A multi-scale CNN-transformer hybrid network with parallel attention mechanism and local linear unit for cross-condition fault diagnosis

E Cai[1], Yangyang Li[2]
[1]School of Automobile, Chang'an University, Xi'an, China
[2]School of Energy and Electrical Engineering, Chang'an University, Xi'an, China
[1]Corresponding author
E-mail: [1]*caie8201@chd.edu.cn*, [2]*thunderrock@126.com*

Received 11 March 2026; accepted 5 May 2026; published online 8 June 2026
DOI https://doi.org/10.21595/jve.2026.26312

Abstract. Domain shift caused by variations in rotational speed and load under cross-condition scenarios may distort the statistical distribution and feature representation of vibration signals, posing a challenge to the reliable deployment of intelligent fault diagnosis systems for rotating machinery. Existing multi-scale CNN–Transformer hybrid architectures generally do not explicitly consider the influence of such condition-induced domain shifts. To address this issue, a multi-scale CNN-Transformer hybrid method integrating a Parallel Attention Mechanism (PAM) and a Local Linear Unit (LLU) is proposed to enhance the stability of feature representations under varying operating conditions. In the proposed method, raw vibration signals are directly used as inputs, and a multi-scale CNN module is employed to capture transient impact features across multiple temporal scales. The PAM then adjusts feature responses in both channel and temporal dimensions through parallel attention branches, enabling adaptive feature reweighting to alleviate condition-induced statistical variations. Furthermore, a Transformer encoder embedded with LLU is adopted as the backbone to incorporate local structural modeling through depthwise separable convolution while preserving the global dependency modeling capability of self-attention. Experiments on three benchmark datasets (PU, PHM09, and CWRU) show that the proposed method achieves average accuracies of 80.15 %, 93.96 %, and 98.17 %, respectively, under unseen operating conditions, consistently outperforming several representative comparative methods. Additional ablation studies further verify the contribution of PAM and LLU to robust feature representation learning. Moreover, noise robustness experiments under multiple signal-to-noise ratio (SNR) levels demonstrate that the proposed method maintains stable performance under moderate noise conditions, highlighting its practical reliability in realistic industrial environments. Our code is available at https://github.com/caie8201/Multi-Scale-CNN-Transformer-Hybrid-Network.

Keywords: cross-condition fault diagnosis, domain shift, CNN–transformer, parallel attention mechanism, local linear unit.

1. Introduction

The operational stability of rotating machinery directly affects the efficiency and reliability of industrial production. The health condition of key components, such as rolling bearings and gears, is critical to the safe and efficient operation of industrial equipment. Therefore, developing accurate and robust intelligent fault diagnosis methods for timely fault identification has become an urgent requirement to guarantee the safe operation of key industrial systems and prevent unplanned downtime [1, 2].

In recent years, with the rapid growth of industrial big data and the significant improvement in the performance of graphics processing units (GPUs), deep learning technologies have sparked a surge of research in the field of intelligent fault diagnosis [3, 4]. Various deep learning models, including convolutional neural networks (CNNs), deep autoencoders (DAEs), deep belief

networks (DBNs), recurrent neural networks (RNNs), generative adversarial networks (GANs), and Transformers, have demonstrated superior performance compared with traditional diagnosis methods. Among these models, CNNs dominated early studies due to their strong capability in local feature extraction [5]. In recent years, however, Transformer-based models and their hybrid architectures have shown greater potential in complex tasks such as cross-condition fault diagnosis, attributed to their inherent ability to model long-range dependencies [6].

Nevertheless, in practical industrial environments, domain shift caused by continuous variations in operating conditions, such as rotational speed and load, is extremely common. These variations lead to significant distribution discrepancies between training data and test data. When deep learning-based diagnosis models are applied to unseen operating conditions, their limited generalization ability often leads to false alarms or missed detections. This issue severely limits their practical engineering value. Essentially, variations in operating conditions alter the statistical characteristics of vibration signals, causing instability in the internal feature responses of deep models and thereby degrading cross-domain diagnostic performance. Currently, research on domain shift in fault diagnosis mainly focuses on domain adaptation (DA) methods [7, 8]. However, DA approaches typically rely on partially labeled or unlabeled samples from the target domain, which not only increases deployment costs in industrial applications but may also lead to overfitting to specific target domains. In contrast, Domain Generalization (DG) aims to capture domain-invariant knowledge while suppressing domain-specific information [9]. Since DG methods do not require any information from the target domain during training, they are more suitable for real industrial deployment. Consequently, DG has been recognized as a more practical and promising technical direction.

Existing DG methods can generally be categorized into three categories [1]: data augmentation, domain-invariant representation learning, and learning strategy-based methods. Among them, domain-invariant representation learning has attracted increasing attention because it can explicitly decouple domain-invariant and domain-specific features. Meanwhile, to simultaneously capture local features and global dependencies, hybrid architectures combining CNNs and Transformers have gradually become an important research direction for invariant representation learning. For instance, W. Liu et al. [10] proposed an Efficient Convolution Transformer Network (ECTN), which integrates the short-time Fourier transform with a standard Transformer encoder to effectively model long-range temporal dependencies under varying operating conditions. S. You et al. [11] designed a Temporal Fusion Transformer (TFT), which improves generalization performance under the unseen conditions through adaptive multi-scale feature fusion and the dynamic block auto-encoding mechanism. Z. Lu et al. [12] developed a CNN–Transformer hybrid method that employs multi-scale CNNs to extract robust local features and utilizes multi-head self-attention in Transformers to establish global temporal relationships, achieving end-to-end cross-domain fault diagnosis. These studies indicate that introducing Transformers as the backbone for global modeling has become the promising trend for improving the diagnostic capability of DG models.

Despite these advances, most existing DG methods (such as CNN-C [13], DANN [14], DIFE [15], DGNIS [16], and IEDGNet [17]) still rely on conventional convolutional neural networks, attention mechanisms, adversarial learning, or explicit feature alignment strategies to learn domain-invariant features, and the potential of Transformers for cross-domain generalization has not been fully explored. Meanwhile, the limited DG approaches based on CNN-Transformer hybrid architectures still suffer from several shortcomings. First, the outputs of multi-scale CNN branches are typically concatenated or averaged directly, lacking dynamic perception and adaptive weight assignment of feature importance across different scales. Second, the standard Transformer block has limited capability in modeling local temporal patterns, making it difficult to effectively capture the critical transient impact features in bearing and gear fault signals. Third, the importance of the channel and temporal features is usually not jointly optimized, resulting in insufficient focus on fault-sensitive information under operating condition disturbances.

To address these issues, this paper focuses on the unstable feature responses caused by domain

shift in cross-condition scenarios and constructs a hierarchical modeling strategy consisting of input response modulation, local structure enhancement, and global dependency modeling. Specifically, after the multi-scale feature extraction, the PAM module is introduced to adaptively regulate feature responses through parallel channel-temporal branches, enabling dynamic feature reweighting. This process alleviates statistical discrepancies caused by operating condition variations. Furthermore, the LLU is embedded into the Transformer encoder. By introducing the local structural modeling capability through depthwise separable convolution, the model maintains the ability to capture long-range dependencies while enhancing the stable perception of transient impact features.

