> TIM-24-09856<

Deep Causal Disentanglement Network with Domain Generalization for Cross-machine Bearing Fault Diagnosis

Chaochao Guo[1], Youchao Sun[*1], Rourou Yu[1], Xinxin Ren[1]

***Abstract*—Domain generalization-based fault diagnosis (DGFD) has gained considerable attention in bearing fault diagnosis due to its ability to extract feature-invariant information from diverse source domains, without requiring direct access to target domain data. However, many existing DGFD approaches primarily rely on statistical models to capture the relationship between time-series data and labels. This often leads to the learning of entangled features, as these models lack prior knowledge to differentiate between task-relevant and task-irrelevant information. To address this limitation, this paper introduces the Deep Causal Disentanglement Network (DCDN), a novel approach tailored for cross-machine bearing fault diagnosis. In this framework, fault data collected from multiple source domains is decomposed into causal factors related to fault representation and non-causal factors associated with domain-specific information, using a structural causal model. This process effectively reconstructs the data generation pathway. By optimizing causal aggregation loss and maximizing information entropy loss, DCDN can distinguish between causal and non-causal features from both direct and indirect perspectives. Furthermore, a contrastive estimation loss is minimized to ensure that the extracted causal features retain most of the essential information from the original dataset. Additionally, a redundancy reduction loss is employed to minimize correlations among the dimensions of the causal vector, further reducing the entanglement between causal and non-causal factors. The effectiveness and superiority of the proposed model are demonstrated across five cross-machine bearing fault datasets. Experimental results show that, compared to other state-of-the-art (SOAT) methods, DCDN achieves superior performance in both estimation accuracy and robustness.**

***Index Terms*—Deep learning, Fault diagnosis, Causal learning, Causal disentanglement.**

List of Abbreviations

Abbreviations	Full Form
CAAN	cycle-consistent adversarial adaptation network
CCDG	conditional contrastive domain generalization
CCN	causal consistency network
CNN	convolutional neural network
CNN-C	center loss to CNN
DAFD	domain adaptation fault diagnosis
DAN	domain adversarial network
DCDN	deep causal disentanglement network
DCFN	deep causal factorization network
DDGFD	deep domain generalization network for fault diagnosis
DGFD	domain generalization-based fault diagnosis
DGNIS	domain generalization network combining invariance and specificity
EDM	electric discharge machining
IEDGNet	intrinsic and extrinsic domain generalization network
LODO	leave-one-domain-out
SCM	structural causal model
SOAT	state-of-the-art

I. Introduction

As industries increasingly transition towards greater informatization and intelligence, advanced intelligent fault diagnosis models have garnered significant attention in predictive maintenance and health management of mechanical systems [1, 2]. Traditional fault diagnosis techniques, critical in fields such as industrial, mechanical, and electronic applications, can be broadly categorized into two types: signal processing-based methods and those based on machine learning algorithms. These methods often depend on expert knowledge and manually designed feature sets, which limits their automation and intelligence. However, with advancements in sensing technology and computational capabilities, data-driven approaches have gained prominence for their high accuracy, straightforward end-to-end modeling, and excellent representation learning abilities. Among these, deep learning has become a leading method due to its competitive performance in fault diagnosis [3]. Techniques such as convolutional neural networks (CNNs) [4, 5], autoencoders [6], graph convolutional networks [7], deep belief networks [8], and sparse autoencoders [9], are capable of automatically extracting relevant features from raw data, establishing connections between these representations and health states with minimal

This work was financially supported by the Joint Fund of National Natural Science Foundation of China and Civil Aviation Administration of China (No. U2033202, U1333119); the National Natural Science Foundation of China (No.52172387); the Fundamental Research Funds for the Central Universities (ILA22032-1A); and the Aeronautical Science Foundation of China (2022Z071052001)

All authors are with the College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China. E-mail: guo_chaochao@126.com, sunyc@nuaa.edu.cn, 15122811703@163.com, ljgdxinxin@163.com.

Corresponding author: Youchao Sun.

reliance on expert input. These methods have shown strong efficacy, positioning them at the forefront of research in health management and early warning systems [10]. However, many traditional deep learning models assume that training and target domain data are collected under identical operational conditions, implying similar data distribution characteristics. In practice, variations in workload, rotational speed, radial force, equipment structure, and wear and degradation can lead to data distribution shifts, a phenomenon known as domain shift. This discrepancy undermines the generalization ability of models trained on a specific dataset when applied to different conditions, posing significant challenges for practical implementation in dynamic industrial environments[11-13]. as illustrated in Fig. 1 a).

To address these challenges, domain adaptation fault diagnosis (DAFD) and domain generalization-based fault diagnosis (DGFD) methods have been introduced [14]. DAFD methods aim to align the source and target domain distributions in high-level feature spaces, facilitating the extraction of domain-invariant features sensitive to faults and consistent across domains. It generally requires access to target domain data during training, which is not always feasible in real-world scenarios, where target domain data may only become available after the model is deployed. Furthermore, DAFD methods typically assume an offline training setting, where models are trained with existing data and then deployed, which is inadequate for environments with rapidly changing conditions. Moreover, traditional DAFD methods often focus on knowledge transfer from a single source domain to a single target domain, increasing the risk of overfitting to the source domain, especially in dynamic environments[15, 16]. (as depicted in Fig. 1b). Retraining models for each new target domain can be time-consuming and impractical, emphasizing the need for more flexible and generalized approaches.

Unlike DAFD, DGFD does not rely on assumptions regarding the availability of target domain data. Instead, it learns generalized knowledge from multiple source domains and applies this knowledge to previously unseen domains, as illustrated in Fig. 1c). DGFD methods generally aim to learn invariant dependencies by enforcing domain invariance in the marginal or conditional distributions of latent representations, thereby minimizing discrepancies between domains. The main advantage of DGFD is its independence from target domain data during model training[15, 17, 18].

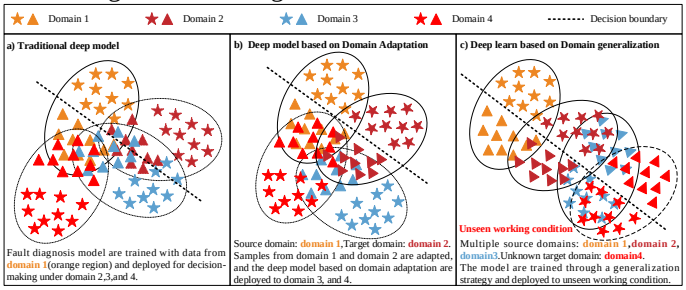

Fig. 1. Illustration of different fault diagnosis methods.

As highlighted in the previous discussion, learning invariant fault representations is essential for domain generalization methods. These fault diagnosis approaches have demonstrated promising results in diagnosing faults across different domains for the same bearing. However, due to the distinct nature of various machines and equipment, these models can introduce individualized biases in cross-machine fault diagnosis scenarios, which can limit their ability to generalize effectively. There are three primary challenges associated with these methods.

1) Fault data and corresponding labels for the test bearing are typically unavailable, while data from similar bearings or equipment is more accessible.
2) The process of extracting domain-invariant features may lack explicit mechanisms that go beyond learning statistical dependencies between variables, which can hinder robust generalization to unseen distributions [19, 20].
3) current DGFD methods often overlook domain-specific information that could improve performance and are susceptible to spurious correlations in the data [21].

As a result, cross-machine domain generalization for fault diagnosis remains a challenging and unresolved issue. One fundamental cause of this is that these methods tend to neglect the intrinsic causal mechanisms underlying data generation, leading to spurious correlations between domain-related factors (such as mechanical structure and operating environment) and fault labels during the data fitting process. Causal learning formalizes the data generation process using a Structural Causal Model (SCM), which allows for reasoning about the effects of changes to this process (i.e., interventions) and facilitates counterfactual reasoning[22-24]. This framework offers deeper insights into the underlying causal mechanisms. By applying this approach, a deep learning architecture can be designed to extract invariant causal relationships from the data, improving the understanding of interactions between variables and faults in complex systems. Although these approaches are grounded in causal learning, they mainly focus on causal factors, overlooking non-causal factors, which can limit their ability to eliminate spurious correlations.

To capture the unique nature of fault causality, this paper introduces a novel deep causal decoupling multi-source domain generalization method known as the Deep Causal disentanglement Network (DCDN). Based on a structural causal model, DCDN learns domain-invariant fault diagnosis knowledge by reconstructing the generation process of bearing vibration data. By optimizing the constructed causal aggregation loss and maximum information entropy loss, fault causal features and non-causal features are skillfully and effectively distinguished from both direct and indirect perspectives. The contrastive estimation loss is then minimized to ensure that the extracted features retain most of the information from the original data. Additionally, the contrastive estimation loss is minimized to retain the maximum amount of information from the original data, while redundancy loss is reduced to suppress potential correlations within the causal vector dimensions, further minimizing the entanglement of causal and non-causal features. The key contributions of this paper are as follows:

1) We introduce the Deep Causal Disentanglement Network (DCDN), a novel multi-source domain generalization

method that addresses the practical challenge of intelligent networks' dependence on target machine data in cross-machine fault diagnosis scenarios. By reconstructing the bearing vibration data generation process through an SCM, it learns domain-invariant fault diagnosis knowledge and achieves generalization by capturing invariant fault causal relationships.

2) The optimization of causal aggregation loss and maximum information entropy loss allows DCDN to effectively distinguish between fault causal and non-causal features, both directly and indirectly. Specifically, causal factors are aggregated for samples with the same category labels, while non-causal factors are aggregated for samples with the same domain labels. The maximum information entropy loss induces confusion between the fault classifier and domain classifier, thereby aiding in the separation of causal fault factors from non-causal domain factors.
3) The contrastive estimation loss is minimized to extract stable and comprehensive causal features, ensuring that the features retain most of the information from the original data. Furthermore, the redundancy loss is reduced to suppress correlations among the dimensions of the causal vector, reducing the entanglement between causal and non-causal features.

The remainder of this paper is structured as follows: Section II presents the related work. Section III discusses the domain generalization problem from a causal perspective. Section IV details the DCDN method. Section V presents the results of ablation experiments to validate DCDN, and Section VI concludes the paper.

II. Related Work

A. Domain adaptation fault diagnosis

Domain adaptation fault diagnosis has gained significant attention due to its ability to address domain shifts that arise in real-world scenarios. This techniques aim to bridge this gap by learning transferable features or aligning the source and target domain distributions. Current DAFD primarily focus on three main approaches: Feature Alignment Approaches, Adversarial-Based Methods, and Representation Learning. To minimize the discrepancy between feature distributions across different domains, methods such as Maximum Mean Discrepancy [25, 26]and Wasserstein Distance [27]are commonly employed. Recent studies have integrated these techniques into deep neural networks to jointly optimize classification performance and domain alignment[28, 29]. Liu et al. proposed a deep adversarial subdomain adaptation network to reduce discrepancies at both the domain and category levels [28]. while Shi et al. introduced an unsupervised multi-source domain adaptation approach with a transferability-based entropy penalty to improve fault diagnosis performance [30]. Additionally, Jiao et al. developed a Cycle-consistent Adversarial Adaptation Network (CAAN) to ensure reliable domain-invariant feature learning through adversarial games and cycle-consistency constraints, demonstrating strong effectiveness in experiments[29]. Li et al. propose an intelligent fault diagnosis method using improved domain adaptation with ensemble learning, effectively addressing domain mismatch in IIoT-based health monitoring for accurate fault diagnosis across varying conditions[31].

