LiteFormer: a lightweight and efficient Transformer for rotating machine fault diagnosis

Wenjun Sun, Ruqiang Yan, *Fellow, IEEE*, Ruibing Jin, Jiawen Xu, Yuan Yang, and Zhenghua Chen, *Senior Member, IEEE*

***Abstract*—Transformer has shown impressive performance on global feature modeling in many applications. However, two drawbacks induced by its intrinsic architecture limit its application, especially in fault diagnosis. Firstly, the quadratic complexity of its self-attention scheme extremely increases the computation cost, which poses a challenge to apply Transformer to a computationally limited platform like an industry system. Additionally, the sequence-based modeling in the Transformer increases the training difficulty and requires a large-scale training dataset. This drawback becomes serious when Transformer is applied in fault diagnosis where only limited data is available. To mitigate these issues, we rethink this common approach and propose a new transformer, which is more suitable for fault diagnosis. In this paper, we first show that the attention module can be actually replaced with or even surpassed by a convolution layer under some conditions in mathematics and experiments. Then, we adopt the convolutions into the transformer, where the computation burden issue is alleviated and the fault classification accuracy is significantly improved. Furthermore, to increase the computation efficiency, a lightweight transformer called LiteFormer, is developed by utilizing the depth-wise convolutional layer. Extensive experiments are carried out on four datasets: CWRU, PU, and two Gearbox datasets of DDS. Through our experiments, our LiteFormer not only reduces the computation cost in model training, but also sets new state-of-the-art results, surpassing other counterparts in both fault classification accuracy and model robustness.**

***Index Terms*—Convolution, Efficient, Fault diagnosis, Lightweight, Transformer**

I. INTRODUCTION

IN industrial systems, the failure of rotating machinery may result in financial losses or even fatalities. To address this issue, rotating machinery health monitoring and fault diagnosis are widely studied [1-5], which is able to facilitate the detection of machine faults and make prompt repairs to avoid losses. Traditional machine fault diagnosis methods mostly use signal processing methods to extract hand-crafted features from fault signals [1-3]. Recently, deep learning (DL) algorithms [6-8] have achieved impressive performances in the field of machine fault diagnosis [9-10] benefitted from the superior automatic feature learning capabilities of deep models. The deep models learn data features with multiple layers, where the low layers learn the edge features of data and the deep layers model the semantic features of data [6]. Benefitted from the deep models, the representation of features is significantly improved and the deep models are widely explored in machine fault diagnosis for transfer learning [11-12], where the deep model is trained on the source dataset and then fine-tuned on the target dataset. To apply the DL methods for machine fault diagnosis with physical interpretability, some methods for interpreting deep models have been proposed [13-14]. Abid et al. [15] implemented a deep-SincNet for fault diagnosis to provide more physical interpretability. Li et al. [16] proposed a novel wavelet-driven deep neural network, which uses a continuous wavelet convolutional layer to replace the first convolutional layer of the standard CNN to obtain a customized kernel bank. Li et al. [17] introduced the attention mechanism to the deep network to locate the informative data segments and visualize the learned diagnosis knowledge. With the research on the interpretability of deep models, the potential of deep learning for intelligent machine fault diagnosis becomes more attractive and valuable.

Among the DL methods, the latest proposed Transformer [8], which is realized completely by the self-attention mechanism, has set off a new round of high tide in the field of natural language processing (NLP) [18] and computer vision (CV) [19-20]. Such success of Transformers has inspired some methods [21-22] proposed for machine fault diagnosis to learn the global features. Although Transformer has shown large model capacity in many tasks, the quadratic complexity of its self-attention scheme extremely increases the computation cost, which limits its applications in some computationally limited industry systems. Moreover, since the sequence-based modeling in Transformer lacks the inductive bias [19, 23-24], it is difficult to train a Transformer with limited training samples. These two drawbacks induced by the intrinsic architecture of Transformer limit its application in fault diagnosis.

Although lots of methods try to improve the attention module [20, 26] or introduce convolutions into Transformer [23-25] to improve the performance of Transformer in the field of NLP or CV, the complex network architecture makes

This work was supported in part by the National Natural Science Foundation of China (Grant No. 51835009) and in part by A*STAR under its AME Programmatic Funds (Grant No. A20H6b0151). *(Corresponding author: Ruqiang Yan and Ruibing Jin.)*

Wenjun Sun, Jiawen Xu and Yuan Yang are with the School of Instrument Science and Engineering at the Southeast University, Nanjing, Jiangsu, China. (e-mail: swjstudent@163.com, jiawen.xu@seu.edu.cn, and yangyuancsi@163.com)

Ruibing Jin and Zhenghua Chen are with the Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore 138632. (e-mail: jin_ruibing@i2r.a-star.edu.sg, chen0832@e.ntu.edu.sg)

Ruqiang Yan is with the School of Instrument Science and Engineering at the Southeast University, Nanjing, Jiangsu, China, and the School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China. (e-mail: ruqiang@seu.edu.cn)

it difficult to optimize on the small datasets in the field of fault diagnosis. Recently, some methods [21-22] propose to improve the original Transformer and devise some Transformer variants which are adapted to fault diagnosis. The existing methods [21-22, 27-28] mostly focus on the attention mechanism of Transformer, ignoring those drawbacks existing in Transformer. Compared with Transformer, convolutional neural networks (CNNs) have been widely investigated in fault diagnosis [29-31] and shown satisfactory performance benefiting from its intrinsic inductive bias and computation efficiency. CNNs are naturally equipped with the intrinsic inductive bias of locality and translation equivariance [32] even for small datasets [29]. Motivated by these characteristics of CNNs, we propose to integrate the advantages of convolution with the superiority of Transformer and devise a new Transformer called LiteFormer for fault diagnosis.

Firstly, we rethink the relationship between the self-attention module and the convolutional layer. We argue that this self-attention module in Transformer actually can be regarded as an enhanced convolution operation in mathematics. Since the self-attention module is more complex, it generally requires more training data than the convolutional layer. In the limited-data regime, the self-attention may perform equally or even inferior to a convolutional layer. To verify our hypothesis, some primary experiments are conducted, where we propose several convolution based Transformers called ConvFormer. In our primary experiments, our proposed ConvFormer performs better than the conventional Transformer. Moreover, in our proposed Convformer, the convolution operations alleviate the computation memory burden of Transformer and improve the fault classification accuracy under limited-sized datasets. Additionally, to further improve the computation efficiency, we propose a lightweight and efficient Transformer called LiteFormer for rotating machine fault diagnosis.

Our proposed LiteFormer adopts the fast and lightweight depth-wise convolutional layer to model the local spatial correlations of fault signals. The patch embedding in vision Transformers [19-20] is utilized to encode the time series into the token embedding. Then the token embeddings are forwarded to our LiteFormer blocks to capture the sequential information. Our LiteFormer block is stacked in multiple layers to effectively learn local and global information. Finally, the outputs of the last LiteFormer block are forwarded to a sequence pooling layer [24] to generate the weighted output representations, and then it is delivered to the multi-layer perceptron (MLP) head for fault classification.

The main contributions of the work can be summarized as follows:

1. Instead of focusing on improving the attention module, we analyze the relationship between the self-attention module and the convolutional layer in mathematics and experiments. Based on this analysis, we find the essence of Transformer and propose a ConvFormer, which inherits the inherent structural superiority of Transformer while alleviating the computation burden in conventional Transformers.

2. Original Transformer based approaches require much computation resources. Compared with them, a lightweight Transformer called LiteFormer is proposed for efficient rotating machine fault diagnosis. Our LiteFormer replaces the multi-head self-attention module with depth-wise convolution, which significantly reduces the computation cost and improves the classification accuracy. With our proposed LiteFormer, the applicability of Transformer is enhanced.

3. Comprehensive experiments are carried out on four fault datasets of three simulation test rigs. According to experimental results, our proposed LiteFormer is more efficient and effective than Transformers. The results also demonstrate the superior classification performance and strong robustness of our proposed LiteFormer, which sets state-of-the-art results.

This paper is organized as follows. In Section II, the previous and related fault diagnosis works are introduced. Then, in Section III, the relationship between self-attention and convolution is analyzed. In Section IV, the framework of the proposed LiteFormer for rotating machine fault diagnosis is provided. In Section V, experimental results and analysis are provided. At last, conclusions are presented in Section VI.

