This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2023.3277956

Lightweight Convolutional Transformers Enhanced Meta Learning for Compound Fault Diagnosis of Industrial Robot

Chong Chen, Tao Wang, Chao Liu, Yuxin Liu, and Lianglun Cheng

***Abstract*—Recent advance of deep learning has seen remarkable progress in compound fault diagnosis modelling for industrial robots. Nevertheless, the data scarcity of compound fault samples jeopardizes the modelling performance of deep learning algorithms. Meta learning has become an effective tool in few-shot fault diagnosis modelling. However, due to the training instability of meta learning, it is challenging to deploy advanced networks such as Transformers as the base learner due to the extremely large model size. Therefore, this study proposes a lightweight convolutional Transformers (LCT) network enhanced meta learning (Meta-LCT) method to achieve accurate compound fault diagnosis with limited compound fault samples. Specifically, the LCT is firstly designed by taking advantage of linear spatial reduction (LSR) attention and spatial pooling mechanism to achieve high computational efficiency. LCT is adopted as the base learner in the Meta-SGD algorithm, and then the meta-training is performed based on the single fault data. Subsequently, the limited compound fault samples are used in the meta testing stage to obtain a compound fault diagnosis model. An experimental study based on the real-world compound fault dataset of industrial robots is presented. The experimental results indicate that the proposed Meta-LCT can achieve the compound fault diagnosis accuracy of 81.1% when only 40 data samples in each compound fault category are available.**

***Index Terms*—Compound fault diagnosis; Meta learning; Industrial robot; Transformers networks; Deep learning.**

I. INTRODUCT

INDUSTRIAL robots have become an essential part of the modern industry since they can automate processes and tasks that would otherwise be completed manually, leading to increased productivity and efficiency. Industrial robots can considerably reduce labor costs and increase the competitiveness of the manufacturing industry [1]. In recent years, the safety and reliability of industrial robots have gained increasing attention. The severe faults of industrial robots can cause costly damage to production facilities, safety hazards for workers, and delays in production. As a result, it is important for industrial robots to diagnose the parts with incipient faults and implement the maintenance in advance. During the operation of industrial robots, multiple faults can happen at the same time due to overload or accidents. The faulty patterns of compound faults are complex and hard to be diagnosed [2]. Therefore, it is necessary to investigate a compound fault diagnosis approach in order to optimize the operational and maintenance management of industrial robots.

With the rapid development of modern manufacturing, the requirement for fault diagnosis has shifted from individual components (e.g. bearings, gears, rotating shafts) to large assets such as industrial robots [3]. Recent advance in deep learning techniques has greatly improved the performance of intelligent fault diagnosis for industrial robots. The existing studies of fault diagnosis for industrial robots mainly focus on single fault diagnosis [1, 3], while few of the studies focused on compound fault diagnosis [2, 4]. However, the main challenge in investigating the compound fault diagnosis of industrial robots is the scarcity of compound fault samples, which are hard to be collected in the actual manufacturing scenarios. When the faulty samples are limited, transfer learning [5] and generative adversarial network (GAN) [6] can be deployed to address this issue. However, transfer learning requires a large number of faulty samples in the source domain to train a model for transfer, which is hard to be satisfied for industrial robots, while GAN requires a certain amount of compound fault samples to achieve Nash equilibrium. Both approaches are hard to get satisfactory performance when only dozens or fewer compound fault samples are available. Meta learning, known as learning to learn, has been widely studied in few-shot fault diagnosis modelling [7-15]. In comparison with GANs and transfer learning, meta learning does not rely on predetermined learning models or pre-trained weights. This makes meta learning highly efficient and effective in few-shot learning, as it allows the model to quickly adapt to new tasks with fewer data samples. Meta learning aims to seek the high-level meta knowledge underlying the given tasks, which enables the quick learning of a base learner. Model-agnostic meta learning (MAML) as a prevailing meta learning algorithm, has been extensively investigated in recent years [16]. MAML is model-agnostic, which is insensitive to the architecture of neural networks. However, since the gradient descent of MAML is performed in both the inner loop and outer loop, the training complexity of MAML is much higher than the supervised learning tasks. Meta-SGD improved MAML by further considering the optimization of the learning rate, which can improve the training stability [17].

Our work is supported by multiple funds in China, including the Key Program of NSFC-Guangdong Joint Funds (U2001201, U1801263), Industrial core and key technology plan of ZhuHai City (ZH22044702190034HJL), the Natural Science Foundation of Guangdong Province (2020B1515120010). Our work is also supported by Guangdong Provincial Key Laboratory of Cyber-Physical System (2020B1212060069). (Corresponding author: Tao Wang)

Chong Chen, Tao Wang, Yuxin Liu, and Lianglun Cheng are with Guangdong Provincial Key Laboratory of Cyber-Physical System, Guangdong University of Technology, Guangzhou 510006, China (email: chenc2021@gdut.edu.cn, 1358113879@qq.com, wangtao_cps@gdut.edu.cn, LLCheng@gdut.edu.cn)

Chao Liu is with the College of Engineering and Physical Sciences, Aston University, Birmingham B47ET, U.K. (e-mail: c.liu16@aston.ac.uk).

This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2023.3277956

Most existing studies of meta learning adopt simple networks such as convolutional neural network (CNN) and autoencoder as the base learner to achieve better training stability. In recent years, researchers have developed lightweight models to achieve satisfactory fault diagnosis performance with lower computational cost, while most of them were developed based on CNN [18]. Transformers networks have emerged as a promising approach for fault diagnosis by effectively identifying faulty patterns from sensory data and capturing temporal dependencies [2]. Compared to CNN and recurrent neural network (RNN) algorithms, Transformers have demonstrated superior performance in identifying faulty patterns. However, Transformers model tends to be complex and has a larger model size, which requires a significant amount of labeled data for training. The lightweight design of Transformers for meta learning is still in its infancy. This may impede the application of Transformers in meta learning tasks, where only a small amount of labeled data is available. Consequently, there is a need to design a lightweight and effective Transformers network specifically for meta learning in compound fault diagnosis.

Therefore, this paper proposes a lightweight convolutional Transformers network that can be adapted in meta-SGD, which can achieve accurate compound fault diagnosis accuracy in few-shot learning. Firstly, the feedback current signals collected from the motors of industrial robots are transformed into time-frequency images via continuous wavelet transformation (CWT). Then, an LCT network is designed to get accurate compound fault diagnosis performance with lower computational costs. Subsequently, LCT is introduced into meta-SGD as a base learner for compound fault diagnosis modelling. The main technical contribution of this study is three-fold: (1) A new lightweight convolutional Transformers network that adopts LSR attention to reduce the computational complexity of multi-head self-attention; (2) A sequence pooling layer is further introduced into LCT to extract sequential patterns and reduce the size of feature maps; (3) A Meta-LCT compound fault diagnosis approach, which introduces LCT into Meta-SGD as the base learner, is proposed for compound fault diagnosis modelling when the compound fault samples are limited. The remainder of this paper is organized as follows. Section 2 reviews the latest research on few-shot learning for fault diagnosis and the research on meta learning for fault diagnosis. Section 3 introduces the preliminaries and Section 4 details the proposed Meta-LCT approach. The experimental setup is presented in Section 5 and the results are demonstrated in Section 6. Finally, Section 7 discusses and Section 8 concludes.

II. Literature Review

A. Few-shots Learning for Fault Diagnosis

Compound faults cannot be recognized by intelligent diagnosis methods with limited data on compound faults. Another solution for the few-shot issues is transfer learning. Wang et al. [19] proposed a transfer learning fault diagnosis approach. Firstly, in order to extract features from the sample pair, twin neural networks were designed for feature learning. Then a metric learning module was developed to predict the similarity of the sample pairs. The similarities of sample pairs in conjunction with the testing samples were then used in fault diagnosis. In order to overcome the challenge of simultaneously aligning cross-domain marginal distribution and conditional distribution, Xiao et al. [20] designed a loss function embedded with joint maximum mean discrepancy. Meanwhile, to prevent the negative transfer, a weight allocation system was proposed for each source-domain sample. Zhang et al. [21] developed an iterative matching network augmented with a selective sample reuse strategy. The unlabeled signals from the unseen conditions were assigned with pseudo labels and reused to train a matching network so to reduce the difference between the source domain and the target domain. Furthermore, a similarity metric based on cosine distance was designed to filter the samples with wrong pseudo labels.

