domain to diagnose the unlabeled data with a relatively satisfying accuracy

4. Experiments

Our proposed WDATL model is a data-driven approach analyzing and mining the time-series data sampled from various sensors onboard spacecraft or telemetry channels. In order to verify the effectiveness and superior diagnosis performance of our proposed approach compared with other state-of-the-art methods in various working circumstances, we conduct experiments on two open and public bearing and gear datasets like many other literatures [15, 31, 32, 42, 46, 47]. However, this does not mean that our method is only suitable for rotating machineries; in fact, any time-series data collected from spacecraft could be used as input data of our method.

4.1. Implementation Details. For comparison, CNN (without transfer) [15, 42], Deep Adaption Network (DAN) [31],

Joint Adaption Network (JAN) [32], Correlation Alignment (CORAL) [46], and Domain Adversarial Neural Network (DANN) [47] are investigated.

The CNN method (without transfer) uses convolutional layers to extract features and use fully connected layers to map the features with corresponding fault models, which are trained in the source domain and then directly adopted to diagnose data in the target domain.

Both DAN and JAN are deep adaption methods, which try to reduce the distribution distance of the source and target domains by mapping the latent features in fully connected layers into Reproducing Kernel Hilbert Space (RKHS) to reduce the MMD distance. The DAN uses MK-MMD to tackle the parameter selection problems and adds the MK-MMD loss to the classification loss forming the final objective function. The JAN utilizes JMMD to measure the distances of empirical joint distributions between source and target domains to align the domain shift. The parameter setting of DAN and JAN can refer to [31, 32]. CORAL aims to decrease the distribution discrepancy of source and target domains by aligning the second-order statistic.

DANN is a deep adversarial-based model, which refers to train a feature extractor to learn characteristics from input data and a domain discriminator to distinguish source and target domains. The DANN attempts to decrease the margin distributions between source and target features.

The structure and hyperparameters in the CNN-based fault diagnosis model directly determine the classification and diagnosis performance, so it is unfair and difficult to decide whose network is better. Therefore, in this paper, the five compared deep models share the same CNN

TABLE 1: Details of proposed CNN deep model in WDATL.

No.	Submodel	Layer	Filters	Kernel size	Output	Stride	Padding	Activation function
1	Feature extractor	Conv1d 1	20	64×1	(64, 20, 964)	1	Yes	ELU
2		Maxpooling 1	/	2×1	(64, 20, 482)	2	No	/
3		Conv1d 2	40	3×1	(64, 40, 482)	1	Yes	ELU
4		Maxpool 2	/	2×1	(64, 40, 241)	2	No	/
5		Conv1d 3	40	3×1	(64, 40, 241)	1	Yes	ELU
6		Maxpool 3	/	2×1	(64, 40, 120)	2	No	/
7		Conv1d 4	60	3×1	(64, 60, 120)	1	Yes	ELU
8		Maxpool 4	/	2×1	(64, 60, 60)	2	No	/
9		Conv1d 5	128	3×1	(64, 128, 58)	1	No	ELU
10		AdaptiveMaxpool	/	/	(64, 128, 6)	/	No	/
11	Fault model	Fully-connected 1	/	/	(64, 256, 1)	/	/	ELU
12		Fully-connected 1	/	/	(64, 128, 1)	/	/	ELU
13		Fully-connected 2	/	/	(64, 10, 1)	/	/	Softmax
14		Fully-connected 1	/	/	(64, 256, 1)	/	/	ELU
15	Critic	Fully-connected 2	/	/	(64, 1, 1)	/	/	Sigmoid

FIGURE 6: The test rig of CWRU.

structure with the proposed WDATL. The CNN-based deep model is illustrated in Figure 3, and the parameters are detailed in Table 1.

As mentioned before, the DAN, JAN, CORAL, and DANN are the TL method, which pretrain 60 epochs in the source domain, then TL methods are activated, and the maximum epoch is 360. We use the minibatch Adam as the optimizer, and the size of minibatch is set to 64 according to the size of the input samples. The learning rates α_1 , α_2 , and α_3 of all the neural networks are initially set to 0.001 with a step decay (multiplied by 0.1) in the epoch 200 and 300, respectively, which not only guarantees the speed of convergence but also avoids oscillation near the optimal point. The balancing coefficient ρ is set to 5, while the λ is set to 1. All the experiments are executed by using Pytorch 1.3 with GPU.