II. Related Works

In fault diagnosis, the DL based methods especially CNN methods [29-31], have been proven to be more effective than the traditional approaches that rely on signal processing methods.

Jiang et al. [10] proposed stacked multilevel-denoising auto-encoders (SMLDAEs) to learn robust and discriminative features from the complex frequency spectra for wind turbine gearbox fault diagnosis. Zhao et al. [33] used handcrafted features as local features for gated recurrent unit (GRU) networks for machine health monitoring. These are methods using deep neural networks (DNNs) or recurrent neural networks (RNNs) for machine fault diagnosis. However, they are not as widely used as those based on CNNs.

Zhang et al. [30] used wide first-layer kernels in a deep convolutional neural network (WDCNN) to extract robust features. Ding et al. [34] adopted a deep ConvNet based on wavelet packet energy (WPE) image for spindle bearing fault diagnosis. Jiang et al. [35] proposed a multi-scale convolutional neural network (MSCNN), which conducts multiple pairs of convolutional layers to extract multiscale features for fault diagnosis. Zhang et al. [31] proposed a method with the deep residual learning algorithm [32] for rotating machinery fault diagnosis.

Since the achievements of Transformers [18-20] have been remarked in NLP and CV field, several approaches [21-22, 36] are proposed to introduce Transformer into the field of machine fault diagnosis. Ding et al. [21] proposed a time-frequency Transformer (TFT) based on the original ViT [19] for fault diagnosis of rolling bearings. Pei et al. [22] proposed a Transformer convolution network (TCN) based on transfer learning for machine fault diagnosis. Fang et al. [36] proposed a CLFormer adopting convolutional embedding and linear self-attention for bearing fault diagnosis. However, most of these methods simply adopt off-the-shelf Transformers like ViT, which are not suitable for the rotating machine fault diagnosis. To solve this issue, we investigate the essence of Transformer and propose a suitable architectural alternative,

LiteFormer for the rotating machine fault diagnosis.

III. Analysis on Self-attention and Convolution

To effectively investigate the essential architecture of Transformer and integrate the convolutional layer with Transformer, we first analyze the relationship between self-attention and convolution. Then, we propose ConvFormer and conduct some primary experiments to verify the conclusions of our mathematical analysis.

A. Revisiting Self-attention and Convolution

The convolution operation has been widely used in many tasks [32, 37] and various types of convolutional layers [37-39] are proposed. In general, a convolutional operation can be defined as:

$$O_c(i)=b(i)+\sum_{\sigma\in\Omega}\mathbf{W}(i,\sigma)\cdot\mathbf{X}(i+\sigma)\,, \tag{1}$$

where b is the bias, Ω denotes the kernel size, and i indicates the index of an input. For an image, i indicates the location in a two-dimensional space. For time series data, i represents the temporal index. When the function $\mathbf{W}$ is independent of the index i, the convolutional operation is equal to a conventionally convolutional layer. When $\mathbf{W}$ is a function of both i and σ, this convolution operation represents some complex types of layers, such as deformable convolution [38] and dynamic convolution [39]. When the bias b is set as zero, this operation can be formulated as:

$$O_c(i)=\sum_{\sigma\in\Omega}\mathbf{W}(i,\sigma)\cdot\mathbf{X}(i+\sigma)\ . \tag{2}$$

According to Ref. [8], a self-attention module can be formulated as:

$$Atten=F(\mathbf{Q},\mathbf{K},\mathbf{V})=soft\max(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V}\,, \tag{3}$$

where $\mathbf{Q}\in\mathrm{R}^{d_{in}\times d_q}$, $\mathbf{K}\in\mathrm{R}^{d_{in}\times d_k}$ and $\mathbf{V}\in\mathrm{R}^{d_{in}\times d_v}$ represents three matrices, which are computed following $\mathbf{Q}=\mathbf{W}_q(\mathbf{X})$, $\mathbf{K}=\mathbf{W}_k(\mathbf{X})$, and $\mathbf{V}=\mathbf{W}_v(\mathbf{X})$, respectively.

Since the definition of convolution in (2) is expressed in a dot-production manner, to clearly compare the self-attention with the convolution, we convert the self-attention defined in (3) into a dot-production manner. Since d_k is a constant, the self-attention in a dot production manner can be formulated as:

$$Atten(i)=\sum_{\sigma\in\psi}\overline{\mathbf{W}}_q(i)\cdot\overline{\mathbf{W}}_k(i+\sigma)\cdot\mathbf{W}_v(i+\sigma)\,, \tag{4}$$

where $\overline{\mathbf{W}}_q$ and $\overline{\mathbf{W}}_k$ represent the normalized function by *softmax* and $\sqrt{d_k}$, ψ indicates the size of the input. Eq. 4 can be further simplified as:

$$Atten(i)=\sum_{\sigma\in\psi}\mathbf{W}_a(i,\sigma)\cdot\mathbf{W}_v(i+\sigma)\,, \tag{5}$$

where $\mathbf{W}_a(i,\sigma)=\overline{\mathbf{W}}_q(i)\cdot\overline{\mathbf{W}}_k(i+\sigma)$.

According to (2) and (5), we can find that the definition of convolution is actually in the same format as the definition of self-attention.

Comparing (2) with (5), we find that there are two different points between the convolution and self-attention: a. The Ω in (2) indicates the kernel size, which includes part of an input. The ψ in (5) represents the whole size of an input. b. In the self-attention defined in (5), $\mathbf{W}_a$ and $\mathbf{W}_v$ are more complex than the corresponding parts in (2).

Based on our analysis, we can find that when the receptive field of a self-attention, ψ is equal to Ω in (2), and the $\mathbf{W}_a$ and $\mathbf{W}_v$ are simplified, the self-attention can be replaced with a convolution layer. Furthermore, through the comparison between (2) with (5), we can explain why the Transformer performs better than a CNN. Firstly, the self-attention in Transformer has a large receptive field than a convolutional layer, which enables the Transformer to fully capture the global information. Additionally, the complex function $\mathbf{W}_a$ and $\mathbf{W}_v$ in (5) help the Transformer to learn the high-order features, which further improves the performance of the Transformer. Hence, the self-attention module actually can be regarded as an enhanced convolutional layer.

However, this complex structure of self-attention requires a large scale of training data and may easily overfit to a limited-data regime. Since the training data in fault diagnosis is much limited, it is challenging to fully train a Transformer and not degenerate its performance.

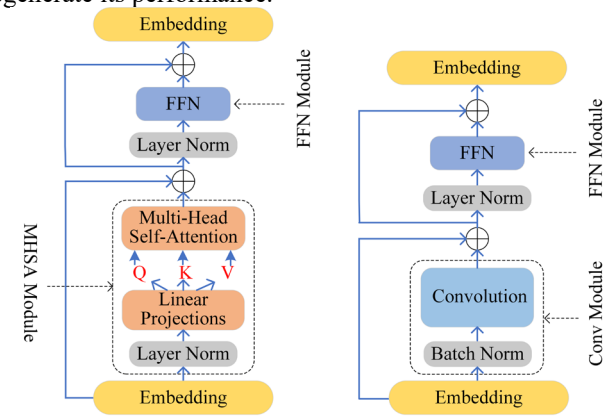

Fig. 1. (a)The architecture of the original Transformer block; (b)The architecture of the proposed ConvFormer block.

B. Primary Experiments

In Sec. 3.1, through our analysis of the relationship between the convolutional layer and the self-attention module, we find that the self-attention actually can be regarded as an enhanced convolution layer with an infinite receptive field. However, this complex self-attention module requires large-scale training data and cannot be fully trained in the limited-data regime like the fault diagnosis task.

In this subsection, to verify our conclusion above, several primary experiments are conducted, where we propose two types of convolution based Transformer (ConvFormer). To solve the issues in Transformer, our ConvFormer replaces the self-attention module with convolutional layers. Our ConvFormer and original Tranformer are illustrated in Fig. 1.

