Inferable Deep Distilled Attention Network for Diagnosing Multiple Motor Bearing Faults

Xiaotian Zhang, *Student Member, IEEE,* Yihua Hu, *Senior Member*, *IEEE*, Aijun Yin, *Member*, *IEEE*, Jiamei Deng, *Senior Member*, *IEEE*, Hui Xu, Jikai Si, *Member*, *IEEE*

***Abstract*—Bearing, as a vital component in electric powertrains, is increasingly used globally such as in electric vehicle (EV). Their damages and faults may bring huge cost loss to the industry and even threaten personal safety. This paper proposes an inferable deep distilled attention network (IDDAN) method which is a self-attention mechanism and transfer learning-based method to diagnose and classify multiple bearing faults in various motor drive systems efficiently and accurately. Compared with convolutional networks, the self-attention-based network can better extract the global feature information and easier to benefit from large amounts of pre-training data. Its significance is to accurately classify various faults of the target machine when the labeled data of the target machine is not enough to directly train the diagnosis model. Firstly, this paper attempt to apply the self-attention-based network to build an advanced fault diagnosis model. Secondly, this paper optimizes the structure of networks through knowledge distillation (KD) technique to require a lighter and fast model. Thirdly, this paper proposes a new data augmentation strategy for 1-D vibration signals to provide large-scale pre-training samples for IDDAN. Experiments show that the self-attention mechanism-based model is more likely to benefit from large-scale data. After testing, compared with many methods and other exist similar methods, the proposed method achieves higher classification accuracy and better performance.**

***Index Terms*—Motor bearing, Fault diagnosis, Fault detection, Neural network applications, Transfer learning**

I. INTRODUCTION

THE electric motor drive system has been widely used in industry and human life. The reliability and safety of its components bear the responsibility of human life and industrial cost. The reliability issues of the electric powertrain may appear on any components. The bearing plays a critical and necessary role in motor drive system. According to incomplete statistics, 40-70% motor and electric powertrain faults are caused by various degrees of rolling bearing damage [1]. Such faults are leading to the higher costs in industrial applications and its maintenance. Therefore, real-time condition monitoring and bearing fault diagnosis (BFD) in all motor drive system is gradually becoming more important and higher-priority work.

The safety and stability of bearings have attracted increasing attention in both academic and engineering. Scholars and engineers have employed many traditional methods to detect which type of bearing fault in some motor drive systems. The conventional method of bearing fault diagnosis mainly relies on advanced signal processing technology to extract effective features for analysis. Paper [2] proposes an adaptive morphological filter (AMF) to analyze the vibration and acoustical signals of the bearing to determine the fault type. Paper [3] proposes the sparse elitist group lasso denoising (SEGLD) algorithm to online diagnose bearing faults in industry. This is because part of the information contained in the motor stator current signal or the bearing vibration signal collected by the measurement is not related to the bearing fault, such as the supply fundamental and its harmonics, noise, etc. The core contributions of the paper [4] and [5] are both to solve this problem. On the other hand, judging from the current trend of big data, artificial intelligence (AI) can bring about better convenience and more advantageous new ideas for bearing fault diagnosis.

The AI-based BFD methods have characteristics of model independence, does not require professional mechanical knowledge, and has excellent performance. Paper [6] uses graph-mapped spectrum (GMS) to represent fault information in bearing vibration signals and applies K-nearest neighbour (K-NN) classifier to identify fault types. Paper [7] combines information fusion (IF) technology with convolutional neural network (CNN) to diagnose bearing faults. Paper [8] is also based on supervised learning of AI. The method proposed in this paper is based on CNN to identify damage to rotor bearings from infrared images. Supervised learning is a classic method in AI-based methods. However, for some machines, it is difficult to obtain sufficient labeled data for supervised training in real situations. People need new ways to solve this problem.

To solve the above problem, transfer learning technology is used in bearing multiple fault diagnosis. It obtains a pre-trained model with rich domain knowledge from a machine, and then adapts it to the target machine with a small amount of data. Due to the difficulty in obtaining bearing fault data, there are many similar methods based on transfer learning recently. Paper [9] proposes an intelligent bearing fault diagnosis system combining AlexNet and transfer learning technology. The deep convolutional transfer learning network (DCTLN) proposed in paper [10] can make the model still effective in the target domain. These methods are almost all proposed based on CNN [11], [12]. Paper [13] uses Bayesian network as a fault diagnosis model for bearing fault detection and proposes varying coefficient transfer learning (VCTL) to obtain knowledge and correlation from the resource domain. Paper [14] uses the bearing vibration data obtained by computer simulation to pre-train the model, and then achieve the general effect of the model through transfer learning. However, traditional network models such as SVM, CNN have the ability of automatic feature learning, but still face many challenges. For example, CNN cannot focus on learning to important discriminate features of faults and ignore useless features. Furthermore, global information cannot be extracted due to the limitation of convolution kernel size. The new type of neural network based on self-attention mechanism will improve the current situation.

Based on the transfer learning and a new self-attention mechanism, an inferable deep distilled attention network (IDDAN) is proposed to diagnose bearing multiple faults. The network can accumulate the advantage of optimized feature mapping across the network through the intervention of the self-attention mechanism, which uses global information to adaptively enhance more discriminative features and suppress irrelevant features. With the proposed data augmentation technique, it is possible to explore how much each method benefits from large amount of data samples. In addition, the knowledge distillation (KD) technique [15] makes the trained network lighter and increases the inference speed. The contributions of this paper are summarized as follows:

1) It originally proposed a new data augmentation method of 1-D vibration signals and applied it in the bearing fault diagnosis framework. Models can benefit from data augmentation to become more generalizable. Experiments show that this strategy cooperates well with the self-attention module to obtain accurate diagnosis results
2) Self-attention mechanism based neural network is introduced in transfer learning-based intelligence diagnosis method. The types of bearing faults are complex and diverse, which poses a higher challenge to the feature recognition ability of the diagnosis model. Self-attention mechanism based can enhance more discriminative features and suppress irrelevant features during training.
3) An advanced bearing multiple fault diagnosis method, IDDAN, is proposed based on data augmentation technique and self-attention networks and works under various working conditions. The model size of the method is lighter than commonly used models so as to increase computing efficiency.

The rest paragraphs are organized as follows. Section II introduces related theories for further understanding the architecture introduced in Section III. Section III describes the proposed method in detail. Section IV presents the case study and experiment result. Finally, Section V provides a conclusion of this paper.

II. Self-Attention Mechanism and Related Theories

The self-attention mechanism is rarely used in engineering and is quite different from the convolution mechanism. These related theories will be briefly introduced in this section.

A. Self-Attention Mechanism

Self-attention is a special form of attention mechanism. The output of attention mechanism (see Fig.1) could be presented as (1) [16].

$$Attention_{Output} = Attention(Q, K, V) \tag{1}$$

In (1), Q, K and V respectively stand for the *Query* matrix, *Key* matrix, and *Value* matrix. Let X be the input and get $Q = W_Q X$, $K = W_K X$, $V = W_V X$ (W_Q , W_K and W_V are parameter matrixes).

