Time-frequency Correspondence-guided Representation Learning for Practical Unsupervised Domain Adaptation in Machinery Fault Detection

Mingrui Li[1]

[1]College of Artificial Intelligence, China University of Petroleum (Beijing), Beijing 102249, China

ABSTRACT

Unsupervised domain adaptation works well for cross-domain mechanical fault diagnosis, but mainstream approaches overlook the intrinsic time-frequency relationships within vibration samples. Global distribution alignment will easily damage delicate local fault characteristics. To remedy these shortcomings, we construct a dual-consistency representation learning framework. It combines internal time-frequency physical matching of single samples with global domain alignment. Equipped with bidirectional token cross-attention and domain-decoupled contrastive regularization, the model can preserve robust diagnostic features against noise and large domain gaps. Under standardized practical UDA experimental settings, cross-dataset tests prove our method greatly reduces false alarms and delivers superior generalization performance compared with competing baselines, with no target-domain labels adopted for training.

Keywords: *Unsupervised domain adaptation, Machinery fault detection, Time-frequency correspondence, Representation learning, Dual-consistency learning.*

1. INTRODUCTION

Intelligent fault detection plays an essential role in guaranteeing the reliable operation of modern rotating machinery [1]. Unsupervised domain adaptation (UDA) offers a feasible way to transfer diagnostic knowledge from labeled source data to unlabeled target operating conditions. A number of UDA methods have achieved decent cross-domain fault classification results by aligning global data distributions [2-5]. Even so, most of these methods process time-domain and frequency-domain vibration signals independently, ignoring the inherent local spectral-temporal dependencies within each signal window [6]. Recent research on multimodal fusion and domain generalization still prioritizes global distribution matching. When faced with significant domain shifts and strong background noise, this kind of global alignment can corrupt fine-grained fault features and produce frequent false alarms in real industrial applications [7]. Against these real-world challenges, we develop a cross-domain fault detection framework rooted in time-frequency correspondence. Rather than merely fusing time- and frequency-domain signals, our work focuses on capturing the intrinsic mapping between temporal impulse components and resonant frequency bands [8]. To retain critical fault-related details during feature aggregation, an adaptive gated residual fusion strategy is adopted in our model [6]. On this basis, we introduce contrastive constraints to enforce physical consistency within individual samples, together with a view-level domain discriminator for global distribution adaptation across domains [9-10].

2. METHODOLOGY

2.1 Problem Formulation and Framework Overview

The framework targets an unsupervised cross-domain binary fault detection task, transferring knowledge from a labeled source domain to an unlabeled target domain:

$$D_s=\{(x_i^{time,s},x_i^{freq,s},y_i^s)\}_{i=1}^{N_s},\quad D_t=\{(x_j^{time,t},x_j^{freq,t})\}_{j=1}^{N_t}$$

Here, x^{time} is the raw time-domain sample, x^{freq} is the corresponding frequency spectrum, and $y \in \{0,1\}$ is the binary health label.

Under the practical UDA protocol, the target domain is split into three mutually isolated subsets: training set $\mathbf{D}_{t,train}$, validation set $\mathbf{D}_{t,val}$ and test set $\mathbf{D}_{t,test}$. No target labels are used during optimization; validation labels are only for checkpoint selection, and the test set is strictly held out for final evaluation.

As shown in "Figure 1, our framework consists of two coordinated learning parts. Internal consistency corresponds to time-frequency matching. Through verification inside each sample, the model extracts robust cross-modal time-frequency relationships within individual vibration windows. External consistency refers to domain alignment on the feature view. Adversarial training is adopted to calibrate cross-domain marginal distributions and strengthen the transfer performance of feature representations.

