IEEE INSTRUMENTATION & MEASUREMENT SOCIETY

Received 27 May 2022; revised 27 June 2022; accepted 28 June 2022. Date of publication 13 July 2022; date of current version 25 July 2022.
The review of this paper was arranged by Associate Editor Peng Wang.

Digital Object Identifier 10.1109/OJIM.2022.3190535

Effective Convolutional Transformer for Highly Accurate Planetary Gearbox Fault Diagnosis

WENJUN SUN[1], HUI WANG[1] (Graduate Student Member, IEEE), JIAWEN XU[1], YUAN YANG[1], AND RUQIANG YAN[1,2] (Fellow, IEEE)

[1]School of Instrument Science and Engineering, Southeast University, Nanjing 210096, Jiangsu, China

[2]School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, Shaanxi, China

CORRESPONDING AUTHOR: R. YAN (e-mail: ruqiang@seu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 51835009.

ABSTRACT To extract the global temporal correlations and local features together to enhance the accuracy for fault diagnosis, this paper proposes an effective convolutional Transformer (ECT), which can learn the global temporal correlations using Transformer and local features with convolution at the same time. The proposed method designs a multi-stage hierarchical structure of Transformer, which utilizes convolutional tokenization to distill dominating sequence features from raw vibration signals while increasing the dimension of token embedding across stages at the same time as that in CNNs. The spatial-reduction attention (SRA) and the linear dimension reduction projections are introduced respectively to Transformer at different stages to reduce the resource consumption of the model. Finally, the proposed method utilizes a sequence pooling strategy on the output of Transformer to eliminate the requirement of the class token and make the model accurate for classification. The specially designed structure makes the model flexible and effective for planetary gearbox fault diagnosis. Experiments performed on planetary gearbox fault simulators indicate that the ECT method has significant effectiveness and high accuracy compared with the state-of-the-art methods for planetary gearbox fault diagnosis.

INDEX TERMS Convolutional tokenization, fault diagnosis, spatial-reduction attention, sequence pooling, transformer.

I. INTRODUCTION

PLANETARY gearbox has been widely used in mechanical systems as a critical transmission part. The operating environment of planetary gearbox is usually harsh, which needs to be exposed to high temperature, heavy load, big shock, friction and other factors long-term. Its bearing and gear, as the core components of planetary gearbox, are frequently suffered from potential defects such as bearing inner race fault, outer race fault, gear root crack, chipped teeth, broken teeth, etc. in the long-running [1]. The potential defects will lead to the failure of the planetary gearbox, resulting in massive economic losses and casualties every year. Thus, effective fault monitoring and diagnosis of planetary gearbox are necessary for the timely identification of mechanical faults. The process of machine fault diagnosis is generally divided into three steps: signals collection, feature extraction, and fault classification. Traditional fault feature extraction mainly relies on the manual selection using mathematical-statistical methods or the signal processing methods to extract time- or frequency- domain features [2]–[3]. The features are then adopted for the health condition classification. Hence, the performance of the final classifier for fault classification is fundamentally depending on the suitability of the manually selected features. Deep learning (DL) methods [4]–[6] have received a lot of attention in recent years because of their ability to learn discriminative features automatically from raw data. And it has been extensively studied and applied in the field of machine fault diagnosis [7]–[9].

DL-based fault diagnosis methods have achieved remarkable achievements, including convolutional neural networks (CNNs) [10]–[12] and recurrent neural

networks (RNNs) [13]–[15]. CNNs have been widely applied in fault diagnosis because of their locality and translation equivariance, which enable CNNs with extraordinary capability to learn the local features and easy to be trained with small datasets [10]–[12]. Luo *et al.* [16] trained the deep convolutional neural network (DCNN) with an explicable training guide for fault diagnosis of planetary gearbox and obtained ideal diagnosis results. Zhang *et al.* [17] proposed a deep convolutional neural network with wide first-layer kernels (WDCNN) for extracting features of raw vibration signals and suppressing high-frequency noise. Jiang *et al.* [18] proposed a multi-scale convolutional neural network (MSCNN) for end-to-end gearbox fault diagnosis of wind turbines. Ma *et al.* [19] apply deep residual learning [6] in planetary gearbox fault diagnosis to construct deeper architecture with demodulated time-frequency features. Although CNNs have promising capability in fault diagnosis, its locality also limits the learning of global correlations. RNNs are proposed to solve the problem of global temporal correlations [13]–[15]. However, the inherent limitations of RNNs, such as the sequential structure which can not be computed in parallel and results in a large amount of computation memory, greatly hinder their applications in machine fault diagnosis.

Transformer [20], which is implemented completely by self-attention mechanisms, has been the dominant standard in natural language processing (NLP) [21]–[23] and has also obtained remarkable achievements in computer vision (CV) domain [24]–[25] for its strong ability to model global correlations with parallel computing. Based on this advantage, this paper proposed a vision Transformer-based method, named effective convolutional Transformer (ECT), for accurate planetary gearbox fault diagnosis with 1D signals. Since the vision Transformer [24] lacks the inductive bias inherent to CNNs and relies heavily on large-sized image datasets to learn it, numerous researchers [26]–[29] introduce convolutions into Transformers to enhance the generalization of various tasks. Inspired by these works and considering the limited-sized datasets in the field of machine fault diagnosis, this paper applies the convolutions to obtain the token embedding in the proposed ECT, which replaces the input patching and positional embedding in the standard Transformer. The convolutional token embedding ensures the proposed ECT obtains the local inductive bias while modeling the global features even with a small dataset, and also allows the ECT more flexible to input length. Besides, the quadratic complexity and luxury resource consumption of large Transformers make them difficult to be directly applied in the field of machine fault diagnosis. To reduce the resource consumption of Transformer and keep the computation balance [30] of the attention mechanism and multi-layer perceptron (MLP) block, our proposed ECT model further adopts the spatial-reduction attention (SRA) [31] and the linear dimension reduction projections to refine the model. Therefore, the proposed ECT makes it possible to learn both global and local features for planetary gearbox fault diagnosis.