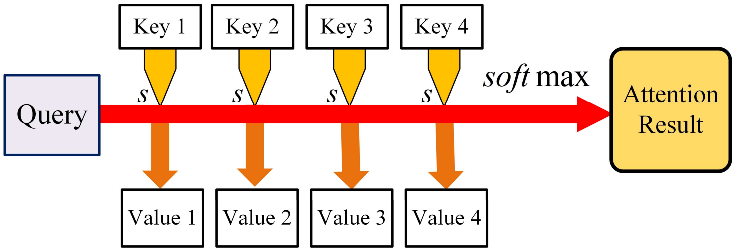


Fig. 1. Diagram of computing process of attention mechanism.

The scaled dot-product is used in the calculation process as shown in (2):

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

Among (2), the *softmax* function is a function that turns a vector of K real values into a vector of K real values that sum to 1, which formula could be presented as (3). The input values can be positive, negative, zero, or greater than one, but the *softmax* transforms them into values between 0 and 1, so that they can be interpreted as probabilities. It is usual to append a *softmax* function as the final layer of the neural network to convert the scores to a normalized probability distribution.

$$\sigma(\vec{Z})_i = \frac{e^{Z_i}}{\sum_{j=1}^{K} e^{Z_j}} \tag{3}$$

B. Multi-Head Attention Mechanism

Multi-head attention is to project the h group Q , K , V through different linear transforms, and connect the final result. In self-attention mechanism, each group of Q, K, V is the same.

$$Multihead(Q, K, V) = Concat(head_1, head_2, \ldots, head_h)W^O \tag{4}$$

$$Head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{5}$$

In both (4) and (5), W represents for the parameter matrix of linear transforms. In detail, $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$ and $W_i^O \in \mathbb{R}^{hd_v \times d_{model}}$ ($d_k = d_v = d_{model}/h$).

C. Classic KD Theory

Neural network models can solve a variety of complex problems, but these models are usually huge and have a large number of parameters, making it difficult or impossible to deploy to edge devices. KD is a new method of compressing neural models. The obtained new smaller network trained through KD technology can achieve the same or similar effect as the original network [17].

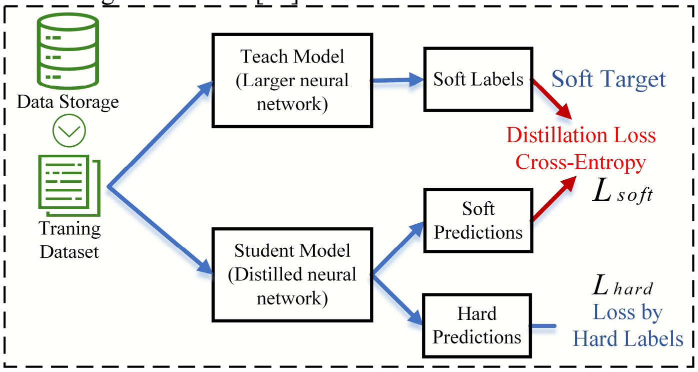


Fig. 2. Schematic diagram of complete KD process.

Fig. 2 shows that the essence of KD is the process of pre-trained larger teacher network teaching smaller student network. The training dataset is applied to both the teacher network and the student network at the same time. The soft label is the output of the teacher network in each layer. The cross-entropy between it and the soft prediction value output by the student model is L_{soft}. L_{hard} is defined as the cross-entropy of the hard prediction value output by the student model and data label. Therefore, the objective function of the KD process is composed of weighted L_{soft} and weighted L_{hard}:

$$L = \alpha L_{soft} + \beta L_{hard} \tag{6}$$

where α and β are weights of L_{soft} and L_{hard}, respectively.

D. Transfer Learning Problem

The essence of the model pre-training is transfer learning [18]. To clearly explain the proposed architecture in Section III, here it is necessary to introduce two types of domains in transfer learning: the source domain D_s and the target domain D_t. Both domains are composed of the feature space X and the probability distribution $P(X)$ of the data. The category spaces of the learning objectives of transfer learning in D_s and D_t are represented by Y_s and Y_t, respectively. When $D_s \neq D_t$, the data distribution before and after transfer is also different. Transfer learning can improve the performance of the target task learning function f_t when $D_s \neq D_t$ or $T_s \neq T_t$. The following Fig. 3 shows the principle of transfer learning to process unlabelled data task and what is different process between transfer learning and traditional neural networks.

This paper aims to looking for a multi-fault diagnosis architecture that can be quickly deployed on any machine and monitor its health conditions. From the perspective of transfer learning, the data used for pre-training $X = \{x_1, x_2 \cdots x_n\}$and the corresponding label space $Y_s = \{y_{x_1}, y_{x_2} \cdots y_{x_n}\}$ compose source domain $D_s = \{X, Y_s\}$. The unlabelled data of the target machine is $X_t = \{x_{t1}, x_{t2} \cdots x_{tn}\}$, where x_{tn} is the data samples of the target task. The more adequate samples in D_s, the larger the category space, the stronger generalization ability, and the better performance of the model after pre-training transfer.

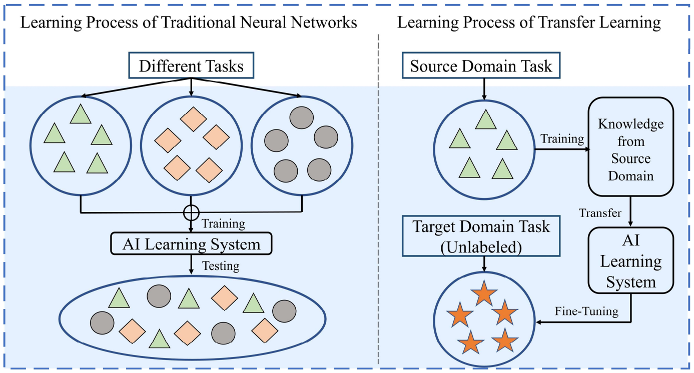

Learning Process of Traditional Neural Networks
Learning Process of Transfer Learning
Different Tasks
Training
AI Learning System
Testing
Source Domain Task
Training
Knowledge from Source Domain
Transfer
Target Domain Task (Unlabeled)
AI Learning System
Fine-Tuning

Fig. 3. The comparison and difference of the learning process between traditional neural networks and transfer learning.

III. Intelligent Diagnosis Framework Based on Inferable Deep Distilled Attention Network (IDDAN)

The proposed diagnosis method consists of four main modules (as shown in Fig. 4): data augmentation, backbone network, distillation strategy and transfer inference. Transfer inference helps IDDAN become more adaptable to the target domain through fine-tuning by the very small amount of data after the pre-training stage. Particularly, the first and second parts of Fig. 4 show how the bearing vibration signal is collected and pre-processed in the context of the current application.

