Collaboration better than Integration: A Novel Time-frequency-assisted Deep Feature Enhancement Mechanism for Few-shot Transfer Learning in Anomaly Detection

Wentao Mao, *Member, IEEE,* Jianing Wu, Shubin Du, Ke Feng, *Senior Member, IEEE,* and Zidong Wang, *Fellow, IEEE*

***Abstract*—Deep transfer learning has achieved significant success in anomaly detection over the past decade, but data acquisition challenges in practical engineering hinder high-quality feature representation for few-shot learning tasks. To address this issue, a novel time-frequency-assisted deep feature enhancement mechanism (TFE) is proposed. Unlike traditional methods that integrate time-frequency analysis with deep neural networks, TFE employs a wavelet scattering transform to establish a parallel time-frequency feature space, where a dual interaction strategy facilitates collaboration between deep feature and time-frequency spaces through two operations: 1) *Enhancement*, where a frequency-importance-driven contrastive learning (FICL) network transfers physically-aware information from wavelet scattering features to deep features, and 2) *Feedback*, which uses a detection rule adaptation module to minimize bias in wavelet scattering features based on deep feature performance. TFE is applied to a domain-adversarial anomaly detection framework and, through alternating training, significantly enhances both deep feature discriminative power and few-shot anomaly detection. Theoretical analysis confirms that the proposed dual interaction strategy reduces the upper bound of classification error. Experiments on benchmark datasets and a real-world industrial dataset from a large steel factory demonstrate TFE's superior performance and highlight the importance of frequency saliency in transfer learning. Thus, *collaboration* is shown to outperform *integration* for few-shot transfer learning in anomaly detection.**

***Index Terms*—Anomaly detection, Time frequency analysis, Transfer learning, Feature enhancement, Few-shot learning.**

I. Introduction

With promising performance in various tasks (e.g. anomaly detection and image recognition), deep learning techniques continue to face challenges in scenarios with limited or insufficient training data. *Few-shot deep learning*, a technique designed to address this issue, aims to develop models that perform effectively with minimal data. In deep transfer learning, sufficient source domain data significantly enhances the representation of domain knowledge, facilitating effective knowledge transfer. Thus, the core challenge of few-shot deep learning lies in developing a feature representation that captures essential information from limited data. To the best of our knowledge, this remains a vast and underexplored research area.

Fault detection in rotating machinery exemplifies the challenges in few-shot deep learning. Mechanical equipment is often prone to unexpected failures in practical engineering, necessitating the early recognition of faults from normal states. As illustrated in Fig. 1, with sufficient normal state data, the detection rule, which is represented by the boundary of the support vector data description (SVDD) hypersphere, becomes well-defined. Despite noise and variations in equipment degradation, models trained on large datasets can effectively separate normal and fault states. Unfortunately, in few-shot scenarios, detection rules are highly sensitive to individual samples, with even minor noise causing shifts in the hypersphere, leading to misclassifications or missed alarms.

To enhance few-shot deep learning, it is imperative to develop feature representations from limited data that effectively capture key information, and such a process is referred to as feature enhancement [1]. Beyond accuracy, few-shot detection results must achieve greater reliability than those derived from large-scale data, which are often presumed to be trustworthy. The main challenges include establishing valid detection rules from limited samples and identifying the most critical information for few-shot tasks. Addressing these challenges requires a robust feature enhancement mechanism that excels in few-shot scenarios and provides a degree of model interpretability.

Generative adversarial networks (GANs) and other generative variants (e.g., diffusion model) are widely regarded as a prevalent method for deep feature enhancement. By establishing adversarial training between a generator and a discriminator, GANs enable the generator to learn the underlying data distribution and produce realistic samples. While GAN-based data augmentation techniques show promise in improving model generalization and feature representation [2], they typically require large training datasets to maintain generative capability. In few-shot scenarios, the negative impact

This work was supported in part by National Natural Science Foundation of China under [Grant 62472146], in part by the Key Technologies Research Development Joint Foundation of Henan Province of China under [Grant 225101610001].

Wentao Mao and Jianing Wu are with the school of Computer and Information Engineering, Henan Normal University, Xinxiang, 453007, China. Wentao Mao is also with the Engineering Lab of Intelligence Business & Internet of Things, Henan Province 453007, China. Shubin Du is with the Department of Equipment and Mechanical Power, HBIS Group Tangsteel Company, Tangshan, 063000, China. Ke Feng is with the State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Shaanxi 710049, China. Zidong Wang is with the Department of Computer Science, Brunel University London, Uxbridge, Middlesex, UB8 3PH, U.K.

Corresponding author: kefeng@xjtu.edu.cn.

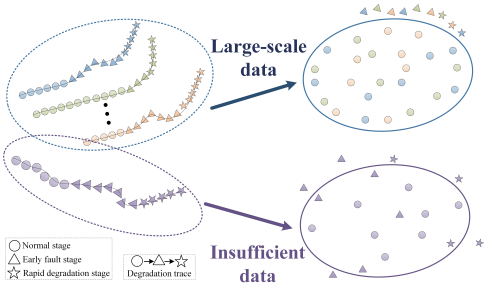


Fig. 1. Schematic of the detection rules with training data of different scales.

of noise cannot be empirically mitigated as it is in scenarios with abundant data, leading to biased simulated data and low-quality generated features. As a result, the performance of GANs significantly deteriorates in few-shot learning tasks.

In recent years, time-frequency analysis methods, such as wavelet transform, have been successfully applied to feature enhancement in few-shot deep learning [3], [4]. With a solid mathematical foundation, wavelet transform captures frequency-level features at different scales, enabling the extraction of complete and physically meaningful information from limited data. However, most current studies use wavelet transform primarily as filters during feature extraction, which can essentially be considered an *integration* strategy [5]. While simple and intuitive, this approach risks information loss during the filtering process. Thus, a critical challenge in few-shot deep learning with time-frequency analysis lies in further reducing information loss and improving the quality of feature representation.

Inspired by the above discussions, this paper presents a novel time-frequency-assisted deep feature enhancement (TFE) mechanism. Unlike the traditional *integration*-based strategy, TFE employs a dual interaction strategy centered on *collaboration* between the time-frequency space and deep feature space. The dual interaction strategy can be essentially viewed as an information-sharing mechanism, hereby realizing the mutual influence between the two spaces. As illustrated in Fig. 2(c), this *collaboration* strategy captures comprehensive and critical features from limited data with minimal information loss, and such a process likened to a "teacher-student" learning model: the wavelet transform acts as the experienced teacher, guiding the deep learning model (the student) to enhance its learning, while the student provides feedback to refine the teacher's guidance. Unlike *integration*, *collaboration* leverages the complementary strengths of the two feature spaces, ensuring complete feature extraction in few-shot scenarios and reducing the potential information loss caused by using wavelet transforms as filters. Consequently, the reliability of the resulting learning model is significantly enhanced, which is a critical aspect of few-shot deep learning.

To illustrate the TFE mechanism, one-class anomaly detection is used as an example, as it clearly defines task objectives, including detection rules and discriminative information. The mechanism consists of two key steps:

1) *Building a Deep Time-Frequency Network*: A time-frequency feature space is created using the wavelet scattering transform (WST) [6], which provides benefits such as norm preservation, non-expansion, translation invariance, and shape stability. These properties support deep features in few-shot scenarios, even with deviations. The time-frequency space operates parallel to the deep feature space, which is expanded by any deep neural network.
2) *Enabling Dual Interaction*: Two operations, *Enhancement* and *Feedback*, facilitate interaction between the time-frequency and deep feature spaces:
   - *Enhancement*: A frequency importance-driven contrastive learning (FICL) network incorporates the frequency importance (*FI*) metric to fuse time-frequency information into deep features through unsupervised contrastive learning. The *FI* metric is designed to quantify the similarity between the two domains, reflecting the contribution of each frequency band during the transfer process.
   - *Feedback*: A detection rule adaptation module transfers discriminative information from deep features to the time-frequency space. This process updates the time-frequency feature space based on detection results in the deep feature space.

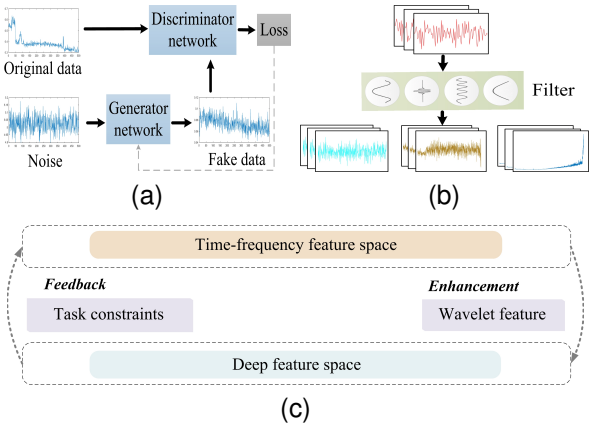


Fig. 2. Schematic of deep feature enhancement methods, where (a) is GANs, (b) is wavelet transforms adopted as filters, and (c) is the TFE mechanism. Quite different from (a) that generates synthetic samples and (b) that integrates fixed filters into deep neural network, TFE employs a dual interaction strategy between the time-frequency space and deep feature space for ensuring complete feature extraction.

The operations of *Enhancement* and *Feedback* are performed alternatively, enabling deep features to incorporate more complete and physically-aware information from wavelet features. Unlike the conventional methods illustrated in Fig. 2, the proposed TFE mechanism preserves the feature extraction capabilities of deep neural networks while mitigating the information loss associated with wavelet transform. TFE is integrated into a domain-adversarial anomaly detection framework to facilitate few-shot transfer learning. An upper bound on TFE's generalization error is derived, revealing two critical factors: the aggregation degree of deep features and wavelet features. Through iterative enhancement and feedback, the representation quality of both feature spaces is improved, which demonstrates that the dual interaction strategy

effectively reduces classification error and enhances model reliability. Experiments on image recognition and early fault detection validate TFE's superior performance and underscore the importance of frequency saliency in the transfer process, as quantified by the *FI* metric. By measuring the significance of each frequency band in the source domain, the detection results in few-shot scenarios are rendered more reliable and trustworthy.

It is noteworthy that TFE extends beyond anomaly detection, with its primary aim being the improvement of feature representation in few-shot scenarios, making it adaptable to any deep neural network. By modifying the detection rule in the feedback step, TFE can be customized for various applications, such as fault prognosis and object detection. The key contributions of this paper are as follows.

1) Proposed TFE Mechanism: A novel deep feature enhancement mechanism using time-frequency analysis is presented. By enabling dual interaction between time-frequency and deep feature spaces, this mechanism significantly improves deep feature representation for few-shot learning tasks. To the best of our knowledge, this is the first approach to apply an interaction strategy for deep feature enhancement with time-frequency analysis.
2) Theoretical Guarantee: An upper bound on the generalization error of TFE is derived, providing a theoretical foundation for the reliability of the interaction strategy. This bound also highlights the critical factors influencing TFE's performance, specifically the aggregation degree of samples in the two feature spaces. Notably, this is the first study to analyze generalization error for deep feature enhancement.

The remainder of this paper is organized as follows. Section II provides a comprehensive analysis of related work. Section III describes the proposed TFE mechanism and the complete anomaly detection transfer learning model. Section IV presents a theoretical analysis of the reliability of the TFE mechanism. Section V discusses the empirical evaluation results, and the paper concludes in Section VI.