The primary experiments are conducted to verify that the MHAS module in Transformer can be equal to or even surpassed by the convolutional layer for machine fault diagnosis. To avoid the increasing difficulty of optimization caused by introducing more parameters, we further use the

TABLE I
THE RESULTS OF PRIMARY EXPERIMENTS

Model	Accuracy(%)±Std	Params (M)	FLOPs (M)	Training GPU Memory (MB)	Test GPU Memory (MB)
Global Transformer	99.64±0.14	0.267	106.513	735	77
Local Transformer	99.82±0.11	0.261	67.994	236	29
ConvFormer(*k*=3)	**99.92±0.08**	**0.230**	**58.819**	**150**	**25**
ConvFormer(*k*=3) (dilated convolution)	**99.97±0.05**	**0.230**	**58.819**	**150**	**25**

dilated convolution [40] instead of standard convolution in ConvFormer block to demonstrate the performance of ConvFormer can be improved by increasing the receptive field. We perform the experiments on the rolling bearing dataset provided by the Bearing Data Center of Case Western Reserve University (CWRU) [41], which is a classic dataset for rotating machine fault diagnosis [22, 30-31, 34]. Detailed dataset division is introduced in the section of the experimental setup. For a more convenient display of the effects, the primary experiments are all performed on the Transformer architecture of 5 depth and the kernel size of the convolutional layer is 3. The training and the test GPU memory are measured by the maximal GPU memory consumption of the model with a batch size of 32 in the model training and test procedure respectively.

Table I shows the comparison results between the self-attention based Transformer and the proposed ConvFormer (*k* indicates the kernel size). As seen in Table I, both MHSA module based Transformer (Global Transformer) and the window MHSA [20] based Transformer (Local Transformer) are surpassed by the convolution based Transformer (ConvFormer) on CWRU dataset for fault diagnosis. The ConvFormer with standard convolution operations obtains 99.92% accuracy, which is much higher than the Global Transformer of 99.64% and the Local Transformer of 99.82%. Since the convolution has the inductive bias, the performance of ConvFormer is significantly improved. The ConvFormer with the dilated convolution, which increases the receptive field without introducing external parameters, has improved the classification accuracy to 99.97%, showing that the bigger receptive field can improve the performance of ConvFormer under equal conditions. Moreover, it can be seen that the convolution operations in the proposed ConvFormer alleviate nearly 80% GPU memory burden of Global Transformer in the training.

The results verify that the self-attention module in Transformer can be viewed as an enhanced convolutional layer with an infinite receptive field. And the convolution operations alleviate the computation memory burden of Transformer and improve the classification accuracy under limited-sized fault datasets. The strong modeling capacity of Transformer may not mainly rely on the self-attention module, but more relies on the inherent architecture design of Transformer, so that the proposed ConvFormer using convolution instead of the self-attention can be more effective than the original Transformer for fault classification.

Although the ConvFormer has an excellent performance in the fault classification task, the computation complexity of convolutions is increasing largely with large kernels. To further improve the computation efficiency, we further propose the LiteFormer in the next section.

IV. THE PROPOSED LITEFORMER FRAMEWORK

Two drawbacks in computation cost and inductive bias of Transformer limit its application in fault diagnosis. Previously, we have analyzed that the convolutional layer can replace the self-attention module in Transformer and alleviate the above drawbacks. To further improve the computation efficiency even under a large receptive field, the LiteFormer approach is proposed in this section.

A. The Architecture of LiteFormer

LiteFormer consists of a patch embedding layer, L LiteFormer blocks, a sequence pooling layer, and a MLP head for classification. The architecture of the proposed LiteFormer can be seen in Fig. 2.

For the given input sequence $N \times 1$, the patch embedding layer first divides it into $\frac{N}{P}$ patches and then linearly projects the patches into an embedded dimension of size C. After that, the embedded patch tokens are delivered to the LiteFormer block with stacked L blocks, and the output feature is the same size as the input embedded patch tokens of size $\frac{N}{P} \times C$. Then the output is weighted by the sequence pooling layer [24] to generate the weighted output representations. Finally, a general MLP head is connected for classification.

B. LiteFormer Block

As illustrated in Fig. 2, the LiteFormer block, which represents the encoder layers in Transformer, contains the depth-wise convolution (DConv) module and the feed-forward network (FFN) module.

In Sec.III, we have verified that the self-attention module in Transformer can be viewed as an enhanced convolutional layer with a global receptive field and can be replaced by a convolutional layer. In our proposed LiteFormer, we utilize the depth-wise convolution with a large kernel size to replace the MHSA module, further improving the computation efficiency compared to the standard convolution. The LiteFormer block not only has the inductive bias of convolution, but also inherits the structural superiority of Transformer.

As in the original Transformer block, the residual connection is also employed around each DConv module and

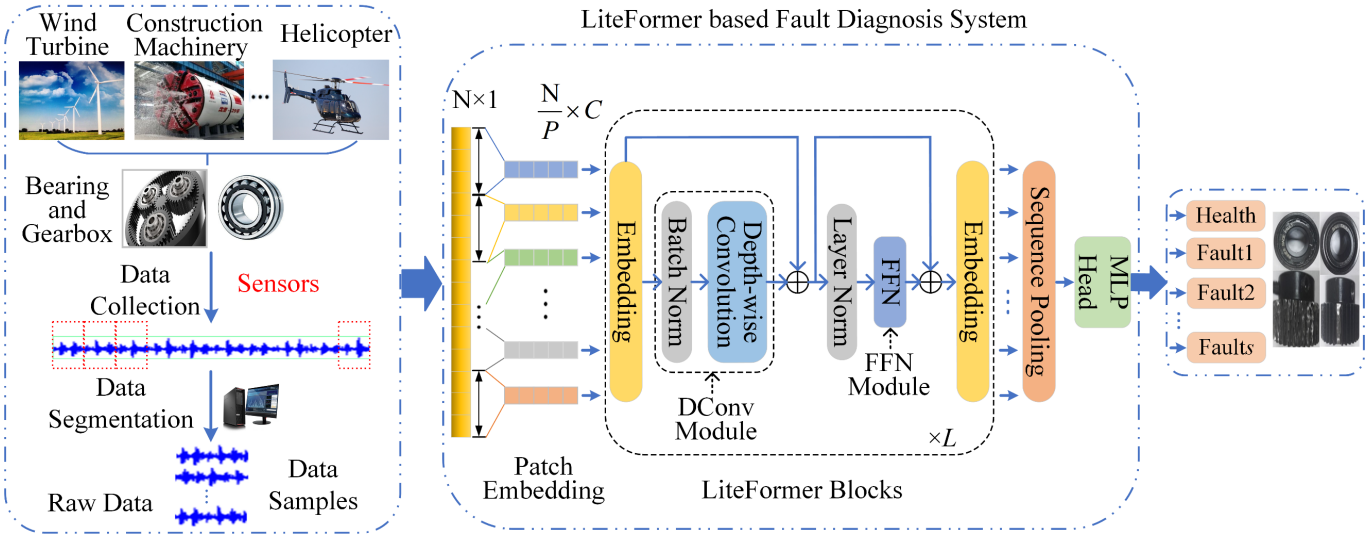

Fig. 2. The flowchart of LiteFormer based fault diagnosis system

FFN module. It means that for input x_i in LiteFormer block of block i, the output y_i of the block can be given as:

$$\begin{aligned} x_i^* &= x_i + \text{DepthwiseConv}(\text{BN}(x_i)) \\ x_i^* &= \text{Reshape}(x_i^*) \\ y_i &= x_i^* + \text{FFN}(\text{LN}(x_i^*)) \\ y_i &= \text{Reshape}(y_i) \end{aligned} \tag{6}$$

Here, DepthwiseConv(·) denotes the depth-wise convolution, BN(·) is the batch normalization (BN), LN(·) represents the layer normalization (LN), and Reshape(·) is the operation to change the shape of input.

1) DConv Module

In our DConv module, a single 1-D depth-wise convolution layer is designed to model the spatial features after batch normalization. Depth-wise convolution [42] is a fast and efficient variant of the standard convolution. Given an input embedding x, $x \in R^{n \times c}$, it performs a convolution independently over each channel. The computation complexity can be reduced from $O(k \cdot n \cdot c^2)$ to $O(k \cdot n \cdot c)$ where k is the kernel size. The depth-wise convolution is defined as:

$$U_{i,d} = \text{DepthwiseConv}(x, W_{d,:}, i, d) = \sum_{j=1}^{k} W_{d,j} \cdot x_{(i+j-\lceil \frac{k+1}{2} \rceil), d} \tag{7}$$

where $W \in R^{c \times k}$ are the learnable parameters. The output $U \in R^{n \times c}$, and $U_{i,d}$ is the output for the element i and channel d. The DConv module significantly decreases the complexity of Transformer, resulting in a lightweight LiteFormer.