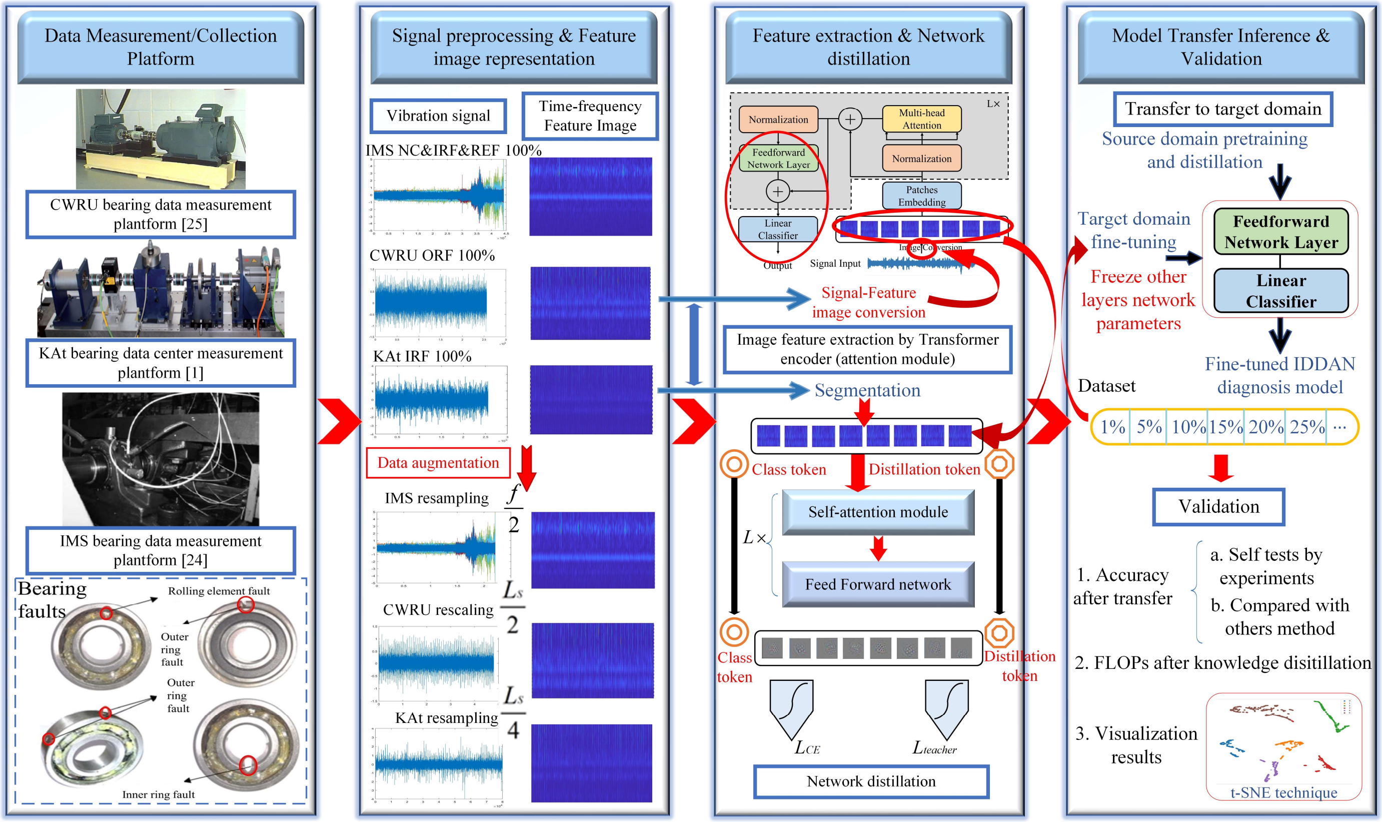

Data Measurement/Collection Platform
CWRU bearing data measurement plantform [25]
KAt bearing data center measurement plantform [1]
IMS bearing data measurement plantform [24]
Bearing faults
Rolling element fault
Outer ring fault
Outer ring fault
Inner ring fault
Signal preprocessing & Feature image representation
Vibration signal
Time-frequency Feature Image
IMS NC&IRF&REF 100%
CWRU ORF 100%
KAt IRF 100%
Data augmentation
IMS resampling
f
2
CWRU rescaling
L_s
2
KAt resampling
L_s
4
Feature extraction & Network distillation
L×
Normalization
Multi-head Attention
Feedforward Network Layer
Normalization
Patches Embedding
Linear Classifier
Output
Signal Input
Image Conversion
Signal-Feature image conversion
Image feature extraction by Transformer encoder (attention module)
Segmentation
Class token
Distillation token
Self-attention module
L×
Feed Forward network
Class token
Distillation token
L_{CE}
$L_{teacher}$
Network distillation
Model Transfer Inference & Validation
Transfer to target domain
Source domain pretraining and distillation
Target domain fine-tuning
Feedforward Network Layer
Linear Classifier
Freeze other layers network parameters
Fine-tuned IDDAN diagnosis model
Dataset
1% 5% 10% 15% 20% 25% ···
Validation
1. Accuracy after transfer
a. Self tests by experiments
b. Compared with others method
2. FLOPs after knowledge disitillation
3. Visualization results
t-SNE technique

Fig. 4. The architecture demonstration of the proposed method.

A. Data Augmentation Strategy

The data augmentation method of IDDAN's pre-training data in the source domain adopts multi-scale and multi-timescale signal conversion which is a special data augmentation method proposed for the collected 1-D bearing vibration signals.

Multiscale signal conversion is similar to the idea of cropping and zoom in computer vision, which can improve the generalization ability of pre-trained IDDAN. In other words, the domain adaptation ability of the IDDAN can be improved when transferring know from the source domain to the target domain. Defining the length of each vibration signal as L_s and the sampling rate of as f_s, the principle of data enhancement can be expressed as Fig. 5.

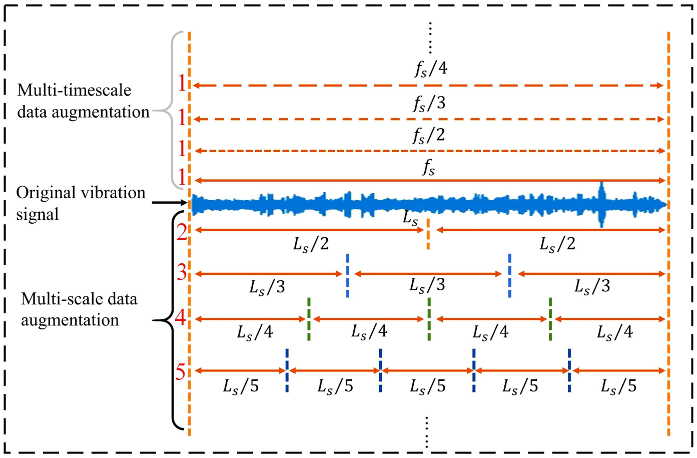


Fig. 5. The data augmentation strategy diagram of the data used for IDDAN pre-training.

It can be obtained from Fig. 5 that the strategy of data augmentation could be presented by equation (7):

$$\begin{cases} k \times \frac{L_s}{k},\ k \in \mathbb{N}^* \\ V_{new}(t) = \sum_{n=-\infty}^{\infty} V_s(nT)\delta(t-nT) \end{cases} \tag{7}$$

where k is the number of segments after each signal cutting, $V_{new}(t)$ is down sampled vibration signal, $V_s(t)$ is the original collected vibration signal, δ represents the impulse function. In this data augmentation step, collected 1-D vibration signals are mainly processed in two ways:

1) Each vibration signal used for IDDAN pre-training step is segmented multiple parts, and the size of each segment is $L_s/2$, $L_s/3$, $L_s/4$, $L_s/5$, etc.
2) Each vibration signal used for IDDAN pre-training step is resampled by sampling rate $f_s/2$, $f_s/3$, $f_s/4$, etc.