Despite significant progress, DAFD methods still face several challenges. For instance, the issue of negative transfer can lead to performance degradation, as aligning unrelated source and target domain features may have adverse effects. Additionally, many existing methods rely on labeled source domain data, which may not always be feasible to obtain in practical applications. Future research could focus on developing more efficient label utilization methods and improving the interpretability of domain adaptation models.

B. Domain generalization-based fault diagnosis

DGFD leverages multiple source domains to extract domain-invariant features, enabling robust performance in unknown operational environments. various DGFD methods have been proposed, all of which focus on modeling the statistical relationships between the observed inputs and corresponding ground truth labels. The three primary approaches in DGFD are domain-invariant representation learning[32, 33], data augmentation [34, 35], and sample selection strategies [36]. These methods generally aim to learn invariant dependencies by enforcing domain invariance in the marginal or conditional distributions of latent representations, thereby minimizing discrepancies between domains. The main advantage of DGFD is its independence from target domain data during model training. For example, Chen et al. developed an adversarial domain-invariant generalization framework specifically designed for fault diagnosis in industrial machinery. This framework uses adversarial learning to extract domain-invariant features, enhancing fault detection in unseen working conditions, and demonstrating significant accuracy and robustness in real-world applications[15]. Wu et al. proposed a convolutional neural network-based method that incorporates multi-sensor signal fusion, input random destruction, and minimal batch normalization. This method achieves over 99.5% accuracy in classifying compound faults, showing strong generalization and noise resilience[17]. In a similar vein, Zhu et al. introduced a novel capsule network that combines inception blocks, regression branches, and dynamic routing, addressing limitations in traditional CNNs related to feature positional relationships, with comparative studies highlighting its superior performance [37]. Yang et al. enhanced traditional CNNs with center loss to create an end-to-end fault diagnosis framework (CNN-C), which minimizes intra-class variations and domain differences across seen operating conditions while maintaining inter-class separability through joint supervision with softmax loss [38]. Zhao et al. introduced the Domain Generalization Network combining invariance and specificity (DGNIS), which aligns global distributions while clustering local class representations, enabling the model to learn both domain-invariant knowledge and discriminative features[18]. Ragab et al. proposed the Conditional Contrastive Domain Generalization (CCDG) method, which maximizes mutual information within similar classes and minimizes it across

different classes. CCDG effectively learns shared class information and environment-independent representations that transfer seamlessly to new, unseen domains[39]. Han et al. presented the Intrinsic and Extrinsic Domain Generalization Network (IEDGNet), which regularizes the deep network's discriminative structure using both intrinsic and extrinsic objectives. By minimizing triplet loss for intra-class compactness and employing adversarial training for domain-level regularization, IEDGNet learns robust features, enhancing its generalization to unseen domains[3].

DGFD encounters distinct challenges, including the limited diversity of source domains, which can hinder its ability to generalize effectively. Additionally, achieving an optimal balance between robustness and specificity remains unresolved.

C. Causal learning for domain generalization

Despite the widespread popularity of high-performance models tailored to specific training data distributions in the industry, these models often merely establish subtle statistical correlations between features. When these models are applied to test data with distributions differing from the training data, their performance tends to decline significantly [40]. This is because these models fail to capture the true causal mechanisms underlying the data [41]. Recently, causal learning has gained significant attention in fault diagnosis using deep learning methods. From a causal perspective, the ideal data representation we seek should fundamentally reveal the true causes of fault labels, enabling predictions to be immune to those features that are statistically correlated with the labels but semantically irrelevant. In statistical terms, when there is a statistical dependency between two observable variables *X* and *Y*, it is often inferred that there exists a latent causal variable Z_C that makes *X* and *Y* conditionally independent, implying that Z_C explains all the statistical dependency between *X* and *Y*. Therefore, causality is regarded as a more profound and nuanced description of variable relationships than mere statistics [34]. Causal learning addresses a significant limitation of traditional statistical methods: conditional probabilities alone cannot always yield accurate predictions when external interventions occur. In the context of domain generalization (DG), domain shift can be viewed as an external intervention that introduces spurious correlations between irrelevant features and class labels [42]. To tackle this issue, we introduce causal learning methods. To accurately capture these causal mechanisms, we utilize Structural Causal Models (SCM) to uncover the intrinsic causal relationships between fault data and corresponding labels [43]. By understanding these causal links, we can develop more robust and generalizable models that are better suited to adapt to data from varying distributions, thereby enhancing the model's ability to generalize across domains. Li et al. proposed the Whitening-Net framework, which utilizes compound domain signals and introduces causal loss for regularization, along with a whitening structure to help the network focus on signal causality instead of domain-specific noise [22]. Similarly, Li et al. introduced the Causal Consistency Network (CCN), which extracts invariant causal features and applies causal consistency loss to transform individualized machine data into consistent representations, facilitating knowledge generalization[23]. Jia et al. proposed the Deep Causal Factorization Network (DCFN), which uses the SCM to define generalized fault representations as causal factors and domain-specific representations as non-causal factors. This method incorporates specialized modules to reconstruct causal mechanisms and separate these two types of factors [24]. However, DCFN has some limitations, particularly concerning the completeness of causal and non-causal factor information, and it has only been tested in vibration-based fault diagnosis, limiting its applicability to other domains.

III. PRELIMINARY

Integrating human prior knowledge of causality into the SCM equips the model with the capability to uncover causal relationships. This approach extends beyond merely analyzing the data itself, as it explores the underlying mechanisms that produce the observed signals. In the case of bearing failure, a series of impulses are generated each time the rolling elements pass over the fault site. These impulses couple with the system's vibration response after traveling through the transmission path and are ultimately captured by accelerometers. From this perspective, the vibration signal $x(t)$ caused by bearing defects can be modeled as follows [19]:

$$x(t) = \delta(t) * h^{\delta}(t) + d(t) * h^{d}(t) \tag{1}$$

Where $\delta(t)$ represents the series of impulse signals generated by the machine fault, while $d(t)$ represents various interference signals caused by factors such as different operating environments or equipment differences (e.g., type of sensors, signal sampling frequency, background noise, rotational speed, load, and structural differences in the machine). The symbol '*' denotes the convolution operation. The terms $h^{\delta}(t)$ and $h^{d}(t)$ represent the transmission path effects for $\delta(t)$ and $d(t)$, respectively, describing the system's response characteristics to input signals. From Equation (1), the signal $x(t)$ can be decomposed into two independent parts: one term $\delta(t) * h^{\delta}(t)$ that is solely related to the fault (causal factor) and another term $d(t) * h^{d}(t)$ that is solely related to operating environment or equipment differences (excluding any information related to the fault itself). The fault's causal influence is concentrated in $\delta(t)$, but this term remains highly coupled with both fault *M* and domain *D* (such as variations in rotational speed, load, and bearing structure). This assumption provides a theoretical basis for separating fault-specific information from complex signals.

In Fig. 2 (a), the generation of fault signals is modeled using a statistical dependency approach. This model captures the direct relationship between signals and fault labels, but it overlooks the underlying causal mechanisms. Consequently, during the data fitting process, the diagnostic model may allow a path from *M* (related to faults) to *Y* (fault labels) due to the influence of *X*, thereby establishing spurious correlations between domain information and fault labels. The causal

relationship underlying the signal generation process can be more accurately described using an SCM, represented as a Directed Acyclic Graph (DAG), where nodes correspond to variables and edges represent functional relationships. In the context of fault diagnosis, if node A is a descendant of node B, then A is considered to be causally influenced by B. According to [24] and[19], the structural causal model of data generation can be represented as Fig. 2 (b). The process by which the data is generated can be expressed as follows:

$$\begin{aligned} X &= f_x(Z_C, Z_D, \varepsilon_x) \\ Y &= h_c(Z_C, \varepsilon_y) \\ Y_D &= h_d(Z_D, \varepsilon_{Y_D}) \\ Z_C &= f_c(M, \varepsilon_c) \\ Z_D &= f_d(D, \varepsilon_d) \end{aligned} \tag{2}$$

Where $\varepsilon_x, \varepsilon_y, \varepsilon_{Y_D}, \varepsilon_c$, and ε_d are independent error terms for X, Y, Y_D, Z_C, and Z_D, respectively. The functions f_x, h_c, h_d, f_c, and f_d, are the general nonparametric functions. Z_C represents the causal information that influences the causal relationship between X and the fault label Y, while Z_D captures the non-causal factors influencing X and the domain label Y_D. The function f_x is assumed to be smooth and invertible, which guarantees the existence of its inverse, f_x^{-1}.

From Fig. 2 (b), two paths connect the class label Y and the domain D. One path is d-separated, while the other is d-connected[44], indicating spurious correlations between Y and D. Consequently, traditional statistical-based predictors may suffer performance degradation due to domain shifts in D. The causality-based DGFD method can mitigate these spurious correlations by introducing unobservable variables Z_C. We can leverage neural networks to approximate the representations of these unobservable variables $[f_x^{-1}]_{Z_C}$ and $[f_x^{-1}]_{Z_D}$, thereby inferring Z_C and Z_D. Considering that joint distribution $P(Y|Z_C)$ remains invariant cross-domains, once Z_M is obtained, the causal predictor f_y: $f_y(Z_C) \to Y$ can be learned in an end-to-end manner.

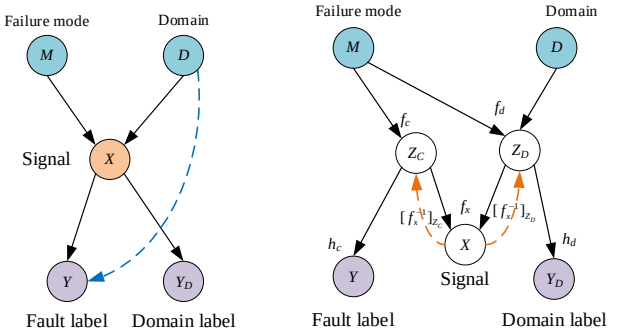

Fig. 2. Illustration of bearing fault signal generation. (a) Statistical dependence model. (b) Structural causal model for fault diagnosis. Solid-colored nodes represent observable variables, while hollow nodes denote unobservable variables. Directed arrows indicate direct causal relationships, blue dashed lines represent spurious correlations, and orange dashed lines indicate inverse functional relationships between unobservable variables and collected signals.

