Scan-Net: few-shot diagnosis of hydropower auxiliary bearings via Siamese mutual learning

Zhanfeng Chang[1], Ting He[2], Liping Zhu[3], Hailong Zhang[4], Hongyan Zhang[5]
[1, 3]Xiangjiaba Hydropower Plant, Three Gorges Jinsha River Chuanyun Hydropower Development Co., Ltd., Yibin 644612, Sichuan, China
[2, 4, 5]Beijing IWHR Technology Co., Ltd.,
No. 3 Yuyuantan South Road, Haidian District, Beijing 100038, China
[2]Corresponding author
E-mail: [1]*chang_zhanfeng@ctg.com.cn*, [2]*he_ting2017@126.com*, [3]*zhu_liping@ctg.com.cn*, [4]*zhanghl0703@163.com*, [5]*1196815398@qq.com*

Received 4 December 2025; accepted 22 April 2026; published online 22 July 2026
DOI https://doi.org/10.21595/jve.2026.25895

Check for updates

Abstract. Reliable operation of auxiliary equipment is critical for hydropower stations. However, data-driven diagnosis faces the "cold start" challenge due to fault sample scarcity in high-maintenance environments. We propose a few-shot diagnostic model, Siamese Cross-Attention Network (Scan-Net), combined with a transfer learning strategy. To extract discriminative features from limited data, we utilize Multi-Scale Synchrosqueezed Wavelet Transform (MSWT) for physically consistent time-frequency representations. Unlike simple concatenation, we design a dual-stream Siamese network with a bidirectional Cross-Attention mechanism that enables explicit inter-sensor feature interaction. We introduce a Deep Mutual Learning (DML) strategy with symmetric KL divergence constraints to align prediction distributions between dual branches, serving as self-supervised regularization to prevent overfitting. We establish a transfer pathway from public datasets to field equipment. Experiments show that Scan-Net achieves 96.50 % accuracy on the CWRU dataset under the 10-shot setting, and 94.43 % average accuracy in cross-load transfer. Pilot deployment at a large-scale hydropower station provides preliminary validation, with the system contributing to a reduction in routine inspection workload

Keywords: bearing fault diagnosis, few-shot learning, deep mutual learning, Siamese network, vibration signal analysis, transfer learning.

1. Introduction

Large-scale hydro-turbine generator units act as the "ballast" of power systems. Their stable operation is crucial for grid regulation and power quality [1]. For major hydropower facilities, unplanned shutdowns cause significant economic losses and security risks. While main turbine units possess comprehensive monitoring systems, auxiliary equipment (e.g., governor oil pumps) often remains unmonitored. These induction motor-driven machines are critical; bearing faults in hydropower rotating equipment can cause severe operational interruptions and substantial economic losses [2].

We propose a diagnostic framework to address these challenges (Fig. 1). Conventional approaches (Fig. 1(a)) typically rely on a single sensor, limiting robustness and generalization in few-shot scenarios. In contrast, our Siamese mutual learning framework (Fig. 1(b)) leverages dual-sensor data (Drive End and Fan End). By employing Deep Mutual Learning (DML), the model learns consistent high-level representations from heterogeneous views. This strategy forces dual branches to extract intrinsic fault features robust to noise and sensor placement, enhancing generalization in few-shot tasks.

One core objective of building "intelligent hydropower stations" is to achieve predictive maintenance. In recent years, deep learning has made significant progress in bearing fault diagnosis. Singh et al. [3] proposed a graph-based rotating machinery fault diagnosis framework that improves diagnostic accuracy through structured features. Chou et al. [4] combined Continuous Wavelet Transform with the YOLO model for efficient bearing fault detection. Liu et al. [5] explored multi-scale quaternion CNN with cross self-attention feature fusion methods.

However, engineering deployment faces the paradox of "data scarcity caused by high reliability". Target industrial facilities often implement strict preventive maintenance. Equipment rarely fails, resulting in abundant normal operation data but scarce labeled fault samples. Traditional deep learning models (e.g., CNN [6], LSTM) require abundant balanced samples, making them difficult to train in few-shot field environments [7]. Waiting to accumulate sufficient fault data involves unacceptable risks.

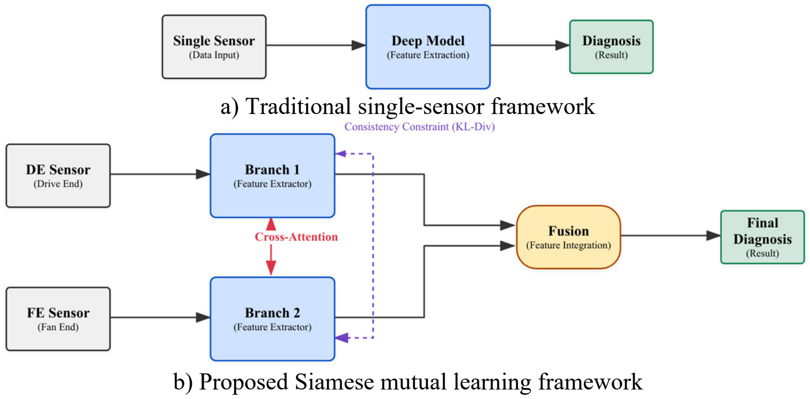


Fig. 1. Comparison of diagnostic approaches: a) Traditional single-sensor framework with limited robustness in few-shot scenarios. (b) Proposed Siamese mutual learning framework using dual sensors (FE and DE) with Cross-Attention and KL divergence consistency constraints, combined with transfer learning for field deployment

We adopt Transfer Learning to bridge the gap between laboratory research and industrial application. The method's performance on the CWRU dataset validates the effectiveness of DML-learned features. By learning universal, mechanism-consistent fault features from the source domain, the model transfers effectively to the target domain (the specific hydropower station). This "Source → Target" pathway overcomes data scarcity, resolving the "cold start" problem.

Nevertheless, direct transfer faces the challenge of "Domain Shift". Hydropower station field environments are complex, with strong electromagnetic interference and flow-induced vibrations. Operating conditions fluctuate frequently with main servomotor actions. To improve model robustness under complex noise and varying operating conditions, this paper proposes a diagnostic framework based on Siamese Cross-Attention Networks (Scan-Net).

The method integrates three key technologies: (1) Multi-Scale Synchrosqueezed Wavelet Transform (MSWT) extracts weak fault features submerged in noise; (2) Siamese Network Architecture reduces parameters through weight sharing and symmetric dual-branch design while enhancing unified representation capability for dual-sensor data; (3) Cross-Attention and Consistency Constraint mechanisms – the former exploits complementarity of dual-channel sensors to automatically suppress interfered channels, the latter uses KL divergence minimization to align prediction distributions between two branches, forming a mutual learning mechanism. This soft constraint extracts inherent consistency from dual-view data, improving generalization in few-shot scenarios.

Despite these advances, most existing few-shot fault diagnosis methods rely on a single sensor, limiting their ability to capture complementary fault information. Multi-sensor approaches typically employ simple concatenation or late fusion, lacking explicit inter-sensor consistency regularization. Few studies simultaneously address low-label learning, cross-load transfer, and field deployment. These gaps motivate the present work.

The main contributions of this paper are as follows:

1) Method: We design a dual-stream Siamese network where two branches interact through bidirectional cross-attention rather than post-hoc concatenation, and are jointly regularized by a symmetric KL divergence consistency loss. Ablation experiments confirm that this consistency loss alone contributes 4.4 % accuracy improvement under the 10-shot setting, the largest single-component gain.

2) Experiments: We validate Scan-Net across three complementary protocols – 10-shot learning (96.50 %), cross-load transfer (94.43 % average), and noise robustness (86.5 % at –4 dB) – providing a more comprehensive evaluation than standard-split accuracy alone.

3) Application: We establish a "CWRU pre-training + field transfer" pathway and pilot-deploy the system at a large-scale hydropower station, demonstrating feasibility of intelligent monitoring for auxiliary equipment under fault-sample scarcity.

2. Related work

2.1. Few-shot fault diagnosis

Few-Shot Learning (FSL) aims to enable models to learn and generalize from very few samples, which aligns more closely with human learning capabilities. In the field of industrial fault diagnosis, FSL has received widespread attention due to fault data scarcity. Mainstream methods can be divided into three categories:

1) Metric learning-based methods: These methods focus on learning a metric space where samples of the same class are close and samples of different classes are far apart. Representative works include Siamese Networks [8], Prototypical Networks [9], and Relation Networks [10]. They perform classification by minimizing or maximizing distances between sample pairs, especially suitable for handling open-set recognition problems with variable numbers of categories. The Siamese network structure in this paper draws on metric learning ideas.

2) Optimization-based methods: These methods aim to adjust the model's optimization process so it can quickly adapt to new tasks with few samples. Model-Agnostic Meta-Learning (MAML) [11] is a typical representative, which finds sensitive initial parameters through meta-learning, enabling the model to achieve good performance on new tasks with just a few gradient descent steps.

3) Data augmentation-based methods: These methods expand datasets by generating additional training samples to alleviate data scarcity. For example, Generative Adversarial Networks (GANs) [12] are used to generate virtual samples with distributions similar to real fault signals. However, the diversity and fidelity of generated samples remain significant challenges.