Domain adaption is important in transfer learning. A multiscale weight-selection adversarial network was proposed to enhance the partial transfer performance in fault diagnosis. In this study, to estimate the possible Gaussian distributions of target-domain samples, a Gaussian mixture model is employed. The Gaussian distribution of each class of source-domain samples is then computed through maximum likelihood estimation. Finally, the Wasserstein distance is utilized to compute the class weights [22]. Qian et al. [23] proposed an improved joint distribution adaptation to enhance domain confusion for cross-machine fault diagnosis. In another study [24], a relationship transfer diagnosis framework was proposed to address domain discrepancy in actual diagnosis scenarios. This framework indirectly measures and reduces the distribution discrepancy between the source and target domains. He et al. [25] proposed a deep transfer multi-wavelet auto-encoder for intelligent fault diagnosis of gearboxes with few training samples. This method started by designing a new type of deep multi-wavelet auto-encoder to capture the key features of the vibration signals gathered from the gearbox. Secondly, similar auxiliary samples were then chosen based on a similarity measure, in order to pre-train a source model that is comparable to the target domain. Finally, the parameter knowledge obtained from the source model was transferred to the target model with only a few target training samples. Xia et al. [26] proposed a digital twin assisted deep transfer learning for few-shot machinery fault diagnosis modelling. Xie et al. [27] proposed a multi-task attention-guided network which can train a multi-task model when the data is limited. In this model, a task-shared network was designed to capture the global features. Subsequently, the extracted features were then sent into two subnetworks to achieve fault type identification and fault severity identification tasks. By taking advantage of multi-labels, the requirement of the labelled data amounts was lowered. He et al. [28] proposed an enhanced deep transfer auto-encoder to achieve bearing fault diagnosis in different machines. In this study, a nonnegative constraint was employed to enhance the reconstruction outcome of the loss function. Subsequently, the parameter weight of the trained network was transferred to a target model.

GAN also have been investigated to address the data imbalance issue in fault diagnosis. Li et al. [29] modified auxiliary classifier GAN by introducing Wasserstein distance into the loss function and used spectral normalization for the parameter update of the discriminator. Yang et al. [30] proposed an approach that combines a conditional GAN (CGAN) and a 2D-CNN for bearing fault diagnosis. To address the lack of sufficient training samples, CGAN was employed to generate fake samples. Afterwards, these training data were converted into time-frequency images and used to train a 2D-CNN for fault diagnosis. In order to achieve the fault diagnosis for the unseen compound fault, Xing et al. [31] proposed a zero-shot intelligent diagnosis method. Firstly, a label description space is constructed to identify the relationship between fault patterns. The label description space is then embedded between the feature space and the health condition label space. Thirdly, a linear supervised autoencoder is built to get the projection between feature space and label description space. Finally, by evaluating the similarity in feature space and label description space, the compound fault can be identified when only single fault training data is available. Oversampling is another effective way to address the data imbalanced issue. Zhang et al. [32] proposed a weighted minority oversampling which contains a data synthesis strategy to avoid generating incorrect or unnecessary samples and a deep autoencoder to select useful features. In order to enhance the performance of the deep autoencoder, the maximum correntropy and sparse penalty were introduced. Meanwhile, a self-adaptive learning rate was adopted to guarantee convergence performance. Finally, the decision tree was built based on the extracted features for fault diagnosis.

B. Meta Learning for Fault Diagnosis

The advance of meta learning has attracted increasing attention in recent years. It has been studied to address the few-shot issue in fault diagnosis. Zhang et al. [16] proposed a few-shot learning framework for bearing fault diagnosis based on MAML. Only limited data is needed to train a classifier whose performance is better than the Siamese network. Zhang et al. [7] introduced discriminant space optimization into MAML to improve the performance of few-shot bearing fault diagnosis. Firstly, tailored networks were designed to get the fault features from the rolling bearings. Then, a discriminant space loss was designed to enhance the clustering performance, which reveals more discriminant classification boundaries. Finally, the MAML was adopted to train a classifier for few-shot bearing fault diagnosis. Yang et al. [8] combined Fourier transform and recurrence plots to get domain-independent data representation. Subsequently, MAML strategy was used to train a cross-domain fault diagnosis model. The residual shrinkage network was adopted as the backbone of the network and the large-margin Gaussian Mixture was used as the loss function. Wu et al. [9] investigated meta-learning for few-shot sample fault diagnosis. Two situations which are conditions transfer and artificial-to-natural transfer were considered in the rotating machinery fault diagnosis modelling. Li et al. [10] proposed a MAML fault diagnosis approach to address the limited data issue. The data was first transformed into time-frequency images, which were then fed into a CNN for modelling. The N-way K-shot protocol was performed in the bearing dataset to evaluate the effectiveness of the meta learning approach.

Aiming at the few-shot issue and multi-label attributes of single-point faults, Yu et al. [11] proposed a fault diagnosis method of rolling bearing based on time-frequency signature matrix (T–FSM) feature and multi-label convolutional neural network with meta-learning (MLCML). Firstly, the T–FSM features sensitive to few-shot fault diagnosis of measured vibration signal were extracted. Then, a multi-label CNN (MLCNN) with a specific architecture was designed to identify the compound faults. Finally, to address the few-shot problem, a meta-learning strategy for learning initial network parameters that are responsive to task changes was incorporated into the MLCNN. Hu et al. [12] proposed a task-sequencing meta-learning algorithm. To begin, a meta-learning model was trained over learning tasks to gain an understanding of how to implement fault diagnosis. The acquired knowledge can be used to adapt and generalize with just a few examples when facing new tasks that have never been seen before. To account for the differences and connections between failures and diagnosis tasks, a task-sequencing algorithm was proposed to order meta-training tasks from easiest to most difficult. After evaluating the level of difficulty of each task, the proposed method started with the simplest tasks and applies the learned knowledge to more complex tasks. By gradually increasing the task difficulty, better knowledge adaptability was achieved. In order to improve the training stability of MAML, Chang et al. [13] introduced adaptive learning rates into MAML-based few-shot fault diagnosis modelling. the proposed adaptive learning rates for meta-training and fine-tuning were adjusted based on the distributions of extracted features can help to address the issues of overfitting and underfitting in few-shot learning.

Besides model-based meta learning, Metric-based meta learning also has gained increasing attention in recent years. Wang et al. [14] presented Feature Space Metric-based Meta-learning Model (FSM3), a combination of conventional supervised learning and episodic metric meta-learning that is able to leverage the attribute information from individual samples as well as the similarity information from sample groups. Meanwhile, a hybrid training strategy was designed to combine global supervised training and episodic training in the learned feature space. Zhang et al. [15] proposed a fault diagnosis method called Meta-Learning with Discriminant Space Optimization (MLDSO) for few-shot bearing fault diagnosis. Firstly, tailored networks are used to extract the fault features of the rolling bearing. Afterwards, the feature extractor was adjusted utilizing the discriminant space loss to improve the clustering of the same category's fault features and differentiate between different types of fault features. Then, a feature extractor and discriminant space optimizer were implemented to create an optimal feature discriminant space with high fault-tolerance, which enables metric-based meta-learning to effectively classify faults in new tasks, even with only a few known samples.

C. A Brief Summary

The few-shot issue has greatly impeded the deployment of the advanced compound gnosis approach. Researchers have investigated various approaches such as transfer learning, GAN, and oversampling strategies to overcome this challenge. However, most of the existing studies were developed based on deep learning, which still requires a certain amount of compound fault samples to achieve satisfactory modelling performance. Since an industrial robot contains a large number of parts, which leads to the number of types of compound faults can be dramatic. When the compound faulty samples are extremely limited, these approaches may not be suitable. Meta learning as a promising tool for few-shot learning, has shown its merits in fault diagnosis. Among various meta learning algorithms, MAML has been widely studied. However, the training process of MAML is more complex since the inner loop and outer loop of MAML need to be trained simultaneously. In order to get stable performance, the base learner of MAML tends to select the simple deep learning algorithm, which may not be able to fully explore the hidden patterns within data. As an improved version of MAML, meta-SGD shows better training stability. In recent years, Transformers network has shown its advantage in fault diagnosis, while its model size and computational complexity tend to be large, which impedes its application in meta learning. In order to achieve better compound fault diagnosis accuracy based on meta-SGD, it is worthwhile to design a lightweight Transformers network.