IV. Proposed Method

A. Problem definition

Let $\mathcal{X}$ denote the sample space and $\mathcal{Y}$ the label space. A domain is represented by a joint distribution P_{XY} defined over $\mathcal{X} \times \mathcal{Y}$. Let $\mathcal{D}^s = \{\mathcal{D}_k^s | k = 1, 2, \cdots, K\}$ represent a training dataset collected from K source domains and $\mathcal{D}^t$ represent the target domain, s represents a source domain, and t represents the target domain. The k-th source domain contains n_k^s labeled samples $\mathcal{D}_k^s = \{(x_{k,i}^s, y_{k,i}^s, d_{k,i})\}_{i=1}^{n_k^s}$, where $x_{k,i}^s \in \mathbb{R}^{N_{input}}$ indicates the original data of length N_{input}, $y_{k,i}^s \in \mathbb{R}^{N_c}$ represents the health state of the i-th sample in the k-th source domain, N_c denotes the number of health states, $d_{k,i} \in \mathbb{R}^{N_s}$ represents the domain label of the i-th sample in the k-th source domain. Similarly, $X^t = \{(x_j^t, y_j^t)\}_{j=1}^{n_t}$ denotes the sample space of the target domain $\mathcal{D}^t$, where $x_j^t \in \mathcal{X} \subset \mathbb{R}^{N_{input}}$ is the raw data in the target domain, $y_j^t \in \mathcal{Y} \subset \{1, 2, \cdots, N_c\}$ represents the health status of the target domain sample, and n_t is the total number of samples in the target domain. Notably, the fault categories to be diagnosed are typically predefined and consistent across different source domains, forming a homogeneous domain. However, since the source and target data are collected under different operating conditions—such as varying load weights and rotational speeds—the data distributions between the multi-source domains and the target domain are not identical, i.e., $P(X_1^s) \neq P(X_2^s) \neq \cdots \neq P(X_K^s) \neq P(X^t)$. Our study focuses on homogeneous DGFD, assuming that both the source and target domains share the same fault modes.

Define $\mathbb{E}_{p^s}[y|\boldsymbol{x}] := \int_{\mathcal{Y}} y p^s(y|\boldsymbol{x}) dy$ for any $(x_{k,i}^s, y_{k,i}^s) \in \mathcal{X} \times \mathcal{Y}$ in the source domain, and denote $\mathbb{E}_{p^t}[y|\boldsymbol{x}] := \int_{\mathcal{Y}} y p^t(y|\boldsymbol{x}) dy$ for any $(x_{k,i}^t, y_{k,i}^t) \in \mathcal{X} \times \mathcal{Y}$ in the target domain. According to the domain-invariance of $P(Y|Z_C)$, we can derive that $\mathbb{E}_{p^s}[y|\boldsymbol{z}_c] = \mathbb{E}_{p^t}[y|\boldsymbol{z}_c] = \mathbb{E}[y|\boldsymbol{z}_c] = \int_{\mathcal{Y}} y p(y|\boldsymbol{z}_c) dy$ and denote $g(Z_C) := \mathbb{E}[y|Z_C]$. Sun et al. provided the generalization error bound $\left|\mathbb{E}_{p^s}[y|\boldsymbol{x}] - \mathbb{E}_{p^t}[y|\boldsymbol{x}]\right|$ [45].

B. Methodology overview

Fig. 3 presents an overview of the DCDN method, detailing its key processes during runtime. Initially, fault datasets from various bearings, representing multiple source domains, are collected. To address differences in sampling frequencies and rotational speeds, the input vibration signals undergo preprocessing using angular resampling and the Hilbert transform. The same preprocessing procedure is applied to the target domain data. The proposed DCDN network is then

trained on the source domain data. Once training is complete, the model's diagnostic performance is evaluated using the target domain data, providing an assessment of its generalization capability.

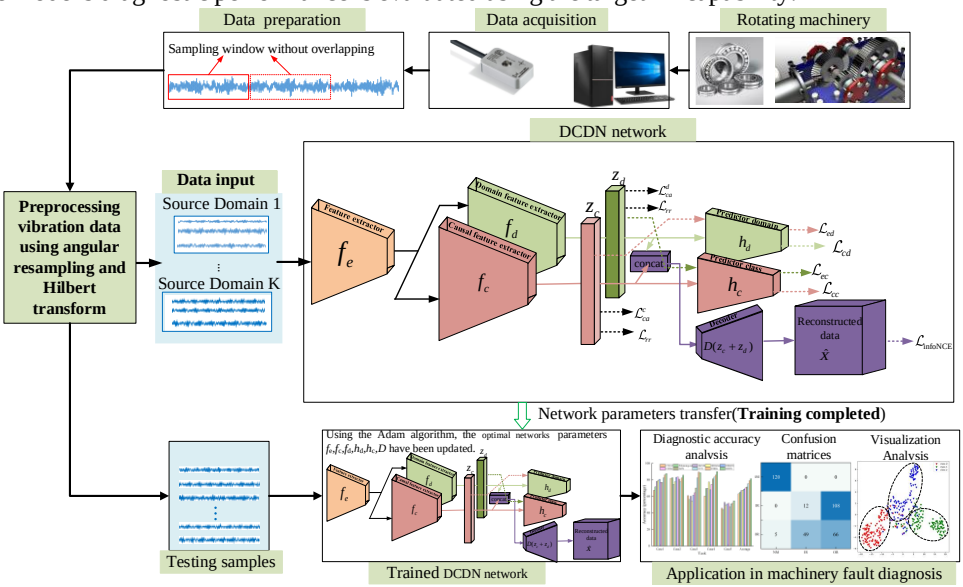

Fig. 3 The framework of the DCDN method.

As shown in the Fig.3, the DCDN network consists of six main components: the feature extraction layer (f_e), domain feature extraction layer (f_d), causal feature extraction layer (f_c), domain classifier (h_d), health state classifier (h_c), and decoder (D). Among these, f_e , f_c , and f_d are convolutional neural networks, with f_c and f_d sharing the same architecture. The health state classifier and domain classifier are fully connected layers. Specifically, multi-source domain vibration signals $\mathcal{D}_k^s = \{(x_{k,i}^s, y_{k,i}^s, d_{k,i})\}_{i=1}^{n_k^s}$ are first fed into f_e for overall feature extraction. The extracted features are then input into the f_c and f_d producing the non-causal factor Z_d and the causal factor Z_c, respectively. Here, Z_d contains domain-specific information, while Z_c captures the health state information of the bearings. To ensure that the extracted causal and domain-specific features effectively preserve the useful information from the original data, Z_c and Z_d are concatenated and fed into the decoder D to reconstruct the original vibration data. Finally, the extracted causal feature Z_c is input into the classifier h_c to predict bearing faults.

C. Disentangling causal and non-causal factors

In cross-machine fault diagnosis, a key assumption is that the learned causal factors Z_c are sufficient for fault classification, while the non-causal factors Z_d are needed for domain classification. Therefore, it is necessary to separate the unobservable causal factors Z_c from the non-causal factors Z_d. To achieve effective disentanglement, regularization must be applied to both the causal and non-causal factor extractors. The overall optimization objective can be formulated as the cross-entropy losses of the two tasks:

$$\min_{\theta_e,\theta_{cc},\theta_{cd}} \mathcal{L}_c(\theta_e,\theta_{cc},\theta_{cd}) = \mathcal{L}_{cc}(\theta_e,\theta_{cc}) + \alpha\mathcal{L}_{cd}(\theta_e,\theta_{cd}) \tag{3}$$

Where, $\theta_e, \theta_{cc}, \theta_{cd}$ represent the parameters of the shared feature extractor f_e , the causal factor extractor f_c , and the non-causal factor extractor f_d , respectively. α is a trade-off parameter, $\mathcal{L}_{cc}$ is the classification loss, and $\mathcal{L}_{cd}$ represents the domain classification loss.

Equation (3) is solved using Stochastic Gradient Descent (SGD), and the parameter updates for each part are computed as follows:

$$\theta_e^{(t+1)} = \theta_e^{(t)} - \eta^{(t)}\left(\frac{\partial\mathcal{L}_{cc}(\theta_e,\theta_{cc})}{\partial\theta_e} + \alpha\frac{\partial\mathcal{L}_{cd}(\theta_e,\theta_{cd})}{\partial\theta_e}\right) \tag{4}$$

$$\theta_{cc}^{(t+1)} = \theta_{cc}^{(t)} - \eta^{(t)}\left(\frac{\partial\mathcal{L}_{cc}(\theta_e,\theta_{cc})}{\partial\theta_{cc}}\right) \tag{5}$$

$$\theta_{cd}^{(t+1)} = \theta_{cd}^{(t)} - \eta^{(t)}\alpha\left(\frac{\partial\mathcal{L}_{cd}(\theta_e,\theta_{cd})}{\partial\theta_{cd}}\right) \tag{6}$$

Where $\eta^{(t)}$ is the learning rate at the t-th iteration, the parameters for the class classification and the domain classification loss, $\mathcal{L}_{cc}$ and $\mathcal{L}_{cd}$ are updated independently, with their respective loss functions adopting the cross-entropy form:

$$\mathcal{L}_{cc} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N_C}1\{y_i^c = j\}\log\frac{\exp(z_{c^{out}},i,j)}{\sum_{k=1}^{N_C}\exp(z_{c^{out}},i,k)} \tag{7}$$

$$\mathcal{L}_{cd} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}1\{y_i^d = j\}\log\frac{\exp(z_{d^{out}},i,j)}{\sum_{k=1}^{N_D}\exp(z_{d^{out}},i,k)} \tag{8}$$

Where $z_{c^{out}} = f_y(f_c(f_e(x)))$ and $z_{d^{out}} = h_d(f_d(f_e(x)))$ are the outputs of the class and domain classification networks, respectively.

Due to domain shift, pseudo-correlations may emerge between domain information and fault labels, which can be mitigated by controlling the domain variable. Thus, separating causal factors from non-causal factors is crucial, as non-causal

factors often inherit diagnostic knowledge with machine-specific biases. This aligns with the principles of SCM, which emphasize decoupling these factors. However, the challenge of effectively separating causal and non-causal factors remains unresolved[46].

To address this, we adopt a method based on metric learning and causal intervention to seek stable and invariant causal representations [20]. First, we introduce a causal aggregation loss that aggregates both causal and non-causal features, effectively disentangling the unobservable causal and non-causal factors. Second, inspired by entropy optimization techniques in[47], we design a maximum information entropy-guided strategy for causal task decomposition, which implicitly extracts highly distinct causal factors.

