Large Model for Rotating Machine Fault Diagnosis Based on a Dense Connection Network With Depthwise Separable Convolution

Yi Qin, *Senior Member, IEEE*, Taisheng Zhang, Quan Qian, and Yongfang Mao, *Member, IEEE*

***Abstract*— Most of the existing intelligent fault diagnosis models are suitable for only a type of rotating machine or equipment. To achieve the intelligent fault diagnosis for various rotating machines, it is significant to construct a diagnostic model with a powerful generalization ability. Thereupon, this work explores a large fault diagnosis model for a variety of rotary machines. To process the big data from a number of rotating machines and mine their fault characteristics effectively, a dense connection network with depthwise separable convolution (DCNDSC) is proposed as the large model. In this network, a dense connection with depthwise separable convolution block (DCDSCB) is designed for representing the complex vibration data and suppressing the over-fitting, and then a series of DCDSCBs are stacked so that DCNDSC can well extract various complicated characteristics caused by different faults and working conditions. A large rotating machine dataset including almost all public rotating machine data and our private data are built to train the large model. For enhancing the diagnostic ability of large model on the new monitoring data, a diminutive network fine-tuning strategy is proposed, while the main feature extraction capability of the pretrained DCNDSC is preserved. Ten fault datasets are applied to verify the high accuracy and strong generalization ability of the developed large model. This model is not only effectively applied to the fault diagnosis of actual rotating machinery but also first provides a pretraining large model for the field of mechanical fault diagnosis. Codes of our work are released at: https://qinyi-team.github.io/2024/04/Dense-connection-network-with-depthwise-separable-convolution/.**

***Index Terms*— Dense connection, depthwise separable convolution, fault diagnosis, fine-tuning, large model.**

I. Introduction

ROTATING machinery plays an important role in various industrial equipment, such as machine tool, robot, automobile, wind turbine, [1], [2], [3]. Bearings and gears are two important types of rotating machines, whose operating states seriously affect the efficiency and reliability of the entire equipment. It is inevitable that there are hash-operating environment, complex operating conditions, and other factors during the operation of the equipment, which may lead to the failure of rotating machinery. In some cases, these faults of rotating machines will result in unimaginable economic losses and casualties. To guarantee the safe operation of the entire industrial equipment, it is important for implementing the fault diagnosis of rotating machinery. With the development of artificial intelligence, the intelligent fault diagnosis based on deep learning has become a hot spot and demonstrated a great potential in the fault diagnosis of actual engineering.

Among the various intelligent fault diagnosis methods, deep learning is most widely researched and obtains the better diagnostic results. The classical deep learning models include deep belief networks, autoencoders, deep neural networks, and convolutional neural networks, all of which achieve the machine fault classification based on the principle that the monitoring signals acquired under different faults have different distributions [4]. To improve the generalization ability of deep learning models, various transfer learning methods have been put forward. Wan et al. [5] designed a ring-based decentralized federated transfer learning for the fault diagnosis of rotating machinery. Yang et al. [6] improved a graph construction strategy to establish a mapping relationship between labels and nodes for the cross-domain rotating machine fault diagnosis. Shao et al. [7] constructed a transfer autoencoder through the particle swarm optimization and shared parameters, which could identify the machine vibration data with large distribution differences. Wang et al. [8] combined subdomain adaptation and domain adaptation to solve the problem of mismatch between class and domain. The aforementioned research has demonstrated the numerous results achieved by various deep neural networks.

Among the various transfer learning methods, fine-tuning is one of the commonly used technologies, which can effectively adapt the model to new tasks or fields [9]. Fine-tuning means the adjustment of the pretrained model's parameters, and it has been successfully applied to the field of fault diagnosis. Jiang et al. [10] proposed a wind turbine gearbox diagnosis method by adding different noises for AE training, which could capture deeper information and useful failure modes. Han et al. [11] constructed a new transfer-learning framework for diagnosing faults under unseen working conditions. These fine-tuning models do not have the adequate generalization ability for the data in a new task, especially for the target

Manuscript received 30 January 2024; revised 1 April 2024; accepted 15 April 2024. Date of publication 6 May 2024; date of current version 15 May 2024. This work was supported by the National Natural Science Foundation of China under Grant 52175075. The Associate Editor coordinating the review process was Dr. Arunava Naha. *(Corresponding author: Yongfang Mao.)*

Yi Qin, Taisheng Zhang, and Quan Qian are with the State Key Laboratory of Mechanical Transmission for Advanced Equipment, Chongqing University, Chongqing 400044, China (e-mail: qy_808@cqu.edu.cn; zts2022@stu.cqu.edu.cn; qian_1998@cqu.edu.cn).

Yongfang Mao is with the School of Automation, Chongqing University, Chongqing 400044, China (e-mail: yfm@cqu.edu.cn).

Digital Object Identifier 10.1109/TIM.2024.3396841

1557-9662 © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: CHONGQING UNIVERSITY. Downloaded on May 19,2024 at 16:58:17 UTC from IEEE Xplore. Restrictions apply.

data that have a big difference with the training data. This issue can be solved by enhancing the ability of pretrained model, that is, base model. As a result, large models have developed rapidly in recent years due to their powerful feature extraction capabilities. Brown et al. [12] expanded the scale of language models to improve the task-agnostic performance and few-shot learning ability. Driess et al. [13] combined the vision transformer (ViT) model with the Pathways Language Model (PaLM) to achieve the comprehensive understanding and generation of multimodal data. These models possess the powerful learning and representation capabilities, enabling them to accomplish various tasks in the related fields and demonstrate high generalization performance [14]. Moreover, they utilize the fine-tuning technology for enhancing their performance. However, these large models are mainly applied in the fields of natural language processing and computer vision. In the field of mechanical fault diagnosis, a large model has not yet been built; meantime, the existing fault diagnosis models address to tackle the specific tasks, without considering the generalization for a variety of rotating machines. Consequently, there is an urgent need to construct a large model for the fault diagnosis of rotating machinery.

The well-known GPT-4 and Gmini have the powerful inference ability since the massive multimodal data are used for training, whereas their backbones cannot be employed directly as the base model for rotating machine fault diagnosis. This is because the vibration data of rotating machines are limited compared to the data used for training GPT-4 and Gmini. Obviously, it is unreasonable to use a too large backbone for processing the relatively little rotating machinery vibration data [15]. Thus, it is of great value to study the basic model suitable for mechanical fault diagnosis. The rotating machinery vibration signals are very complex due to the varied load and speed, the complex mechanical structure, and the strong background noise. Aiming at the difficulty of fault feature extraction from the complex data, a dense connection network with depthwise separable convolution (DCNDSC) is proposed to adaptively learn various complex fault features. The proposed DCNDSC adopts the dense connection with strong memory ability, which can allow the complex data features to flowthrough the network while minimizing the information loss and improving the utilization of model parameters [16]. On the other hand, depthwise separable convolution can better capture the local sequence correlation of mechanical vibration signals and prevent overfitting, thereby improving the training accuracy of the model. The DCNDSC is first trained by a large number of rotating machine vibration samples, then a fine-tuning strategy is used to reoptimize the model parameters, further enhancing the diagnostic ability and generalization performance of pretrained base model. Finally, the effectiveness and advantage of the proposed large model are verified by ten diagnostic tasks for various rotating machines.

The following are the contributions of this work.

1) To enhance the representation ability of complex data, a DCNDSC is constructed as the base model. The DCNDSC is trained by the public rotating machine datasets collected as best we could and our own private rotating machine dataset; hence, the trained model possesses the ability to diagnose faults of various rotating machines. More importantly, it is the first large model in the field of rotating machine fault diagnosis.
2) A fine-tuning strategy is developed for mining more fault features from new data while maintaining the feature extraction ability for the pretraining dataset, which can continuously enhance the diagnostic ability and generalization performance of the proposed large model.
3) The proposed large model is validated by ten rotating machine datasets and compared with the typical large diagnostic networks. The experimental results indicate that the proposed large model has higher diagnostic accurately and stronger generalization ability. Meanwhile, it is successfully applied to the fault diagnosis of actual rotating machines.

II. Proposed Large Model for Rotating Machine Fault Diagnosis

In this section, the depthwise separable convolution and dense connection network are introduced, then the proposed DCNDSC model is elaborated. Additionally, a diminutive network fine-tuning strategy is proposed.