III. PRELIMINARIES

A. Vision Transformers Network

Vision Transformers (ViT) network is a type of transformer-based neural network that has been used to achieve image classification. In ViT, the input is first processed by patch embedding which divides an image into smaller patches and then represents those patches as vectors. Subsequently, the patches are then sent into Transformers blocks for sequence learning. In the Transformers block, the multi-head self-attention (MSA) layer is used to learn important features from various aspects. In the MSA layer, three matrices named Q, K and V are used to get the attention score. We denote input data as:

$$X = [x_1,\ x_2, \dots,\ x_n] \tag{1}$$

Then linear transformation is then implemented on the input data to yield the matrices Q, K and V:

$$Q = XW^q \tag{2}$$

$$K = XW^k \tag{3}$$

$$V = XW^v \tag{4}$$

, where W^q, W^k and W^v are trainable projection matrices.

Subsequently, the matrices are sent to scaled dot-product attention to calculating the attention score, which can be expressed as:

$$Head\_Score_1 = \frac{softmax\ (\mathrm{Q} \times K^T)\ \times \mathrm{V}}{\sqrt{d}} \tag{5}$$

, where d is a scalable factor.

B. Meta-SGD

Meta-SGD is a type of meta-learning algorithm for training machine learning models that aims to allow a single model to quickly adapt to new tasks using only a small number of training samples. Meta-SGD operates by learning a set of initial parameters, learning rates and optimization directions that are then used to quickly adapt to the new task by making small adjustments. In order to achieve accurate compound fault diagnosis accuracy based on limited compound fault samples, the single fault samples are first used in meta training to obtain a well-trained learner. In the meta testing stage, the limited data samples are then used for model adaption to get a fine-tuned model.

Meta-SGD assumes that the model's parameters are represented by an initial parameter vector θ and that the loss function is differentiable enough near θ to enable the use of a gradient-based learning approach. The base learner in Meta-SGD is denoted as f_θ and the task distribution is denoted as $p(T)$.

When a well-trained learner (model) is adapted to a new task T_i, the model parameters are updated to $\acute{\theta_\iota}$ after gradient descent in the training process. For each task in the tasks set, k samples are selected and the loss is computed and updated in the gradient, which can be expressed as:

$$\acute{\theta_\iota} = \theta - \alpha \circ \nabla_\theta L_{T_i}(f_\theta) \tag{6}$$

During the training process, the learnable parameter α is also updated. The adaptation term $\alpha \circ \nabla_\theta L_{T_i}(f_\theta)$ is a vector whose direction represents the update direction and its length represents the learning rate. Subsequently, the initial model parameters are updated by optimizing the performance of $f_{\acute{\theta_\iota}}$ across tasks. The meta-optimization objective in the training process can be expressed as:

$$\min_\theta \sum_{T_i \sim p(T)} L_{T_i} f_{\acute{\theta_\iota}} = \sum_{T_i \sim p(T)} L_{T_i}\left(f_\theta - \alpha \circ \nabla_\theta L_{T_i}(f_\theta)\right) \tag{7}$$

The stochastic gradient descent is then used to optimize the outer loop in the meta learning. The update of the initial model parameter θ can be expressed as:

$$\theta = \theta - \beta \nabla_\theta L_{T_i}\big(f_{\acute{\theta_\iota}}\big) \tag{8}$$

After the meta training and meta testing, a fine-tuned model can be obtained and used for the actual deployment.

IV. METHODOLOGY

A. Overall Flowchart

In this section, the details of the proposed Meta-LCT compound fault diagnosis approach are elaborated. The overall flowchart of the proposed approach is illustrated in Fig 1. In the first stage, the data is collected and pre-processed. The feedback current signals of the industrial robot are collected from the servo drivers of the transmission system. The feedback current data can be directly collected from the existing device, which does not require additional sensors for collection. Subsequently, the data is split into the compound fault dataset and the single fault dataset. The CWT is then implemented in both datasets to get the time-frequency images. Both meta training and meta testing stages require a training process and therefore both stages require support set for training and query set for testing. In this study, the single fault dataset is used for meta training and the compound fault dataset is used for meta testing. In the meta training stage, the single fault dataset is used to train a Meta-LCT model. The support set is used as the training data and the query set is used as the testing data. In the meta testing stage, the dataset of compound fault which has limited samples is then fed into the well-trained model yielded in the meta training stage. Finally, the model trained in meta testing stage can be used for compound fault diagnosis for industrial robots.

This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2023.3277956

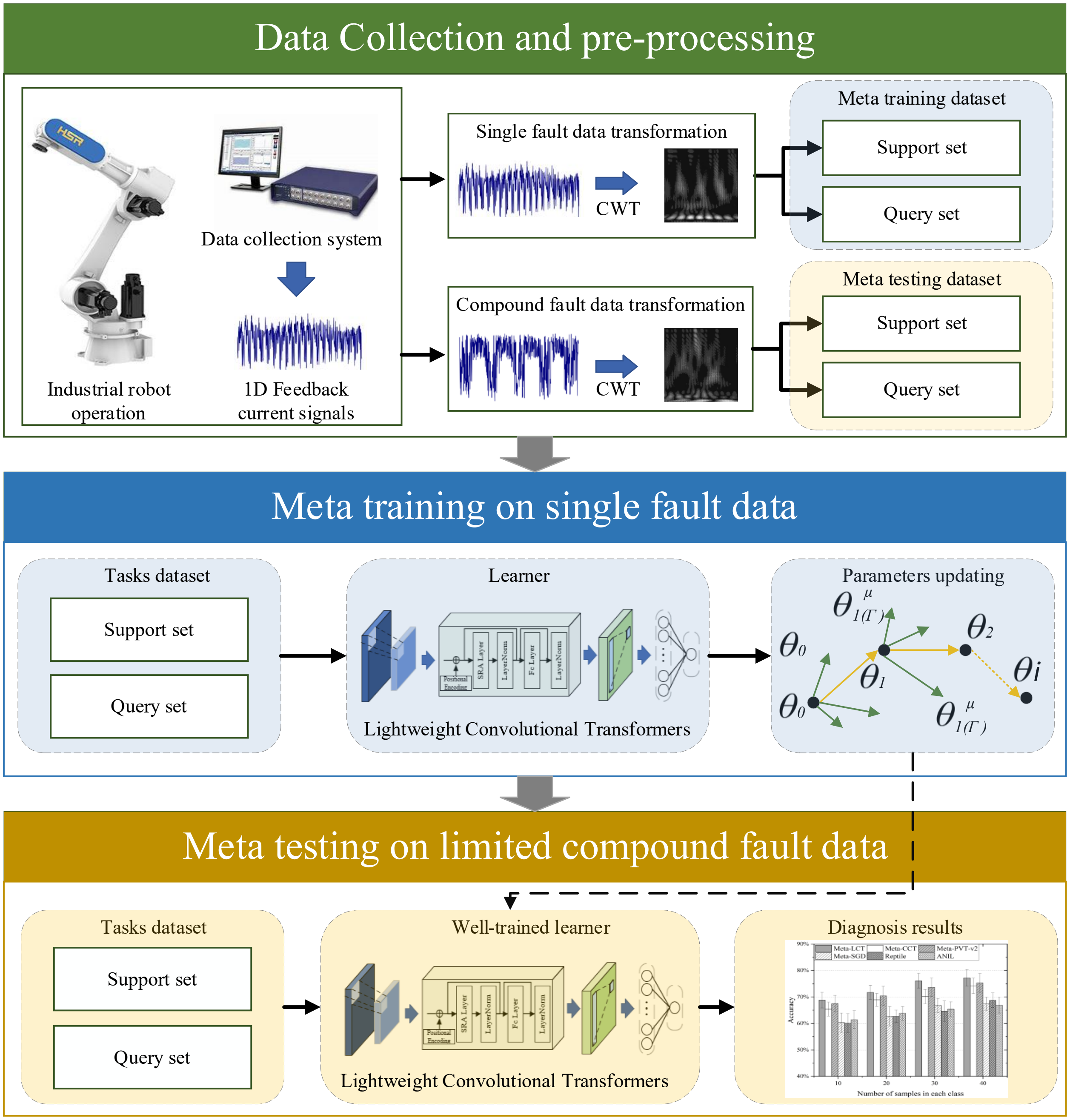

Fig 1. The flowchart of the proposed Meta-LCT compound fault diagnosis approach.

