Compound Fault Diagnosis for Industrial Robots based on Dual-Transformer Networks

Chong Chen[a], Chao Liu[b], Tao Wang[a, c*], Ao Zhang[a], Wenhao Wu[a], and Lianglun Cheng[a]

[a] Guangdong Provincial Key Laboratory of Cyber-Physical System, Guangdong University of Technology, Guangzhou 510006, China

[b] College of Engineering and Physical Sciences, Aston University, Birmingham B47ET, UK

[c] School of Automation, Guangdong University of Technology, Guangzhou 510006, China.

[*]Corresponding author. E-mail: wangtao_cps@gdut.edu.cn

Abstract

The accurate diagnosis of the compound fault of industrial robots can be highly beneficial to maintenance management. In the actual noisy working environment of industrial robots, the mixed and feeble failure features are easy to be overwhelmed, which poses a major challenge for the industrial robot compound fault diagnosis. Meanwhile, in the existing studies, a large-size deep learning model is the guarantee of decent denoising and fault diagnosis performance. However, this demands expensive computational costs and large data samples, which are not always available. In order to address both challenges, in this study, an integrated approach that contains two compact Transformer networks is proposed to achieve accurate compound fault diagnosis for industrial robots. In this approach, the feedback current signals collected from a six-axis industrial robot are first transformed into time-frequency image representation via continuous wavelet transformation (CWT). Secondly, a novel deep learning algorithm called compact Uformer is proposed to denoise the time-frequency image. Subsequently, the denoised time-frequency images are fed into compact convolutional Transformer (CCT) for compound fault diagnosis. An experimental study based on a real-world industrial robot compound fault dataset was conducted. The experimental results reveal that the proposed method can achieve satisfactory compound fault diagnosis accuracy based on the data collected from the noisy environment in comparison with the state-of-the-art algorithms.

***Keywords**:* Compound fault diagnosis; Industrial robot; signal denoising; Deep learning; Transformer network.

1. Introduction

Industrial robots have long been employed in automated manufacturing processes in order to improve productivity, quality, and safety [1]. They have been widely used in various scenarios of the industry, including assembling [2] and welding [3]. In the actual industrial application, multiple components in an industrial robot may fail at the same time due to overload or accidents. The transmission system is the core of an industrial robot. The components such as the motor and reducer in the transmission system are tightly coupled, which poses a great challenge in compound fault diagnosis. If the compound fault cannot be diagnosed accurately, it is likely that the robot with unidentified fault cannot perform its task and would lead to the stoppage of the production line [4]. With the accurate diagnosis of compound fault, appropriate maintenance can be scheduled so to lower the maintenance cost. Hence, it is worthwhile to investigate an accurate robotic compound fault diagnosis approach.

Among numerous research about compound fault diagnosis, most research focuses on rotating machinery such as bearings [5-7] and gearboxes [8, 9]. These studies were conducted based on the vibration signal collected from the accelerators. Through the modelling using deep learning and advanced signal processing approaches, the failure features of compound fault can be identified, and therefore achieve accurate diagnosis performance. However, in the actual implementation, it is challenging to install the accelerators into a mechanical system. Furthermore, there are various machines in the workshop and each of them has multiple components. The vibration signal collected by the accelerator is mixed with strong noise generated by other coupling components and the machines nearby, which overwhelm the actual failure features. For the industrial robot, the existing studies have barely focused on the compound fault diagnosis. Besides conducting single fault diagnosis using vibration data [10], researchers have focused on single fault diagnosis based on attitude data [11], simulated joint angle and current data [12-14], power consumption data [15] and so on. The choice of the data source depends on the specific diagnosis part and the data availability. For instance, To et al. [16] proposed a fault diagnosis approach for grit-blasting based on the data collected from RGB-D camera, audio and pressure transducers to diagnose grit-blasting spot position and the state of blasting. Servo motor, due to its high precision and controllability, has been widely deployed in industrial robots. The feedback current of the servo motor can be collected from the driver, which is more accessible in comparison with most data types mentioned above. The feedback current may record the failure patterns when the failures of the motor or reducer occur. However, there are no existing studies that report the fault diagnosis based on the feedback current data.

The denoising of the monitoring data is required to achieve compound fault diagnosis based on the data collected from the noisy environment. The prevailing approach for the signal denoising approach can be classified into two types: signal-processing approaches [17-20] and data-driven approaches [21, 22]. The signal processing approaches such as CWT, empirical wavelet transforms (EWT) and empirical mode decomposition (EMD) tends to separate the noisy component from the signal, which requires extra domain knowledge. In contrast, the data-driven approach mainly focuses on learning the patterns of noisy components and achieved automatically denoising via autoencoder structure neural networks. However, most deep learning algorithms proposed in recent years require a large training parameter to achieve decent performance, which requests a high computation cost. Transformer network, as a prevailing deep learning algorithm, has been widely investigated in natural language processing, computational vision and speech recognition [23]. Its powerful parallel computing and global patterns learning capability can identify the fault related patterns in the monitoring data. Identifying these fault related patterns is the key issue in signal denoising and compound fault diagnosis missions, while the technical path for applying the Transformers network towards both targets has yet to be studied. Meanwhile, computational efficiency is a major concern in deep learning modelling. Hence, it is worthwhile to investigate lightweight Transformer networks for the signal denoising and compound fault diagnosis of the industrial robot.

In this paper, an approach that composes of two compact Transformer networks is presented to establish an accurate compound fault diagnosis model for industrial robots operating in noisy environments. In the proposed approach, the feedback current signals collected from the industrial robot are first converted to a time-frequency image via CWT. Among different signal processing approaches, CWT is a prevailing tool to reveal the time-frequency patterns of the signal [24]. The time-frequency image generated by CWT is partially resilient to environmental noise. Subsequently, a powerful and lightweight denoising network called compact Uformer is proposed to denoise the time-frequency images. UNet is a powerful structure based on convolutional neural networks, which is a prevailing tool in image denoising [25]. Compact Uformer takes advantage of both UNet and Transformer networks, and it has higher computation efficiency due to its compact structure. Then, the denoised image is then fed into

CCT to establish a compound fault diagnosis model. CCT is a novel and compact Vision Transformer (ViT) [26] structure that has strong learning capability with fewer trainable parameters [27]. The main contribution of this study is three-fold: (1) Different from UNet and ViT which require high computational cost, a compact Uformer is proposed to denoise the time-frequency images; (2) As a lightweight ViT, this is the first time that CCT is introduced for robotic compound fault diagnosis modelling; (3) An experimental study based on the real-word robotic compound fault dataset is implemented to reveal the effectiveness of the proposed approach. The overall structure of the paper is organised as follows: Section 2 reviews the related works on recent advances in signal denoising and compound fault diagnosis. Section 3 details the methodology of this paper. Section 4 introduces the experimental setup and the experimental results are demonstrated in Section 5. Finally, Section 6 discusses, and Section 7 concludes.

2. Literature Review

2.1. Recent advances of Compound Fault Diagnosis

Recent advances in data analytics and smart manufacturing demonstrate that data-driven approaches are prevailing in fault diagnosis in rotation machinery. As manufacturing equipment is increasingly equipped with sensors and communication capabilities, real-time data collection is becoming easier and easier for fault diagnosis. The use of data-driven methods, especially deep learning approaches, has gained significant attention for compound fault diagnosis in recent years.

Component level fault diagnosis is the main stream in the existing studies [28, 29]. Among different deep learning structures, CNN is widely used in the studies of compound fault diagnosis. Huang et al. [30] proposed a decoupling CNN to identify the relationship between different faults. In this approach, a deep CNN is first used as the model for learning the features, which can effectively learn the discriminative features from raw vibration data. Additionally, multi-stack capsules can be used as a decoupling classifier to locate and isolate compound faults. Finally, the proposed model is trained and optimized using the routing by agreement algorithm and the margin loss cost function. Wang et al. [8] proposed an ensemble extreme learning machine model to diagnose the compound fault of rotating machinery. Specifically, the proposed model consists of two subnetworks, a clustering network and a multi-label classification network. The first network computes the Euclidean distance between each point and every centroid using unsupervised clustering, and the latter network identifies potential output labels using multiple-output-node multi-label learning. Xu et al. [31] developed a method for encoding the fault semantics according to the fault characteristics. The time-frequency features of the compound fault signal are extracted with a CNN before the semantic features of the fault are embedded into the visual space of the fault data. To identify the categories of unknown compound faults, Euclidean distance is adopted to measure the distance between signal features and semantic features of the compound faults. In order to estimate the compound fault severity, Dibaj et al. [32] proposed an end-to-end fault diagnosis method based on a fine-tuned variational node decomposition and CNN. To solve the complex mapping relationship between vibration signals and bearing faults, Lyu et al. [5] proposed a deep learning method that combines residual connection, soft thresholding and global context. The proposed approach integrates the working mechanisms of soft threshold and global context to achieve effective noise reduction and feature extraction. By integrating CNN with wavelet transform and multi-label classification, Liang et al. [9] proposed a new compound fault diagnosis method is proposed for gearbox compound fault diagnosis. Wavelet transform is first adopted to extract the 2D-time-frequency image based on the vibration signal before a multi-label CNN model is built to diagnose compound gearbox faults based on extracted features. Jin et al. [33] presented a novel decoupling attentional residual network for compound fault diagnosis. The original signal is processed by the short-time Fourier transform (STFT), and the output is fed into the attention-enhanced CNN for feature extraction. Then a multi-label decoupling classifier is used to get the compound fault diagnosis results. In order to address the insufficiency of the labelled data, active learning is introduced. Besides CNN, Huang et al. [34] adopted a decoupling capsule network (DCN) is constructed as the basic model. Then, a DCN model can be pretrained using multiple sensor data, which can be used to generate multiple pretrained DCN models. Finally, by combining ensemble learning skills with pretrained DCN models, the deep ensemble CN model is yielded for intelligent compound fault detection and diagnosis. The compound fault diagnosis also can be achieved by other machine learning algorithms. By introducing quantum genetic algorithms (QGA) to improve maximum correlated kurtosis deconvolution (MCKD), Lyu et al. [35] proposed a QGA-MCKD method that can be used to diagnose gear and bearing compound faults. In this approach, QGA adaptively selects the filter length and deconvolution period of MCKD, which correspond to each fault. Based on the key parameters obtained, MCKD processes the compound fault signal, and each individual fault feature related to the individual failed part can be extracted. Li et al. [36] proposed a multiple enhanced sparse decomposition algorithm to simultaneously isolate and extract the harmonic components and transient features from the raw signals.

2.2. The Studies of Signal Denoising

Recently, the integration of deep learning and advanced signal processing techniques has become mainstream in signal denoising. Wang et al. [21] presented an attention-guided joint learning convolutional neural network for the denoising and fault diagnosis of mechanical equipment. The fault diagnosis task and signal denoising task are integrated into an end-to-end CNN architecture, which can achieve decent noise robustness through dual-task joint learning. Zhang et al. [17] proposed an ensemble empirical model decomposition-convolutional deep belief network to denoise the signal collected from reciprocating compressors and extract more robust features for fault diagnosis. Jiang et al. [37] presented a stacked multi-level-denoising autoencoder for denoising and fault diagnosis. In the training stage of autoencoders, multiple levels of denoising schemes were adopted. Zou et al. [22] proposed an adversarial denoising CNN to denoise the signal and get accurate fault diagnosis results. In order to boost the anti-noise performance of training samples, maximum moving is applied to the frequency spectrum of the vibration signal. Zhao et al. [38] developed an interpretable denoising layer for neural networks based on reproducing kernel Hilbert space, which can integrate traditional signal processing technology with physical interpretation into network training. Other prevailing signal processing methods also can be used for signal denoising. Chegini and Najafi [18] reported an approach that uses empirical wavelet transform to isolate the noisy component and utilise a thresholding function to remove the noise. Similarly, Guo et al. [19] adopted a wavelet scattering transform and an improved soft threshold denoising algorithm for noise reduction of the vibration signal. Meng et al. [39] adopted adaptive sparse denoising to determine regularization parameters adaptively, which can denoise the raw vibration signal and reveal fault types.