A. Depthwise Separable Convolution

The commonly used convolutional neural network consists of nonlinear activation functions, and multiple batch normalization (BN) layers and convolutional layers. The BN is expressed as

$$\mathrm{BN}_{\gamma,\boldsymbol{\beta}} = \gamma \widehat{\boldsymbol{X}} + \boldsymbol{\beta} \tag{1}$$

where $\widehat{\boldsymbol{X}}$ is the normalization result of input $\boldsymbol{X}$, γ denotes the scale, and $\boldsymbol{\beta}$ represents the bias parameter. After performing the convolution operation, BN and nonlinear activation on $\boldsymbol{X}$, the output $\boldsymbol{Y}_i$ is obtained as

$$\boldsymbol{Y}_i = \sigma(\mathrm{BN}_{\gamma,\boldsymbol{\beta}}(\boldsymbol{W}_i \boldsymbol{X} + \boldsymbol{\beta}_i)) \tag{2}$$

where $\boldsymbol{W}_i$ is the weight matrix of the convolution layer, $\boldsymbol{\beta}_i$ denotes the bias matrix, and σ is the nonlinear activation function.

In contrast to the traditional convolutional operation, 1-D depthwise separable convolution is based on the idea of grouping and is divided into two portions: hierarchical convolution and channel convolution [17], as shown in Fig. 1. The hierarchical convolution can mine the features from each input signal, while the channel convolution can fuse the features from different input signals and mine the features between channels effectively. As 1-D depthwise separable convolution allows the formation of convolutional sparse structures with multiple sets of convolution kernels corresponding to multiple groups of channels, it can better extract the useful characteristics from the complex data than the traditional convolutional operation. Suppose that K_s, K_n, C, respectively, denote the size of the convolution kernel, the number of kernels, and the number of channels. It then follows that the number of parameters in the 1-D depth-separable convolution is just $K_s \times C + K_n \times C$, while that in the traditional 1-D convolution is $K_s \times K_n \times C$. Obviously, the 1-D depth-separable convolution

Authorized licensed use limited to: CHONGQING UNIVERSITY. Downloaded on May 19,2024 at 16:58:17 UTC from IEEE Xplore. Restrictions apply.

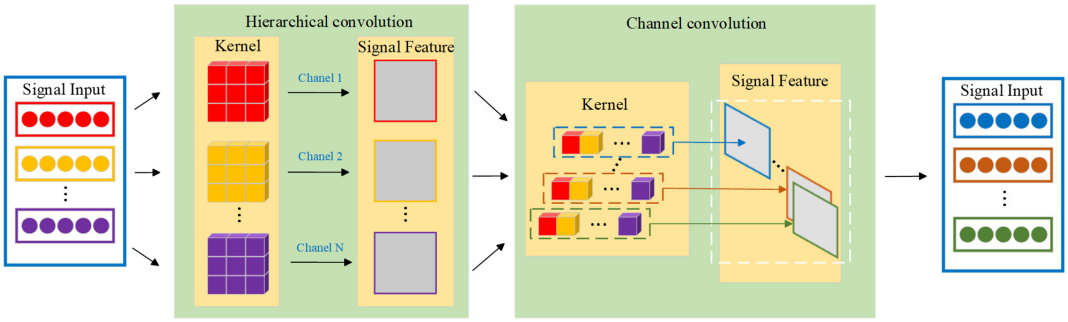

Fig. 1. 1-D depthwise separable convolution structure.

has much fewer parameters than the traditional 1-D convolution; thus, it can reduce the complexity of network and avoid overfitting, that is, it brings the regularization effect to the network and enhances the generalization ability. According to the above analysis, our method uses 1-D depthwise separable convolution instead of the traditional 1-D convolution.

B. Dense Connection Network

The dense connection means that each layer is directly connected to all subsequent layers [18], that is, the input of a layer is fed into all subsequent layers. Assume that $\boldsymbol{X_i}$ represents the input of the ith layer, and $\boldsymbol{A}_j$ denotes the output of the jth layer. $\boldsymbol{X_i}$ can be written as

$$\boldsymbol{X_i} = \text{combine}(\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_{i-1}) \tag{3}$$

where combine $(\cdot)$ means the combination of several outputs. Obviously, the dense connection promotes the information flow and feature propagation throughout the network; thus, it mitigates the problem of gradient vanishing and enhances the utilization rate of feature information; moreover, it deepens the network architecture.

With the dense connection, the dense connection network can be built, and it primarily consists of two sections: a dense block and a transition block. The basic structure of dense block is composed of two BN layers: two nonlinear activation layers, and two convolutional layers. The transition block includes an average pooling layer and a convolutional layer, which can reduce the dimensionality of features and decrease the amount of output data processed by the subsequent layers.

C. DCNDSC

Suppose that the vibration sample of rotating machinery and the corresponding label are respectively, denoted by x_i and y_i. Then, we define the labeled rotating machinery dataset as $\boldsymbol{X} = \{x_i, y_i\}_{i=1}^{n}$. This dataset is constructed by the public rotating machine datasets collected by us as far as possible and our own private rotating machine dataset, where there are totally nine fault categories. Most of these subsets have different operating conditions, sampling frequencies, fault categories, and severities, bringing great complexity and difficulty to the fault diagnosis. The commonly used deep networks, such as DenseNet, ResNet, cannot be suitable for the complex diagnostic task; therefore, a DCNDSC is proposed.

First, by stacking multiple Conv_blocks, a dense connection with depthwise separable convolution block (DCDSCB) is designed, and its structure diagram is depicted in Fig. 2. The Conv_block is composed of two 1×1 convolutions, and one 3×3 group convolution, as shown in Fig. 2. These Conv_blocks adopt the dense connection, that is, the input of each Conv_block includes the output of all previous Conv_blocks. Suppose that convb$(\cdot)$ denotes the operation of Conv_block, the output of the nth Conv_Block can be written as

$$\boldsymbol{L}_n = \sum_{i=1}^{n-1} \boldsymbol{L}_{i-1} + \text{convb}\left(\sum_{i=1}^{n-1} \boldsymbol{L}_{i-1}\right) \tag{4}$$

where $\boldsymbol{L}_i$ is the output of the ith Conv_block. Note also that these blocks use depthwise separable convolution to extract the fault features and capture the channel information. Due to the special structure and operation of DCDSCB, it has powerful feature representation ability for the complex machine vibration data and is beneficial to suppress the over-fitting.

Next, by considering the strong feature extraction ability of deep CNN [19], [20], [21], we design a quite deep network structure for DCNDSC. In addition, DCDSCB is used to construct DCNDSC. The architecture of DCNDSC is illustrated in Fig. 3. It can be seen that a Conv_block is first used to mine the shallow detail features in the vibration signals, then 130 DCDSCBs and three transition Conv_blocks are utilized for the extraction of deep semantic features in the vibration signals, finally a classifier is built to achieve the fault classification. The first Conv_block consists of a convolutional layer, a BN layer, ReLU activation function, and a max pooling layer, and it can be regarded as an initial extractor $\boldsymbol{Y}_{\text{fcb}}$. The extractor for mining the deep semantic feature is denoted by $\boldsymbol{Y}_{\text{dsf}}$, and the DCDSCB n is defined as $\boldsymbol{L}_n$. The transition Conv_block $\boldsymbol{Y}_{\text{tc}}$ comprises a BN layer, ReLU activation function, and an average pooling layer, which can achieve the feature dimension reduction. The classifier $\boldsymbol{Y}_c$ is composed of an adaptive pooling layer, a fully connected (FC) layer, and softmax function. According to Fig. 3, the calculation procedure $\boldsymbol{Y}_{\text{dcndsc}}$ of DCNDSC is formulated as

$$\boldsymbol{Y}_{\text{fcb}}(\boldsymbol{X}) = \text{Maxpooling}(\sigma(\text{BN}_{\gamma,\beta}(\boldsymbol{W}_{\text{fcb}}\boldsymbol{X} + \boldsymbol{b}_{\text{fcb}}))) \tag{5}$$

$$\boldsymbol{Y}_{\text{dsf}} = \boldsymbol{L}_{48}(\boldsymbol{Y}_{\text{tc}}(\boldsymbol{L}_{64}(\boldsymbol{Y}_{\text{tc}}(\boldsymbol{L}_{12}(\boldsymbol{Y}_{\text{tc}}(\boldsymbol{L}_6)))))) \tag{6}$$

$$\boldsymbol{Y}_{\text{tc}}(\boldsymbol{X}) = \text{Averagepooling}(\sigma(\text{BN}_{\gamma,\beta}(\boldsymbol{W}_{\text{tc}}\boldsymbol{X} + \boldsymbol{b}_{\text{tc}}))) \tag{7}$$

$$\boldsymbol{Y}_c(\boldsymbol{X}) = \text{softmax}(\text{FC}(\text{AdaptivePooling}(\boldsymbol{X}))) \tag{8}$$

$$\boldsymbol{Y}_{\text{dcndsc}}(\boldsymbol{X}) = \boldsymbol{Y}_c(\boldsymbol{Y}_{\text{dsf}}(\boldsymbol{Y}_{\text{fcb}}(\boldsymbol{X}))) \tag{9}$$

Authorized licensed use limited to: CHONGQING UNIVERSITY. Downloaded on May 19,2024 at 16:58:17 UTC from IEEE Xplore. Restrictions apply.