II. Preliminary Works

Traditional one-class anomaly detection methods (e.g. iForest [7], OC-SVM [8], and SVDD [9]) often rely on handcrafted features. For instance, SVDD aims to find the smallest hypersphere that contains most normal data in a high-dimensional space, using the hypersphere boundary as the detection rule. Recently, deep learning has provided new solutions with its end-to-end feature extraction capabilities [10]–[13]. Ruff et al. [14] introduced the Deep SVDD algorithm, integrating classical SVDD with a deep convolutional neural network. To address the challenge of insufficient training samples, transfer learning techniques, such as fine-tuning (e.g., Baireddy et al. [15], Bergmann et al. [16], Han et al. [17], Subramanian et al. [18], Swati et al. [19], and Zhou et al. [20]), and domain-adversarial training strategies (e.g., Farahani et al. [21], Han et al. [22], Ma et al. [23], Ragab et al. [24], Shermin et al. [25], Wang et al. [26], and Yang et al. [27]), have been employed to facilitate knowledge transfer. While these methods have achieved promising results, they remain sensitive to low-quality target domain data. Mao et al. [28] tackled this issue by developing a deep domain-adversarial anomaly detection (DAAD) model that uses hypersphere adversarial training to transfer detection rules. However, this approach still relies heavily on large-scale source domain data, and their performance degrades significantly when source data is limited or small-scale.

Feature enhancement techniques, such as GANs (e.g., Bowles et al. [29], Frid-Adar et al. [30], Goodfellow et al. [31], Kang et al. [32], Liu et al. [33], and Tran et al. [34]), diffusion models (e.g., Ho et al. [35]), and time-frequency analysis methods, including wavelet and Fourier transforms, are commonly used to reduce data dependency and address data scarcity concerns. However, GAN-based feature generation relies heavily on large training datasets. In recent years, wavelet transform has gained popularity as a powerful tool for feature enhancement in few-shot scenarios due to its robust mathematical foundation [36]. Studies such as Behmanesh et al. [37], Bouny et al. [38], Fu et al. [39], Mishra et al. [40], and Pan et al. [41] have incorporated wavelet transform into deep neural networks to extract key features. However, as wavelet transform function is a fixed filter, it inevitably causes some degree of information loss, which highlights a significant opportunity for further advancements in feature enhancement techniques for few-shot learning.

Interpretability analysis, which explains the decision-making process, is crucial for improving the trustworthiness of learning results. It typically focuses on aspects such as sample saliency (e.g., Tirinzoni et al. [42] and Zhang et al. [43]), feature significance (e.g., Bugata et al. [44], Jeyasothy et al. [45], and Wu et al. [46]), and network structure (e.g., Dhebar et al. [47], Singh et al. [48], and Zeiler et al. [49]). In transfer learning, transferability analysis seeks to elucidate transfer strategies, thereby enhancing the interpretability of the transfer process [50], [51]. In few-shot learning, the reliance on large datasets reduces result reliability due to insufficient data and poor feature extraction. Time-frequency analysis, capable of decomposing signal data, offers potential for extracting and transferring valuable information from a frequency perspective. To the best of our knowledge, no existing research has explored transferability analysis based on frequency saliency.

III. Methodology

This section provides a detailed explanation of the proposed TFE mechanism. To demonstrate its training process, a few-shot transfer learning model for anomaly detection incorporating TFE is developed, as illustrated in Fig. 3. The detailed implementation steps are outlined below.

A. Problem description

Assume that the source data $X^S = \left\{x^n_{i\times j}\right\}_{n=1}^{M}$ contains M normal samples, where the superscript S denotes the source domain. The target data $X^T = \left\{x^n_{i\times j}\right\}_{n=1}^{N}$ contains N normal samples, where the superscript T indicates the target domain. The test data $X^{T_{test}} = \left\{x^n_{i\times j}\right\}_{n=1}^{Test}$ originates from the target domain and includes both normal and anomalous samples.

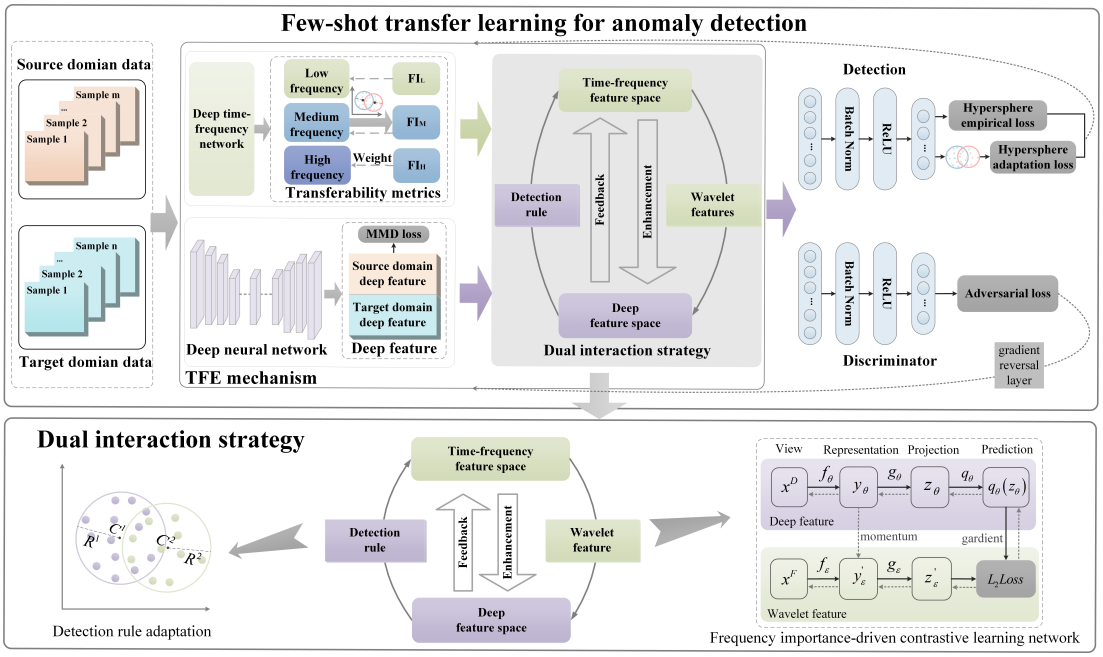


Fig. 3. Structure diagram of the few-shot transfer learning for anomaly detection model with the TFE mechanism. Running with domain adversarial training, the model integrates the proposed TFE mechanism into feature extractor. The *FI* metric guides purposeful knowledge transfer in terms of frequency importance.

The anomaly detection problem addressed in this paper adheres to the following conditions:

1) The source domain $D^S = \{X^S, P(X^S)\}$ consists of X^S and its corresponding distribution $P(X^S)$.
2) The target domain $D^T = \{X^T, P(X^T)\}$ consists of X^T and its corresponding distribution $P(X^T)$.
3) The data distributions are distinct, i.e., $P(X^S) \neq P(X^T)$, but the anomaly detection rules in the source and target domains share inherent similarities.

Transfer learning for anomaly detection seeks to leverage the normal data from X^S and X^T to train an anomaly detection model capable of identifying anomalies in $X^{T_{test}}$. In this context, X^T is small-scale, posing challenges for transfer learning. By transferring discriminative information from X^S, the detection performance on $X^{T_{test}}$ can be enhanced. However, in few-shot scenarios, X^S is also insufficient. Therefore, this paper addresses the transfer learning problem for anomaly detection where both X^S and X^T are small-scale.

B. Proposed TFE mechanism

The TFE mechanism consists of three main components: feature extraction, enhancement operation, and feedback operation. The overall algorithmic framework is summarized as follows. 1) First, X^S and X^T are simultaneously input into a deep neural network $f(\cdot, W_D)$ and a deep time-frequency network $f(\cdot, W_T)$, which generate a deep feature set X_d and a wavelet feature set X_w for the two domains. Here, $f(\cdot, W_D)$ can be any deep feature extractor, while $f(\cdot, W_T)$ is constructed by cascading convolutional filters and WST. 2) Next, X_d and X_w are fed into the FICL network for the enhancement operation, producing enhanced deep features X_{d_e} and enhanced wavelet features X_{w_e} respectively. These enhanced features are then passed to the detection rule adaptation module to perform the feedback operation. 3) The enhancement and feedback operations run alternatively until the deviation between the two feature sets falls below a predefined threshold. The detailed implementation process is introduced in the following sections.

1) Feature extraction: Without loss of generality, a deep convolutional network is utilized as the deep feature extractor, denoted as $f(\cdot, W_D) : X \to F \in \mathbb{R}^m$, where W_D represents the network weights. Using $f(\cdot, W_D)$, the deep feature set $X_d = \{X_d^S, X_d^T\}$ is obtained, where X_d^S and X_d^T correspond to the deep features from the source and target domains, respectively.

The learnable deep time-frequency network $f(\cdot, W_T)$ is designed to extract $X_w = \{X_w^S, X_w^T\}$, where X_w^S and X_w^T are the wavelet features from the source and target domains, respectively. WST is utilized to decompose the deep features obtained from convolutional filters into a full frequency band. Compared to classical wavelet transform, WST provides superior properties, including norm preservation, non-expansion, translation invariance, and shape stability [6]. These characteristics allow WST to produce a complete and physically-aware feature set, enhancing classical deep features and benefiting few-shot feature representation. Due to space limitation, the detailed implementation of WST is provided in Appendix.

Although multiple WST kernels can be cascaded to extract fine-granularity wavelet features, the energy of wavelet scattering coefficients diminishes as the network depth increases, eventually approaching zero. Consequently, this paper employs a cascaded three-layer WST network, as the energy iteration for three layers achieves over 98% [6]. The output of each layer is approximately defined as a frequency band. Thus, X_w can be divided into three frequency bands: low frequency, medium frequency and high frequency, i.e., $X_w^S =$

$\{X_L^S, X_M^S, X_H^S\}$, $X_w^T = \{X_L^T, X_M^T, X_H^T\}$ for the source and target domains, respectively.

2) Enhancement operation:

a) Frequency Importance Metric: With the frequency bands specified, the wavelet features become physically-aware. To quantify the significance of each frequency band in the transfer process, this paper introduces a new *FI* metric based on a frequency hypersphere matching strategy. The construction of *FI* is rooted in the principle of prioritizing the transition of anomaly detection rules over feature adaptation for anomaly detection tasks. In few-shot scenarios, feature representations may vary significantly in terms of edges, textures, shapes, and other attributes, making anomaly detection rules (e.g., the SVDD hypersphere) more effective in capturing intrinsic patterns that distinguish normal from anomalous samples. To implement this, we propose a frequency hypersphere matching strategy that quantifies the deviation between the hyperspheres of different domains within the same frequency band.

Here, we use the low-frequency band as an example to illustrate the specific implementation. Hyperspheres are constructed for X_L^S and X_L^T in the high-dimensional space respectively, with their center and radius denoted as $C_L^S, C_L^T, R_L^S, R_L^T$. By calculating the matching divergence between the two hyperspheres, the *FI* metric of low frequency band can be calculated as:

$$FI_L = \frac{R_L^S + R_L^T}{\left\|C_L^S - C_L^T\right\|^2 + \left|R_L^S - R_L^T\right|} \tag{1}$$

where $\|\cdot\|$ represents l_2-norm, $|\cdot|$ is absolute value.

In (1), the denominator is designed to measure the disparity between the two hyperspheres, while the numerator is the radius of the two hyperspheres. Since different frequency bands pose divergent energy in WST, the radius sum is used to normalize the disparity for mitigating the magnitude difference. A reciprocal operation is further carried out for measuring the similarity between the two domains.

Similarly, the *FI* metrics for the medium and high-frequency bands, FI_M and FI_H, can also be computed. A higher *FI* value indicates greater importance of the corresponding frequency band, and vice versa.

$\{FI_i\}_{i=L,M,H}$ is linearly scaled to produce valid weights, as follows:

$$\widetilde{FI_i} = \frac{FI_i}{FI_L + FI_M + FI_H} \tag{2}$$

Clearly, the scaled weights satisfy $\widetilde{FI_L} + \widetilde{FI_M} + \widetilde{FI_H} = 1$.

b) Frequency importance-driven contrastive learning network: Using the weights $\left\{\widetilde{FI_i}\right\}_{i=L,M,H}$, purposeful alignment between X_d and X_w is achieved based on the importance of each frequency band. Given that wavelet features obtained through WST are theoretically guaranteed to be complete, this alignment facilitates the full supplementation of discriminative information from X_w to X_d. To implement this purposeful alignment in an unsupervised manner, the classic contrastive learning network BYOL [52] is adopted as the foundational architecture. The specific implementation of FICL is outlined as follows.