In the research of compound fault diagnosis, signal processing approaches such as CWT [9] or STFT [33] have been adopted to get the time-frequency image which can reveal the failure pattern of the signal. Recent advances in deep learning have shown decent performance in image denoising [40]. The signal noise in image representation can be removed using the image denoising approach. With the development of CNN, the structure of UNet was proposed. UNet is a powerful deep learning structure that has been widely explored in image segmentation [25], restoration [41] and denoising [42]. In order to address the issue that UNET is not robust to noise in the training process, Thesia et al. [43] modified UNet approach by using a special neural network called Valve, which uses latent features of different UNets as control signals to analyse noise distribution. In recent years, since the Transformer network has become prevailing in computer vision, innovative algorithms that take the advantage of UNet and the Transformer network have emerged. Fan et al. [42] proposed an algorithm called SUNet which combines Swin Transformer and UNet for image denoising. In this structure, three Swin Transformer blocks were adopted, which enables the network the learn the hidden patterns effectively. Yao et al. [44] investigated a dense residual Transformer for image denoising. In this approach, a depth-wise convolutional layer is added to a Transformer block to enhance the local information learning. Then the dense connection is introduced to multiple modified

Transformer blocks to get a comprehensive feature representation.

2.3. A Brief Summary

From the existing studies, it can be seen that the recent advances in compound fault basically focus on the rotation machinery based on the vibration data. Several studies regarding robotic single fault diagnosis were conducted, while there is no research that concerns the compound fault diagnosis for industrial robots. Meanwhile, signal denoising is a key issue in fault diagnosis. The signal processing-based approach heavily relies on expert knowledge, while the deep learning-based approach is complex in modelling. The integration of advanced signal processing and deep learning techniques has shown merits in signal denoising. Transformers network, as a powerful deep learning structure, has shown its merits in denoising and fault diagnosis. However, the existing Transformer networks-based approaches are high in model size, which requires large computational costs. Hence, it is worthwhile to explore an integrated approach that can achieve both high computational efficiency as well as satisfactory denoising and compound fault diagnosis performance.

3. Methodology

In this study, an integrated approach for the robotic compound fault diagnosis is proposed. As illustrated in Figure 1, there are three stages in the proposed approach. Taking a six-axis industrial robot as an example of this study. Firstly, in the data collection and transformation stage, the feedback current signals are collected from the motors of the industrial robot. The sliding window approach is adopted to segment the dataset. Then the collected signals within each time window are transformed into time-frequency images via CWT, which can lower the noise impact and reveal the fault-related features. The time-frequency images generated by CWT contain rich information on both the time domain and frequency domain, which is helpful for compound fault diagnosis. In the second stage, the image mixup [45] is first implemented to increase the number of training samples. It is well known that Transformer models tend to be data-hungry, which requires a large dataset to yield better performance. Mixup is a simple but effective approach that can generate fake samples based on the linear combination of the existing samples. In order to ease the data-hungry issue, mixup strategy is adopted in this study to boost the algorithm performance of both compact Transformer networks. Subsequently, the time-frequency images are used to train a denoising model via compact Uformer. The noisy component in the time-frequency images can be identified and removed by the compact Uformer. The details of Compact Unformer are introduced in Section 3.1. After the time-frequency images are denoised, CCT is adopted to train a compound fault diagnosis model. The diagnosis results are then used to provide decision support for maintenance. The details of CCT are elaborated in Section 3.2.

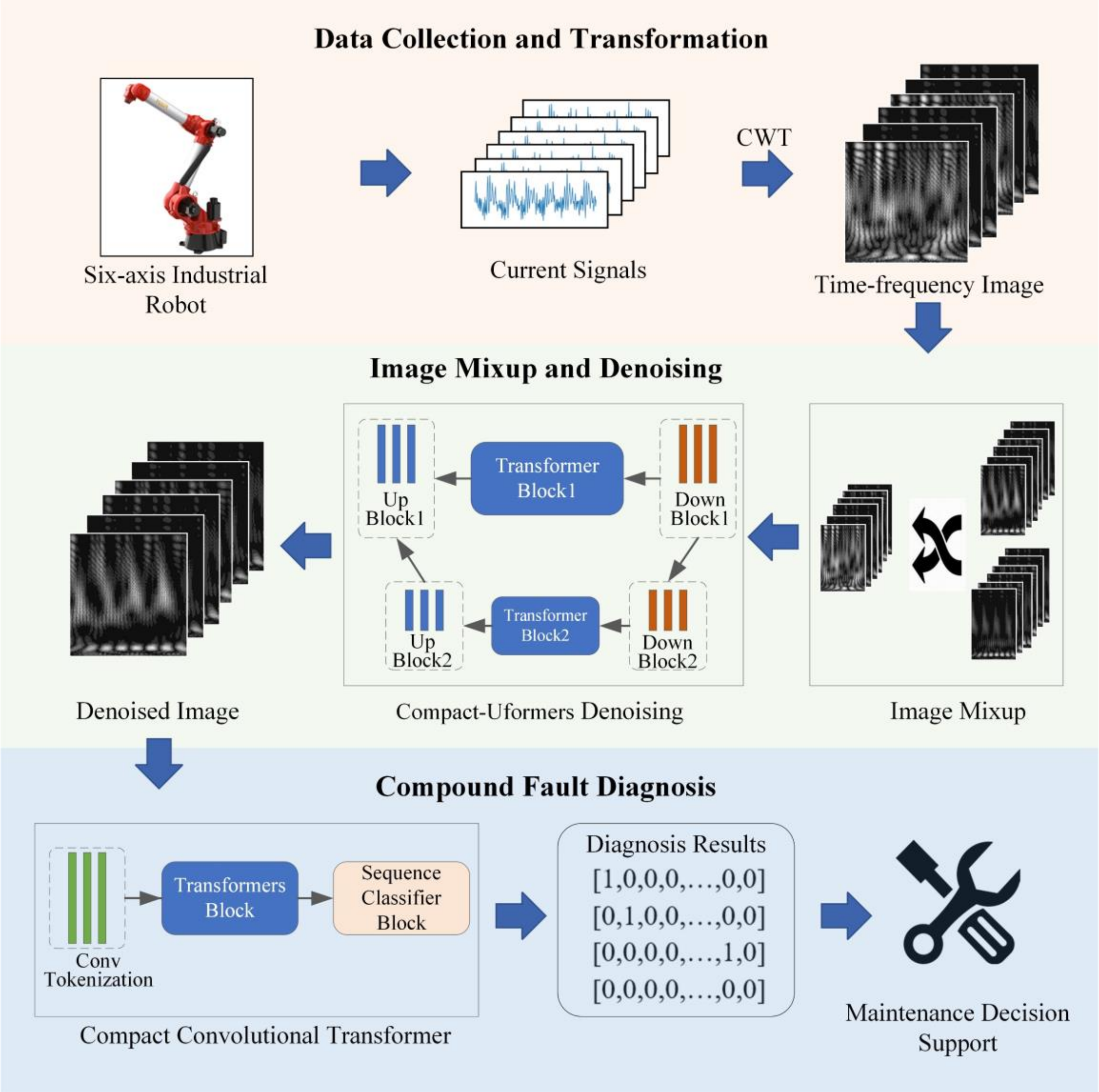


Figure 1. The overall architecture of the proposed approach

3.1.Compact Uformers for Time-frequency Image Denoising

Data denoising is an essential task in signal processing. With the deployment of CWT, the fault-related features can be partially revealed. However, the noise in time-frequency images needs to be further identified and removed. As a prevailing denoising algorithm, the vanilla UNet is composed of four convolutional blocks as an encoder to compress and extract the key features within the images, and another four convolutional blocks as a decoder for image restoration. Moreover, the skip connection between encoders and decoders is used to improve the reconstruction of images. The proposed network architecture is illustrated in the lower part of Figure 2.

The proposed network consists of a feature contracting part, a feature expansive part, and two Transformer blocks. In order to achieve high cost-effectiveness, only two levels of transformers are designed in the proposed compact Uformer. Similar to the Vanilla UNet, the feature contracting part contains two convolutional blocks, which are composed of two 3x3 convolutions, a rectified linear unit (ReLU), and a 2x2 max pooling operation. In contrast, each convolutional block in the expansive part consists of an upsampling of the feature map followed by a 2x2 up-convolution. After that, a concatenation with the correspondingly cropped feature map from the contracting path is applied, followed by two 3x3 convolutions, before the outputs are sent to a ReLU. Different from the Vanilla UNet that establishes a skip connection between the same level convolutional down block and up block, we introduce Transformer block to replace the skip connection so to enhance the global feature extraction capability of the compact Uformer. It can be seen that the feature map size yielded in the second convolutional down block is a quarter of that in the second convolutional down block. Because of that, both Transformer blocks can learn the global patterns in different scales, which can provide more useful information for noise removal. The image datasets such as Urban100 and SIDD datasets that are applied in the existing image restoration or denoising task are the picture collected from the real world, which are rich in detail. In contrast, the time-frequency image generated by CWT only contains limited information. Hence, only two levels of feature down and up blocks are adopted in the proposed network to keep the model compact and efficient in computation.

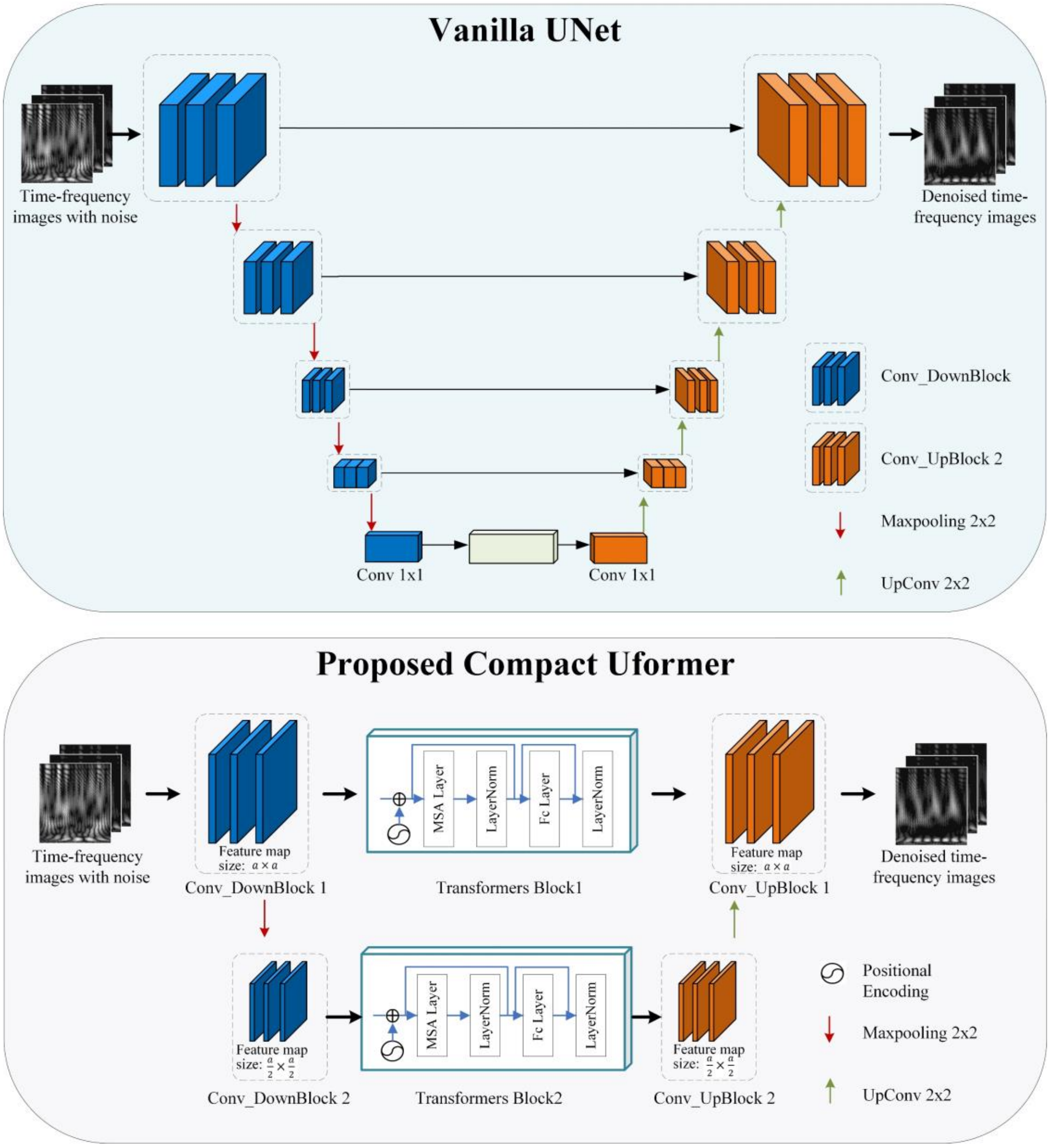

Vanilla UNet
Time-frequency images with noise
Denoised time-frequency images
Conv 1x1
Conv 1x1
Conv_DownBlock
Conv_UpBlock 2
Maxpooling 2x2
UpConv 2x2
Proposed Compact Uformer
Time-frequency images with noise
Feature map size: $a \times a$
Conv_DownBlock 1
MSA Layer
LayerNorm
Fc Layer
LayerNorm
Transformers Block1
Feature map size: $a \times a$
Conv_UpBlock 1
Denoised time-frequency images
Feature map size: $\frac{a}{2} \times \frac{a}{2}$
Conv_DownBlock 2
Transformers Block2
Feature map size: $\frac{a}{2} \times \frac{a}{2}$
Conv_UpBlock 2
Positional Encoding
Maxpooling 2x2
UpConv 2x2