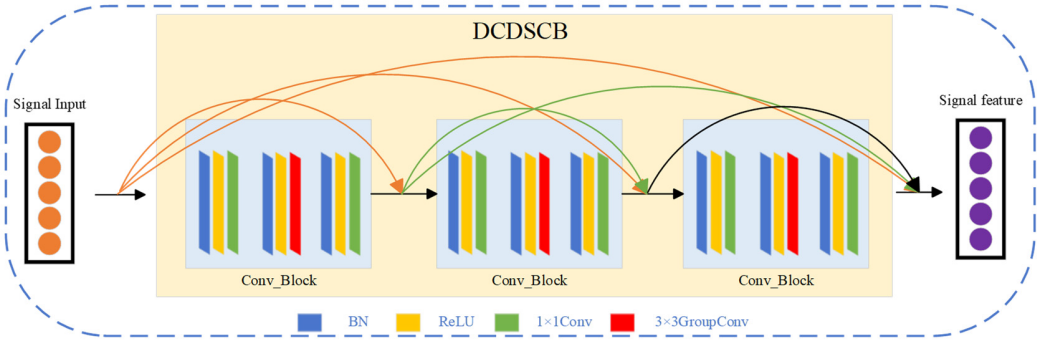

Fig. 2. Structure of DCDSCB.

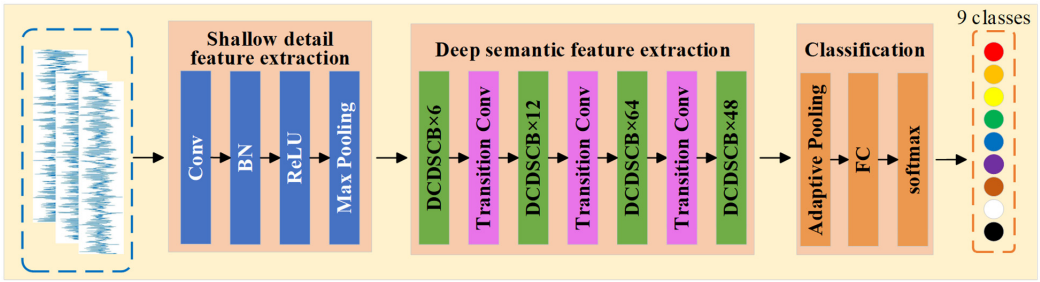

Fig. 3. Architecture of DCNDSC.

where $\boldsymbol{W}_{\text{fcb}}$ and $\boldsymbol{b}_{\text{fcb}}$ are, respectively, the weight and bias parameters of $\boldsymbol{Y}_{\text{fcb}}$, $\boldsymbol{W}_{\text{tc}}$, and $\boldsymbol{b}_{\text{tc}}$ are, respectively, the weight and bias parameters of $\boldsymbol{Y}_{\text{tc}}$; x denotes the input data.

The specific constitution of DCNDSC is shown in Table I. From this table and the foregoing analysis, we can know that the proposed DCNDSC can efficiently represent various complex characteristics caused by different faults and working conditions in a large number of samples, thereby improving the generalization ability and accuracy of fault diagnosis model.

D. Fine-Tuning Strategy for the Optimization of Diagnostic Model

Fine-tuning is a key step in the construction of large classification model [22]. The common fine-tuning method is to freeze the feature extraction layers and unfreeze the classification layer for adapting to new tasks. However, the research shows that only when the distributions of different samples are very similar, this method will have a certain effect [11], namely, its generalization ability is limited. Therefore, we propose a diminutive network fine-tuning strategy. Its implementation scheme is illustrated in Fig. 4. It can be seen that the parameters of the shallow detail feature extractor and the first 127 DCDSCBs in the deep semantic feature extractor are frozen, while the parameters of the last three DCDSCBs and classifier are unfrozen.

Evidently, the proposed fine-tuning strategy avoids to learn the repetitive knowledge in the pretrained model, meantime, focusing on mining some important characteristics of the new fault samples. After the fine-tuning, the trained DCNDSC can not only possess the ability to extract general fault features but also obtain the capability to extract the discriminative features from the new data; therefore, this strategy effectively improves the generalization ability of DCNDSC.

TABLE I
SPECIFIC CONSTITUTION OF DCNDSC

Network	Number of block/ number of channel	Input size	Output size
First Conv block	1/64	(3072,1)	(768,64)
DCDSCB×6	6/256	(768,64)	(768,256)
Transition block1	1/128	(768,256)	(384,128)
DCDSCB×12	12/512	(384,128)	(384,512)
Transition block2	1/256	(384,512)	(192,256)
DCDSCB×64	64/2304	(192,256)	(192,2304)
Transition block3	1/1152	(192,2304)	(96,1152)
DCDSCB×48	48/2688	(96,1152)	(96,2688)
Adaptive pooling	1/2688	(96,2688)	(1,2688)
Flatten	/	(1,2688)	2688
FC	1/9	2688	9

III. EXPERIMENT AND DISCUSSION

To demonstrate the robust generalization capability of the proposed DCNDSC model, this section conducts several diagnostic experiments on ten datasets. The fault diagnosis experiments on the target testing sets and all the testing sets are performed.

A. Data Description

The large model for rotating machine fault diagnosis requires a great deal of data to train. At present, there is not a large-scale dataset for the rotating machine fault diagnosis.

Authorized licensed use limited to: CHONGQING UNIVERSITY. Downloaded on May 19,2024 at 16:58:17 UTC from IEEE Xplore. Restrictions apply.

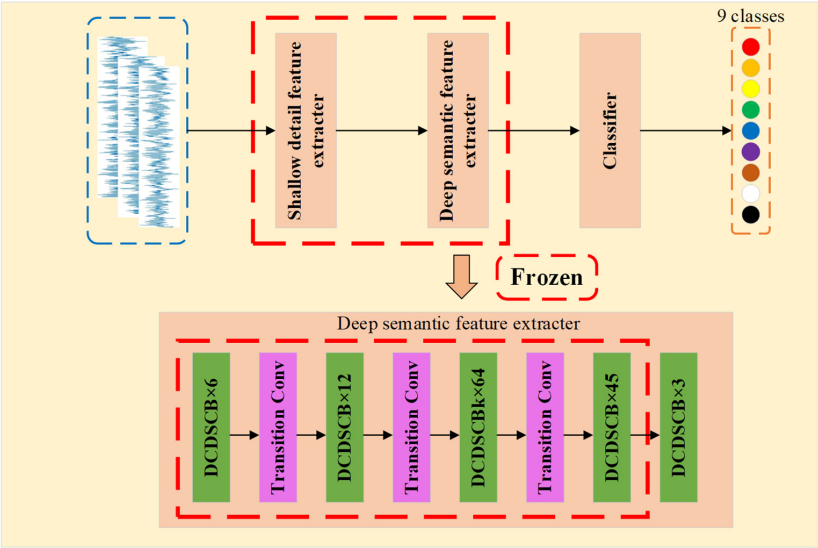

Fig. 4. Diminutive network fine-tuning strategy.

Thereupon, we collect the public rotating machine datasets as far as possible. These datasets and our own private rotating machine dataset constitute the large-scale dataset. They are Case Western Reserve University (CWRU), Data Castle (DC), Drivetrain Diagnostics Simulator (DDS), Planetary Gearbox Test Rig (PGTR), Industrial Innovation Platform (IIP), Intelligent Maintenance Systems (IMS), Iowa State University (ISU), Jiang Nan University (JNU), Maintenance Fault Prognostics Testing (MFPT), Northeast Electric Power University (NEPU), Paderborn University (PU), Politecnico Di Torino University (PDTU), Rotor Dynamics Simulator (RDS), SouthEast University (SEU), Southwest JiaoTong University (SWJTU), University Of Connectic (UOC), Vibration, Acoustic, Temperature, Current Dataset (VATCD), Wind Turbine Gearbox (WTG), XJTU_Gearbox, and XJTU_SY. Among them, the data from UOC utilized the time synchronous averaging (TSA) technique, resulting in both time domain and angular frequency domain data. For this experiment, the time-domain data were employed. A total of 20 subdatasets are involved in the training dataset, while ten subdatasets are involved in the test. The nine fault categories normal condition (NC), bearing inner race fault (IF), bearing ball fault (BF), bearing outer race fault (OF), ring gear fault (RF), output shaft gear fault (OSF), planetary gear fault (PF), sun gear fault (SF), and input shaft gear fault (ISF) are denoted as $y_{tr} \in \{NC, \text{IF}, \text{BF}, \text{OF}, \text{RF}, \text{OSF}, \text{PF}, \text{SF}, \text{ISF}\}$.