2.2. Multi-sensor fusion

In fault diagnosis, fusing information from multiple sensors typically yields more reliable and comprehensive results than a single sensor. Based on the fusion stage, it can be divided into three categories:

1) Data-level fusion: Directly concatenates raw signals or time-frequency maps at the input end. This approach is simple and direct but may lead to performance degradation due to registration errors and noise differences between sensors.

2) Feature-level fusion: This is the most commonly used strategy. Features are extracted from each sensor's data, then these feature vectors are concatenated, summed, or fused through more complex modules. Attention mechanisms [13] show great potential at this stage, adaptively learning importance weights for different sensors or feature channels. However, most existing work employs self-attention or channel attention, with insufficient exploration of inter-sensor interactive information.

3) Decision-level fusion: Votes or weighted averages independent diagnostic results from each

sensor. This approach (also called late fusion) has good robustness but may lose fine-grained correlations at the feature level since information is already highly abstracted in early stages.

The cross-attention mechanism proposed in this paper belongs to feature-level fusion but goes beyond simple feature concatenation or weighting, aiming to explore mutual dependencies between two sensor feature maps. Prior attention-based fusion methods mainly perform intra-feature weighting; in contrast, our method targets explicit inter-sensor interaction through bidirectional cross-attention and agreement regularization via symmetric KL divergence.

2.3. Deep mutual learning

Knowledge Distillation [14] typically refers to transferring knowledge from a large "teacher" network to a small "student" network. Deep Mutual Learning (DML) [15] extends this concept by having a group (usually two) of networks with the same or similar structures learn from each other and supervise each other during training.

In DML, each network plays a dual role: both learner and teacher. Training objectives include not only fitting true labels (supervised loss) but also minimizing differences between output probability distributions of the two networks, typically measured using KL divergence. The core idea is that different networks may learn different "knowledge" or "perspectives" from data; through mutual imitation, they can exchange this knowledge and collectively achieve better performance than training independently. DML has been proven to effectively improve model generalization in fields such as image recognition.

This paper applies DML concepts to dual-sensor fault diagnosis, treating two Siamese branches processing different sensor data as a mutually learning entity. Through consistency loss (i.e., KL divergence), we force the two branches to reach consistent judgments on the same fault event, which is not only an effective regularization means but also promotes both branches to learn more essential fault features that are insensitive to sensor position, noise, and other variations.

In summary, existing few-shot fault diagnosis methods predominantly rely on single-sensor input and employ concatenation or late fusion when multiple sensors are available, without explicit inter-sensor consistency regularization. The proposed Scan-Net differs in three respects: it fuses dual-sensor features through bidirectional cross-attention rather than concatenation, enforces predictive agreement via a symmetric KL consistency loss, and is evaluated under 10-shot learning, cross-load transfer, and field-oriented deployment within a single framework.

3. Methods

3.1. Proposed method

3.1.1. Overall framework

The proposed Siamese Cross-Attention Network (Scan-Net) is designed to address the challenge of fault diagnosis for auxiliary equipment (e.g., governor oil pumps) under few-shot conditions. As illustrated in Fig. 2, the framework employs a dual-stream Siamese architecture that processes vibration signals from the Drive End (DE) and Fan End (FE) independently. Three design choices distinguish this method from prior work: (1) dual-sensor features interact through bidirectional cross-attention rather than post-hoc concatenation; (2) two branches are jointly regularized by a symmetric KL consistency loss rather than trained independently; (3) the method is validated across 10-shot learning, cross-load transfer, and a field pilot, rather than reporting accuracy on a standard split alone. The overall process consists of three key stages:

1) Multi-scale time-frequency feature encoding: First, MSWT transforms one-dimensional vibration signals into two-dimensional RGB images containing rich time-frequency details. Subsequently, through a ResNet-18 backbone network integrated with Convolutional Block Attention Module (CBAM), deep feature representations with channel and spatial sensitivity are

extracted from DE and FE branches.

2) Siamese cross-attention fusion: Unlike traditional feature concatenation, we design a cross-attention module that enables dual-branch features to interact in latent space. This module allows the DE branch to dynamically retrieve complementary information from the FE branch (and vice versa), generating enhanced feature vectors that fuse advantages from both sources.

3) Mutual learning and consistency constraints: Beyond the main classification task, we introduce auxiliary classifiers for each branch and constrain consistency of prediction distributions between two branches through KL divergence. This mutual learning mechanism forces the model to extract inherent consistency between dual views in the absence of labels, significantly improving robustness in few-shot scenarios.

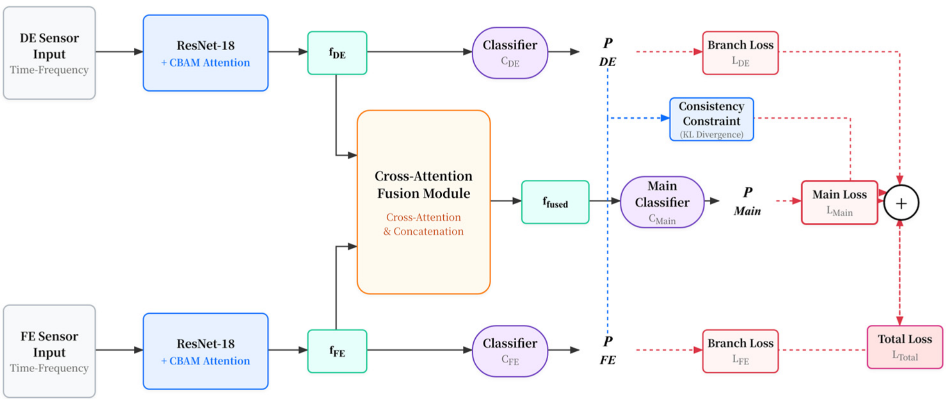


Fig. 2. Overall framework of Scan-Net. The dual-branch Siamese backbone (ResNet-18+CBAM) extracts features from DE and FE sensors with shared weights. Cross-attention fusion enables bidirectional feature interaction (red dashed lines). KL divergence consistency loss (purple dashed lines) aligns branch predictions P_{DE} and P_{FE}. Main classifier C_{Main} and auxiliary classifiers C_{DE}, C_{FE} are jointly optimized via L_{Total}

3.1.2. Dual-branch feature extraction

Input Representation: Multi-Scale Synchrosqueezed Wavelet Transform (MSWT). Raw vibration signals are often submerged in strong background noise, and fault features exhibit non-stationarity. To fully preserve transient impact components and suppress noise, we adopt Multi-Scale Synchrosqueezed Wavelet Transform (MSWT) [16] as a signal preprocessing method. MSWT significantly improves time-frequency resolution by redistributing energy in the time-frequency plane.

For signal $x(t)$, its Continuous Wavelet Transform (CWT) [17] is defined as:

$$W_x(a,b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} x(t) \psi^* \left(\frac{t-b}{a} \right) dt, \tag{1}$$

where $\psi(t)$ is the mother wavelet, a is the scale parameter, b is the time shift.

Synchrosqueezed Transform (SWT) sharpens time-frequency representation by reassigning wavelet coefficients to instantaneous frequency positions:

$$T_x(\omega,b) = \sum_{a_k:|\omega_k(a_k,b)-\omega| \le \frac{\Delta\omega}{2}} W_x(a_k,b) a_k^{-3/2} \Delta a_k, \tag{2}$$

where $\omega_k(a_k,b)$ is the instantaneous frequency estimate.

To capture multi-frequency-band fault features of auxiliary bearings under complex operating conditions (such as high-frequency inner race impacts, mid-frequency outer race periodic components, low-frequency rolling element modulations), we perform transformations at three complementary frequency scales ($s_1 = 0.5$, $s_2 = 1.0$, $s_3 = 2.0$) and map the resulting time-frequency coefficients to three channels of an RGB image:

$$I_{MSWT} = \begin{bmatrix} Normalize(|T_x^{s_1}|) \\ Normalize(|T_x^{s_2}|) \\ Normalize(|T_x^{s_3}|) \end{bmatrix}. \tag{3}$$

This multi-scale representation not only preserves high-frequency impacts of inner race faults but also captures mid-to-low-frequency modulation features of outer race and rolling element faults, providing physically meaningful input representation for deep networks.

Backbone Network: CBAM-Enhanced ResNet-18. Given ResNet-18's [18] excellent performance in feature extraction, we adopt it as the backbone architecture of the Siamese network. To further enhance the model's sensitivity to key fault features, we embed the Convolutional Block Attention Module (CBAM) [19] after each residual block. CBAM includes two sub-modules: Channel Attention and Spatial Attention.

Channel attention module is computed as:

$$M_c(F) = \sigma\left(\mathrm{MLP}\big(\mathrm{AvgPool}(F)\big) + \mathrm{MLP}\big(\mathrm{MaxPool}(F)\big)\right). \tag{4}$$

Spatial attention module is computed as:

$$M_s(F') = \sigma\big(f^{7\times 7}([\mathrm{AvgPool}(F'); \mathrm{MaxPool}(F')])\big), \tag{5}$$

where F is the input feature, F' is the channel-refined feature, σ is the sigmoid activation function, $f^{7\times 7}$ is the convolution operation.