Figure 2. The comparison between vanilla UNet and the proposed compact Uformer

The Transformer network is a powerful deep learning architecture that is expertise in learning the global patterns within the data. The key component of the Transformer network is the multi-head self-attention layer. To learn the important features from different perspectives, the extracted information is first sent to the multi-head self-attention (MSA) layer, which adopts multiple self-attention layers. The output of the attention is obtained by determining the relationship between a query and a set of key-value pairs to output. In the attention output, the values are weighted and combined to indicate the location of the important features. The weights are obtained via the computation of a compatibility function of the query with the corresponding key. By computing the query vector q, key vector k and value vector v, the attention score of the standard self-attention layer can be obtained. When an input i is sent into a self-attention attention layer, the attention score is calculated as:

$$Score = softmax(q_i * k_i) * v_i \tag{1}$$

In the MSA, three matrices which are Q, K and V are adopted to replace the vectors q, k and v. By combining rich information from different perspectives of the matrix, the importance of various features can be determined comprehensively. Denoting an input data as:

$$X = [x_1, x_2, \dots, x_n] \tag{2}$$

the matrices Q, K and V are obtained via a linear transformation of the input data, which can be expressed as:

$$Q = XW^q \tag{3}$$

$$K = XW^k \tag{4}$$

$$V = XW^v \tag{5}$$

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

Then the obtained matrices are further fed into scaled dot-product attention to get the attention score, which can be expressed as:

$$Head\_Score_1 = \frac{softmax（\mathrm{Q} \times K^T）\times \mathrm{V}}{\sqrt{d}} \tag{6}$$

, where d is a scalable factor.

Besides using the features captured from the convolutional blocks, positional encoding is also adopted to provide sequential information to the Transformer blocks. The global features captured by Transformer block 2 are then sent into the convolutional up block 2 for image restoration. Then the up-convolutional feature maps and the output of Transformer block 1 are jointly fed into convolutional up block 1 to get the denoised image.

3.2. Compact Convolutional Transformer for Compound Fault Diagnosis

Compound fault diagnosis is a challenging task for industrial robots. In this section, a compound fault approach that adopts CCT as the backbone is elaborated. The details of CCT compound fault diagnosis modelling are illustrated in Figure 3. The denoised images are fed into CCT. The standard CCT adopts convolution blocks for feature extraction. The extracted features are then sent into Transformer blocks for global patterns mining. Finally, the output of the Transformer block is then fed into a sequence pooling layer to further locate the essential information. Meanwhile, the high computation cost of the Transformer block is caused by the large size of the feature map. The convolutional block is effective in the reduction of feature map size, which can lower the computation cost of the Transformer block. Because of the combination of the convolutional block and the Transformer block, both local and global patterns can be extracted with lower model sizes in comparison with the prevailing deep learning algorithms.

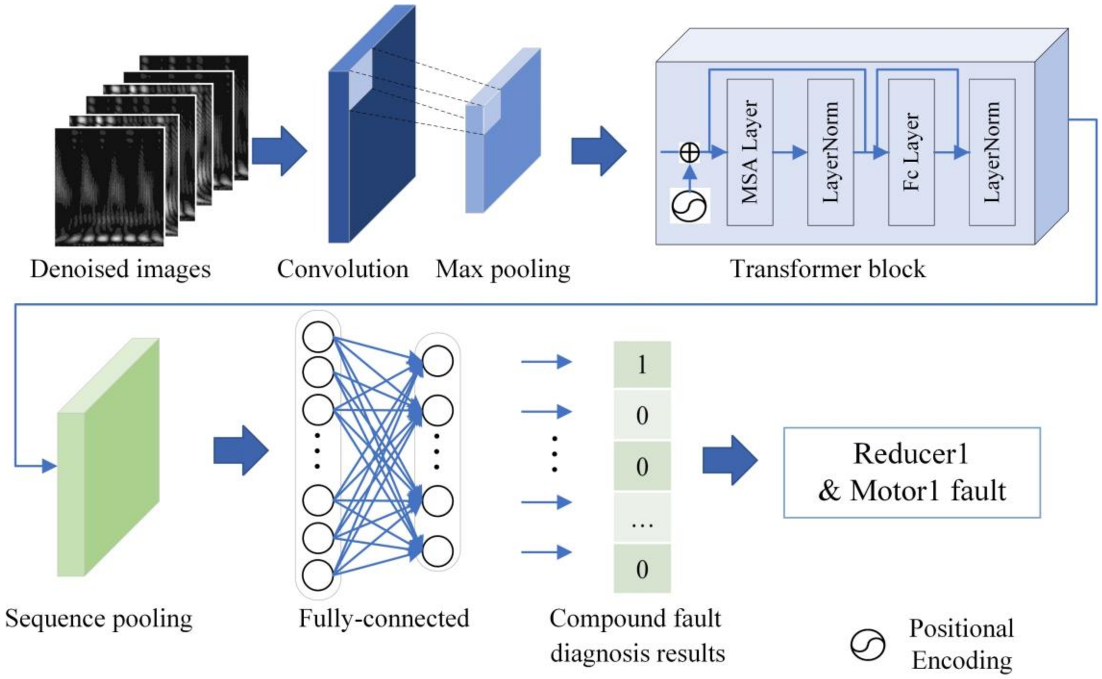


Figure 3. The flow chart of Compact convolutional Transformer for compound fault diagnosis

In CCT, the convolution block is composed of convolutional layers with small strides allowing efficient tokenization and the extraction of local spatial relationships. Meanwhile, because of the downsampling of the feature maps, the computation cost of the subsequent Transformer block is reduced due to the size of the feature maps being decreased. Different from ViT, CCT replaces the "image patching" and "encoding" layers in ViT with simple convolutional blocks to introduce the inductive bias. A convolution block consists of a convolution layer, a ReLU activation and a max-pooling layer. The process can be expressed as:

$$F = MaxPooling(ReLU(Conv2d(x)) \tag{7}$$

where x and F are the input and output of the convolution block.

Then the features extracted by the convolutional block and the positional encoding are jointly fed into a Transformer block for global pattern learning. The details of the Transformer block are demonstrated in Section 3.1. Subsequently, the global features yielded by the Transformer block are then processed by the sequence pooling layer. The sequence pooling can process the sequential-based information yielded by the Transformer block, which eliminates the requirement for the extra class token. Due to the fact that the sequence of data contains relevant information across various parts of the input image, the entire sequence of data is pooled to get the essential features. The output of a Transformer block is expressed as:

$$x_T = f(x_0) \in \mathrm{R}^{b \times n \times d} \tag{8}$$

where the dimension of x_T is $b \times n \times d$, b is the mini-batch size, n is the sequence length and d is the embedding dimension. Subsequently, a linear layer with SoftMax activation is used to further process x_T, which can be expressed as:

$$x_P = x_0 \times x_T = SoftMax(Linear(x_T)) \times x_T \in \mathrm{R}^{b \times 1 \times d} \tag{9}$$

The second dimension of x_T is the sequence length. By pooling the data in this dimension, the importance of the sequence can be revealed and the dimension of x_P becomes $b \times 1 \times d$. By utilizing the sequence pooling strategy and convolution, CCT can eliminate the requirement for class tokens.

After the output of sequence pooling is obtained. It is further used to get the compound fault diagnosis result. By directly sending x_P into two linear layers, the classification result can be obtained.

4. Experimental Setup

4.1.Data Collection

It is well known that Industrial robot is high in reliability. The overload or accident issues in manufacturing can cause compound faults in the industrial robot. Collecting the compound fault data in a run-to-failure manner is time-consuming. In order to address this issue, a fault injection experiment was implemented for the data collection. The faulty motors and reducers were collected from actual industrial robots from manufacturing works. The faulty parts used in this experiment can continue to operate despite minor abnormal sounds or oil leakage. This type of fault is in its infancy stage which is hard to diagnose. In the experiment, the faulty parts were used to replace the normal parts in a healthy six-axes robot. The robot with a faulty part was then used to perform different tasks. The structure of the six-axes robot is demonstrated in Figure 4. The three-phase currents in a motor are monitored by the Hall element in the driver. The collected signals are the currents of q-axis, which frequency is 1Hz.

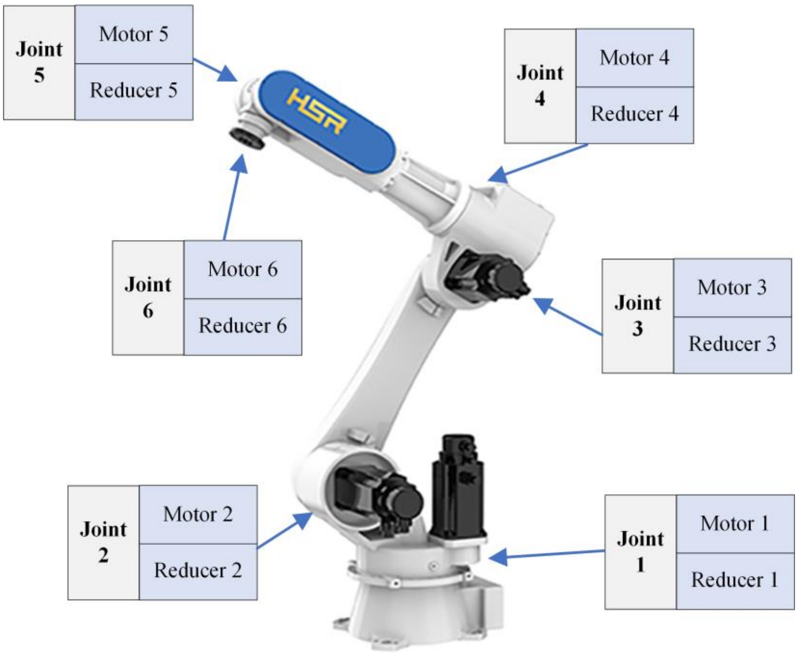


Figure 4. The structure of the six-axis industrial robot adopted in the experiment

In the fault injection experiment, there are three types of single faults and three types of compound faults were performed, and the relevant data was collected. The details of the collected data are shown in Table 1. In the era of Industry 4.0, collecting the monitoring data of all the axis requires a large transmission cost, which may not be affordable for a company that possess a large number of industrial robots. Meanwhile, industrial robot is a type of closed coupled asset, which the faulty parts may affect the operation of the coupled components. Hence, faulty patterns also can be found in the monitoring signals of other components. Based on this consideration, only the feedback current of motor 3 was used in the experiment for modelling to realise compound fault diagnosis with lower resources. Each category contains 300,000 data entries. In the data pre-processing stage, the data was segmented by the length of 100. Then the data in each window is transformed to time-frequency images via CWT. The size of the time-frequency image is 32×32.

Table 1. The details of the collected dataset.

No	Fault type	Faulty parts
1		Normal
2	Compound fault	Reducer 1 and Motor 2
3		Reducer 1 and Reducer 3
4		Reducer 3 and Reducer 4
5	Single fault	Reducer 3
6		Motor 2
7		Reducer 4

After CWT, each data category contains 3,000 time-frequency images, for a total of 21,000 data samples. Examples of time-frequency images for each fault category are shown in Figure 5. Since deep learning is a type of data-hungry algorithm, the number of data samples is essential to the performance of deep learning. In order to ease the effect of data insufficiency, an image mixup [45] was adopted to augment the training dataset. In this study, the mixup ratio of two images was set as 0.5, which indicates that 50% of the two images are adopted to generate a pseudo image. The number of newly generated images for each category is 3,000. After the mixup, the training data is increased by 21,000.