Our own dataset is a WTG dataset collected from the six actual wind turbine gearboxes. The structure of wind turbine gearbox and vibration signal acquisition system are shown in Fig. 5. The wind turbine gearbox is composed of two planetary-stage and parallel-stages. There are two gear faults occurring on the secondary ring and the high-speed shaft, respectively. Two acceleration sensors were placed on the cases of the secondary ring and the high-speed shaft for collecting the fault vibration signal, and the sampling frequencies are 12.8 and 25.6 kHz, respectively. The details of all used datasets are listed in Table II. In this table, taking 16–30 Hz as an example, it means that the rotational frequency continuously varies from 16 to 30 Hz.

To obtain more samples, we use the sliding sampling technology, and the length of a sample is set to 3072 for obtaining the sufficient fault characteristic information. According to different working conditions, all the samples are divided into two types of training sets and two types of testing sets. The number of total (unprocessed) samples in the pretraining phase and the fine-tuning phase are 328 606 and 115 558, respectively. The samples acquired under some working conditions are used for pretraining DCNDSC, whose number is 303 576. On the other hand, the samples acquired under other working conditions in some datasets are used for fine-tuning, whose number is 90 528. The samples used for the target testing have the same working conditions with those for fine-tuning, and their number is 4970. Meantime, the test samples acquired under all the working conditions is employed to check the comprehensive performance of the trained large model, and their number is 20 060. The sizes of the above sample sets in the two phases are listed in Table III.

In addition, for reducing the distribution discrepancy in different datasets, it is necessary to normalize the original signals. Assume that $(x_1, x_2, \ldots, x_n)$ represent the original input samples, and $\{f(x_1), f(x_2), \ldots, f(x_n)\}$ denote the normalized results. This work employs the max-min normalization method, which is formulated as

$$f(x_i) = \frac{x_i - \min_{1\le i\le n}(x_i)}{\max_{1\le i\le n}(x_i) - \min_{1\le i\le n}(x_i)}. \tag{10}$$

B. Description of Fault Diagnosis Experiment

The training of the large model can be divided into two stages: pretraining and fine-tuning, and the training procedure is illustrated in Fig. 6. In this process, the goal of pretraining stage is to make the backbone obtain a powerful feature extraction ability, and fine-tuning is used to enhance the generalization ability of model. We use the Adam optimizer [41] for

Authorized licensed use limited to: CHONGQING UNIVERSITY. Downloaded on May 19,2024 at 16:58:17 UTC from IEEE Xplore. Restrictions apply.

TABLE II
DETAILS OF ALL USED DATASETS

Dataset	Dataset division based on operating condition				Fault type	Rotational speed	Sampling frequency
	Pre-training set	Fine-tuning set	Target testing set	All testing set			
CWRU[23]	0, 1, 3HP	2HP	2HP	All working conditions	NC, IF BF, OF	1720-1797rpm	12kHz, 48kHz
DC[24]	/	/	/	-	NC, IF BF, OF	/	/
DDS[25]	0, 1.4, 25.2Nm	2.8Nm	2.8Nm	-	NC, SF	1500rpm	5.1kHz
PTGR[26]	0, 5Nm	/	/	-	NC, RF PF, SF	600, 1200, 2400rpm	16.3kHz
IIP[27]	1800, 2100, 2400, 3000rpm	/	/	-	NC, IF BF, OF	/	/
IMS[28]	6000 lbs	/	/	-	NC, IF, OF	2000rpm	20kHz
ISU-MSF[29]	16-30, 15.5, 17.5-29.5Hz	16.5Hz	16.5Hz	-	NC, IF, OF	/	12.8kHz
JNU[30]	600, 800, 1000rpm	/	/	-	NC, IF, BF, OF	/	50kHz
MFPT[31]	0, 25, 50, 100, 150, 200, 300lbs	250lbs	250lbs	-	NC, IF, OF	1500rpm	48.8kHz, 97.7kHz
NEPU[32]	0, 0.2, 0.3Nm	0.1Nm	0.1Nm	-	NC, IF, BF, OF	1443-1487rpm	12kHz
PU[33]	0.7Nm,1000N; 0.1Nm,1000N	0.7Nm, 400N	0.7Nm, 400N	-	NC, IF, OF	900-3000rpm	64kHz
PDTU[34]	0, 1000, 1400, 1800N 200, 300, 400, 500Hz	1000N, 100Hz	1000N, 100Hz	-	NC, IF, BF	/	51.2kHz
RDS[35]	0, 2, 3kN	1kN	1kN	-	NC, IF, BF, OF	1000, 2000, 3000rpm	8kHz
SEU[36]	0, 2V	/	/	-	NC, SF	1200, 1800rpm	/
SWJTU[35]	1, 2, 3kN	/	/	-	NC, IF, BF, OF	896rpm	10kHz
UOCU[37]	/	/	/	-	NC, ISF	/	20kHz
VATCD[38]	0, 4Nm	2Nm	2Nm	-	NC, IF, OF	680-2460rpm	25.6kHz
WTG	500-1800rpm	650-1300rpm	650-1300rpm	-	NC, RF, OSF	/	12.8kHz, 25.6kHz
XJTU_Gearbox [39]	/	/	/	-	NC, PF	1800rpm	20.5kHz
XJTU_SY[40]	10, 11, 12kN	/	/	-	NC, IF, OF	2100, 2250, 2400rpm	25.6kHz

training. The target testing set is employed for verifying the effect of fine-tuning, while all the testing sets are applied to validate the overall performance of the final model. To make a fair comparison, seven deep large network models, including DCNDSC(ours), ResNet-152 [42], VGG-19 [43], and four dense connection networks (DenseNet-264, DenseNet-201, DenseNet-169, DenseNet-121) [44] are compared. For a fair comparison, their hyperparameter configurations are identical, and their sizes are 26.5, 9.7, 24.1, 27.5, 15.7, 10.5, and 5.5 M, respectively. Note that the size of the proposed model is slightly smaller than that of DenseNet-264. This is because the depthwise separable convolution is used in the proposed model. In the experiment, all fault diagnosis models were repeated three times for testing their diagnostic capabilities. In the fine-tuning stage, the last three convolutional layers and FC layers of the comparative models are fine-tuned.

The computing hardware is composed of Intel E5-2687W CPU, 32 GB memory, and 4090 GPU, and Pytorch is used as

Authorized licensed use limited to: CHONGQING UNIVERSITY. Downloaded on May 19,2024 at 16:58:17 UTC from IEEE Xplore. Restrictions apply.

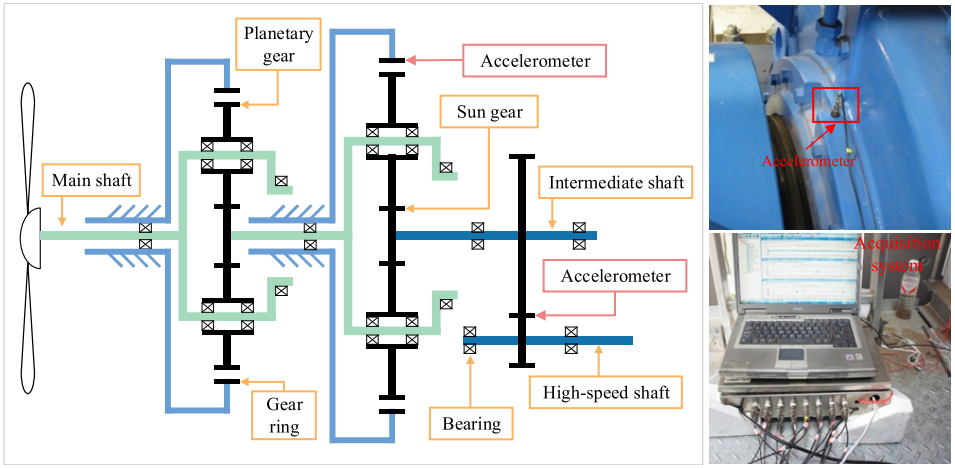

Fig. 5. Wind turbine gearbox acquisition equipment and gearbox structure.

TABLE III
SIZES OF VARIOUS SAMPLE SETS

Stage	Total	Training	Target testing	All testing
Pre-training	328606	303576	4970	20060
Fine-tuning	115558	90528	4970	20060

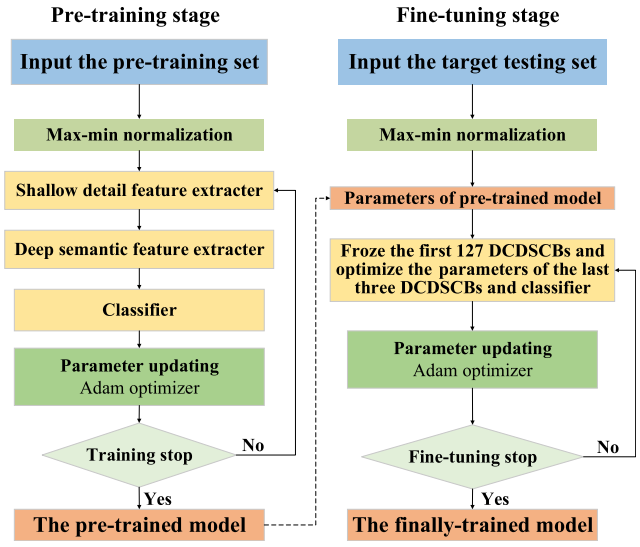

Fig. 6. Procedure of pretraining and fine-tuning.

programming software. The batch size is set to 64. The number of pretraining epochs is 50, while the number of fine-tuning epochs is 20.