The proposed ECT method introduces the convolution into Transformer to obtain convolutional token embedding for Transformer input tokens. In particular, the ECT utilizes convolutional tokenization that performs the overlapping convolution operation with stride on 1D sequences to construct Transformer into multiple stages to form a hierarchical structure as that in CNNs. The multi-stage structure allows the ECT to capture local features and progressively decrease the sequence length while increasing the dimension of token features across stages at the same time. The increasing dimension of token embedding enhances the learned attention maps. Then the SRA in Transformer block learns the global features. Finally, a novel sequence pooling strategy [28] is applied to remove the need for the conventional class token design in Transformer and make the identification more accurate. The special structure makes the ECT effective and suitable for small datasets of fault signals. ECT model combines the advantages of both Transformer and CNN which guarantee the model to be robust and flexible to learn the global and local features simultaneously for highly accurate fault diagnosis.

The main contributions of our work are as follows:

1) An ECT model is proposed for highly accurate planetary gearbox fault diagnosis. In particular, the vision Transformer is introduced into planetary gearbox fault diagnosis with 1D vibration signals.

2) A multi-stage hierarchical structure by convolutional tokenization is designed for Transformer to learn both local and global features. The SRA technique and the linear dimension reduction projections are adopted to reduce resource consumption. The sequence pooling strategy is combined in Transformer to simplify the classification. All the above make the ECT model more flexible and effective for planetary gearbox fault diagnosis.

3) Extensive experiments confirm the accuracy and generalizability of our ECT method. In addition, the ECT model outperforms the state-of-the-art CNN-based models in planetary gearbox fault diagnosis under long-range sequences.

The paper is organized as follows. In Section II, the framework of our proposed ECT model is introduced. In Section III, the procedure of our proposed ECT method for planetary gearbox fault diagnosis is presented. Then, in Section IV, the experimental setup and results are illustrated. At last, concluding remarks are provided in Section V.

II. THE FRAMEWORK OF OUR ECT MODEL

The proposed methodology of Transformer is introduced in this section. Our proposed ECT approach leverages convolutional tokenization to construct the multi-stage hierarchical structure of Transformer. In particular, the Transformer adopts the SRA and the linear dimension reduction projections in two stages of our ECT model, respectively. The

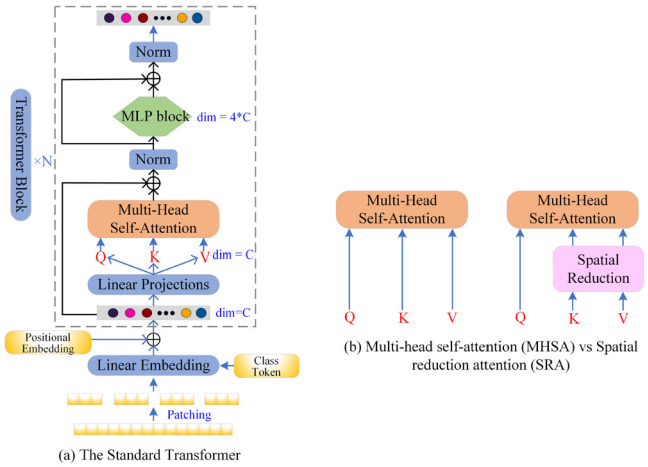

FIGURE 1. The pipeline of Transformer model: (a) The standard Transformer. (b) Multi-head self-attention (MHSA) vs Spatial reduction attention (SRA).

improvement of the model is potential to reduce resource consumption effectively. In the final stage, the sequence pooling strategy is used for the output of Transformer to make the classification more accurate. The methodology details of the ECT model are presented as follows.

A. CONVOLUTIONAL TOKENIZATION

The input of a standard Transformer [20], [24] is a sequence of token embeddings. As shown in Fig. 1(a), the ViT [24] divides the image into patches and flattens each patch into a sequence to form the token embedding. The positional embedding and a class token will be added to the sequence respectively as the input tokens. In ECT, the token embedding is generated by convolutional tokenization which performs the overlapping convolution operation with stride on 1D vibration sequence at the beginning of each stage of the Transformer. Then a layer normalization [32] is designed following the convolution operation. For a 1D time sequence of length H_{i-1} or an output token sequence from the previous stage $x_{i-1} \in R^{H_{i-1} \times C_{i-1}}$. The convolution operation $f(\cdot)$ with C_i kernels of size k_i and stride s_i is used to map x_i into new tokens $f(x_i) \in R^{H_i \times C_i}$, and the length H_i can be denoted as

$$H_i = \frac{H_{i-1} - k_i}{s_i} + 1 \tag{1}$$

Then the convolutional token embedding is normalized by layer normalization to fit in the input token of Transformer blocks of stage i.

Convolutional tokenization enables the model to adjust the length H of sequences and the dimension C of token embeddings at the start of each stage by varying the convolution operation parameters. In this case, our model can learn the multi-scale features and the local inductive bias for Transformer. Furthermore, convolutional tokenization aids in the construction of a multi-stage hierarchical structure, which simplifies the Transformer with flexible input size and transforms the token embedding into high-dimensional space to provide comprehensive information.

B. TRANSFORMER BLOCK

The encoder module of Transformer is used for classification tasks. The pipeline of the conventional encoder module of the original Transformer [20] and the ViT model [24] is shown in Fig. 1(a). The encoder module is made up of a series of stacked encoders consisting of multi-head self-attention (MHSA) and MLP layers. Each MHSA and MLP layer is surrounded by a residual connection [6], which is then accompanied by layer normalization (LN) [32]. Each encoder layer's output ports are listed as

$$y = \mathrm{LN}\big(x' + \mathrm{MLP}(x')\big), \textit{ and } x' = \mathrm{LN}(x + \mathrm{MHSA}(x)) \tag{2}$$

It is noted that the stacked multiple encoder layers of Transformer take the same structure, but do not share the same parameters. Therefore, the attention mechanism and other structures used in Transformer are illustrated.