4.2. Results and Analysis in CWRU Dataset

4.2.1. Data Description and Preprocessing. The CWRU bearing dataset is provided by the Electrical Engineering Laboratory of Case Western Reserve University [48] and has become one of the most well-known public datasets in the field of fault diagnosis. Nowadays, many literatures conducted experiments on CWRU to compare the effectiveness of different diagnosis algorithms [36]. The test rig of CWRU is shown in Figure 6.

Following most of the published articles, we use the vibration data collected from the drive end and the sampling frequency is equal to 12 kHz. One healthy condition and three kinds of single-point bearing faults containing inner race fault (IF), outer race fault (OF), and ball fault (BF) are separated into 10 categories (one normal category and 9

faulty categories) according to the size of faults, i.e., 7 mils, 14 mils, and 21 mils (1mil = 0.001 in.). The description of the CWRU dataset is illustrated in Table 2. In addition, this dataset is sampled under four working circumstances with different motor loads and operating speeds, including 0 hp/1797 rpm, 1 hp/1772 rpm, 2 hp/1750 rpm, and 3 hp/1730 rpm, which is shown in Table 3. In this paper, each working condition represents a task, so $0 \rightarrow 1$ represents that the source domain contains data sampled in 0 hp/1797 rpm and the target domain contains data collected in 1 hp/1797 rpm. Totally, there are 12 transfer scenarios in CRWU.

The data in CWRU are 1-D time series, and we split the data into small slices without any overlapping, and each slice denotes a sample with the length of 1024. If the length of raw signal is $L (L \geq 120000)$, then we can get N samples, where $N = \text{floor}(L/1024)$. We randomly select 80% of the total samples as the training dataset, and the remaining 20% of total samples as the test dataset. The feature extractor directly processes the raw time-serial slices avoiding labor-intensive artificial feature extraction.

Input normalization which limits the input data into a certain range can speed up the training procedure; therefore, it is very vital for deep models. So, we use the Z-score normalization to process the input data. The Z-score normalization is formulated as follows:

$$\hat{x}_i = \frac{x_i - \mu}{\sigma}, \quad (11)$$

where x_i is the input data and μ and σ represent the mean value and standard deviation of x_i , respectively.

4.2.2. Result Analysis. In this paper, we use classification accuracy defined as below to compare the performance of different algorithms.

$$\text{Accuracy} = \frac{|x : x \in D_{t'} \wedge \hat{y}_{t'} = y_{t'}|}{|x : x \in D_{t'}|}, \quad (12)$$

where $D_{t'}$ represents the test dataset of target domain.

We carry out the experiments six times and select the mean accuracy for each algorithm to avoid randomness. The classification accuracies of 12 transfer scenarios in CWRU are provided in Table 4, from which we can draw the following conclusions:

- (1) The proposed WDATL approach outperforms other compared methods in the total 12 transfer diagnosis tasks with a mean diagnosis accuracy of 99.91%
- (2) The 1-D convolutional neural network-based feature extractor proposed in this paper has strong feature extraction and generalization capabilities without any TL. Therefore, the CNN method could get high diagnosis accuracy in target domain by directly applied network trained in source domain. Namely, for the transfer task $1 \rightarrow 2$, the diagnosis accuracy on the target domain is as high as 99.87% by simply using the CNN method

TABLE 2: The details of CWRU.

Label	0	1	2	3	4	5	6	7	8	9
Fault label	NA	IF	BF	OF	IF	BF	OF	IF	BF	OF
Fault size (mils)	0	7	7	7	14	14	14	21	21	21

TABLE 3: Transfer scenarios in CRWU.

Task	0	1	2	3
Load (HP)	0	1	2	3
Speed (rpm)	1797	1772	1750	1730

TABLE 4: Accuracy (%) of different algorithms for different working conditions.