C. Experiment on the Target Testing Set

Except for the accuracy of fault classification, Micro-F1 is used to evaluate the diagnostic results. To define it, Precision and Recall are first defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{11}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

where TP is the number of true positive samples, FP denotes the number of false positive samples, and FN is the number of false negative samples. As these two indexes are contradictory, Micro-F1 is employed based on them for comprehensively evaluating the performance of fault diagnosis model, and it is given by

$$\text{Micro-F1} = \frac{\sum_{i=1}^{n} \frac{2\text{TP}_i}{2\text{TP}_i + \text{FN}_i + \text{FP}_i}}{n} \tag{13}$$

where n is the number of classes.

With the target testing set, the pretrained DCNDSC and the finally trained DCNDSC are validated, respectively, and DenseNet-264, DenseNet-201, DenseNet-169, DenseNet-121, ResNet-152, and VGG-19 are used for comparison. The diagnostic accuracies and Micro-F1s obtained by the pretrained DCNDSC and contrast models are, respectively, shown in Tables IV and V. In these tables, the former value denotes the mean diagnosis accuracy, while the latter value means the standard deviation of diagnostic results. It can be seen from these two tables that the pretrained DCNDSC and Micro-F1 has higher accuracy than other pretrained large-scale networks for most of the datasets. Especially, the pretrained DCNDSC and Micro-F1 has the highest average accuracy. We can also conclude that the proposed DCNDSC has a stronger generalization ability for various rotating machines with different operating conditions than the current advanced large-scale models.

Next, to verify the necessity of fine-tuning, the diagnostic accuracies and Micro-F1s obtained by the finally trained DCNDSC and contrast models are shown in Tables VI and VII. Comparing Tables IV and V with Tables VI and VII, a noticeable improvement in accuracy and Micro-F1 is achieved by fine-tuning in most datasets. For instance, the accuracy for the RDS dataset increases from 94.40% to 99.73% after fine-tuning, and the Micro-F1 achieves a rise of 4.94%. Note in particular that there is a substantial improvement in both accuracy and Micro-F1 for the WTG dataset. Namely, the accuracy increases from 55.89% to 91.11%, and the improvement of

Authorized licensed use limited to: CHONGQING UNIVERSITY. Downloaded on May 19,2024 at 16:58:17 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
ACCURACIES OBTAINED BY VARIOUS PRETRAINED MODELS FOR THE TARGET TESTING SET

	DCNDSC	DenseNet-264	DenseNet-201	DenseNet-169	DenseNet-121	ResNet-152	VGG-19
CWRU	92.40±0.28	92.93±0.84	92.73±1.04	93.06±1.11	**94.20±1.40**	91.20±1.18	91.40±1.61
DDS	98.00±1.63	98.67±0.94	98.00±1.63	**100.00±0.00**	97.33±0.94	96.00±3.27	96.47±2.71
ISU	99.60±0.57	**100.00±0.00**	98.92±1.52	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	98.38±1.32
MFPT	**99.19±0.33**	95.55±1.29	94.55±2.14	96.70±0.76	93.67±3.54	94.88±1.01	99.12±1.24
NEPU	100.00±0.00	99.93±0.09	99.87±0.09	99.93±0.09	**100.00±0.00**	**100.00±0.00**	99.93±0.09
PU	**98.11±0.81**	96.30±0.25	96.97±0.66	95.21±0.19	92.32±0.49	95.42±0.94	87.47±4.39
PDTU	**91.04±0.50**	91.04±0.50	88.82±2.38	91.58±0.81	84.78±0.19	84.18±3.06	71.04±12.15
RDS	**94.40±0.16**	87.87±1.05	89.13±0.34	91.06±3.40	82.33±1.72	79.67±4.64	78.13±6.61
VATCD	**100.00±0.00**	100.00±0.00	**100.00±0.00**	99.93±0.10	**100.00±0.00**	99.93±0.10	80.40±7.14
WTG	55.89±2.04	55.01±0.83	54.61±4.10	54.48±2.63	53.13±2.16	**59.46±4.14**	54.75±3.88
Average	**92.86±12.71**	91.73±12.86	91.36±12.86	92.20±13.02	89.78±13.55	90.07±12.09	85.71±13.98

TABLE V
MICRO-F1S OBTAINED BY VARIOUS PRETRAINED MODELS FOR THE TARGET TESTING SET

	DCNDSC	DenseNet-264	DenseNet-201	DenseNet-169	DenseNet-121	ResNet-152	VGG-19
CWRU	93.48±0.41	94.02±0.54	94.08±0.63	93.78±1.01	**95.20±1.22**	91.91±1.01	92.66±1.68
DDS	**98.06±1.57**	98.69±0.92	98.06±1.57	**100.00±0.00**	97.41±0.89	96.23±3.04	96.60±2.61
ISU	99.60±0.56	**100.00±0.00**	99.45±0.95	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	98.44±1.26
MFPT	**99.43±0.19**	95.85±1.24	94.89±1.90	96.86±0.84	94.21±4.02	95.21±0.92	99.13±1.23
NEPU	100.00±0.00	99.93±0.09	99.87±0.09	99.93±0.09	**100.00±0.00**	**100.00±0.00**	99.93±0.09
PU	**98.42±0.75**	96.46±0.31	97.16±0.68	94.93±0.95	91.50±0.12	95.63±1.42	88.27±5.64
PDTU	**92.16±0.43**	91.42±0.55	89.64±2.32	91.90±0.87	85.29±0.46	84.79±2.74	73.40±12.18
RDS	**94.79±0.11**	87.89±0.90	89.86±0.10	91.60±3.01	83.91±1.05	79.49±4.92	79.81±5.91
VATCD	**100.00±0.00**	100.00±0.00	**100.00±0.00**	99.97±0.05	**100.00±0.00**	99.97±0.05	87.11±4.59
WTG	51.58±11.74	45.23±0.95	45.21±5.00	44.91±2.61	**55.29±11.75**	55.20±8.54	58.75±5.19
Average	**92.75±13.99**	90.95±15.71	90.82±15.63	91.39±15.82	90.28±12.90	89.84±13.23	87.41±12.67

Micro-F1 is 39.76%. Unfortunately, the diagnostic accuracy and Micro-F1 only for the PU dataset decrease. In general, via fine-tuning, the average accuracy increases from 92.86% to 97.54%, while the average Micro-F1 increases from 92.75% to 97.73%. This indicates that the efficacy of the proposed diminutive network fine-tuning strategy can further enhance the performance of the pretrained model. The comparative results in Tables VI and VII further show that the finally trained DCNDSC possesses a higher performance in terms of diagnostic accuracy and Micro-F1 compared with other existing finally trained networks, and it has a more powerful generalization ability.

D. Experiment on All the Testing Sets

In order to validate the overall performance of the proposed DCNDSC on all data, all the testing sets are used in this experiment. Similarly, seven diagnostic models are compared. First, the pretrained models are used for implementing the fault diagnosis of samples acquired under all the working conditions, and the obtained diagnostic accuracies and Micro-F1s are, respectively, shown in Tables VIII and IX. As can be seen from these two tables, in all testing sets, the pretrained DCNDSC has higher diagnostic accuracy and Micro-F1 than other pretrained models apart from three datasets (CWRU, ISU, and NEPU), and it possesses higher average diagnostic accuracy and Micro-F1.

Second, seven finally trained models are tested and compared. Tables X and XI listed the computed accuracies and Micro-F1s. Comparing Tables VIII–XI, it can be seen that the finally trained DCNDSC achieves a significant improvement in terms of accuracy (increase by 11.28%) and Micro-F1 (increase by 7.24%) for the difficult and complex WTG

Authorized licensed use limited to: CHONGQING UNIVERSITY. Downloaded on May 19,2024 at 16:58:17 UTC from IEEE Xplore. Restrictions apply.