1) MULTI-HEAD SELF-ATTENTION (MHSA)

Transformer relies on the self-attention (SA) mechanism to compute a representation of a sequence that relates to features at different positions. In other words, the self-attention mechanism enables the extracting of dependencies ignoring the distance in input sequences. The sequences of input tokens $x_i \in R^{H_i \times C_i}$for a self-attention module are linearly transformed into qkv space, i.e., queries $Q_i \in R^{H_i \times C_i}$, keys $K_i \in R^{H_i \times C_i}$and values $V_i \in R^{H_i \times C_i}$. A single-head computes scaled dot-product attention for all queries and keys, divides each by scaling factor $\sqrt{C_i}$, and applies a softmax function to the values to obtain weights. The representation of computed weighted values is given as

$$\mathrm{Attention}(Q_i, K_i, V_i) = \mathrm{softmax}\left(\frac{Q_i K_i^T}{\sqrt{C_i}}\right) V \tag{3}$$

Multi-head self-attention (MHSA) is proven to be more advantageous than a single attention function, which linearly projects the queries, keys, and values h times using distinct, learnt linear projections to d_k, d_k and d_v dimensions. MHSA enables the model to simultaneously attend to information from various representation subspaces at various points [20]. It is a SA extension that separates queries, keys, and values for h times and runs the attention function in parallel, as follows.

$$\begin{aligned} \mathrm{MultiHead}(Q_i, K_i, V_i) &= \mathrm{Concat}(\mathrm{head}_{i1}, \ldots, \mathrm{head}_{ih}) W^o \\ \text{where } \mathrm{head}_{ij} &= \mathrm{Attention}\left(Q_i W_{ij}^Q, K_i W_{ij}^K, V_i W_{ij}^V\right) \end{aligned} \tag{4}$$

where the parameter matrices $W_{ij}^Q \in R^{C_i \times d_k}$, $W_{ij}^K \in R^{C_i \times d_k}$, $W_{ij}^V \in R^{C_i \times d_v}$ and $W^o \in R^{h d_v \times C_i}$ are for linear projections. Normally, d_k has the same value as d_v and is set as $d_k = d_v = d_{head} = C_i/h$. The total computing cost is comparable to single-head attention with full dimensionality. And a linear aggregation is performed for corresponding values V_i.

The similarity between different tokens is determined using the MHSA dot-product, resulting in long-range and global attention.

C. SPATIAL-REDUCTION ATTENTION (SRA)

In SRA [31], the spatial scales of the key K and the value V are reduced before the multi-head attention operation (shown in Fig. 1(b)), which results in a significant reduction in attention computation/memory. Details of the SRA in stages can be formulated as

$$\begin{aligned}\mathrm{SRA}(Q_i, K_i, V_i) &= \mathrm{Concat}(\mathrm{head}_{i1}, \ldots, \mathrm{head}_{ih})W^o \\ \text{where } \mathrm{head}_{ij} &= \mathrm{Attention}\Big(Q_i W_{ij}^Q, \mathrm{SR}(K_i)W_{ij}^K, \mathrm{SR}(V_i)W_{ij}^V\Big) \quad (5)\end{aligned}$$

where the parameter matrices $W_{ij}^Q \in R^{C_i \times d_k}$, $W_{ij}^K \in R^{C_i \times d_k}$, $W_{ij}^V \in R^{C_i \times d_v}$ and $W^{\mathrm{o}} \in R^{hd_v \times C_i}$ are for the linear projections, $d_k = d_v = C_i/h$. $\mathrm{SR}(\cdot)$ is the operation to reduce the spatial dimension of the Kor V, which is written as

$$\mathrm{SR}(x) = \mathrm{LN}(\mathrm{Reshape}(x, r_i)W^s \quad (6)$$

where, $x \in R^{H_i \times C_i}$ represents the input sequence, and r_i denotes the reduction ratio of the attention layers in stages. $\mathrm{Reshape}(x, r_i)$ is a reshaping operation that transforms the input sequence x into a sequence of size $\frac{H_i}{r_i} \times r_i C_i$. $W^s \in R^{(r_i C_i) \times C_i}$is a linear projection that reduces the sequence dimension to C_i. Then the attention operation is as the original MHSA. It can be found that the computation cost of attention operation in SRA is r_i times lower than that of MHSA.

D. MLP BLOCK

The MLP block with residual connection is integrated after the attention layers. The MLP block contains a fully connected feed-forward network composed of two linear transformations separated by a rectified linear unit (ReLU) activation layer. This feed-forward layer can be denoted as

$$\mathrm{MLP}(x) = \mathrm{ReLU}(0, xW_1 + b_1)W_2 + b_2 \quad (7)$$

where $W_1 \in R^{C_i \times d_{\mathrm{f}}}$, $W_2 \in R^{d_{\mathrm{f}} \times C_i}$, $b_1 \in R^{d_{\mathrm{f}}}$, $b_2 \in R^{C_i}$ is the weights and bias of two layers, respectively. Normally, $d_{\mathrm{f}} = 4C_i$ is chosen as the dimensionality of MLP inner-layer.

E. SEQUENCE POOLING

A class token (like BERT [21]) is used in previous transformer-based classifiers, including ViT, to map the sequential outputs to a single class index. The sequence pooling (SeqPool) strategy [28] has been shown to simplify the model for classification by pooling the Transformer outputs across the sequence and improve model accuracy.