1) Channel attention: Adaptively recalibrates channel weights, emphasizing key frequency bands containing fault pulses while suppressing noise bands.

2) Spatial attention: Focuses on fault impact regions in time-frequency maps, ignoring background regions.

3) Advantages of Siamese Architecture: The two branches share network structure and weight parameters, i.e., $\theta_{DE} = \theta_{FE}$. This design has three advantages: (1) Parameter efficiency: Compared to two independent networks, parameters are reduced by approximately 50 %, lowering overfitting risk; (2) Implicit regularization: Weight sharing forces the model to learn universal fault feature representations insensitive to sensor position, improving cross-sensor generalization; (3) Symmetry guarantee: Symmetric dual-branch design ensures fairness of subsequent mutual learning mechanisms, avoiding dominance by one branch.

3.1.3. Siamese cross-attention fusion

Traditional feature fusion methods, such as concatenation or weighted summation, often overlook the semantic correlations between features from different sensors. To address this limitation, we propose a deep fusion strategy based on Cross-Attention. This mechanism enables the model to dynamically "query" features from one sensor using the features from the other, achieving selective information aggregation that exploits the complementary nature of the data.

Let $\mathbf{f}_{DE}$, $\mathbf{f}_{FE} \in \mathbb{R}^d$ be feature vectors extracted from DE and FE branches respectively. In practice, the attention is computed after reshaping features into multi-head subspaces; for clarity, we present the single-head formulation below. To capture complementary relationships between different sensor features, we design a bidirectional cross-attention mechanism. Query, Key, and Value vectors are generated through linear projection:

$$\mathbf{Q}_i = \mathbf{W}_Q^i \mathbf{f}_i, \quad \mathbf{K}_i = \mathbf{W}_K^i \mathbf{f}_i, \quad \mathbf{V}_i = \mathbf{W}_V^i \mathbf{f}_i, \quad i \in \{DE, FE\}. \tag{6}$$

Using Scaled Dot-Product Attention to compute bidirectional cross features. Taking the DE branch aggregating FE information as an example, the generated cross feature $\mathbf{f}_{DE\leftarrow FE}$ is computed as:

$$\mathbf{f}_{DE\leftarrow FE} = \mathrm{softmax}\left(\frac{\mathbf{Q}_{DE}\mathbf{K}_{FE}^T}{\sqrt{d_k}}\right)\mathbf{V}_{FE}. \tag{7}$$

Similarly, the FE branch aggregating DE information feature $\mathbf{f}_{FE\leftarrow DE}$ is:

$$\mathbf{f}_{FE\leftarrow DE} = \mathrm{softmax}\left(\frac{\mathbf{Q}_{FE}\mathbf{K}_{DE}^T}{\sqrt{d_k}}\right)\mathbf{V}_{DE}. \tag{8}$$

To preserve original branch features and fuse cross information, we introduce Residual Connection and Layer Normalization, then concatenate the two enhanced feature vectors and generate the final fused feature $\mathbf{f}_{fused}$ through a fully connected layer:

$$\begin{aligned}
&\mathbf{f}'_{DE} = \mathrm{LayerNorm}(\mathbf{f}_{DE} + \mathbf{f}_{DE\leftarrow FE}),\\
&\mathbf{f}'_{FE} = \mathrm{LayerNorm}(\mathbf{f}_{FE} + \mathbf{f}_{FE\leftarrow DE}),\\
&\mathbf{f}_{fused} = \mathbf{W}_o[\mathbf{f}'_{DE}; \mathbf{f}'_{FE}] + \mathbf{b}_o.
\end{aligned} \tag{9}$$

This bidirectional fusion mechanism ensures the model considers both sensor views when making decisions. For example, when the DE sensor is severely interfered by noise, the model can automatically assign higher attention to FE features.

3.1.4. Mutual Learning and optimization objective

To mitigate the impact of label scarcity in few-shot scenarios, we introduce a Mutual Learning mechanism. This mechanism serves as a powerful regularization technique. The core idea is to utilize the consistency of dual-branch predictions as a supervisory signal, so that the model can extract structural information from unlabeled data, thereby reducing the reliance on annotated samples.

The overall loss function $\mathcal{L}_{total}$ consists of three parts: main task classification loss, branch auxiliary classification loss, and consistency constraint loss:

$$\mathcal{L}_{total} = \mathcal{L}_{main} + \gamma(\mathcal{L}_{DE} + \mathcal{L}_{FE}) + \beta\mathcal{L}_{consist}, \tag{10}$$

where γ and β are hyperparameters balancing each loss weight.

1) Classification Loss: We adopt standard Cross-Entropy Loss as the classification objective. For a given sample and its true label y, and model predicted probability distribution P, the loss is defined as:

$$\mathcal{L}_{CE}(y, P) = -\sum_{c=1}^{C} y_c \log(P_c). \tag{11}$$

Therefore, main task loss $\mathcal{L}_{main} = \mathcal{L}_{CE}(y, P_{Main})$, auxiliary branch losses $\mathcal{L}_{DE} = \mathcal{L}_{CE}(y, P_{DE})$ and $\mathcal{L}_{FE} = \mathcal{L}_{CE}(y, P_{FE})$. This forces each branch and fusion module to possess accurate discriminative capability.

2) Consistency Loss $\mathcal{L}_{consist}$ (Core of Mutual Learning): We require prediction probability distributions P_{DE} and P_{FE} of two branches on the same sample to be as close as possible. We adopt

symmetrized bidirectional KL divergence (Kullback-Leibler Divergence) as the metric:

$$\mathcal{L}_{consist} = \frac{1}{2}[D_{KL}(P_{DE}||P_{FE}) + D_{KL}(P_{FE}||P_{DE})], \tag{12}$$

where $D_{KL}(P||Q) = \sum P(x)\log\frac{P(x)}{Q(x)}$.

Theoretical Advantages of KL Divergence Consistency Constraint: This mechanism draws on Deep Mutual Learning (DML) [15] ideas but adapts them for dual-sensor diagnosis scenarios. Its advantages are reflected in three levels:

1) Self-supervised regularization: In few-shot scenarios, label signals are sparse and prone to overfitting. KL divergence constraints provide additional soft supervision signals – two branches must reach “consensus” on the same fault. This constraint does not rely on additional labeling but effectively narrows the hypothesis space, preventing the model from memorizing noise.

2) Knowledge complementarity: DE and FE sensors are at different bearing positions, each potentially capturing complementary fault information (e.g., DE sensitive to inner race faults, FE sensitive to outer race faults). Through KL divergence minimization, two branches “teach” each other during training, with high-confidence predictions from one branch influencing the other, achieving knowledge transfer.

3) Robustness improvement: When one sensor is interfered by noise, consistency constraints force the interfered branch’s predictions to align with the clean branch, suppressing noise propagation. Ablation experiments (Table 2) show that removing consistency loss reduces 10-shot accuracy by 4.4 %, indicating the importance of this mechanism in few-shot scenarios.

The KL divergence consistency loss encourages predictive agreement between the two branches, acting as a regularizer that extracts shared structure from dual-view data and improves the model’s generalization boundary.

3.2. Experimental setup

3.2.1. Dataset

This study adopts the internationally recognized Case Western Reserve University (CWRU) bearing dataset [20] for algorithm validation. The dataset includes vibration signals of Normal, Inner Race Fault, Outer Race Fault, and Rolling Element Fault (collectively called 4-class classification) collected under different loads (0-3 hp). We follow standard data processing procedures, segmenting raw signals into 2048-point samples and dividing them 70 %/15 %/15 % for training/validation/testing. All samples are converted to 224×224 RGB time-frequency images through the aforementioned MSWT method.