2) Feed-Forward Module

After the DConv module with residual connection, a FFN module with residual connection is integrated. FFN module contains two fully connected layers, with the middle layer activated by a rectified linear unit (ReLU). There is an expansion ratio f_i set in FFN to control the dimension of the inner layer. The FFN performs point-wise operations to mix features in the channel dimension.

C. Framework for Fault Diagnosis

Our LiteFormer is proposed for efficient rotating machine fault diagnosis. The LiteFormer learns features from raw sensor data and finishes classification tasks in an end-to-end manner.

TABLE II
THE DETAILED CONFIGURATION OF LITEFORMER MODEL

Input Size	Output Size	Layer Name	LiteFormer
$N \times 1$	$\frac{N}{8} \times 64$	Patch Embedding	P=8; C=64 (proj): Conv1d(1, C, P, P, bias=False)
$\frac{N}{8} \times 64$	$\frac{N}{8} \times 64$	LiteFormer Blocks (depth L=7)	DConv: k=16 (norm): BatchNorm1d(64) (dconv): Conv1d(64, 64, k, 1, groups=64, bias=False) FFN: f=4 (norm): LayerNorm(64) (fc1): Linear(64, f*64) (act): ReLU() (fc2): Linear(f*64, 64) (drop): Dropout(p=0.2)
$\frac{N}{8} \times 64$	64	Sequence Pooling	(pool): Linear(64, 1)
64	s	MLP Head	(head): Linear(64, s)

Firstly, the vibration data sample is collected and then divided into patches. The patches are transformed to the token embeddings of C dimension through the patch embedding layer. Then, the token embeddings are forwarded to LiteFormer blocks for feature encoding. The outputs of the final block are forwarded to a sequence pooling layer to weigh the output tokens. Finally, the weighted representation is delivered to a MLP head for fault classification of s classes. The cross-entropy loss is used for optimizing the LiteFormer model. Dropout with a 0.2 rate is applied. The detailed

configuration of LiteFormer is provided in Table II.

The LiteFormer based machine fault diagnosis system is presented in Fig. 2. The raw vibration data are utilized as input and its condition labels are served as output in the training. The LiteFormer model is optimized through the optimizer Adam.

V. Experiment Analysis

A. Experimental Setup

To evaluate the efficiency of our proposed LiteFormer for rotating machine fault diagnosis, four experimental datasets are investigated. One is the CWRU dataset [41]. The second one is the bearing dataset from the Paderborn University Bearing Data Center [43-44], which is called PU dataset. The third one is the planetary gearbox dataset acquired from the drivetrain dynamic simulator (DDS) and it is called Gearbox A dataset. The last one is the parallel gearbox vibration dataset also acquired from the DDS and it is called Gearbox B dataset. The test rig of DDS is shown in Fig. 3. The two-stage planetary gearbox is connected to the driving motor, and its rotating speed is reduced by the parallel gearbox. The data length of each sample in these four datasets is all 2048.

The bearing data of CWRU dataset is measured by acceleration transducers from the drive-end bearings at a sampling frequency of 12 kHz under four operational conditions (load 0, 1, 2, and 3 hp). The rotating speed changes between 1730 and 1797 rpm based on the applied load. Single point faults with fault diameters of 0.007, 0.014, and 0.021 are set on the rolling element, the inner raceway, and the outer raceway, respectively. The CWRU dataset chooses 100 samples (50 for training and 50 for testing) for each condition under four loads. Thus, there are 10 different working conditions under the four loads. There are 2000 training samples and 2000 testing samples in total.

The bearing data of PU dataset is measured by a piezoelectric accelerometer at the top end of the rolling bearing module with a sampling frequency of 64 kHz. PU datasets consist of 32 sets of current signals and vibration signals, caused by bearings that include six undamaged bearings, twelve artificially damaged bearings, and fourteen bearings with real damages caused by accelerated lifetime tests. Each set of signals is collected under four working conditions. In this paper, the vibration signals of 13 bearings (KA04, KA15, KA16, KA22, KA30, KB23, KB24, KB27, KI14, KI16, KI17, KI18, and KI21) with real damages caused by accelerated lifetime tests and 1 healthy bearing (K001) under the working condition N15_M07_F10 are used to verify the performances. The PU dataset chooses 1000 samples (500 for training and 500 for testing) for each fault condition, so there are 7000 training samples and 7000 testing samples in total for 14 different classes.

The planetary gearbox data of Gearbox A dataset and the parallel gearbox data of Gearbox B dataset are all acquired by the 608A11 vibrating sensors placed on the planetary gearbox and the parallel gearbox, respectively under various speed-load conditions. The sampling frequency is 5120 Hz. The Gearbox A dataset collected the mixed planetary gearbox data samples of bearing-gear faults from the working conditions of 20Hz_0 (20Hz denotes the working speed of a motor, 0 indicates the corresponding load size), 30Hz_2, 40Hz_0, and 30Hz_4 for experiments. The Gearbox B dataset collected the mixed parallel gearbox data samples from the working conditions of 20Hz_0, 30Hz_1, 40Hz_0, and 50Hz_0 for experiments. The various bearing-gearbox fault descriptions are listed in Table III. Gearbox datasets A and B can be regarded as the 9-class condition data which includes 8 fault conditions listed in Table III and 1 health condition. The two datasets both consist of 400 samples (half of the samples for training and half for testing) for each fault condition of each working condition. Both Gearbox A and B have 7200 samples for training and 7200 samples for testing, respectively.

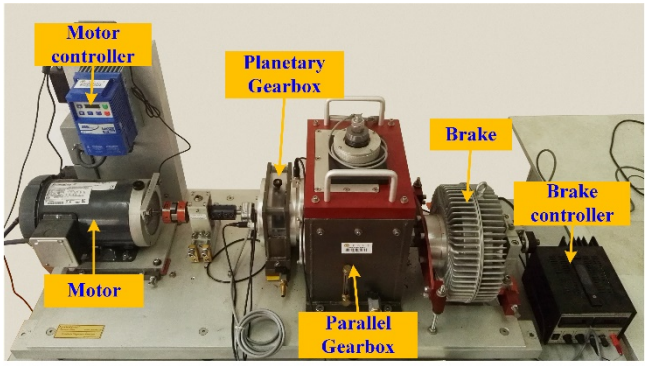

Fig. 3. The test rig of DDS

TABLE III
Gearbox Condition Descriptions

Component	Type	Description
Gear	Chipped	Crack occurs in the feet
	Miss	One of the feet is missing
	Root	Crack occurs in the root of the feet
	Surface	Wear occurs in the surface
Bearing	Ball	Crack occurs in the ball
	Combo	Crack occurs in the both inner and outer ring
	Inner	Crack occurs in the inner ring
	Outer	Crack occurs in the outer ring

B. Comparison Approaches

In our experiments, the original Transformer [8, 19] is used as our baseline, and here is called Global Transformer. The proposed ConvFormer is also constructed for comparison to verify that our proposed LiteFormer further improves the computation efficiency and performance for machine fault diagnosis. For a fair comparison, the architecture of Global Transformer and ConvFormer are the same as our LiteFormer. Global Transformer utilizes the MHSA module and adds the learnable positional embedding [19]. The head number of MHSA module is 4. ConvFormer utilizes the standard convolution in Conv module.

We also re-implement several state-of-the-art approaches including CNN based and Transformer based methods for rotating machine fault diagnosis. The ViT [19], CCT [24], Conformer [25] models are conducted for 1D sequence fault diagnosis as comparisons. The WDCNN [30], MSCNN [35],

TABLE IV
CLASSIFICATION ACCURACY OF ABLATION EXPERIMENTS

Model	CWRU dataset	PU dataset	Gearbox A dataset	Gearbox B dataset	AVG (%)	Params (M)	FLOPs (M)	Training GPU Memory (MB)
	Accuracy(%)±Std							
Global Transformer	99.73±0.14	93.30±0.26	73.81±1.11	90.29±0.46	89.28	0.366	149.046	1015
ConvFormer(3)	99.98±0.03	99.84±0.02	98.98±0.28	99.87±0.09	99.67	0.321	82.216	205
ConvFormer(16)	99.87±0.08	99.53±0.20	99.86±0.12	99.70±0.14	99.74	0.694	177.636	211
LiteFormer	**100±0.00**	**99.94±0.03**	**99.90±0.05**	**99.82±0.06**	**99.92**	**0.242**	**62.03**	**204**

and ResNet18 [32] are adopted as the general and state-of-the-art CNN methods of fault diagnosis for comparison. The details of these methods are illustrated as follows.