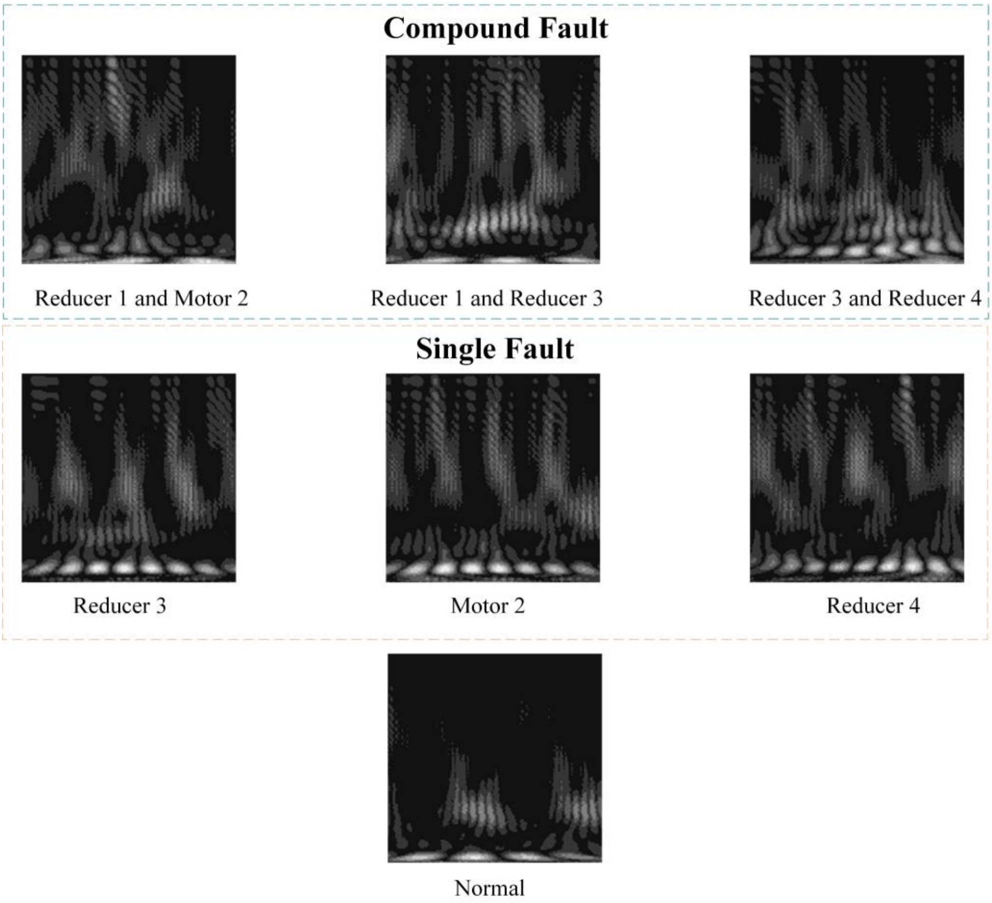


Figure 5. The time-frequency images generated by motor 3’s feedback current signals of different fault categories

4.2. Benchmarking Experiments

In this study, the denoising performance of compact Uformer is firstly explored. Different noises from the environment, assets, and rotating components can be regarded as following a Gaussian distribution. In order to get the noisy data for benchmarking experiments, different levels of Gaussian white noise were separately added to the raw feedback current signals to generate a time-frequency image. Figure 6 illustrated three examples of adding noises and time-frequency image generation. It can be seen that when the noise level is 30dB or lower, the noise in the time-frequency image is weak. When the noise level is up to 2dB, the key patterns in the time-frequency image are partially overwhelmed. When the noise level becomes -10dB, it is challenging to identify useful patterns. In this study, there are 10 different levels of noise, which range from -10dB to 30dB, were added to the feedback current signals to evaluate the denoising performance of the proposed compact Uformer.

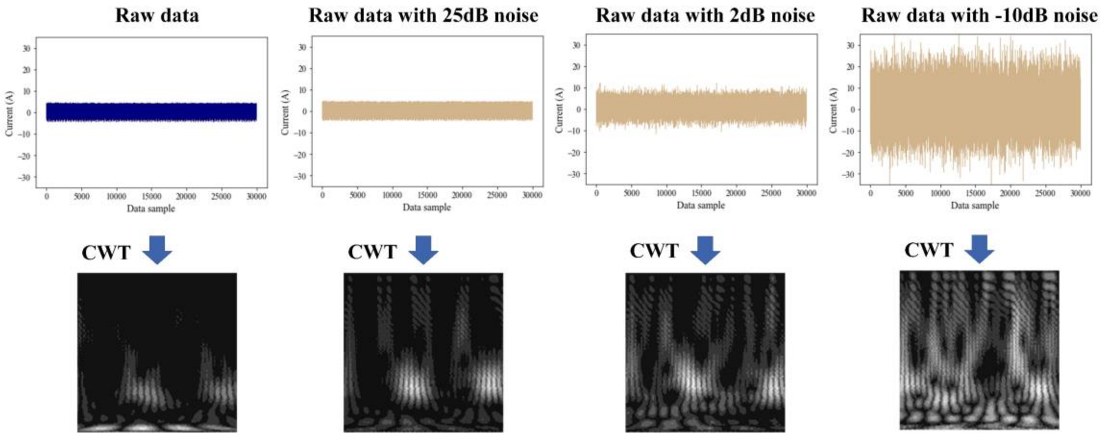


Figure 6. The time-frequency images of feedback current signal under different levels of Gaussian white noise

The parameters setting is important to the model training. In this study, Adam is adopted as the optimiser with the learning rate set as 0.001. Mean square loss is used as the loss function and the training loss is set as 300 epochs. The number of heads in both Transformer block's MSA layers was set as 6. The number of nodes in the linear layer of Transformer block 1 was set as 256 and the number of nodes in the linear layer of Transformer block 2 was set as 64. In order to reveal the effeteness of the compact Uformer, three benchmarking algorithms were adopted. The key parameters setting such as learning rate and training epoch is the same as the compact Uformer. The details of these benchmarking algorithms are demonstrated as follows:

1. **UNet** [25]: proposed in 2015, it is the classical algorithm based on CNN, which contains

four convolutional down sampling blocks and four convolutional up sampling blocks.

2. **SwinIR** [46]: proposed in 2021, it is a novel image restoration algorithm that uses CNN for shallow feature extraction and image reconstruction and Swin Transformer blocks for deep features extraction.
3. **MPRNet** [47]: proposed in 2021, it is a multi-stage architecture that progressively learns the hidden patterns for the degraded inputs, which are then used to restore the image. In this algorithm, three UNets are adopted as the backbone of each stage.

In the image-denoising experiment, the image denoising of all the data samples was performed. The data from seven categories were mixed together for model training. The images denoised by Compact Uformer were then fed into CCT for compound fault modelling. CCT adopted convolutional blocks with 3×3 convolutions for feature extraction. Then the extracted features with the shape of 64×256 were sent into stacked Transformer blocks. The number of heads in MSA layers was set as 6 and the number of nodes of the linear layer in Transformer blocks was set as 64. Adam and mean square error were adopted as the optimiser and loss function. The learning rate and training epoch were set as 0.001 and 300, respectively. The key parameters of CCT are the number of convolutional layers and Transformer layers, which is investigated in the experiment.

In order to reveal the merits of CCT, several prevailing deep learning algorithms were adopted. The benchmarking algorithms include two large and powerful networks which are Swin Transformer and ResNet50. Meanwhile, two lightweight networks which are DeIT and

MobileNet V2, were adopted in this experiment. The details of the benchmarking algorithms are presented below:

1. **Swin Transformer** [48]: is a modified version of VIT. It adopts shifted windows self-attention and patch merging to achieve lower computation complexity and higher image classification accuracy.
2. **DeIT** [49]: is a data-efficient image Transformer network. The distillation strategy is deployed in this model to reduce the network size and keep satisfactory performance. In this experiment, the tiny version DeIT, which the number of parameters is 5M, was adopted.
3. **MobileNet V2** [50]: is a lightweight version of ResNet. In this algorithm, the depth separable convolution and inverted residual connections are used to lower the computational cost and improve the effeteness of feature extraction.

The 5-fold cross-validation was implemented to get comprehensive results. All tests were conducted on an Intel i9-10920X 3.50Ghz CPU with Nvidia GeForce RTX 3090 graphics card. Besides, the t-SNE technique was adopted in this experiment to further study the insights of the modelling results.

4.3.Evaluation Metrics

In this study, two prevailing metrics, which are peak signal-to-noise ratio (PSNR) and structure similarity index measure (SSIM), were used to evaluate the image denoising performance. The PSNR can be calculated as follows:

$$PSNR = 10\log_{10}\left(\frac{(2^n-1)^2}{\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}(x(i,j)-y(i,j))^2}\right) \tag{10}$$

, where $H \times W$ is the size of the image. $x(i,j)$ and $y(i,j)$ are the pixels in both images. n

is the bit number.

The expression of SSIM is shown as follows:

$$\mathrm{SSIM}(x,y)=\frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)} \tag{11}$$

, where μ_x and μ_y are the mean value of x and y. σ_x and σ_y are the variance of x and y. C_1 and C_2 are two constants.

For the evaluation of compound fault diagnosis, classification accuracy was adopted since the compound fault diagnosis modelling is a multi-classes classification task.

5. Experimental Results

5.1. Experiment on Time-Frequency Image Denoising

The denoising performance of the proposed compact Uformer is revealed in this experiment. The model parameters and the training time of all four algorithms are listed in Table 2. It can be seen that the number of parameters of the proposed compact Uformer is far less than the benchmarking algorithms. In the experiment, the denoising experiment that used the data of all seven categories is performed.

Table 2. The comparison of model computation cost.

	Compact Uformer	UNet	SwinIR	MPRNet
#Params (M)	**0.81**	4.1	3.2	5.6

It can be seen from Figures 7 and 8 that compact Uformer shows better PSNR in all the experiments under different noise levels, while compact Uformer achieves the best SSIM when the noise level is lower than 10dB. All the algorithms show better PSNR and SSIM when the noise level increases from -10dB to 30dB. In the extremely heavy noise scenario, which the noise level is lower than 5dB, the merit of compact Uformer in terms of PSNR and SSIM is obvious. However, when the noise level is above 5dB, SwinIR and MPRNet also can get decent denoising performance. The advantage of the proposed compact Uformer is mere. When the noise level is 30dB, the PSNR of the proposed compact Uformer is slightly better than that of MPRNet. When the noise level arrives at 30dB, SwinIR and MPRNet outperform the compact Uformer in SSIM.

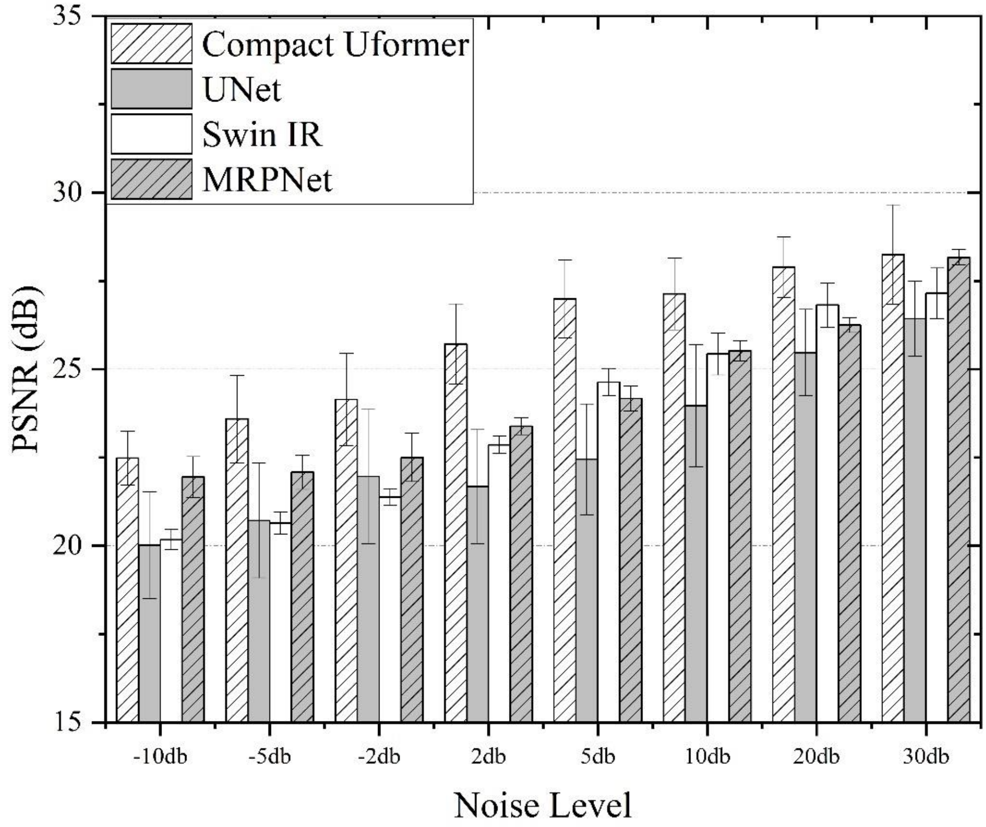


Figure 7. The denoising performance in terms of PSNR of algorithms based on data from all categories in different noise levels