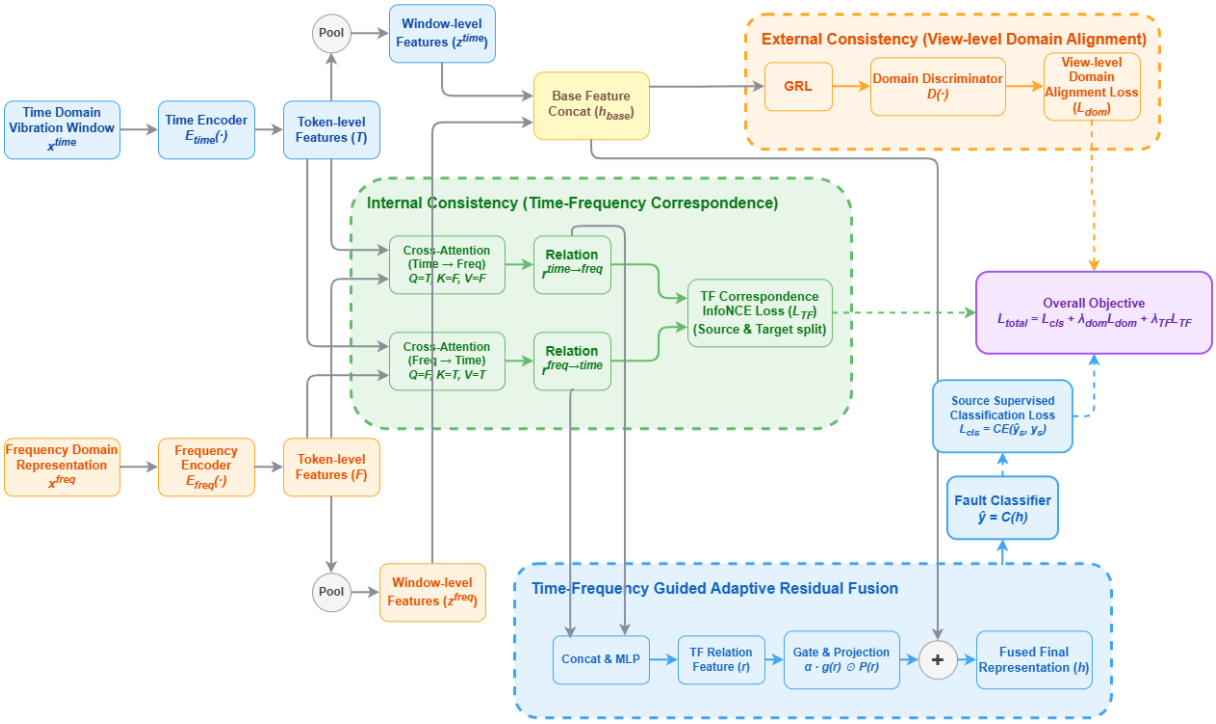


Figure 1 Global architecture of the proposed framework, balancing micro-level internal consistency and macro-level external domain alignment.

2.2 Dual-path Feature Encoder

A dual-path encoder processes temporal and spectral inputs in parallel to capture multi-modal characteristics. Both branches adopt a 1D convolutional backbone with a downsampling layer to generate token-level features of consistent dimensions. The time encoder $E_{\text{time}}(\cdot)$ extracts transient impulses and anomalous waveforms from raw vibration sequences x^{time}, consisting of 4 stacked 1D convolutional layers with kernel sizes 64, 32, 16, 8, outputting 64 tokens of dimension 256. The frequency encoder $E_{\text{freq}}(\cdot)$ captures spectral energy distributions and resonant responses from x^{freq}, sharing the same depth and output dimension with kernel sizes 16, 8, 4, 2 adapted to spectral features.

For an arbitrary input, the extracted token-level features are:

$$T = E_{\text{time}}(x^{\text{time}}), \quad F = E_{\text{freq}}(x^{\text{freq}})$$

where $T \in \mathbb{R}^{N_{\text{time}} \times d}$ and $F \in \mathbb{R}^{N_{\text{freq}} \times d}$ are token-level feature maps for the two domains. A global

average pooling operator $\mathrm{Pool}(\cdot)$ aggregates tokens into window-level global representations:

$$z^{\mathrm{time}}=\mathrm{Pool}(T),\quad z^{\mathrm{freq}}=\mathrm{Pool}(F)$$

Retaining token-level features preserves fine-grained structural information, avoiding the loss of local correlations caused by sole reliance on pooled statistics.

2.3 Token-level Bidirectional Cross-attention

Our bidirectional cross-attention module adopts 8 attention heads with 32 dimensions each. It constructs subtle temporal-spectral correlations guided by bearing fault physical characteristics. Local bearing defects trigger periodic transient impulses in the time domain and excite resonant peaks in the frequency domain. Rather than relying solely on data to merge features, this bidirectional architecture explicitly captures the natural physical correspondence behind vibration signals.

2.3.1 Time-to-frequency Direction

$$(R^{time\to freq})$$

Time-domain tokens serve as Query, while frequency-domain tokens act as Key and Value:

$$R^{\mathrm{time}\to\mathrm{freq}}=\mathrm{CrossAttn}(Q=T,K=F,V=F)$$

This direction allows temporal impulse patterns to dynamically query correlated spectral resonant bands, mimicking the physical rule that mechanical impacts excite specific resonant frequencies.