The complete data sequence is pooled across a sequence that includes important data from various parts of the input sequence. This procedure is shown as the mapping transformation $T : R^{b \times h \times \mathrm{c}} \rightarrow R^{b \times c}$. The output sequences are given as

$$x_t = f(x) \in R^{b \times h \times \mathrm{c}} \quad (8)$$

where x_t or $f(x)$ is the output of the last transformer encoder, b denotes the mini-batch size, h denotes the sequence length, and $c$denotes the embedding dimension. After that, x_t is fed to a linear function $\mathrm{g}(x_t) \in R^{c \times 1}$ and activated with softmax as

$$x_t^* = \mathrm{softmax}\Big(g(x_t)^T\Big) \in R^{b \times 1 \times h} \quad (9)$$

Hence, the final weighted embedding can be computed as

$$x_s = x_t^* x_t = \mathrm{softmax}\Big(g(x_t)^T\Big) \times x_t \in R^{b \times 1 \times c} \quad (10)$$

A weighted embedding output $x_s \in R^{b \times c}$ is obtained through pooling the second dimension. Then this output can be sent to the MLP head for classification as that in the previous studies.

SeqPool is the process of assigning learnable weights across a sequence of data, allowing the model to weigh the sequential embeddings of the latent space produced by the Transformer encoder. Furthermore, it enables better correlation of data across input data. As a result, our model can not only give more weight to tokens containing more information relevant to the classifier, but it can also better utilize information across spatially sparse data.

F. THE FRAMEWORK OF OUR ECT MODEL

Our proposed ECT model designs a two-stage hierarchical structure that utilizes convolutional tokenization to generate convolutional token embedding from a 1D sequence for the Transformer in different stages. The beginning of each stage consists of the convolutional token embedding generated by convolutional tokenization, followed by layer normalization. The convolutional token embedding is used as the input token of Transformer block in each stage. We use wide kernels in the first stage to capture the important information of vibration signals in the intermediate and low-frequency bands, and then small kernels in the second stage to achieve finer feature representation. Hence, the ECT model employs kernel size $k_1 = 64$, kernel number $C_1 = 64$, and stride $s_1 = 8$ for the convolution operation in the first stage, kernel size $k_2 = 7$, kernel number $C_2 = 256$, and stride $s_2 = 2$ for the convolution operation in the second stage.

Then Transformers use self-attention mechanisms to learn global features of sequences. In the first stage, we replace the original MHSA with the SRA to reduce the computation in Transformer block and the reduction ratio r_i of SRA is set at the value of 4. Attention layers are the major context capturing unit in the Transformer. Since the dimension C_2 of token embedding in the second stage is 4 times increased than that in the first stage, the total computation cost of Transformer block in the second stage is increasing largely. To reduce the corresponding computation and keep the computation balance of attention layers and MLP block, we reduce the dimension of token embedding by half via the linear projections to Q, K and V space, and reduce the inner layer dimension of the MLP block by 1/4 instead of 4 times expansion in the Transformer block of the second stage (shown in Fig. 2). In our ECT model, two Transformer blocks are stacked for the first stage and three Transformer blocks stacked for the second stage. Parallel

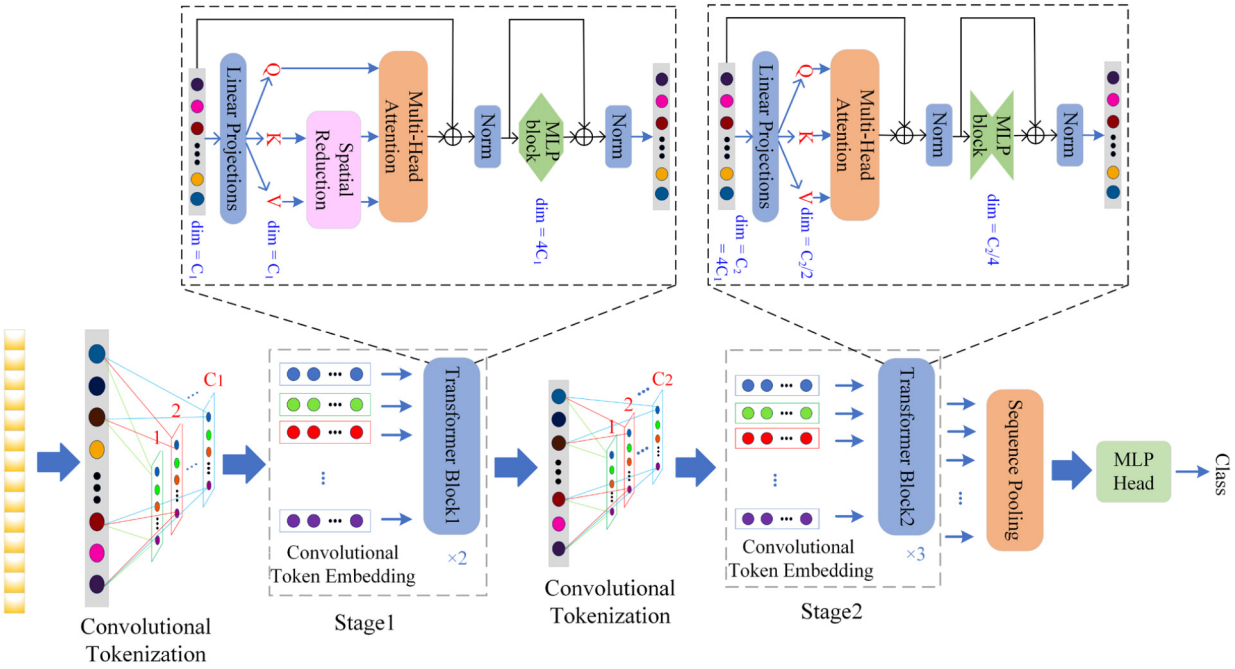

FIGURE 2. The framework of our ECT model.

attention heads $h = 2$ are employed for the first stage and use $d_k = d_v = C_1/h = 32$. For the second stage, parallel attention heads $h = 4$ are employed and we reduce the dimension of the token embedding via the linear projections to Q, K and V space by half to still obtain $d_k = d_v = C_2/2h = 32$.

Finally, sequence pooling is applied to handle the output of Transformer block for the subsequential MLP head which consists of the layer normalization and one linear layer. And the sequence pooling removes the need for the conventional class token in general transformers and makes the model more accurate. For the output of the last MLP head, the MSE loss is calculated for optimizing the model. Our proposed ECT model combines the advantages of both Transformer and CNN which guarantee the model in a flexible size to learn global and local features simultaneously. The framework of our ECT model is shown in Fig. 2.