First, the *FI*-weighted wavelet feature set X_{w_FI} is obtained by means of $\left\{\widetilde{FI_i}\right\}_{i=L,M,H}$:

$$\begin{cases} X_{w\_FI} = \left\{X_{w\_FI}^S, X_{w\_FI}^T\right\} \\ X_{w\_FI}^S = \widetilde{FI_L} * X_L^S + \widetilde{FI_M} * X_M^S + \widetilde{FI_H} * X_H^S \\ X_{w\_FI}^T = \widetilde{FI_L} * X_L^T + \widetilde{FI_M} * X_M^T + \widetilde{FI_H} * X_H^T \end{cases} \tag{3}$$

Second, X_d and X_{w_FI} are fed into the BYOL network to generate the enhanced feature X_{d_e} and X_{w_e}. The loss of FICL can be calculated as follows:

$$L_{FICL} = \left\|q_\theta\left(f_\theta\left(X_d, W_\theta\right)\right) - f_\varepsilon\left(X_{w\_FI}, W_\varepsilon\right)\right\| \tag{4}$$

where W_θ and W_ε represent the weights of the deep and wavelet layers, respectively, while q_θ is a regression mapping.

By minimizing L_{FICL}, the divergence between X_d and X_w is progressively reduced. Consequently, the deep features are effectively enhanced by the wavelet features through the FICL mechanism.

3) Feedback operation: Although the wavelet features extracted by WST are mathematically complete, they do not account for data-specific characteristics and are therefore suboptimal for anomaly detection tasks. To address this, the wavelet features need to be updated based on the learned detection rule in the deep feature space. In this section, we design a detection rule adaptation module to enable the feedback operation. This module allows discriminative information in deep features to be transferred to the deep time-frequency network, thereby updating the wavelet features. Specifically, X_{d_e} and X_{w_e} are mapped to an RKHS space using a Gaussian kernel function to construct their respective SVDD hyperspheres, which serve as the detection rule. Let the hypersphere centers and radii for the deep features and wavelet features be denoted as C_d, C_w, R_d, R_w, respectively. The detection rule adaptation loss is then calculated as:

$$L_{adapt} = \left\|C_d - C_w\right\|^2 + \left|\left(R_d\right)^2 - \left(R_w\right)^2\right| \tag{5}$$

Intuitively, the detection rule adaptation can be visualized as the overlapping of the two hyperspheres in the RKHS space. This ensures that the wavelet features are effectively updated to align with the detection rule learned in the deep feature space.

C. Few-shot transfer learning for anomaly detection

It is worth noting that the enhancement operation and feedback operation need to run alternatively until the two feature sets remain identical. Such iterative training should be placed in the training process of the whole network. Therefore, we first build a transfer learning for anomaly detection model and then present the training process of TFE.

We adopt one of our previous works, the deep domain-adversarial anomaly detection method (DAAD for short) [28], as the base framework. Since DAAD relies on sufficient source domain data for model training, this paper applies the TFE mechanism into the feature extraction part of DAAD to realize a few-shot learning task. With domain-adversarial training, the loss function of DAAD [28] is:

$$\min_{W_f, R^S, R^T} \max_{W_{DC}} L_D = L_C + L_{DC} + \lambda L_{HA} + \mu L_{MMD} \tag{6}$$

where:

$$\begin{aligned} L_C(R, W_\theta) =& (R^S)^2 + (R^T)^2 \\ &+ \frac{1}{M}\sum_{i=1}^{M}\max\left\{0, \left\|g_i^S - C^S\right\|^2 - (R^S)^2\right\} \\ &+ \frac{v}{N}\sum_{j=1}^{N}\max\left\{0, \left\|g_j^T - C^T\right\|^2 - (R^T)^2\right\} \\ L_{DC} =& \mathbb{E}_{x\sim X^S}\left[\log DC(X_{d\_e}^S)\right] \\ &+ \mathbb{E}_{x\sim X^T}\left[\log(1 - DC(X_{d\_e}^T))\right] \\ L_{HA} =& \left\|C^S - C^T\right\|^2 + \left|(R^T)^2 - (R^S)^2\right| \end{aligned} \tag{7}$$

In (7), L_C is the hypersphere empirical loss, L_{DC} is domain-adversarial loss, L_{HA} is the hypersphere adaptation loss, L_{MMD} is MMD loss. Minimizing L_C is devoted to establishing proper detection rules based on the two hyperspheres. Maximizing L_{DC} makes the domain discriminator unable to distinguish whether the sample comes from the source domain or the target domain. Minimizing L_{HA} pushes the two hyperspheres to be geometrically overlapped, then realizing rule adaptation. Minimizing L_{MMD} dedicates to feature alignment in an RKHS space.

After integrating the TFE mechanism, the deep convolutional network within TFE serves as the feature extractor in DAAD to construct few-shot transfer learning for anomaly detection model, referred to as few-shot DAAD. The updated loss function L_D is defined as:

$$\min_{W_D, W_T, W_\theta, W_\varepsilon, R^S, R^T}\max_{W_{DC}} L_{F\_D} = L_{FICL} + L_{adapt} + L_D \tag{8}$$

With the model parameters determined, anomaly detection is performed by calculating the distance between the test sample x^{Test} and the center of the hypersphere in the target domain, as follows:

$$Score\left(x^{Test}\right) = \left\|f\left(x^{Test}, W_D\right) - C^T\right\| - R^T \tag{9}$$

If $Score(x^{Test}) > 0$, the test sample x^{Test} is classified as an anomalous sample.

D. Training algorithm

The training process of (8) involves several network weights: W_θ and W_ε from the FICL network, W_T from the deep time-frequency network in the TFE mechanism, and W_D, W_{DC}, and $R = \left\{R^S, R^T\right\}$ from DAAD. Since the feature extractor in DAAD has been replaced by the deep convolutional network in the TFE mechanism, the feedback operation and domain-adversarial training in DAAD can be trained jointly. However, the enhancement operation (i.e., the FICL network) must be trained separately.

Given the coupling of these weights in solving (8), an alternating training strategy is adopted. The training proceeds as follows. First, fix $W = \{W_D, W_T, W_{DC}\}$ and $R = \left\{R^S, R^T\right\}$, train W_θ and W_ε. Then, fix W_θ and W_ε, train W and R. These two steps are alternated iteratively until the entire model converges.

To optimize W_θ and W_ε, the stochastic gradient descent (SGD) is utilized to update W_θ and $q_\theta\left(Q_\theta\right)$, as follows:

$$\begin{cases} (W_\theta^*, Q_\theta^*) = \arg\min_{W_\theta, Q_\theta} L_{FICL} \\ W_\theta \leftarrow W_\theta - \eta\dfrac{L_{FICL}}{\partial W_\theta} \\ Q_\theta \leftarrow Q_\theta - \eta\dfrac{L_{FICL}}{\partial Q_\theta} \end{cases} \tag{10}$$

where η is the step length. Being identical to the classical BYOL, W_ε is further updated by using exponential moving average (EMA) strategy:

$$W_\varepsilon \leftarrow \mu W_\varepsilon + (1-\tau)\, W_\theta^* \tag{11}$$

where $\tau \in [0, 1]$ represents the weight decay rate.

For W and R, the solving process is decomposed as follows:

$$\begin{aligned} \left(W_D^*, W_T^*, R^{S*}, R^{T*}\right) =& \arg\min_{W_D, W_T, R^S, R^T}(L_{adapt} + L_C + \lambda L_{HA} \\ &+ \mu L_{MMD}) \\ \left(W_{DC}^*\right) =& \arg\max_{W_{DC}}(L_{DC}) \end{aligned} \tag{12}$$

In (12), the gradient update is:

$$\begin{aligned} W_D^* &\leftarrow W_D - \eta\left(\frac{\partial L_C}{\partial W_D} + \lambda\frac{\partial L_{HA}}{\partial W_D} + \mu\frac{\partial L_{MMD}}{\partial W_D} - \frac{\partial L_{DC}}{\partial W_D}\right) \\ W_T^* &\leftarrow W_T - \eta\frac{\partial L_{adapt}}{\partial W_T} \\ W_{DC}^* &\leftarrow W_{DC} - \eta\frac{\partial L_{DC}}{\partial W_{DC}} \\ R^{S*} &\leftarrow R^S - \eta\left(\frac{\partial L_C}{\partial R^S} + \lambda\frac{\partial L_{HA}}{\partial R^S}\right) \\ R^{T*} &\leftarrow R^T - \eta\left(\frac{\partial L_C}{\partial R^T} + \lambda\frac{\partial L_{HA}}{\partial R^T}\right) \end{aligned} \tag{13}$$

In (13), we have

$$\begin{aligned} \frac{\partial L_C}{\partial R^S} &= 2\left(1 - \frac{n_{out}^S}{n^S}\right)R^S, \frac{\partial L_{HA}}{\partial R^S} = 2R^S \\ \frac{\partial L_C}{\partial R^T} &= 2\left(1 - \frac{n_{out}^T}{vn^T}\right)R^T, \frac{\partial L_{HA}}{\partial R^T} = 2R^T \end{aligned} \tag{14}$$

where n_{out}^S and n_{out}^T are the numbers of samples located outside the source and target domain hyperspheres respectively.

The entire training process is summarized in Algorithm 1.

E. Analysis of Computational Complexity

For the domain-adversarial training architecture in Algorithm 1, the time complexity primarily involves the feature extractor with the TFE mechanism, the label classifier, and the domain discriminator. The TFE mechanism itself consists of a deep neural network, a deep time-frequency network, an FICL network, and a detection rule adaptation module. In the deep time-frequency network, the wavelet kernel of WST is fixed and does not participate in the training process. Therefore, its time complexity depends solely on the deep convolutional layers. The FICL network comprises a deep convolutional network and a fully connected (FC) layer. Similarly, the label classifier and domain discriminator are constructed with FC layers, but the label classifier additionally involves solving an SVDD classifier based on the FC layer's output. Furthermore,

Algorithm 1 Training process of few-shot DAAD with the TFE mechanism.

Input: Training sample sets $X^S = \{x_{i\times j}^n\}_{n=1}^M$ and $X^T = \{x_{i\times j}^n\}_{n=1}^N$, test sample set $X^{T_{test}} = \{x_{i\times j}^n\}_{n=1}^{Test}$, the regularization parameter λ and μ, and the learning rates η, τ;

Output: Anomaly score of test sample set $X^{T_{test}}$;

Initialize randomly the network weight W_D, W_T, W_{DC}, W_θ, W_ε and calculate the starting values of hypersphere centers C^S, C^T and radiuses R^S, R^T;
for $epoch = 1$ to $epoch\_max$ **do**
  for $batch\_id = 1$ to $batch\_max$ **do**
    Extract the deep feature set X_d and the wavelet feature set X_w by feeding X^S and X^T into $f(W_D)$ and $f(W_T)$;
    Calculate the importance metric $\left\{\widetilde{FI}_i\right\}_{i=L,M,H}$ by (1-2);
    if (the change of L_{FICL} is greater than a threshold between two consecutive steps) **then**
      Update W_θ and W_ε by feeding X_d, X_w, $\left\{\widetilde{FI}_i\right\}_{i=L,M,H}$ into (3-4) and (10-11);
    end if
    Calculate the total loss L_{F_D} by (8);
    Update W_D, W_T, W_{DC}, R^S and R^T by (12-14);
  end for
end for
Get the optimal network weight $W_D{}^*$, $W_T{}^*$, $W_{DC}{}^*$, $W_\theta{}^*$, $W_\varepsilon{}^*$, R^{S*}, R^{T*};
Obtain the anomaly score of $X^{T_{test}}$ by (9);

the detection rule adaptation module incurs computational cost due to the determination of the SVDD boundary. Consequently, the overall time complexity of Algorithm 1 is determined by the costs of convolutional operations, FC layer operations, and solving the SVDD boundary.