TABLE VI
ACCURACIES OBTAINED BY VARIOUS FINALLY TRAINED MODELS FOR THE TARGET TESTING SET

	DCNDSC	DenseNet-264	DenseNet-201	DenseNet-169	DenseNet-121	ResNet-152	VGG-19
CWRU	94.46±0.25	89.93±0.19	93.60±0.00	82.67±1.20	**96.47±0.37**	86.00±0.43	88.53±0.25
DDS	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	97.40±0.00	97.80±0.28	98.07±0.94
ISU	99.60±0.57	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	98.38±0.00
MFPT	**99.60±0.16**	89.16±0.50	94.28±0.09	95.22±0.25	97.78±0.16	97.58±0.44	99.60±0.00
NEPU	99.60±0.16	99.40±0.00	**100.00±0.00**	100.00±0.00	99.93±0.09	99.00±0.33	99.80±0.00
PU	93.00±0.69	95.82±0.25	96.97±0.00	95.35±1.31	91.78±0.48	**97.44±0.67**	90.30±0.29
PDTU	**98.25±0.34**	64.18±1.58	74.34±0.59	70.64±2.20	64.38±1.37	77.98±1.29	70.44±0.67
RDS	**99.73±0.25**	91.33±0.41	96.20±0.28	95.60±0.91	90.66±0.84	96.73±0.38	97.07±0.19
VATCD	**100.00±0.00**	99.79±0.00	100.00±0.00	96.43±0.67	98.92±0.34	99.60±0.29	88.75±0.19
WTG	**91.11±1.19**	65.39±0.42	78.85±0.25	72.53±0.66	68.35±0.34	67.61±0.85	75.62±0.10
Average	**97.54±3.19**	89.50±13.00	93.42±8.78	90.84±10.75	90.57±12.49	91.97±10.56	90.66±9.81

TABLE VII
MICRO-F1S OBTAINED BY VARIOUS FINALLY TRAINED MODELS FOR THE TARGET TESTING SET

	DCNDSC	DenseNet-264	DenseNet-201	DenseNet-169	DenseNet-121	ResNet-152	VGG-19
CWRU	94.82±0.21	93.31±0.10	94.68±0.13	97.61±0.75	**97.07**±0.26	88.97±0.26	89.80±0.22
DDS	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	98.38±0.00	98.60±0.14	98.08±0.94
ISU	99.60±0.56	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	98.42±0.00
MFPT	**99.80±0.10**	93.54±0.31	94.90±0.14	96.37±0.20	98.58±0.14	98.23±0.29	99.70±0.00
NEPU	99.80±0.16	99.40±0.00	**100.00±0.00**	100.00±0.00	99.93±0.09	99.02±0.31	99.80±0.00
PU	93.55±0.52	96.28±0.40	97.20±0.16	96.14±1.07	92.88±0.48	**97.58±0.52**	91.18±0.28
PDTU	**98.62±0.31**	70.96±1.40	77.70±0.58	74.14±1.65	69.00±1.07	79.13±1.03	76.24±0.53
RDS	**99.73±0.25**	92.06±0.34	96.20±0.28	95.80±0.84	90.71±0.74	96.77±0.35	97.08±0.19
VATCD	100.00±0.00	99.80±0.00	100.00±0.00	97.61±0.19	99.29±0.25	99.80±0.14	90.72±0.16
WTG	**91.34±1.16**	65.76±0.33	79.91±0.03	73.32±0.56	68.99±0.44	69.67±1.26	76.93±0.20
Average	**97.73±3.06**	91.11±11.79	94.06±7.90	93.10±9.81	91.48±11.61	92.78±9.89	91.80±8.40

TABLE VIII
ACCURACIES OBTAINED BY VARIOUS PRETRAINED MODELS FOR ALL THE TESTING SETS

	DCNDSC	DenseNet-264	DenseNet-201	DenseNet-169	DenseNet-121	ResNet-152	VGG-19
CWRU	98.10±0.07	98.22±0.19	98.13±0.26	98.23±0.27	**98.52±0.36**	97.72±0.31	96.90±0.71
DDS	**97.97±0.46**	97.90±0.11	97.57±0.26	98.67±0.10	97.37±0.62	94.48±1.95	95.93±1.36
ISU	99.92±0.02	99.97±0.02	99.97±0.05	99.97±0.02	99.98±0.02	99.95±0.00	99.55±0.18
MFPT	**99.88±0.10**	99.50±0.08	99.30±0.26	99.56±0.17	98.97±0.14	99.12±0.38	99.73±0.31
NEPU	99.95±0.04	99.95±0.00	99.87±0.05	99.88±0.10	99.93±0.06	**100.00±0.00**	99.93±0.06
PU	**92.56±0.16**	91.59±0.34	91.34±0.13	91.73±0.36	90.59±0.58	91.11±0.16	91.25±0.70
PDTU	**98.92±0.21**	98.43±0.15	98.42±0.16	98.52±0.33	98.08±0.16	97.37±0.61	86.60±3.51
RDS	**99.32±0.06**	98.95±0.15	99.22±0.09	99.27±0.20	98.67±0.21	98.27±0.36	95.68±1.25
VATCD	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	99.95±0.04	**100.00±0.00**	99.92±0.09	93.47±2.29
WTG	**69.46±2.36**	63.11±0.23	64.66±1.25	65.03±0.91	64.88±0.87	67.90±0.67	52.09±3.22
Average	**95.61±8.97**	94.76±10.81	94.85±10.35	95.08±10.28	94.70±10.28	94.58±9.29	91.11±13.60

Authorized licensed use limited to: CHONGQING UNIVERSITY. Downloaded on May 19,2024 at 16:58:17 UTC from IEEE Xplore. Restrictions apply.

TABLE IX
MICRO-F1S OBTAINED BY VARIOUS PRETRAINED MODELS FOR ALL THE TESTING SETS

	DCNDSC	DenseNet-264	DenseNet-201	DenseNet-169	DenseNet-121	ResNet-152	VGG-19
CWRU	98.30±0.14	98.44±0.12	98.43±0.16	98.36±0.27	**98.74±0.32**	97.80±0.29	97.41±0.33
DDS	**98.01±0.45**	97.95±0.12	97.62±0.25	98.68±0.10	97.44±0.59	94.79±2.17	96.09±1.28
ISU	99.95±0.02	99.98±0.02	99.98±0.03	99.97±0.02	**99.99±0.02**	99.97±0.00	99.72±0.09
MFPT	**99.89±0.12**	99.50±0.08	99.30±0.25	99.56±0.17	99.02±0.20	99.17±0.40	99.73±0.31
NEPU	99.95±0.04	99.95±0.00	99.87±0.05	99.88±0.10	99.93±0.06	**100.00±0.00**	99.93±0.06
PU	**92.75**±0.14	91.97±0.37	92.02±0.18	92.13±0.41	91.09±0.31	91.84±0.35	91.38±0.69
PDTU	**98.96±0.20**	98.44±0.15	98.44±0.16	98.52±0.34	98.09±0.16	97.38±0.62	86.87±3.58
RDS	**99.33±0.07**	98.95±0.15	99.23±0.09	99.28±0.19	98.68±0.20	98.29±0.36	95.88±1.20
VATCD	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	99.97±0.02	**100.00±0.00**	99.96±0.05	95.72±1.45
WTG	**76.21±2.36**	70.25±0.80	71.18±1.28	72.00±0.82	71.71±0.64	75.02±0.61	55.60±4.28
Average	**96.34±7.02**	95.54±8.73	95.61±8.44	95.84±8.25	95.47±8.30	95.41±7.24	91.83±12.68

TABLE X
ACCURACIES OBTAINED BY VARIOUS FINALLY-TRAINED MODELS FOR ALL THE TESTING SETS

	DCNDSC	DenseNet-264	DenseNet-201	DenseNet-169	DenseNet-121	ResNet-152	VGG-19
CWRU	98.33±0.28	95.80±0.19	98.32±0.02	93.98±0.52	**98.82±0.09**	94.90±0.35	89.92±0.35
DDS	**98.68±0.06**	95.42±0.25	99.03±0.02	99.12±0.02	98.47±0.19	95.70±0.25	90.38±0.44
ISU	**98.82±0.21**	99.78±0.05	99.88±0.02	99.90±0.04	99.78±0.02	97.76±0.43	98.74±0.00
MFPT	**99.81±0.05**	97.81±0.37	99.43±0.05	99.43±0.05	99.33±0.02	98.84±0.33	99.85±0.00
NEPU	99.83±0.62	99.45±0.00	**99.85±0.00**	99.98±0.02	99.83±0.02	99.30±0.19	99.82±0.02
PU	90.00±0.19	**91.57±0.17**	90.74±0.04	90.69±0.21	90.40±0.37	88.06±0.45	88.13±0.31
PDTU	**96.13±0.27**	92.39±0.90	94.78±0.10	93.60±0.88	89.31±0.23	92.98±0.50	84.98±0.17
RDS	**99.57±0.08**	98.40±0.00	99.35±0.04	99.32±0.02	97.82±0.12	98.92±0.14	92.80±0.12
VATCD	**99.68±0.06**	99.92±0.02	100.00±0.00	98.52±0.17	99.58±0.12	99.02±0.23	90.96±0.07
WTG	**80.74±0.60**	70.81±0.27	76.70±0.37	75.98±0.38	72.71±0.30	67.90±0.38	68.00±0.09
Average	**96.16±5.87**	94.14±8.27	95.81±6.96	95.05±7.07	94.61±8.18	93.34±9.12	90.36±8.89