Specifically, the causal aggregation loss for causal factor Z_c and non-causal factor Z_d is computed using the following formula:

$$\mathcal{L}_{ca}^{c} = -\frac{\sum_{i_1=1}^{N_c}\sum_{j_1=1}^{n_b}\sum_{i_2=1}^{N_c}\sum_{j_2=1}^{n_b} 1\{y_{j_1}^{i_1}=y_{j_2}^{i_2}\}\left(z_{c\,j_1}^{\,i_1}\right)^{\top} z_{c\,j_2}^{\,i_2}}{\sum_{i_1=1}^{N_c}\sum_{j_1=1}^{n_b}\sum_{i_2=1}^{N_c}\sum_{j_2=1}^{n_b} 1\{y_{j_1}^{i_1}=y_{j_2}^{i_2}\}} + \frac{\sum_{i_1=1}^{N_c}\sum_{j_1=1}^{n_b}\sum_{i_2=1}^{N_c}\sum_{j_2=1}^{n_b} 1\{y_{j_1}^{i_1}\neq y_{j_2}^{i_2}\}\left(z_{c\,j_1}^{\,i_1}\right)^{\top} z_{c\,j_2}^{\,i_2}}{\sum_{i_1=1}^{N_c}\sum_{j_1=1}^{n_b}\sum_{i_2=1}^{N_c}\sum_{j_2=1}^{n_b} 1\{y_{j_1}^{i_1}\neq y_{j_2}^{i_2}\}} \tag{9}$$

$$\mathcal{L}_{ca}^{d} = -\frac{\sum_{j_1=1}^{n_b}\sum_{j_2=1}^{n_b}\sum_{i_1=1}^{N_c}\left(z_{d\,j_1}^{\,i}\right)^{\top} z_{d\,j_2}^{\,i}}{N_c n_b^2} + \frac{\sum_{i_1=1}^{N_c}\sum_{j_1=1}^{n_b}\sum_{i_2=1}^{N_c}\sum_{j_2=1}^{n_b} 1\{i_1\neq i_2\}\left(z_{d\,j_1}^{\,i_1}\right)^{\top} z_{d\,j_2}^{\,i_2}}{(N_c-1)N_c n_b^2} \tag{10}$$

$$\mathcal{L}_{ca} = \mathcal{L}_{ca}^{c} + \mathcal{L}_{ca}^{d} \tag{11}$$

Where n_b represents the batch size for each domain in the multi-source domains. The data $\{(\boldsymbol{x}_j^i, y_j^i)\,|\,i=1,2,\cdots,N_c; j=1,2,\cdots,n_b\}$ passes through a causal encoder f_c to obtain the causal factor $\boldsymbol{z}_{c\,j}^{\,i} = f_c(\boldsymbol{x}_j^i)$, and the same data $\{(\boldsymbol{x}_j^i, y_j^i)\,|\,i=1,2,\cdots,N_c; j=1,2,\cdots,n_b\}$ is processed by a domain encoder f_d to obtain the non-causal factor $\boldsymbol{z}_{d\,j}^{\,i} = f_d(\boldsymbol{x}_j^i)$.

We propose a maximum information entropy-guided causal task decomposition method designed to implicitly extract well-separated causal factors. Specifically, we assume that the extracted causal factors Z_c are free from domain-specific information. When these causal factors are input into the domain classifier f_d, the output should exhibit a certain degree of uncertainty, i.e., f_d should be confused by these factors. Similarly, if the non-causal factors Z_d extracted from the data completely exclude fault-related information, then when input into the fault classifier f_c, the output will be disorganized and chaotic. Therefore, we can assess the independence between causal and non-causal factors by measuring the degree of output chaos, with the specific calculation formula as follows:

$$\mathcal{L}_{ec} = -\frac{1}{N}\sum_{i=1}^{n_b}\sum_{j=1}^{K}\frac{\exp(z_{cd,i,j})}{\sum_{k=1}^{K}\exp(z_{cd,i,k})}\log\frac{\exp(z_{cd,i,j})}{\sum_{k=1}^{K}\exp(z_{cd,i,k})} \tag{12}$$

$$\mathcal{L}_{ed} = -\frac{1}{N}\sum_{i=1}^{n_b}\sum_{j=1}^{N_c}\frac{\exp(z_{dc,i,j})}{\sum_{k=1}^{K}\exp(z_{dc,i,k})}\log\frac{\exp(z_{dc,i,j})}{\sum_{k=1}^{K}\exp(z_{dc,i,k})} \tag{13}$$

$$\mathcal{L}_{et} = \mathcal{L}_{ec} + \beta\mathcal{L}_{ed} \tag{14}$$

Where β is the trade-off coefficient, $z_{cd} = h_c(z_d)$ denotes the process of inputting the obtained non-causal factors z_d into the fault classifier h_c, and $z_{dc} = h_d(z_c)$ represents the process of inputting the obtained causal factors z_c into the domain classifier h_d. As illustrated in Fig. 4, due to the entanglement of causal and non-causal factors in the original signal, we first use causal aggregation loss to aggregate samples with the same category labels. Then, by maximizing information entropy loss, we further obtain a purer causal mechanism, which indirectly ensures the separation of Z_c and Z_d. At this stage, entropy can be interpreted as a measure of the classifier's "confidence" in its predictions for a given input.

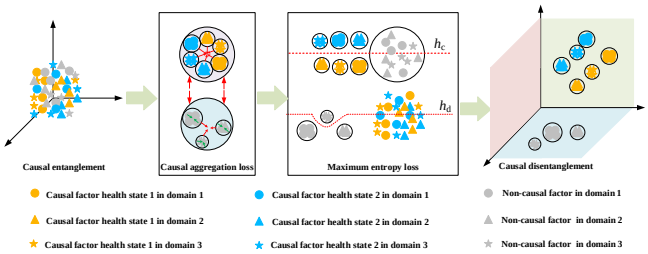

Fig. 4. The process of disentangling causal factors from non-causal factors via causal aggregation and maximum entropy optimization.

D. Completeness of features via contrastive learning

Although the aforementioned casual aggregation loss and maximum entropy loss assist in disentangling causal from non-causal factors, they may yield trivial solutions (for instance $\mathbf{z}_c = \mathbf{0}$ or $\mathbf{z}_d = \mathbf{0}$), resulting in the loss of valuable information. To ensure that both causal and non-causal features are representative while preserving as much useful information as possible, we draw inspiration from [48] and introduce contrastive learning to enhance feature learning. Specifically, we concatenate the output $\mathbf{z}_c$ from the causal feature extractor f_c and the output $\mathbf{z}_d$ from the non-causal feature extractor f_d. These concatenated features are then processed through a decoder D to reconstruct the original input. The decoding process is formulated as follows:

$$\hat{x}_i = D(z_{c,i} + z_{d,i}) \tag{15}$$

The mutual information between the original input signal and the decoded signal is given by:

$$I(x_i, \hat{x}_i) = \sum_{x_i,\hat{x}_i}\log\frac{p(x_i\,|\,\hat{x}_i)}{p(x_i)} \tag{16}$$

Due to the difficulty in directly modeling mutual information, we construct a density ratio function to represent it, as shown below:

$$f(x_i, \hat{x}_i) \propto \frac{p(x_i\,|\,\hat{x}_i)}{p(x_i)} \tag{17}$$

By maximizing mutual information, the extracted causal and non-causal features can retain as much information as possible from the original data. We approximate the density ratio function using the following formula:

$$f(x_i, \hat{x}_i) = x_i^{\top} \hat{x}_i \tag{18}$$

To maximize the density ratio function, the optimization objective is given by:

$$\mathcal{L}_{InfoNCE} = -\mathop{\mathbb{E}}_{x_i \in \mathcal{B}}\left[\log \frac{\exp(f(x_i, \hat{x}_i))}{\sum_{i \neq j, x_j \in \mathcal{B}} \exp(f(x_j, \hat{x}_i))}\right] \tag{19}$$

Where $\mathcal{B}$ denotes a mini-batch, $(x_i, \hat{x}_i)$ denotes positive pairs, and $(x_j, \hat{x}_i), i \neq j$ denotes negative pairs.

E. Independence analysis of causal and non-causal features

After disentangling the causal and non-causal features, two limitations remain. First, due to potential correlations among the feature dimensions in the causal vector, these correlations may negatively impact model predictions, leading to suboptimal discriminative representations. According to the Independent Causal Mechanisms principle [49], it is essential to pursue causal factors with highly separable features to improve the generalization of diagnostic knowledge. Second, separating causal and non-causal factors may impact the domain invariance of the causal features. To address this, we introduce a redundancy reduction loss, inspired by the redundancy reduction principle from [41], to regularize both the causal and non-causal factors. Let $z_c = \mathbb{R}^{n_b \times N_d}$ denote the causal feature matrix, and $z_d = \mathbb{R}^{n_b \times N_d}$ the non-causal feature matrix. The redundancy reduction loss is defined as:

$$\mathcal{L}_{rr} = \frac{\left\|(\mathbf{1} - \boldsymbol{E}) \odot (\mathbf{z}_c^{\top} \mathbf{z}_c)\right\|_F^2}{N_d (N_d - 1)} + \frac{\left\|(\mathbf{1} - \boldsymbol{E}) \odot (\mathbf{z}_d^{\top} \mathbf{z}_d)\right\|_F^2}{N_d (N_d - 1)} + \frac{\left\|(\mathbf{z}_c^{\top} \mathbf{z}_d)\right\|_F^2}{N_d^2} \tag{20}$$

Where N_d is the dimension of the feature vectors, $\mathbf{1} \in \mathbb{R}^{N_d \times N_d}$ is a matrix with all elements set to 1, $\boldsymbol{E} \in \mathbb{R}^{N_d \times N_d}$ is the identity matrix, $\odot$ denotes the Hadamard product, and $\|\bullet\|_F$ denotes the Frobenius norm.

F. Overall framework

The overall architecture of DCDN is illustrated in Fig. 5, and the overall optimization objective can be expressed as:

$$\mathcal{L}_{all} = \mathcal{L}_c + \gamma_1 \mathcal{L}_{ca} + \gamma_2 \mathcal{L}_{et} + \gamma_3 \mathcal{L}_{\text{infoNCE}} + \gamma_4 \mathcal{L}_{rr} \tag{21}$$

Where $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ are trade-off parameters. By minimizing this objective, the weight parameters of the DCDN can be updated.

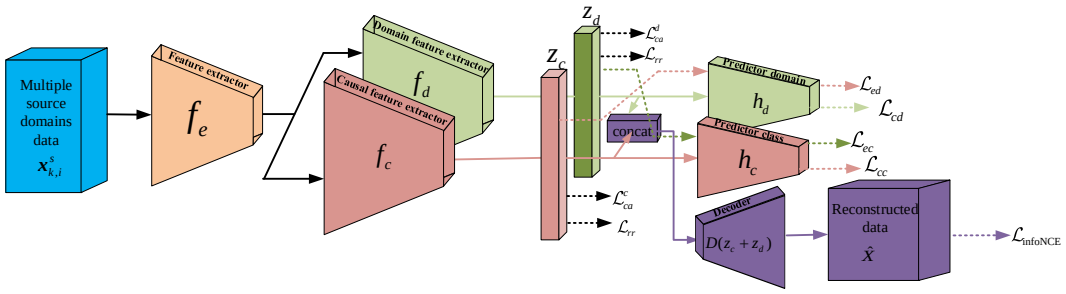

Fig. 5. The diagram of the proposed DCDN.