III. THE PROCEDURE FOR FAULT DIAGNOSIS

The paper proposed an ECT for planetary gearbox fault diagnosis. The framework of the proposed ECT is illustrated in Section II-D. Here we use the ECT method for planetary gearbox fault diagnosis. Similar to the majority of the deep learning methods for machine fault diagnosis, ECT method also needs sufficient fault samples for training to learn the excellent features and exhibit strong recognition ability. Thus, the sensor data is first collected for training and testing the model. Fig. 3 presents a brief planetary gearbox fault diagnosis procedure of ECT method. In the training procedure, abundant labeled fault and health datasets are employed to train the ECT model, learning discriminative features for classification. During the training procedure, the parameters of ECT model are gradually fine-tuned according to the label information. Additionally, the raw vibration signal is utilized to train the model for end-to-end fault diagnosis. The general procedure is summarized as follows:

Step 1: Collect the vibration signals under different working conditions of the experimental facilities to construct the fault datasets.

Step 2: Divide the training data and test data from the fault datasets, respectively, and normalize the data samples.

Step 3: Construct the ECT model with convolutional tokenization, Transformer blocks, sequence pooling, and MLP head for classification according to Section II.

Step 4: Train the ECT model with training data and fine-tune it adequately to verify the performance.

Step 5: Verify the effectiveness of our proposed ECT method using test data for fault diagnosis.

IV. EXPERIMENT ANALYSIS

A. EXPERIMENTAL SETUP

Experiments were carried out on the planetary gearbox vibration signals acquired from the drivetrain dynamic simulator (DDS), as shown in Fig. 4, to validate the effectiveness of our ECT model for planetary gearbox fault diagnosis. The 608A11 vibrating sensors with a sampling frequency of 5120 Hz are chosen to collect vibration signals under various speed-load conditions. Table 1 lists the descriptions of various types of planetary gearbox faults.

Here the vibration signal of the planetary gearbox under the working conditions (20Hz_0, 30Hz_2, 30Hz_4 and 40Hz_0) are collected as the experimental dataset. The 20 Hz, 30 Hz, and 40 Hz denote the working speed of

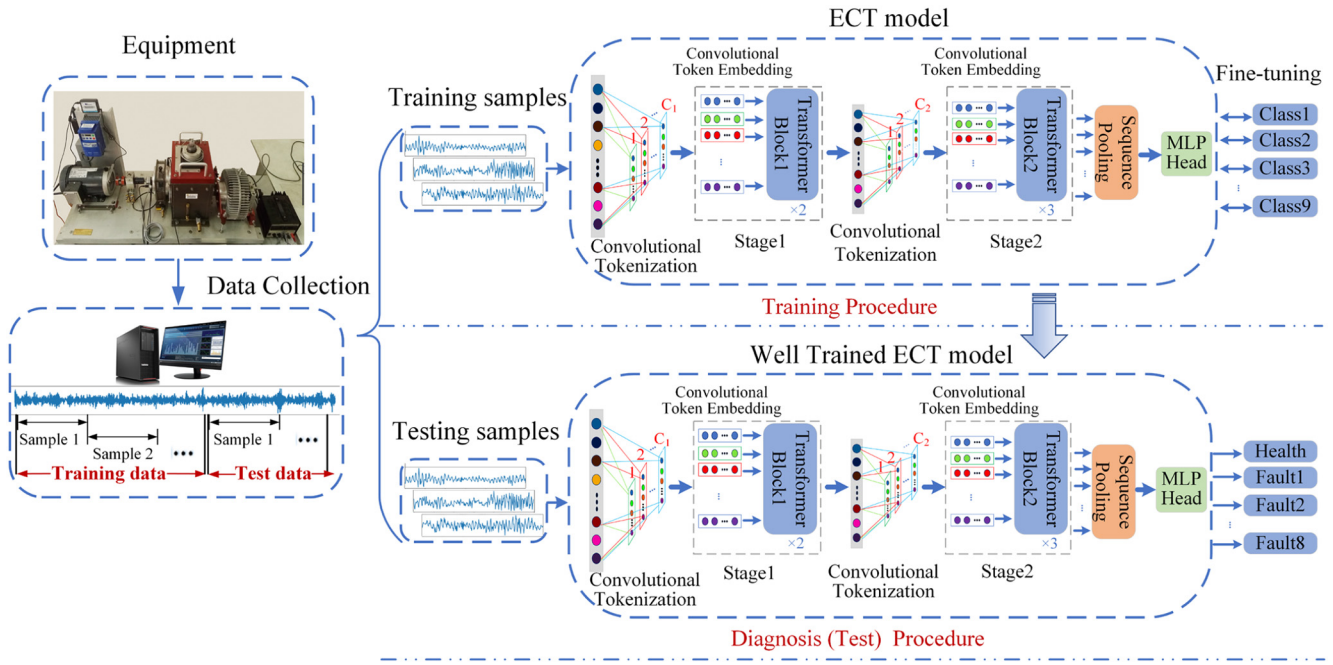

FIGURE 3. The procedure for planetary gearbox fault diagnosis.

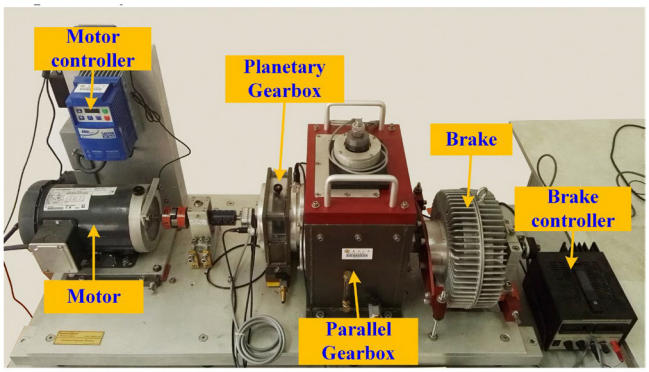

FIGURE 4. Experimental setup of DDS.