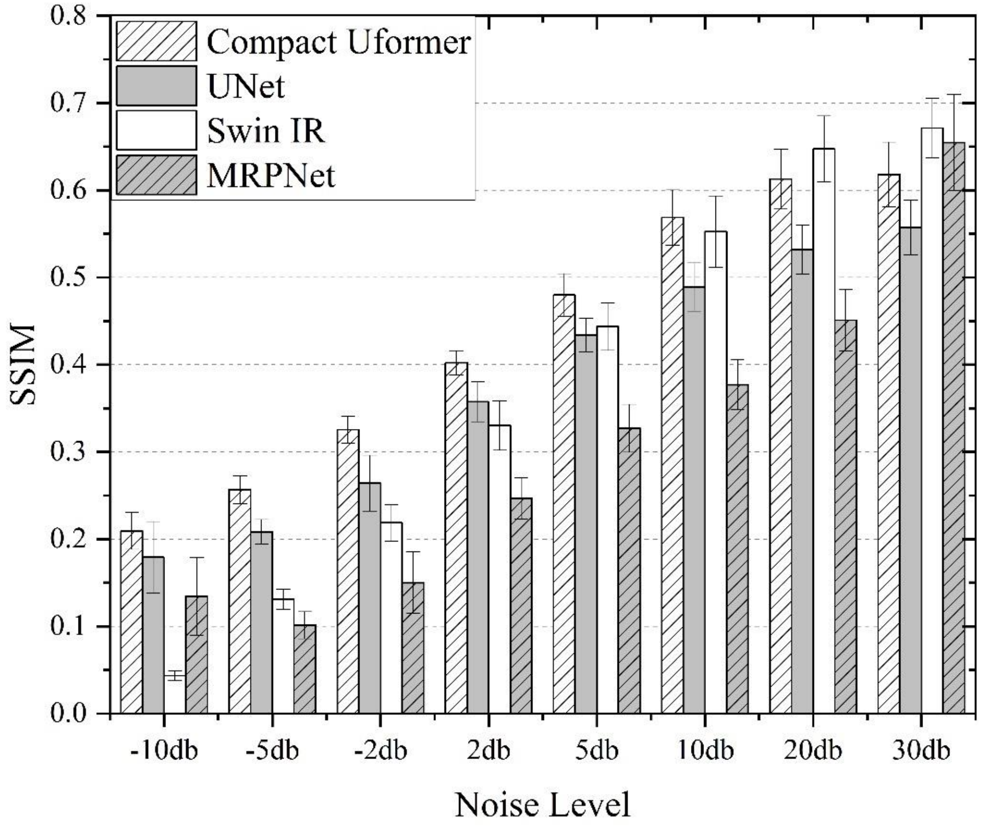


Figure 8. The denoising performance in terms of SSIM of algorithms based on data from all categories in different noise levels

After the denoising experiment using all the data from each category, it is worthwhile to study the denoising performance of algorithms of each category separately. It can be seen from Tables 3 and 4 that the denoising performance of all the algorithms in terms of PSNR and SSIM decreases with the enhancement of noise. When the noise level becomes -10dB, the PSNR obtained by all the algorithms is lower than 20dB. Among all the seven categories, the denoising of the Normal category achieved the best performance in most scenarios, while three single fault categories show moderate performances. The denoising performance in the compound fault categories is worse than that of the normal and the single fault categories. The results of denoising of Reducer1and Motor2 are the worst in all experiments. When the noise level is 30dB, the best denoising of Reducer1and Motor2 in terms of PSNR and SSIM are only 19.65dB and 0.3477. Similar to the results derived from the data of all the categories, when the added noise is at a low level such as 20dB or 30dB, the results of Compact Uformer in terms of PSNR and SSIM are not advantageous when compared to SwinIR and NPRnet. In contrast, when the noise level is -5dB or -10dB, the algorithm performance of SwinIR and NPRnet deteriorates rapidly in terms of PSNR and SSIM.

Table 3. The denoising performance in terms of PSNR of algorithms based on data from different categories.

Noise (dB)	Algorithms	Normal	Reducer1 and Motor2	Reducer1 and Reducer3	Reducer3 and Reducer4	Reducer3	Motor2	Reducer4
-10	Compact Uformer	**19.47**	**18.56**	**19.63**	**19.69**	**19.27**	**19.26**	**19.31**
	UNet	18.54	17.32	17.39	17.94	17.42	17.45	17.48
	SwinIR	17.63	17.16	17.39	17.21	17.40	17.41	17.37
	NPRnet	18.82	17.91	19.51	18.30	17.89	18.32	18.70
-5	Compact Uformer	**20.47**	**19.07**	**19.82**	**20.34**	**19.67**	**19.59**	**19.65**
	UNet	19.87	18.48	19.48	19.68	18.52	18.84	18.95
	SwinIR	18.41	17.56	17.41	18.26	17.65	17.70	17.78
	NPRnet	19.87	18.48	19.48	19.68	18.52	18.84	18.95
-2	Compact Uformer	**21.56**	**19.21**	**20.10**	**21.79**	**19.96**	**20.24**	**20.43**
	UNet	19.45	17.63	17.46	18.63	17.98	18.21	18.36
	SwinIR	19.40	17.48	17.44	18.08	17.84	18.03	18.28
	NPRnet	20.71	18.54	18.95	17.97	18.79	19.44	19.75
2	Compact Uformer	22.93	**19.70**	**20.71**	**21.25**	**20.84**	**21.41**	**21.94**
	UNet	21.25	18.68	19.68	20.32	20.14	20.99	21.55
	SwinIR	**23.06**	17.44	17.57	19.21	18.70	19.06	19.43
	NPRnet	21.02	17.78	18.86	18.74	18.33	19.77	20.74
5	Compact Uformer	24.12	**19.44**	**20.34**	**23.96**	**21.77**	**22.33**	**22.93**
	UNet	22.11	18.22	18.68	22.52	19.87	20.69	20.10
	SwinIR	22.06	17.75	17.49	19.59	19.64	19.96	20.28
	NPRnet	**24.32**	19.18	19.66	18.53	21.02	21.88	22.64
10	Compact Uformer	**25.79**	**19.65**	**20.06**	**24.52**	**22.83**	**23.50**	24.46
	UNet	23.41	18.32	18.97	22.46	20.69	22.24	23.32
	SwinIR	23.14	17.88	17.45	19.84	20.79	21.07	21.19
	NPRnet	25.70	19.29	19.96	23.64	22.46	23.41	**24.55**
20	Compact Uformer	26.95	19.53	19.96	**22.58**	**23.83**	24.64	25.99
	UNet	23.68	18.58	19.12	22.48	21.63	21.78	21.21
	SwinIR	24.30	17.87	17.68	20.25	22.10	21.85	22.44
	NPRnet	**27.03**	**19.55**	**20.13**	23.26	23.82	**25.03**	**26.53**
30	Compact Uformer	26.97	**19.65**	19.63	**22.93**	24.03	**24.98**	26.81
	UNet	23.14	18.55	19.01	21.75	22.30	21.41	21.97
	SwinIR	24.19	18.25	17.79	20.37	22.35	22.71	23.51
	NPRnet	**27.51**	19.52	**20.02**	22.89	**24.53**	25.70	**27.34**

Table 4. The denoising performance in terms of SSIM of algorithms based on data from different categories.

Noise (dB)	Algorithms	Normal	Reducer1 and Motor2	Reducer1 and Reducer3	Reducer3 and Reducer4	Reducer3	Motor2	Reducer4
-10	Compact Uformer	**0.4173**	**0.2238**	**0.4798**	**0.4047**	**0.4150**	**0.4080**	**0.4291**
	UNet	0.1641	0.1330	0.1378	0.2524	0.1941	0.2045	0.2553
	SwinIR	0.0540	0.0327	0.0335	0.4060	0.1356	0.1482	0.1432
		0.2593	0.0479	0.4435	0.3918	0.0564	0.1518	0.2712
	NPRnet	0.3766	0.0877	0.3162	0.3276	0.2654	0.3943	0.4357
-5	Compact Uformer	**0.4940**	**0.3156**	**0.5130**	**0.4341**	**0.4940**	**0.4681**	**0.4742**
	UNet	0.1761	0.1871	0.2422	0.3306	0.1079	0.1901	0.2398
	SwinIR	0.1630	0.1313	0.2345	0.3117	0.1893	0.2111	0.2269
	NPRnet	0.3530	0.1568	0.4399	0.2612	0.2065	0.2880	0.3063
-2	Compact Uformer	**0.5214**	**0.3340**	**0.5537**	**0.4754**	**0.5148**	**0.5294**	**0.5477**
	UNet	0.2708	0.2384	0.2543	0.2866	0.1911	0.2218	0.2460
	SwinIR	0.2589	02317	0.2398	0.2108	0.1805	0.2046	0.2285
	NPRnet	0.3766	0.0877	0.3162	0.3276	0.2654	0.3943	0.4357
2	Compact Uformer	**0.5830**	**0.4350**	**0.6083**	**0.4842**	**0.5930**	0.6276	**0.6567**
	UNet	0.5796	0.1514	0.4178	0.2065	0.4716	0.5670	0.6087
	SwinIR	0.3890	0.0363	0.0443	0.3411	0.3129	0.3272	0.3747
	NPRnet	0.5835	0.3487	0.4969	0.3995	0.3253	**0.6307**	0.6317
5	Compact Uformer	**0.6685**	**0.3483**	**0.5656**	**0.5278**	**0.6839**	**0.7018**	**0.7227**
	UNet	0.6492	0.2247	0.4440	0.2773	0.6005	0.6536	0.5940
	SwinIR	0.4796	0.3436	0.0567	0.4374	0.4222	0.4330	0.4670
	NPRnet	0.5797	0.3388	0.0000	0.4630	0.5283	0.6166	0.7044
10	Compact Uformer	**0.7330**	**0.3567**	**0.5274**	0.5735	**0.7597**	**0.7746**	**0.8004**
	UNet	0.7228	0.2922	0.5006	0.4407	0.7356	0.7663	0.8000
	SwinIR	0.5566	0.2618	0.3838	0.5016	0.5675	0.5616	0.5616
	NPRnet	0.6971	0.2754	0.4259	**0.6252**	0.5961	0.6929	0.7303
20	Compact Uformer	**0.7330**	**0.3567**	**0.5274**	0.5735	**0.7597**	**0.7746**	**0.8004**
	UNet	0.7228	0.2922	0.5006	0.4407	0.7356	0.7663	0.8000
	SwinIR	0.5566	0.2618	0.3838	0.5016	0.5675	0.5616	0.5616
	NPRnet	0.6971	0.2754	0.4259	**0.6252**	0.5961	0.6929	0.7303
30	Compact Uformer	0.7899	0.3460	0.4625	**0.6200**	**0.8265**	0.8483	**0.8818**
	UNet	0.7243	0.3372	0.4110	0.5473	0.8414	0.8665	0.6991
	SwinIR	0.6290	02315	0.1421	0.5140	0.6947	0.7252	0.7620
	NPRnet	**0.8193**	**0.3477**	**0.4900**	0.6056	0.7599	**0.8476**	0.8680

The denoised examples when the noise level is -2dB are demonstrated in Figure 9. These examples were randomly selected from the testing set. It can be seen that the information in the CWT image of the normal category is less than the single fault and the compound fault categories. Hence, the denoising performance of the normal category is obviously better than all fault categories. The contour of the key features in the denoised images of the single faults categories is still obvious. For the CWT images of the compound fault, it can be seen that the patterns are more complex, which poses a challenge of denoising. The details of the features were not fully restored in the denoising process.