The ViT model divides the input data sample into patches of size 64. The head number of MHSA is 4, the expansion ratio in FFN is set at 4 and the depth is set at 7, which is the same set as our baseline. The CCT model utilizes convolution operations with the kernel size of 64 and a stride of 8 to obtain convolutional token embeddings. The other hyperparameters are set the same as ViT above. The Conformer model consisting of the MHSA and convolution modules alternately uses the same patch embedding layers as in our LiteFormer to obtain input tokens. Its convolution kernel size is also set at 16, which is the same as our LiteFormer. The other hyperparameters are also set the same as our baseline. For a fair comparison, all Transformer models adopt sequence pooling to weight the outputs of the final Transformer block.

The WDCNN adopts the same structure in Ref. [30]. The convolution kernel size of MSCNN is 128 and the other hyperparameters are set the same as in Ref. [35]. The ResNet18 [32] for 1D sequence firstly utilizes the convolution operations to increase the channel number to 64 and then the rest structure remains in the original configuration for fault diagnosis.

Our LiteFormer runs 30 epochs and the learning rate is set at 0.001. The learning rate reducing per epoch based on cosine annealing is adopted. As set up in Ref. [35], MSCNN runs 50 epochs and its learning rate is 0.001. The other models all run 30 epochs as our LiteFormer. The learning rate for WDCNN and ResNet18 is 0.01. The learning rate of all Transformer models is 0.001, the same as our LiteFormer. All models use the Adam optimizer for training from scratch. For CWRU dataset and PU dataset, the training batch size is both 32, and for Gearbox datasets, the training batch size is 128. The experimental results are the average of 10 random experiments for our work. Our works are realized in Python 3.8.0 with torch 1.8.0, Cuda version 11.1 and executed on Computer operating system Windows 10, Intel(R) Core(TM) i9-10940X CPU @ 3.30 GHz, 96.0GB RAM, and GPU NVIDIA GeForce RTX 3080, 10GB.

C. Results and Analysis

1) Ablation Study

In the previous primary experiments in section III(B), we have proved that the convolution operations can alleviate the GPU memory burden compared with the self-attention module and improve the performance of the model. Here, we will verify that the proposed LiteFormer further improves the computation efficiency and the performance for rotating machine fault diagnosis. The performance evaluation and the results are presented in Table IV. Table IV lists the classification accuracies with standard deviation (Std) on all four datasets, the average accuracy (AVG) of four datasets, the trainable parameters (Params), the floating-point of operations (FLOPs) and the maximal training GPU memory consumption (batch size of 32) of each method.

According to Table IV, it can be seen that the ConvFormer with a kernel size of 16 (ConvFormer(16)) brings a 0.88% performance gain over ConvFormer with a kernel size of 3 (ConvFormer(3)) on Gearbox A dataset, but it suffers performance losses of 0.11%, 0.31% and 0.17% on the CWRU dataset, PU dataset and Gearbox B dataset, respectively. Additionally, ConvFormer(16) has a computation complexity that is more than twice as high as ConvFormer (3). It can be explained that the increased computation complexity affects the performance of ConvFormer and ConvFormer with a large kernel size may easily overfit. Our proposed LiteFormer decreases the computation complexity by nearly 65% compared to ConvFormer(16) and obtains the highest average accuracy (AVG) on the four datasets, outperforming ConvFormer. We can also see that the GPU memory consumption of our LiteFormer is similar to ConvFormer, which alleviates the computation memory burden of Transformer greatly. The results indicate that our proposed LiteFormer outperforms ConvFormer under a large receptive field for rotating machine fault diagnosis.

2) Visualization Study

The input data of our LiteFormer model is the samples of the 1-D complex vibration signals. The LiteFormer aims to learn the intrinsic vibration characteristics of signals for fault diagnosis. In order to show the learning process of our proposed LiteFormer, we have drawn the learning weights of the depth-wise convolution kernel in DConv module under Gearbox A dataset. The learning weight maps are shown in Fig. 4. Since there are 64 kernels of the same size in convolutional layers, the size of one learning weight map of the depth-wise convolution kernels is 16×64 , where 16 means the kernel size and 64 means the number of channels. Fig. 4 presents the convolutional learning weight maps of LiteFormer block 1, block 5, and block 7. It can be seen that the total values of learning weights are increasing as the model learns layer by layer, and the weight maps become

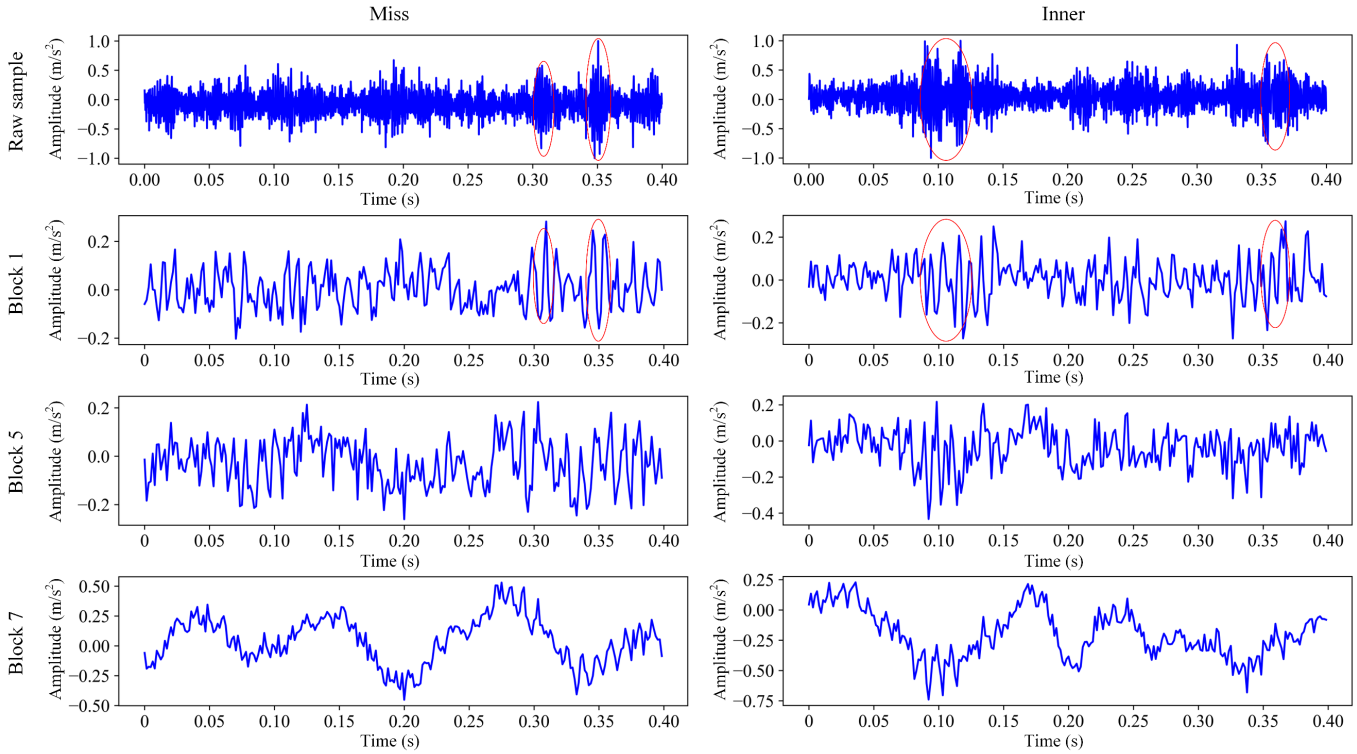

Fig. 5. The feature maps of different blocks of our proposed LiteFormer

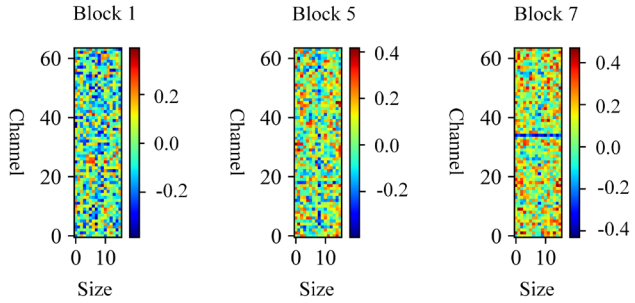

Fig. 4. The learning weight maps in the depth-wise convolution of our LiteFormer

brighter and brighter as the depth of the model increases.