2.3.2 Frequency-to-time Direction

$$(R^{freq\to time})$$

Conversely, frequency tokens act as Query and time tokens as Key and Value:

$$R^{\mathrm{freq}\to\mathrm{time}}=\mathrm{CrossAttn}(Q=F,K=T,V=T)$$

This reverse direction allows spectral anomalies to back-verify associated transient impacts in the time domain. Such bidirectional coupling conforms to the physical nature of mechanical failures, where time-domain impacts and frequency-domain resonances are mutually causal.

2.4 Time-frequency Guided Adaptive Residual Fusion

The directional relation matrices are pooled into window-level relationship representations:

$$r^{\mathrm{time}\to\mathrm{freq}}=\mathrm{Pool}(R^{\mathrm{time}\to\mathrm{freq}}),\quad r^{\mathrm{freq}\to\mathrm{time}}=\mathrm{Pool}(R^{\mathrm{freq}\to\mathrm{time}})$$

The directional features are concatenated and projected via a two-layer MLP with ReLU activation to obtain the integrated time-frequency correspondence feature r:

$$r=\mathrm{MLP}\left([r^{\mathrm{time}\to\mathrm{freq}}\,\|\,r^{\mathrm{freq}\to\mathrm{time}}]\right)$$

where $\|$ denotes concatenation.

To avoid overriding baseline diagnostic details, an adaptive gated residual fusion strategy is adopted. The baseline multi-modal representation is formed by concatenating global features:

$$\mathrm{h}_{\mathrm{base}}=[z^{\mathrm{time}}\,\|\,z^{\mathrm{freq}}]$$

The correspondence feature r is injected as a controlled residual to generate the enhanced representation:

$$\mathrm{h}=\mathrm{h}_{\mathrm{base}}+\alpha\cdot g(r)\odot P(r)$$

where $P(\cdot)$ is a linear projection head, $g(\cdot)$ is a sigmoid gating function with output range $[0,1]$, $\odot$ is element-wise multiplication, and $\alpha=1.0$ is a scaling factor. The final prediction is obtained via a health classifier:

$$\hat{y}=C(\mathrm{h})$$

2.5 Optimization Objectives and Multi-loss Constraints

A multi-loss strategy is designed to jointly optimize classification accuracy, internal correspondence stability and external domain alignment.

2.5.1 Source Supervised Classification Loss

The source supervised classification loss includes a main head on fused features $\mathbf{h}$ and two auxiliary heads on single-modality features z^{time} and z^{freq}. Let y^s_{main}, y^s_{time}, y^s_{freq} denote the corresponding predictions; the cross-entropy losses are:

$$L_{\text{main}}=\text{CE}(y^s_{\text{main}},y^s),\quad L_{\text{time}}=\text{CE}(y^s_{\text{time}},y^s),\quad L_{\text{freq}}=\text{CE}(y^s_{\text{freq}},y^s)$$

The total classification loss is a weighted combination:

$$L_{\text{cls}}=L_{\text{main}}+0.3L_{\text{time}}+0.3L_{\text{freq}}$$

The auxiliary weights are empirically set to balance supervision strength.

2.5.2 Time-frequency Correspondence InfoNCE Loss

An InfoNCE-based contrastive constraint is introduced to ensure the cross-attention captures physically meaningful time-frequency structures. For the i-th sample, the two directional relationship vectors form a positive pair, while other samples in the batch serve as negatives. Vectors are projected and L_2 -normalized via a two-layer non-linear projection head $\phi(\cdot)$:

$$u_i=\text{Norm}\left(\phi(r^{\text{time}\to\text{freq},i})\right),\quad v_i=\text{Norm}\left(\phi(r^{\text{freq}\to\text{time},i})\right)$$

The contrastive loss for the $u\to v$ direction is:

$$L_{u\to v}=-\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\text{sim}(u_i,v_i)/\tau)}{\sum_{j=1}^{B}\exp\left(\text{sim}(u_i,v_j)/\tau\right)}$$

where B is batch size, $\tau=0.1$ is temperature, and $\text{sim}(\cdot,\cdot)$ is cosine similarity. The symmetric loss $L_{v\to u}$ is computed similarly, giving $L_{\text{TF-single}}=\frac{1}{2}(L_{u\to v}+L_{v\to u})$.