Let $\theta_{f_e}, \theta_{f_c}, \theta_{f_d}, \theta_{h_d}, \theta_{h_c}, \theta_D$ denote the parameters of $f_e, f_c, f_d, h_d, h_c, D$, respectively. Based on the chain rule of derivation, the parameter updating processes for each part of DCDN at *t*-th iteration can be expressed as:

$$\begin{aligned}
\theta_{f_e}^{(t+1)} &\leftarrow \theta_{f_e}^{(t)} - \eta^{(t)}\left(\frac{\partial \mathcal{L}_c}{\partial \theta_{f_e}} + \gamma_1 \frac{\partial \mathcal{L}_{ca}}{\partial \theta_{f_e}} + \gamma_2 \frac{\partial \mathcal{L}_{et}}{\partial \theta_{f_e}} + \gamma_3 \frac{\partial \mathcal{L}_{\text{infoNCE}}}{\partial \theta_{f_e}} + \gamma_4 \frac{\partial \mathcal{L}_{rr}}{\partial \theta_{f_e}}\right) \\
\theta_{f_c}^{(t+1)} &\leftarrow \theta_{f_c}^{(t)} - \eta^{(t)}\left(\frac{\partial \mathcal{L}_c}{\partial \theta_{f_c}} + \gamma_1 \frac{\partial \mathcal{L}_{ca}}{\partial \theta_{f_c}} + \gamma_2 \frac{\partial \mathcal{L}_{et}}{\partial \theta_{f_c}} + \gamma_3 \frac{\partial \mathcal{L}_{\text{infoNCE}}}{\partial \theta_{f_c}} + \gamma_4 \frac{\partial \mathcal{L}_{rr}}{\partial \theta_{f_c}}\right) \\
\theta_{f_d}^{(t+1)} &\leftarrow \theta_{f_d}^{(t)} - \eta^{(t)}\left(\frac{\partial \mathcal{L}_c}{\partial \theta_{f_d}} + \gamma_1 \frac{\partial \mathcal{L}_{ca}}{\partial \theta_{f_d}} + \gamma_2 \frac{\partial \mathcal{L}_{et}}{\partial \theta_{f_d}} + \gamma_3 \frac{\partial \mathcal{L}_{\text{infoNCE}}}{\partial \theta_{f_d}} + \gamma_4 \frac{\partial \mathcal{L}_{rr}}{\partial \theta_{f_d}}\right) \\
\theta_{h_d}^{(t+1)} &\leftarrow \theta_{h_d}^{(t)} - \eta^{(t)} \frac{\partial \mathcal{L}_c}{\partial \theta_{h_d}} \\
\theta_{h_c}^{(t+1)} &\leftarrow \theta_{h_c}^{(t)} - \eta^{(t)} \frac{\partial \mathcal{L}_c}{\partial \theta_{h_c}} \\
\theta_D^{(t+1)} &\leftarrow \theta_D^{(t)} - \eta^{(t)} \frac{\partial \mathcal{L}_{\text{infoNCE}}}{\partial \theta_D}
\end{aligned} \tag{22}$$

Where $\eta^{(t)}$ is the learning rate.

The pseudocodes for the proposed DCDN algorithm is given in Algorithm 1.

Algorithm 1 Deep Causal Disentanglement Network(DCDN)

Training stage:

Require: Training data $\mathcal{D}_k^s = \{(x_{k,i}^s, y_{k,i}^s, d_{k,i})\}_{i=1}^{n_k^s}$ from K multiple source domains, the batch size n_b, η is the learning rate, N_c denotes the number of health states, $\theta_{f_e}, \theta_{f_c}, \theta_{f_d}, \theta_{h_d}, \theta_{h_c}, \theta_D$ are the trade-off parameters.

Module: The untrained network (1)feature extractor f_e, (2) causal feature extractor f_c, (3) domain feature extractor f_d,(4) domain classifier h_d, (5) health states classifier h_c, (6) decoder D.

For i=1 in epoch **do**:

Step 1: Randomly sample a batch of data from training data $\mathcal{D}_k^s = \{(x_{k,i}^s, y_{k,i}^s, d_{k,i})\}_{i=1}^{n_k^s}$.

Step 2: Obtain the common information by a feature extractor f_e.

Step 3: Calculate the classification loss $\mathcal{L}_c$ by Eq.(3).

Step 4: Calculate the causal aggregation loss $\mathcal{L}_{ca}$ by Eq.(11).

Step 5: Calculate the maximizing entropy loss $\mathcal{L}_{et}$ by Eq.(14).

Step 6: Calculate the contrastive estimation loss $\mathcal{L}_{InfoNCE}$ by Eq.(19).

Step 7: Calculate the redundancy reduction loss by Eq. (20)

Step 8: Update the parameter of the proposed framework.

End

Return the trained network DCDN.

Testing stage:

Require: Test data $\mathcal{D}^t$ from the target domain.

Module: The optimal networks $f_e, f_c, f_d, h_d, h_c, D$ in the training stage.

Output: Classification results of $\mathcal{D}^t$.

V. Experiments

In this section, we validate the effectiveness and advancement of the proposed DCDN algorithm in the task of cross-bearing fault diagnosis. We conduct extensive comparisons with other SOTA methods and perform ablation studies, parameter sensitivity analyses, and visualizations of the proposed framework to further demonstrate the superiority of our method.

A. Dataset description

To evaluate the proposed method, we use five distinct time-series datasets of bearing faults. Since inner race and outer race faults are the most common types of bearing faults [24], we focus on three health states: normal condition (NC), inner race fault (IF), and outer race fault (OF). Notably, during model training, the target bearing remains unseen, meaning that the model has no prior information about the target bearing's data. Additionally, while domain shifts under different operating conditions for the same bearing are important, domain shifts between different bearings are even more challenging and critical. Therefore, we treat data from the same bearing under different operating conditions as one domain.

1) **CWRU Dataset**

The CWRU dataset[50], widely used in rolling bearing fault diagnosis, serves as a benchmark. Vibration signals in this dataset were collected via accelerometers mounted on the motor drive end and fan end, with a sampling rate of 12 kHz. For each operational condition, there is one healthy state and three faulty states: IF, OF, and ball fault (BF). Each faulty state is further classified into three severity levels based on fault dimensions of 0.007 inches, 0.014 inches, and 0.021 inches. In this paper, we use fault data related to IF and OF only.

2) **MFPT Dataset**

The MFPT dataset[51], provided by the Machinery Failure Prevention Technology (MFPT) Society, includes NC samples collected under a constant load of 270 lbs, with a sampling frequency of 97,656 Hz. The IF and OF samples were collected under three different loads (200 lbs, 250 lbs, and 300 lbs) at a sampling frequency of 48,828 Hz. Throughout the data collection, the machine maintains a constant rotational speed of 25 Hz.

3) **JNU Dataset**

The JNU dataset[52], provided by Jiangnan University, contains vibration signals from single-row spherical roller bearings in a centrifugal fan system. It includes one healthy condition and three fault modes: IF, OF, and rolling element defect. In this study, we focus on fault data related to IF and OF.

4) **UOTTAWA Dataset**

The UOTTAWA dataset[53], released by the University of Ottawa, contains vibration signals obtained from the Spectra Quest mechanical fault simulator. It includes data collected under various operating conditions, such as acceleration, deceleration, acceleration followed by deceleration, and deceleration followed by acceleration. The dataset covers bearings with three health conditions: healthy, IF, and OF. The sampling frequency is 200 kHz.

5) **PU Dataset**

The PU dataset[54], provided by the Paderborn University Bearing Data Center, consists of 32 sets of current and vibration signals. In this dataset, bearing faults are classified into real and artificial damage. Artificial damages are caused by processes such as electric discharge machining (EDM), drilling, and electric engraving. Accelerated life failure primarily manifests as fatigue pitting corrosion. The dataset includes vibration signals from seven bearing fault modes across four operating conditions. It is worth noting that detecting and diagnosing PU bearing faults presents significant challenges.

The detailed information for the five bearing datasets is presented in Table I. Due to variations in sampling frequencies and rotational speeds across these datasets, it is crucial to apply appropriate signal preprocessing steps. Preprocessing the input vibration signals through angular resampling and Hilbert transform can, to a certain extent, ameliorate the measurement challenges in cross-machine DGFD tasks. In the context of vibration signal analysis across different machines or operating conditions, vibration signals from various devices or under different operational statuses may exhibit differing sampling frequencies. Angular resampling, by converting the signal's sampling mode, allows for the comparison and analysis of diverse signals under a unified sampling frequency, thereby significantly mitigating the reduction in model prediction capability caused by domain shift. In practical measurements of vibration signals, discrepancies in signal distribution may arise due to sensor response and machine resonant frequencies, often obscuring the true characteristics of the signals and rendering signal processing and analysis difficult. Extracting the signal envelope using the Hilbert transform can alleviate such distribution discrepancies to a certain degree. This is because the envelope reflects the instantaneous amplitude variations of the signal, further elucidating the true dynamic characteristics of the signal.

TABLE I

Detailed Information of The Five Datasets.

Code	Dataset	Bearing types	Speeds(Hz)	Sampling rate(Hz)	Number of training samples	Number of test samples

A	CWRU	SKF6205	29.95/29.53/29.17/28.83	12,000	560	560
B	MFPT	/	25	48,828	300	300
C	JNU	N205 and NU205	600/800/1000	50,000	270	270
D	UOTTAWA	ER16K	time-varying	20,000	360	360
E	PU	ball bearing6203	900/1500	64,000	360	360

B. Compared methods

To demonstrate the superiority of the proposed method, we compared it with nine SOTA approaches:

1) **CNN-C**[38]: CNN-C minimizes intra-class variation and maximizes inter-class separability, addressing poor model performance under unseen operating conditions. It constructs an end-to-end fault diagnosis framework.

2) **CCDG**[39]: CCDG maximizes mutual information for similar categories across domains while minimizing it for different categories. This approach learns domain-independent class representations that can be transferred to new, unseen domains.

3) **DGNIS**[18]: DGNIS enables deep models to leverage domain-invariant features while preserving domain-specific structures, thereby enhancing predictive power across various domains.

4) **DDGFD**[55]: The first method to integrate prior diagnostic knowledge with deep domain generalization networks for cross-domain fault diagnosis of rolling bearings. This combination improves diagnostic accuracy and model generalization.

5) **IEDGNet**[3]: IEDGNet regularizes the discriminative structure of deep networks by incorporating both intrinsic and extrinsic generalization objectives. It uses triplet loss minimization on intrinsic multi-source data to achieve intra-class compactness and inter-class separability, leading to more generalized decision boundaries.

6) **ADN**[56]: ADN uses domain augmentation to expand the dataset and implements domain adversarial training to learn universal features. It also applies distance metric learning to further enhance robustness in fault classification without requiring test data availability.

7) **Whitening-Net**[22]: This framework introduces causal loss to impose regularization constraints on the network, enhancing its ability to mine causal relationships. To avoid the interference of domain-specific information, a whitening structure is proposed to eliminate domain noise, enabling the network to focus more on the causal relationships of signals.

8) **CCN**[23]: CCN emphasizes the consistency of fault causal representations over domain invariance. It introduces causal consistency loss to enforce consistency in deep latent variables, along with a collaborative training loss to convert personalized data into consistent representations.

9) **CDDG**[41]: CDDG designs causal aggregation loss to disentangle causal and non-causal factors, applies reconstruction loss to maintain information integrity, and uses redundancy reduction loss to learn efficient features.