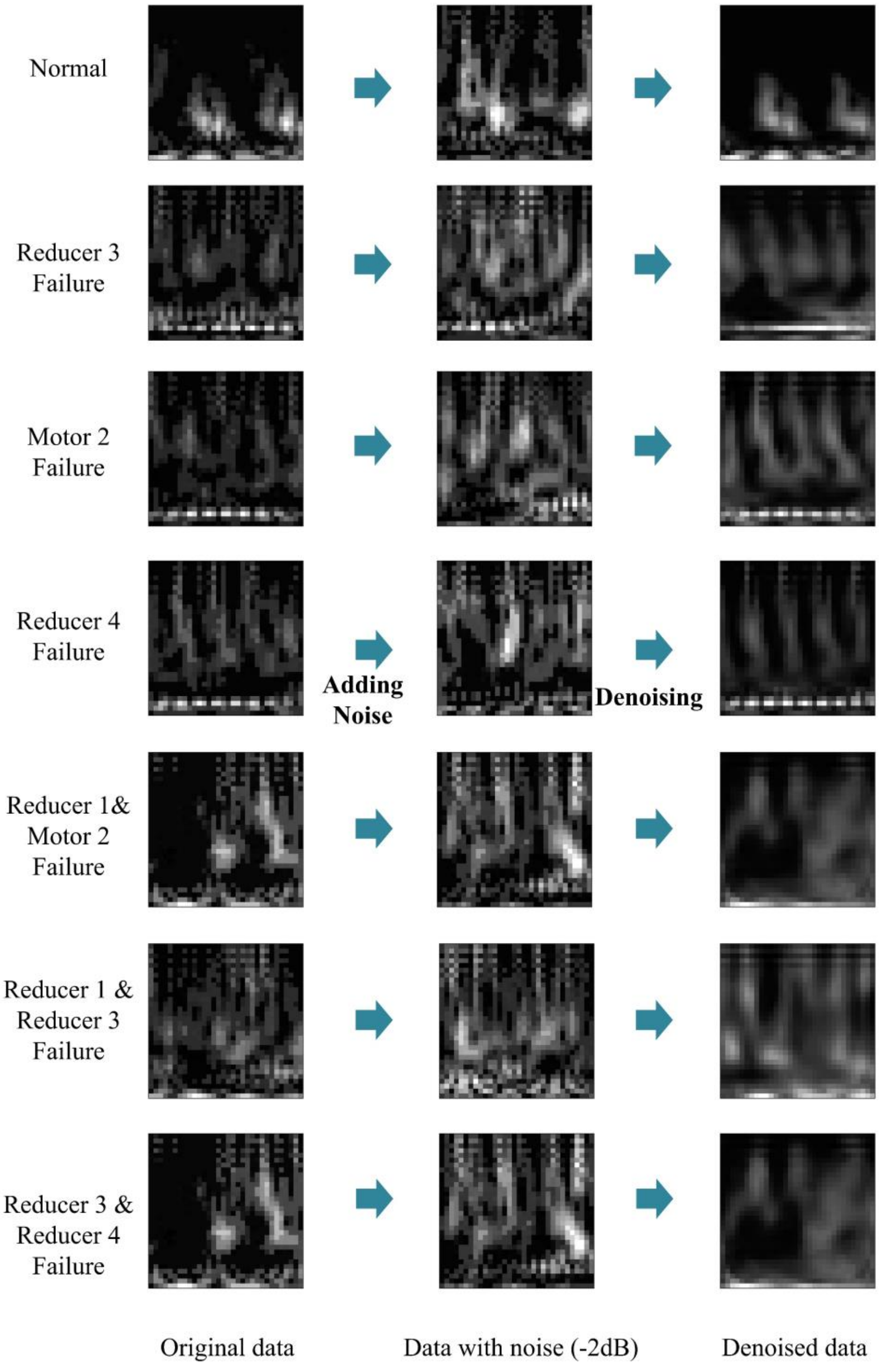


Figure 9. The denoised results of compact Uformer when the noise level is -2dB

5.2. Experiment on Compound Fault Diagnosis for Industrial Robot

In this experiment. The impact of the key parameters of CCT on the compound fault diagnosis modelling performance is firstly revealed. The number of Transformer layers and the convolutional layers are the key parameters in CCT, which can affect the model size and the performance. Table 6 demonstrated the compound fault diagnosis results in different parameter settings. The overuse of convolutional layers leads to extremely small feature maps, which is challenging for the Transformer layers to further learn the hidden patterns. Hence, the maximum of two convolutional layers was adopted. It can be seen from the results that when the Transformer layers were set at six and two convolutional layers were adopted, CCT can achieve the best accuracy which is 86.29%. When the number of Transformer layers increased to 8, the classification accuracy was compromised while the model size increase rapidly. Hence, this setting of six Transformer layers and two convolutional layers was adopted in the following experiments.

Table 6. The compound fault diagnosis results of CCT under different parameters settings

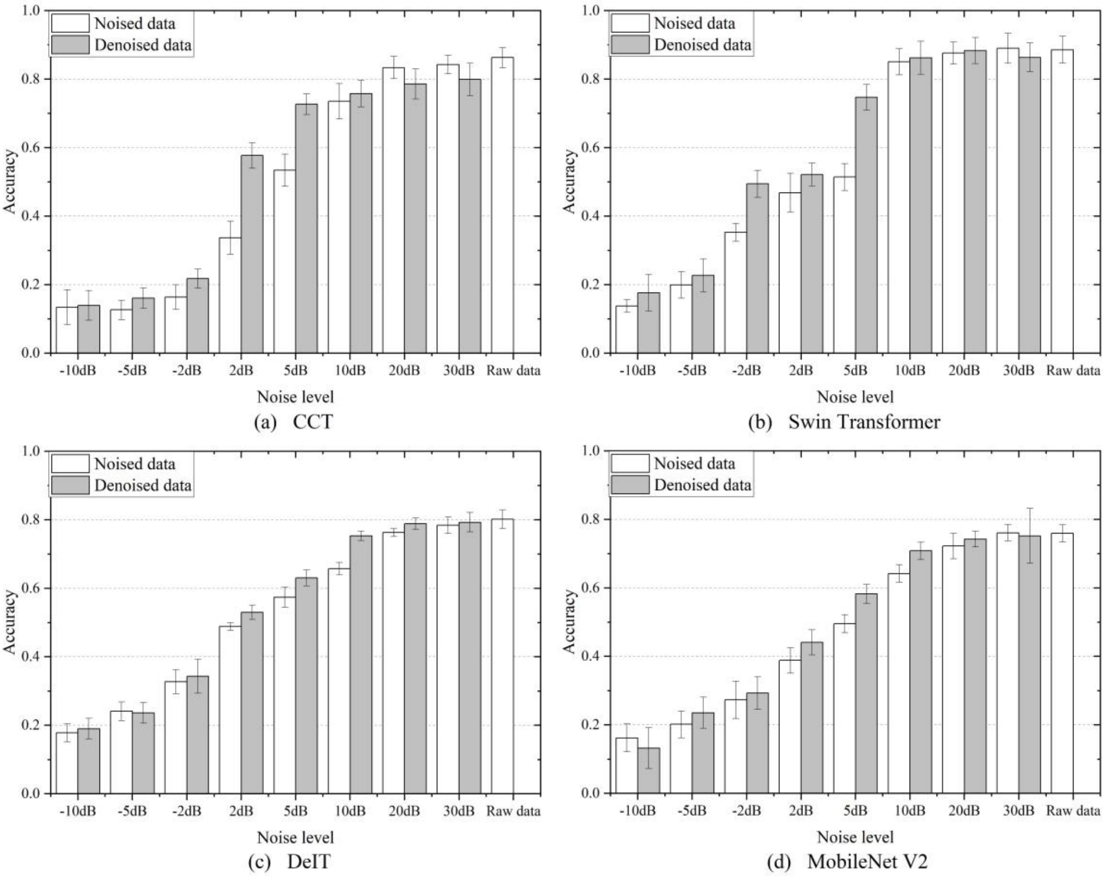
#Transformer layers	# Convolutional layers	Accuracy (%)	#Pramas (M)
2	1	76.15±2.97	0.20
2	2	81.13±2.35	0.28
4	1	83.73±2.87	0.40
4	2	84.12±2.65	0.48
6	1	85.72±3.26	3.1
6	2	**86.29±2.95**	3.3
8	1	86.18±2.83	4.2
8	2	86.07±2.77	4.4

For the benchmarking experiment, three benchmarking algorithms which are Swin Transformer, DeIT and MobileNet V2, were adopted to reveal the algorithm performance of CCT. The CWT images without noise were for modelling and the results can be seen in table 6. It can be seen that Swin Transformer get the best classification accuracy which is 88.61%. The classification accuracy of CCT is 86.29%, which is slightly lower than that of Swin Transformer. The compound fault classification accuracy of DeIT and MobileNet V2 is 80.15% and 75.94% respectively. Though Swin Transformer achieved the highest accuracy, its number of parameters is the largest, which is up to 139.71M. In striking contrast, the number of parameters is far less than all the benchmarking algorithms, which is 3.3M.

Table 6. The comparison between the model size and the Compound fault diagnosis accuracy

	CCT	Swin Transformer	DeIT	MobileNet V2
#Params (M)	**3.3**	139.71	20.93	8.52
Accuracy (%)	86.29±2.95	**88.61±1.27**	80.15±2.74	75.94±2.52

Subsequently, the denoised images generated by compact Uformer based on data from all categories were adopted for modelling. From Figure 10 (a), it is obvious that when the added noise is lower than 2dB, the accuracy improvement with the help of image denoising is not obvious, which is approximately 5%. The classification accuracy based on denoised data is higher than that based on noisy data when the noise level is 2dB to 10dB. When the noise level exceeds 10dB, the classification accuracy based on the denoised is worse than that based on the noisy data. What is evident in Figure 10 (b) is that the classification result based on the denoised data is advantageous when the noise level locates at -5dB or 2dB. The classification accuracy of Swin Transformer is higher than that of CCT when the noise level is above 10dB. Figures 10 (c) and (d) show a similar trend to CCT and Swin Transformer, while the accuracy is lower than CCT and Swin Transformer in all the stages. It also can be seen from both figures that the difference between the accuracy obtained based on the denoised data and noisy data is not significant when the noise level surpasses 10dB.



(a) CCT

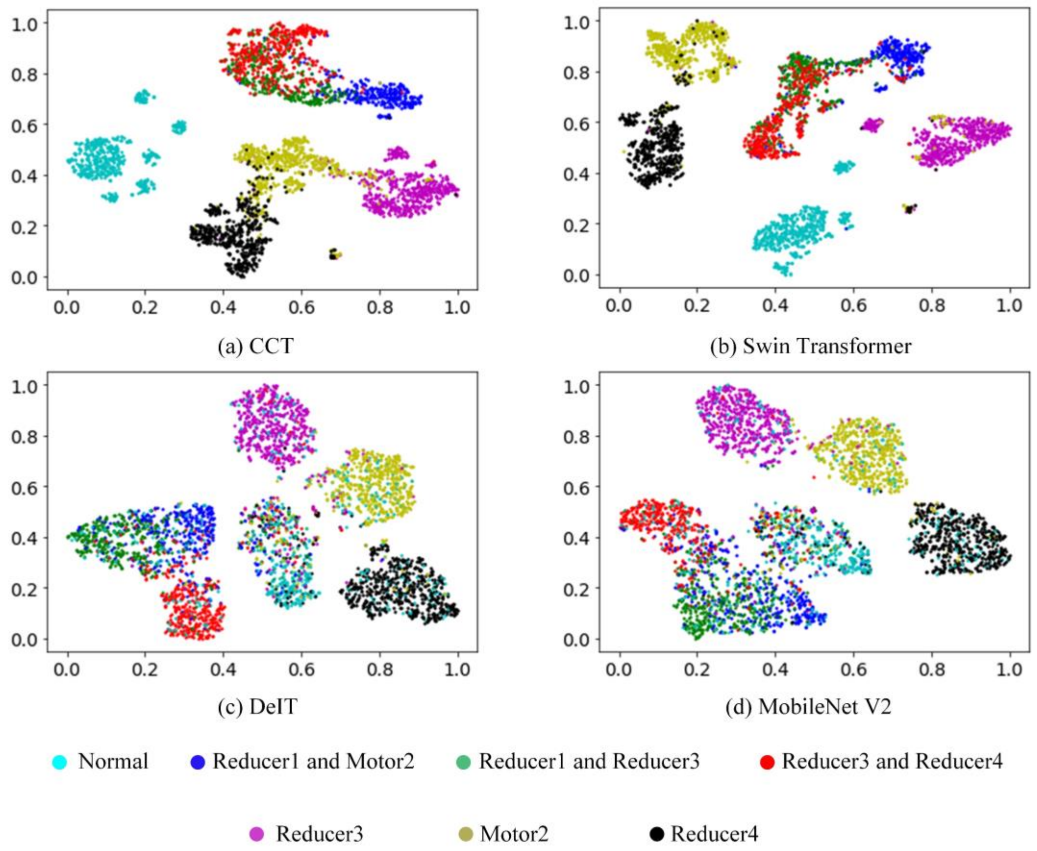
(b) Swin Transformer

(c) DeIT

(d) MobileNet V2

Figure 10. The compound fault diagnosis accuracy of all algorithms based on the denoised and noised images

For the purpose of better understanding the performance of the algorithms in this experiment, the t-SNE technique was employed to visualize the distribution of learned features. One of the five-fold results of t-SNE were illustrated in Figure 11. In Figure 11, the coordinates of each data sample are computed via t-SNE and then plotted as a scatter. In all sub-figures, it is evident that the single faulty categories are easy to be classified, while the samples in Reducer1_and_Reducer3_faulty and Reducer3_and_Reducer4_faulty are challenging to be separated. CCT and Swin Transformer shows good performance in the classification of Reducer1_and_Motor2_faulty. Figure 11 (b) demonstrates a better classification boundary of different groups. For the result of DiIT and MobileNet V2, the boundaries of all three compound fault categories are vague and the classification performance between the normal class and faulty classes is relatively poor.