B. *Lightweight Convolutional Transformers Network*

Transformers network has been investigated in the fault diagnosis modelling and achieved satisfactory performance. However, as a type of large model, it requires extremely high computational cost and a large number of labelled data for training, which hinders its development in meta-learning. In order to introduce the Transformers network into meta learning, a lightweight design is needed. In this study, a new lightweight Transformers network is proposed, which is illustrated in Fig 2. The proposed LCT differs from the standard VIT in three ways. Firstly, a convolutional layer is adopted for overlap patch embedding. Secondly, LSR attention is introduced to increase the efficiency of multi-head attention [33]. Thirdly, a sequence pooling layer [34] is utilized to locate and extract sequential patterns from the hidden features. Patch embedding is essential in ViT to separate an image as different patches for feature extraction. However, the relation between the adjacent patches has not been considered during the patch embedding, which leads to information loss. In order to effectively capture the relation between the adjacent patches, a convolutional layer with zero padding is adopted.

Given a CWT image whose size is $h \times w \times c$. A convolutional layer with the stride S, the kernel size $2S-1$, and c' kernels. The output size of the convolutional layers becomes $(h/S) \times (w/S) \times c'$. After the patch features were obtained, layer normalization is then processed to reduce the

This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2023.3277956

internal covariate shift in model training. Subsequently, the patch features and the positional encoding are combined and sent to a lightweight Transformer block for global pattern learning.

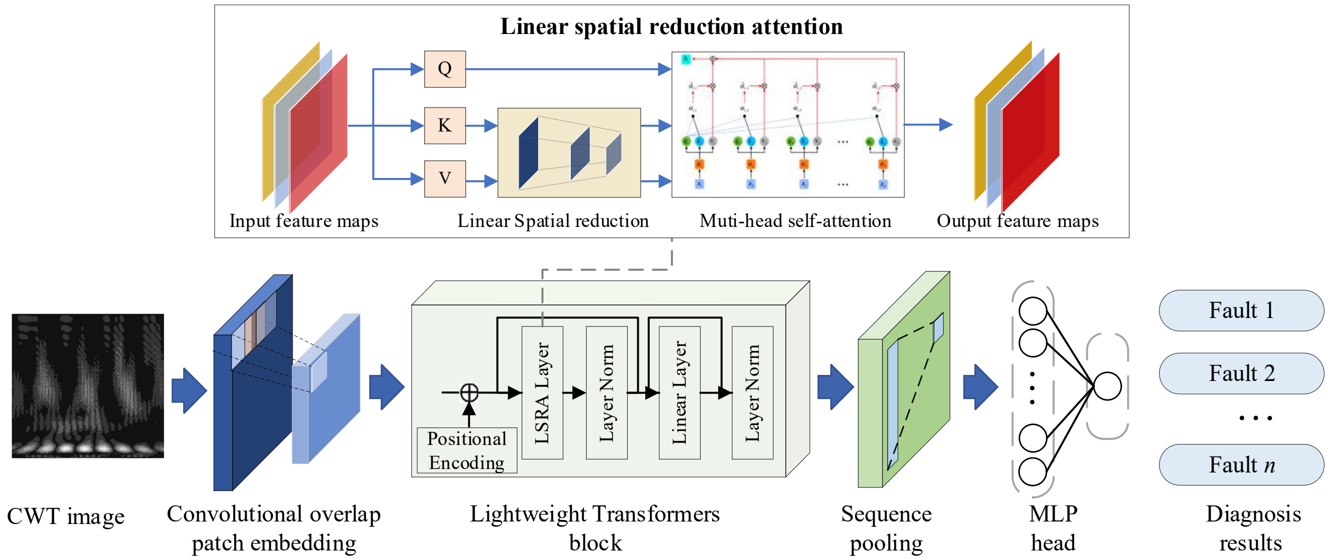

Fig 2. The structure of lightweight convolutional Transformers network.

The strong pattern learning capability of MSA allows Transformers to get the salient features from the data. As the number of heads grows, the performance tends to be better, while the computational complexity also grows rapidly. Hence, LSR attention was proposed to lower the computational complexity by reducing the dimension of K and V. Both matrices are then processed by the LSR module. The K and V are used as input in the LSR module, which is denoted as x_{KV}. The operation of the LSR module can be expressed as:

$$X_{KV} = PointwiseConv\big(AveragePool(x_{KV})\big) \quad (9)$$

where X_{KV} is the output of the LSR module and $PointwiseConv$ denotes the pointwise convolution operation. Other than that, the Gaussian error linear units (GELU) and layer normalization are adopted to improve the performance.

The LSR module can reduce the input feature map dimension from $h_0 \times w_0 \times c_0$ to the size $p \times p \times c_0$. The value of p can be determined in the experiment according to the algorithm performance. The output of the LSR module is then sent into the MSA layer for feature extraction. The output of the lightweight Transformer block can be expressed as:

$$x_T = f(x_0) \in \mathrm{R}^{b \times n \times d} \quad (10)$$

where the output dimension is $b \times n \times d$, b is the mini-batch size, n is the sequence length and d represents the embedding dimension. Finally, x_T is sent into the sequence pooling layer, which can be expressed as:

$$x_P = x_0 \times x_T = SoftMax(Linear(x_T)) \times x_T \in \mathrm{R}^{b \times 1 \times d} \quad (11)$$

The dimension n in x_T represents the sequence length. Standard vision Transformers flatten x_T, and then sent it to the MLP head for further processing, which loses a part of the sequential information. By pooling the data in sequence dimension, the sequence information can be kept and the computation cost in the subsequent MLP head can be reduced. Finally, the compound diagnosis results can be obtained.

V. Experimental Setup

A. Data Collection

To reveal the effectiveness of the proposed Meta-LCT compound fault diagnosis approach, a compound fault injection experiment was conducted to collect the faulty data of a six-axis industrial robot. In the fault injection experiment, the normal components were replaced with faulty reducers and motors which were collected from the actual manufacturing scenarios. The feedback current of the third axis was collected from the motor drivers of the transmission system. The faulty parts used in the experimental industrial robot exhibit incipient faults, such as minor abnormal noises and oil leakage. The incipient fault could potentially lead to more serious equipment failures in the future, while it is hard to diagnose since the faulty patterns are feeble. The robot with faulty parts was then deployed for the operation. The three-phase current signals were subsequently collected using a Hall element in the motor driver. The signals that were used for modelling were the currents of the q-axis, which frequency is 1Hz. The feedback current signals are demonstrated in Fig 3.

The collected dataset comprises seven categories, which were divided into three types of single faults, three types of compound faults and a normal category. The faulty components responsible for these faults were identified as reducers and motors in industrial robots. To label the data, one-hot encoding technique was employed. One-hot encoding is a popular method used in machine learning to convert categorical data into numerical data. It involves creating a binary vector where each category is represented by a unique index in the vector. By using one-hot encoding, each of the seven categories in the dataset was labeled for further analysis and modelling.

To generate the time-frequency images, the data is divided into segments of 100. Each segment was then used to get a time-frequency image whose dimension is 32*32 via CWT. After that, the dataset contains more than 2,000 samples. Specifically, there are 3000 samples in the normal and single fault categories, and there are over 1,800 samples in each compound fault category.