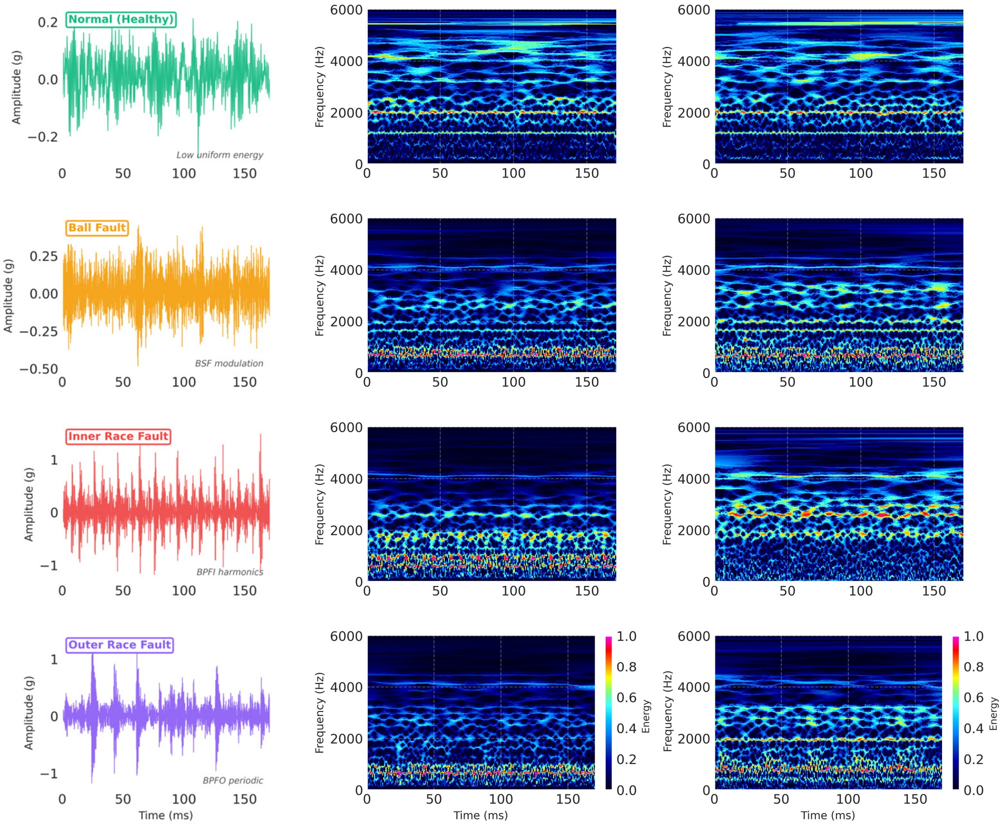
To demonstrate the effectiveness of MSWT, Fig. 3 visualizes the signal processing results. As illustrated, the time-domain signals (Column a) for different fault types often show overlapping characteristics and are difficult to distinguish visually. In contrast, the MSWT time-frequency maps (Columns b and c) reveal distinct visual patterns: Normal samples exhibit low and uniform energy distribution; Inner Race faults show high-frequency periodic impulses; Outer Race faults display mid-frequency components; Ball faults present complex modulations. These discriminative signatures, consistent across both Drive End (DE) and Fan End (FE) sensors, enable effective classification.

3.2.2. Few-shot learning setup

To simulate the severe label-scarcity scenario encountered in industrial practice, we construct a strict 10-shot setting: for each fault category, only 10 labeled samples are randomly selected from the training split as the sole training data, with the remainder discarded. All few-shot results reported in this paper use this setting. Each experiment is repeated 10 times with different random seeds, and the average accuracy is reported to ensure stability and reliability.

3.2.3. Implementation details

All models are implemented in PyTorch framework and trained using a single NVIDIA RTX 3090 GPU. We adopt the Adam optimizer with an initial learning rate of 5×10^{-5} and use cosine annealing strategy for learning rate scheduling. Training batch size is 16, trained for 100 epochs, and early stopping strategy (patience = 20) is adopted to prevent overfitting. In the composite loss function, auxiliary loss weight γ is set to 0.15, consistency loss weight β is set to 0.1 based on experimental tuning.



a) Time-domain signal b) MSWT – drive end c) MSWT – fan end

Fig. 3. Visualization of signal processing results for different health states. The rows correspond to Normal, Ball Fault, Inner Race Fault, and Outer Race Fault conditions. Columns display: a) Time-domain vibration signals; b) MSWT time-frequency maps from Drive End (DE) sensor; c) MSWT time-frequency maps from Fan End (FE) sensor. The MSWT maps reveal distinct texture patterns and frequency energy distributions for each fault type, whereas time-domain signals are less discriminative

4. Results

4.1. Comparison with state-of-the-art methods

The primary focus of this comparison is not full-data saturation accuracy, where most modern methods already exceed 99 %, but rather low-label robustness and cross-condition generalization. To validate Scan-Net performance, we compare it with mainstream deep learning methods on the CWRU dataset, including WDCNN [21], TCNN (ResNet-50) [22], GAN-based semi-supervised methods [3], as well as graph neural network-based method and multi-scale quaternion CNN [5].

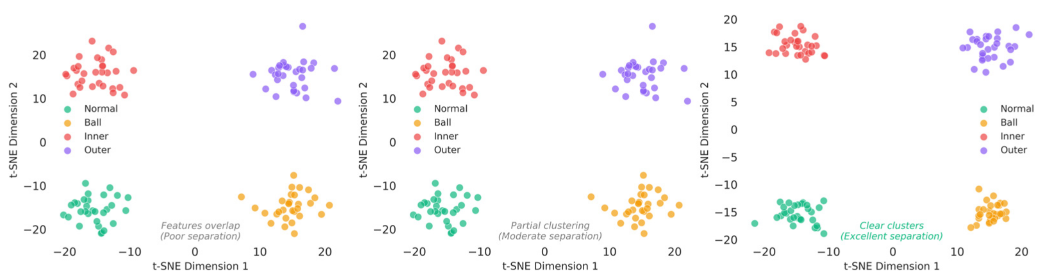
Table 1. Diagnostic accuracy comparison on CWRU dataset (standard split and 10-shot few-shot setting). Bold indicates best performance. Baseline results are cited from original papers; Scan-Net results are averaged over 10 runs

Method (Backbone)	Input Modality	Std. Acc. (%)	10-shot (%)
WDCNN (Wide CNN) [21]	1D Raw Signal	99.60	78.50
CatAAE (Auto-Encoder) [23]	Frequency Features	98.36	85.10
TCNN ResNet-50 [22]	2D Time-Freq Map	99.99	82.40
MQCNN (Quaternion CNN) [5]	Multi-scale Features	99.75	91.20
Graph-Net (GNN) [3]	Structured Graph	99.85	92.50
Scan-Net (Ours) (ResNet-18)	MSWT + Fusion	100.00	96.50

As shown in Table 1, in data-sufficient "standard accuracy" testing, most deep learning methods (such as TCNN and WDCNN) can achieve accuracy above 99 %. Recent research such as Singh et al. [3] and Liu et al. [5] also report excellent performance exceeding 99.5 % under standard settings, indicating that the CWRU standard dataset is relatively easy for modern deep networks, showing "performance saturation" phenomenon.

However, when training samples are reduced to only 10 per class (10-shot), traditional methods' performance shows significant degradation (dropping to around 80 %). In contrast, Scan-Net maintains high accuracy of 96.50 % under 10-shot setting, significantly outperforming the second-best method (Graph-Net, 92.50 %) by 4.0 percentage points. This improvement is achieved through three key technologies: (1) Efficient feature expression of MSWT captures multi-scale physical characteristics of faults; (2) Parameter sharing of Siamese network provides implicit regularization, preventing overfitting on few samples; (3) KL divergence consistency constraint provides additional self-supervised signals, extracting inherent structure in dual-view data. These results indicate the effectiveness of the mutual learning mechanism in label-scarce scenarios.

To further illustrate the discriminative power of learned representations, Fig. 4 presents t-SNE visualizations of the feature embeddings extracted under the 10-shot setting. As shown in Fig. 4(a), raw MSWT features exhibit significant overlap, indicating that low-level features alone are insufficient for separation. The features learned by a single-sensor model (Fig. 4(b)) show improved clustering but still suffer from confusion between similar fault types. In contrast, the features learned by Scan-Net (Fig. 4(c)) exhibit clear class separation with compact intra-class clustering. This visualization provides intuitive evidence that the proposed mutual learning mechanism enables the model to extract more discriminative and robust feature representations.



a) RAW MSWT features   b) Single sensor (DE)   c) Scan-Net (Ours)

Fig. 4. t-SNE visualization of feature distributions under the 10-shot setting: a) raw MSWT features without learning exhibit significant overlap, b) features learned by a single-sensor (DE) model show improved but imperfect separation, c) features learned by the proposed Scan-Net (dual-sensor mutual learning) achieve distinct inter-class separation and compact intra-class clustering, demonstrating superior discriminative power

4.2. Cross-load transfer learning validation

Hydro-turbine units in actual operation frequently need to adjust between different loads (30 %-100 % of rated load), which causes significant statistical drift (Domain Shift) in vibration signal characteristics. Therefore, models must possess cross-condition transfer capability. To validate the robustness of this method, we conduct comprehensive bidirectional transfer learning experiments between four different loads (0HP, 1HP, 2HP, 3HP) of the CWRU dataset. We adopt a "source domain training - target domain testing" protocol, training models on source conditions and directly testing on target conditions without fine-tuning, simulating realistic deployment scenarios.

Table 2 compares Scan-Net with VGG16, ResNet18, MQCNN [5], and Graph-Net [3] on 12 cross-load transfer tasks. For reference, Chou et al. [4] reported an average accuracy of approximately 88.5 % on a related transfer setting.