In the design of the comparative experiments, we included several multi-source transfer learning methods. The first category consists of domain generalization models for unseen target data, where the source and target domains correspond to different operating conditions of the same machine. The models include ADN, CNN-C, IEDGNet, CCDG, and DGNIS algorithms. Although these methods are designed for unseen target data, they primarily focus on fault diagnosis of a single bearing under different operating conditions. While the probability distributions of fault types vary across different conditions, the differences are relatively small. The second category focuses on cross-machine fault diagnosis models, where the source and target domains come from different machines. The models include DDGFD, Whitening-Net, CCN, and CDDG, with the latter three algorithms being based on causality. For a fair comparison, all methods employed the same network architecture as DCDN. Each case was repeated five times to mitigate randomness in the diagnostic outcomes.

C. Experiment details

Batch size, learning rate, and epoch number significantly impact model accuracy. Additionally, it is important to explore the effect of output dimensionality on the performance of the causal feature extractor, as all four regularization losses are linked to the causal feature extraction layer (specifically, the number of channels in the last convolutional layer). Using Case 4 in Table IV as an example, each of the four parameters was tested 10 times while keeping the other hyperparameters fixed. The experimental results are presented in the Fig. 6.

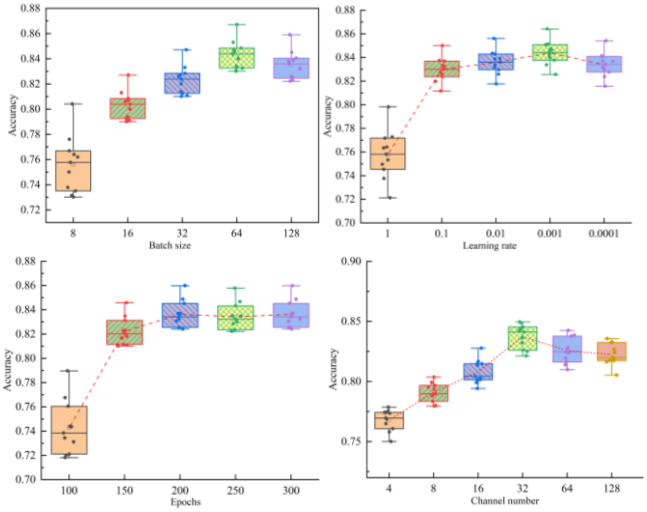

Fig. 6. Sensitivity of DCDN to the three training parameters and channel number

From the figure, it is evident that a batch size of 64, an initial learning rate of 0.001, 200 epochs, and a channel number of 32 yield the best performance. Table II presents the structural parameters of the causal/domain feature extractor, decoder, and health/domain classifier. Specifically, the notation (1, 32, 32) indicates that the input has 1 channel, the output has 32 channels, and the convolution kernel size is 32×1. The data length is set at 2560, with a batch size of 64. The model is trained for 400 iterations, starting with a learning rate of 0.001. To ensure training stability, the learning rate is reduced by a factor of 0.15 every 50 iterations. All parameters are set empirically using a train-validation-test split. The Adam optimizer is chosen, with hyperparameters set to 0.1, 0.1, 0.1, and 0.01, respectively. To further improve the reliability of the model and reduce dependence on specific data splits, we adopted a combination of leave-one-domain-out(LODO) and k-fold cross-validation. Specifically, for each domain, we introduced the concept of k-fold cross-validation by dividing each domain into five equally sized subsets (folds). Four of these subsets were randomly selected as the training set, and the model was trained and validated five times, ensuring stable performance across different training and test splits. The DCDN model is implemented using PyTorch 2.4.1 and runs on a system with an Nvidia GeForce RTX 4070S GPU.

TABLE II
THE NETWORK ARCHITECTURE PARAMETERS OF CDDG FOR CROSS-MACHINE BEARING FAULT DIAGNOSIS

	layer	Normalization	parameter	Activation	Output shape
Feature extraction	Conv1	BN	(1,32,32)	LReLU	(B,32,2529)
	MaxPool	/	2	/	(B,32,1264)
	Conv2	BN	(32,32,32)	LReLU	(B,32,1262)
	MaxPool	/	2	/	(B,32,616)
	Flatten	/	/	/	(B,32,19712)
Causal and domain feature extraction	Conv1	BN	(32,32,128)	LReLU	(B,32,19585)
	MaxPool	/	2	/	(B,32,9792)
	Conv2	BN	(32,32,64)	LReLU	(B,32,9729)
	MaxPool	/	2	/	(B,32,4864)
	Conv3	BN	(32,32,32)	LReLU	(B,32,4833)
	MaxPool	/	2	/	(B,32,2416)
	Conv4	BN	(32,32,16)	LReLU	(B,32,2401)
	MaxPool	/	2	/	(B,32,1200)
	Conv5	BN	(32,32,16)	LReLU	(B,32,1185)
	MaxPool	/	2	/	(B,32,592)
	Conv6	BN	(32,32,3)	LReLU	(B,32,590)
	MaxPool	/	2	/	(B,32,295)
	Conv7	BN	(32,32,7)	LReLU	(B,32,145)
	Flatten	/	/	/	(B,4640)
Decoder	Upsample	/	2	/	(B,64,290)
	Conv1	BN	(64,32,15)	LReLU	(B,32,284)
	Upsample	/	2	/	(B,32,568)
	Conv2	BN	(32,32,15)	LReLU	(B,32,584)
	Upsample	/	2	/	(B,32,1168)
	Conv3	BN	(32,32,31)	LReLU	(B,32,1200)
	Upsample	/	2	/	(B,32,2400)
	Conv4	BN	(32,32,63)	LReLU	(B,32,2464)
	Upsample	/	2	/	(B,32,4928)
	Conv5	BN	(32,32,3)	LReLU	(B,32,2560)
	Conv6	BN	(32,1,127)	LReLU	(B,1,2560)
Health class classifier	Linear1	/	(4640,2000)	/	(B,2000)
	Linear2	/	(2000,1000)	/	(B,1000)
	Linear3	/	(1000,300)	/	(B,300)
	Linear4	/	(300,3)	/	(B,3)
Domain classifier	Linear1	/	(4640,2000)	/	(B,2000)
	Linear2	/	(2000,1000)	/	(B,1000)
	Linear3	/	(1000,500)	/	(B,500)
	Linear4	/	(500,4)	/	(B,4)

D. Evaluation metrics

To evaluate the performance of different models in intelligent fault diagnosis, we use three primary metrics: accuracy, precision, and F1-score. Accuracy gives an overall measure of the model's diagnostic effectiveness, reflecting the proportion of correctly identified instances. Precision focuses on minimizing false positives, ensuring that the model's predictions are trustworthy and not too inclusive. The F1-score, which is the harmonic mean of precision and recall, provides a more balanced assessment by considering both the accuracy of predictions and the model's ability to detect faults comprehensively. This metric is particularly useful for refining fault diagnosis models, as it emphasizes both the reliability and completeness of the model's performance.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \times 100\%$$
$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (23)$$
$$F_1 = \frac{2TP}{2TP + FP + FN} \times 100\%$$

Where TP, TN, FP, and FN represent the number of true positive, true negative, false positive, and false negative samples, respectively.

E. Diagnosis result and discussion

1) **Hyperparameter Sensitivity Analysis**

The performance of the DCDN model is highly dependent on the selection of hyperparameters $\gamma_1, \gamma_2, \gamma_3, \gamma_4$, which control the weights of the causal aggregation loss, maximum entropy loss, contrastive estimation loss, and redundancy reduction loss. To examine the sensitivity of these hyperparameters to model performance, we conducted extensive experiments. The four hyperparameters of DCDN were tested across five different scenarios, and the results of the sensitivity analysis are shown in Fig. 7.

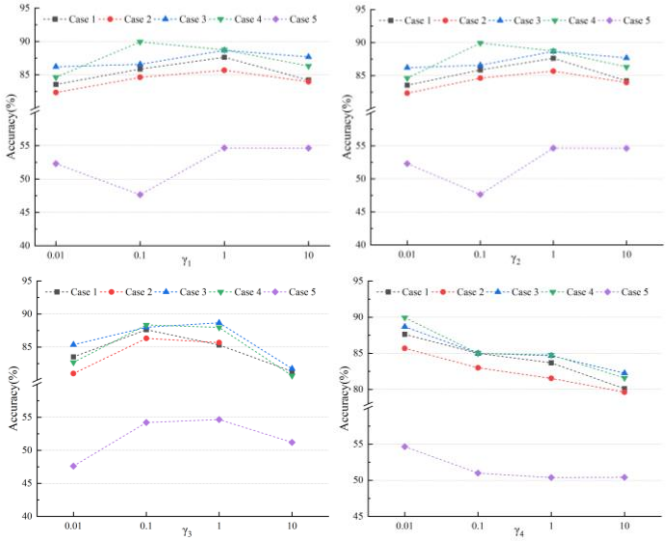

Fig. 7. Hyperparameter sensitivity analysis across five cases.

From the above figure, it is evident that the accuracy of the DCDN model decreases as γ_4 increases, suggesting that the redundancy reduction regularization should not be overly strong. This could be because certain redundant information is highly correlated with the model's generalization ability. Secondly, except for Case 5, the choice of γ_1 has minimal impact on most cases. When γ_1 is set to 1, the model achieves the highest accuracy. In general, different cases exhibit distinct patterns as γ_2, γ_3 varies, and the model performs optimally when $\gamma_2 = 0.1, \gamma_3 = 0.1$. In summary, the optimal values for the hyperparameters $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ are determined to be 1, 0.1, and 0.01, respectively. In summary, the overall performance does not follow a simple monotonic trend with respect to each balancing parameter. Instead, achieving optimal results requires precise calibration of the interaction among these parameters. This highlights the critical importance of well-balanced parameter configurations in fully leveraging the DCDN architecture's potential for cross-domain fault diagnosis.

2) **Cross-condition experiment on PU Dataset**

The PU dataset consists of four operating conditions, namely 'N09_M07_F10', 'N15_M01_F10', 'N15_M07_F04', and 'N15_M07_F10', where 'N' represents rotational speed, 'M' indicates load torque, and 'F' denotes radial force. For example, 'N15_M07_F04' corresponds to a rotational speed of 1500 rpm, a load torque of 0.7 Nm, and a radial force of 400 N. Each task categorizes the data based on fault type, causation, and severity. Eight classes are selected for analysis, including one NC class (K003), IR damage(KI01) and OR damage(KA03) of EDM, IR damage(KI07) and OR damage(KA06) of electric engraver, OR damage(KA08) of drilling, IR damage(KI18) and OR damage(KA16) of pitting. Table III presents the experimental results of the proposed DCDN along with those of other fault diagnosis methods, with the best results highlighted in bold.

From Table III and **Fig. 8**, the DCDN method demonstrates exceptional performance, surpassing the second-best method (CDDG) by an average accuracy margin of 3.8%. Due to the significant differences in data distribution across domains, most methods achieve an accuracy below 85%. Notably, even in the more complex operating environment of 'N09_M07_F10', the DCDN consistently exhibits better performance and maintains stability, achieving a 12.4% improvement in accuracy compared to the second-best CDDG method. Next section, we will further examine its adaptability to more challenging cross-machines fault diagnosis scenario to further validate the effectiveness.