To avoid cross-domain distribution interference in negative pair construction, contrastive losses for source and target domains are computed independently, with pairs only built within the same domain:

$$L_{\text{TF}}=\frac{1}{2}(L^s_{\text{TF}}+L^t_{\text{TF}})$$

2.5.3 View-level Domain Alignment Loss

Domain adversarial training is applied to baseline features $\mathbf{h}_{\text{base}}$ instead of the enhanced representation $\mathbf{h}$, to avoid destroying learned time-frequency correspondences while aligning marginal distributions.

The domain discriminator $D(\cdot)$ is a 3-layer LeakyReLU MLP (256/128 hidden dimensions) that predicts domain identity. A Gradient Reversal Layer (GRL) with linearly increasing gradient scale is inserted between the encoder and discriminator:

$$\hat{d}=D(\text{GRL}(\mathbf{h}_{\text{base}}))$$

The adversarial adaptation is optimized via binary cross-entropy:

$$L_{\text{dom}}=\text{BCE}(\hat{d},d)$$

where $d\in\{0,1\}$ is the domain label.

2.5.4 Total Objective Function

The overall training objective combines the above constraints:

$$L_{\text{total}}=L_{\text{cls}}+\lambda_{\text{dom}}L_{\text{dom}}+\lambda_{\text{TF}}L_{\text{TF}}$$

where hyperparameters $\lambda_{\text{dom}}=0.025$ and $\lambda_{\text{TF}}=0.0125$ are selected via grid search on the target validation set.

3. EXPERIMENTAL VERIFICATION

3.1 Dataset and Preprocessing

A challenging cross-dataset transfer task is constructed to evaluate generalization performance: the Case Western Reserve University (CWRU)

bearing dataset [11] as the source domain $\mathbf{D}_s$, and the MFPT dataset [12] as the target domain $\mathbf{D}_t$. This CWRU → MFPT task involves severe domain shifts from different testbeds, sensor placements, operating speeds and manufacturing tolerances.

We formulate the task as binary Normal/Fault classification, which matches the practical demand of industrial early warning and avoids multi-class ambiguity in cross-domain transfer.

For the source domain derived from the CWRU dataset, we employ drive-end vibration signals sampled at 48 kHz collected under loads ranging from 0 to 3 hp. The data covers healthy bearing conditions alongside three distinct fault modes: inner race fault, outer race fault and rolling element fault, with fault diameters spanning 0.1778 mm to 0.5334 mm; all faulty samples are collectively categorized into a unified fault class, and 12,000 samples are prepared for source-domain training.

The target domain adopts the MFPT dataset. Its raw vibration signals are resampled to 48 kHz to align with the sampling frequency of the source data, and this dataset is also configured for binary fault detection. To avoid sample overlap between sliding windows and prevent unintended data leakage, we partition the MFPT dataset strictly on the basis of original signal files, which yields 6,000 training samples, 1,500 validation samples and 2,500 test samples respectively.

Signals are sliced via sliding windows (length 2048, hop 1024), with per-window z-score normalization to standardize amplitude variations.

3.2 Practical UDA Protocol

Our experiments adhere to the standardized practical unsupervised domain adaptation paradigm detailed in reference [13]. No labeled samples from the target domain are accessible throughout the entire training process. We utilize the validation set solely to adjust hyperparameters and pick optimal model checkpoints, with Macro-F1 adopted as the assessment metric. The test dataset is completely reserved unseen throughout tuning; we conduct final evaluation merely one time using the selected best checkpoint.

3.3 Implementation Details and Baselines

We split model training into two stages. We first pre-train the network on source-domain labeled data for 50 epochs, and then fine-tune the whole model for 100 epochs under the unsupervised domain adaptation setting with two domains jointly involved. We choose the Adam optimizer with an initial learning rate of 1×10^{-4} and a batch size of 32, and loss weights are fixed during fine-tuning. All experiments were completed on a single NVIDIA RTX 3090 GPU.