Compared with existing CNN-Transformer hybrid approaches that primarily emphasize architectural fusion, this work focuses on improving the stability of feature representations under varying operating conditions. In the proposed method, the PAM regulates feature responses across different dimensions, while the LLU enhances the capability of the Transformer to model transient impact features through improved local structural modeling. These two modules cooperate from the perspectives of feature response regulation and internal representation enhancement, forming a structured modeling strategy to mitigate the impact of domain shift in cross-condition fault diagnosis. The main contributions of this paper are summarized as follows:

1) A feature response regulation strategy is introduced to alleviate cross-condition domain shift. The proposed PAM jointly models the importance of features in both the channel and temporal dimensions, enabling adaptive feature reweighting and improving the robustness of feature representations under varying operating conditions.

2) A locally enhanced Transformer encoding unit is developed by embedding the LLU into the self-attention architecture. This design introduces controllable local structural modeling capability through depthwise separable convolution, allowing the model to better capture transient impact features while preserving global dependency modeling.

3) An end-to-end domain generalization fault diagnosis framework is constructed by integrating the multi-scale CNN feature extractor, PAM-based feature response regulation, and the LLU-enhanced Transformer encoder. Experiments on multiple cross-condition datasets demonstrate the effectiveness and robustness of the proposed method. In addition, noise robustness experiments under different SNR levels are conducted to evaluate the reliability of the proposed framework in noisy environments, and the model complexity and inference efficiency are further analyzed to assess its practical applicability.

The remainder of this paper is organized as follows: Section 2 briefly reviews the relevant theoretical background. Section 3 describes the overall framework and key technical components of the proposed method. Section 4 presents comparative experiments and ablation studies on three public bearing and gear datasets, along with noise robustness evaluations conducted on selected datasets, to verify the effectiveness and reliability of the proposed method. Finally, Section 5 concludes the paper and discusses future research directions.

2. Theoretical background

This section briefly reviews the theoretical fundamentals related to the Parallel Attention Mechanism (PAM), multi-scale CNNs, Transformers, Local Linear Unit (LLU), and domain generalization (DG).

2.1. Parallel attention mechanism (PAM)

The channel attention mechanism is an effective approach for improving convolutional operations and has demonstrated significant potential in enhancing the performance of deep CNNs. Traditional attention mechanisms typically focus only on feature responses along the channel dimension, while their capability to model temporal dependencies in one-dimensional time-series signals is limited. To overcome this limitation, S. Woo et al. [18] proposed the

Convolutional Block Attention Module (CBAM), a lightweight structure that sequentially integrates channel attention and spatial attention to enhance feature selection capability. Recently, Y. Sun et al. [19] further developed the PAM, which models the importance of channel and temporal dimensions through two parallel branches. The outputs of these branches are integrated to enhance the discriminative capability of one-dimensional temporal features. Its typical mathematical formulation is given by:

$$X_{out} = GELU\big(AMMP(X)\big) \odot X + GELU\big(Conv_{1\times1}(X)\big) \odot X, \tag{1}$$

where $X_{out} \in \mathbb{R}^{L\times C}$ denotes the output feature map, $AMMP(\cdot)$ represents adaptive mixing mean pooling, $\mathrm{Conv}_{1\times1}(\cdot)$ denotes the one-dimensional convolution, $GELU(\cdot)$ is the Gaussian Error Linear Unit activation function, and $\odot$ represents element-wise multiplication.

2.2. Multi-scale CNNs

As one of the most representative deep learning models, CNNs are naturally suitable for extracting local temporal features from one-dimensional vibration signals, owing to their properties of local connectivity, weight sharing, and spatial pooling [20]. However, under cross-condition scenarios, the time–frequency structures of fault signals are highly complex. Consequently, vibration signals often exhibit pronounced multi-scale characteristics. A single convolutional kernel has a limited receptive field and is therefore insufficient to comprehensively capture multi-band transient impact features induced by different operating conditions and fault classes (e.g., bearing pitting or gear tooth breakage). To address this issue, multi-scale CNN has been widely adopted. By deploying multiple convolution kernels of different sizes in parallel, the network can simultaneously capture fine-grained local details and coarse-grained contextual information within a single forward propagation, enabling the automatic learning of complementary and discriminative multi-scale diagnostic features from complex vibration signals [21]. Given an input signal $x \in R^L$, the output of the multi-scale CNN branch is defined as:

$$F_{ms} = \big[\, Conv_{k_1\times1}(x) \oplus Conv_{k_2\times1}(x) \oplus \cdots \oplus\ Conv_{k_n\times1}(x)\big], \tag{2}$$

where $Conv_{k_i\times1}(\cdot)$ denotes a one-dimensional convolution with kernel size k_i, n is the number of parallel convolution branches, and $\oplus$ represents channel-wise concatenation.

2.3. Transformer

The Transformer [22] was originally designed for natural language processing tasks. It consists of a stack of Transformer blocks, forming a deep network architecture that includes multi-head self-attention (MHSA), feed-forward network (FFN), and layer normalization (LN). The core component, the multi-head self-attention mechanism, is capable of modeling long-range dependencies between any two-time steps in a sequence. For an input sequence $X \in \mathbb{R}^{L\times d}$, the MHSA operation is calculated as:

$$MHSA(X) = \mathrm{Concat}(head_{1,\cdots,}head_h)W^Q, \tag{3}$$

$$head_i = \mathrm{Attention}\big(QW_i^Q, KW_i^K, VW_i^V\big), \tag{4}$$

$$\mathrm{Attention}(Q,K,V) = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{5}$$

where W_i^Q, W_i^K, and W_i^V are the learnable parameters, and d_k is the dimension of query and key vectors.

2.4. Local linear unit (LLU)

Although Transformers possess strong global modeling capability, directly using a conventional Transformer architecture to extract fault features may result in the loss of detailed fault information due to the absence of the inherent inductive bias provided by CNNs [23]. To address this issue, C. Weng et al. [23] proposed introducing the LLU into the Transformer architecture to enhance its ability to capture local features. The output feature embedding is defined as:

$$z_t = GELU(f^{k\times 1;D}(U_{t-1}(X))), \tag{6}$$

where $f^{k\times 1;D}(\cdot)$ denotes the one-dimensional convolution function, $k \times 1$ and D represent the kernel size and number of filters, respectively, $U_{t-1}(\cdot)$ denotes the input transformation function at layer $t-1$, and $GELU(\cdot)$ denotes the activation function.

2.5. DG Learning

The primary objective of DG is to learn transferable representations from multiple labeled source domains, enabling the model to maintain stable performance on unseen target domains, where target-domain data are completely unavailable during training [24]. Accordingly, the learning objective of DG is formulated as minimizing the expected risk on the unknown target-domain distribution D_t:

$$\min_f \mathbb{E}_{(x,y)\sim\mathcal{D}_t}[\ell(f(x),y)], \tag{7}$$

where f denotes the learned function and ℓ represents the loss function.

3. Proposed method

To address the domain shift challenge in cross-condition fault diagnosis of rotating machinery, this paper proposes an end-to-end domain-generalized fault diagnosis method using raw vibration signals as input. First, multi-scale one-dimensional convolutions are employed to capture local fault features in parallel across different temporal scales. Subsequently, the PAM module is introduced to the extracted feature representations to simultaneously optimize feature responses in both the channel and temporal dimensions, thereby suppressing interference induced by operating condition variations. Finally, the LLU is embedded into the Transformer blocks to compensate for the limitations of the self-attention mechanism in modeling transient impact features. This design enables collaborative learning of global dependencies and local details. The entire model is trained using only the source-domain data without requiring any target-domain information, facilitating high-accuracy fault diagnosis under unseen operating conditions.