To better present the learning process, the corresponding learned features of our proposed LiteFormer blocks and the raw signal samples are illustrated in Fig. 5. The LiteFormer is a sequence-based model and the feature learning actually works along the time axis, so that the feature matrixes are averaged along the channel dimension to plot the feature maps. Fig. 5 presents the raw signals and the feature maps of the gear fault "Miss" and the bearing fault "Inner". It can be seen from Fig. 5 that the feature maps in the lower block (block 1) learn the basic waveform of the vibration signals, where the amplitudes of features are bigger when the impulses are obvious in raw signals. With the model depth increasing, the learned feature maps become more and more abstract and discriminative, so that the LiteFormer model can finally attach different features to different fault types.

To illustrate the features more intuitively, we present the feature maps using the t-SNE method in Fig. 6. Fig. 6 clearly shows the learned feature distributions of all fault types of LiteFormer blocks. The features learned by the first block cannot be separated, so as the raw samples do. The features of the higher blocks (block 5 and block 7) are gradually separated, which is consistent with the trend of the feature maps shown in Fig. 5. The fault features of the last block of our proposed LiteFormer are clustered well. The visualization study on Gearbox A dataset demonstrates the effectiveness of the feature learning of our proposed LiteFormer model.

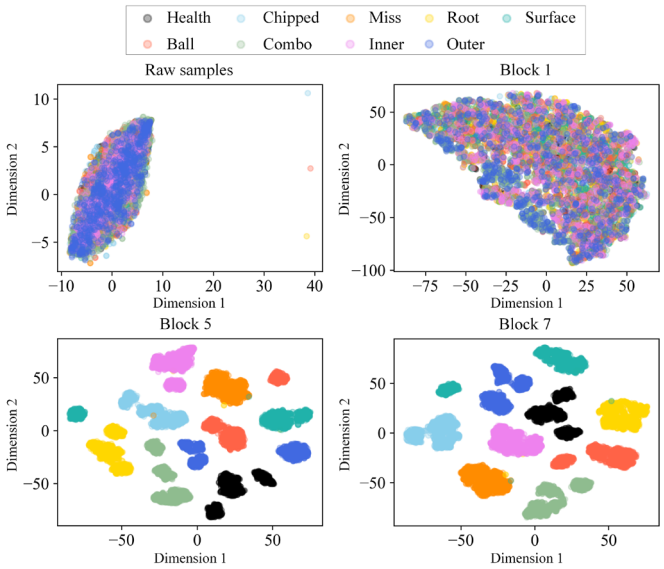

Fig. 6. The t-SNE feature maps of different blocks of our proposed LiteFormer

3) *Comparison Experiments*

To investigate the effectiveness of our proposed LiteFormer

TABLE V
COMPARISON EXPERIMENTS FOR FAULT DIAGNOSIS

Model	CWRU dataset	PU dataset	Gearbox A dataset	Gearbox B dataset	AVG (%)	Params (M)	FLOPs (M)	Training GPU Memory (MB)
	Accuracy(%)±Std							
ViT	98.80±0.41	95.08±0.59	74.83±3.65	93.21±0.65	90.48	0.356	12.217	49
CCT	99.92±0.06	99.47±0.11	96.06±1.00	97.83±0.34	98.32	0.354	143.551	978
Conformer	100.00±0.00	99.91±0.05	99.43±0.20	99.09±0.23	99.61	0.79	286.938	1322
WDCNN	99.69±0.12	98.72±0.20	95.33±1.89	96.02±0.63	97.44	0.055	0.755	4
MSCNN	99.79±0.07	98.97±0.17	98.27±0.51	98.37±0.46	98.85	21.546	113.709	496
ResNet18	99.58±0.17	99.13±0.11	99.18±0.32	98.35±0.34	99.06	3.857	89.132	119
LiteFormer	**100.00±0.00**	**99.94±0.03**	**99.90±0.05**	**99.82±0.06**	**99.92**	**0.242**	**62.03**	**204**

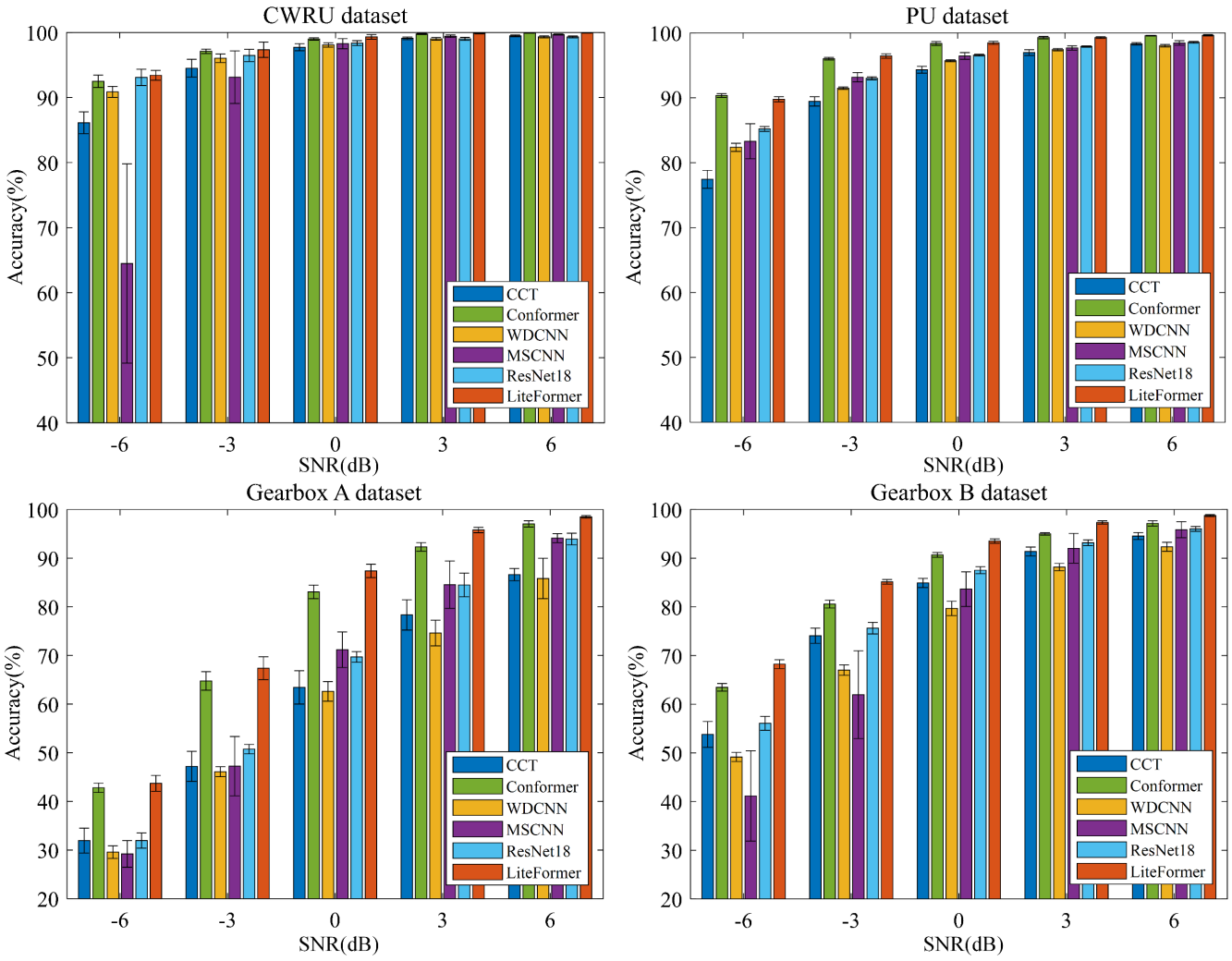

Fig. 7. Accuracies under different SNRs of white Gaussian noise

method for rotating machine fault diagnosis, comparisons between the Transformer-based models and the state-of-the-art CNN-based fault diagnosis methods are presented in Table V. ViT model, the pure Transformer of self-attention, is far less accurate than the proposed LiteFormer. CCT model, which uses convolutional tokenization, improves the performance a lot on all four datasets compared to ViT. However, the FLOPs number of CCT is increasing largely since its token number is also growing. Although Conformer obtains a high average accuracy (AVG) of 99.61%, which is only -0.31% lower than that of our LiteFormer, its Params and FLOPs are about 4 times that of our LiteFormer, and its training GPU memory is more than 6 times that of our LiteFormer. The relatively high computation complexity and high memory consumption limit its application in engineering for fault diagnosis. The results also show that the convolutions introduced into Transformers

largely enhance the performance for small datasets, especially on Gearbox A dataset. And our proposed LiteFormer method, achieving the test accuracy of 100%, 99.94%, 99.90%, and 99.82% respectively on the four datasets, outperforms the other Transformer models for rotating machine fault diagnosis.