TABLE 1. Planetary 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

the motor. Besides, 0, 2 and 4 denote the corresponding load size.

In our study, the diagnostic task of the planetary gearbox with mixture fault diagnosis is carried out. This task can be regarded as the 9-class classification task which includes the 8 type faults and one Health type. There are four type faults of Bearing and gear respectively, as presented in Table 1.

TABLE 2. The detail of data samples.

Working condition	Number of samples		
	Training	Test	Total
20_0 Hz	450×9	50×9	500×9
30_2 Hz	450×9	50×9	500×9
30_4 Hz	450×9	50×9	500×9
40_0 Hz	450×9	50×9	500×9
Mixture	16200	1800	18000

Firstly, we mixed the data of multiple working conditions (20Hz_0, 30Hz_2, 30Hz_4 and 40Hz_0) to construct the mixture dataset, which consists of 500 data samples chosen from each planetary gearbox fault type of each working condition, respectively. And nine-tenth samples of each fault type from each working condition are arbitrarily chosen for training and the rest one-tenth for testing. In other words, there are 1800×9 samples and 200×9 samples under the multiple working conditions kept for training and testing, respectively. The detail can be seen in Table 2.

B. COMPARISON APPROACHES

To prove that our proposed ECT model is effective and highly accurate for fault diagnosis of planetary gearbox, our method is compared with the state-of-the-art classification methods for fault diagnosis, including the 1DCNN model which has the same number of layers as in [16], the WDCNN in [17], the MSCNN in [18] and the convolutional Bi-LSTM

network like in [15], and the Transformer-based models as ViT in [24] and CCT like in [28].

The 1DCNN model utilizes the structure of [Conv1d (1, 64, 64, 8) -> ReLu -> Conv1d (64, 128, 7, 1) -> ReLu -> MaxPool1d (4, 4) -> Conv1d (128, 256, 3, 1) -> ReLu -> MaxPool1d (4, 4) -> Linear (256, 9), which obtain the same convolution operation as our ECT model in the first stage. The WDCNN uses the same structure as in [17]. The MSCNN utilizes the convolution kernel size of 128 and the other hyperparameters set the same as in [18].

The convolutional Bi-LSTM network also uses the same convolution operation like ours to generate features for the Bi-LSTM network. Then the Bi-LSTM network is completed by two hidden layers with hidden_dim =128 and one Linear (256, 9) layer for classification.

The Transformer-based models include ViT [24], Compact Convolutional Transformer (CCT) [28] for 1D sequence and the variant standard convolutional Transformer (SCT). The ViT model divides the whole 2048 points sequence into patches of size 64, which means the token embedding dimension is reshaped to 64 without convolutions. The depth of ViT is set 5 and the head of MHSA is set 2 with head dimension of 32 the same as our ECT model for the fair contrast. The CCT utilizes the same convolutional tokenization as in the first stage of our ECT model to obtain the convolutional token embedding as the input tokens of Transformer and sets the depth of 5 and the MHSA head of 2 with head dimension of 32 for a fair comparison. The SCT model here means the standard convolutional Transformer model which is the same two-stage hierarchical structure as our ECT model while using the standard Transformer block in both two stages. The CCT model is a "columnar" structure the same as ViT and SCT model is the same two-stage structure as our ECT model like the "pyramid" structure. The Transformer-based models and our ECT model all use a dropout for the attention and the MLP block in Transformer block with a probability of 0.2.

In all experiments, the average accuracy (AVG) and standard deviation (Std) of 10 random measurements are chosen as the performance evaluation indexes. 1DCNN and WDCNN model run 20 epochs, MSCNN and ViT model run 50 epochs. The convolutional Bi-LSTM, CCT, SCT model and our ECT run for 30 epochs for best performance and all these models run with a batch size of 128. The learning rate for 1DCNN, WDCNN and the convolutional Bi-LSTM model is set as 0.01, and for MSCNN, it is set as 0.001. The value of the ViT, CCT, SCT and our ECT model is set as 0.0005, and the learning rate of all models reduces per epoch based on cosine annealing [33]. And these models are all warmed up for 3 epochs and use the primary optimizer Adam for the training from scratch. Our works are programmed in python 3.6.2 with torch 1.1.0, Cuda version 9.0 and executed on the Ubuntu 16.04 operating system.

TABLE 3. Classification accuracy of Transformer models.

Model	Head	Pool	AVG ± Std (%)	Params (M)	FLOPs (M)
ViT	2	CT	88.61±5.34	0.256	9.045
ViT	2	SP	95.32±1.30	0.256	8.758
CCT	2	SP	99.29±0.22	0.254	102.803
CCT	4	SP	99.48±0.10	0.999	327.985
SCT	2--4	SP	99.72±0.15	0.757	132.732
SCT	2--8	SP	99.83±0.09	2.589	367.616
ECT	2--4	CT	99.79±0.11	0.752	116.028
ECT	**2--4**	**SP**	**99.89±0.05**	**0.752**	**115.402**

C. RESULTS AND ANALYSIS

Firstly, we compare the ECT model with Transformer-based models such as ViT, CCT and SCT, because the proposed ECT model is improved based on these models. Here, the ViT, CCT, SCT and our ECT are evaluated, as shown in Table 3. The column "Head" means the head number of MHSA with head dimension of 32 (2–4 indicates 2 heads in the first stage and 4 heads in the second stage), which also corresponds to different token embedding dimensions. The column "Pool" denotes the use of the class token (CT) or the SeqPool (SP). In these experiments, it can be found that convolutional tokenization is more effective than the patching from the results of ViT and CCT, for convolutions providing the right inductive bias. The "pyramid" structure is superior to the "columnar" structure from the comparison between CCT and SCT, ECT model. The SCT(2-4) has fewer trainable parameters (Params) and floating point of operations (FLOPs) than that of CCT(4) while achieving higher accuracy, and it is more accurate than CCT(2), which indicates the effectiveness of the two-stage structure. Our ECT method adopting the SRA and the linear dimension reduction projection techniques further improves the accuracy compared to SCT(2-4) model and is even more accurate than the SCT(2-8) model, with Params and FLOPs reduced above 60% when obtaining the same token embedding dimension of 256 in the second stage. It indicates that our ECT model has better performance and lower computation complexity than the standard Transformer with the same structure. From the results about the column "Pool", it indicates that the SeqPool simplifies the classification and improves the classification accuracy of Transformer compared to the class token. The results have proved the effectiveness of our ECT model and our proposed method helps save the resource consumption while remaining high accuracy for planetary gearbox diagnosis.