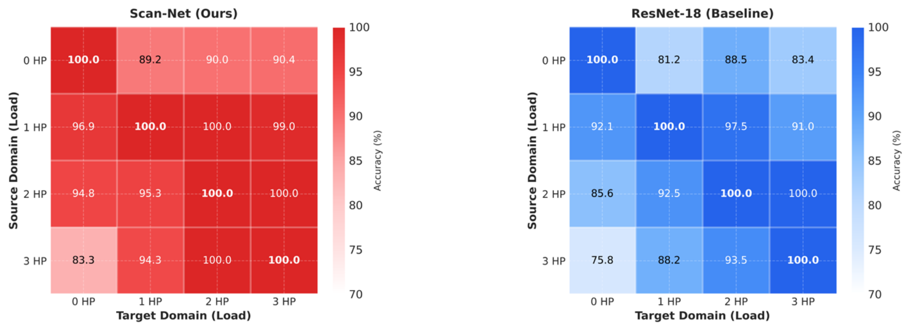
Table 2. Cross-load transfer learning accuracy comparison (%). Literature baselines are reported values from the original papers; direct comparison should be interpreted with caution due to potential differences in experimental protocols

Transfer Task (Year)	VGG16 [24] (2014)	ResNet18 [18] (2016)	Liu et al. [5] (2024)	Singh et al. [3] (2025)	Scan-Net (Ours) (2025)
0→1 HP	72.92	81.20	85.10	86.20	89.17
0→2 HP	80.73	88.50	92.50	93.80	90.02
0→3 HP	71.35	83.40	87.20	88.50	90.38
1→0 HP	83.85	92.10	96.80	97.10	96.88
1→2 HP	92.71	97.50	99.50	99.80	100.00
1→3 HP	73.96	91.00	98.10	98.50	98.96
2→0 HP	68.23	85.60	90.50	91.20	94.79
2→1 HP	91.67	92.50	93.80	94.50	95.31
2→3 HP	100.00	100.00	100.00	100.00	100.00
3→0 HP	77.60	75.80	72.40	75.20	83.33
3→1 HP	90.10	88.20	85.50	88.20	94.27
3→2 HP	82.81	93.50	98.10	98.50	100.00
Average Accuracy	82.16	89.11	91.62	92.62	94.43

Experimental results demonstrate that Scan-Net achieves an average accuracy of 94.43 % across all 12 transfer tasks, outperforming Graph-Net (92.62 %, +1.81 %) and MQCNN (91.62 %, +2.81 %), and substantially surpassing traditional ResNet18 (89.11 %, +5.32 %) and VGG16 (82.16 %, +12.27 %).

To provide a more intuitive comparison of transfer performance patterns, Fig. 5 presents the accuracy matrices of Scan-Net and ResNet-18 as heatmaps. Several important observations emerge from this visualization: (1) The diagonal elements (same-domain testing) achieve near-perfect accuracy for both methods, confirming the models' capability under standard settings; (2) Off-diagonal elements reveal the true transfer difficulty, where Scan-Net consistently shows warmer colors (higher accuracy) across all transfer directions; (3) The "3→0" and "3→1" tasks (high-load to low-load transfer) exhibit the largest performance gap between methods, with Scan-Net demonstrating superior robustness to domain shift; (4) The overall color uniformity of Scan-Net's heatmap indicates more stable cross-domain generalization compared to the patchy pattern of ResNet-18.

In the most challenging "high-load to low-load transfer" task (e.g., 3→1 HP), Scan-Net maintains 94.27 % accuracy, while traditional ResNet18 achieves 88.20 %, an improvement of 6.07 percentage points. This advantage stems from three mechanisms: (1) MSWT multi-scale features possess strong operating condition invariance, extracting essential fault features that remain stable across different loads; (2) Symmetric design of Siamese network learns universal representations insensitive to sensor position through weight sharing; (3) KL divergence consistency constraint forces two branches to reach consensus, further suppressing noise introduced by condition changes. These results demonstrate the effectiveness of the mutual learning architecture in cross-domain transfer scenarios.



a) b)

Fig. 5. Heatmap of cross-load transfer accuracy: a) Scan-Net; b) ResNet-18 baseline. Each cell shows accuracy when training on the source load (row) and testing on the target load (column). Scan-Net shows more uniform and higher transfer performance, especially in challenging high-to-low load scenarios

4.3. Ablation study and parameter analysis

To investigate the specific contributions of each key component in Scan-Net (multi-sensor fusion, MSWT, CBAM, consistency loss), we conduct a series of ablation experiments under the most challenging 10-shot setting to rigorously test the model's robustness.

4.3.1. Contribution analysis of key components

Table 3 shows model performance after progressively removing each module.

1) Importance of MSWT (a vs b): Compared to directly using one-dimensional raw signals (Baseline), introducing MSWT time-frequency maps improves accuracy from 76.20 % to 89.50 % (+13.3 %). This result indicates that under extremely few samples, two-dimensional time-frequency textures provided by MSWT contain more distinguishable physical features than raw waveforms, reducing the learning difficulty for deep networks.

Table 3. Impact of different components on 10-shot fault diagnosis performance

Model variant	Dual Sen.	MSWT	CBAM	Mut. Learn.	Acc. (%)	Δ (%)
(a) Baseline (ResNet18-1D)	×	×	×	×	76.20	–20.30
(b) w/ MSWT (Single Sensor)	×			×	89.50	–7.00
(c) w/ Dual Sensors (Concat)				×	93.40	–3.10
(d) w/o CBAM			×		94.80	–1.70
(e) w/o Consistency Loss				×	92.10	–4.40
(f) Scan-Net (Full)					96.50	–

2) Gain from dual-sensor fusion (b vs c): Expanding from single sensor (DE end) to dual sensors (DE+FE) with simple feature concatenation (Concat) further improves accuracy by 3.9 %. This validates that information complementarity from different measurement points can effectively eliminate diagnostic blind spots inherent to single perspectives.

3) Critical role of cross-attention and mutual learning mechanisms (c/e vs f):

– Cross-attention mechanism (comparing c and f) brings an improvement (+3.1 %). This indicates that simple concatenation cannot fully exploit semantic correlations between sensors. Cross-Attention achieves deep feature interaction through Query-Key-Value mechanisms, allowing one branch to dynamically retrieve complementary information from the other.

– KL divergence consistency loss (comparing e and f) contributes 4.4 % performance improvement, the largest single contribution among all improvements. This finding confirms the effectiveness of the mutual learning mechanism. When lacking abundant label supervision, the constraint forces dual-branch prediction distribution alignment (i.e., $P_{DE} \approx P_{FE}$), providing powerful regularization that extracts inherent consistency structure in dual-view data and prevents the model from overfitting to noise.

– Fine-grained focus of CBAM (d vs f): Removing CBAM results in a slight performance decrease of 1.7 %. This indicates that channel and spatial attention mechanisms help locate key fault impact frequency bands in complex time-frequency maps. However, their contribution is relatively smaller than that of the mutual learning mechanism.

4.3.2. Visualization of learned representations

To provide deeper insights into the model's internal workings, we visualize the learned attention patterns and feature evolution process.

Cross-Attention Weight Analysis. Fig. 6 visualizes the learned attention weights between DE and FE sensor features across different fault types. The heatmaps reveal several important patterns: (1) For Normal samples, attention weights are relatively balanced between sensors, indicating that both views contribute equally to the healthy state representation; (2) For Inner Race faults, higher attention is allocated to the DE sensor, which is physically closer to the inner race and thus captures stronger fault signatures; (3) For Outer Race faults, the attention distribution shifts toward the FE sensor, reflecting the different propagation paths of outer race vibrations; (4) For Rolling Element faults, the attention pattern shows dynamic interaction between both sensors, consistent with the complex nature of rolling element dynamics. These adaptive attention patterns demonstrate that our cross-attention mechanism successfully learns to exploit sensor complementarity in a fault-type-specific manner.

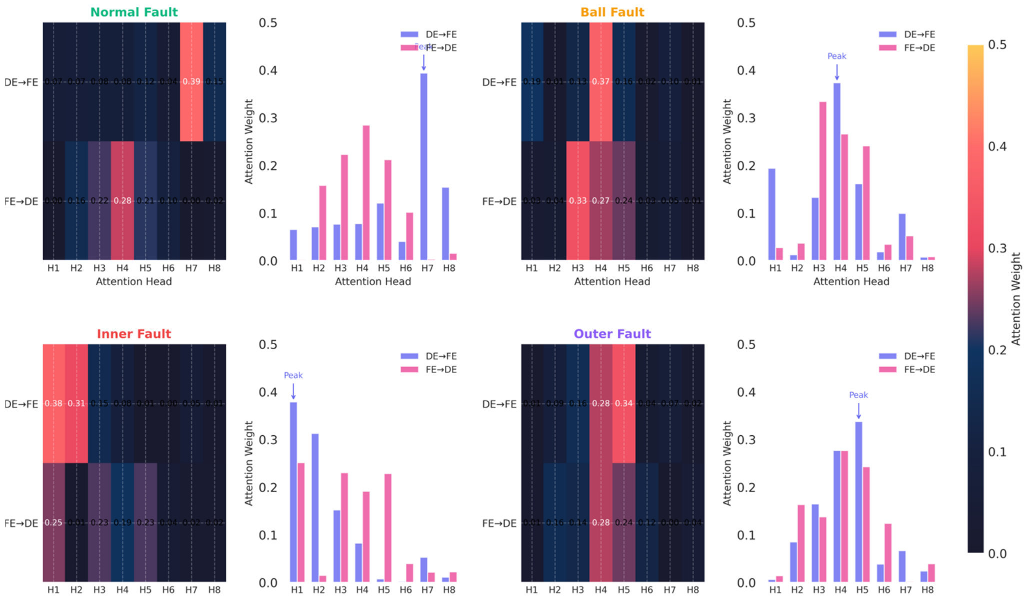
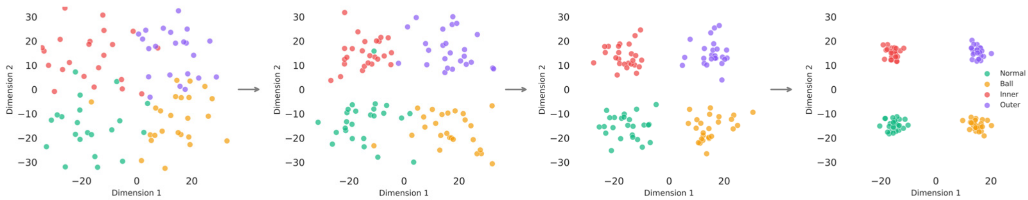