Compared to CNN-based methods, our proposed LiteFormer also achieves the best classification performance on all four datasets. ResNet18 obtains 99.06% AVG, which is -0.86% lower than that of our LiteFormer, and its number of Params is much higher. Although WDCNN has the lowest computation complexity, its fault classification accuracies are all much lower than that of our LiteFormer. The MSCNN improves its accuracy by increasing the scales of CNNs. However, its performance is still inferior to our LiteFormer and its Params number is the largest. For CWRU dataset, PU dataset and Gearbox datasets, the proposed LiteFormer method surpasses the other counterparts and sets new state-of-the-art results, which indicates the strong feature learning ability and generalization of LiteFormer for rotating machine fault diagnosis.

4) Analysis of Noise Effect

To investigate the robustness of our proposed LiteFormer to resist background noise, we add the white Gaussian noise with different signal-to-noise ratios (SNRs) to the original data samples of all four datasets, respectively. White Gaussian noise of SNRs ranging from -6 dB to 6 dB is added to data samples with a stride of 3 dB. CCT, Conformer, WDCNN, MSCNN, and Resnet18 models are chosen for comparison. Fig. 7 presents the results of models with different SNRs on the four datasets.

From the noise experimental results on the four datasets in Fig. 7, it can be observed that the proposed LiteFormer obtains excellent anti-noise performance in each dataset. The classification accuracy of LiteFormer on the four datasets with SNR of 6dB is 99.96%, 99.64%, 98.50%, and 98.76%, respectively. When the SNR is set to -6dB, the classification accuracy of LiteFormer is down to 93.42%, 89.77%, 43.71%, and 68.24%, respectively, which is still higher than the other methods on the CWRU and Gearbox datasets, and is comparable to the Conformer on PU dataset. The results indicate that the proposed LiteFormer is of strong robustness under noise. The LiteFormer shows obvious advantages for rotating machine fault diagnosis even under noise compared with the CNN-based fault diagnosis methods. The ResNet18 obtains lower performance than our LiteFormer and MSCNN is the most unstable model under noise for its large number of Params. The overall performance of Conformer is a little inferior to that of LiteFormer, indicating that the proposed DConv module, which is lightweight for fault diagnosis, can perform even better than the combination of MHSA module and convolution module. The results prove that our proposed LiteFormer has strong robustness and outperforms the state-of-the-art methods under noise for rotating machine fault diagnosis.

5) Analysis of Class Imbalance

The fault data in real applications exists the class imbalance issue, which affects the performances of intelligent fault diagnosis models. To investigate the robustness of our proposed LiteFormer on the unbalanced dataset, we simulate the moderate class imbalance experiments on the PU dataset and Gearbox A dataset. We set three groups of datasets with different imbalance ratios on both datasets, which are shown in Table VI and Table VII, respectively. Group 1 is the balanced dataset used in our paper. Group 2 and Group 3 are imbalanced datasets constructed by reducing the training samples of some fault types. Group 3 is more imbalanced than Group 2. The three groups of datasets are used to simulate the imbalanced experiments and the experimental results are shown in Fig. 8.

TABLE VI
THREE GROUPS OF UNBALANCED DATASETS ON PU DATASET

Fault mode	Training samples			Testing samples
	Group 1	Group 2	Group3	Group 1/2/3
K001	500	500	500	500
KA04	500	500	500	500
KA15	500	300	250	500
KA16	500	300	250	500
KA22	500	300	250	500
KA30	500	180	125	500
KB23	500	180	125	500
KB24	500	180	125	500
KB27	500	100	62	500
KI14	500	100	62	500
KI16	500	100	62	500
KI17	500	50	31	500
KI18	500	50	31	500
KI21	500	50	31	500

TABLE VII
THREE GROUPS OF UNBALANCED DATASETS ON GEARBOX A DATASET

Fault mode	Training samples			Testing samples
	Group 1	Group 2	Group3	Group 1/2/3
Health	800	800	800	800
Chipped	800	600	500	800
Miss	800	600	500	800
Root	800	400	300	800
Surface	800	400	300	800
Ball	800	200	120	800
Combo	800	200	120	800
Inner	800	80	60	800
Outer	800	80	60	800

It can be observed from Fig. 8 that the imbalanced datasets cause the performance degradation of models and the overall accuracy of Group 3 is much lower than that of Group 2, which indicates that the class imbalance of the dataset will degrade the model performance and the dataset with higher imbalance ratio degrades the performance more. Nevertheless, our proposed LiteFormer still performs the best under certain imbalance ratios of the datasets compared with other methods. The results prove that our proposed LiteFormer has strong

robustness under moderate class imbalance and outperforms the state-of-the-art methods on imbalanced datasets for rotating machine fault diagnosis.

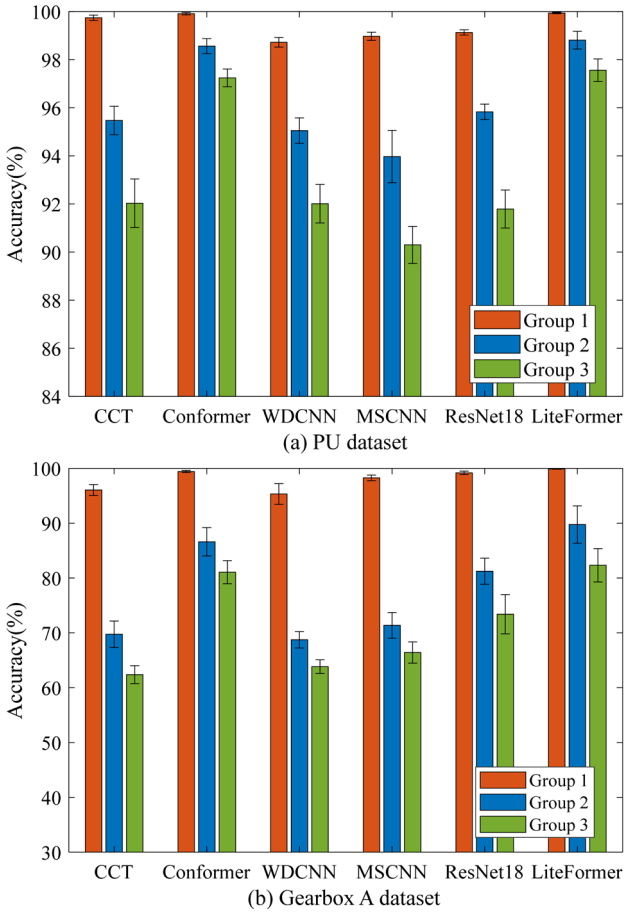

Fig. 8. Results of three groups of datasets for class imbalance experiments.