Through this data augmentation method, the amount of data can be increased efficiently, providing sufficient data for IDDAN pre-training to improve the effect of knowledge transfer.

B. Backbone Network

The bearing health status recognition is realized through self-attention based network module as the main body.

The backbone of IDDAN is a modified transformer network, which includes one transformer encoder and a linear classifier. The transformer encoder is constructed by one multi-head attention module, one feed-forward module and two normalization layers, which can automatically learn global features with the help of the self-attention mechanism [19]. A linear classifier is used to identify and distinguish bearing health conditions. As mentioned in Section I, the self-attention mechanism can be processed parallelly with global capabilities, long-distance information will not be weakened [20]. Self-attention could be considered as a CNN with a learnable receptive field. In other words, self-attention can learn receptive field automatically, but the receptive field of CNN needs manual adjustment and optimization of parameters.

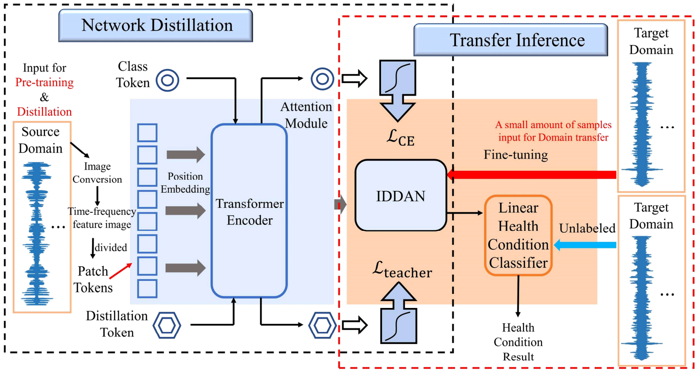


Fig. 6. The structure presentation of IDDAN.

As shown in Fig. 6, the self-attention module includes a position embedding module, a transformer encoder (feature extraction in Fig.4), and a linear layer. The linear layer can be regarded as a condition classifier to classify the global features extracted by the self-attention mechanism in the transformer encoder.

The motor signal of a certain length is first converted into the time-frequency map representation. For handling these 2D signal representations, the transformer encoder reshapes the image $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ into a sequence of flattened 2D patches $\mathbf{x}_p \in \mathbb{R}^{N\times(P^2\cdot C)}$, where (H, W) is the resolution of the signal representation, C is the number of channels, (P, P) is the resolution of each image patch. The effective input sequence length could be calculated through:

$$N = HW/P^2 \tag{8}$$

where N stands for the resulting number of patches.

These patches representing machine health information have positions in the original time-series signal. Position embedding can achieve the effect of abstracting data in time series and represent relative or absolute position information in the input sequence. The implementation process of position embedding in this paper is as follows:

$$\mathbf{z}_0 = \left[\mathbf{x}_{\text{class}};\ \mathbf{x}_p^1\mathbf{E};\ \mathbf{x}_p^2\mathbf{E};\ \cdots;\ \mathbf{x}_p^N\mathbf{E}\right] + \mathbf{E}_{pos} \tag{9}$$

where $\mathbf{E} \in \mathbb{R}^{(P^2\cdot C)\times D}$, $\mathbf{E}_{pos} \in \mathbb{R}^{(N+1)\times D}$, $\mathbf{E}$ is the patch embedding projection and is D the constant latent vector size through all of transformer layers.

The patches with marked positional information need to perform a normalization operation before entering the multi-head self-attention (MSA) mechanism. This paper adopts layer normalization (LN) [21]. The state at the output of the Transformer encoder $\mathbf{z}_L^0$ serves as the image representation $\mathbf{y} = \text{LN}(\mathbf{z}_L^0)$. Both during pre-training and fine-tuning, a classification head is attached to $\mathbf{z}_L^0$.

$$\mathbf{z}'_\ell = \text{MSA}\big(\text{LN}(\mathbf{z}_{\ell-1})\big) + \mathbf{z}_{\ell-1} \tag{10}$$

where $\ell = 1, \cdots, L$. MSA is an extension of standard self-attention (SA), which runs SA operations k times (The calculation principle has shown in Section II-A).

$$\text{MSA}(\mathbf{z}) = [\text{SA}_1(z);\ \text{SA}_2(z);\ \cdots;\ \text{SA}_k(z)]\mathbf{U}_{msa} \tag{11}$$

For calculating MSA in this paper, we set $\text{SA}(\mathbf{z}) = A\mathbf{v}$ and $\mathbf{U}_{msa} \in \mathbb{R}^{k\cdot D_h\times D}$,where $A = \text{softmax}\big(\mathbf{q}\mathbf{k}^\top/\sqrt{D_h}\big)$ $(A \in \mathbb{R}^{N\times N}$, $D_h = D/k$ and $[\mathbf{q}, \mathbf{k}, \mathbf{v}] = \mathbf{z}\mathbf{U}_{sa}$, $\mathbf{U}_{sa} \in \mathbb{R}^{D\times 3D_h})$.

As shown in Fig. 6, LN is employed before both MSA block and FFN block, and residual connections after every block. Our FFN block contains two layers: one hidden layer at pre-training time and by a single linear layer at fine-tuning time:

$$FFN(x) = \max(0, xW_1 + b_1)\, W_2 + b_2 \tag{12}$$

where two different parameters W_1 and W_2 from layer to layer are used, and a gaussian error linear unit (GELU) activation function is applied between layers [22]:

$$GELU(x) = x \cdot \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right] \approx 0.5x\left(1 + \tanh\left[\sqrt{2/\pi}\,(x + 0.044715x^3)\right]\right) \tag{13}$$

Both the class and the distillation embeddings of IDDAN are associated with linear condition classifier. The final health condition prediction result is determined by the addition of the *softmax* [23]outputs of the two routes:

$$Class_{predict} = Linear[softmax(O_{ce}) + softmax(O_{de})] \tag{14}$$

where O_{ce} and O_{de} are the output of the two classifiers respectively.

C. Knowledge Distillation Strategy

Backbone network's distillation process is carried out simultaneously with pre-training. In the output of the *softmax* layer, in addition to positive examples, negative labels also carry a lot of information. This training method of KD makes each sample bring more information to student model than the traditional training method.

Firstly, in order to improve the recognition accuracy of health conditions, the teacher model is chosen as VGG-16. It is a strong feature extractor and classifier. In this distillation strategy, the hard decisions of the teacher model $y_{\mathrm{t}} = \mathrm{argmax}_c Z_{\mathrm{t}}(c)$ (Z_{t} is the logits of the teacher model) are true labels. The applied hard-label distillation objective should be defined as:

$$\mathcal{L}_{global}^{HardDistill} = \frac{1}{2}\mathcal{L}_{\mathrm{CE}}(\psi(Z_{\mathrm{s}}), y) + \frac{1}{2}\mathcal{L}_{\mathrm{CE}}(\psi(Z_{\mathrm{s}}), y_{\mathrm{t}}) \tag{15}$$

where $\mathcal{L}_{\mathrm{CE}}$ represents the cross-entropy, and Z_{s} is the logits of the student model (IDDAN). For the image represents of machine signal, it is possible that hard labels will change according to the DA (Section III-*A* in this paper). In this hard-label distillation loss function, the teacher prediction y_{t} has the same role with the true label y.