	CNN	DAN	JAN	CORAL	DANN	WDATL
$0 \rightarrow 1$	97.01	100	100	99.03	100	100
$0 \rightarrow 2$	89.67	100	100	100	100	100
$0 \rightarrow 3$	84.13	99.68	100	100	100	100
$1 \rightarrow 0$	96.55	93.87	95.79	96.55	99.62	100
$1 \rightarrow 2$	99.87	100	100	100	100	100
$1 \rightarrow 3$	92.16	100	100	99.35	100	100
$2 \rightarrow 0$	95.25	89.66	99.62	93.87	98.85	99.81
$2 \rightarrow 1$	93.24	99.68	99.03	99.03	98.70	99.87
$2 \rightarrow 3$	87.11	100	100	100	100	100
$3 \rightarrow 0$	78.77	86.21	90.55	91.57	99.22	99.61
$3 \rightarrow 1$	84.28	99.35	84.74	92.53	97.73	99.61
$3 \rightarrow 2$	90.97	100	100	100	99.68	100
Average	90.75	97.37	97.48	97.66	99.48	99.91

FIGURE 7: Radar diagram of diagnosis accuracy for different algorithms.

FIGURE 8: Training process of task 3 $\rightarrow$ 0 using WDATL.

- (3) For tasks such as 1 $\rightarrow$ 2 and 2 $\rightarrow$ 3 with small distribution shifts between source and target domains, DAN, JAN, CORAL, DANN, and WDATL algorithms can achieve 100% diagnosis accuracies in the target domain, which means only simple transfer strategies could get excellent transfer effectiveness
- (4) Due to the consideration of conditional distribution, the average diagnosis accuracy (97.48%) of JDA is slightly higher than the average diagnosis accuracy (97.37%) of DAN
- (5) The diagnosis accuracy of adversarial TL algorithms (DANN, WDATL) is generally higher than that of MMD-based transfer algorithms, which indicates that adversarial learning can help align the domain shift better

The radar diagram of diagnosis accuracy for different algorithms in the 12 transfer tasks is shown in Figure 7.

4.2.3. Visualization Analysis. The WDATL algorithm proposed in this paper has good performance of convergence. The loss functions of the domain critic l_d and feature extractor l_e during the training process of the transfer task 3 $\rightarrow$ 0 are shown in Figure 8.

In order to further prove the superiority of the fault diagnosis performance of the WDATL under various working circumstances, the confusion matrix and t-SNE [49] are used to visualize the results, respectively. The confusion matrixes of test data in task 3 $\rightarrow$ 0 with different methods are detailed in Figure 9.

The coordinates in Figure 9 indicate the different health states of the bearings. The figures in Figures 9(a), 9(c), 9(e), 9(g), 9(i), and 9(k) show the confusion matrixes of the CNN, DAN, JAN, CORAL, DANN, and WDATL algorithms on

the source test data, respectively. From Figure 9, we can conclude that the discussed six methods can reach an accuracy of 100% when diagnosing the data in the source test dataset, verifying that the feature extractor based on the improved 1-D convolutional neural network proposed can extract effective feature representations. The figures in Figures 9(b), 9(d), 9(f), 9(h), 9(j), and 9(l) show the results of diagnosis accuracies on the target test dataset. DAN erroneously classifies the 14 mils inner race fault into the normal state, indicating that the MMD-based method cannot distinguish these two types of data well. The classification accuracies of adversarial learning methods such as DANN and WDATL are significantly higher than other algorithms. Among them, WDATL mistakes one 21 mils ball fault sample to 7 mils ball fault and one 21 mils ball fault sample to 14 mils ball fault. Figure 9 demonstrates that the WDATL algorithm can achieve a highest diagnosis accuracy in the target domain among all the compared methods.