From [53], the time complexity of a convolution operation is $O\left((M+N)P^2K^2C_{in}C_{out}\right)$, where $(M+N)$ is the total number of training samples, P and K are the sizes of the input features and convolution kernel, respectively, and C_{in} and C_{out} denote the number of input and output channels, respectively. The time complexity of an FC layer operation is $O\left(L(M+N)F_{in}F_{out}\right)$, where L is the number of FC layers, and F_{in} and F_{out} are the feature dimensions of input and output respectively of the FC layer. Due to down-sampling, the input to the FC layer typically has a much smaller dimension than the output of the convolution operation, making the computational cost of the FC layer negligible. According to [9], determining the SVDD boundary is equivalent to solving a convex quadratic programming problem, which has a time complexity of approximately $O\left((M+N)^2\right)$. Thus, the total time complexity of Algorithm 1 can be approximated as $O\left((M+N)P^2K^2C_{in}C_{out}\right) + O\left((M+N)^2\right)$.

IV. Reliability Analysis

Since the TFE mechanism relies on exchanging discriminative information between the time-frequency and deep feature spaces, there is a concern about potential information loss during the dual interaction. A risk to TFE's reliability arises if the amount of retained discriminative information is uncertain. Before conducting the empirical evaluation in Section V, it is essential to perform a theoretical analysis of the upper bound of classification error for the TFE mechanism.

For simplicity, we approximate the SVDD classifier as $G_f(x) = \|f(x) - c\| - R$, where $c = E_{x\in X}[f(x)]$ and R are the center and radius of hypersphere, respectively. The model performance is quantified using the classification error. When $G_f(x) > 0$, the sample lies outside the hypersphere, leading to a misclassification. Hence, the classification error can be defined as the probability that a normal sample lies outside the hypersphere:

$$Err(G_f) = \sum_{i=1}^{M+N} P[G_f(x_i) > 0] \tag{15}$$

To facilitate the derivation, we simplify the anomaly detection task (i.e., DAAD in Section III-C) as a downstream task with its feature extraction network denoted as f, where $\|f\| = r$. Here, $\|\cdot\|$ represents the l_2-norm for vectors or the Frobenius norm for matrices, as appropriate. Inspired by reference [54], we provide the following definition for wavelet features.

Definition 1. *(σ_k-Augmentation). Given $\sigma_k \in (0,1]$, for the feature set X_k extracted from the k-th frequency band of the original data X, if there exists a subset $X_k^0 \subseteq X_k$ such that $P[x \in X_k^0] \geq \sigma_k P[x \in X_k]$ holds, X_k is called σ_k-Augmentation of X.*

From Definition 1, σ_k represents the aggregation degree of wavelet features. A larger subset X_k^0 results in a higher value of σ_k, indicating a more clustered distribution of features in X_k. Samples with distributions that are highly clustered in both the time-frequency and deep feature spaces, i.e., $X_L^0 \cap X_M^0 \cap X_H^0 \cap S_\delta$, are strongly related to the detection performance. Here, $S_\delta := \{x \in X_d : \forall x_1, x_2 \in X_d(x),\ \|f(x_1) - f(x_2)\| \leq \delta\}$ defines the set of samples in X_d with pairwise distances bounded by δ. These samples should exhibit aggregation across all three frequency bands and possess deep features with strong discriminative capabilities. Furthermore, as discussed in Section III-B2, the importance of each frequency band in the transfer process is expressed by $FI = [FI_L, FI_M, FI_H]$. Based on this, we present the following lemma.

Lemma 1. *If the samples belonging to $\left(FI_L X_L^0\right) \cap \left(FI_M X_M^0\right) \cap \left(FI_H X_H^0\right) \cap S_\delta$ can be classified correctly by SVDD, the upper bound of classification error is given by $(1 - FI_L \times \sigma_L - FI_M \times \sigma_M - FI_H \times \sigma_H) + P\left[\overline{S_\delta}\right]$.*

Proof. See the proof in Appendix.

Lemma 1 establishes a sufficient condition for determining the upper bound of the classification error for the SVDD model with the TFE mechanism. Based on Lemma 1, we further analyze when the upper bound holds and derive the following.

Lemma 2. *If the radius of SVDD hypersphere satisfies* $R > r\sqrt{\left(1-\frac{\delta P\left[\overline{S_\delta}\right]}{r}+\frac{\delta\alpha}{r}\right)}$, *then the samples* $x_0 \in \left\{\left(FI_L X_L^0\right) \cap \left(FI_M X_M^0\right) \cap \left(FI_H X_H^0\right) \cap S_\delta\right\}$ *can be correctly classified, where* $\alpha = FI_L \times \sigma_L + FI_M \times \sigma_M + FI_H \times \sigma_H$.

Proof. See the proof in Appendix.

Lemma 2 provides a sufficient condition for correctly classifying the targeted samples. Using Lemmas 1 and 2, we derive the error bound for the TFE mechanism, as follows.

Theorem 1. *Given the wavelet feature set based on* σ_k*-Augmentation in the TFE mechanism, if*

$$R > r\sqrt{\left(1-\frac{\delta P\left[S_\delta\right]}{r}+\frac{\delta\alpha}{r}\right)} \tag{16}$$

then the classification error satisfies:

$$Err(G_f) \leq (1-\alpha) + P\left[S_\delta\right] \tag{17}$$

where $\alpha = FI_L \times \sigma_L + FI_M \times \sigma_M + FI_H \times \sigma_H$.

Theorem 1 demonstrates that the classification error of the TFE mechanism is theoretically constrained by a determined upper bound, provided the hypersphere radius is not too small (e.g., avoiding hypersphere collapse [14]). Moreover, Theorem 1 highlights two key factors from (17) that influence the classification error upper bound:

1) *Aggregation Degree in Deep Feature Space* The first factor is $P\left[S_\delta\right]$ which represents the aggregation degree of samples in the deep feature space. A higher aggregation degree of normal samples with deep features results in a smaller value of $P\left[S_\delta\right]$. In the TFE mechanism, by incorporating time-frequency information, the deep features of normal samples become more discriminative with respect to anomalies, achieving greater geometric aggregation in the feature space. Consequently, $P\left[S_\delta\right]$ is reduced.
2) *Aggregation Degree in Wavelet Feature Space* The second factor is α, which reflects the aggregation degree of wavelet features. A higher aggregation degree of normal samples with wavelet features corresponds to a larger value of σ_k. As discussed in Section III-B3 , the feedback operation employs the detection rule adaptation strategy to update the deep time-frequency network. The wavelet features are encouraged to become more concentrated within each frequency band. This concentration increases the value of σ_k, thereby reducing the upper bound on the classification error.

To better understand Theorem 1, consider an extreme case where $\sigma_L = 1, \sigma_M = 1, \sigma_H = 1$, which implies that normal samples are highly aggregated in all three frequency bands, meaning the wavelet features exhibit excellent discriminative capability. In this scenario, the wavelet features significantly outperform the deep features, and the upper bound of classification error becomes solely dependent on $P\left[S_\delta\right]$. In this case, minimizing $P\left[S_\delta\right]$ can be achieved exclusively through the enhancement operation in the TFE mechanism. This analysis demonstrates that (17) provides a theoretical and robust reliability guarantee for the proposed TFE mechanism. For few-shot learning tasks, the theoretical analysis serves as a strong complement to the empirical results that will be provided in the next section. Moreover, it reveals the key factors influencing the model's generalization capability, being helpful to understand the operational mechanism of the proposed dual interaction strategy.

V. Experiments

In this section, two typical anomaly detection experiments, namely, image recognition detection and early fault detection of rolling bearings, are conducted to evaluate the effectiveness of the proposed TFE mechanism. The *image recognition detection* experiment includes handwritten character and object recognition anomaly detection, serving as baseline experiments to intuitively validate the TFE mechanism. The *early fault detection of rolling bearings* experiment is designed to assess the TFE mechanism in real engineering scenarios. The experiments are conducted in a programming environment comprising Matlab2019a and Python3.7, using a computer configured with an Intel i5-8400 CPU and an NVIDIA REX1080Ti GPU.

A. Handwritten character anomaly detection

1) Experimental settings: The widely-used MNIST and USPS [55] datasets, which consist of handwritten digits from 0 to 9, are selected for the transfer learning task. Compared to MNIST, the image quality in USPS is significantly lower, as shown in Fig. 4. Accordingly, MNIST is used as the source domain with 100 clear images per digit, while USPS serves as the target domain with 25 lower-quality images per digit. Ten experiments are designed, each focusing on one digit from 0 to 9. In each experiment, a specific digit (e.g., 0) is chosen as the normal class, and images of this digit from both domains are used to train the model. New images of the same digit from the target domain are used for testing. To simulate anomalies, a different digit (e.g., 1) is selected as the anomaly class.

Fig. 4. Introduction to the MNIST∼USPS.

2) Experimental results: In this section, eight representative anomaly detection methods are introduced for comparative analysis to verify the effectiveness of the TFE mechanism. These methods are categorized into three groups: shallow models, deep models, and deep transfer learning models.

1) *Shallow Models:* OC-SVM [8] and SVDD [9] use kernel functions to detect anomalies in high-dimensional spaces, while iForest [7] is an unsupervised anomaly detection algorithm based on data segmentation.
2) *Deep Models:* Deep SVDD [14] is trained with only target domain data, referred to as Deep SVDD (1), and Deep SVDD is trained with both source and target domain data, referred to as Deep SVDD (2).

3) *Deep Transfer Learning Models:* DCAE (pre-train) [20] utilizes autoencoders for pre-training before fine-tuning Deep SVDD for anomaly detection. SSL (pre-train) [56] employs self-supervised learning for pre-training before fine-tuning Deep SVDD. DAAD [28] offers a domain-adversarial anomaly detection method that adapts detection rules through domain-adversarial training.

The comparison results in Table I demonstrate that the TFE mechanism achieves the highest accuracy across all classes. This indicates that the effective interaction between wavelet and deep features significantly enhances the transferability of deep models, enabling them to capture discriminative information more effectively in few-shot scenarios.

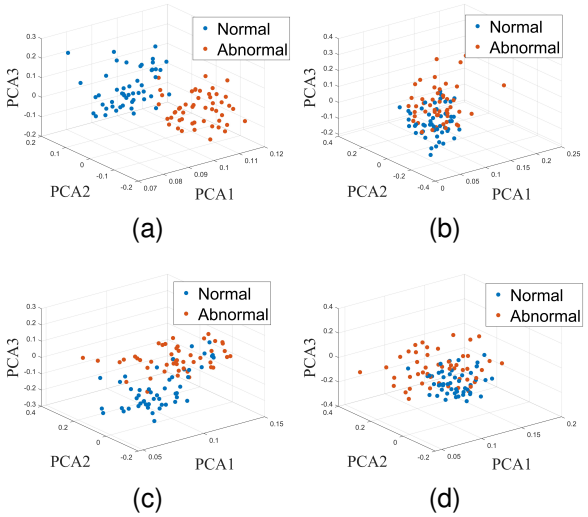


Fig. 5. Visual detection results on character 1 by (a) the proposed TFE mechanism, (b) Deep SVDD (2), (c) SSL (pre-train), (d) DAAD. Here the test data are from the USPS dataset, containing 50 images of character 1 and 50 images of character 2. Clearer separation indicates stronger discriminative capability, as shown in (a).

To provide an intuitive comparison, Fig. 5 displays the visual detection results for character 1 using three state-of-the-art (SOTA) methods. The proposed TFE mechanism produces a clearer decision boundary between normal and anomalous classes compared to the other methods. This improvement is due to the reliance of the compared methods on large data volumes for effective model training. In this experiment, the training set consisted of 100 images in the source domain and 25 images in the target domain, which was insufficient for reliable feature extraction, resulting in biased detection models. In contrast, the TFE mechanism leverages the time-frequency feature space to clarify the importance of information for transfer learning and optimizes deep features on small-scale samples through its interactive strategy.