Fig. 6 demonstrates that the distillation token, the class token, and patch tokens interact through the self-attention layers in the transformer encoder. The output of the distillation token is the hard-label predicted by the teacher model. In this method, the class token $\mathbf{w}_{class}$ and the distillation token $\mathbf{w}_{distill}$ are trained by back-propagation algorithm:

$$\mathbf{w}(m+1) = \mathbf{w}(m) - \eta\frac{\partial J(\mathbf{w}(m))}{\partial \mathbf{w}(m)} \tag{16}$$

$J(\mathbf{w})$ represents the training error at any instance, m denotes the number of iterations, and $\eta > 0$ is the pre-set learning rate before training. This paper employs the gradient-based parameter optimizer AdamW, replacing L2 regularization of Adam with weight decay.

D. Transfer Inference Objectives

The transfer inference stage of IDDAN includes fine-tuning and condition classification. Fine-tuning is a pivotal step to infer the pre-trained model to the target task through the transfer learning method. Condition classification is to use the inference completed mode to classify and diagnose fault conditions.

The backbone network can be divided into feature extractors and classifiers. The feature extractor extracts the low-level features of the image. In the pre-training stage, the proposed vibration signal data augmentation strategy can obtain large-scale pre-training samples of the source domain, and the pre-model trained with large-scale data has a higher generalization ability to extract the underlying features. Therefore, during the transfer process, the bottom layer weights are frozen, and the high layer weights are opened. In this paper, the FFN layers and linear classifier layer in the pre-trained IDDAN will be updated with parameters in fine-tuning (has been demonstrated in Fig. 4). This is because the previous self-attention layers and layer normalization layer are used to obtain a general representation from the image, and the latter FFN layer and linear classifier are more relevant to downstream special fault diagnosis tasks.

IV. Case Study and Result

To verify the approach proposed in section III, this paper uses three professionally measured bearing datasets.

A. Experiment Data and Description

1) Dataset A: KAt-DataCenter bearing dataset [1] contains a variety of faults to perform fault diagnosis experiments. This dataset focuses on not only artificial bearing damages but also real damages. It could prove better than the proposed approach in this paper is competent for different kinds of bearing fault diagnosis. The tested motor is a 425W permanent magnet synchronous motor (PMSM) which has the nominal torque $T = 1.35\ Nm$, the nominal speed $n = 3000\ rpm$, the nominal current $I = 2.3\ A$ and the number of pole pair $p = 4$.
2) Dataset B: The IMS bearing dataset is measured by provided by the Center for Intelligent Maintenance Systems (IMS), University of Cincinnati [24]. Recorded vibration signals include normal condition, rolling element fault, inner race fault, outer race fault. Each data describes a test-to-failure experiment and consists of individual files that are 1-second vibration signal snapshots recorded at specific intervals. Each file consists of 20,480 points with the sampling rate set at 20 kHz and collected by NI DAQ Card 6062E.
3) Dataset C: Experiments of the CWRU bearing dataset were conducted using a 2 hp Reliance electric motor, and acceleration data were measured at locations near to and remote from the motor bearings [25]. Motor bearings were seeded with faults using electro-discharge machining. Faults ranging from 0.007 inches in diameter to 0.040 inches in diameter were introduced separately at the rolling element fault, inner race fault, and outer race fault. Vibration data was collected at 12,000 samples per second, and data was also collected at 48,000 samples per second for drive end bearing faults. Speed and horsepower data were collected using the torque transducer.

These three bearing datasets include vibration signals while they are obtained from different machines and different operation conditions. Their detailed information is displayed in Table I.

Table I

Detailed Information of Various Experiment Bearing Datasets

Dataset names	Bearings	Conditions	Speed (rpm)	Load conditions
KAt	PMSM bearing	Normal	1500	0.1Nm, 0.7Nm
		Inner ring fault	1500	0.1Nm, 0.7Nm
		Outer ring fault	1500	0.1Nm, 0.7Nm

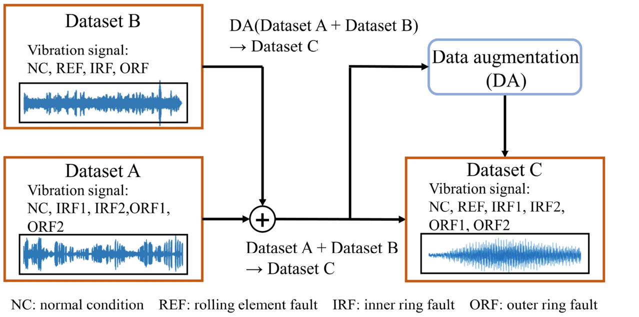
IMS	Rexnord ZA-2115 bearing	Normal	2000	6000 lbs
		Inner ring fault	2000	6000 lbs
		Rolling element fault	2000	6000 lbs
		Outer ring fault	2000	6000 lbs
CWRU	Motor bearing	Normal	1750	0, 1,2,3 HP
		Inner ring fault	1750	0, 1,2,3 HP
		Rolling element fault	1750	0, 1,2,3 HP
		Outer ring fault	1750	0, 1,2,3 HP

B. Transfer Multiple Fault Diagnosis of the IDDAN

1) Experiments Setting

The planned transfer diagnosis experiment is shown in Fig. 7, which includes the use of each dataset and the partition of training data and test data. In this experiment, all used data are bearing vibration signals collected from different machines or devices. Therefore, this experiment put the target on testing the ability and performance of inferencing trained IDDAN to a new machine. It can be found that the dataset A and B are mixed as pretraining data. This is because the dataset A does not include vibration signal samples of rolling element fault. In No. 2 experiments, samples of two damage levels are also not included in the dataset B. No.1 experiment tests the accuracy of the proposed method with standard data amount (2000) which come from two different datasets, while No. 2 experiment tests with a larger amount of data (40000).

In this paper, the percentage of the dataset means that the data is picked evenly from each fault and each working condition in the dataset. 20% of the target domain dataset is used in this section as the fine-tuning dataset.



No.	IDDTN transfer experiments	Training data	Test data
1	Dataset A + Dataset B → Dataset C	100% labeled Dataset A + 100% labeled Dataset B + 20% labeled Dataset C	50% unlabeled Dataset C
2	**DA**(Dataset A + Dataset B) → Dataset C	DA(100% labeled Dataset A + 100% labeled Dataset B) + 20% labeled Dataset C	50% unlabeled Dataset C

Fig.7. Diagram of different transfer diagnosis experiments.

2) Validation

We evaluate the proposed IDDAN through two designed experiments shown in Fig. 7. In the first experiment, B refers to the source domain dataset and C refers to the target domain dataset. In the second experiment, A plus B turn into the source domain data, and C also refers to the target domain dataset. In each experiment, the training data includes all the labeled samples from the source domain dataset and the fine-turning data uses 20% labeled data from the target domain dataset. Then, we randomly take 50% of unlabeled samples as the test data.