In order to understand the way that the 1-D convolutional neural network processes the raw time-series signals, the features extracted by the feature extractor are visualized by the t-SNE technology, as shown in Figure 10. The different colors in Figure 10 denote different health states of samples. Taking the transfer task 3 $\rightarrow$ 0 for example, figures in Figures 10(a), 10(c), 10(e), 10(g), 10(i), and 10(k) are learned by the source feature extractor on source training datasets, and all the six algorithms can distinguish the features well. For comparison, figures in Figures 10(b), 10(d), 10(f), 10(h), 10(j), and 10(l) are the features learned by the target feature extractor on target training datasets. CNN-based algorithms cannot distinguish category 2 from category 5 and category, and the distribution of features in each category is relatively scattered. Therefore, the CNN-based method could diagnose the source dataset well but perform poorly in target domain. The WDATL algorithm can divide

FIGURE 9: The confusion matrixes of test data in task 3 $\rightarrow$ 0: (a) CNN source; (b) CNN target; (c) DAN source; (d) DAN target; (e) JAN source; (f) JAN target; (g) CORAL source; (h) CORAL target; (i) DANN source; (j) DANN target; (k) WDATL source; (l) WDATL target.

10 categories into 10 nonoverlapping clusters, and each cluster is relatively far away. It is convenient for the classifier to classify them better.

4.3. Results and Analysis in SEU Dataset

4.3.1. Data Description. To further verify the performance of our proposed WDATL method, the aforementioned deep

models have been applied to the SEU dataset [50] under various working circumstances.

The SEU dataset provided by Southeast University includes a gear dataset and a bearing dataset. In our experiment, we use the dataset collected from channel 2. There are five fault models for each subdataset, one normal state and four failure states. Totally, there are nine health states in the SEU dataset. The description of the SEU dataset is depicted in Table 5.

FIGURE 10: Continued.

FIGURE 10: Visualization of features for CRWU dataset of task 3 $\rightarrow$ 0: (a) CNN source; (b) CNN target; (c) DAN source; (d) DAN target; (e) JAN source; (f) JAN target; (g) CORAL source; (h) CORAL target; (i) DANN source; (j) DANN target; (k) WDATL source; (l) WDATL target.

TABLE 5: The description of the SEU dataset.

Label	0	1	2	3	4	5	6	7	8
Location	Gear/bearing	Bearing	Bearing	Bearing	Bearing	Gear	Gear	Gear	Gear
Type	Health	Ball	Outer	Inner	Combination	Chipped	Miss	Surface	Root

There are only two kinds of working circumstances denoted as 0 and 1 in the SEU dataset; therefore, there exist two transfer tasks $0 \rightarrow 1$ and $1 \rightarrow 0$.

We use the same data preprocessing method in Section 4.2.1 to process the data in the SEU dataset.

4.3.2. Comparative Analysis. We still use the diagnosis accuracy to compare the performance of different methods in transfer tasks. The diagnosis accuracies under different transfer tasks using various algorithms are listed in Table 6. The accuracies of the CNN-based method for the two transfer tasks are relatively low (average 33.65%), indicating that the distribution discrepancies between source and target domain are large caused by different working speed and load. The DAN, JAN, CORAL, DANN, and WDATL could improve the diagnosis accuracy by using different transfer strategies, among which the WDATL perform best, increasing from 33.65% to 76.17% in the two transfer tasks averagely.

TABLE 6: Accuracy (%) for different methods in the SEU dataset.

	$0 \rightarrow 1$	$1 \rightarrow 0$	Average
CNN	44.72	22.58	33.65
DAN	60.26	43.55	51.91
JAN	59.97	57.04	58.51
CORAL	48.24	38.56	43.40
DANN	56.74	53.93	55.34
WDATL	73.59	78.75	76.17

4.3.3. Visualization Analysis. The features extracted by the feature extractor are also visualized by the technology of t-SNE, which maps the high-dimensional representation space into a 2-D space for clear visualization. The features of source and target training datasets in task $1 \rightarrow 0$ are shown in Figure 11.

The CNN, DAN, and WDATL could align the feature representations of source training data to reduce distribution

FIGURE 11: Visualization of features for the SEU dataset of task 1 $\rightarrow$ 0: (a) CNN source; (b) CNN target; (c) DAN source; (d) DAN target; (e) WDATL source; (f) WDATL target.

discrepancies. However, the CNN could not separate the target features well, which explains why the CNN model performs poorly in diagnosing the target training data. There is a bit of improvement in diagnosing target training data using the DAN method with MMD-based domain adaption. The proposed WDATL method could separate the target training data well, which demonstrates that the WDATL can learn the best features with discrimination and promote the diagnosis accuracy significantly.