TABLE III

THE ACCURACY (%) OF THE DIAGNOSTIC RESULTS FOR THE PROPOSED METHOD AND THE SOAT METHODS. (Note that the operating conditions 'N15_M01_F10', 'N15_M07_F04', 'N15_M07_F10', and 'N09_M07_F10', are represented by the letters I, J, K, and L, respectively.)

Method	Unseen Target Domain				
	I	J	K	L	Average
CNN-C	94.32	60.56	94.23	44.31	73.36
Whitening-net	92.7	56.06	93.56	37.47	69.95
DGNIS	96.49	73.37	97.09	56.68	80.91
IEDGNet	96.62	67.63	96.94	55.91	79.28

DDGFD	94.93	69.69	95.95	59.82	80.10
CCDG	95.94	65.21	96.71	49.59	76.86
ADN	95.14	73.92	96.12	63.07	82.06
CCN	95.41	76.73	95.79	60.2	82.03
CDDG	96.52	81.35	96.87	65.54	85.07
DCDN	**98.73**	**82.92**	**97.89**	**73.68**	**88.31**

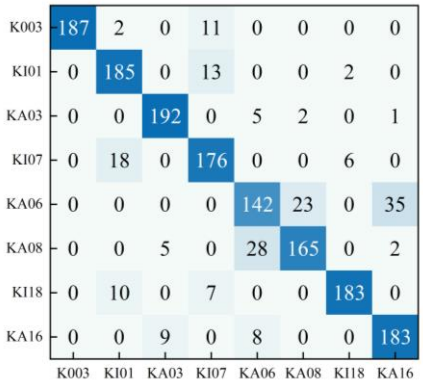

Fig. 8 The confusion matrix of DCDN algorithm on the PU dataset.

3) **Cross-machine fault diagnosis on five datasets**

The diagnostic performance of the proposed DCDN method, along with other current SOTA methods, is summarized in TABLE IV and Fig. 9, which show the average accuracy and corresponding standard deviations. The DCDN model achieves the highest average accuracy of 81.32%, surpassing all the compared approaches. Across all test cases, DCDN consistently ranks within the top two for accuracy. After applying LODO with k-fold cross-validation, we observed that the model's performance became more consistent across different training-test splits. Although the accuracy showed a slight decrease compared to the LODO method, the overall standard deviation of accuracy was reduced, indicating an improvement in the model's reliability.

Specifically, the CDDG method achieves an average accuracy of 78.96%, superior to other SOTA methods. This superior performance can be attributed to our method focuses on causal disentanglement, enabling a clear distinction between causal and non-causal factors across different domains. In contrast, traditional multi-source transfer learning relies more on extracting features from multiple source domains, which may contain substantial noise and redundant information. On the other hand, methods like CCN and Whitening-Net, which are also causal-based, focus more on maintaining consistency between latent variables and predictions, which limits their performance. Unlike the above three causal-based methods, our method effectively distinguishes between causal and non-causal features by optimizing the causal aggregation loss and maximizing the entropy loss, considering both direct and indirect perspectives. This approach significantly reduces the impact of domain-specific features, thereby enhancing the model's generalization ability and robustness. Additionally, by minimizing the contrastive estimation loss, the extracted causal features retain most of the key information from the original dataset. Meanwhile, the redundancy reduction loss is employed to minimize the correlation between the dimensions of the causal vectors, further mitigating the entanglement between causal and non-causal factors. CNN-C has the lowest average accuracy of only 63.23%, as it merely introduces center loss into the traditional CNN. Notably, the DGNIS method achieves the best result only in Case 4 but shows a significant gap compared to DCDN in the other four tasks, particularly in Case 3, where DGNIS is 28.5% lower than the proposed DCDN. This is due to the mixing of domain-specific and causal information, which leads to suboptimal feature extraction. Relying solely on domain-invariant regularization can cause the network to ignore key features, reducing generalization. DDGFD, with an average accuracy of 68.79%, demonstrates that penalizing only the samples near the decision boundary may lead to misguided model training. IEDGNet and ADN, both adversarial-based methods for DGFD, perform similarly, but ADN outperforms IEDGNet by improving consistency via deep metric learning. While ADN, CCN, and CDDG show relatively good results, their large standard deviations indicate high sensitivity to network initialization and data input order, which diminishes their robustness in industrial applications.

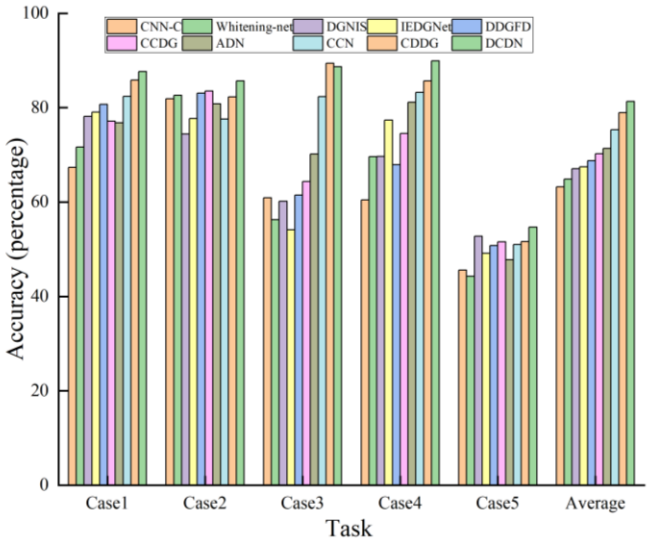

Fig. 9. Diagnostic accuracies of different methods for different task

4) **Ablation study**

In this section, ablation experiments were conducted to verify the effectiveness of each regularization loss. By removing $\mathcal{L}_{ca}$, $\mathcal{L}_{et}$, $\mathcal{L}_{\text{infoNCE}}$ and $\mathcal{L}_{rr}$ from DCDN, we obtained four ablation models: R1-DCDN_no_ca, R2-DCDN_no_et, R3-DCDN_no_infoNCE, and R4-DCDN_no_rr. The results of these ablation studies are shown in Table IV and **Fig. 10**.

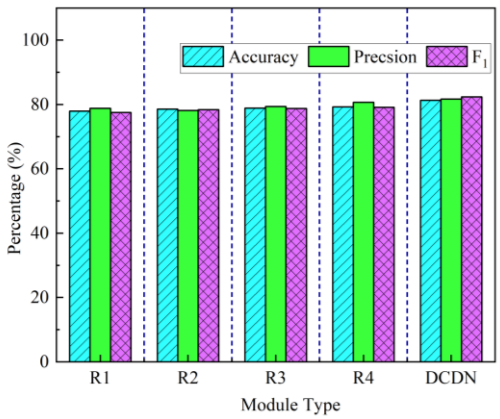

Fig. 10. The influence of different modules in the DCDN model on fault diagnosis effect.

As shown in the Table IV and Fig. 10, the R1 module—removal of the causal aggregation—has the most significant impact on the model's performance. Removing this module resulted in a decrease in accuracy, precision, and F1-score by 4.3%, 3.7%, and 6.27%, respectively. This is because, by introducing causal aggregation loss, causal and non-causal features are aggregated separately, effectively disentangling the unobservable causal and non-causal factors. Increasing the number of constraints imposed on the model leads to improved performance, highlighting the effectiveness of each proposed component. The various loss functions guide the model within a more constrained solution space, making it easier for the model to converge to the optimal solution. DCDN achieves the highest average accuracy, outperforming DCDN_no_ca, DCDN_no_et, DCDN_no_infoNCE, and DCDN_no_rr by 3.36%, 2.78%, 2.47%, and 2.10%, respectively. The four regularization losses have a positive impact on all five tasks. Thus, the results validate that the proposed causal aggregation loss, maximum entropy loss, contrastive estimation loss, and reconstruction loss contribute to the model's ability to learn stable causal features and generalized diagnostic knowledge. This is mainly because the four proposed regularization losses prompt the model to learn a good representation, thereby achieving higher-precision fault causal feature extraction. It is noteworthy that the contrastive prediction loss used in this paper is the InfoNCE loss rather than the traditional MSE loss. This is because MSE loss forces the model to reconstruct the original data accurately, ensuring minimal error for each data point. Such a strict constraint makes the model highly sensitive to outliers, potentially leading to over-adjustment of parameters to accommodate outliers, resulting in model instability and less ideal learned features. Unlike traditional point-wise error minimization, InfoNCE loss prioritizes maximizing the mutual information between the decoded and original data. This characteristic makes InfoNCE loss more robust to outliers, enabling the model to learn more stable and high-quality features.

TABLE IV

THE ACCURACY AND STANDARD DEVIATION (%) OF THE DIAGNOSTIC RESULTS FOR THE PROPOSED METHOD AND THE SOAT METHODS.(Note that "BCDE→A" means that datasets B, C, D, and E serve as source domains, while dataset A is the target domain. Other configurations follow a similar structure.) The top two results are highlighted in bold.

Method	Case1 BCDE→A	Case2 ACDE→B	Case3 ABDE→C	Case4 ABCE→D	Case5 ABCD→E	Average
CNN-C	67.36 (1.62)	81.88 (2.38)	60.88 (8.99)	60.46 (2.35)	45.57 (7.3)	63.23
Whitening-net	71.64 (3.12)	82.61 (0.99)	56.27 (2.58)	69.64 (7.65)	44.28 (6.01)	64.89
DGNIS	78.13 (2.65)	74.44 (3.73)	60.18 (3.58)	69.68 (6.38)	**52.77** (10.99)	67.04
IEDGNet	79.06 (1.82)	77.71 (5.68)	54.13 (5.27)	77.37 (6.71)	49.15 (3.97)	67.48
DDGFD	80.69 (3.43)	83.07 (1.83)	61.46 (1.46)	67.93 (4.57)	50.79 (5.34)	68.79
CCDG	77.16 (3.12)	83.51 (1.87)	64.35 (1.75)	74.51 (7.84)	51.57 (7.67)	70.22
ADN	76.79 (4.68)	80.82 (3.41)	70.17 (7.57)	81.14 (9.57)	47.79 (5.95)	71.34
CCN	82.38 (2.28)	77.61 (2.98)	82.36 (5.48)	83.27 (2.37)	51.02 (7.96)	75.33
CDDG	**85.83** (1.84)	**82.31** (2.63)	**89.39** (3.67)	**85.65** (3.72)	51.64 (3.89)	78.96
DCDN_no_ca	85.04 (1.21)	82.78 (1.78)	85.65 (3.52)	84.94 (4.52)	51.37 (3.64)	77.96
DCDN_no_et	86.19 (1.32)	83.98 (1.65)	85.63 (3.26)	86.21 (3.98)	50.68 (3.98)	78.54

DCDN_no_infoNCE	86.73 (1.08)	84.76 (1.98)	87.87 (2.68)	88.65 (4.65)	46.23 (4.32)	78.85
DCDN_no_rr	87.01 (0.65)	85.02 (1.38)	87.61 (2.54)	88.79 (4.02)	47.69 (4.01)	79.22
DCDN(LODO)	**87.63** (0.98)	**85.69** (1.32)	**88.68** (2.31)	**89.94** (3.58)	**54.66** (3.65)	**81.32**
DCDN(LODO with k-fold)	87.95 (0.32)	84.32 (0.56)	87.97 (1.32)	89.32 (1.54)	55.30 (1.48)	80.97

5) **Visualization**

To assess the classification performance of DCDN across different categories, Fig. 11 presents the confusion matrices for five cases. The figure reveals that DCDN can accurately distinguish between the NC and IF of the target dataset (CWRU) in Case 1, but a portion of OF samples (approximately 42%) are misclassified as NC and IF. In Case 2, nearly all normal and faulty samples in the target dataset (MFPT) are correctly classified, though about 38 inner race fault samples are misclassified as OF. From another perspective, the proposed method still demonstrates strong fault detection capabilities. When using the JNU dataset as the target domain, DCDN almost perfectly differentiates between NC and OF samples, with only 4 OF samples misclassified as IF. For the UOTTAWA dataset, nearly all samples are correctly identified. However, DCDN performs least well on the target dataset (PU), where it struggles to classify IF and OF correctly but still successfully distinguishes normal from faulty samples. The model achieves excellent diagnostic results primarily because our method focuses on learning causal representations, making the model becomes more stable and reliable when faced with different fault conditions. This is crucial for understanding and explaining fault patterns, particularly in industrial applications where operators need to clearly understand the basis for the model's decisions.