The detailed information of training parameters is demonstrated as follows. The CPU and GPU devices for training are I9-12900K and RTX 3080Ti respectively. The epoch for pre-training and fine-tuning is set to be 500 and 150 respectively. In addition, the patch size is set as 32, the number of attention heads is 6. This step is trained using the AdamW optimizer with a learning rate 3×10^{-4}. In the fine-tuning step, the dimension of the MLP block is set as 2048, the resolution of target domain time-frequency maps increases to 256 from 224. The training loss of experiments is plotted in Fig. 8. As shown in Fig. 8, the initial loss in the fine-tuning stage is significantly lower than in the pre-training stage.

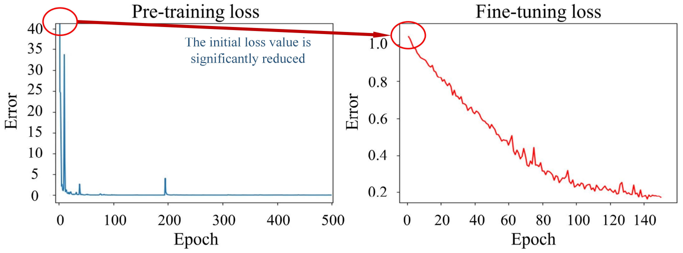


Fig. 8. The pre-training loss (500 epoch) and the fine-tuning loss (150 epoch) carve.

The No.1 experiment is repeated ten times with those pre-set parameters. In each experiment, all accuracies of the transfer diagnosis are over 82% and the average accuracy is around 85%. For obtaining a baseline testing accuracy, we set a control transfer test using the CNN model to condition recognition. In this CNN module, the number of convolutional layers is 5, and the size of the convolution kernel is set to be 3. The 5 pooling layers are followed by each convolutional layer separately and the size of the pooling kernel is 2. The CNN model is also pre-trained and fine-tuned using the same data samples. According to the same experiments times with IDDAN, the average accuracy is around 89% and the lowest accuracy is 85%. It means that the proposed IDDAN method can effectively diagnose the normal condition and three faults of bearing, but the accuracy of the CNN-based model is slightly higher than IDDAN with a small number of pre-training samples.

C. Fault diagnosis analysis of the proposed method

1) Effect Analysis of Data Augmentation

Due to the difference between the self-attention mechanism and the convolutional network, self-attention-based network structures benefit more from large-scale data [19]. The experiments of this part take the same settings as Fig. 7 and the data augmentation strategy in Section III-A is applied in the experiment. We use MATLAB (signal processing toolbox) to perform the proposed data augmentation method on all vibration signals in the source domain dataset, followed by the signal-to-image conversion. The samples used in the No. 2 experiment were expanded from 2000 to 40000, reaching 20 times the original pre-training samples.

In this section, the No.2 experiments in Fig. 7 should be repeated several times. In these experiments, the amount of data used in the fine-tuning stage as a percentage of the total target domain data (P_T) is kept as 20%. It can be found that all accuracies of the transfer diagnosis are over 92% and the average accuracy is around 95%. To provide visual insights into the effects of features transferring from the source datasets and target dataset, we use the *t*-distributed stochastic neighbor embedding (*t*-SNE) technique to map the high-dimensional features into a 2-D space. Fig. 9 demonstrates the feature recognition ability of the last layers before *softmax* layer when we completed the fine-tuning step.

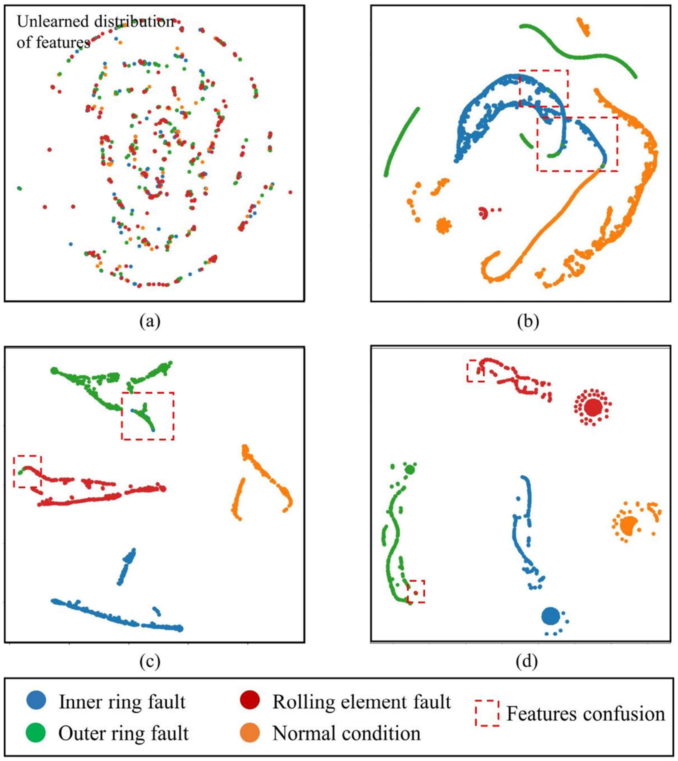


Fig. 9. t-SNE visualization of features after fine-tuning (P_T=20%). (a) Before training. (b) Small amount of pre-training samples. (c) CNN method with large amount of pre-training samples. (d) IDDAN with large amount of pre-training samples.

2) FLOPs and Parameters Analysis of KD

When deploying the model in a device such as an electric vehicle (EV), the floating point operations (FLOPs) of the method must be considered. FLOPs describe the computing power required by the deployed device, and the number of parameters describes the required memory size.

In this paper, the KD strategy is used to further reduce the model parameters and required FLOPs, reducing the burden on equipment and memory. Fig. 10 illustrates FLOPs of IDDAN, which compares the number of model parameters and FLOPs between IDDAN and common deep learning frameworks.

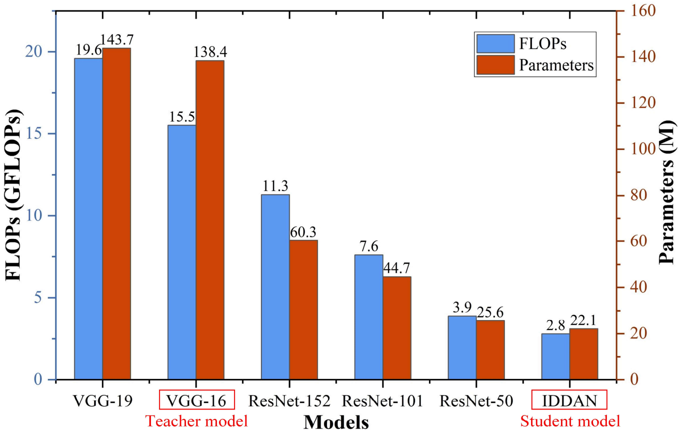


Fig. 10. t-SNE visualization of features after fine-tuning (P_T=20%).

As shown in Fig. 10, the IDDAN achieves a relatively minimum requirement for FOLPs and parameter quantities compared to some commonly used frameworks. In detail, the required FLOPs of IDDAN are 2.8GFLOPs and the number of parameters is 22.1M, which is obviously less than the teacher model VGG-16.