(a) CCT

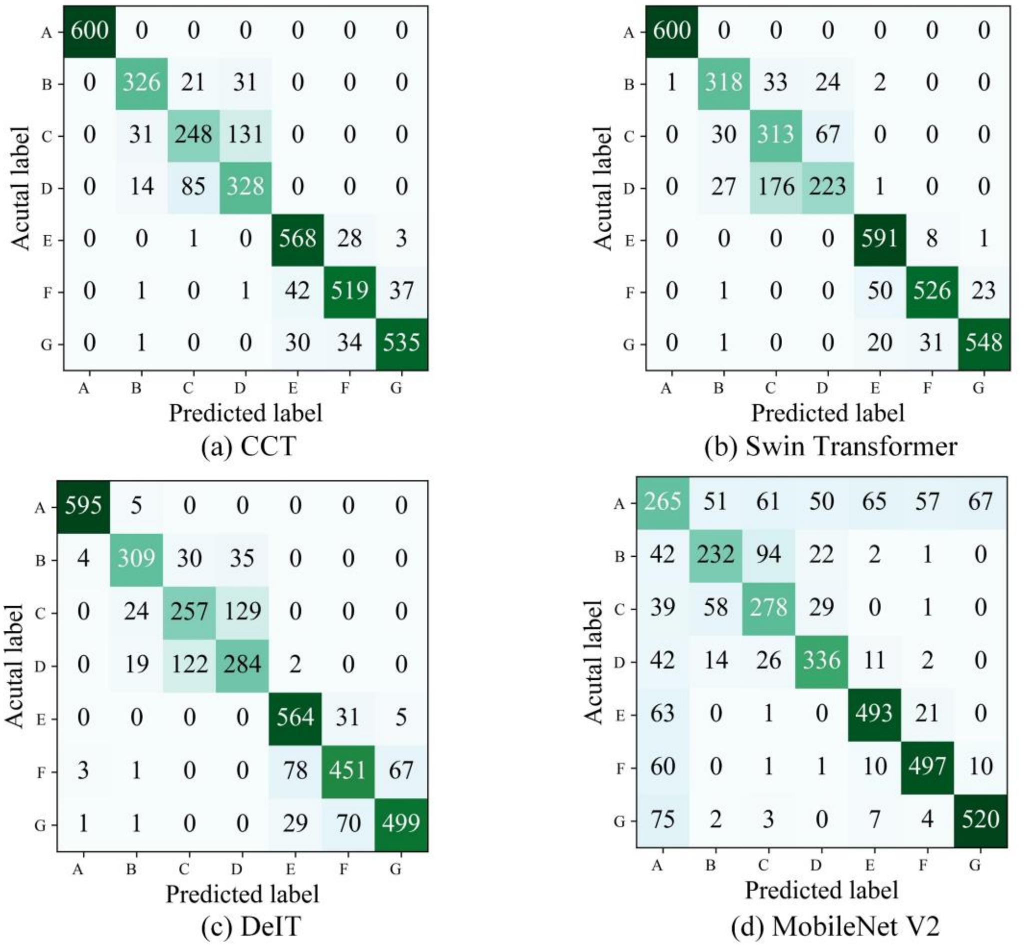
(b) Swin Transformer

(c) DeIT

(d) MobileNet V2

Figure 11. The results of t-SNE for compound fault diagnosis using different algorithms

The corresponding confusion matrixes to the results of Figure 11 are illustrated in Figure 12. The confusion matrixes (a) and (b) reveal that the CCT and Swin Transformer can correctly identify the normal class and show excellent performance in the classification of single fault. For the compound fault classification, CCT shows worse performance in the classification of Reducer1 and Reducer3 faulty, while Swin Transformer performs badly in the classification of Reducer3 and Reducer4 faulty. The result of DeIT indicates that its performance is slightly lower than that of CCT. Finally, the result of MobileNet V2 indicates that it only has the capability to accurately identify a single fault, while its performance in compound fault and Normal classification is not acceptable.



	A	B	C	D	E	F	G
A	600	0	0	0	0	0	0
B	0	326	21	31	0	0	0
C	0	31	248	131	0	0	0
D	0	14	85	328	0	0	0
E	0	0	1	0	568	28	3
F	0	1	0	1	42	519	37
G	0	1	0	0	30	34	535

(a) CCT

	A	B	C	D	E	F	G
A	600	0	0	0	0	0	0
B	1	318	33	24	2	0	0
C	0	30	313	67	0	0	0
D	0	27	176	223	1	0	0
E	0	0	0	0	591	8	1
F	0	1	0	0	50	526	23
G	0	1	0	0	20	31	548

(b) Swin Transformer

	A	B	C	D	E	F	G
A	595	5	0	0	0	0	0
B	4	309	30	35	0	0	0
C	0	24	257	129	0	0	0
D	0	19	122	284	2	0	0
E	0	0	0	0	564	31	5
F	3	1	0	0	78	451	67
G	1	1	0	0	29	70	499

(c) DeIT

	A	B	C	D	E	F	G
A	265	51	61	50	65	57	67
B	42	232	94	22	2	1	0
C	39	58	278	29	0	1	0
D	42	14	26	336	11	2	0
E	63	0	1	0	493	21	0
F	60	0	1	1	10	497	10
G	75	2	3	0	7	4	520

(d) MobileNet V2

A= Normal **B**= Reducer1 and Motor2 **C**= Reducer1 and Reducer3

D= Reducer3 and Reducer4 **E**= Reducer3 **F**= Motor2 **G**= Reducer4

Figure 12. The confusion matrix for compound fault diagnosis using different algorithms

In order to reveal the relationship between training data size and the algorithm performance, a data amount test was performed. The results were illustrated in Figure 13. The total number of training data is 37800. When 50% or lower of the training data was adopted for model training, the fault diagnosis accuracy of all the algorithms is lower than 60%. When the training data percentage locates in the range of 50% to 90%, it can be seen that the performance of CCT is better than the benchmarking algorithms. When 100% of the training data is adopted, the fault diagnosis accuracy of Swin Transformer is better than that of CCT. Meanwhile, the increment of the accuracy of Swin Transformers in the second half is larger than the rest algorithms.

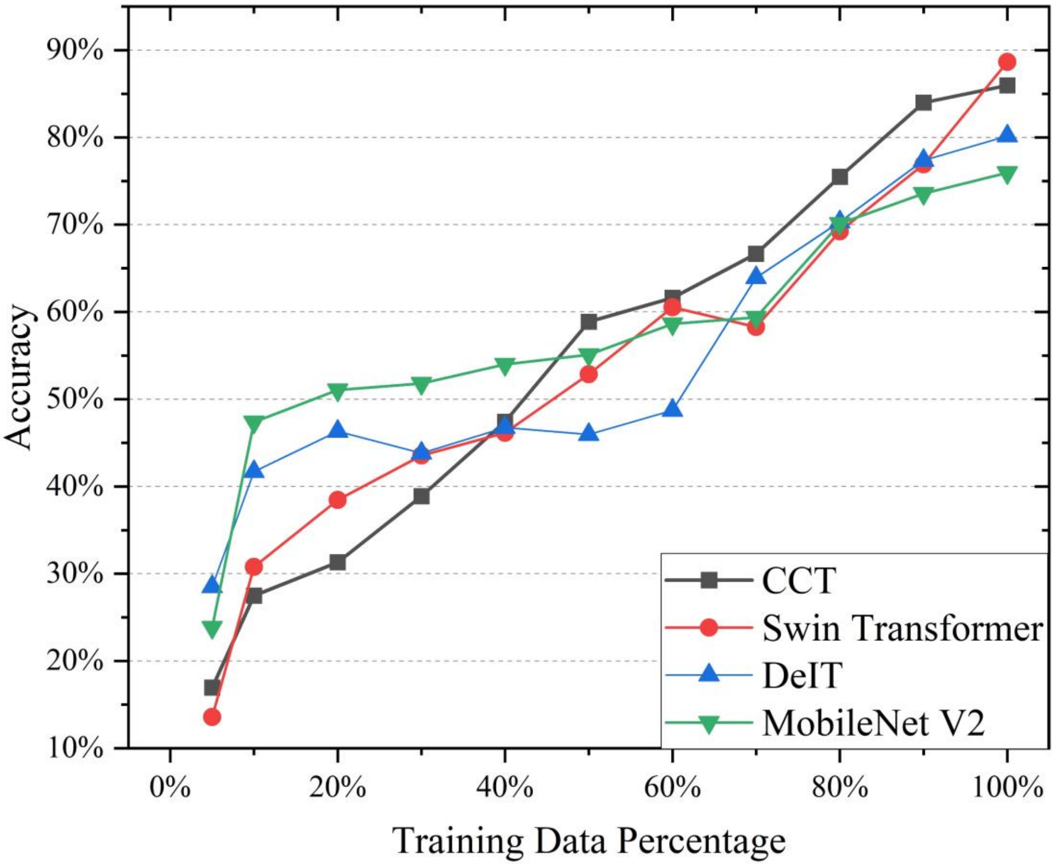


Figure 13. The relation between training data size and fault diagnosis accuracy

6. Discussions

Due to the high reliability of industrial robots, a large volume of failure data is hard to collect. When the failure data samples are limited, it is challenging to apply those large and powerful deep learning algorithms to train a compound fault diagnosis model with superior performance, especially when the data is collected from a noisy environment. Under this limitation, the lightweight efficient neural networks are easier to be trained and achieve better data efficiency. In our experiments, both compact Uformer and compact Transformer shows better performance in comparison with the state-of-the-art algorithms. When the noise level is low, state-of-the-art algorithms may yield better performance. From the experimental results of section 5.1, it is obvious that the compact Uformer can achieve better denoising performance in terms of PSNR and SSIM with higher computational efficiency in comparison with other state-of-the-art algorithms. When the data from all the categories were used for modelling, the compact Uformer shows merits when the noise level is lower than 5dB. When the noise level surpasses this threshold, the low-level noise is easy to be identified and removed by state-of-the-art algorithms. The results indicate that the proposed compact Uformer is suitable for the denoising of middle- and high-level noise. Furthermore, our experiments also evaluated the denoising performance of different algorithms based on the data from different categories. The results demonstrated that the noises in all three compound faults are challenging to remove in comparison with the results of the normal class and single fault class. The main reason is the time-frequency image of compound faults is rich in information. In a noisy environment, the boundary between the original time-frequency information and the noise becomes unclear, which obstacles the algorithms to identifying the noise.

The data used in this study were collected at the asset level. Deep learning modelling based on vibration data is the mainstream in fault diagnosis. At the asset level, the vibration signal is hard to be collected. In contrast, the feedback current data can be easily collected from the motor driver. Hence, the fault diagnosis approach based on feedback current data is easier to be deployed in the actual manufacturing scenario. The data was collected when an industrial robot perform the assigned task. In comparison with the data collected from the component level such as bearing or gearbox testing platform, the existing noise in the collected data of industrial robots is larger. When the extra Gaussian white noise was added, the fault-related features within the data were further overwhelmed. The results of compound fault diagnosis indicate that the denoise of time-frequency images can promote the algorithm performance when the noise level is lower than 10dB. When it is over this threshold, the denoise process is not necessary. The main reason is that the time-frequency images were obtained via CWT, which is able to restrain the noise to some degree. For the time-frequency images with 20 or 30dB noise, the fault patterns are still obvious so that deep learning algorithms are able to establish an accurate classification model based on these images with low-level noise. Meanwhile, when the noise level is lower than 2dB, the classification accuracy of all the algorithms is decreased dramatically. Hence, the suitable range of noise level for the compact Uformer denoising algorithm is 2dB to 10dB.