5. Conclusion

In this article, a novel WGAN-GP-based deep adversarial transfer learning (WDATL) model is proposed to decrease distribution discrepancy between source and target domains

resulting from different working conditions in the fault diagnosis of aerospace. Firstly, WDATL pretrains in the source domain with sufficient labeled data to initialize the parameters of feature extractor and fault model classifier. Then, the domain critic tries to learn domain-invariant and fault model discriminative feature representations by minimizing the Wasserstein distance between the source and target feature distributions through adversarial training. Lastly, the fault model classifier could be generalized to target domain to diagnose faults with unlabeled or insufficient labeled data. Extensive experiments conducted on two open datasets demonstrate that the proposed WDATL algorithm shows a better diagnosis performance than those of other baseline approaches on transfer diagnosis tasks under diverse working conditions. In the future, we will concentrate on the

more challenging and practical transfer scenario, i.e., single source and multiple targets in the fault diagnosis of aerospace.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of 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.

Acknowledgments

The study was supported by the National Defense Basic Scientific Research under Grant No. JCKY2016203A003 and the Equipment Pre-research Key Laboratory Funds under Grant No. 61425010102.

References

- [1] Y. G. Lei, B. Yang, X. W. Jiang, F. Jia, N. Li, and A. K. Nandi, "Applications of machine learning to machine fault diagnosis: a review and roadmap," *Mechanical Systems and Signal Processing*, vol. 138, article 106587, 2020.
- [2] C. C. Che, H. W. Wang, Q. Fu, and X. M. Ni, "Combining multiple deep learning algorithms for prognostic and health management of aircraft," *Aerospace Science and Technology*, vol. 94, article 105423, 2019.
- [3] P. Liang, C. Deng, J. Wu, and Z. Yang, "Intelligent fault diagnosis of rotating machinery via wavelet transform, generative adversarial nets and convolutional neural network," *Measurement*, vol. 159, article 107768, 2020.
- [4] X. Guo, M. Tian, Q. Li, C. K. Ahn, and Y. H. Yang, "Multiple-fault diagnosis for spacecraft attitude control systems using RBFNN- based observers," *Aerospace Science and Technology*, vol. 106, article 106195, 2020.
- [5] V. Suo, B. Zhu, R. An, H. Sun, S. Xu, and Z. Yu, "Data-driven fault diagnosis of satellite power system using fuzzy Bayes risk and SVM," *Aerospace Science and Technology*, vol. 84, pp. 1092–1105, 2019.
- [6] W. Fan, Q. Zhou, J. Li, and Z. Zhu, "A wavelet-based statistical approach for monitoring and diagnosis of compound faults with application to rolling bearings," *IEEE Transactions on Automation Science and Engineering*, vol. 15, no. 4, pp. 1563–1572, 2018.
- [7] R. N. Toma, A. E. Prosvirin, and J. M. Kim, "Bearing fault diagnosis of induction motors using a genetic algorithm and machine learning classifiers," *Sensors*, vol. 20, no. 7, p. 1884, 2020.
- [8] J. Y. Jiao, M. Zhao, J. Lin, and K. X. Liang, "A comprehensive review on convolutional neural network in machine fault diagnosis," *Neurocomputing*, vol. 417, pp. 36–63, 2020.
- [9] P. Tamilselvan and P. F. Wang, "Failure diagnosis using deep belief learning based health state classification," *Reliability Engineering and System Safety*, vol. 115, pp. 124–135, 2013.