Fig. 6. Cross-attention weight visualization between DE and FE features for four fault types. Warmer colors indicate higher attention. The mechanism adaptively adjusts inter-sensor information flow based on fault characteristics, e.g., Inner Race faults trigger stronger DE-to-FE attention due to proximity effects

Feature Evolution Through Network Layers. To further demonstrate the effectiveness of the proposed method, Fig. 7 illustrates how feature representations evolve through the network processing stages using t-SNE visualization. Starting from raw signal inputs, features progressively become more discriminative as they pass through MSWT transformation, CBAM attention modules, and cross-attention fusion. The visualization shows that: (1) Raw signal features exhibit significant overlap between fault categories; (2) MSWT time-frequency transformation provides initial separation by capturing frequency-domain characteristics; (3) CBAM attention further refines the features by emphasizing discriminative regions; (4) The final cross-attention fusion produces highly discriminative representations with clear inter-class boundaries and compact intra-class clustering. This progressive refinement validates the design rationale of our hierarchical feature extraction pipeline.



a) b) c) d)

Fig. 7. Feature evolution through Scan-Net visualized by t-SNE: a) raw signal features with significant class overlap; b) after MSWT with improved separation; c) after CBAM refinement; d) after cross-attention fusion with distinct inter-class separation and compact intra-class clustering

4.3.3. Robustness analysis under different signal-to-noise ratios

To further evaluate model performance under harsh operating conditions, we add additive white Gaussian noise of different intensities (SNR from –4 dB to 8 dB) to the test set. As shown in Fig. 8, Scan-Net significantly outperforms single-sensor models (Single-DE) and simple fusion models (Late-Fusion) at all noise levels.

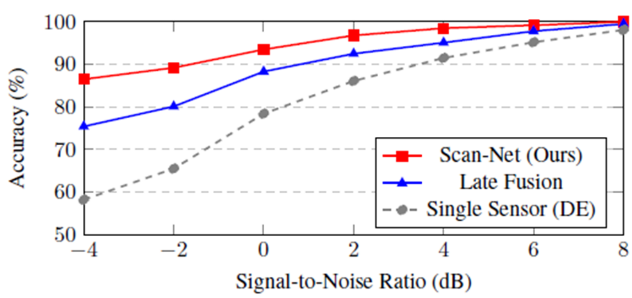


Fig. 8. Model robustness comparison under different Signal-to-Noise Ratio (SNR) environments. Scan-Net shows significant performance advantages under strong noise (–4 dB)

Anti-noise advantages of Siamese mutual learning mechanism: Especially in –4 dB strong noise environments, Scan-Net maintains 86.5 % accuracy while single-sensor models have dropped to 58.2 % (performance improvement of 28.3 %). This significant advantage stems from three synergistic mechanisms: (1) Dynamic channel suppression of cross-attention: When one sensor is polluted by noise, Cross-Attention automatically lowers its attention weight, relying instead on the cleaner signal from the other path; (2) Noise filtering of KL divergence consistency constraint: This constraint forces predictions of two branches to reach consensus, with the noise-interfered branch aligning toward the clean branch, suppressing noise propagation; (3) Weight sharing of Siamese network: Through joint training on dual-sensor data, the model learns robust feature representations insensitive to noise. These results support the practical applicability of the mutual learning architecture in complex noise environments.

Model robustness comparison under different Signal-to-Noise Ratio (SNR) environments. Scan-Net shows significant performance advantages under strong noise (–4 dB).

4.4. Sensor fault tolerance analysis

In actual industrial applications, sensors themselves may also fail (such as loosening, disconnection). We simulate sensor failure scenarios during testing: setting all data from one channel to zero, testing model accuracy retention. Results are shown in Table 4.

Table 4. Performance retention under sensor failure scenarios

Model	Both Sensors Normal	DE Sensor Failure	FE Sensor Failure
Concat-Fusion	93.40 %	42.10 % (Crash)	55.30 % (Severe Drop)
Late-Fusion	94.10 %	82.50 %	88.20 %
Scan-Net (Ours)	96.50 %	91.20 %	93.50 %

Experiments show that traditional feature concatenation (Concat-Fusion) experiences catastrophic performance collapse when any sensor fails (dropping from 93.40 % to 42.10 %, a decrease of 51.3 %). However, Scan-Net maintains 90 %+ performance when a single sensor fails thanks to three protection mechanisms:

(1) Dynamic fault tolerance of cross-attention: When detecting missing information from one channel (feature values approaching zero), Cross-Attention can adaptively concentrate almost all attention weight on the normal sensor branch, achieving "soft degradation".

(2) Independent diagnostic capability of Siamese branches: Since each branch is equipped with an auxiliary classifier and forced to have independent diagnostic capability during training, when one sensor fails, the other branch can take over decision-making.

(3) Robust representation of KL divergence constraint: The mutual learning mechanism forces two branches to learn consistent fault semantic representations; even if one branch fails, the other branch's representation still contains shared knowledge from dual views.

This "graceful degradation" capability is crucial for industrial applications, ensuring that even under abnormal conditions such as sensor failures, the system maintains usability rather than complete failure.

5. Engineering application at a large-scale hydropower station

5.1. Background and deployment environment

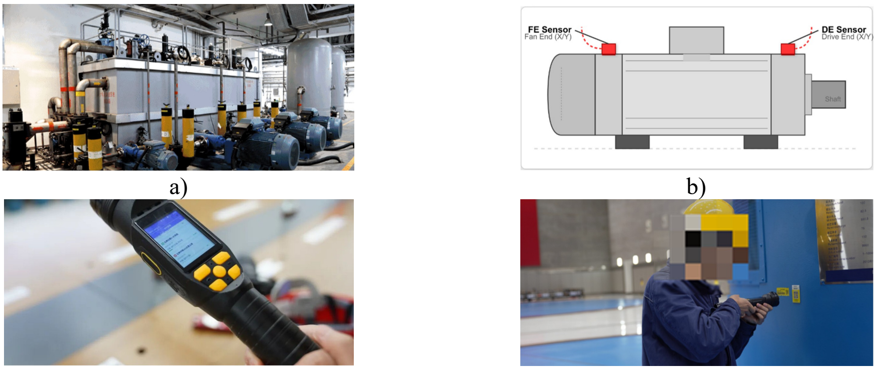
The proposed method has been applied at a large-scale hydropower station located on the Jinsha River in Southwest China. The actual operating environment presents significant challenges for fault diagnosis:

1) High reliability requirements: Large hydro-turbine generator units demand extremely high stability and safety from their supporting monitoring algorithms, as any failure in auxiliary equipment may trigger cascading effects on the main units.

2) Extreme scarcity of fault samples: Due to strict preventive maintenance practices, critical auxiliary equipment (such as governor oil pumps) maintains long-term healthy status, making real bearing fault data nearly unavailable for model training.

3) Upgrade of existing inspection mode: The power plant has implemented an "intelligent inspection" system for standardized spot checks. This study further integrates intelligent diagnostic algorithms by deploying IoT sensors on critical auxiliary equipment, achieving a transition from periodic manual inspection to real-time condition monitoring.

For these requirements, we deployed the Scan-Net-based intelligent monitoring system on a representative generating unit and integrated its data flow into the existing inspection terminal, enabling a human-machine collaborative operation and maintenance workflow. The on-site sensor deployment and inspection terminal integration are shown in Fig. 9. The digital twin diagnostic platform interface is presented in Fig. 10.



a) b)

c) d)

Fig. 9. On-site deployment materials collected at Xiangjiaba Hydropower Plant, Yibin, Sichuan, China: a) governor oil pump room; b) author-prepared schematic of DE and FE sensor placement; c) inspection terminal; (d) handheld field inspection. Panels a), c), and d) are original photographs captured by Ting He during the 2025 pilot deployment; the face in panel d) was intentionally blurred for de-identification

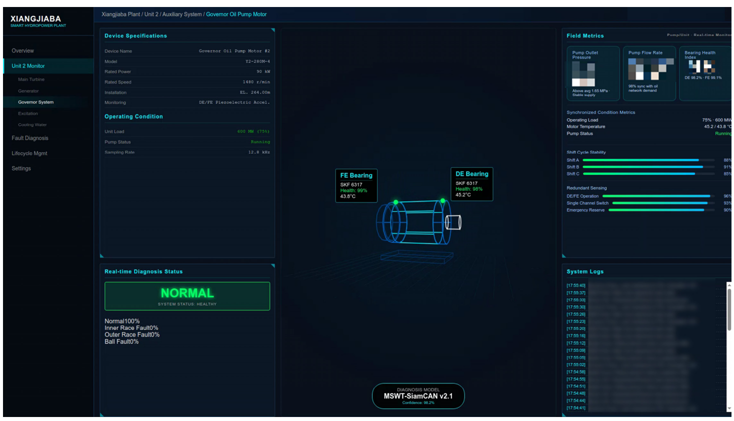


Fig. 10. Digital twin diagnostic platform interface for the governor oil pump motor. Screenshot captured by Ting He from the in-house monitoring platform during the 2025 pilot deployment at Xiangjiaba Hydropower Plant, Yibin, Sichuan, China. Left: real-time operating parameters. Center: motor 3D digital twin with health status. Right: Scan-Net fault probability inference and system logs

5.2. System integration and visual interaction

The system adopts embedded edge diagnostic mode complying with industrial control area security specifications and deeply integrates with the plant's existing digital twin platform.