3.1. Overall architecture

The overall framework proposed in this paper is illustrated in Fig. 1. The model takes raw vibration signals as input and first processes them through parallel multi-scale CNN branches. Specifically, one-dimensional convolutions with kernel sizes of 31, 63, and 127 are used to capture local fault features at different temporal scales. The outputs of all branches are concatenated along the channel dimension and subsequently compressed through a 1×1 convolution layer to map them into a unified embedding dimension. The compressed high-dimensional features are then fed into the PAM module, which simultaneously models the channel importance and key temporal weights to enhance the discriminative capability of feature representations under cross-condition scenarios. The PAM-enhanced feature sequence is further compressed into a fixed-length token

sequence using the adaptive max pooling (AMP). The resulting sequence is concatenated with the learnable classification token ([CLS] token) and added to the learnable positional embeddings before being fed into a stack of deep Transformer encoder layers with embedded LLU modules. Within each Transformer block, the depthwise separable convolution explicitly models local temporal dependencies and is fused with the self-attention pathway through residual connections, enabling collaborative learning of local details and global contextual information. Finally, the [CLS] is normalized using Layer Normalization (LN) and mapped to the predefined fault categories through a linear classifier to produce the final diagnosis result.

This architecture effectively combines the local inductive bias of CNNs with the long-range dependency modeling capability of Transformers. Meanwhile, the PAM module strengthens feature focusing under cross-condition scenarios, and the LLU module enhances the modeling of transient impact responses, thereby effectively alleviating performance degradation caused by domain shift and enabling robust and accurate fault diagnosis under unseen operating conditions.

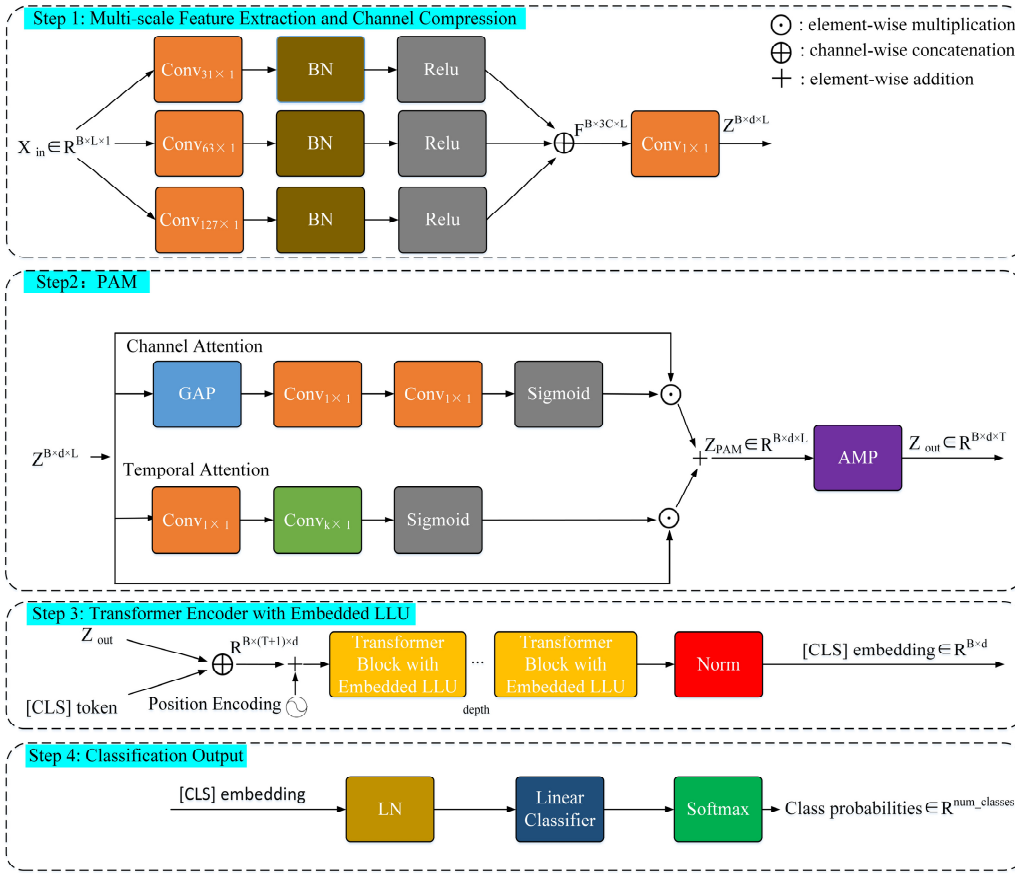


Fig. 1. Structure of the proposed domain-generalizable fault diagnosis method

3.2. Multi-Scale feature extraction and channel compression

To comprehensively capture multi-band transient fault features excited under different operating conditions, three parallel one-dimensional convolutional branches are adopted for feature extraction. The kernel sizes are 31, 63, and 127, which are intended to capture high-frequency impulsive features, mid-frequency modulation patterns, and low-frequency trend information. To avoid domain-specific bias caused by complex fusion mechanisms, the outputs of all branches are directly concatenated along the channel dimension. A 1×1 convolution layer is

then applied to compress the concatenated feature map into a unified embedding dimension d, producing a high-dimensional feature map $Z \in \mathbb{R}^{B\times d\times L}$, where B denotes the batch size and L represents the sequence length.

The resulting feature map is subsequently input into the PAM module described in Section 3.3 to perform joint feature reweighting along both the channel and temporal feature dimensions. After attention enhancement, the Adaptive Max Pooling (AMP) operation is applied to compress the temporal length into a predefined number of tokens T, yielding the token sequence $Z_{tok} \in \mathbb{R}^{B\times d\times T}$. This strategy effectively reduces computational complexity while preserving the peak amplitude points of transient impact signals, which are crucial for accurate fault diagnosis.

3.3. PAM in the proposed method

Under cross-condition scenarios, different channel features and temporal signal segments often exhibit different response intensities to fault information. Consequently, some discriminative fault features may be weakened or even overwhelmed, thereby degrading the model's capability to characterize critical fault patterns. To mitigate this issue, the modified PAM module is designed to adaptively reweight feature responses related to faults along both the channel and temporal dimensions.

Existing PAM approaches typically rely on the parameterized Adaptive Mixing Mean Pooling (AMMP) combined with a single convolution layer. However, such designs may introduce domain-specific bias and are often inadequate for modeling local temporal signal structures associated with transient impact. To address these limitations, this study proposes a simplified and more physically interpretable PAM variant. Specifically, the Global Average Pooling (GAP) is used to replace AMMP, eliminating additional parameterization; the temporal branch employs two consecutive 1×1 convolution layers to better capture local temporal patterns; the final attention outputs are normalized using a sigmoid activation function and fused through weighted summation. The proposed PAM operates on the high-dimensional features extracted by the multi-scale CNN. Given the input feature map $Z \in \mathbb{R}^{B\times d\times L}$, where B is the batch size, L is the sequence length, and d denotes the embedding dimension, the channel and temporal attention branches are defined as:

$$w_c = \sigma\left(Conv_{1\times1}^{(2)}\left(Conv_{1\times1}^{(1)}\left(GAP(Z)\right)\right)\right), \tag{8}$$

$$w_t = \sigma(Conv_{k\times1}(Conv_{1\times1}(Z))), \tag{9}$$

where $GAP(\cdot)$ denotes Global Average Pooling, $Conv_{k\times1}$ represents a one-dimensional convolution with kernel size k, $\sigma(\cdot)$ denotes the Sigmoid activation function, $Conv_{1\times1}^{(1)}$ and $Conv_{1\times1}^{(2)}$ represent the first and second 1×1 convolution layers in the channel attention branch.