- [10] X. S. Lin, B. W. Li, and X. Y. Yang, "Engine components fault diagnosis using an improved method of deep belief networks," in *7th International Conference on Mechanical and Aerospace Engineering (ICMAE)*, pp. 454–459, London, UK, 2016.
- [11] M. Ma, C. Sun, and X. Chen, "Deep coupling autoencoder for fault diagnosis with multimodal sensory data," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 3, pp. 1137–1145, 2018.
- [12] X. Lin, B. Li, X. Yang, and J. Wang, "Fault diagnosis of aero-engine bearing using a stacked auto-encoder network," in *IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC)*, pp. 545–548, Chongqing, China, 2018.
- [13] K. B. Lee, S. Cheon, and C. Kim, "A convolutional neural network for fault classification and diagnosis in semiconductor manufacturing processes," *IEEE Transactions on Semiconductors Manufacturing*, vol. 30, no. 2, pp. 135–142, 2017.
- [14] F. Jia, Y. Lei, N. Lu, and S. Xing, "Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization," *Mechanical Systems and Signal Processing*, vol. 110, pp. 349–367, 2018.
- [15] S. Guo, T. Yang, W. Gao, and C. Zhang, "A novel fault diagnosis method for rotating machinery based on a convolutional neural network," *Sensors*, vol. 18, no. 5, p. 1429, 2018.
- [16] H. Shi, S. Hu, and J. Zhang, "LSTM based prediction algorithm and abnormal change detection for temperature in aerospace gyroscope shell," *International Journal of Intelligent Computing and Cybernetics*, vol. 12, no. 2, pp. 274–291, 2019.
- [17] M. Yuan, Y. Wu, and L. Lin, "Fault diagnosis and remaining useful life estimation of aero engine using lstm neural network," in *2016 IEEE International Conference on Aircraft Utility Systems (AUS)*, pp. 135–140, Beijing, China, 2016.
- [18] D. T. Hoang and H. J. Kang, "A survey on deep learning based bearing fault diagnosis," *Neurocomputing*, vol. 335, pp. 327–335, 2019.
- [19] D. Ezzat, A. E. Hassanien, A. Darwish, M. Yahia, A. Ahmed, and S. Abdelghafar, "Multi-objective hybrid artificial intelligence approach for fault diagnosis of aerospace systems," *IEEE Access*, vol. 9, pp. 41717–41730, 2021.
- [20] S. Pan and Q. Yang, "A survey on transfer learning," *Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.
- [21] C. Li, S. Zhang, Y. Qin, and E. Estupinan, "A systematic review of deep transfer learning for machinery fault diagnosis," *Neurocomputing*, vol. 407, no. 24, pp. 121–135, 2020.
- [22] F. Li, T. Tang, B. Tang, and Q. He, "Deep convolution domain-adversarial transfer learning for fault diagnosis of rolling bearings," *Measurement*, vol. 169, article 108339, 2021.
- [23] X. Li, Y. Hu, M. Li, and J. Zheng, "Fault diagnostics between different type of components: a transfer learning approach," *Applied Soft Computing Journal*, vol. 86, article 105950, 2020.
- [24] W. Zhang, X. Li, H. Ma, Z. Luo, and X. Li, "Open-set domain adaptation in machinery fault diagnostics using instance-level weighted adversarial learning," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 11, pp. 7445–7455, 2021.
- [25] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. Lawrence, *Dataset Shift in Machine Learning [M]*, The MIT Press, 2009.
- [26] L. Mihalkova, T. Huynh, and R. Mooney, *Mapping and revisiting markov logic networks for transfer learning. AAAI. Volume 7*, Vancouver, Canada, 2007.
- [27] L. Wen, L. Gao, and X. Li, "A new deep transfer learning based on sparse auto-encoder for fault diagnosis," *IEEE Transactions*