When only 50% of training data is available, the diagnosis accuracy of all the algorithms is lower than 60%, which is hard to be applied in the actual industrial scenario. However, when 60% to 90% of the training dataset is available, the performance of CCT is obviously better than the benchmarking algorithms. Even though Swin Transformer achieved better diagnosis accuracy in comparison with CCT when 100% training data is available, it requires a far larger computational cost. In the existing studies, it is well known that the larger network parameters are the guarantee of extraordinary performance. The classification accuracy of the Swin

Transformer is likely to be further promoted when a large dataset and more training epoch is available. However, such a large dataset and extremely high computational costs are not always affordable. In contrast, CCT as a lightweight neural network is easier to be trained and deployed. The compound fault diagnosis accuracy of CCT is only 2.32% lower than that of Swin Transformer, while its model parameters are 97.64% less than that of Swin Transformer. This promising performance enables CCT to be quickly deployed and transferred in the actual maintenance management of industrial robots.

In the current stage, we have collected the data from the fault injection experiment, in which faulty parts have incipient faults. However, the root causes of the faulty parts have not been identified due to the challenges in disassembling the faulty parts. In future works, it is worthwhile to investigate the root causes of the compound fault, which can provide the fault diagnosis task. Meanwhile, the proposed compact Uformer and CCT are only used in supervised modelling based on noisy data. The denoising and fault diagnosis performance of the compact Uformer and CCT can be further improved if the denoising and fault diagnosis knowledge can be obtained in advance. The denoising and fault diagnosis knowledge can be obtained via transfer learning and knowledge distillation techniques, which will be investigated in our future works. Meanwhile, only three types of compound faults were investigated in this study. When unseen compound faults happen, the well-trained fault diagnosis model may not be able to diagnose it accurately. Hence, it is worthwhile to investigate how to achieve the unseen compound fault diagnosis approach in our future works.

7. Conclusions

The maintenance management of industrial robots can gain considerable benefits from an accurate compound fault diagnosis. In order to achieve decent denoising and compound fault diagnosis performance without large computational cost, an integrated approach composed of two lightweight Transformer networks was proposed. An experimental study based on a real-world industrial robot dataset demonstrated the effectiveness of the proposed approach. The key findings of this study are: (1) Compact Uformer shows merits in the time-frequency images denoising in comparison with the prevailing benchmarking algorithms, especially in the middle and strong level noise; (2) When limited training data (60% -90% in this case) is available, CCT shows better fault diagnosis performance; (3) Both models are lightweight, which are computational efficient compared with those prevailing Transformer networks; (4) The proposed approach is able to achieve satisfactory compound fault diagnosis accuracy when the noise level in the range from 2dB to 10dB. In the future, the knowledge-informed neural networks and the unseen fault diagnosis approach will be further investigated to achieve better compound fault diagnosis accuracy for the industrial robot.

Acknowledgement

Our work is supported by multiple funds in China, including the Key Program of NSFC-Guangdong Joint Funds (U1801263), the Natural Science Foundation of Guangdong Province (2020B1515120010), Dedicated Fund for Promoting High-Quality Economic Development in Guangdong Province (Marine Economic Development Project) GDNRC [2021]44, Industrial core and key technology plan of ZhuHai City (ZH22044702190034HJL). Our work is also supported by Guangdong Provincial Key Laboratory of Cyber-Physical System (2020B1212060069).

References

[1] U. Izagirre, I. Andonegui, I. Landa-Torres, U. Zurutuza, A practical and synchronized data acquisition network architecture for industrial robot predictive maintenance in manufacturing assembly lines, Robotics and Computer-Integrated Manufacturing, 74 (2022) 102287.

[2] R. Zhang, J. Lv, J. Li, J. Bao, P. Zheng, T. Peng, A graph-based reinforcement learning-enabled approach for adaptive human-robot collaborative assembly operations, Journal of Manufacturing Systems, 63 (2022) 491-503.

[3] B. Wang, S.J. Hu, L. Sun, T. Freiheit, Intelligent welding system technologies: State-of-the-art review and perspectives, Journal of Manufacturing Systems, 56 (2020) 373-391.

[4] X. Zhou, H. Zeng, C. Chen, H. Xiao, Z. Xiang, An attention-enhanced multi-modal deep learning algorithm for robotic compound fault diagnosis, Measurement Science and Technology, 34 (2022) 014007.

[5] P. Lyu, K. Zhang, W. Yu, B. Wang, C. Liu, A novel RSG-based intelligent bearing fault diagnosis method for motors in high-noise industrial environment, Advanced Engineering Informatics, 52 (2022) 101564.

[6] J. Luo, H. Shao, H. Cao, X. Chen, B. Cai, B. Liu, Modified DSAN for unsupervised cross-domain fault diagnosis of bearing under speed fluctuation, Journal of Manufacturing Systems, 65 (2022) 180-191.

[7] W. Li, X. Zhong, H. Shao, B. Cai, X. Yang, Multi-mode data augmentation and fault diagnosis of rotating machinery using modified ACGAN designed with new framework, Advanced Engineering Informatics, 52 (2022) 101552.

[8] X.-B. Wang, X. Zhang, Z. Li, J. Wu, Ensemble extreme learning machines for compound-fault diagnosis of rotating machinery, Knowledge-Based Systems, 188 (2020) 105012.

[9] P. Liang, C. Deng, J. Wu, Z. Yang, J. Zhu, Z. Zhang, Compound fault diagnosis of gearboxes via multi-label convolutional neural network and wavelet transform, Computers in Industry, 113 (2019) 103132.

[10] Y. Wu, Z. Fu, J. Fei, Fault diagnosis for industrial robots based on a combined approach of manifold learning, treelet transform and Naive Bayes, Review of Scientific Instruments, 91 (2020) 015116.

[11] J. Long, J. Mou, L. Zhang, S. Zhang, C. Li, Attitude data-based deep hybrid learning architecture for intelligent fault diagnosis of multi-joint industrial robots, Journal of Manufacturing Systems, 61 (2021) 736-745.

[12] L. Chen, J. Cao, K. Wu, Z. Zhang, Application of Generalized Frequency Response Functions and Improved Convolutional Neural Network to Fault Diagnosis of Heavy-duty Industrial Robot, Robotics and Computer-Integrated Manufacturing, 73 (2022) 102228.

[13] Y. Liu, C. Chen, T. Wang, L. Cheng, An attention enhanced dilated CNN approach for cross-axis industrial robotics fault diagnosis, Autonomous Intelligent Systems, 2 (2022) 1-11.

[14] K. Lu, C. Chen, T. Wang, L. Cheng, J. Qin, Fault diagnosis of industrial robot based on dual-module attention convolutional neural network, Autonomous Intelligent Systems, 2 (2022) 1-12.

[15] A.H. Sabry, F.H. Nordin, A.H. Sabry, M.Z.A. Ab Kadir, Fault detection and diagnosis of industrial robot based on power consumption modeling, IEEE Transactions on Industrial Electronics, 67 (2019) 7929-7940.

[16] A.W.K. To, G. Paul, D. Liu, A comprehensive approach to real-time fault diagnosis during automatic grit-blasting operation by autonomous industrial robots, Robotics and Computer-Integrated Manufacturing, 49 (2018) 13-23.

[17] Y. Zhang, J. Ji, B. Ma, Fault diagnosis of reciprocating compressor using a novel ensemble empirical mode decomposition-convolutional deep belief network, Measurement, 156 (2020) 107619.

[18] S.N. Chegini, A. Bagheri, F. Najafi, Application of a new EWT-based denoising technique in bearing fault diagnosis, Measurement, 144 (2019) 275-297.

[19] J. Guo, Z. Si, J. Xiang, A compound fault diagnosis method of rolling bearing based on wavelet scattering transform and improved soft threshold denoising algorithm, Measurement, (2022) 111276.

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

[21] H. Wang, Z. Liu, D. Peng, Z. Cheng, Attention-guided joint learning CNN with noise robustness for bearing fault diagnosis and vibration signal denoising, ISA transactions, (2021).

[22] L. Zou, Y. Li, F. Xu, An adversarial denoising convolutional neural network for fault diagnosis of rotating machinery under noisy environment and limited sample size case, Neurocomputing, 407 (2020) 105-120.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, arXiv, (2017).

[24] G. Chen, K. Li, Y. Liu, Applicability of continuous, stationary, and discrete wavelet transforms in engineering signal processing, Journal of Performance of Constructed Facilities, 35 (2021) 04021060.

[25] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234-241.

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929, (2020).

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

[28] A. Rohan, I. Raouf, H.S. Kim, Rotate vector (RV) reducer fault detection and diagnosis system: towards component level Prognostics and Health Management (PHM), Sensors, 20 (2020) 6845.

[29] A. Rohan, Holistic Fault Detection and Diagnosis System in Imbalanced, Scarce, Multi-Domain (ISMD) Data Setting for Component-Level Prognostics and Health Management (PHM), Mathematics, 10 (2022).

[30] R. Huang, Y. Liao, S. Zhang, W. Li, Deep decoupling convolutional neural network for intelligent compound fault diagnosis, IEEE Access, 7 (2018) 1848-1858.

[31] J. Xu, L. Zhou, W. Zhao, Y. Fan, X. Ding, X. Yuan, Zero-shot learning for compound fault diagnosis of bearings, Expert Systems with Applications, 190 (2022) 116197.

[32] A. Dibaj, M.M. Ettefagh, R. Hassannejad, M.B. Ehghaghi, A hybrid fine-tuned VMD and CNN scheme for untrained compound fault diagnosis of rotating machinery with unequal-severity faults, Expert Systems with Applications, 167 (2021) 114094.

[33] Y. Jin, C. Qin, Y. Huang, C. Liu, Actual bearing compound fault diagnosis based on active learning and decoupling attentional residual network, Measurement, 173 (2021) 108500.

[34] R. Huang, J. Li, W. Li, L. Cui, Deep ensemble capsule network for intelligent compound fault diagnosis using multisensory data, IEEE Transactions on Instrumentation and Measurement, 69 (2019) 2304-2314.

[35] X. Lyu, Z. Hu, H. Zhou, Q. Wang, Application of improved MCKD method based on QGA in planetary gear compound fault diagnosis, Measurement, 139 (2019) 236-248.

[36] N. Li, W. Huang, W. Guo, G. Gao, Z. Zhu, Multiple enhanced sparse decomposition for gearbox compound fault diagnosis, IEEE Transactions on Instrumentation and Measurement, 69 (2019) 770-781.

[37] G. Jiang, H. He, P. Xie, Y. Tang, Stacked multilevel-denoising autoencoders: A new representation learning approach for wind turbine gearbox fault diagnosis, IEEE Transactions on Instrumentation and Measurement, 66 (2017) 2391-2402.

[38] B. Zhao, C. Cheng, G. Tu, Z. Peng, Q. He, G. Meng, An interpretable denoising layer for neural networks based on reproducing kernel Hilbert space and its application in machine fault diagnosis, Chinese Journal of Mechanical Engineering, 34 (2021) 1-11.

[39] J. Meng, H. Wang, L. Zhao, R. Yan, Adaptive sparse denoising and periodicity weighted spectrum separation for compound bearing fault diagnosis, Measurement Science and Technology, 32 (2021) 085011.

[40] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, A survey on vision transformer, IEEE transactions on pattern analysis and machine intelligence, (2022).

[41] Q. Yang, Y. Liu, J. Tang, T. Ku, Residual and Dense UNet for Under-Display Camera Restoration, in: European Conference on Computer Vision, Springer, 2020, pp. 398-408.

[42] C.-M. Fan, T.-J. Liu, K.-H. Liu, SUNet: Swin Transformer UNet for Image Denoising, arXiv preprint arXiv:2202.14009, (2022).

[43] Y. Thesia, M. Suthar, T. Pandya, P. Thakkar, Image Denoising with Self-adaptive Multi-UNET Valve, in: Soft Computing for Problem Solving, Springer, 2021, pp. 647-659.

[44] C. Yao, S. Jin, M. Liu, X. Ban, Dense residual Transformer for image denoising, Electronics, 11 (2022) 418.

[45] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, arXiv preprint arXiv:1710.09412, (2017).

[46] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, R. Timofte, Swinir: Image restoration using swin transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1833-1844.

[47] S.W. Zamir, A. Arora, S. Khan, M. Hayat, F.S. Khan, M.-H. Yang, L. Shao, Multi-stage progressive image restoration, in: Proceedings of the IEEE/CVF Conference on Computer

Vision and Pattern Recognition, 2021, pp. 14821-14831.

[48] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012-10022.

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

[50] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510-4520.