Table 4 compares the classification accuracy realized by popular methods introduced in Section IV-B to further demonstrate that our proposed ECT method has significant performance gains for planetary gearbox fault diagnosis than the state-of-the-art methods for fault diagnosis. It can be seen that, when compared to 1DCNN, WDCNN, and MSCNN

TABLE 4. Comprehensive comparison for fault diagnosis.

Method	Bearing-gear Dataset (mixture)									AVG ± Std
	Health	Chipped	Miss	Root	Surface	Ball	Comb	Inner	Outer	
ViT	95.35	84.15	84.10	89.80	93.15	89.15	96.15	79.05	86.60	88.61±5.34
Bi-LSTM	99.60	97.90	97.90	97.90	98.05	99.25	99.60	98.15	98.30	98.52±1.14
WDCNN	99.75	99.40	99.20	99.35	99.80	99.85	99.90	99.15	99.65	99.56±0.22
1DCNN	99.95	99.45	99.30	99.20	99.40	99.75	100.00	99.40	99.45	99.54±0.36
MSCNN	100	99.55	99.40	99.80	99.65	99.90	99.85	99.30	99.50	99.66±0.18
CCT(2)	99.95	99.45	99.00	99.25	98.90	99.30	99.60	99.20	98.95	99.29±0.22
ECT	**100.00**	**99.90**	**99.80**	**99.70**	**99.80**	**99.95**	**100.00**	**100.00**	**99.85**	**99.89±0.05**

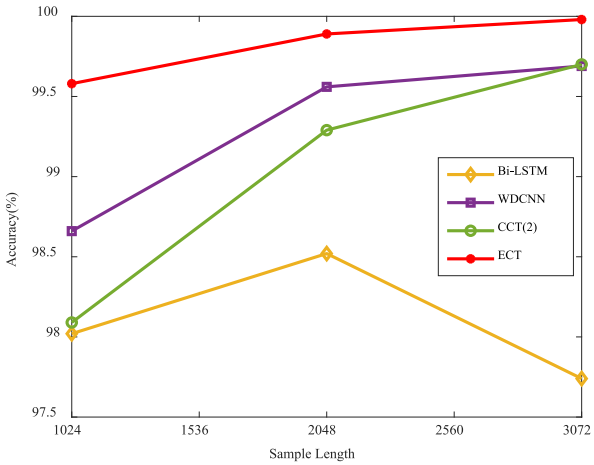

FIGURE 5. Performance under different input sample lengths.

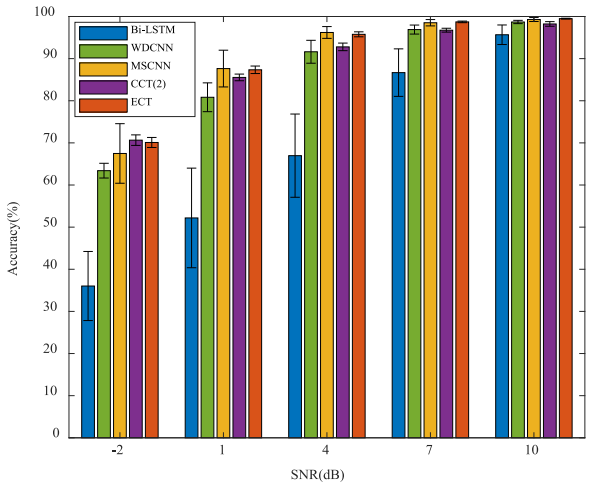

FIGURE 6. Accuracy for fault diagnosis with Gaussian noise.

CNN-based models, our ECT method achieves the highest fault classification accuracy, demonstrating that our proposed ECT model outperforms the CNN-based methods for fault diagnosis. The convolutional Bi-LSTM method's fault classification accuracy is also lower than that of our ECT method. In other words, Transformer outperforms LSTM for sequence learning. For comparison, the ViT and CCT(2) are chosen as the Transformer baselines. Our ECT method is more accurate, demonstrating the effectiveness of our proposed structure. Furthermore, our ECT model achieves the highest fault classification accuracy of 99.89%, showing that our proposed ECT method is effective for planetary gearbox fault diagnosis.

To verify the flexibility of our ECT model and further investigate the effectiveness of our ECT model, we change the input data length from 1024 to 2048 and 3072 to train the model. In addition, we chose the CCT(2) model, the convolutional Bi-LSTM, WDCNN for comparison for they are all flexible to the input size. Figure 5 depicts the results. As shown in Fig. 5, the classification accuracy of our ECT model improves as the length of the input data sequence increases. It demonstrates that our ECT model is also flexible to input size and performs better on long-range sequence learning. When the sample length is 3072, the ECT model can achieve 99.98% for planetary gearbox fault classification tasks under mixed multiple working conditions. When compared to the convolutional Bi-LSTM model, the ECT model is much more robust with varying input sequence lengths, whereas the Bi-LSTM model's performance drops when the length reaches 3072. The ECT model outperforms both the CCT(2) model and the WDCNN model a lot under different input lengths. Based on these findings, we can conclude that our proposed ECT model can be used for effective and accurate planetary gearbox fault diagnosis and the model is flexible for learning long-range dependencies.