B. Object recognition anomaly detection

1) Experimental settings: The Office-Home dataset [57] contains diverse object images from real-world scenarios. For this experiment, "bike" is selected as the normal class (as shown in Fig. 6). Specifically, 40 bike images from the "Produce" category are chosen as the source domain, while 20 bike images from the "Clip" category are used as the target domain. To simulate anomalies, images from other categories (e.g., pan, hammer, scissors) are included in the test data as anomaly classes.

Fig. 6. Instances of the bike images in the Office-Home dataset.

2) Experimental results: Fig. 7 illustrates the feature distribution before and after transfer using the proposed few-shot DAAD with the TFE mechanism. For comparison, the results of a classical domain adversarial neural network (DANN) are also shown. The results indicate that DANN struggles to adapt features effectively, highlighting the limitations of classical adversarial training in ensuring reliable domain knowledge transfer in few-shot tasks. In contrast, the proposed few-shot DAAD demonstrates smaller differences in feature distributions and probability density curves, showcasing its ability to successfully transfer discriminative information.

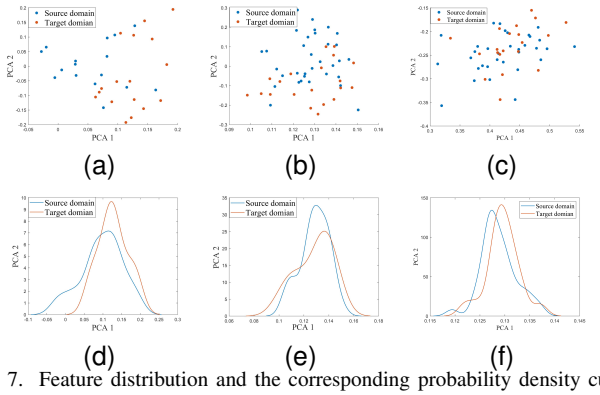


Fig. 7. Feature distribution and the corresponding probability density curve of the two domains' data, where (a) and (d) are before the transfer, (b) and (e) are of the transfer by the classical DANN, (c) and (f) are of the transfer by the proposed few-shot DAAD.

Table II presents the detection results for the "bike" class. The proposed TFE mechanism achieves the highest accuracy among all methods. Similarly to the results in Table I, these experimental findings further validate the strong performance of the TFE mechanism in few-shot tasks.

C. Early fault detection of rolling bearings

1) Experimental settings: The public IEEE PHM 2012 dataset [58] (referred to as PHM) and a real-world industrial dataset from a large steel factory in China are selected for this study. The PHM dataset contains complete life-cycle data for 17 bearings under three different operating conditions: 1800 r/min, 1650 r/min, and 1500 r/min, with corresponding loads of 4000 N, 4200 N, and 5000 N. The real-world industrial dataset (referred to as FWM) was collected from the main motor bearings FAG 23056-B-MB of an industrial fixed-width machine using the factory's official health monitoring system. Each sampling lasts 3s, and the sample frequency is 80.9 kHz.

TABLE I
CLASSIFICATION ACCURACY OF MNIST∼USPS.

Normal class	Anomalous class	OC-SVM	SVDD	iForest	Deep SVDD (1)	Deep SVDD (2)	DCAE (pre-train)	SSL (pre-train)	DAAD	Our
0	1	0.40	0.59	0.48	0.64	0.57	0.52	0.65	0.60	**0.84**
1	2	0.80	0.38	0.61	0.69	0.70	0.71	0.79	0.78	**0.89**
2	3	0.70	0.36	0.69	0.63	0.65	0.78	0.62	0.71	**0.76**
3	4	0.58	0.49	0.48	0.59	0.64	0.63	0.65	0.64	**0.72**
4	5	0.76	0.54	0.64	0.57	0.63	0.62	0.73	0.60	**0.78**
5	6	0.67	0.52	0.51	0.52	0.51	0.57	0.68	0.46	**0.72**
6	7	0.59	0.56	0.47	0.59	0.71	0.61	0.59	0.58	**0.73**
7	8	0.69	0.40	0.66	0.62	0.59	0.66	0.72	0.69	**0.77**
8	9	0.47	0.48	0.48	0.55	0.57	0.58	0.53	0.58	**0.70**
9	0	0.59	0.48	0.58	0.56	0.51	0.67	0.56	0.67	**0.68**

TABLE II
CLASSIFICATION ACCURACY ON BIKE CLASS IN THE OFFICE-HOME DATASET.

Normal class	Anomalous class	OC-SVM	SVDD	iForest	Deep SVDD (1)	Deep SVDD (2)	DCAE (pre-train)	SSL (pre-train)	DAAD	Our
Bike	Pan, Hammer, Scissors	0.47	0.53	0.56	0.58	0.52	0.54	0.61	0.64	**0.74**

The FWM dataset spans operational data from June 2022 to April 2024. The two datasets' details are shown in Fig. 8.

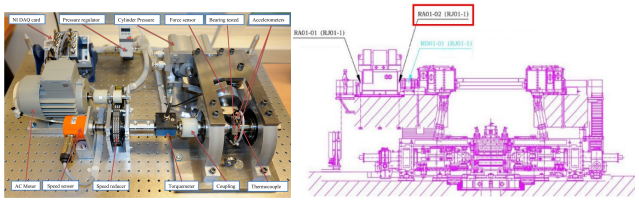


Fig. 8. Schematic of the testbed: (a) PHM dataset; (b) FWM dataset. For reasons of commercial confidentiality, the physical pictures of FWM will no longer be provided here.

To simulate a few-shot scenario, two experimental tasks were set up, as detailed in Table III. The 500 initial normal samples from Bearing 1_1 have been verified as normal in [59].

TABLE III
EXPERIMENTAL SETTING FOR THE EARLY FAULT DETECTION OF ROLLING BEARINGS.

Task	Domain	Bearing	Number
Task1	Source domain	PHM (Bearing1_1)	500
	Target domain	PHM (Bearing2_1)	100
Task2	Source domain	PHM (Bearing1_1)	500
	Target domain	FWM	100

2) Experimental results of Task 1:

a) Experimental results: Fig. 9 illustrates the feature distribution and probability density curves for the source and target domain data before and after transfer. The results show that the feature distributions and probability density curves of the two domains align closely after applying the proposed transfer learning model with the TFE mechanism. This alignment is attributed to the effective fusion of wavelet and deep features, which enhances the representation of detection rules and facilitates better transfer of discriminative information in few-shot scenarios.

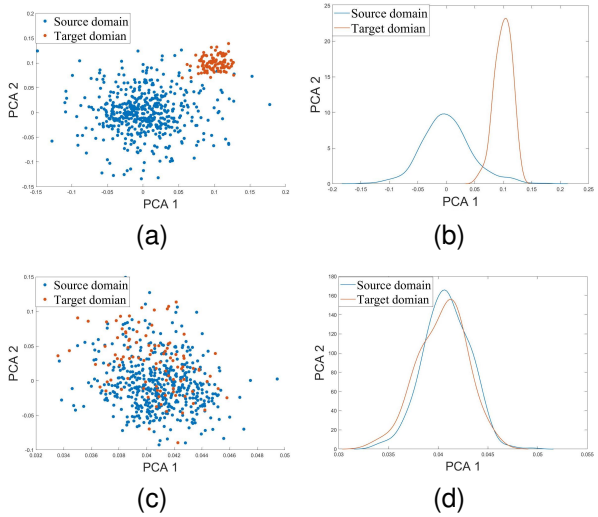


Fig. 9. Feature distribution and the corresponding probability density curve of the two domains' data, where (a) and (b) are before the transfer, (c) and (d) are after the transfer.

To verify the effectiveness of the interaction strategy in the TFE mechanism, the deep and wavelet features from the first and last epochs are shown in Fig. 10. In the initial epoch (Figs. 10(a) and 10(b)), noise interference prevents the features from exhibiting significant discriminative ability. However, after training (Figs. 10(c) and 10(d)), both deep and wavelet features display strong fault discriminability, with noise effectively suppressed. These results confirm the effectiveness of the interaction strategy in enhancing feature quality.

To emphasize the significance of each frequency band, Fig. 11 presents the *FI* change curve for each frequency band in the source domain during the transfer process. The low-frequency bands consistently exhibit higher importance due to their stable degradation trends, which facilitate the extraction of common discriminative information.

Fig. 12 displays the detection results for the target bearing, Bearing 2_1, in Task 1. The alarm strategy used in this experiment identifies three consecutive anomaly samples as

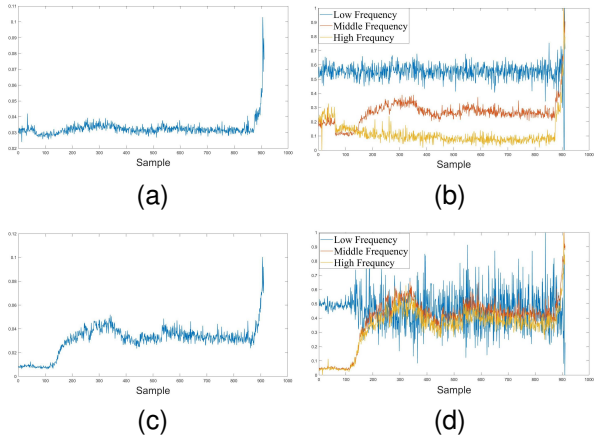


Fig. 10. Visual output of the TFE training process for Bearing 2_1 in Task1, where (a) and (b) are the deep features and wavelet features in the first epoch, respectively, (c) and (d) are those in the last epoch.

an early fault. The proposed few-shot DAAD with the TFE mechanism detects the early fault at the 137th sample. Moreover, a significant increase in the anomaly score is observed around the early fault, highlighting the model's strong ability to effectively discriminate early faults.

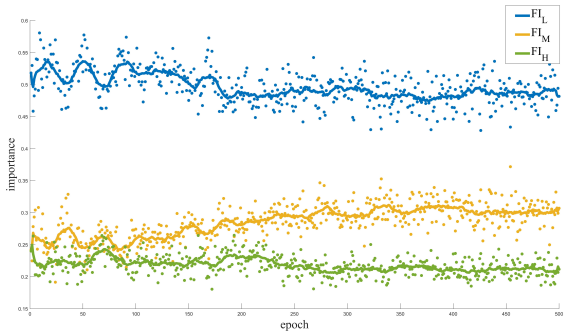


Fig. 11. Change curve of *FI* in each frequency band of Task1.

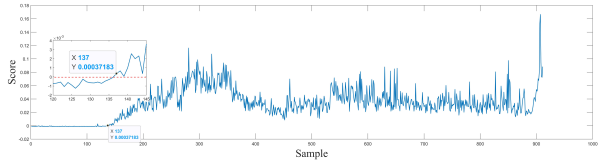


Fig. 12. Detection result on the bearing Bearing 2_1 in Task1, the samples whose score is below 0 are recognized in normal state, while the samples with scores above 0 are judged in early fault state.

To verify the necessity of the dual interaction in the TFE mechanism, two ablation experiments were conducted: (1) removing the feedback operation and (2) removing the enhancement operation. The results, shown in Fig. 13, reveal the following:

1) In the first experiment, where the feedback operation was removed, early fault detection was delayed (red line), although the model retained some discrimination ability. This is because only the task constraint loss was removed, while the enhancement of deep features by wavelet features remained intact.
2) In the second experiment, where the enhancement operation was removed, the anomaly score showed no change at the early fault (yellow line), indicating that the model lost its ability to extract discriminative information.

These findings highlight the critical role of both feedback and enhancement operations in achieving effective fault detection.