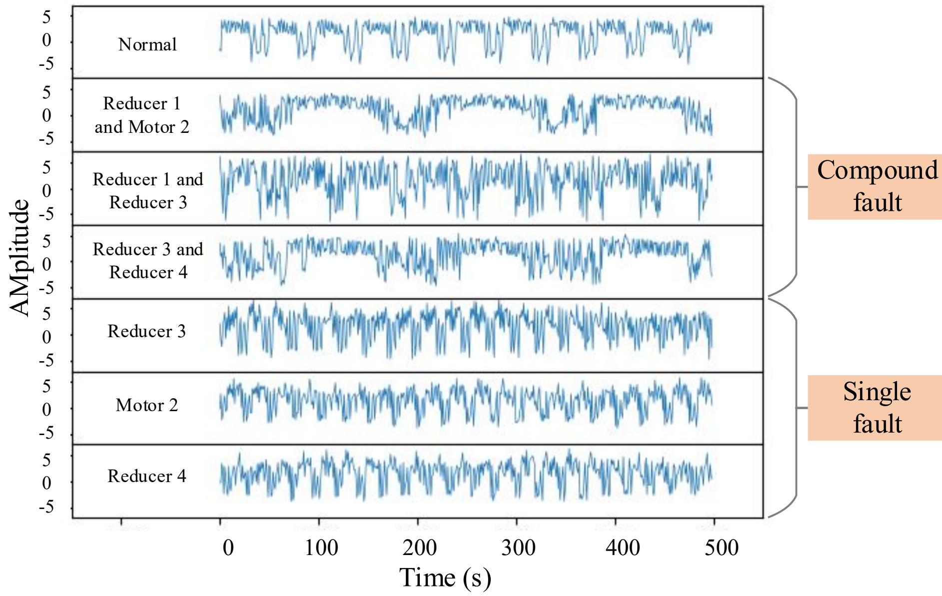

Fig 3. The comparison of signals from different faulty categories.

B. Benchmarking Experiments

In this study, the performance of LCT and Meta-LCT was evaluated via two experiments. In the first experiment, a benchmarking experiment was set up to reveal the cost-effectiveness of LCT. Seven prevailing algorithms were adopted. The training parameters of benchmarking algorithms were setup as the same as the LCT. The details of the benchmarking algorithms are introduced below:

1 **Pyramid Vision Transformer v2 (PVT-v2)** [33]: is a variant of VIT with lower computational cost. PVT adopts a progressive shrinking pyramid to reduce the computations of large feature maps. The lightweight version PVT v2-tiny was used in this study, which embedding dimensions in all four stages are 64, 128, 320, and 512.

2 **Compact Convolutional Transformer (CCT)** [34]**:** is a lightweight Transformers network which adopts convolutional block for feature reduction, Transformers block for global patterns learning and sequence pooling for sequential patterns learning. The number of convolutional layers was set as 2 and the number of Transformer encoders was set as 2.

3 **ViT-Tiny:** is a ViT whose hyperparameters are much smaller than the standard ViT. In order to get the similar size to LCT, the patch size was set as 4 and the depth of the network was set as 2.

4 **Swin Transformer (Swin-T)** [35]: is a variant of VIT. It uses shifted windows self-attention and patch merging to achieve higher computational efficiency. The number of heads in multi-head self-attention layers in all four stages were set as 2, 4, 6, and 8.

5 **DeIT** [36]: is a data-efficient Transformer network. Model distillation is used to reduce the network size and without sacrificing the model performance. In this experiment, the tiny version DeIT was adopted.

6 **Convolutional Block Attention Module-CNN (CBAM-CNN)** [37]: is a CNN network enhanced by CBAM. The CBAM module combines convolutional layers, attention layers, and a block structure to improve the performance of CNN.

7 **ResNet** [38]: is a deep CNN algorithm. ResNet uses residual connections to address the problems of gradient vanishing and gradient explosion while training deep convolutional neural networks.

LCT and the above algorithms were used for compound fault diagnosis modelling based on the full-size dataset which contains seven categories. The parameters of the networks such as learning rate, training epoch and the number of nodes were determined after multiple trials. The hyperparameters of LCT are detailed in Table I. Furthermore, Adam was adopted as the optimizer and label smooth cross entropy was used as the loss function. The overall accuracy for all the categories and the accuracy for compound fault categories were used as the evaluation metrics. Moreover, 5-fold cross-validation was adopted to get comprehensive results. Specifically, the mean and standard deviation of the outcomes from the 5-fold experiments were recorded as the result.

In the second experiment, the performance of Meta-LCT on compound fault diagnosis modelling was revealed. CCT and PVTv2 were used to replace Meta-LCT to reveal the performance of Meta-LCT. The combination of both benchmarking algorithms and Meta-SGD were denoted as Meta-CCT and Meta-PVTv2. Besides three prevailing meta learning algorithms which are standard Meta-SGD, Almost No

This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2023.3277956

Inner Loop (ANIL) [39], and Reptile [40] were adopted as the benchmarking algorithms. ANIL is a type of meta learning algorithm that requires minimal gradient updates during the meta-learning phase. Instead of updating the weights of the inner loop at every iteration, ANIL meta learning only updates the weights once per episode of training. Reptile is an optimization-based meta-learning algorithm which is designed as a variant of MAML, Reptile Meta Learning is designed to find a good initialization for a model that can be quickly adapted to new tasks by performing a few gradient updates on the new task.

In this experiment, the inner loop learning rate of meta learning was set as 0.001 and the outer loop learning rate of meta learning was set as 0.0001. There are three types of single faults and three types of compound faults in the collected dataset. The single fault dataset was used to construct the tasks set for meta training. The compound fault dataset was used to construct the tasks set for meta testing. The 3-way 1-shot and 3-way 5-shot experiments were implemented to reveal the algorithm performance in few-shot learning. The impact of dataset size on meta learning performance was revealed by using different sizes of the dataset for modelling. The data description for meta learning is presented in Table II. Besides the data used for modelling, the rest data was used for validation. Each experiment was conducted five times and the mean and standard deviation of the outcomes were recorded as the result. All tests were performed on an Ubuntu 16.0 server featuring an Intel i9-10920X 3.50Ghz CPU and an Nvidia GeForce RTX 3090 graphics card. The experimental environment is Python 3.8.12 and Pytorch 1.10.1 package was used to develop the algorithms. Besides, the overall accuracy, which is the accuracy of all the faulty categories and the accuracy for compound fault categories was used to reveal the algorithm performance. The model size and FLOPs (floating-point operations per second) were adopted to compare the computational efficiency in this experiment.

TABLE I
THE DETAILS OF HYPERPARAMETERS OF LCT.

Hyperparameter	Value
Learning rate	0.01
Embedding size	64
Number of epochs	200
Batch size	128
Number of Transformers encoder	2
Number of the head in MSA	4
Convolutional kernel size for patch embedding	7*7
Average pooling size	3

TABLE II
THE DATA DESCRIPTION FOR META LEARNING.

Stage	Meta training			Meta testing		
Faulty type	Reducer 3	Motor 2	Reducer 4	Reducer 1 and Motor 2	Reducer 1 and Reducer 3	Reducer 3 and Reducer 4
Sample number	1,000	1,000	1,000	10/20/30/40	10/20/30/40	10/20/30/40

TABLE III
THE COMPARISON OF THE COMPUTATIONAL COST AND ALGORITHM PERFORMANCE

Algorithms	#Params (M)	FLOPs (M)	Overall accuracy (%)	Accuracy for compound fault categories (%)
LCT (proposed)	**0.95**	**13.6**	**81.9±2.4**	**62.3±2.8**
PVT-v2	13.24	41.3	81.5±2.2	56.3±3.4
CCT	1.05	28.9	79.1±2.8	58.9±1.9
ViT-Tiny	1.26	22.5	76.9±2.7	55.3±3.3
Swin Transformers	18.91	2815.4	80.9±3.9	59.4±3.7
DeiT	20.92	1081.6	80.2±2.7	60.7±2.5
CBAM-CNN	7.14	16.6	81.6±1.9	56.2±2.5
ResNet	8.18	105.9	81.3±2.1	49.3±1.3

VI. EXPERIMENTAL RESULTS

A. The Comparison between LCT and the Benchmarking Algorithms

The algorithm performance, the model size and the FLOPs of LCT are revealed in this experiment. Table III displays the modelling results of all the algorithms based on the full-size training set. It can be seen that LCT achieves the highest overall accuracy and accuracy for compound fault categories. The leading margin of overall accuracy is marginal, while the accuracy for compound fault categories achieved by LCT is 3.4% higher than the second place in all the algorithms. Moreover, the model size of LCT is smaller than other algorithms, which is mere 0.95m. In striking contrast, the tiny version of PVT-v2 requires 13.2M parameters for modelling. The model size of LCT is similar to tiny-VIT, while its FLOPs is 65% lower than that of LCT. However, the performance tiny-VIT is the worst in this experiment, in which overall

This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2023.3277956

accuracy and the accuracy for compound fault categories are 76.9% and 55.3%, respectively. Swin Transformers and DeIT are large models. The FLOPs of both algorithms are 2815.4M and 1081.6M, which is far higher than the rest of the algorithms. The overall accuracy of Swin Transformers and DeIT are 80.9% and 80.2%, which is slightly worse than that of LCT, PVT-v2 and CCT. Besides, the overall accuracy of CBANM-CNN and ResNet are 81.6% and 81.3%, respectively. However, the accuracy for compound fault categories are lower than most of the Transformers-based algorithms.