The final output is obtained by element-wise summation of the channel-weighted and temporal-weighted features. This lightweight design adopts a fixed structure without introducing additional adaptive parameters, enabling effective joint focusing on fault-sensitive channels and key temporal segments, thereby providing a robust feature basis for subsequent domain-invariant representation learning.

3.4. Transformer encoder with embedded LLU

Although the multi-scale CNN effectively captures local transient fault features, its receptive field is still constrained by fixed convolution kernel sizes, making it difficult to model long-term fault evolution patterns across cycles. To overcome this limitation, the Transformer encoder is introduced to model long-range dependencies in the sequence using the Multi-Head Self-Attention (MHSA) mechanism. However, standard Transformers rely entirely on data-driven attention weights and lack local inductive bias, which may cause them to overlook critical local waveform

details when processing short-duration high-energy impact signals such as bearing pitting.

Inspired by the work of C. Weng et al. [23], this study introduces the LLU based on depthwise separable convolution, which is embedded within each Transformer block in a residual manner, as shown in Fig. 2. This design enhances the Transformer's capability to model local temporal structures, such as the waveform details of transient impact, while maintaining low computational cost. The forward computation of the LLU is given by:

$$X_{local} = GELU\left(BN\left(Conv_{depthwise}^{(k)}(X)\right)\right) \in \mathbb{R}^{B\times T\times d}, \tag{10}$$

where $Conv_{depthwise}^{(k)}$ denotes the depthwise separable convolution with kernel size k, BN represents Batch Normalization, and $GELU(\cdot)$ denotes the Gaussian Error Linear Unit activation function. The extracted local features are then added to the input of the current Transformer block through a residual connection. After the Layer Normalization (LN), the features are further processed by the MHSA and the feed-forward network (FFN). The detailed computation procedure is as follows:

$$X_1 = X + X_{local}, \tag{11}$$
$$X_2 = X_1 + MHSA\big(LN(X_1)\big), \tag{12}$$
$$X_{out} = X_2 + FFN\big(LN(X_2)\big). \tag{13}$$

In addition, the Pre-LayerNorm structure is adopted, where normalization is performed before both the MHSA and FFN modules. This design enhances training stability, particularly in small-sample DG scenarios.

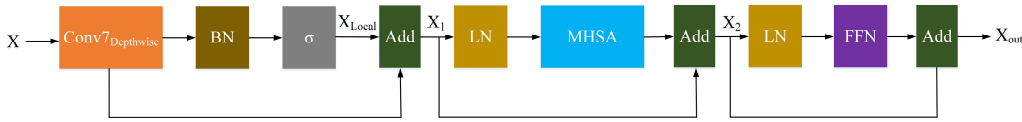


Fig. 2. Structure of the proposed Transformer block with embedded LLU

3.5. Classification output

Finally, the model maps the learned feature representation to predefined fault categories through a linear classifier. Specifically, the [CLS] token in the Transformer output is adopted as the global feature representation. After the Layer Normalization (LN), it is input into a linear classification layer to generate classification logits, which are used for the final fault diagnosis.

4. Experiments and results

To evaluate the effectiveness of the proposed method for cross-condition DG fault diagnosis, experiments were conducted on three publicly available bearing and gear fault datasets. All experiments strictly follow the DG experimental setting, where only source-domain operating condition-related data are available during training, while the target-domain operating conditions remain completely unseen.

4.1. Datasets and task settings

Three benchmark datasets were used in this study, namely Case Western Reserve University (CWRU), Paderborn University (PU), and PHM09 datasets. The operating condition variations considered in this study include variable rotational speed (PU and PHM09 datasets) and variable load (CWRU dataset). The key characteristics of each dataset are summarized in Table 1.

Table 1. Details of datasets

Dataset	Number of fault classes	Fault description
CWRU	10	Bearing faults including inner race, outer race, and rolling element defects introduced by electrical discharge machining (EDM), with three defect sizes (0.007 inch, 0.014 inch, 0.021 inch and 0.028 inch) [25]
PU	14	Real bearing faults generated through accelerated lifetime testing; detailed descriptions of the fault categories can be found in Table 5 of [26]
PHM09	8	Compound faults involving spur gear, rolling bearing, and transmission shaft: normal, gear wear, gear tooth breakage, gear eccentricity, bearing inner race fault, outer race fault, rolling element fault, and shaft imbalance [27]

To ensure fair comparison with existing methods, the task partition strategy and naming convention strictly follow the experimental settings in Reference [2]. For the PU and PHM09 datasets, four diagnostic tasks are constructed for each dataset, resulting in eight tasks in total. For each task, three operating conditions from the same machine are selected as the source domains, and the remaining condition is used as the target domain. The operating conditions and task settings of the PU dataset are shown in Tables 2 and 3, respectively.

The PHM09 dataset task configuration is shown in Table 4.

Table 2. Operating conditions for the PU dataset

Condition ID	Speed (rpm)	Load (Nm)	Radial force (N)
C0	900	0.7	1000
C1	1500	0.1	1000
C2	1500	0.7	400
C3	1500	0.7	1000

Table 3. Diagnostic task settings for the PU dataset

Task ID	Source domains	Target domain
T0	C1, C2, C3	C0
T1	C0, C2, C3	C1
T2	C0, C1, C3	C2
T3	C0, C1, C2	C3

Table 4. Diagnostic task settings for the PHM09 dataset

Task ID	Source domains	Target domain
T0	30 Hz, 35 Hz, 40 Hz	45 Hz
T1	30 Hz, 35 Hz, 45 Hz	40 Hz
T2	30 Hz, 40 Hz, 45 Hz	35 Hz
T3	35 Hz, 40 Hz, 45 Hz	30 Hz

In addition, the proposed method is compared with a representative CNN–Transformer hybrid architecture, namely the Efficient Convolution Transformer Network (ECTN) [10], to further evaluate its performance in small-sample DG scenarios. ECTN also adopts a CNN-Transformer hybrid architecture and has demonstrated effectiveness in cross-condition tasks on the CWRU dataset. Under the same DG task settings, the multi-source to single-target (3→1) tasks are further extended by constructing six single-source to single-target tasks, as shown in Table 5.

4.2. Experimental settings and comparison methods

The input data for all datasets consist of single-channel raw vibration signals. Each signal segment is first preprocessed using fixed-length truncation and sample-wise z-score normalization before being fed into the model. The signal lengths adopted in the experiments are 3200, 6660, and 3200 for the PU, PHM09, and CWRU datasets respectively. To ensure fairness across different datasets, the core hyperparameters of the model are kept consistent across all

experiments. The embedding dimension is set to $d = 72$, the number of stacked Transformer blocks is depth = 3, the batch size is set to 64, the token length is 64, and the number of attention heads in the multi-head self-attention mechanism is 4. In addition, to control model complexity and prevent overfitting, dropout with a rate of 0.1 is applied in the feed-forward network (FFN) within the Transformer blocks.