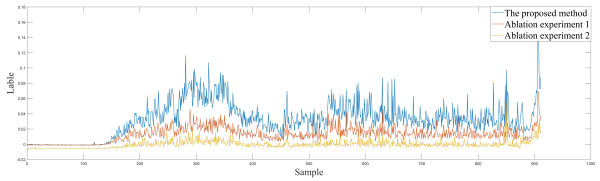


Fig. 13. Results of ablation experiments on the bearing Bearing 2_1.

b) Comparative experiments: This section introduces 24 anomaly detection methods for comparative analysis, covering signal analysis-based approaches and early fault detection methods with and without deep transfer learning. The details of the newly-added methods are introduced here.

1) Method 1 [60]: A state-of-the-art approach using bandwidth empirical mode decomposition (B-EMD) for anomaly detection.
2) Methods 2–7: Combine three classical anomaly detection methods (LOF, SVDD, and iForest) with two feature types (kurtosis and stacked denoising autoencoder (SDAE)).
3) Method 10 [56]: A self-supervised anomaly detection method for small-scale data, leveraging data augmentation and contrastive learning to extract discriminative information.
4) Method 12 [16]: Pre-train a network using knowledge distillation and then fine-tune it for anomaly detection.
5) Methods 14–16: Deep transfer anomaly detection methods utilizing domain adaptation.
6) IRAD [27]: Train a cross-domain encoder with adversarial learning to extract domain-invariant features.
7) PANDA [61]: Adapt the target distribution using pre-trained features to avoid feature degradation.
8) LogTAD [22]: Map data into the same hypersphere using domain adversarial training to extract domain-invariant features.
9) Methods 17–18: Recognize early fault using deep learning algorithms.
10) Method 17 [62]: Matche SDAE deep fault features between two domains using a sliding window.
11) Method 18 [63]: Detect anomalies based on deviation.
12) Methods 19–21: Early fault detection methods using deep transfer learning.
13) SRD [64]: Utilize sparse dictionary coding and K-nearest for fault detection.
14) OD-DTL [59]: Employ fine-tuning strategies for fault detection.

15) DAFD [65]: Leverage MMD distance for bearing fault diagnosis.
16) Methods 23–24: Method 23 and Method 24 adopt WSN and GANs to extract time-frequency features and to generate synthetic samples for small-scale data, respectively. The outputs from both methods are then fed into the DAAD model for anomaly detection in few-shot scenarios.

Two evaluation metrics are used for analysis: detection location and the number of false alarms [28]. Table IV presents the detection results of the proposed few-shot DAAD with the TFE mechanism compared to the 24 other methods.

Table IV shows that the proposed few-shot DAAD with the TFE mechanism detects early faults earlier and with the lowest false alarm rate. Although Methods 23-24 reduce the false alarm rate, they still detect faults later, demonstrating that traditional feature enhancement can't fully address the data limitations in few-shot scenarios. In contrast, the few-shot DAAD with TFE mechanism ensures earlier detection and fewer false alarms by enhancing deep features with wavelet-based time-frequency information, allowing for more accurate detection rules in few-shot tasks.

TABLE IV
DETECTION RESULTS OF TOTAL 25 METHODS ON THE PHM DATASET.

Method type	Method name	Detection location	Number of false alarms
Signal analysis method	1. BEMD + AMMA [60]	185	-
Anomaly detection methods without deep transfer learning	2. Kurtosis + LOF	606	35
	3. SDAE + LOF	280	17
	4. Kurtosis + SVDD	819	53
	5. SDAE + SVDD	880	4
	6. Kurtosis + iForset	311	36
	7. SDAE + iForest	876	20
	8. Deep SVDD (1)	253	2
	9. Deep SVDD (2)	201	0
	10. SSL [56]	885	7
Anomaly detection methods with deep transfer learning	11. DCAE pre-train [20]	156	16
	12. KD pre-train [16]	155	25
	13. SSL pre-train [56]	150	24
	14. IRAD [61]	150	7
	15. PANDA [59]	156	5
	16. LogTAD [22]	154	14
Early fault detection methods without deep transfer learning	17. SDFM [62]	175	7
	18. FDDA [63]	879	4
Early fault detection methods with deep transfer learning	19. SRD [64]	880	2
	20. OD-DTL [59]	166	26
	21. DAFD [65]	147	28
	22. DAAD [28]	169	24
	23. WSN [6] + DAAD	170	6
	24. GAN [2] + DAAD	154	3
	25. Our approach	**137**	**0**

3) Experimental results of Task 2: Fig. 14 illustrates the results of feature enhancement after applying the TFE mechanism for the FWM bearing in Task 2. Similarly to Fig. 10, both feature representations demonstrate strong discriminative ability, further validating the effectiveness of the TFE mechanism.

Fig. 15 presents the detection results for the FWM bearing in Task 2, using the same alarm strategy as in Task 1. The proposed few-shot DAAD with the TFE mechanism detected

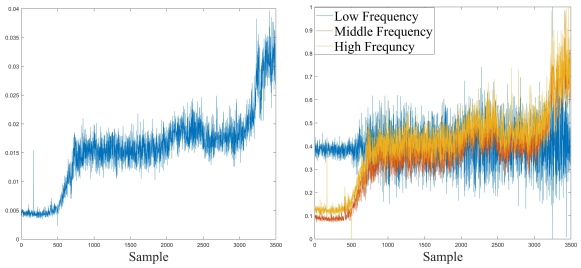


Fig. 14. Output of the last TFE training epoch for the FWM bearing in Task2, with (a) deep features and (b) wavelet features.

the fault at the 579th sample, corresponding to *January 13, 2023*. In comparison, the factory's official bearing condition monitoring system identified the early fault on *February 18, 2023*, as shown in Fig. 16, which means that the proposed few-shot DAAD provided an alarm approximately one month earlier.

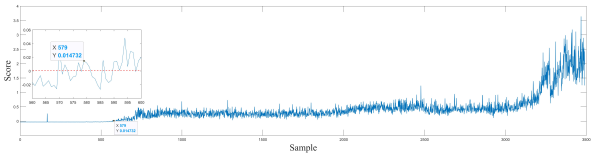


Fig. 15. Detection result on the bearing FWM, where the samples whose score is below 0 are recognized in normal state, while the samples with scores above 0 are judged in early fault state.

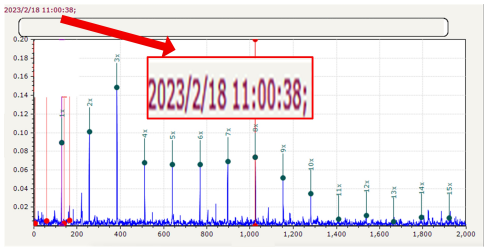


Fig. 16. Fault frequency-based detection results for the bearing FWM, provided by the factory's official bearing condition monitoring system.

Three representative anomaly detection methods are used for comparative analysis, as shown in Fig. 17. The results indicate that, regardless of the transfer strategy, existing models show delayed alarm locations in few-shot scenarios.

VI. CONCLUSION

In this paper, a TFE mechanism has been proposed to address the challenge of feature enhancement in few-shot transfer learning for anomaly detection. The dual interaction between wavelet and deep features, guided by the *FI* metric, has clarified the importance of each frequency band and improved result interpretability. Experiments on two anomaly detection problems have validated its effectiveness. It has been shown that 1) sharing complete yet physically-aware time-frequency information has enhanced deep feature representation, demonstrating that *collaboration* outperforms *integration*;

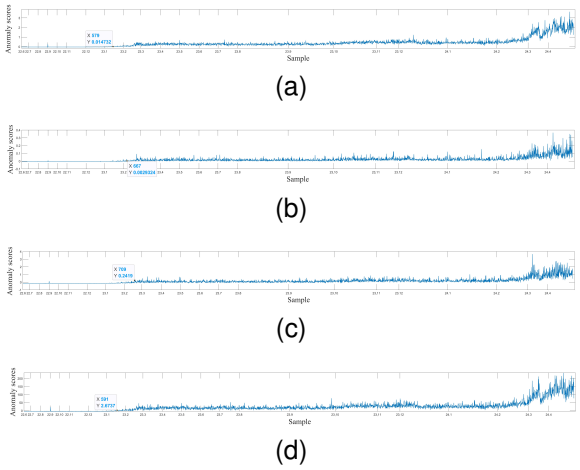


Fig. 17. Comparative results of the target bearing FWM, where (a) is the result of few-shot DAAD with the TFE mechanism, (b) is the result of DAAD, (c) is the result of preKD, and (d) is the result of preSSL.

2) the TFE mechanism has served as a flexible framework, adaptable to tasks like fault prognosis and object detection by adjusting detection rules; and 3) theoretical upper bound for classification error has been established to ensure the reliability of the dual interaction strategy and identify key performance factors.

As a theoretical exploration, this paper unavoidably has some limitations. Applying the TFE mechanism to other task types, such as fault prognosis, still remains a significant challenge, particularly in formulating the task objective analogous to the detection rule in anomaly detection tasks. Moreover, future research will focus on streaming data anomaly detection, tackling concept drift and incremental updates, and exploring advanced time-frequency analysis methods to further enhance few-shot learning.

REFERENCES

[1] A. Laine, S. Schuler, J. Fan, and W. Huda. Mammographic feature enhancement by multiscale analysis, *IEEE Transactions on Medical Imaging*, vol. 13, no. 4, pp. 725–740, 1994.

[2] F. Tanaka and C. Aranha. Data augmentation using GANs, *arXiv preprint arXiv: 1904.09135*, 2019.

[3] X. Chen, L. Zeng, M. Gao, et al. DiffWT: Diffusion-based pedestrian trajectory prediction with time-frequency wavelet transform, *IEEE Internet of Things Journal*, vol. 12, no. 5, pp. 5109–5121, 2024.

[4] L. Jia, T. Chow, Y. Yuan. GTFE-Net: A gramian time frequency enhancement CNN for bearing fault diagnosis, *Engineering Applications of Artificial Intelligence*, vol. 119, pp. 105794, 2023.

[5] S. Zhang, Z. Tao, S. Lin. Waveletformernet: A transformer-based wavelet network for real-world non-homogeneous and dense fog removal. *Image and Vision Computing*, vol. 146, pp. 105014, 2024.

[6] J. Shi, Y. Zhao, W. Xiang, V. Monga, X. Liu, and R. Tao. Deep scattering network with fractional wavelet transform, *IEEE Transactions on Signal Processing*, vol. 69, pp. 4740–4757, 2021.

[7] Y. Chen, Z. Zhao, and H. Wu, et al. Fault anomaly detection of synchronous machine winding based on isolation forest and impulse frequency response analysis, *Measurement*, vol. 188, pp. 110531, 2022.

[8] J. Saari, J. Lundberg, and A. Thomson. Detection and identification of windmill bearing faults using a one-class support vector machine (SVM), *Measurement*, vol. 137, pp. 287–301, 2019.

[9] Y. Zhao, S. Wang, and F. Xiao. Pattern recognition-based chillers fault detection method using Support Vector Data Description (SVDD), *Applied Energy*, vol. 112, pp. 1041–1048, 2013.

[10] X. Li, M. Li, P. Yan, G. Li, Y. Jiang, H. Luo and S. Yin, Deep learning attention mechanism in medical image analysis: Basics and beyonds, *International Journal of Network Dynamics and Intelligence*, vol. 2, no. 1, pp. 93–116, 2023.

[11] C. Ma, P. Cheng and C. Cai, Localization and mapping method based on multimodal information fusion and deep learning for dynamic object removal, *International Journal of Network Dynamics and Intelligence*, vol. 3, no. 2, no. 100008, 2024.

[12] D. Wang, C. Wen and X. Feng, Deep variational Luenberger-type observer with dynamic objects channel-attention for stochastic video prediction, *International Journal of Systems Science*, vol. 55, no. 4, pp. 728–740, 2024.

[13] J. Wang, Y. Zhuang and Y. Liu, FSS-Net: A fast search structure for 3D point clouds in deep learning, *International Journal of Network Dynamics and Intelligence*, vol. 2, no. 2, no. 100005, 2023.