The algorithm performance of LCT is further explored under different training data sizes. The results were plotted in Fig 4. With the increase in training data size, the overall accuracy of LCT increases steadily. However, the accuracy for compound fault categories does not increase until the training data size reaches 8,000. After that, the accuracy for compound fault categories improves rapidly.

B. The Comparison between Meta-LCT and Other Meta Learning Algorithms

The algorithm performance of Meta-LCT was evaluated in this section. Firstly, the experiment of 3-way 1-shot was performed and the results are presented in Fig 5. It can be seen that the accuracy achieved by Meta-LCT is the highest in all the stages. When the data samples in each compound fault category arrive at 40, Meta-LCT can achieve the compound fault diagnosis accuracy which is around 77%. With the introduction of advanced algorithms such as LCT, CCT and PVT-v2, the algorithm performance of meta learning are obviously higher than that of the prevailing meta learning algorithms which are Meta-SGD, reptile and ANIL. The leading accuracy is up to 5%. Meanwhile, with the increase in the number of samples in each class, the performance of all the algorithms were promoted.

The experimental results of 3-way 5-shots are reported in Fig 6. The performance of all the algorithms was promoted in this experiment in comparison with the experiment of 3-way 1-shot. When 10 samples in each category are available, Meta-LCT achieves the compound fault diagnosis accuracy which is 71.8%, which is 3% higher than that of the result obtained from the 3-way 1-shot experiment. When the data samples in each category reach 40, Meta-LCT achieves the best compound fault diagnosis accuracy which is 81.1%. The performance of Meta-CCT is slightly lower than that of Meta-LCT in most stages. When the data samples in each category are 20, the performance of Meta-CCT marginally surpasses Meta-LCT.

In order to further reveal the algorithm performance of all the algorithms, the accuracy and loss curve in the meta testing stage of all the algorithms under 3-way 1-shot condition are plotted. The number of samples in each class is 40. It is obvious that the loss of Meta- LCT decreases smoothly in the training stage, with the compound fault diagnosis accuracy improving to around 77% when the training epoch reaches 800. The accuracy curve of Meta-CCT is similar to that of Meta-LCT, while its loss is relatively unstable in the last stage of training. Meta-PVT v2 shows the most moderate varying trend in the loss and accuracy in this experiment. Meanwhile, the loss curves of Meta-SGD, reptile and ANIL are struggling in the beginning stage, before dropping rapidly when the training epoch reaches 100. The change in both loss and accuracy of these algorithms was marginal in the second half of the figures.

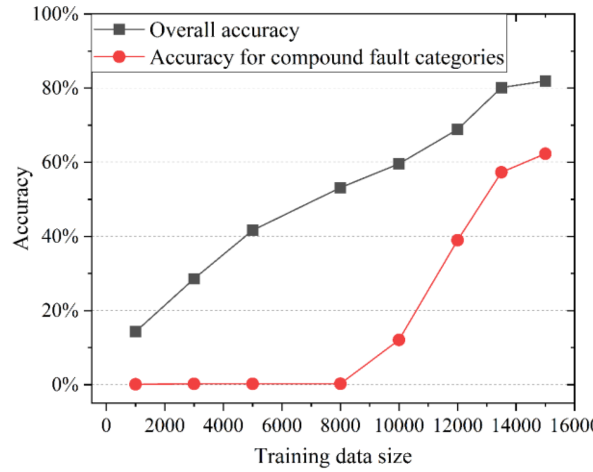

Fig 4. The algorithm performance of LCT along with the increase of training data size.

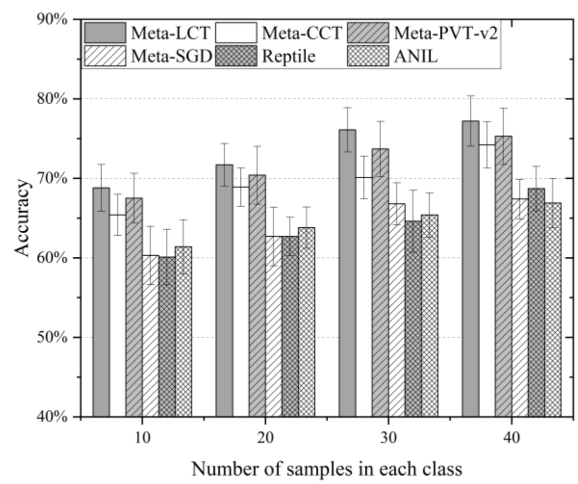

Fig 5. The algorithm performance of different algorithms under 3-way 1-shot condition along with different training data sizes in each class.

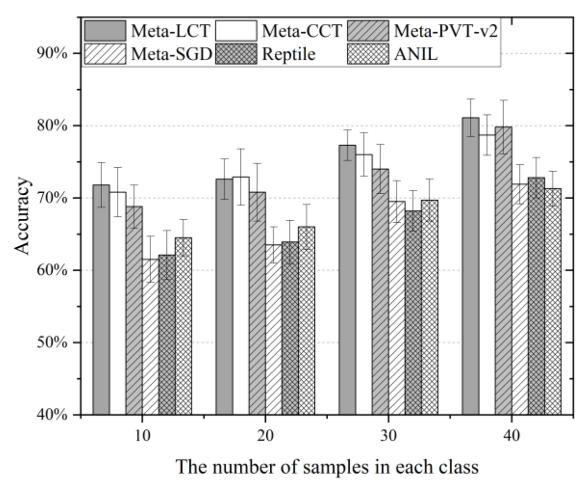

Fig 6. The algorithm performance of different algorithms under 3-way 5-shots condition along with different training data size in each class.