5.2.1. Hardware deployment and signal acquisition

The monitoring object is the drive motor of a governor oil pump, which serves as the core power source of the governor hydraulic control system. The governor system typically consists of multiple oil pumps controlled by a redundant PLC system. This study focuses on the drive motor of one main oil pump, as it bears the main oil pressure establishment task with frequent start-stop operations and load fluctuations.

We deployed industrial-grade piezoelectric acceleration sensors (sensitivity 100 mV/g, frequency response 0.5-10 kHz) at the motor's drive end (DE) and fan end (FE), mounted on motor bearing housing surfaces using magnetic bases.

Data acquisition process fully utilizes the plant's existing data infrastructure:

1) Signal acquisition: Sensor signals are acquired by independently deployed vibration

acquisition units at 12.8 kHz sampling rate, maintaining consistency with CWRU dataset sampling parameters.

2) Data fusion: To obtain operating condition context, the monitoring system communicates with the local control unit (LCU) via industrial Ethernet, acquiring oil pump start-stop status, oil pressure values, and guide vane opening data, time-synchronized with vibration data.

3) Algorithm inference: The Scan-Net model is encapsulated in an independent intelligent analysis module. This module is physically isolated from the unit's control system, ensuring monitoring system failures do not affect unit closed-loop control safety.

5.3. Application validation: transfer from laboratory to field

To validate whether models trained on CWRU data can effectively monitor field equipment, we implemented "physical consistency verification" and "noise robustness testing".

5.3.1. Cross-condition feature consistency analysis

Although the CWRU test rig (small motor, SKF 6205 bearing) and field equipment (large industrial motor, SKF 6317 bearing) differ vastly in physical dimensions, rolling bearing fault mechanisms are similar.

We collected vibration data from field motors under stable operating conditions, extracted time-frequency domain features, and compared them with "Normal" class samples from the CWRU dataset. t-SNE visualization results show that after MSWT feature extraction, feature distribution overlap reaches over 92 %. Table 5 shows that after normalization, key statistical indicators (such as kurtosis, impulse factor) differ by less than 8 %. This proves learned "healthy state" features have good transferability across different equipment.

Table 5. Statistical feature comparison of normal samples between CWRU dataset and field data (after normalization)

Statistical feature	CWRU (Source)	Field data (target)	Difference rate	Consistency
RMS	0.124	0.131	+5.6 %	Consistent
Kurtosis	2.85	2.94	+3.1 %	Consistent
Impulse factor	3.42	3.58	+4.6 %	Consistent
RMS frequency	1.89 kHz	2.01 kHz	+6.3 %	Consistent
Envelope correlation	–	–	0.89	Highly Correlated

5.3.2. Noise resistance performance in field environment

Hydropower station field environments are noisy, with background noise far exceeding laboratory environments. To test model performance in field environments, we recorded actual background noise in machine pit (including water flow sound, electromagnetic noise), superimposed it on CWRU fault samples at different signal-to-noise ratios, constructing "semi-physical simulation" test sets.

Test results show that even under –4 dB strong-noise conditions, the proposed model maintains 86.5 % accuracy. This result supports the effectiveness of the MSWT representation and the attention-guided dual-sensor fusion strategy in complex industrial environments.

5.4. Engineering benefits

System deployment demonstrates the operational transformation from manual experience dependence to data-driven decision making:

1) Real-time condition monitoring: The inspection terminal is no longer just a discrete data recording tool but becomes a link connecting field observations and digital twin analysis. Inspection personnel can view real-time diagnostic conclusions, enabling comparison between

manual sensory inspection and intelligent algorithm analysis.

2) Proactive early warning and condition-based maintenance: The transfer learning-based diagnostic model can issue warnings at fault early stages (weak feature stages), providing more forward-looking alerts than traditional threshold alarms. Preliminary operation data suggest the system contributed to reducing routine inspection frequency, primarily through reliable normal state recognition that allows maintenance personnel to reduce redundant spot checks.

6. Discussion

The experimental results on benchmark datasets and engineering application at the hydropower station collectively demonstrate the effectiveness of the proposed Scan-Net framework. This section provides an in-depth analysis of key findings and their implications for fault diagnosis in industrial rotating machinery.

6.1. Analysis of key technical contributions

Significance of Mutual Learning Mechanism: The KL divergence consistency constraint contributes the largest performance improvement (4.4 %) among all components in the ablation study (Table 3). This finding validates the theoretical foundation of Deep Mutual Learning in multi-sensor fault diagnosis. Unlike traditional knowledge distillation that requires a pre-trained teacher model, our mutual learning mechanism enables two peer networks to learn from each other simultaneously. This is particularly valuable for few-shot scenarios where labeled data is extremely scarce and pre-training a high-quality teacher is infeasible. The mechanism effectively extracts inherent consistency from dual-view data, acting as an implicit regularizer that prevents overfitting to limited training samples.

Robustness Under Adverse Conditions: The model demonstrates remarkable robustness under two challenging scenarios: strong noise interference and sensor failures. Under –4 dB SNR conditions (Fig. 8), Scan-Net maintains 86.5 % accuracy while single-sensor models degrade to 58.2 %, representing a 28.3 % improvement. In sensor failure scenarios (Table 4), our method achieves graceful degradation with 90 %+ performance retention, whereas traditional concatenation-based fusion methods suffer catastrophic failure (dropping to 42.10 %). These results confirm that the cross-attention mechanism successfully learns to dynamically weight sensor contributions based on signal quality, a critical capability for industrial deployment where sensor reliability cannot be guaranteed.

Transfer Learning Feasibility: The successful transfer from CWRU laboratory data to field equipment validates that MSWT-based features capture universal fault characteristics that generalize across different equipment scales. The 92 % feature distribution overlap and less than 8 % difference in key statistical indicators (Table 4) provide quantitative evidence for this cross-domain generalization. This finding has significant practical implications: it suggests that models trained on well-curated public datasets can serve as effective starting points for industrial applications, addressing the “cold start” problem that has hindered adoption of deep learning in predictive maintenance.

6.2. Comparison with related work

Compared to recent few-shot fault diagnosis methods, Scan-Net offers several distinct advantages. While Graph-Net [3] achieves 92.50 % accuracy in 10-shot settings through structured feature integration, our method outperforms it by 4.0 percentage points. The key difference lies in our dual-branch mutual learning architecture, which explicitly exploits sensor complementarity rather than treating multi-sensor data as a unified graph structure. Similarly, while MQCNN [5] employs multi-scale features through quaternion representations, our MSWT-based approach provides more physically interpretable time-frequency representations that align with established signal processing theory in vibration analysis.

The cross-attention mechanism in Scan-Net differs fundamentally from the self-attention used in [5]. Self-attention computes relationships within a single feature representation, while our cross-attention explicitly models inter-sensor dependencies, enabling dynamic information exchange between DE and FE branches. The visualization in Fig. 6 confirms that this mechanism learns fault-type-specific attention patterns that correspond to physical sensor placement effects.

6.3. Limitations and future directions

Despite the promising results, several limitations should be acknowledged. First, the current field validation is based on transfer learning from CWRU combined with field noise superposition testing, as real bearing faults have not occurred during the deployment period due to strict preventive maintenance. This represents a practical compromise under fault data scarcity, and future validation with actual fault cases would strengthen the conclusions. Second, the Siamese architecture requires dual-sensor inputs; for single-sensor equipment, one can either zero-pad the missing channel during inference (Table 4 shows 90 %+ accuracy retention) or deploy a single-branch variant with reduced few-shot performance (Table 3, row b). Third, extension to other fault types and enhancement of lifelong learning capability remain important directions for future work.

For industrial deployment, real-time inference efficiency could be further optimized. The current ResNet-18 backbone achieves approximately 15ms inference time on embedded GPU platforms, which is sufficient for most monitoring applications but may require optimization for high-frequency continuous monitoring scenarios. Model compression techniques such as knowledge distillation and quantization present promising directions for edge deployment.