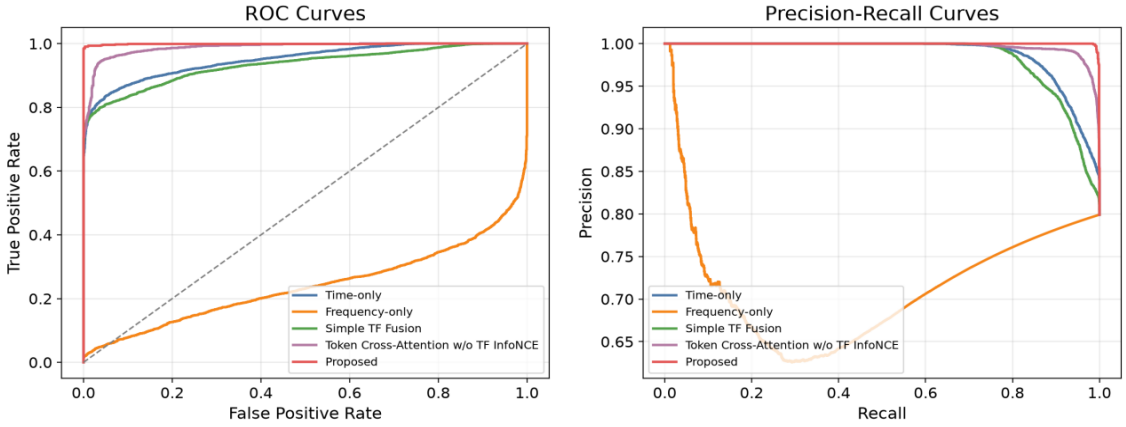
We prepare two sets of comparative models for evaluation. The ablation group covers four simplified variants: time-only input, frequency-only input, naive time-frequency fusion, and cross-attention without InfoNCE regularization. All these variants share the same encoder backbone and training setup.

For mainstream comparison benchmarks, source-only CNN and Transformer are regarded as the baseline lower bounds. We also incorporate DANN [3] and CDAN [4], and equip them with consistent encoder architecture to enable equitable performance comparison.

All measured metrics are averaged across 10 independent runs with different random seeds and reported as mean ± standard deviation. Typical ROC and PR curves plotted from the 11th seed are supplied for visual demonstration.

3.4 Ablation Study and Visual Dynamics Analysis

“Figure 2” shows ROC and PR trajectories for a representative trial, and “Table 1” summarizes quantitative results over 10 seeds.



(a)ROC curves (seed 11). (b)PR curves (seed 11).

Figure 2 Representative diagnostic trajectories for ablation variants under seed 11.

Table 1. Ablation study results across 10 random seeds (mean ± std).

Method	Macro-F1 (%)	Fault Recall (%)	FAR (%)	MCC	ROC-AUC (%)	PR-AUC (%)
Time-only	84.51 ± 13.75	87.52 ± 16.78	8.71 ± 21.40	0.7455 ± 0.1898	99.30 ± 1.18	99.83 ± 0.29
Frequency-only	25.25 ± 3.25	12.08 ± 5.38	19.50 ± 10.26	-0.0840 ± 0.0714	23.97 ± 4.71	72.01 ± 2.93
Simple TF Fusion	77.86 ± 4.86	77.98 ± 6.41	2.96 ± 7.06	0.6283 ± 0.0746	95.50 ± 3.79	98.93 ± 0.89
Token Cross-Attn w/o InfoNCE	87.13 ± 12.52	87.04 ± 13.65	1.54 ± 4.86	0.7856 ± 0.1996	99.37 ± 1.91	99.84 ± 0.48
Proposed	96.62 ± 3.13	97.52 ± 2.91	1.57 ± 4.72	0.9359 ± 0.0571	99.93 ± 0.09	99.98 ± 0.02

Our ablation experiments deliver three core observations. Models using only frequency information perform even worse than random guessing. Cross-testbed domain gaps alter resonant frequencies and degrade spectral discriminability, and temporal information is unavailable to remedy this issue. Plain time-frequency fusion brings moderate performance gains but hits an obvious ceiling, as it cannot dynamically match local temporal impulses and resonant frequency bands. Bidirectional cross-attention helps refine the classification boundary. Combined with domain-decoupled contrastive constraints, time-frequency correspondence remains stable under domain shifts, attaining the optimal overall performance.

3.5 Comparative Evaluation Against State-of-the-art Methods

"Table 2" presents the comparison with representative domain adaptation methods on the target test set.

Table 2. Performance comparison of domain adaptation methods

Method	Accuracy (%)	Macro-F1 (%)	Fault Recall (%)	FAR (%)	MCC
CNN source-only	68.98 ± 16.13	43.12 ± 5.74	82.64 ± 25.87	85.36 ± 31.79	-0.0627 ± 0.2168
Transformer source-only	84.18 ± 3.54	77.37 ± 5.42	86.59 ± 4.04	25.39 ± 15.91	0.5631 ± 0.1149
DANN UDA	87.25 ± 7.83	73.63 ± 21.66	93.97 ± 8.23	39.50 ± 46.01	0.6678 ± 0.2911
CDAN UDA	89.53 ± 10.23	80.11 ± 21.61	93.39 ± 11.93	25.86 ± 42.92	0.7256 ± 0.3276
Proposed	97.70 ± 2.24	96.62 ± 3.13	97.52 ± 2.91	1.57 ± 4.72	0.9359 ± 0.0571