This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2023.3277956

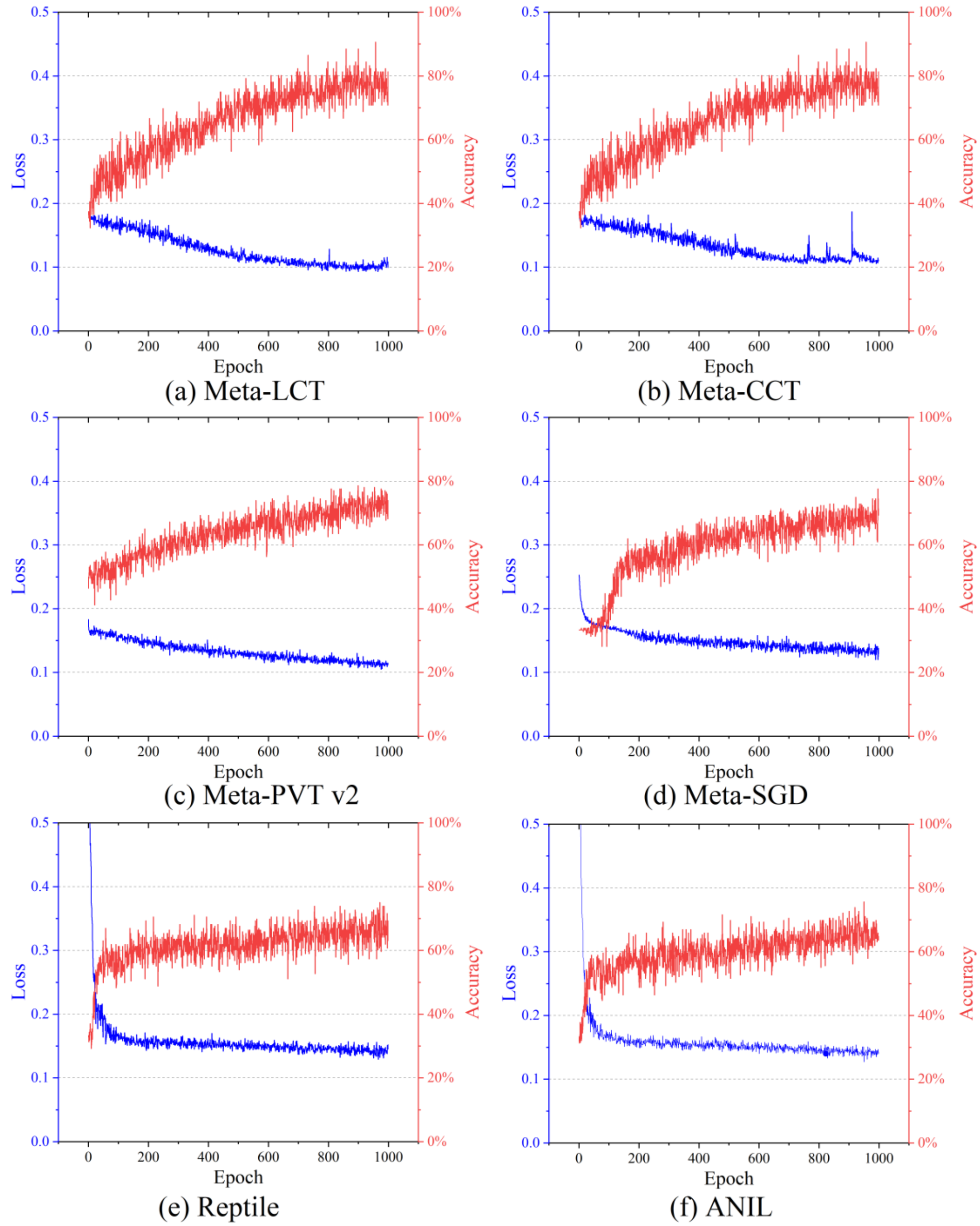

Fig 7. The accuracy and loss curves of different algorithms under 3-way 1-shots condition.

VII. Discussion

The lightweight design of Transformers is helpful to its deployment in meta learning. From the results of Table III, the proposed LCT shows merits in compound fault diagnosis modelling in comparison with the prevailing algorithms. Hence, LCT is suitable to be adopted in meta learning. Meanwhile, the training data size test in the first experiment indicates that LCT cannot get satisfactory compound fault diagnosis accuracy when the training data is lower than 8,000. Transformers is a type of data-hungry algorithm that requires a large number of diversified samples for training. In the supervised modelling using the Transformers network, millions of samples can be needed to ensure the performance of modelling. However, it is hard to directly deploy LCT in the supervised learning task to establish an accurate compound fault diagnosis model. Meanwhile, the compound fault samples of industrial robots are hard to be collected, which further impedes the deployment of LCT in compound fault diagnosis modelling.

The results of 3-way 1-shot and 3-way 5-shots indicate that the Meta-LCT can achieve better compound fault diagnosis accuracy in comparison with other benchmarking algorithms. When meta learning was performed, only 10 samples in each compound fault category are needed to get over 70% compound fault diagnosis accuracy. The results obtained from the condition of 3-way 5-shots are generally better than that of 3-way 1-shot. The reason is that the tasks in 3-way 5-shots contain more data instances, which can provide more information to the algorithms. Accordingly, the computational cost of 3-way 5-shots modelling is higher than that of 3-way 1-shot modelling. Meanwhile, compared with the prevailing meta learning algorithms that used CNN as the backbone model, Meta-LCT, Meta-CCT and Meta-PVT v2 show better performance, which indicates that lightweight and effective networks can greatly promote the performance of meta learning.

Even though the proposed Meta-LCT approach can achieve satisfactory compound fault diagnosis accuracy when the labelled data is limited, there are two limitations that need to be further addressed. Firstly, the meta learning consists of meta training and meta-testing stages. The meta training in this study requires 3,000 epochs and the meta testing requires 1,000 training epochs. The reason for adopting such a large number of training epochs is that training a Transformers network is time-consuming. Transformers network is complex, which increases its training instability. Hence, it is worthwhile to investigate an optimization approach that can achieve rapid convergence in meta training and meta testing. Meanwhile, Meta-LCT was only deployed in the compound fault diagnosis modelling without considering the working conditions. Industrial robot is a type of asset that can work in various conditions. The performance of Meta-LCT in cross-condition compound fault diagnosis modelling will be revealed in the future.

VIII. Conclusions

The maintenance of industrial robots can be greatly improved by the correct diagnosis of the compound fault. Transformers network is hard to be deployed in meta learning due to its large size and training instability. In order to overcome this challenge, a lightweight and effective Transformers network called LCT was proposed in this study and it was introduced into the Meta-SGD algorithm as the baseline model. A real-world industrial robot compound fault dataset was used to conduct an experimental study, which reveals the effectiveness of the proposed Meta-LCT compound fault diagnosis approach. The key findings of this research are: (1) In comparison with those state-of-the-art algorithms, the algorithm performance of LCT in compound fault diagnosis modelling is better and its model size is smaller; (2) With the introduction of LCT, Meta-LCT shows merits in compound fault diagnosis in comparison with the prevailing meta learning algorithms; (3) Meta-LCT can achieve over 70% compound fault diagnosis accuracy when only limited compound fault samples are available. The proposed Meta-LCT can bring tangible benefits to engineers for the maintenance of industrial robots.

References

[1] B. Xia, K. Wang, A. Xu, P. Zeng, N. Yang, and B. Li, "Intelligent Fault Diagnosis for Bearings of Industrial Robot Joints Under Varying Working Conditions Based on Deep Adversarial Domain Adaptation," *IEEE Transactions on Instrumentation and Measurement,* vol. 71, pp. 1-13, 2022, doi: 10.1109/TIM.2022.3158996.

[2] C. Chen, C. Liu, T. Wang, A. Zhang, W. Wu, and L. Cheng, "Compound fault diagnosis for industrial robots based on dual-transformer networks," *Journal of Manufacturing Systems,* vol. 66, pp. 163-178, 2023.

[3] K. Lu, C. Chen, T. Wang, L. Cheng, and J. Qin, "Fault diagnosis of industrial robot based on dual-module attention convolutional neural network," *Autonomous Intelligent Systems,* vol. 2, no. 1, pp. 1-12, 2022.

[4] X. Zhou, H. Zeng, C. Chen, H. Xiao, and Z. Xiang, "An attention-enhanced multi-modal deep learning algorithm for robotic compound fault diagnosis," *Measurement Science and Technology,* vol. 34, no. 1, p. 014007, 2022.

[5] Y. Liu, C. Chen, T. Wang, and L. Cheng, "An attention enhanced dilated CNN approach for cross-axis industrial robotics fault diagnosis," *Autonomous Intelligent Systems,* vol. 2, no. 1, pp. 1-11, 2022.

[6] M. Chen, H. Shao, H. Dou, W. Li, and B. Liu, "Data augmentation and intelligent fault diagnosis of planetary gearbox using ILoFGAN under extremely limited samples," *IEEE Transactions on Reliability,* vol. 15, pp. 1-9, 2022.

[7] D. Zhang, K. Zheng, B. Yin, D. Yao, D. Yang, and S. Wang, "Few-shot bearing fault diagnosis based on meta-learning with discriminant space optimization," *Measurement Science and Technology,* 2022.

[8] T. Yang, T. Tang, J. Wang, C. Qiu, and M. Chen, "A novel cross-domain fault diagnosis method based on model agnostic meta-learning," *Measurement,* vol. 199, p. 111564, 2022.

[9] J. Wu, Z. Zhao, C. Sun, R. Yan, and X. Chen, "Few-shot transfer learning for intelligent fault diagnosis of machine," *Measurement,* vol. 166, p. 108202, 2020.

[10] C. Li, S. Li, A. Zhang, Q. He, Z. Liao, and J. Hu, "Meta-learning for few-shot bearing fault diagnosis under complex working conditions," *Neurocomputing,* vol. 439, pp. 197-211, 2021.

[11] C. Yu, Y. Ning, Y. Qin, W. Su, and X. Zhao, "Multi-label fault diagnosis of rolling bearing based on meta-learning," *Neural Computing and Applications,* vol. 33, no. 10, pp. 5393-5407, 2021.

[12] Y. Hu, R. Liu, X. Li, D. Chen, and Q. Hu, "Task-sequencing meta learning for intelligent few-shot fault diagnosis with limited data," *IEEE Transactions on Industrial Informatics,* vol. 18, no. 6, pp. 3894-3904, 2021.

[13] L. Chang and Y.-H. Lin, "Meta-Learning With Adaptive Learning Rates for Few-Shot Fault Diagnosis," *IEEE/ASME Transactions on Mechatronics,* 2022.