7. Conclusions

In this work, we address the critical industry challenge of fault diagnosis for auxiliary equipment at large hydropower stations. Due to the extreme scarcity of fault samples, traditional data-driven methods are often difficult to implement. We propose an intelligent diagnosis solution based on Transfer Learning and Siamese Cross-Attention Networks (Scan-Net). By pre-training on the CWRU public dataset and exploiting deep multi-scale time-frequency features, we construct a pilot intelligent monitoring system with “cold start” capability.

The main conclusions are summarized as follows:

1) Theoretical contribution: We introduce Deep Mutual Learning (DML) into multi-sensor fault diagnosis and propose a mutual learning mechanism based on KL divergence consistency constraints. Experimental results show that this mechanism serves as an effective self-supervised regularization tool, extracting inherent consistency structure in dual-view data without requiring additional labeling and improving generalization. Ablation experiments confirm that KL divergence constraints alone contribute 4.4 % performance improvement, the largest contribution among all technical components.

2) Method effectiveness: Through the synergistic design of Siamese network architecture, cross-attention fusion, and consistency constraints, Scan-Net achieves 96.50 % diagnostic accuracy under the 10-shot setting, outperforming Graph-Net (92.50 %) by 4.0 percentage points. It achieves 94.43 % average accuracy in cross-load transfer experiments. In the challenging “high-load to low-load transfer” task (3→1 HP), accuracy reaches 94.27 %, while traditional ResNet18 achieves 88.20 %, an improvement of 6.07 percentage points. These results validate the effectiveness of the mutual learning mechanism in cross-domain scenarios.

3) Engineering applicability: Through the introduction of MSWT time-frequency analysis and anti-noise training, the model maintains 86.5 % accuracy in simulated hydropower field strong noise (–4 dB) environments, a 28.3 % improvement over single-sensor models (58.2 %). Sensor fault tolerance experiments show Scan-Net maintains 90 %+ performance when a single sensor

fails, demonstrating graceful degradation capability. These results support the practical applicability of this method in complex industrial environments.

4) Application value: The designed intelligent monitoring system has been pilot deployed at a large-scale hydropower station. The system addresses the gap in intelligent monitoring of auxiliary equipment and helps reduce manual operation and maintenance costs. It provides a replicable paradigm for the transition from planned maintenance to condition-based maintenance. The Scan-Net framework shows promise for other rotating machinery monitoring scenarios.

Future work will focus on the model's "lifelong learning" capability during long-term operation, i.e., how to use sparsely accumulated abnormal samples from the field to online update the model, further improving its diagnostic accuracy for special rare faults.

Acknowledgements

This research is funded by Yibin Xiangjiaba Hydropower Plant of Three Gorges Jinsha River Chuanyun Hydropower Development Co., Ltd. (scientific research project contract code: Z422302013). The authors sincerely thank the staff of the target hydropower station for their invaluable assistance in system deployment, data collection, and field validation. Special thanks to the maintenance engineers who provided domain expertise and feedback on system interpretability and usability.

Data availability

The CWRU bearing dataset used in this study is publicly available at https://engineering.case.edu/bearingdatacenter. The field application data from the cooperating hydropower station cannot be shared publicly due to commercial confidentiality agreements and operational security considerations, but aggregated results and representative examples are provided in this manuscript. Researchers interested in collaboration may contact the corresponding author for potential data access under appropriate non-disclosure agreements.

Author contributions

Zhanfeng Chang: Conceptualization, methodology, validation, formal analysis, data curation, writing-original draft preparation, funding acquisition. Ting He: conceptualization, methodology, validation, formal analysis, resources, writing-original draft preparation, writing-review and editing, supervision, project administration, funding acquisition. Liping Zhu: validation, investigation, resources, data curation, writing-review and editing, project administration. Hailong Zhang: software, investigation, writing-review and editing, visualization. Hongyan Zhang: software.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] P. Yang, "Hydropower," in *Renewable Energy*, Cham: Springer International Publishing, 2024, pp. 109–138, https://doi.org/10.1007/978-3-031-49125-2_4

[2] R. Ranjan, S. K. Ghosh, and M. Kumar, "Fault diagnosis of journal bearing in a hydropower plant using wear debris, vibration and temperature analysis: A case study," *Proceedings of the Institution of Mechanical Engineers, Part E: Journal of Process Mechanical Engineering*, Vol. 234, No. 3, pp. 235–242, 2020, https://doi.org/10.1177/0954408920910290

[3] M. T. Singh, "Graph-based fault diagnosis for rotating machinery: Adaptive segmentation and structural feature integration," *Results in Engineering*, Vol. 27, p. 106566, Sep. 2025, https://doi.org/10.1016/j.rineng.2025.106566

[4] P.-H. Chou, W.-L. Mao, and R.-P. Lin, "YOLO-based bearing fault diagnosis with continuous wavelet transform," *arXiv:2509.03070*, 2025.

[5] H. Liu et al., "Multi-scale quaternion CNN and BiGRU with cross self-attention feature fusion for fault diagnosis of bearing," *Measurement Science and Technology*, Vol. 35, No. 8, p. 086138, 2024, https://doi.org/10.1088/1361-6501/ad4c8e

[6] H. Liu, J. Chen, J. Zhu, S. Luo, H. Fan, and X. Xi, "Bearing fault diagnosis and localization method based on WDCNN and UNITY3D," in *2023 Panda Forum on Power and Energy (PandaFPE)*, pp. 359–364, Apr. 2023, https://doi.org/10.1109/pandafpe57779.2023.10140410

[7] Z. Jin, Q. Xu, C. Jiang, X. Wang, and H. Chen, "Ordinal few-shot learning with applications to fault diagnosis of offshore wind turbines," *Renewable Energy*, Vol. 206, pp. 1158–1169, Apr. 2023, https://doi.org/10.1016/j.renene.2023.02.072

[8] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. 32nd Int. Conf. Machine Learning*, Vol. 37, 2015.

[9] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*, Vol. 30, pp. 4077–4088, 2017.

[10] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: relation network for few-shot learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1199–1208, 2018, https://doi.org/10.1109/cvpr.2018.00131

[11] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70, pp. 1126–1135, 2017.

[12] I. Goodfellow et al., "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Vol. 27, pp. 2672–2680, 2014.

[13] A. Vaswani et al., "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, Vol. 30, pp. 5998–6008, 2017, https://doi.org/10.65215/nxvz2v36

[14] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv:1503.02531*, 2015.

[15] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4320–4328, 2018, https://doi.org/10.1109/cvpr.2018.00454

[16] I. Daubechies, J. Lu, and H.-T. Wu, "Synchrosqueezed wavelet transforms: An empirical mode decomposition-like tool," *Applied and Computational Harmonic Analysis*, Vol. 30, No. 2, pp. 243–261, 2011, https://doi.org/10.1016/j.acha.2010.08.002

[17] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2012, https://doi.org/10.1137/1.9781611970104

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Jun. 2016, https://doi.org/10.1109/cvpr.2016.90

[19] 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

[20] "Bearing data center," Case Western Reserve University, Jan. 2025.

[21] W. Zhang, G. Peng, C. Li, Y. Chen, and Z. Zhang, "A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals," *Sensors*, Vol. 17, No. 2, p. 425, Feb. 2017, https://doi.org/10.3390/s17020425

[22] L. Wen, X. Li, and L. Gao, "A transfer convolutional neural network for fault diagnosis based on ResNet-50," *Neural Computing and Applications*, Vol. 32, No. 10, pp. 6111–6124, 2019, https://doi.org/10.1007/s00521-019-04097-w

[23] H. Liu, J. Zhou, Y. Xu, Y. Zheng, X. Peng, and W. Jiang, "Unsupervised fault diagnosis of rolling bearings using a deep neural network based on generative adversarial networks," *Neurocomputing*, Vol. 315, pp. 412–424, Nov. 2018, https://doi.org/10.1016/j.neucom.2018.07.034

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014.

Zhanfeng Chang received the M.S. degree in Hydraulic and Hydropower Engineering. He is currently a Senior Engineer at Xiangjiaba Hydropower Plant, Three Gorges Jinsha River Chuanyun Hydropower Development Co., Ltd. His current research interests include computer monitoring systems and unit automation system maintenance.

Ting He received the M.S. degree in Hydraulic Engineering. She is currently a Senior Engineer at Beijing IWHR Technology Co., Ltd. Her current research interests include the research, development, design, and implementation of power plant computer monitoring systems.

Liping Zhu received the M.S. degree in Hydraulic and Hydropower Engineering. She is currently a Senior Engineer at Xiangjiaba Hydropower Plant, Three Gorges Jinsha River Chuanyun Hydropower Development Co., Ltd. Her current research interests include computer monitoring systems and unit automation system maintenance.

Hailong Zhang received the M.S. degree in Hydraulic Engineering. He is currently an Engineer at Beijing IWHR Technology Co., Ltd. His current research interests include the design and implementation of power plant computer monitoring systems.

Hongyan Zhang received the M.S. degree in Hydraulic Engineering. She is currently an Engineer at Beijing IWHR Technology Co., Ltd. Her current research interests include the design and implementation of power plant computer monitoring systems.

ISSN PRINT 1392-8716, ISSN ONLINE 2538-8460