D. GENERALIZATION ANALYSIS UNDER BACKGROUND NOISE

To validate the efficacy of our ECT model in planetary gearbox fault diagnosis with background noise, white Gaussian noise with signal-to-noise ratios (SNRs) ranging from −2 dB to 10 dB with a stride of 3 dB was added to the data samples. In this part, we compare our proposed ECT model to the CCT(2) model, the convolutional Bi-LSTM model, the WDCNN model, and the MSCNN model. Fig. 6 shows the detailed results of the models with different SNRs of noise. The classification accuracy of our ECT model improves as SNR increases, and the ECT model outperforms the CCT(2)

model, the convolutional Bi-LSTM model, and the WDCNN model with minor standard deviation fluctuations as SNR increases. Although the classification accuracy of MSCNN under noise is also very high, its standard deviation fluctuates greatly, indicating that MSCNN is not as stable and robust as our ECT model. The results prove that ECT model has good generalization and can be robust to resist noise to some extent.

V. CONCLUSION

In this paper, an effective convolutional Transformer (ECT) for planetary gearbox fault diagnosis is developed. The model employs convolutional tokenization in the construction of the multi-stage Transformer to combine the benefits of both CNN and Transformer for local and global feature learning. And our proposed ECT method uses spatial-reduction attention (SRA) and linear dimension reduction projections to refine the model in two stages to reduce the Transformer's resource consumption while maintaining the high classification accuracy. Finally, the special structure combined with the sequence pooling strategy further simplifies the model by removing the class token and improves the accuracy of the model. The proposed ECT model allows for variable input data size and is effective for planetary gearbox fault diagnosis. Experiment results validate the effectiveness and flexibility of our ECT model, which is suitable for planetary gearbox fault diagnosis even under noise. Compared to other state-of-the-art methods, our ECT model is more effective and robust for planetary gearbox fault diagnosis, especially in learning long-range temporal dependencies.

REFERENCES

[1] Y. Lei, J. Lin, M. J. Zuo, and Z. He, "Condition monitoring and fault diagnosis of planetary gearboxes: A review," *Measurement*, vol. 48, pp. 292–305, Feb. 2014.

[2] Y. Lei, J. Lin, Z. He, and M. J. Zuo, "A review on empirical mode decomposition in fault diagnosis of rotating machinery," *Mech. Syst. Signal Process.*, vol. 35, pp. 108–126, Feb. 2013.

[3] R. Yan, R. X. Gao, and X. Chen, "Wavelets for fault diagnosis of rotary machines: A review with applications," *Signal Process.*, vol. 96, pp. 1–15, Mar. 2014.

[4] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.

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

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

[7] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao, "Deep learning and its applications to machine health monitoring," *Mech. Syst. Signal Process.*, vol. 115, pp. 213–237, Jan. 2019.

[8] Z. Chen and W. Li, "Multisensor feature fusion for bearing fault diagnosis using sparse autoencoder and deep belief network," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 7, pp. 1693–1702, Jul. 2017.

[9] M. Zhao, M. Kang, B. Tang, and M. Pecht, "Deep residual networks with dynamically weighted wavelet coefficients for fault diagnosis of planetary gearboxes," *IEEE Trans. Ind. Electron.*, vol. 65, no. 5, pp. 4290–4300, May 2018.

[10] 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, Jun. 2017.

[11] 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, Aug. 2017.

[12] J. Chen, R. Huang, K. Zhao, W. Wang, L. Liu, and W. Li, "Multiscale convolutional neural network with feature alignment for bearing fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–10, May 2021.

[13] R. Zhao, D. Wang, R. Yan, K. Mao, F. Shen, and J. Wang, "Machine health monitoring using local feature-based gated recurrent unit networks," *IEEE Trans. Ind. Electron.*, vol. 65, no. 2, pp. 1539–1548, Feb. 2018.

[14] Q. Xing-Yu, Z. Peng, X. Chengcheng, and F. Dong-Dong, "RNN-based method for fault diagnosis of grinding system," in *Proc. IEEE Int. Conf. Cyber Technol. Autom. Control Intell. Syst. (CYBER)*, 2017, pp. 673–678.

[15] R. Zhao, R. Yan, J. Wang, and K. Mao, "Learning to monitor machine health with convolutional bi-directional LSTM networks," *Sensors*, vol. 17 no. 2, p. 273, 2017.

[16] P. Luo, N. Hu, G. Shen, L. Zhang, and Z. Cheng, "DCNN with explicable training guide and its application to fault diagnosis of the planetary gearboxes," *IEEE Access*, vol. 8, pp. 122641–122653, 2020.

[17] 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.

[18] G. Jiang, H. He, J. Yan, and P. Xie, "Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox," *IEEE Trans. Ind. Electron.*, vol. 66, no. 4, pp. 3196–3207, Apr. 2019.

[19] S. Ma, F. Chu, and Q. Han, "Deep residual learning with demodulated time-frequency features for fault diagnosis of planetary gearbox under nonstationary running conditions," *Mech. Syst. Signal Process.*, vol. 127, pp. 190–201, Jul. 2019.

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

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

[22] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskeve, *Language Models Are Unsupervised Multitask Learners*, vol. 1, OpenAI Blog, San Francisco, CA, USA, 2019, p. 9.

[23] T. B. Brown *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33. Red Hook, NY, USA: Curran, 2020, pp. 1877–1901.

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

[25] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[26] H. Wu *et al.*, "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 22–31.

[27] K. Yuan, S. Guo, Z. Liu, A. Zhou, F. Yu, and W. Wu, "Incorporating convolution designs into visual transformers," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 579–588.

[28] 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*.

[29] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, "Early convolutions help transformers see better," in *Advances in Neural Information Processing Systems*, vol. 34. Red Hook, NY, USA: Curran, 2021.

[30] Z. Wu, Z. Liu, J. Lin, Y. Lin, and S. Han, "Lite transformer with long-short range attention," 2020, *arXiv:2004.11886*.

[31] W. Wang *et al.*, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 568–578.

[32] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[33] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*.