Comparison of diagnostic performance against representative domain-adaptation baselines under the CWRU → MFPT cross-dataset transfer task. All reported metrics are aggregated from ten independent runs with different random seeds. Comparative results reveal that direct transfer without domain adaptation yields severe performance degradation and high false-alarm rates, which demonstrates the necessity of domain-adaptive strategies. Although conventional adversarial adaptation methods can boost fault recall, they still suffer from excessive false alarms since global distribution alignment may corrupt fine-grained local diagnostic structures. By contrast, our proposed method obtains the optimal overall performance and substantially suppresses false alarms. The relatively large standard deviation of FAR originates from the limited fraction of fault samples within the target domain, which magnifies result variance across different random seeds. On the whole, the dual-consistency framework achieves favourable prediction accuracy and reliable practical performance.

4. CONCLUSION

We develop a time-frequency correspondence driven representation learning framework to realize unsupervised cross-domain mechanical fault diagnosis. By jointly optimizing sample-level time-frequency consistency and global adversarial domain alignment, our method learns physically interpretable invariant features robust to large domain gaps and noise. Tests on the CWRU-to-MFPT unsupervised transfer task verify that our method obtains competitive diagnostic accuracy and greatly reduces false alarms, without introducing target test data leakage.

This study still has room for improvement. The bidirectional cross-attention leads to increased inference computation cost, and our validation is limited to one cross-domain task only. In future work, we will adopt lightweight sparse attention to accelerate inference for industrial online applications, and extend the dual-consistency framework to multi-class and composite fault diagnosis across various transfer scenarios.

REFERENCES

[1] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, and A. K. Nandi, “Applications of machine learning to machine fault diagnosis: A review and roadmap,” *Mechanical Systems and Signal Processing*, vol. 138, p. 106587, 2020.

[2] W. Lu, B. Liang, Y. Cheng, D. Meng, J. Yang, and T. Zhang, “Deep adversarial network with multilayer domain adaptation for bearing fault diagnosis,” *IEEE Transactions on Industrial Electronics*, vol. 64, no. 8, pp. 6800–6809, 2017.

[3] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, “Domain-adversarial training of neural networks,” *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[4] M. Long, Z. Cao, J. Wang, and M. I. Jordan, “Conditional adversarial domain adaptation,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, pp. 1647–1657, 2018.

[5] K. Zhang, Y. Xu, and P. Wang, “Joint alignment of marginal and conditional distributions for cross domain fault diagnosis,” *IEEE Transactions on Industrial Informatics*, vol. 19, no. 4, pp. 5120–5130, 2023.

[6] J. Wang, S. Li, and Z. Zhao, “Multi-scale time-frequency fusion network with contrastive learning for machinery fault diagnosis,” *Mechanical Systems and Signal Processing*, vol. 208, p. 111050, 2024.

[7] X. Chen, Y. Tang, and H. Zhou, “Physics-informed time-frequency contrastive learning for domain generalized fault detection,” *IEEE/ASME Transactions on Mechatronics*, vol. 29, no. 3, pp. 2105–2116, 2024.

[8] L. Xu, H. Niu, and G. Dong, “Cross-modal dynamic feature fusion network for intelligent fault diagnosis of rolling bearings,” *IEEE Transactions on Industrial Informatics*, vol. 19, no. 9, pp. 9812–9822, 2023.

[9] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.

[10] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshkin, “Momentum contrast for unsupervised visual rep resentation learning,” in *Proceedings of the IEEE/CVF Conference on Computer*

Vision and Pattern Recognition (CVPR), pp. 9729–9738, 2020.

[11] W. A. Smith and R. B. Randall, "Rolling element bearing diagnostics using the Case Western Reserve University data: A review," *Mechanical Systems and Signal Processing*, vol. 64, pp. 100–131, 2015.

[12] E. Bechhoefer, "Machinery Failure Prevention Technology (MFPT) fault data sets," *Society for Ma chinery Failure Prevention Technology*, 2013.

[13] H. Zheng, R. Yan, and X. Chen, "Practical unsupervised domain adaptation for intelligent fault diag nosis: Protocols, benchmarks and guidelines," *IEEE Transactions on Industrial Electronics*, vol. 70, no. 12, pp. 12780–12790, 2023.