Table 5. Diagnostic task settings for the CWRU dataset

Task ID	Source domain (Horsepower)	Target domain (Horsepower)
T0	1	2
T1	1	3
T2	2	1
T3	2	3
T4	3	1
T5	3	2

The model is trained using the Adam optimizer with an initial learning rate of $lr = 0.003$ and a minimum learning rate of $lr_{min} = 0.0001$. A cosine annealing learning rate scheduler with warm restarts is adopted, where the initial cycle length is set to $T_0 = 30$ and the cycle multiplication factor is $T_{mult} = 2$, and the maximum number of training epochs is 90. To ensure reproducibility, the random seeds of both NumPy and PyTorch libraries are fixed in all experiments. Each task is independently conducted five times, and the average diagnostic accuracy along with the corresponding standard deviation is reported. All experiments are conducted on an NVIDIA GeForce RTX 4060 Laptop GPU with the PyTorch 2.4.1 framework.

To comprehensively and fairly evaluate the performance of the proposed method, comparisons are conducted with several representative approaches, whose implementations are provided in the survey paper [2]. These methods include the baseline approach AGG [2]; domain alignment methods such as Domain-Adversarial Neural Network (DANN) [14], Maximum Mean Discrepancy (MMD) [28], CORAL [29], and Triplet loss [30]; DG methods including Multi-domain Mixup [31] and Meta-Learning DG (MLDG) [32]; and the ensemble-based method Domain Adaptive Ensemble Learning (DAEL) [33].

To further validate the effectiveness of the key components of the proposed method, four ablation variants are designed as internal baseline models. The first variant, denoted as Ours-A, is a CNN-only baseline that retains only the multi-scale CNN and global average pooling, with the PAM, LLU, and Transformer modules removed. The second variant, denoted as Ours-B, is constructed by adding a standard Transformer encoder to Ours-A. The third variant, denoted as Ours-C, further integrates the PAM module into Ours-B. The fourth and final variant, denoted as Ours-D, corresponds to the complete proposed model, incorporating the Transformer encoder along with both the PAM and LLU modules.

4.3. Main results analysis

4.3.1. Cross-condition diagnostic accuracy analysis

Tables 6, 7, and 8 present the detailed diagnostic results on the PU, PHM09, and CWRU datasets respectively. It can be seen that the proposed method achieves the best performance across all datasets and demonstrates high stability and strong DG capability under different equipment types, transfer directions, and task settings.

On the PU dataset (14 fault classes), the complete model (Ours-D) achieves an average diagnostic accuracy of 78.46 %, achieving an improvement of 11.56 % compared with the best competing method, Mixup (66.90 %). On the PHM09 dataset (8 fault classes), Ours-D achieves an average diagnostic accuracy of 93.96 %, exceeding the best competing method Mixup (83.90 %) by 10.06 %. For the CWRU dataset (10 fault classes), six single-source-domain diagnostic tasks were constructed. The proposed method achieves an average accuracy of

98.17 %, outperforming the comparison model ECTN (94.69 %). Moreover, a 100 % diagnostic accuracy is achieved in several transfer tasks. These results indicate that even under extremely challenging scenarios where the training data are derived from only a single operating condition and the available sample size is limited, the proposed method maintains excellent diagnostic robustness and reliability.

Table 6. Diagnostic results (%) on PU dataset

Methods	T0	T1	T2	T3	Average
AGG	35.10	90.40	45.90	88.80	65.05
DANN	34.50	88.30	43.90	86.80	63.38
MMD	35.80	88.70	44.30	87.80	64.15
CORAL	38.90	91.10	45.00	88.40	65.85
Triplet loss	38.40	87.90	42.90	86.40	63.90
Mixup	47.10	88.30	48.70	83.50	66.90
MLDG	35.50	88.50	44.50	88.60	64.28
DAEL	36.20	80.00	35.30	73.50	56.25
Ours-A	41.21±5.47	96.04±0.04	52.68±3.44	98.46±0.19	72.19
Ours-B	51.71±4.49	96.04±0.32	60.02±4.83	97.66±1.21	76.36
Ours-C	51.29±4.63	96.16±0.13	65.91±6.86	98.45±0.16	77.95
Ours-D	50.93±5.86	96.66±0.15	67.57±3.26	98.66±0.42	78.46

Table 7. Diagnostic results (%) on PHM09 dataset

Methods	T0	T1	T2	T3	Average
AGG	72.80	89.50	90.80	79.90	83.25
DANN	75.10	88.40	90.00	80.40	83.48
MMD	73.80	89.50	90.50	79.90	83.43
CORAL	73.00	90.40	91.30	80.20	83.73
Triplet loss	73.50	87.70	87.20	79.10	81.88
Mixup	69.20	91.70	92.10	82.60	83.90
MLDG	73.40	90.00	91.10	79.30	83.45
DAEL	59.90	92.20	92.30	68.10	78.13
Ours-A	81.00±3.71	97.88±0.41	99.19±0.58	84.75±4.27	90.71
Ours-B	72.31±4.58	92.50±3.96	90.81±3.68	76.00±4.39	82.91
Ours-C	81.62±4.82	95.19±2.36	90.38±2.56	80.06±4.31	86.81
Ours-D	88.94±3.17	99.56±0.54	99.69±0.48	87.63±3.11	93.96

Table 8. Diagnostic results (%) on CWRU dataset

Methods	T0	T1	T2	T3	T4	T5	Average
ECTN	99.37	92.38	96.67	97.60	85.53	96.57	94.69
Ours-A	100.00±0	91.50±3.26	98.00±2.47	100.00±0	81.50±3.00	92.10±3.73	93.85
Ours-B	100.00±1.11	98.30±1.03	98.00±0.63	98.90±1.07	90.60±2.20	98.20±1.44	97.33
Ours-C	99.10±0	96.30±2.20	97.80±0.68	99.00±0.32	90.00±0	98.90±0.86	96.85
Ours-D	100.00±0	99.90±0.20	98.50±0.84	100±0	90.70±0.93	99.90±0.20	98.17

4.3.2. Visualization analysis

To visually verify the discriminative capability of the features learned by the proposed method under unseen operating conditions, t-SNE visualization was performed on the PU dataset for Task T3, accompanied by confusion matrix analysis to further examine the classification behavior.

As shown in Fig. 3, the features learned by Ours-D exhibit clearly separable cluster structures in the feature space across the 14 fault categories, which include inner-race and outer-race faults with single-point, multi-point, and repeated distributions, as well as distributed fatigue pitting and plastic indentation. The intra-class samples are tightly clustered with clear boundaries maintained between different fault categories, thus achieving an overall diagnostic accuracy of 98.66 %. The main confusion occurs when Fault 7 is occasionally misclassified as Fault 1.