TABLE XI
MICRO-F1S OBTAINED BY VARIOUS FINALLY TRAINED MODELS FOR ALL THE TESTING SETS

	DCNDSC	DenseNet-264	DenseNet-201	DenseNet-169	DenseNet-121	ResNet-152	VGG-19
CWRU	98.46±0.16	97.45±0.10	98.56±0.06	96.03±0.29	**99.11±0.07**	96.25±0.18	91.03±0.29
DDS	**98.88±0.04**	97.42±0.10	99.46±0.02	99.30±0.03	99.06±0.13	97.18±0.12	91.91±0.36
ISU	**99.40±0.13**	99.89±0.02	99.93±0.02	99.95±0.02	99.89±0.02	98.82±0.22	99.29±0.00
MFPT	**99.91±0.02**	98.81±0.20	99.46±0.04	99.55±0.03	99.49±0.04	99.37±0.21	99.87±0.00
NEPU	99.83±0.06	99.48±0.00	99.85±0.00	99.98±0.02	**99.83±0.02**	99.31±0.18	99.82±0.02
PU	90.78±0.20	**91.80±0.18**	91.27±0.05	91.57±0.12	90.80±0.32	88.53±0.43	88.59±0.27
PDTU	**96.27±0.25**	93.20±0.89	95.05±0.11	93.85±0.84	90.00±0.22	93.02±0.51	85.44±0.13
RDS	**99.72±0.09**	98.77±0.01	99.48±0.03	99.45±0.02	98.20±0.12	99.03±0.14	95.24±0.10
VATCD	**99.84±0.03**	99.92±0.03	100.00±0.00	99.01±0.04	99.71±0.08	99.13±0.21	93.46±0.03
WTG	**83.45±0.53**	71.96±0.28	77.96±0.35	77.45±0.33	73.85±0.27	70.81±0.10	69.73±0.10
Average	**96.65±5.15**	94.87±8.08	96.10±6.61	95.61±6.65	94.99±7.89	94.15±8.46	91.44±8.59

dataset, although its diagnostic effect becomes slightly worse in some datasets. Thus, it has higher average accuracy and Micro-F1 than the pretrained DCNDSC, that is, the average accuracy increases from 95.61% to 96.16%, and the average Micro-F1 increases from 96.34% to 96.65%. With regard to other models, we can also see the obvious improvement for the WTG dataset, whereas their average accuracies and Micro-F1s become slightly smaller. It follows that the proposed diminutive network fine-tuning strategy not only enhances the diagnostic ability of large model on the new target data but also retains the main feature extraction capability of the pretrained DCNDSC. From the comparison of Tables VIII–XI,

Authorized licensed use limited to: CHONGQING UNIVERSITY. Downloaded on May 19,2024 at 16:58:17 UTC from IEEE Xplore. Restrictions apply.

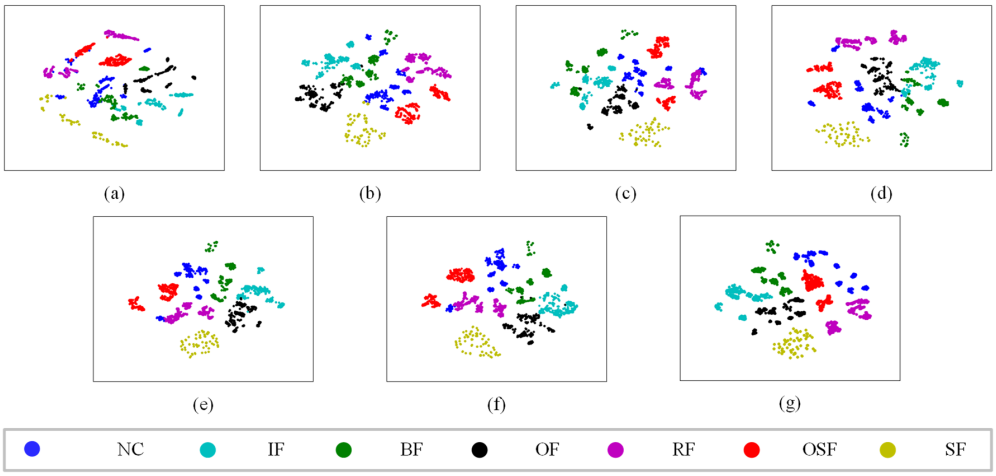

Fig. 7. t-SNE maps obtained by various models for some testing datasets: (a) VGG19, (b) ResNet-152, (c) DenseNet-121, (d) DenseNet-169, (e) DenseNet-201, (f) DenseNet-256, and (g) DCNDSC.

we also note that the diminutive network fine-tuning strategy can make the proposed method have stronger generalization ability than other methods. In addition, it can be known from Tables X and XI that DCNDSC has higher average accuracy and Micro-F1 than the typical and advanced large models. The high diagnostic performance of DCNDSC arises from strong feature extraction capability of DCDSCB and the stacking of many DCDSCBs.

For the purpose of clearly observing the difference between the learned features, the t-distributed stochastic neighbor embedding (t-SNE) [45] maps obtained by various models for some testing datasets are drawn in Fig. 7. It can be noted that the features of different faults extracted by DCNDSC have a better divisibility; hence, the proposed DCNDSC has a higher classification accuracy than the existing large-scale networks.

IV. Conclusion

To accurately diagnose the various types of faults in different rotating machines and obtain the strong generalization ability, a large model for rotating machine fault diagnosis is investigated and provided in this work for the first time. First, a DCNDSC is constructed as the backbone of a large model. As the designed DCDSCB can effectively capture the features across different channels and layers and a large number of DCDSCBs are stacked. The proposed DCNDSC can well mine a variety of complicated features caused by different faults and working conditions. Second, we collect the public rotating machine fault data and our own fault data as far as possible to set up a large rotating machine fault dataset, and it is used to pretrain and fine-tune the proposed DCNDSC. Third, a diminutive network fine-tuning strategy is developed for enhancing the feature extraction ability on the new task data and generalization ability of pretrained DCNDSC. Finally, a series of experiments are performed to check the performance of the pretrained and finally tuned models, and the proposed large model is compared with the current typical and advanced large-scale networks. The comparative results indicate that the developed large model has a higher diagnostic accuracy and strong generalization ability; meantime, the experiment results prove the significance and necessity of the diminutive network fine-tuning strategy. In future work, we will collect more rotating machine fault samples to train the large model, thus further improving the ability of feature extraction and generalization. Additionally, a fine-tuning strategy based on reinforcement learning is worth exploring to further enhance the generalization capability of large model and the lightweight techniques will be investigated to achieve the model deployment in industrial scenarios.

References

[1] H. Pu, S. Teng, D. Xiao, L. Xu, Y. Qin, and J. Luo, "Compound fault diagnosis of rotating machine through label correlation modeling via graph convolutional neural network," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–10, 2024.

[2] Y. Qin, Q. Li, S. Wang, and P. Cao, "Dynamics modeling of faulty planetary gearboxes by time-varying mesh stiffness excitation of spherical overlapping pittings," *Mech. Syst. Signal Process.*, vol. 210, Mar. 2024, Art. no. 111162.

[3] C. Fan, K. Xiahou, L. Wang, and Q. H. Wu, "Hybrid fault diagnosis of multiple open-circuit faults for cascaded H-bridge multilevel converter based on perturbation estimation convolution network," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–12, 2024.

[4] Q. Qian, J. Zhou, and Y. Qin, "Relationship transfer domain generalization network for rotating machinery fault diagnosis under different working conditions," *IEEE Trans. Ind. Informat.*, vol. 19, no. 9, pp. 9898–9908, Sep. 2023.

[5] L. Wan, J. Ning, Y. Li, C. Li, and K. Li, "Intelligent fault diagnosis via ring-based decentralized federated transfer learning," *Knowl.-Based Syst.*, vol. 284, Jan. 2024, Art. no. 111288.

[6] C. Yang, J. Liu, K. Zhou, M.-F. Ge, and X. Jiang, "Transferable graph features-driven cross-domain rotating machinery fault diagnosis," *Knowl.-Based Syst.*, vol. 250, Aug. 2022, Art. no. 109069.