- on Systems, Man, and Cybernetics: Systems, vol. 49, no. 1, pp. 136–144, 2019.
- [28] J. Zhu, N. Chen, and C. Shen, “A new multiple source domain adaptation fault diagnosis method between different rotating machines,” *IEEE Transactions on Industrial Informatics*, vol. 17, no. 7, pp. 4788–4797, 2021.
 - [29] G. Matasci, M. Volpi, M. Kanevski, L. Bruzzone, and D. Tuia, “Semisupervised transfer component analysis for domain adaptation in remote sensing image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 7, pp. 3550–3564, 2015.
 - [30] S. Schwendemann, Z. Amjad, and A. Sikora, “Bearing fault diagnosis with intermediate domain based layered maximum mean discrepancy: a new transfer learning approach,” *Engineering Applications of Artificial Intelligence*, vol. 105, article 104415, 2021.
 - [31] M. Long, Y. Cao, J. Wang, and M. Jordan, “Learning transferable features with deep adaptation networks,” in *International conference on machine learning*, pp. 97–105, Lille, France, 2015.
 - [32] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, “Transfer feature learning with joint distribution adaptation,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2200–2207, Sydney, NSW, Australia, 2013.
 - [33] W. Zhang, X. Li, H. Ma, Z. Luo, and X. Li, “Universal domain adaptation in fault diagnostics with hybrid weighted deep adversarial learning,” *IEEE Transactions on Industrial Informatics*, vol. 17, no. 12, pp. 7957–7967, 2021.
 - [34] X. Li, W. Zhang, Q. Ding, and X. Li, “Diagnosing rotating machines with weakly supervised data using deep transfer learning,” *IEEE Transactions on Industrial Informatics*, vol. 16, no. 3, pp. 1688–1697, 2020.
 - [35] C. Cheng, B. Zhou, G. Ma, D. Wu, and Y. Yuan, “Wasserstein distance based deep adversarial transfer learning for intelligent fault diagnosis with unlabeled or insufficient labeled data,” *Neurocomputing*, vol. 409, pp. 35–45, 2020.
 - [36] Z. Zhao, Q. Zhang, X. Yu et al., “Unsupervised deep transfer learning for intelligent fault diagnosis: an open source and comparative study,” <https://arxiv.org/abs/1912.12528>.
 - [37] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” 2016, <https://arxiv.org/abs/1511.06434v1>.
 - [38] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in *Proceedings of the 34th International Conference on Machine Learning, PMLR*, pp. 214–223, Sydney, NSW, Australia, 2017.
 - [39] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumonlin, and A. C. Courville, “Improved training of wasserstein GANs,” 2017, <https://arxiv.org/abs/1704.00028>.
 - [40] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, “Spectral normalization for generative adversarial networks,” 2018, <https://arxiv.org/abs/1802.05957>.
 - [41] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., “Generative adversarial nets,” *Advances in Neural Information Processing Systems*, vol. 27, pp. 2672–2680, 2014.
 - [42] W. Zhang, G. Peng, C. Li, Y. Chen, and Z. Zhang, “A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals,” *Sensors*, vol. 17, no. 2, p. 425, 2017.
 - [43] A. L. Maas, A. Y. Hannun, and A. Y. Ng, “Rectifier nonlinearities improve neural network acoustic models,” in *ICML Workshop on Deep Learning for Audio, Speech and Language Processing: volume 30*, Citeseer, 2013.
 - [44] D. Clevert, T. Unterthiner, and S. Hochreiter, “Fast and accurate deep network learning by exponential linear (elus),” 2016, <https://arxiv.org/abs/1511.07289>.
 - [45] D. Kingma and J. Ba, “Adam: a method for stochastic optimization,” 2015, <https://arxiv.org/abs/1412.6980>.
 - [46] B. Sun and K. Saenko, “Deep coral: correlation alignment for deep domain adaption,” in *European Conference on Computer Vision*, Springer, Cham, 2016.
 - [47] Y. Ganin, E. Ustinova, H. Ajakan et al., “Domain-adversarial training of neural networks,” *Journal of Machine Learning Research*, vol. 17, pp. 1–35, 2016.
 - [48] W. A. Smith and R. B. Randall, “Rolling element bearing diagnostics using the Case Western Reserve University data: a benchmark study,” *Mechanical Systems and Signal Processing*, vol. 64–65, pp. 100–131, 2015.
 - [49] L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, article 25792605, 2008.
 - [50] T. Jin, C. Yan, C. Chen, Z. Yang, H. Tian, and S. Wang, “Light neural network with fewer parameters based on CNN for fault diagnosis of rotating machinery,” *Measurement*, vol. 181, article 109639, 2021.