In contrast, the feature distributions of CORAL (88.4 %) and AGG (88.8 %) exhibit noticeable overlap among multiple fault categories. CORAL exhibits severe interference between Fault 7 and Fault 1 and also misclassifies a large number of Fault 9 samples (inner-race fatigue pitting with repeated distribution) as Fault 7, indicating its difficulty in distinguishing between inner-race and outer-race damage. Furthermore, its modeling capability for different fault distribution patterns, such as “repeated distribution” and “distributed defects,” remains limited.

Although AGG performs slightly better, it still misclassifies many Fault 9 samples as Fault 6 (multi-point fatigue pitting on both inner and outer races) and Fault 7. Additionally, Fault 13 (inner-race fatigue pitting with double points) is sometimes misclassified as Fault 9 (inner-race fatigue pitting with repeated distribution), suggesting insufficient sensitivity to subtle morphological differences between double-point and repeated-distribution pitting patterns.

These observations demonstrate that the PAM, by jointly modeling channel and temporal attention, effectively focuses on fault-sensitive regions that are invariant to operating condition variations. Meanwhile, the LLU enhances the perception of transient fault details such as indentation morphology, the number of pitting defects, and distribution patterns through local linear modeling. The synergy between these two modules enables the model to accurately distinguish between different fault categories even under the unseen operating conditions, significantly outperforming existing DG approaches.

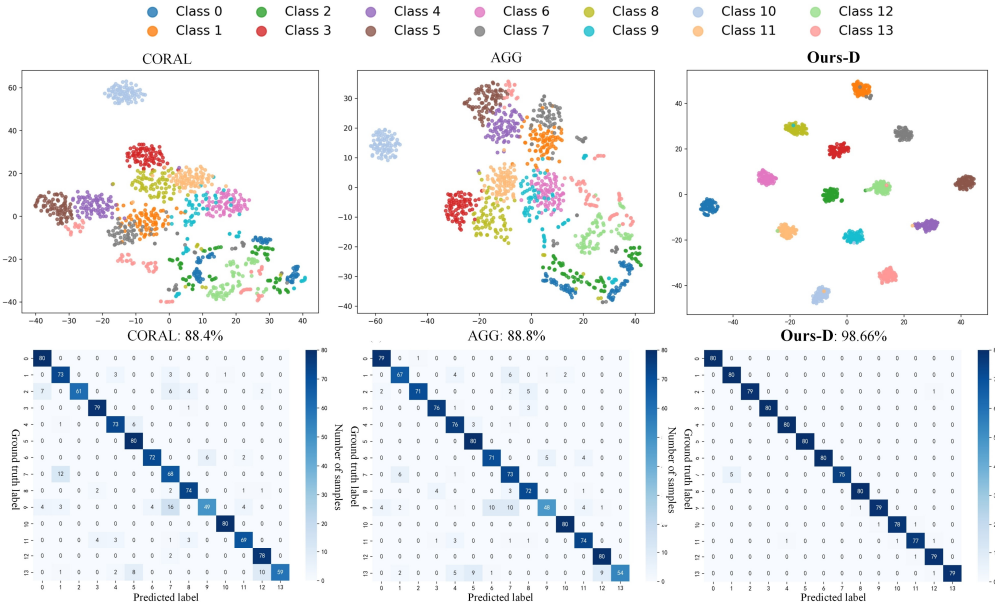


Fig. 3. t-SNE feature visualization and the corresponding confusion matrices for Task T3 on the PU dataset

4.3.3. Hierarchical ablation analysis

The ablation results presented in Tables 6-8 systematically validate the effectiveness and collaborative mechanism of each module on the PU, PHM09, and CWRU datasets. On the PU dataset, the diagnostic performance improves steadily as modules are progressively introduced (72.58 % → 77.15 % → 79.17 % → 80.15 %), indicating that the Transformer, PAM, and LLU components all contribute positively to cross-condition fault diagnosis. On the PHM09 dataset, the performance of Ours-B decreases significantly compared with Ours-A (90.71 % → 82.91 %). This result can be attributed to the lack of strong local inductive bias inherent in the standard Transformer architecture. When processing complex gearbox composite fault signals, the Transformer struggles to capture critical transient structures and becomes more susceptible to

condition-related noise, which ultimately leads to degraded generalization capability under cross-condition scenarios.

After introducing the PAM (Ours-C), the diagnostic accuracy increases again to 86.81 %, demonstrating that the PAM effectively suppresses irrelevant condition disturbances by jointly modeling channel and temporal attention, thereby strengthening the response to fault-sensitive frequency bands and key temporal segments. When the LLU is further embedded (Ours-D), the performance increases significantly to 93.96 %, even surpassing the original CNN-only model (Ours-A). This result indicates that the LLU explicitly enhances the model's local modeling capability for transient impact components such as gear tooth breakage and bearing pitting faults, effectively compensating for the lack of local perception in self-attention mechanisms. This ablation chain clearly demonstrates that the proposed PAM and LLU are not simply stacked modules. Instead, they collaboratively enhance the Transformer representation capability from two complementary perspectives: input feature focusing and local structural modeling.

For the single-source-domain diagnostic tasks in the CWRU dataset, Ours-B improves the performance of Ours-A by 3.48 %, while the final model Ours-D reaches 98.17 %, with standard deviations generally below 1 %, indicating strong model stability under single-source DG scenarios. Overall, the results suggest that Transformers do not always outperform CNNs in DG fault diagnosis tasks. Their full potential can only be realized when combined with the cooperative enhancement effects provided by PAM and LLU.

4.3.4. Sensitivity analysis of key hyperparameters

To investigate the influence of local modeling scales in PAM and LLU on diagnostic performance, a total of 20 different parameter combinations were tested on Task T0 of the PU dataset, covering temporal kernel sizes in the PAM $k_t \in \{3,5,7,9\}$ and depthwise kernel sizes in LLU $k_l \in \{3,5,7,9,11\}$. The average diagnostic accuracies are shown in Fig. 4.

The results indicate that $k_t = 7$ and $k_l = 11$ achieve the optimal diagnostic performance for the current task, reaching an accuracy of 57.16 %. Compared with the kernel size in PAM, the kernel size in LLU has a more pronounced impact on model performance. When k_t is fixed, the diagnostic accuracy generally increases with larger k_l, suggesting that larger LLU kernels are beneficial for capturing long-period modulation envelope components in bearing fault signals. On the other hand, the PAM exhibits an optimal range of kernel sizes. Small kernels can precisely localize transient impacts but are more susceptible to noise interference, whereas excessively large kernels tend to oversmooth temporal attention weights, weakening the model's ability to respond to critical impulses.

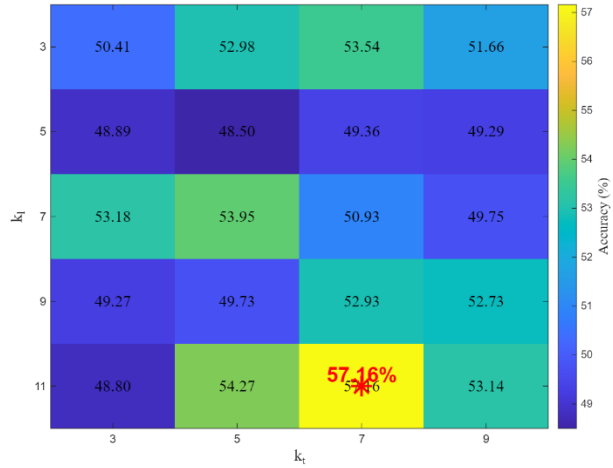


Fig. 4. Impact of local modeling scales in the PAM and LLU on diagnosis performance (T0 of PU dataset)

The optimal combination $k_t = 7$ and $k_l = 11$ reflects the collaborative effect between PAM and LLU. This configuration maintains high sensitivity to transient fault features while enhancing the perception of fault evolution trends. In practical applications, the convolution kernel sizes of the PAM and LLU should be selected according to the physical characteristics of rotating machinery fault signals, such as impact duration and modulation frequency. Although the current study adopts $k_t = 7$ and $k_l = 7$ (accuracy 50.93 %) to balance computational efficiency and robustness, further performance improvements may still be achieved through appropriate kernel size tuning.