[14] D. Wang, M. Zhang, Y. Xu, W. Lu, J. Yang, and T. Zhang, "Metric-based meta-learning model for few-shot fault diagnosis under multiple limited data conditions," *Mechanical Systems and Signal Processing,* vol. 155, p. 107510, 2021.

[15] D. Zhang, K. Zheng, Y. Bai, D. Yao, D. Yang, and S. Wang, "Few-shot bearing fault diagnosis based on meta-learning with discriminant space optimization," *Measurement Science and Technology,* vol. 33, no. 11, p. 115024, 2022.

[16] S. Zhang, F. Ye, B. Wang, and T. G. Habetler, "Few-shot bearing fault diagnosis based on model-agnostic meta-learning," *IEEE Transactions on Industry Applications,* vol. 57, no. 5, pp. 4754-4764, 2021.

[17] Z. Li, F. Zhou, F. Chen, and H. Li, "Meta-sgd: Learning to learn quickly for few-shot learning," *arXiv preprint arXiv:1707.09835,* 2017.

[18] Z. Fan, X. Xu, R. Wang, and H. Wang, "Fan Fault Diagnosis Based on Lightweight Multiscale Multiattention Feature Fusion Network," *IEEE Transactions on Industrial Informatics,* vol. 18, no. 7, pp. 4542-4554, 2022, doi: 10.1109/TII.2021.3121294.

[19] C. Wang and Z. Xu, "An intelligent fault diagnosis model based on deep neural network for few-shot fault diagnosis," *Neurocomputing,* vol. 456, pp. 550-562, 2021.

[20] Y. Xiao, H. Shao, S. Han, Z. Huo, and J. Wan, "Novel Joint Transfer Network for Unsupervised Bearing Fault Diagnosis From Simulation Domain to Experimental Domain," *IEEE/ASME Transactions on Mechatronics,* 2022.

[21] K. Zhang, J. Chen, T. Zhang, S. He, T. Pan, and Z. Zhou, "Intelligent fault diagnosis of mechanical equipment under varying working condition via iterative matching network augmented with selective Signal reuse strategy," *Journal of Manufacturing Systems,* vol. 57, pp. 400-415, 2020.

[22]Q. Qian, Y. Qin, J. Luo, and S. Wang, "Partial Transfer Fault Diagnosis by Multiscale Weight-Selection Adversarial Network," *IEEE/ASME Transactions on Mechatronics,* vol. 27, no. 6, pp. 4798-4806, 2022, doi: 10.1109/TMECH.2022.3166977.

[23]Q. Qian, Y. Qin, J. Luo, Y. Wang, and F. Wu, "Deep discriminative transfer learning network for cross-machine fault diagnosis," *Mechanical Systems and Signal Processing,* vol. 186, p. 109884, 2023.

[24]Q. Qian, J. Zhou, and Y. Qin, "Relationship Transfer Domain Generalization Network for Rotating Machinery Fault Diagnosis Under Different Working Conditions," *IEEE Transactions on Industrial Informatics,* 2023.

[25]Z. He, H. Shao, P. Wang, J. J. Lin, J. Cheng, and Y. Yang, "Deep transfer multi-wavelet auto-encoder for intelligent fault diagnosis of gearbox with few target training samples," *Knowledge-Based Systems,* vol. 191, p. 105313, 2020.

[26]M. Xia, H. Shao, D. Williams, S. Lu, L. Shu, and C. W. de Silva, "Intelligent fault diagnosis of machinery using digital twin-assisted deep transfer learning," *Reliability Engineering & System Safety,* vol. 215, p. 107938, 2021.

[27]Z. Xie, J. Chen, Y. Feng, K. Zhang, and Z. Zhou, "End to end multi-task learning with attention for multi-objective fault diagnosis under small sample," *Journal of Manufacturing Systems,* vol. 62, pp. 301-316, 2022.

[28]H. Zhiyi, S. Haidong, J. Lin, C. Junsheng, and Y. Yu, "Transfer fault diagnosis of bearing installed in different machines using enhanced deep auto-encoder," *Measurement,* vol. 152, p. 107393, 2020.

[29]W. Li, X. Zhong, H. Shao, B. Cai, and X. Yang, "Multi-mode data augmentation and fault diagnosis of rotating machinery using modified ACGAN designed with new framework," *Advanced Engineering Informatics,* vol. 52, pp. 1-16, 2022.

[30]J. Yang, J. Liu, J. Xie, C. Wang, and T. Ding, "Conditional GAN and 2-D CNN for Bearing Fault Diagnosis With Small Samples," *IEEE Transactions on Instrumentation and Measurement,* vol. 70, pp. 1-12, 2021.

[31]J. Xu, L. Zhou, W. Zhao, Y. Fan, X. Ding, and X. Yuan, "Zero-shot learning for compound fault diagnosis of bearings," *Expert Systems with Applications,* vol. 190, p. 116197, 2022/03/15/ 2022, doi: https://doi.org/10.1016/j.eswa.2021.116197.

[32]Y. Zhang, X. Li, L. Gao, L. Wang, and L. Wen, "Imbalanced data fault diagnosis of rotating machinery using synthetic oversampling and feature learning," *Journal of manufacturing systems,* vol. 48, pp. 34-50, 2018.

[33]W. Wang *et al.*, "Pvt v2: Improved baselines with pyramid vision transformer," *Computational Visual Media,* vol. 8, no. 3, pp. 415-424, 2022.

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

[35] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012-10022.

[36] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*, 2021: PMLR, pp. 10347-10357.

[37]J. Li, Y. Liu, and Q. Li, "Intelligent fault diagnosis of rolling bearings under imbalanced data conditions using attention-based deep learning method," *Measurement,* vol. 189, p. 110500, 2022.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[39]A. Raghu, M. Raghu, S. Bengio, and O. Vinyals, "Rapid learning or feature reuse? towards understanding the effectiveness of maml," *arXiv preprint arXiv:1909.09157,* 2019.

[40] A. Raghu, M. Raghu, S. Bengio, and O. Vinyals, "Rapid Learning or Feature Reuse? Towards Understanding the Effectiveness of MAML," in *International Conference on Learning Representations*, 2019.

Chong Chen received PhD degree from the School of Engineering, Cardiff University, UK in 2021.

He is currently a research associate with Guangdong Provincial Key Laboratory of Cyber-Physical System, Guangdong University of Technology. He has published over 20 papers in international journals and conferences. His research interests mainly focus on prognostic and health management, artificial intelligence, and the industrial big data. He is a Managing Guest Editor for Autonomous Intelligent Systems.

Tao Wang received PhD degree from Sun Yat-sen University, in 2010.

He is currently an Associate Professor with the School of Automation, Guangdong University of Technology. His research interests include manufacturing the Internet of Things, industrial big data and knowledge acquisition, intelligent equipment, and robot systems.

Chao Liu is a Lecturer in the College of Engineering and Physical Sciences at Aston University. He received his Ph.D. in Mechanical Engineering at the University of Auckland, New Zealand. He obtained his Bachelor's degree and Master's degree in Mechanical Engineering both from Northeastern University, China.

Prior to joining Aston University, Dr. Liu has worked as a Research Assistant Professor in the Department of Industrial and Systems Engineering at The Hong Kong Polytechnic University in Hong Kong, and as a Postdoctoral Research Associate in the School of Engineering at Cardiff University, UK. He is a member of SME and a Managing Guest Editor for Robotics and Computer-Integrated Manufacturing. His research interests include Industry 4.0, smart manufacturing, machine tools, digital twin, CPS, IoT, big data analytics, AR, machine learning, metal additive manufacturing, etc.

Yuxin Liu received the B.Eng degree in Guangdong University of Technology , Guangdong, China in 2021.

He is currently pursuing the M.S. degree in computer technology with the school of Computer, Guangdong University of Technology, Guangdong, China. He current research interests include process monitoring, fault diagnosis and meta learning.

Lianglun Cheng received the master's degree in automation from the Huazhong University of Science and Technology and the Ph.D. degree in machinery manufacturing and automation from the Changchun Institute of Optical Precision Machinery and Physics, Chinese Academy of Sciences.

He is currently a Professor with the school of Computer, Guangdong University of Technology. His research interests include the IoT, CPS, and sensor networks.