[7] S. Haidong, D. Ziyang, C. Junsheng, and J. Hongkai, "Intelligent fault diagnosis among different rotating machines using novel stacked transfer auto-encoder optimized by PSO," *ISA Trans.*, vol. 105, pp. 308–319, Oct. 2020.

[8] Z. Wang, X. He, B. Yang, and N. Li, "Subdomain adaptation transfer learning network for fault diagnosis of roller bearings," *IEEE Trans. Ind. Electron.*, vol. 69, no. 8, pp. 8430–8439, Aug. 2022.

[9] Q. Qian, J. Luo, and Y. Qin, "Adaptive intermediate class-wise distribution alignment: A universal domain adaptation and generalization method for machine fault diagnosis," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 21, 2024, doi: 10.1109/TNNLS.2024.3376449.

[10] G. Jiang, H. He, P. Xie, and Y. Tang, "Stacked multilevel-denoising autoencoders: A new representation learning approach for wind turbine gearbox fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 9, pp. 2391–2402, Sep. 2017.

[11] T. Han, C. Liu, W. Yang, and D. Jiang, "Learning transferable features in deep convolutional neural networks for diagnosing unseen machine conditions," *ISA Trans.*, vol. 93, pp. 341–353, Oct. 2019.

Authorized licensed use limited to: CHONGQING UNIVERSITY. Downloaded on May 19,2024 at 16:58:17 UTC from IEEE Xplore. Restrictions apply.

[12] T. B. Brown et al., "Language models are few-shot learners," in *Proc. NIPS*, 2020, pp. 1877–1901.
[13] D. Driess et al., "PaLM-E: An embodied multimodal language model," 2023, *arXiv:2303.03378*.
[14] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–26.
[15] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–22.
[16] Y. Zhang, N. Qin, D. Huang, B. Wu, and Z. Liu, "High-accuracy and adaptive fault diagnosis of high-speed train bogie using dense-squeeze network," *IEEE Trans. Veh. Technol.*, vol. 71, no. 3, pp. 2501–2510, Mar. 2022.
[17] X. Li, J. Li, C. Zhao, Y. Qu, and D. He, "Gear pitting fault diagnosis with mixed operating conditions based on adaptive 1D separable convolution with residual connection," *Mech. Syst. Signal Process.*, vol. 142, Aug. 2020, Art. no. 106740.
[18] M. Miao, Y. Sun, and J. Yu, "Deep sparse representation network for feature learning of vibration signals and its application in gearbox fault diagnosis," *Knowl.-Based Syst.*, vol. 240, Mar. 2022, Art. no. 108116.
[19] M. M. Bejani and M. Ghatee, "A systematic review on overfitting control in shallow and deep neural networks," *Artif. Intell. Rev.*, vol. 54, no. 8, pp. 6391–6438, Dec. 2021.
[20] M. Bianchini and F. Scarselli, "On the complexity of neural network classifiers: A comparison between shallow and deep architectures," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 8, pp. 1553–1565, Aug. 2014.
[21] O. Delalleau and Y. Bengio, "Shallow vs. deep sum-product networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 666–674.
[22] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI, Tech. Rep., 2018.
[23] W. A. Smith and R. B. Randall, "Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study," *Mech. Syst. Signal Process.*, vols. 64–65, pp. 100–131, Dec. 2015.
[24] *Data Castle., Bearing Fault Data*. Accessed: 2022. [Online]. Available: https://www.datacastle.cn/dataset_description.html?type=dataset&id=539
[25] Q. Qian, Y. Wang, T. Zhang, and Y. Qin, "Maximum mean square discrepancy: A new discrepancy representation metric for mechanical fault transfer diagnosis," *Knowl.-Based Syst.*, vol. 276, Sep. 2023, Art. no. 110748.
[26] Q. Yao, Y. Qin, X. Wang, and Q. Qian, "Multiscale domain adaption models and their application in fault transfer diagnosis of planetary gearboxes," *Eng. Appl. Artif. Intell.*, vol. 104, Sep. 2021, Art. no. 104383.
[27] *Industrial Innovation Platform., Gearbox Failure Test Data*. Accessed: 2022. [Online]. Available: https://www.industrial-bigdata.com/Dataset
[28] H. Qiu, J. Lee, J. Lin, and G. Yu, "Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics," *J. Sound Vibrat.*, vol. 289, nos. 4–5, pp. 1066–1090, Feb. 2006.
[29] H. Lu et al., "A physics-informed feature weighting method for bearing fault diagnostics," *Mech. Syst. Signal Process.*, vol. 191, May 2023, Art. no. 110171.
[30] K. Li, X. Ping, H. Wang, P. Chen, and Y. Cao, "Sequential fuzzy diagnosis method for motor roller bearing in variable operating conditions based on vibration analysis," *Sensors*, vol. 13, no. 6, pp. 8013–8041, Jun. 2013.
[31] E. Bechhoefer. (2016). *Condition Based Maintenance Fault Database for Testing of Diagnostic and Prognostics Algorithms*. [Online]. Available: https://mfpt.org/fault-data-sets/
[32] H. Fang et al., "You can get smaller: A lightweight self-activation convolution unit modified by transformer for fault diagnosis," *Adv. Eng. Informat.*, vol. 55, Jan. 2023, Art. no. 101890.
[33] C. Lessmeier, J. K. Kimotho, D. Zimmer, and W. Sextro, "Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification," in *Proc. PHM Soc. Eur. Conf.*, 2016.
[34] A. P. Daga, A. Fasana, S. Marchesiello, and L. Garibaldi, "The Politecnico di Turin rolling bearing test rig: Description and analysis of open access data," *Mech. Syst. Signal Process.*, vol. 120, pp. 252–273, Apr. 2019.
[35] Q. Qian, Y. Qin, J. Luo, Y. Wang, and F. Wu, "Deep discriminative transfer learning network for cross-machine fault diagnosis," *Mech. Syst. Signal Process.*, vol. 186, Mar. 2023, Art. no. 109884.
[36] S. Shao, S. McAleer, R. Yan, and P. Baldi, "Highly accurate machine fault diagnosis using deep transfer learning," *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2446–2455, Apr. 2019.
[37] P. Cao, S. Zhang, and J. Tang, "Preprocessing-free gear fault diagnosis using small datasets with deep convolutional neural network-based transfer learning," *IEEE Access*, vol. 6, pp. 26241–26253, 2018.
[38] W. Jung, S.-H. Kim, S.-H. Yun, J. Bae, and Y.-H. Park, "Vibration, acoustic, temperature, and motor current dataset of rotating machine under varying operating conditions for fault diagnosis," *Data Brief*, vol. 48, Jun. 2023, Art. no. 109049.
[39] T. Li, Z. Zhou, S. Li, C. Sun, R. Yan, and X. Chen, "The emerging graph neural networks for intelligent fault diagnostics and prognostics: A guideline and a benchmark study," *Mech. Syst. Signal Process.*, vol. 168, Apr. 2022, Art. no. 108653.
[40] B. Wang, Y. G. Lei, and N. Li, "A hybrid prognostics approach for estimating remaining useful life of rolling element bearings," *IEEE Trans. Rel.*, vol. 69, no. 1, pp. 401–412, Mar. 2018.
[41] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.
[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[44] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput Vis. Pattern Recognit.*, Jul. 2017, pp. 4700–4708.
[45] L. van der Maaten, "Accelerating t-SNE using tree-based algorithms," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3221–3245, Oct. 2014.

Yi Qin (Senior Member, IEEE) received the B.Eng. and Ph.D. degrees in mechanical engineering from Chongqing University, Chongqing, China, in 2004 and 2008, respectively.

Since January 2009, he had been with Chongqing University, where he is currently a Professor with the College of Mechanical Engineering. He has published over 190 articles. His current research interests include fault diagnosis, life prediction, and digital twin.

Taisheng Zhang received the B.Eng. degree in mechanical engineering from Chongqing University, Chongqing, China, in 2022, where he is currently pursuing the master's degree in mechanical engineering.

His research interests mainly include deep learning and fault diagnosis.

Quan Qian was born in Chongqing, China, in 1998. He received the B.Eng. degree in mechanical engineering from Southwest Jiaotong University, Chengdu, China, in 2020. He is currently pursuing the Ph.D. degree in mechatronic engineering with Chongqing University, Chongqing, China.

His research interests mainly include signal processing, transfer learning, and intelligent fault diagnosis.

Yongfang Mao (Member, IEEE) received the B.Eng. and Ph.D. degrees in mechanical engineering from Chongqing University, Chongqing, China, in 2004 and 2008, respectively.

Since March 2009, she has been with Chongqing University, where she is currently an Associate Professor with the College of Automation. Her current research interests include fault diagnosis and prognosis.

Authorized licensed use limited to: CHONGQING UNIVERSITY. Downloaded on May 19,2024 at 16:58:17 UTC from IEEE Xplore. Restrictions apply.