D. Test Results comparison

1) Multiple methods comparison

The dataset of bearing vibration signals under various working conditions is shown in Table II. Therefore, a very harsh transfer learning test environment is formed, which can well detect the performance of the proposed method. However, the number of labeled samples provided by the target domain is also the key to performance. We divide nine grades according to P_T from less to more: 1%, 5%, 10%, 15%, 20%, 25%, 30%, 40% and 50%. The test data is still 50% unlabeled data of the target domain. It is necessary to compare the performance of the different methods. In this comparison, the historically outperforming classical algorithms support vector machine (SVM) and CNN are replaced with IDDAN, and short-time Fourier transform (STFT) and Hilbert-Huang transform (HHT) will be replaced with continuous wavelet transform (CWT). Data augmentation (DA) is also added as a controllable condition. Results of all tests are summarized in Table II, which shows that the proposed method can achieve the best performance if $P_T \geq 10\%$.

Table II
Fault Diagnosis Results of Various Methods

P_T / Methods	1%	5%	10%	15%	20%	25%	30%	40%	50%
STFT+SVM	69.8%	73.9%	75.5%	77.2%	78.8%	82.7%	83.2%	87.3%	88.6%
STFT+CNN	67.9%	71.7%	78.4%	79.7%	85.3%	90.9%	91.0%	91.5%	92.4%
STFT+IDDAN	66.4%	66.8%	77.6%	78.1%	80.7%	84.4%	86.6%	93.2%	92.6%
HHT+SVM	**71.6%**	74.0%	80.9%	82.5%	80.4%	83.2%	81.5%	84.1%	89.8%
HHT+CNN	66.9%	75.5%	80.2%	84.6%	87.9%	91.3%	91.9%	90.7%	92.7%
HHT+IDDAN	60.1%	64.9%	77.1%	79.4%	84.2%	85.3%	90.6%	92.5%	93.6%
CWT+SVM	71.0%	79.4%	82.2%	83.7%	86.1%	87.8%	88.9%	90.5%	90.4%
CWT+CNN	69.9%	75.9%	80.3%	82.2%	89.5%	92.3%	92.2%	93.2%	93.9%
CWT+CNN+DA	71.4%	**79.3%**	86.7%	90.9%	94.1%	95.6%	95.7%	96.1%	96.2%
CWT+IDDAN	67.5%	69.2%	76.0%	80.5%	84.4%	87.9%	92.2%	92.7%	93.2%
CWT+IDDAN+DA	70.4%	85.5%	**87.2%**	**92.2%**	**95.9%**	**96.7%**	**97.3%**	**99.0%**	**99.5%**

2) Comparison with other methods

To demonstrate the performance of the proposed IDDAN, three different existing bearing diagnosis methods are used for comparison. Table III shows the comparison of the diagnostic results of other methods collected from the paper with the proposed IDDAN in transfer fault diagnosis experiments (P_T=50%). All methods are tested on CWRU dataset.

Table III
Fault diagnosis results compared with other existing methods (P_T=50%)

Methods	**Transfer method**	**Target domain dataset**	**Average accuracy**
CNN	Fine-tuning	CWRU	93.9%
CNN-DA	Fine-tuning	CWRU	96.2%
DDC [26]	Maximum mean discrepancy	CWRU	78.2%
DCTLN [10]	Multiple domain adaptation	CWRU	86.8%
DANN [27]	Domain adversarial	CWRU	80.9%
IDDAN	Fine-tuning	CWRU	**93.2%**
IDDAN-DA	Fine-tuning	CWRU	**99.5%**

The results demonstrate that fine-tune-based methods outperform all compared methods. The accuracy of the CNN model is better than IDDAN in experiments without DA. However, IDDAN overtakes the CNN model after going through DA supporting. According to application conditions, it could be divided into two categories:

1) There is not any labeled data collected from the target domain machine. In this condition, domain adaptation-based methods are widely used for solving transfer learning problems. The most advanced research is based on the maximum mean discrepancy (MMD) of data samples. For example, the deep domain confusion (DDC) is to add an adaptation layer and an MMD module to the traditional CNN structure [26]. The domain adversarial training of neural networks (DANN) is the use of deep neural networks capable of domain discrimination [27]. The deep convolutional transfer learning networks (DCTLN) add two domain adaptation losses to the CNN loss function to optimize the MMD distance between the source domain and target domain [10]. However, this method will not reach the highest accuracy when diagnosing faults in the target domain.
2) There are a few labeled data collected from the target domain machine. Fine-tuning is currently widely used in computer vision to further adapt to the target domain after pre-training. And the accuracy of the model after fine-tuning is often related to the amount of pre-trained data. The application scenario that requires the target domain can provide a small amount of label data. At the same time, the self-attention mechanism included in IDDAN is proved to be more dependent on large-scale pre-training samples.

E. Multi-Level Fault Detection and Results

For further testing the ability of the proposed diagnostic method to identify fault features, we set two damage levels of inner ring fault and outer ring fault. In detail, the inner fault and the outer fault in CWRU bearing dataset are divided into 0.007 inches and 0.021 inches, while the inner fault and outer fault in KAt bearing dataset are also subdivided into two severity levels: within 2mm, between 2mm and 4.5mm.

The experiments are repeated several times with the proposed data augmentation process in the pre-training stage. The result shows that faults recognition accuracies of the CNN-based transfer method and IDDAN are around 84% and 92% (P_T =50%). The t-SNE features visualizations are demonstrated in Fig. 11. According to the information from parts *C, D* and *E*, we can further compare and analyze the effectiveness of IDDAN and other methods. We can observe the following points:

1) Compared with the classical CNN based transfer learning method, IDDAN achieves higher classification accuracy when given enough pre-training data. This means that transfer learning-based diagnosis accuracy has been further improved. Its purpose is to accurately classify various faults of the target machine when the labeled data of the target machine is not enough to directly train the diagnosis model.
2) When classifying the inter ring fault and outer ring fault among the four fault types into two damage levels, the classification accuracy of the transfer learning-based diagnosis model decreases significantly. This is due to the high similarity between features of different damage levels. As can be intuitively seen in Fig. 11, the classification model produces confusion between different fault types and levels.

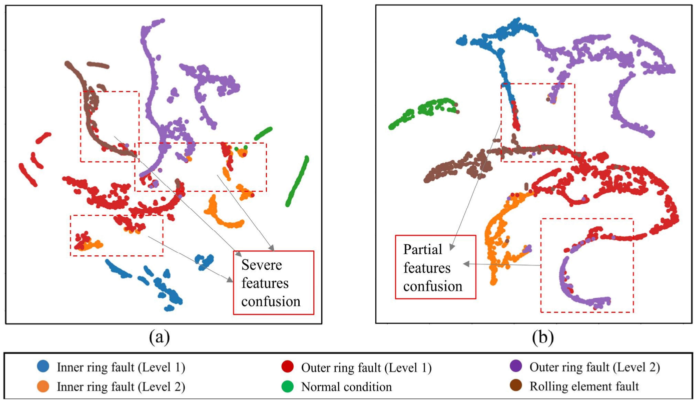


Fig. 11. t-SNE visualization of features after fine-tuning (P_T=50%). (a) CNN-based transfer method with large amount of pre-training samples. (b) IDDAN with large amount of pre-training samples.