4.3.5. Model complexity and inference time analysis

To evaluate the engineering feasibility of the proposed model, its number of parameters (Params), floating-point operations (FLOPs), and average single-sample inference time were evaluated. Taking Task T0 of the PU dataset as an example, the proposed model contains only 0.21M parameters and requires 34.53M FLOPs. When deployed on an NVIDIA GeForce RTX 4060 Laptop GPU, the average inference time per sample is 2.409 milliseconds. These results indicate that the proposed method is a lightweight deep learning model, and its inference efficiency can meet the latency and throughput requirements of industrial fault diagnosis applications in typical offline batch processing and quasi-real-time monitoring scenarios.

4.3.6. Noise robustness analysis under different SNR levels

To further evaluate the robustness of the proposed method in noisy industrial environments, a signal-to-noise ratio (SNR) sensitivity analysis was conducted by introducing additive Gaussian noise into the vibration signals at different SNR levels.

Gaussian noise was added to the raw vibration signals of the PU dataset under five SNR levels, namely Clean, 20 dB, 15 dB, 10 dB, and 5 dB. The experiments were conducted on all four cross-condition diagnostic tasks (T0-T3), and the average diagnostic accuracy under different SNR levels is summarized in Table 9.

As shown in Table 9, the proposed method maintains relatively stable diagnostic performance under moderate noise conditions. Specifically, when the SNR decreases from Clean to 15 dB, the average diagnostic accuracy decreases only slightly from 78.46 % to 76.61 %, indicating that the learned feature representations exhibit strong tolerance to moderate noise interference. When the SNR further decreases to 10 dB, the accuracy shows a noticeable degradation (70.50 %), suggesting that increasing SNR levels begin to distort transient fault features and reduce the effectiveness of feature extraction. Under severe noise conditions (5 dB), the average accuracy drops significantly to 44.16 %, indicating that excessive noise severely affects the discriminative characteristics of fault signals and leads to substantial performance degradation.

Overall, the results confirm that the proposed method demonstrates satisfactory robustness under moderate noise conditions, while its performance degrades as noise intensity increases. These findings provide additional evidence of the practical reliability of the proposed method in noisy industrial environments.

Table 9. Diagnostic results (%) under different SNR levels on PU dataset

Task	Clean	20 dB	15 dB	10 dB	5 dB
T0	50.93±5.86	47.83±6.18	45.95±6.51	31.30±7.62	17.82±2.29
T1	96.66±0.15	96.46±0.12	96.45±0.22	94.43±1.02	50.96±3.96
T2	67.57±3.26	67.38±3.40	66.20±3.54	62.23±3.68	47.48±6.43
T3	98.66±0.42	98.29±0.35	97.82±0.42	94.05±1.77	60.36±10.01
Average	78.46	77.49	76.61	70.50	44.16

5. Conclusions

This paper addresses the performance degradation problem of fault diagnosis models under cross-condition scenarios by proposing a lightweight CNN-Transformer hybrid method integrating multi-scale CNN, PAM, and a Transformer encoder with an embedded LLU. The multi-scale CNN module captures transient fault features across different temporal scales, while the PAM enhances discriminative feature learning by jointly modeling channel and temporal dependencies. Furthermore, the LLU introduces local inductive bias into the Transformer to better capture transient impact features and complement global self-attention.

Extensive experiments on three widely used rotating machinery datasets (PU, PHM09, and CWRU) validate the effectiveness. The results show that the proposed method achieves superior performance across multiple domain generalization tasks and outperforms several representative baseline methods. Visualization and ablation studies further confirm that PAM and LLU jointly improve the robustness and domain-invariant representation capability of the model. Noise robustness experiments under different SNR levels further validate the reliability of the proposed method under noisy conditions.

Future work will focus on extending the proposed method to more complex industrial scenarios, including cross-machine diagnosis, multi-sensor fusion, and partially labeled domain generalization settings.

Acknowledgements

This work was financially supported by the Shaanxi Provincial Natural Science Foundation of China (Grant No. 2024JC-YBQN-0021).

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Author contributions

E Cai: conceptualization, investigation, methodology, writing-original draft, writing-review and editing, validation, formal analysis, data curation. Yangyang Li: conceptualization, supervision, project administration, formal analysis, data curation, funding acquisition.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] C. Zhao, E. Zio, and W. Shen, "Domain generalization for cross-domain fault diagnosis: An application-oriented perspective and a benchmark study," *Reliability Engineering and System Safety*, Vol. 245, p. 109964, 2024, https://doi.org/10.1016/j.ress.2024.109964

[2] Y. Xiao, H. Shao, S. Yan, J. Wang, Y. Peng, and B. Liu, "Domain generalization for rotating machinery fault diagnosis: A survey," *Advanced Engineering Informatics*, Vol. 64, p. 103063, Mar. 2025, https://doi.org/10.1016/j.aei.2024.103063

[3] J. Jiao, M. Zhao, J. Lin, and K. Liang, "A comprehensive review on convolutional neural network in machine fault diagnosis," *Neurocomputing*, Vol. 417, pp. 36–63, Dec. 2020, https://doi.org/10.1016/j.neucom.2020.07.088

[4] D. Neupane, M. R. Bouadjenek, R. Dazeley, and S. Aryal, "Data-driven machinery fault diagnosis: A comprehensive review," *Neurocomputing*, Vol. 627, p. 129588, Apr. 2025, https://doi.org/10.1016/j.neucom.2025.129588

[5] J. Hou, X. Lu, Y. Zhong, W. He, D. Zhao, and F. Zhou, "A comprehensive review of mechanical fault diagnosis methods based on convolutional neural network," *Journal of Vibroengineering*, Vol. 26, No. 1, pp. 44–65, Feb. 2024, https://doi.org/10.21595/jve.2023.23391

[6] W. Xu, Y. Si, A. Lei, L. Kong, and W. Guo, "A multi-scale adaptive transformer with feature enhancement for fault diagnosis of rolling bearings under imbalanced small-sample and cross-condition scenarios," *Structural Health Monitoring*, p. 14759217251371293, Nov. 2025, https://doi.org/10.1177/14759217251371293

[7] B. Lu, Y. Zhang, Z. Liu, H. Wei, and Q. Sun, "A novel sample selection approach based universal unsupervised domain adaptation for fault diagnosis of rotating machinery," *Reliability Engineering and System Safety*, Vol. 240, p. 109618, 2023, https://doi.org/10.1016/j.ress.2023.109618