VI. Conclusion

Our proposed LiteFormer approach has been developed for rotating machine fault diagnosis in this paper. Firstly, the paper has provided the mathematical analysis between self-attention and convolution, demonstrating that the self-attention module can be regarded as an enhanced convolutional layer. Then the ConvFormer has been proposed to show that the convolutional layer replacing the self-attention scheme in Transformer mitigates the issues of computation burden and the training difficulty with small datasets, while keeping the high fault classification accuracy. The exploration of ConvFormer shows the inherent structural superiority of Transformer and investigates the essence of Transformer. Finally, the paper has further proposed the LiteFormer, which utilizes depth-wise convolution to reduce the computation complexity and improve the efficiency and generalization of ConvFormer for fault diagnosis. The proposed LiteFormer not only inherits the inherent structural superiority of Transformer, but also has the inductive bias of convolution and reduces the computation cost. Experimental studies have verified the effectiveness and robustness of the proposed LiteFormer for rotating machine fault diagnosis. As compared to other state-of-the-art methods, the proposed LiteFormer has been more accurate and robust for rotating machine fault diagnosis, even under noise and class imbalance scenarios. In the future, we will further investigate the structural superiority of Transformer and design a more lightweight and robust deep model based on Transformer for machine fault diagnosis.

References

[1] Z. Wu, W. Lin, B. Fu, J. Guo, Y. Ji, and M. Pecht, "A local adaptive minority selection and oversampling method for class-imbalanced fault diagnostics in industrial systems," *IEEE Trans. Rel.*, vol. 69, no. 4, pp. 1195-1206, 2019.

[2] P. Li, F.-R. Kong, Q.-B. He, and Y.-B. Liu, "Multiscale slope feature extraction for rotating machinery fault diagnosis using wavelet analysis," *Measurement*, vol. 46, pp. 497-505, 2013.

[3] Y. Li, X. Wang, S. Si, and S. Huang, "Entropy based fault classification using the Case Western Reserve University data: A benchmark study," *IEEE Trans. Rel.*, vol. 69, no. 2, pp. 754-767, 2019.

[4] J. Miao, J. Wang, and Q. Miao, "An enhanced multifeature fusion method for rotating component fault diagnosis in different working conditions," *IEEE Trans. Rel.*, vol. 70, no. 4, pp. 1611-1620, 2021.

[5] Z. Chen, K. Gryllias, and W. Li, "Intelligent fault diagnosis for rotary machinery using transferable convolutional neural network," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 339-349, 2019.

[6] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015.

[7] R. Jin, M. Wu, K. Wu, K. Gao, Z. Chen, and X. Li, "Position Encoding Based Convolutional Neural Networks for Machine Remaining Useful Life Prediction," *IEEE-CAA J AUTOMATIC*, vol. 9, no. 8, pp. 1427-1439, 2022.

[8] A. Vaswani *et al*., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30. Red Hook, NY, USA: Curran, 2017.

[9] A. He, and X. Jin, "Deep variational autoencoder classifier for intelligent fault diagnosis adaptive to unseen fault categories," *IEEE Trans. Rel.*, vol. 70, no. 4, pp. 1581-1595, 2021.

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

[11] W. Long, L. Gao, and X. Li, "A new deep transfer learning based on sparse auto-encoder for fault diagnosis," *IEEE Trans. Syst., Man, Cybern., Syst*., vol. 49, no. 1, pp. 136-144, 2017.

[12] Z. Chen, K. Gryllias, and W. Li, "Intelligent fault diagnosis for rotary machinery using transferable convolutional neural network," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 339-349, 2019.

[13] J. Grezmak, J. Zhang, P. Wang, K. A. Loparo and R. X. Gao, "Interpretable convolutional neural network through layer-wise relevance propagation for machine fault diagnosis," *IEEE Sens. J.*, vol. 20, no. 6, pp. 3172-3181, Mar. 2020.

[14] H. Yang, X. Li, and W. Zhang, "Interpretability of deep convolutional neural networks on rolling bearing fault diagnosis," *Measurement Science and Technology*, vol. 33, no. 5, p. 055005, 2022.

[15] F. B. Abid, M. Sallem and A. Braham, "Robust interpretable deep learning for intelligent fault diagnosis of induction motors," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 6, pp. 3506-3515, Jun. 2020.

[16] T. Li, Z. Zhao, C. Sun, L. Cheng, X. Chen, R. Yan, and R. X. Gao, "WaveletKernelNet: An interpretable deep neural network for industrial intelligent diagnosis, *IEEE Trans. Syst., Man, Cybern., Syst*., vol. 52, no. 4, pp. 2302-2312, 2021.

[17] X. Li, W. Zhang, and Q. Ding, "Understanding and improving deep learning-based rolling bearing fault diagnosis with attention mechanism," *Signal processing*, vol. 161, pp. 136-154, 2019.

[18] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[19] A, Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[20] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 10012-10022.

[21] Y. Ding, M. Jia, Q. Miao, and Y. Cao, "A novel time-frequency Transformer based on self-attention mechanism and its application in

fault diagnosis of rolling bearings," *Mech. Syst. Signal Process.*, no. 168, p. 108616, 2022.

[22] X. Pei, X. Zheng, and J. Wu, "Rotating Machinery Fault Diagnosis Through a Transformer Convolution Network Subjected to Transfer Learning," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-11, 2021.

[23] Y. Xu, Q. Zhang, J. Zhang, and D. Tao, "Vitae: Vision transformer advanced by exploring intrinsic inductive bias," in *Advances in Neural Information Processing Systems*, vol. 34, Red Hook, NY, USA: Curran, 2021.

[24] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the big data paradigm with compact transformers," 2021, *arXiv: 2104.05704*.

[25] A. Gulati, J. Qin, C. C. Chiu, N. Parmar, Y. Zhang, J. Yu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," 2020, *arXiv:2005.08100*.

[26] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," 2019, *arXiv:1904.10509*.

[27] L. Zhang, Z. Song, Q. Zhang, and Z. Peng, "Generalized transformer in fault diagnosis of Tennessee Eastman process," *Neural. Comput. Appl.*, vol. 34, no. 11, pp. 8575-8585, 2022.

[28] Z. Wei, X. Ji, L. Zhou, Y. Dang, and Y. Dai, "A novel deep learning model based on target transformer for fault diagnosis of chemical process," *Process. Saf. Environ.*, vol. 167, pp. 480-492, 2022.

[29] W. Sun, R. Zhao, R. Yan, S. Shao, and X. Chen, "Convolutional discriminative feature learning for induction motor fault diagnosis," *IEEE Trans. Ind. Informat.*, vol. 13, no. 3, pp. 1350-1359, 2017.

[30] W. Zhang, G. Peng, C. Li, Y. Chen, and Z. Zhang, "A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals," *Sensors*, vol. 17, no. 3, p. 425, 2017.

[31] W. Zhang, X. Li, and Q. Ding, "Deep residual learning-based fault diagnosis method for rotating machinery," *ISA transactions*, vol. 95, pp. 295-305, 2018.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, Jun. 2016, pp. 770-778.

[33] R. Zhao, D. Wang, R. Yan, et al., "Machine health monitoring using local feature-based gated recurrent unit networks," *IEEE Trans. Ind. Electron.*, vol. 65, no. 2, pp. 1539-1548, Feb. 2018.

[34] X. Ding, and Q. He, "Energy-fluctuated multiscale feature learning with deep convnet for intelligent spindle bearing fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 66, no.8, pp. 1926-1935, 2017.

[35] G. Jiang, H. He, J. Yan, P. Xie, "Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox," *IEEE Trans. Ind. Electron.*, vol. 66, no. 4, pp. 3196-3207, 2018.

[36] H. Fang, J. Deng, Y. Bai, B. Feng, S. Li, S. Shao, and D. Chen, "CLFormer: A Lightweight Transformer Based on Convolutional Embedding and Linear Self-attention with Strong Robustness for Bearing Fault Diagnosis under Limited Sample Conditions," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1-8, 2021.

[37] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 3431-3440.

[38] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 764-773.

[39] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 11030-11039.

[40] F. Yu, V. Koltun, and T. Funkhouser, "Dilated Residual Networks," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 472-480.

[41] Case Western Reserve University Bearing Data Center. [Online]. Available:http://csegroups.case.edu/bearingdatacenter/pages/welcomecase-western-reserve-university-bearing-data-center-website

[42] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 1251-1258.

[43] 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. Eur. Conf. Prognostics Health Manage. Soc.*, 2016, pp. 5-8.

[44] C. Lessmeier KAt-DataCenter, Chair of Design and Drive Technology. Paderborn, Germany: Paderborn University, 2019. [Online]. Available: https://mb.uni-paderborn.de/kat/forschung/datacenter/bearingdatacenter