[14] L. Ruff, R. Vandermeulen, N. Goernitz, and L. Deecke, etc. Deep one-class classification, *Proceedings of the International conference on machine learning(PMLR)*, pp. 4393–4402, 2018.

[15] S. Baireddy, S. Desai, J. Mathieson, etc. Spacecraft time-series anomaly detection using transfer learning, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1951-1960, 2021.

[16] P. Bergmann, M. Fauser, and D. Sattlegger, etc. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4183–4192, 2020.

[17] H. Han, H. Liu, and C. Yang, etc. Transfer Learning Algorithm With Knowledge Division Level, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 11, pp. 8602–8616, 2023.

[18] M. Subramanian, K. Shanmugavadivel, and P. Nandhini. On fine-tuning deep learning models using transfer learning and hyper-parameters optimization for disease identification in maize leaves, *Neural Computing and Applications*, vol. 34, no. 16, pp. 13951–13968, 2022.

[19] Z. Swati, Q. Zhao, and M. Kabir, etc. Brain tumor classification for MR images using transfer learning and fine-tuning, *Computerized Medical Imaging and Graphics*, vol. 75, pp. 34–46, 2019.

[20] Y. Zhou, X. Liang, and W. Zhang, etc. VAE-based deep SVDD for anomaly detection, *Neurocomputing*, vol. 453, pp. 131–140, 2021.

[21] H. Farahani, A. Fatehi, and A. Nadali, etc. Domain Adversarial Neural Network Regression to design transferable soft sensor in a power plant, *Computers in Industry*, vol. 132, pp. 103489, 2021.

[22] X. Han, S. Yuan, and A. Claims. Unsupervised cross-system log anomaly detection via domain adaptation, *Proceedings of the 30th ACM international conference on information & knowledge management*, pp. 3068–3072, 2021.

[23] G. Ma, Z. Wang, W. Liu, etc, Estimating the state of health for lithium-ion batteries: a particle swarm optimization-assisted deep domain adaptation approach, *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 7, pp. 1530–1543, Jul. 2023.

[24] M. Ragab, E. Eldele, and Z. Chen, etc. Self-Supervised Autoregressive Domain Adaptation for Time Series Data, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 1, pp. 1341-1351, 2024.

[25] T. Shermin, G. Lu, and S. Teng, etc. Adversarial network with multiple classifiers for open set domain adaptation, *IEEE Transactions on Multimedia*, vol. 23, pp. 2732-2744, 2020.

[26] C. Wang, Z. Wang, W. Liu, Y. Shen and H. Dong, A novel deep offline-to-online transfer learning framework for pipeline leakage detection with small samples, *IEEE Transactions on Instrumentation and Measurement*, vol. 72, no. 3503913, 2023.

[27] Z. Yang, I. Soltani, and E. Darve. Anomaly detection with domain adaptation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2957–2966, 2023.

[28] W. Mao, G. Wang, and L. Kou, etc. Deep domain-adversarial anomaly detection with one-class transfer learning, *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 2, pp. 524–546, 2023.

[29] C. Bowles, L. Chen, and R. Guerrero, etc. Gan augmentation: Augmenting training data using generative adversarial networks, *arXiv preprint arXiv:1810.10863*, 2018.

[30] M. Frid-Adar, E. Klang, and M. Amitai, etc. Synthetic data augmentation using GAN for improved liver lesion classification, *Proceedings of the 15th international symposium on biomedical imaging (ISBI 2018)*, pp. 289–293, 2018.

[31] I. Goodfellow, J. Pouget-Abadie, and M. Mirza, etc. Generative adversarial networks, *Communications of the ACM*, vol. 53, no. 11, pp. 139-144, 2020.

[32] M. Kang, J. Shin, and J. Park. StudioGAN: a taxonomy and benchmark of GANs for image synthesis, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 15725-15742, 2023.

[33] K. Liu, Z. Ye, and H. Guo, etc. FISS GAN: A Generative Adversarial Network for Foggy Image Semantic Segmentation, *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 8, pp. 1428-1439, 2021.

[34] N. Tran, V. Tran, N. Nguyen, etc. On data augmentation for GAN training, *IEEE Transactions on Image Processing*, vol. 30, pp. 1882-1897, 2021.

[35] J. Ho, A. Jain, P. Abbeel. Denoising diffusion probabilistic models, *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[36] Y. Fu, X. Xie, and Y. Fu, etc. Wave-SAN: Wavelet based Style Augmentation Network for Cross-Domain Few-Shot Learning, *Proceedings of the 2022 Computer Vision and Pattern Recognition*, pp. 2203.07656, 2022.

[37] M. Behmanesh, P. Adibi, and S. Ehsani, etc. Geometric multimodal deep learning with multiscaled graph wavelet convolutional network, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 5, pp. 6991–7005, 2022.

[38] L. Bouny, M. Khalil, A. Adib. ECG heartbeat classification based on multi-scale wavelet convolutional neural networks, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3212–3216, 2020.

[39] X. Fu, J. Tao, and K. Jiao, etc. A novel semi-supervised prototype network with two-stream wavelet scattering convolutional encoder for TBM main bearing few-shot fault diagnosis, *Knowledge-Based Systems*, vol. 286, pp. 111408, 2024.

[40] M. Mishra, P. Dash, and J. Nayak, etc. Deep learning and wavelet transform integrated approach for short-term solar PV power prediction, *Measurement*, vol. 166, pp. 108250, 2020.

[41] E. Pan, Y. Ma, and X. Mei, etc. D2Net: Deep Denoising Network in Frequency Domain for Hyperspectral Image, *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 3, pp. 813-815, 2023.

[42] A. Tirinzoni, A. Sessa, and M Pirotta, etc. Importance weighted transfer of samples in reinforcement learning, *International Conference on Machine Learning. PMLR*, pp. 4936–4945, 2018.

[43] J. Zhang, Z. Ding, and W. Li, etc. Importance weighted adversarial nets for partial domain adaptation, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8156–8164, 2018.

[44] P. Bugata and P. Drotar. Feature selection based on a sparse neural-network layer with normalizing constraints, *IEEE transactions on cybernetics*, vol. 53, no. 1, pp. 161–172, 2021.

[45] A. Jeyasothy, S. Suresh, and S. Ramasamy, etc. Development of a novel transformation of spiking neural classifier to an interpretable classifier, *IEEE Transactions on Cybernetics*, vol. 54, no. 1, pp. 3–12, 2022.

[46] J. Wu, W. Mao, and Y. Zhang, etc. A Novel Few-shot Deep Transfer Learning Method for Anomaly Detection: Deep Domain-Adversarial Contrastive Network With Time-Frequency Transferability Analytics, *IEEE Internet of Things Journal*, vol. 11, no. 17, pp. 28809–28823, 2024.

[47] Y. Dhebar and K. Deb. Interpretable rule discovery through bilevel optimization of split-rules of nonlinear decision trees for classification problems, *IEEE Transactions on Cybernetics*, vol. 51, no. 11, pp. 5573–5584, 2020.

[48] C. Singh, W. Murdoch, and B. Yu. Hierarchical interpretations for neural network predictions. *arXiv preprint arXiv:1806.05337*, 2018.

[49] M. Zeiler and R. Fergus. Visualizing and understanding convolutional networks, *Computer Vision–ECCV 2014: 13th European Conference*, pp. 818–833, 2014.

[50] C. Nguyen, L. Ho, and V. Dinh, etc. Transferability Between Regression Tasks, *Proceedings of the NeurIPS 2022 Workshop on Distribution Shifts:Connecting Methods and Applications*.

[51] Y. Xue, R. Yang, and X. Chen, etc. A Review on Transferability Estimation in Deep Transfer Learning, *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 12, pp. 5894–5914, 2024.

[52] J. Grill, F. Strub, and C. Tallec, etc. Bootstrap your own latent-a new approach to self-supervised learning, *Advances in neural information processing systems*, vol. 33, pp. 21271–21284, 2020.

[53] K. He and J. Sun. Convolutional neural networks at constrained time cost, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5353–5360, 2015.

[54] W. Huang, M. Yi, and X. Zhao, etc. Towards the generalization of contrastive self-supervised learning, *arXiv preprint arXiv:2111.00743*, 2021.

[55] G. Pang, C. Shen, and L. Cao, etc. Deep learning for anomaly detection: A review, *Proceedings of the ACM computing surveys (CSUR)*, vol. 54, no. 2, pp. 1–38, 2021.

[56] K. Sohn, C. Li, and J. Yoon, etc. Learning and evaluating representations for deep one-class classification, *arXiv preprint arXiv:2011.02578*, 2020.

[57] H. Venkateswara, J. Eusebio, and S. Chakraborty, etc. Deep hashing network for unsupervised domain adaptation, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5018–5027, 2017.

[58] P. Nectoux, R. Gouriveau, and K. Medjaher, etc. PRONOSTIA: An experimental platform for bearings accelerated degradation tests, *Proceedings of the IEEE International Conference on Prognostics and Health Management, PHM'12*, pp. 1–8, 2012.

[59] W. Mao, L. Ding, and S. Tian, etc. Online detection for bearing incipient fault based on deep transfer learning, *Measurement*, vol. 152, pp. 107278, 2020.

[60] Y. Li, M. Xu, and X. Liang, etc. Application of bandwidth EMD and adaptive multiscale morphology analysis for incipient fault diagnosis of rolling bearings, *IEEE Transactions on Industrial Electronics*, vol. 64, no. 8, pp. 6506–6517, 2017.

[61] T. Reiss, N. Cohen, and L. Bergman, etc. Panda: Adapting pretrained features for anomaly detection and segmentation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2806–2814, 2021.

[62] W. Mao, J. Chen, and X. Liang, etc. A new online detection approach for rolling bearing incipient fault via self-adaptive deep feature matching, *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 2, pp. 443–456, 2019.

[63] W. Lu, Y. Li, and Y. Cheng, etc. Early fault detection approach with deep architectures, *IEEE Transactions on instrumentation and measurement*, vol. 67, no. 7, pp. 1679–1689, 2018.

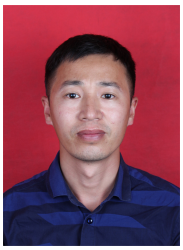
[64] X. Guo, S. Liu, and Y. Li. Fault detection of multi-mode processes employing sparse residual distance, *Acta Automatica Sinica*, vol. 45, no. 3, pp. 617–625, 2019.

[65] W. Lu, B. Liang, and Y. Cheng, etc. Deep model based domain adaptation for fault diagnosis, *IEEE Transactions on Industrial Electronics*, vol. 64, no. 3, pp. 2296–2305, 2016.

[66] S. Mallat. Group invariant scattering, *Communications on Pure and Applied Math ematics*, vol. 65, no. 10, pp. 1331–1398, 2012.

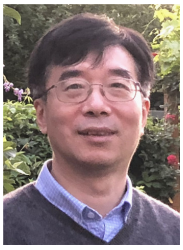
Wentao Mao (Member, IEEE) received his M.S. degree in Computer Science from Chongqing University of Posts and Telecommunications in 2006, and the Ph.D degree in engineering mechanics from Xi'an Jiaotong University, China, in 2011. He is currently serving as a Full Professor at the School of Computer and Information Engineering, Henan Normal University, China. His current research interests include machine learning, time series analysis and fault prognostics. Now he have conducted and is conducting about 10 research projects such as National Natural Science Foundation of China as project principal or main researcher.

Jianing Wu received her M.S. degree in Computer Science and Technology from Henan Normal University in 2025. Her research interests include time series forecasting and fault prognostics with deep learning techniques.



Shubin Du obtained a Bachelor's degree in Engineering from Yanshan University in Qinhuangdao, China in 2008. He is currently engaged in the diagnosis of mechanical and electrical equipment faults at Hebei Iron and Steel Tangshan Company, serving as a senior engineer and holding the qualification of International Vibration Analyst Level III.