[8] R. Wang, W. Huang, J. Wang, C. Shen, and Z. Zhu, "Multisource domain feature adaptation network for bearing fault diagnosis under time-varying working conditions," *IEEE Transactions on Instrumentation and Measurement*, Vol. 71, pp. 1–10, Jan. 2022, https://doi.org/10.1109/tim.2022.3168903

[9] Y. He and W. Shen, "MSiT: A cross-machine fault diagnosis model for machine-level CNC spindle motors," *IEEE Transactions on Reliability*, Vol. 73, No. 1, pp. 792–802, Mar. 2024, https://doi.org/10.1109/tr.2023.3322417

[10] W. Liu, Z. Zhang, J. Zhang, H. Huang, G. Zhang, and M. Peng, "A novel fault diagnosis method of rolling bearings combining convolutional neural network and transformer," *Electronics*, Vol. 12, No. 8, p. 1838, Apr. 2023, https://doi.org/10.3390/electronics12081838

[11] Y. Keshun, L. Zengwei, C. Ronghua, and G. Yingkui, "A novel rolling bearing fault diagnosis method based on time-series fusion transformer with interpretability analysis," *Nondestructive Testing and Evaluation*, pp. 1–27, Nov. 2024, https://doi.org/10.1080/10589759.2024.2425813

[12] Z. Lu, L. Liang, J. Zhu, W. Zou, and L. Mao, "Rotating machinery fault diagnosis under multiple working conditions via a time-series transformer enhanced by convolutional neural network," *IEEE Transactions on Instrumentation and Measurement*, Vol. 72, pp. 1–11, Jan. 2023, https://doi.org/10.1109/tim.2023.3318707

[13] Y. Yang, J. Yin, H. Zheng, Y. Li, M. Xu, and Y. Chen, "Learn generalization feature via convolutional neural network: A fault diagnosis scheme toward unseen operating conditions," *IEEE Access*, Vol. 8, pp. 91103–91115, Jan. 2020, https://doi.org/10.1109/access.2020.2994310

[14] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5400–5409, 2018, https://doi.org/10.1109/cvpr.2018.00566

[15] Z. Hua, J. Shi, and P. Dumond, "Domain-invariant feature exploration for intelligent fault diagnosis under unseen and time-varying working conditions," *Mechanical Systems and Signal Processing*, Vol. 224, p. 112193, Feb. 2025, https://doi.org/10.1016/j.ymssp.2024.112193

[16] C. Zhao and W. Shen, "A domain generalization network combing invariance and specificity towards real-time intelligent fault diagnosis," *Mechanical Systems and Signal Processing*, Vol. 173, p. 108990, 2022, https://doi.org/10.1016/j.ymssp.2022.108990

[17] T. Han, Y.-F. Li, and M. Qian, "A hybrid generalization network for intelligent fault diagnosis of rotating machinery under unseen working conditions," *IEEE Transactions on Instrumentation and Measurement*, Vol. 70, pp. 1–11, Jan. 2021, https://doi.org/10.1109/tim.2021.3088489

[18] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Lecture notes in computer science*, Cham: Springer International Publishing, 2018, pp. 3–19, https://doi.org/10.1007/978-3-030-01234-2_1

[19] Y. Sun, H. Tao, and V. Stojanovic, "End-to-end multi-scale residual network with parallel attention mechanism for fault diagnosis under noise and small samples," *ISA Transactions*, Vol. 157, pp. 419–433, Feb. 2025, https://doi.org/10.1016/j.isatra.2024.12.023

[20] M. Hakim, A. A. B. Omran, A. N. Ahmed, M. Al-Waily, and A. Abdellatif, "A systematic review of rolling bearing fault diagnoses based on deep learning and transfer learning: Taxonomy, overview, application, open challenges, weaknesses and recommendations," *Ain Shams Engineering Journal*, Vol. 14, No. 4, p. 101945, Apr. 2023, https://doi.org/10.1016/j.asej.2022.101945

[21] G. Jiang, H. He, J. Yan, and P. Xie, "Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox," *IEEE Transactions on Industrial Electronics*, Vol. 66, No. 4, pp. 3196–3207, Apr. 2019, https://doi.org/10.1109/tie.2018.2844805

[22] A. Vaswani et al., "Attention is all you need," in *31st Conference on Neural Information Processing Systems*, Vol. 30, 2025, https://doi.org/10.65215/nxvz2v36

[23] C. Weng, B. Lu, Q. Gu, and X. Zhao, "A novel multisensor fusion transformer and its application into rotating machinery fault diagnosis," *IEEE Transactions on Instrumentation and Measurement*, Vol. 72, pp. 1–12, Jan. 2023, https://doi.org/10.1109/tim.2023.3244822

[24] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, No. 4, pp. 1–20, Jan. 2022, https://doi.org/10.1109/tpami.2022.3195549

[25] W. A. Smith and R. B. Randall, "Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study," *Mechanical Systems and Signal Processing*, Vol. 64-65, pp. 100–131, Dec. 2015, https://doi.org/10.1016/j.ymssp.2015.04.021

[26] C. Lessmeier, J. K. Kimotho, D. Zimmer, and W. Sextro, "Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification," *European Conference*, Vol. 3, No. 1, 2016, https://doi.org/10.36001/phme.2016.v3i1.1577

[27] "PHM 2009 data challenge," PHM Society, 2019.

[28] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: maximizing for domain invariance," *arXiv:1412.3474*, 2014, https://doi.org/10.48550/arxiv.1412.3474

[29] Z. Zhao et al., "Applications of unsupervised deep transfer learning to intelligent fault diagnosis: A survey and comparative study," *IEEE Transactions on Instrumentation and Measurement*, Vol. 70, pp. 1–28, Jan. 2021, https://doi.org/10.1109/tim.2021.3116309

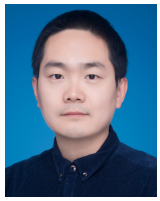
[30] B. Shen, M. Zhang, L. Yao, and Z. Song, "Novel triplet loss-based domain generalization network for bearing fault diagnosis with unseen load condition," *Processes*, Vol. 12, No. 5, p. 882, Apr. 2024, https://doi.org/10.3390/pr12050882

[31] Y. Wang, H. Li, and A. C. Kot, "Heterogeneous domain generalization via domain mixup," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3622–3626, 2020, https://doi.org/10.1109/icassp40776.2020.9053273

[32] D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales, "Learning to generalize: Meta-learning for domain generalization," *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, No. 1, Apr. 2018, https://doi.org/10.1609/aaai.v32i1.11596

[33] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain adaptive ensemble learning," *IEEE Transactions on Image Processing*, Vol. 30, pp. 8008–8018, Jan. 2021, https://doi.org/10.1109/tip.2021.3112012

E Cai received bachelor's degree in mechanical engineering from Northwestern Polytechnical University, Xi'an, China, in 2002, and master's degree in mechanical and Electronic Engineering from Northwestern Polytechnical University, Xi'an, China, in 2005. He is currently researching signal acquisition and processing, as well as the application of artificial intelligence in fault diagnosis.



Yangyang Li received bachelor's degree in transportation engineering from Chang'an University, Xi'an, China, in 2008 and his Ph.D. degree in new energy vehicle engineering from Chang'an University, Xi'an, China, in 2018. He is currently researching the clean and efficient development of new energy systems for automotive power equipment and the fault diagnosis of related devices.