V. Conclusion

This paper proposes the self-attention mechanism in the field of online fault diagnosis of motor bearing with higher accuracy and proposed a new diagnosis framework based on IDDAN for solving the problem that it is hard to obtain enough labeled data to train a diagnosis model for a new target machine. Our experiment results present that the fine-tune-based transfer learning method could get better accuracy on the same dataset and the IDDAN has a better performance by pre-training using large-scale data. The mentioned DA method provides sufficient pre-training samples for IDDAN. Meanwhile, when IDDAN consumes large-scale data for pre-training, its diagnostic accuracy could surpass the CNN-based transfer learning model. The following points could conclude from this paper:

1) The paper proposes a self-attention mechanism-based intelligent fault diagnosis method IDDAN for deploying on new machines with a small number of labeled data by transfer learning.
2) The proposed DA method in Section III-A effectively expands the number of pre-training samples and has an excellent effect on IDDAN.
3) The IDDAN obtain a higher recognition accuracy of multiple bearing fault conditions with a large amount of pre-training data than the classic CNN-based method.

References

[1] 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," *PHME_CONF,* vol. 3, no. 1, Jul. 2016.

[2] R. Duan, Y. Liao and S. Wang, "Adaptive Morphological Analysis Method and Its Application for Bearing Fault Diagnosis," *IEEE Transactions on Instrumentation and Measurement,* vol. 70, pp. 1-10, 2021.

[3] K. Zheng, T. Li, Z. Su and B. Zhang, "Sparse Elitist Group Lasso Denoising in Frequency Domain for Bearing Fault Diagnosis," *IEEE Transactions on Industrial Informatics,* vol. 17, no. 7, pp. 4681-4691, July 2021.

[4] Y. Qin, L. Jin, A. Zhang and B. He, "Rolling Bearing Fault Diagnosis With Adaptive Harmonic Kurtosis and Improved Bat Algorithm," *IEEE Transactions on Instrumentation and Measurement,* vol. 70, pp. 1-12, 2021.

[5] F. Dalvand, S. Dalvand, F. Sharafi and M. Pecht, "Current Noise Cancellation for Bearing Fault Diagnosis Using Time Shifting," *IEEE Transactions on Industrial Electronics,* vol. 64, no. 10, pp. 8138-8147, Oct. 2017.

[6] T. Wang, Z. Liu, G. Lu and J. Liu, "Temporal-Spatio Graph Based Spectrum Analysis for Bearing Fault Detection and Diagnosis," *IEEE Transactions on Industrial Electronics,* vol. 68, no. 3, pp. 2598-2607, March 2021.

[7] D. T. Hoang and H. J. Kang, "A Motor Current Signal-Based Bearing Fault Diagnosis Using Deep Learning and Information Fusion," *IEEE Transactions on Instrumentation and Measurement,* vol. 69, no. 6, pp. 3325-3333, June 2020.

[8] H. Shao, M. Xia, G. Han, Y. Zhang and J. Wan, "Intelligent Fault Diagnosis of Rotor-Bearing System Under Varying Working Conditions With Modified Transfer Convolutional Neural Network and Thermal Images," *IEEE Transactions on Industrial Informatics,* vol. 17, no. 5, pp. 3488-3496, May 2021.

[9] T. Lu, F. Yu, B. Han and J. Wang, "A Generic Intelligent Bearing Fault Diagnosis System Using Convolutional Neural Networks With Transfer Learning," *IEEE Access,* vol. 8, pp. 164807-164814, 2020.

[10] L. Guo, Y. Lei, S. Xing, T. Yan and N. Li, "Deep Convolutional Transfer Learning Network: A New Method for Intelligent Fault Diagnosis of Machines With Unlabeled Data," *IEEE Transactions on Industrial Electronics,* vol. 66, no. 9, pp. 7316-7325, Sept. 2019.

[11] M. Zhao, S. Zhong, X. Fu, B. Tang and M. Pecht, "Deep Residual Shrinkage Networks for Fault Diagnosis," *IEEE Transactions on Industrial Informatics,* vol. 16, no. 7, pp. 4681-4690, July 2020.

[12] S. Shao, S. McAleer, R. Yan and P. Baldi, "Highly Accurate Machine Fault Diagnosis Using Deep Transfer Learning," *IEEE Transactions on Industrial Informatics,* vol. 15, no. 4, pp. 2446-2455, April 2019.

[13] Y. Hou et al., "Bearing Fault Diagnosis Under Small Data Set Condition: A Bayesian Network Method With Transfer Learning for Parameter Estimation," *IEEE Access,* vol. 10, pp. 35768-35783, 2022.

[14] C. Liu and K. Gryllias, "Simulation-Driven Domain Adaptation for Rolling Element Bearing Fault Diagnosis," *IEEE Transactions on Industrial Informatics,* vol. 18, no. 9, pp. 5760-5770, Sept. 2022.

[15] Z. Shen and W. Guo, "An Intelligent Bearing Fault Diagnosis based on Modified Probabilistic Knowledge Distillation," in *2021 Global Reliability and Prognostics and Health Management (PHM-Nanjing)*, 2021.

[16] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate,," arXiv.org, 19 May 2016. [Online]. Available: https://arxiv.org/abs/1409.0473. [Accessed 13 Jul 2021].

[17] C. Blakeney, X. Li, Y. Yan and Z. Zong, "Parallel Blockwise Knowledge Distillation for Deep Neural Network Compression," *IEEE Transactions on Parallel and Distributed Systems,* vol. 32, no. 7, pp. 1765-1776, 1 July 2021.

[18] H. Mao, "A Survey on Self-supervised Pre-training for Sequential Transfer Learning in Neural Networks," arXiv.org, 2020. [Online]. Available: https://arxiv.org/abs/2007.00800. [Accessed 7 Sep 2021].

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale,," arXiv.org, 03 June 2021. [Online]. Available: https://arxiv.org/abs/2010.11929.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," arXiv.org, 06 Dec 2017. [Online]. Available: https://arxiv.org/abs/1706.03762v5. [Accessed 16 Jul 2021].

[21] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 21 Jul 2016. [Online]. Available: https://arxiv.org/abs/1607.06450.

[22] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," arXiv.org, 11 Nov 2018. [Online]. Available: https://arxiv.org/abs/1606.08415v3.

[23] Z. -X. Hu, Y. Wang, M. -F. Ge and J. Liu, "Data-Driven Fault Diagnosis Method Based on Compressed Sensing and Improved Multiscale Network," *IEEE Transactions on Industrial Electronics,* vol. 67, no. 4, pp. 3216-3225, April 2020.

[24] H. Qiu, J. Lee, J. Lin and G. Yu, "Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics," *Journal of Sound and Vibration,* vol. 289, no. 4, pp. 1066-1090, 2006.

[25] "Case Western Reserve University Bearing Data Center Website," [Online]. Available: https://engineering.case.edu/bearingdatacenter/.

[26] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," arXiv:1412.3474, 2014. [Online].

[27] Y. Ganin et al., "Domain-adversarial training of neural networks," *Journal of Machine Learning Research,* vol. 17, no. 59, pp. 1-35, 2016.