Ke Feng (Senior Member, IEEE) is a Full Professor at Xi'an Jiaotong University, China. He is a Marie Curie Fellow (Imperial College London & Brunel University London). He received a Ph.D. degree from the University of New South Wales, Australia, in 2021. He worked at the University of British Columbia and the National University of Singapore in 2022 and 2023, respectively. His main research interests include digital twins, vibration analysis, structural health monitoring, dynamics, tribology, signal processing, and machine learning. He is recognized as the Emerging Leader (2023) by the Measurement Science and Technology journal. He is the Associate Editor and Editorial Board Member of several journals, including IEEE Transactions on Industrial Informatics, Information Fusion, IEEE Internet of Things Journal, Journal of Intelligent Manufacturing, Structural Health Monitoring, IEEE Transactions on Instrumentation and Measurement, IEEE Sensors Journal, IET Collaborative Intelligent Manufacturing, Measurement Science and Technology, Proc Inst Mech Eng B J Eng Manuf, etc.



Zidong Wang (Fellow, IEEE) received the B.Sc. degree in mathematics in 1986 from Suzhou University, and the M.Sc. degree in applied mathematics in 1990 and the Ph.D. degree in electrical engineering in 1994, both from Nanjing University of Science and Technology.

He is currently a Professor of dynamical systems and computing in the Department of Computer Science' Brunel University London, U.K, From 1990 to 2002, he held teaching and research appointments inuniversities in China, Germany and the U.K.. His research interests include dynamical systems, signal processing, bioinformatics, control theory and applications. He has published a number of papers in international journals. He is a Holder of the Alexander von Humboldt Research Fellowship of Ger-many, the JSPS Research Fellowship of Japan, William Mong Visiting Research Fellowship of Hong Kong.

Prof. Wang serves (or has served) as the Editor-in-Chief for International Journal of Systems Science, the Editor-in-Chief for Neurocomputing, the Editor-in-Chief for Systems Science and Control Engineering, and an Associate Editor for 12 international journals including IEEE Transactions on Automatic Control, IEEE Transactions on Control Systems Technology, IEEE Transactions on Neural Networks, IEEE Transactions on Signal Processing, and IEEE Transactions on Systems, Man, and Cybernetics-Part C. He is a Member of the Academia Europaea, a Member of the European Academy of Sciences and Arts, an Academician of the International Academy for Systems and Cybernetic Sciences, a Fellow of the IEEE, a Fellow of the Royal Statistical Society and a member of program committee for many international con-ferences.

APPENDIX

A. Specific implementation of Wavelet Scattering Transform

First, $x_{i\times j}$ is transformed by the classical wavelet transform to obtain the wavelet coefficients $W_{x_{i\times j}}(t,\lambda_{p,k})$:

$$W_{x_{i\times j}}(t,\lambda_{p,k}) = x_{i\times j} * \psi_{\lambda_{p,k}}(t) \tag{18}$$

where t is the translation variable, $*$ represents the convolution operator, and $\psi_{\lambda_{p,k}}(t)$ denotes the orientation-scale filter obtained by rotating and scaling the mother wavelet $\psi(t)$. $\lambda_{p,k}$ is calculated as follows:

$$\lambda_{p,k} = 2^p r_k, r_k \triangleq \begin{pmatrix} \cos\theta_k & -\sin\theta_k \\ \sin\theta_k & \cos\theta_k \end{pmatrix} \tag{19}$$

where $p = 0,1,2,...J-1$ and $k = 0,1,2,...K-1$ represent the scale factor and orientation factor respectively. Here, J and K are the numbers of scale factors and orientation factors, respectively. By adjusting p and k, the features of $x_{i\times j}$ at different scales and orientations can be obtained. Referenced by [6], the numbers of scale factor and orientation factor in WST are set to $J=4$ and $K=8$, where this setting retains over 98% of the signal energy and is effective in most scenarios.

Second, the modulus operation is applied on $W_{x_{i\times j}}(t,\lambda_{p,k})$ to ensure the integrity of the discriminative information:

$$Q_{x_{i\times j}}(t,\lambda_{p,k}) = \left|W_{x_{i\times j}}(t,\lambda_{p,k})\right| = \left|x_{i\times j} * \psi_{\lambda_{p,k}}(t)\right| \tag{20}$$

Finally, $Q_{x_{i\times j}}(t,\lambda_{p,k})$ is modified by introducing a low-pass filter to reduce the sensitivity to local shift [66]:

$$S_{x_{i\times j}}(t,\lambda_{p,k}) = Q_{x_{i\times j}}(t,\lambda_{p,k}) * \phi_{J-1}(t) \tag{21}$$

where $\phi_{J-1}(t)$ is the maximum scale function, $S_{x_{i\times j}}(t,\lambda_{p,k})$ represents the wavelet scattering coefficient.

B. Proof of Lemma 1

Assume that each sample $x_0 \in \left\{\left(FI_L X_L^0\right) \cap \left(FI_M X_M^0\right) \cap \left(FI_H X_H^0\right) \cap S_\delta\right\}$ can be correctly classified. Then, classification error is given by:

$$\begin{aligned} Err(G_f) &= \sum_{i=1}^{M+N} P[G_f(x_i) > 0] \\ &\le P\left[\overline{(FI_L X_L^0) \cap (FI_M X_M^0) \cap (FI_H X_H^0) \cap S_\delta}\right] \\ &= P\left[\overline{FI_L X_L^0} \cup \overline{FI_M X_M^0} \cup \overline{FI_H X_H^0} \cup \overline{S_\delta}\right] \\ &\le P\left[\overline{FI_L X_L^0}\right] + P\left[\overline{FI_M X_M^0}\right] + P\left[\overline{FI_H X_H^0}\right] + P\left[\overline{S_\delta}\right] \\ &\le FI_L \times (1-\sigma_L) + FI_M \times (1-\sigma_M) + FI_H \times (1-\sigma_H) + P\left[\overline{S_\delta}\right] \\ &= 1 - FI_L \times \sigma_L - FI_M \times \sigma_M - FI_H \times \sigma_H + P\left[\overline{S_\delta}\right] \end{aligned} \tag{22}$$

The proof is complete.

C. Proof of Lemma 2

Let $B = \left(FI_L X_L^0\right) \cap \left(FI_M X_M^0\right) \cap \left(FI_H X_H^0\right)$. According to (22), we have:

$$P[B \cap S_\delta] = 1 - P\left[\overline{B \cap S_\delta}\right] \ge \alpha - P\left[\overline{S_\delta}\right] \tag{23}$$

where $\alpha = FI_L \times \sigma_L + FI_M \times \sigma_M + FI_H \times \sigma_H$.

From (23), we have:

$$-P\left[\overline{B \cap S_\delta}\right] \ge \ \alpha - 1 - P\left[\overline{S_\delta}\right] \tag{24}$$

For any $f^T(x_0)$ and $f(x)$, where $x \in \overline{B \cap S_\delta}$, $x_0 \in B \cap S_\delta$, it follows from the Cauchy-Schwarz inequality $|u \cdot v| \le \|u\| \times \|v\|$ and the normalization of f (by $\|f\| = r$) that

$$\left|f^T(x_0) f(x)\right| \le \left\|f^T(x_0)\right\| \times \|f(x)\| \le r^2 \tag{25}$$

where $\|\cdot\|$ represents l_2-norm, $|\cdot|$ is absolute value.

Notice that $x_0, x \in S_\delta$, $\|f(x) - f(x_0)\| \le \delta$. Therefore, for any $f^T(x_0)$ and $f(x)$, where $x_0, x \in B \cap S_\delta$, we have:

$$\left|f^T(x_0)\left(f(x) - f(x_0)\right)\right| \le \left\|f^T(x_0)\right\| \times \|f(x) - f(x_0)\| \le r\delta \tag{26}$$

With (25) and (26), we have:

$$\begin{cases} \underset{x \in \overline{B \cap S_\delta}}{E}\left[f^T(x_0) f(x)\right] \ge -r^2 \\ \underset{x \in B \cap S_\delta}{E}\left[f^T(x_0)\left(f(x) - f(x_0)\right)\right] \ge -r\delta \end{cases} \tag{27}$$

Based on the above analysis, we obtain:

$$\begin{aligned}
f^T(x_0)c &= f^T(x_0)\underset{x}{E}\left[f(x)\mathbb{I}\left(x\in X\right)\right] \\
&= f^T(x_0)\underset{x}{E}\left[f(x)\mathbb{I}\left(x\in B\cap S_\delta\right)\right]+f^T(x_0)\underset{x}{E}\left[f(x)\mathbb{I}\left(x\in \overline{B\cap S_\delta}\right)\right] \\
&= P\left[B\cap S_\delta\right]f^T(x_0)\underset{x\in B\cap S_\delta}{E}\left[f\left(x\right)\right]+P\left[\overline{B\cap S_\delta}\right]\underset{x\in \overline{B\cap S_\delta}}{E}\left[f^T(x_0)f(x)\right] \\
&\geq P\left[B\cap S_\delta\right]f^T(x_0)\underset{x\in B\cap S_\delta}{E}\left[f\left(x\right)\right]-r^2P\left[\overline{B\cap S_\delta}\right] \\
&= P\left[B\cap S_\delta\right]\underset{x\in B\cap S_\delta}{E}\left[f^T(x_0)\left(f\left(x\right)-f\left(x_0\right)+f\left(x_0\right)\right)\right]-r^2P\left[\overline{B\cap S_\delta}\right] \\
&\geq P\left[B\cap S_\delta\right]\left(r^2+\underset{x\in B\cap S_\delta}{E}\left[f^T(x_0)\left(f\left(x\right)-f\left(x_0\right)\right)\right]\right)-r^2P\left[\overline{B\cap S_\delta}\right] \\
&\geq \left(\alpha-P\left[\overline{S_\delta}\right]\right)\left(r^2-r\delta\right)+r^2\left(\alpha-1-P\left[\overline{S_\delta}\right]\right)
\end{aligned} \tag{28}$$

where $\mathbb{I}$ is the indicator function.

Since f can be normalized by $\|f\|=r$, we know that $\|f\left(x_0\right)\|^2+\|c\|^2\leq 2r^2$ holds. Then, we have:

$$\begin{aligned}
\|f\left(x_0\right)-c\|^2 &= \|f\left(x_0\right)\|^2-2f^T(x_0)c+\|c\|^2 \\
&\leq 2r^2-\left(\alpha-P\left[\overline{S_\delta}\right]\right)\left(r^2-r\delta\right)+r^2\left(\alpha-1-P\left[\overline{S_\delta}\right]\right) \\
&= r^2\left(\alpha+1-P\left[\overline{S_\delta}\right]\right)-\left(\alpha-P\left[\overline{S_\delta}\right]\right)\left(r^2-r\delta\right) \\
&= r^2\left(\alpha+1-P\left[\overline{S_\delta}\right]-\left(\alpha-P\left[\overline{S_\delta}\right]\right)\left(1-\frac{\delta}{r}\right)\right) \\
&= r^2\left(1-\frac{\delta P\left[\overline{S_\delta}\right]}{r}+\frac{\delta\alpha}{r}\right)
\end{aligned} \tag{29}$$

From the SVDD principle [14], if $\sqrt{\|f\left(x_0\right)-c\|^2}<R$, the sample $x_0\in\left\{\left(FI_LX_L^0\right)\cap\left(FI_MX_M^0\right)\cap\left(FI_HX_H^0\right)\cap S_\delta\right\}$ can be correctly classified. Therefore, according to (29), if the radius of SVDD hypersphere satisfies $R>r\sqrt{\left(1-\frac{\delta P\left[\overline{S_\delta}\right]}{r}+\frac{\delta\alpha}{r}\right)}$, any $x_0\in\left\{\left(FI_LX_L^0\right)\cap\left(FI_MX_M^0\right)\cap\left(FI_HX_H^0\right)\cap S_\delta\right\}$ can be correctly classified.

The proof is now complete.