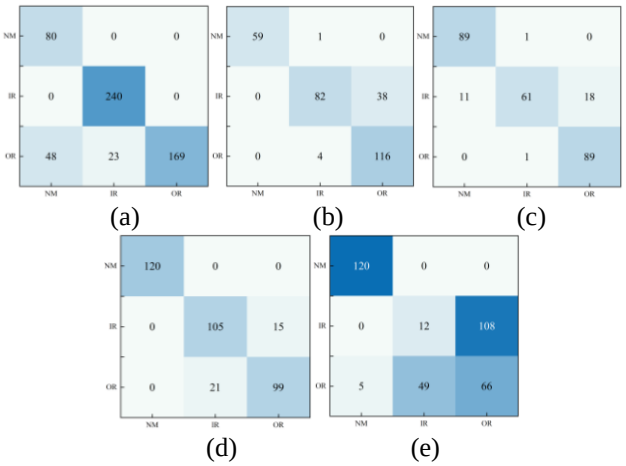

Fig. 11. The confusion matrices of DCDN. (a) Case1; (b) Case2; (c) Case 3; (d) Case4; (e) Case5.

To visually assess the effect of the proposed DGNSCA method on DGFD tasks, we utilized the t-SNE technique [57] to plot the fault features extracted by various methods. As illustrated in Fig. 12, the t-SNE plots show that the NC class consistently forms well-defined clusters, whereas the clustering of OF and IF is less distinct. In particular, in Fig. 9(e), OF and IF are indistinguishable, which corresponds to the findings in the confusion matrix presented in Fig. 8(e).

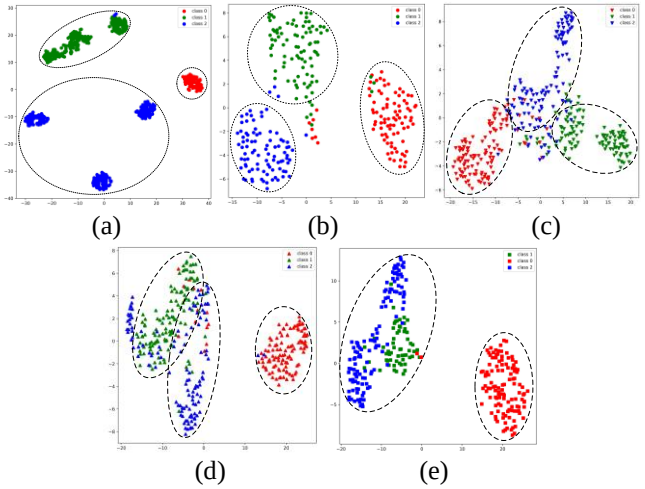

Fig. 12. t-SNE visualization results on five cases. (a) Case1; (b) Case2; (c) Case 3; (d) Case4; (e) Case5

VI. Conclusion

This paper presents a novel fault diagnosis approach, DCDN (Deep Causal Disentanglement Network), based on domain generalization. Unlike traditional statistical models, DCDN offers a distinctive approach by reconstructing the data generation process through a SCM, focusing on causal learning. The architecture of DCDN, with its causal and domain feature extractors, effectively decouples fault-related causal factors from domain-related non-causal factors. To enhance the model's performance, causal aggregation loss is applied to cluster samples with identical health and domain labels, while the maximum information entropy loss further refines the disentanglement of causal and non-causal factors. The InfoNCE loss is employed to avoid trivial solutions and preserve critical information from the original data. Additionally, the redundancy reduction loss suppresses feature redundancy and reduces spurious correlations between causal and non-causal factors. Extensive experiments across five bearing fault diagnosis cases demonstrate that DCDN outperforms nine SOTA methods, showcasing its superior capability in fault diagnosis.

References

[1] S. Lu, Z. Gao, Q. Xu, C. Jiang, A. Zhang, and X. Wang, "Class-Imbalance Privacy-Preserving Federated Learning for Decentralized Fault Diagnosis With Biometric Authentication,"

IEEE Transactions on Industrial Informatics, vol. 18, no. 12, pp. 9101-9111, 2022.

[2] Y. Chen, M. Rao, K. Feng, and M. J. Zuo, "Physics-Informed LSTM hyperparameters selection for gearbox fault detection," *Mechanical Systems and Signal Processing,* vol. 171, 2022.

[3] T. Han, Y. Li, and M. Qian, "A Hybrid Generalization Network for Intelligent Fault Diagnosis of Rotating Machinery Under Unseen Working Conditions," *IEEE Transactions on Instrumentation and Measurement,* vol. 70, pp. 1-11, 2021.

[4] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1D convolutional neural networks and applications: A survey," *Mechanical Systems and Signal Processing,* vol. 151, 2021.

[5] L. Jia, T. W. S. Chow, and Y. Yuan, "GTFE-Net: A Gramian Time Frequency Enhancement CNN for bearing fault diagnosis," *Engineering Applications of Artificial Intelligence,* vol. 119, 2023.

[6] X. Wu, Y. Zhang, C. Cheng, and Z. Peng, "A hybrid classification autoencoder for semi-supervised fault diagnosis in rotating machinery," *Mechanical Systems and Signal Processing,* vol. 149, 2021.

[7] X. Yu, B. Tang, and K. Zhang, "Fault Diagnosis of Wind Turbine Gearbox Using a Novel Method of Fast Deep Graph Convolutional Networks," *IEEE Transactions on Instrumentation and Measurement,* vol. 70, pp. 1-14, 2021.

[8] H. Zhao *et al.*, "Intelligent Diagnosis Using Continuous Wavelet Transform and Gauss Convolutional Deep Belief Network," *IEEE Transactions on Reliability,* vol. 72, no. 2, pp. 692-702, 2023.

[9] Z. Chen and W. Li, "Multisensor Feature Fusion for Bearing Fault Diagnosis Using Sparse Autoencoder and Deep Belief Network," *IEEE Transactions on Instrumentation and Measurement,* vol. 66, no. 7, pp. 1693-1702, 2017.

[10] 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, 2020.

[11] M. Ragab *et al.*, "Adversarial Multiple-Target Domain Adaptation for Fault Classification," *IEEE Transactions on Instrumentation and Measurement,* vol. 70, pp. 1-11, 2021.

[12] J. Jiao, J. Lin, M. Zhao, and K. Liang, "Double-level adversarial domain adaptation network for intelligent fault diagnosis," *Knowledge-Based Systems,* vol. 205, 2020.

[13] C. Zhao and W. Shen, "Dual adversarial network for cross-domain open set fault diagnosis," *Reliability Engineering & System Safety,* vol. 221, 2022.

[14] J. Tian, D. Han, H. R. Karimi, Y. Zhang, and P. Shi, "Deep learning-based open set multi-source domain adaptation with complementary transferability metric for mechanical fault diagnosis," *Neural Networks,* vol. 162, pp. 69-82, 2023.

[15] L. Chen, Q. Li, C. Shen, J. Zhu, D. Wang, and M. Xia, "Adversarial Domain-Invariant Generalization: A Generic Domain-Regressive Framework for Bearing Fault Diagnosis Under Unseen Conditions," *IEEE Transactions on Industrial Informatics,* vol. 18, no. 3, pp. 1790-1800, 2022.

[16] Y. Chen, S. Schmidt, P. S. Heyns, and M. J. Zuo, "A time series model-based method for gear tooth crack detection and severity assessment under random speed variation," *Mechanical Systems and Signal Processing,* vol. 156, 2021.

[17] P. Wu, X. Nie, and G. Xie, "Multi-sensor signal fusion for a compound fault diagnosis method with strong generalization and noise-tolerant performance," *Measurement Science and Technology,* vol. 32, no. 3, 2020.

[18] C. Zhao and W. Shen, "A domain generalization network combing invariance and specificity towards real-time intelligent fault diagnosis," *Mechanical Systems and Signal Processing,* vol. 173, 2022.

[19] J. Li, Y. Wang, Y. Zi, H. Zhang, and Z. Wan, "Causal Disentanglement: A Generalized Bearing Fault Diagnostic Framework in Continuous Degradation Mode," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 34, no. 9, pp. 6250-6262, 2023.

[20] F. Lv *et al.*, "Causality inspired representation learning for domain generalization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8046-8056.

[21] Q. Miao, J. Yuan, and K. Kuang, "Domain generalization via contrastive causal learning," *arXiv preprint arXiv,* 2022.

[22] J. Li, Y. Wang, Y. Zi, and Z. Zhang, "Whitening-Net: A Generalized Network to Diagnose the Faults Among Different Machines and Conditions," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 33, no. 10, pp. 5845-5858, 2022.

[23] J. Li, Y. Wang, Y. Zi, H. Zhang, and C. Li, "Causal Consistency Network: A Collaborative Multimachine Generalization Method for Bearing Fault Diagnosis," *IEEE Transactions on Industrial Informatics,* vol. 19, no. 4, pp. 5915-5924, 2023.

[24] S. Jia, Y. Li, X. Wang, D. Sun, and Z. Deng, "Deep causal factorization network: A novel domain generalization method for cross-machine bearing fault diagnosis," *Mechanical Systems and Signal Processing,* vol. 192, 2023.

[25] L. Guo, Y. Lei, S. Xing, T. Yan, and N. Li, "Deep Convolutional Transfer Learning Network: A New Method for Intelligent Fault Diagnosis of Machines With Unlabeled Data," *IEEE Transactions on Industrial Electronics,* vol. 66, no. 9, pp. 7316-7325, 2019.

[26] B. Yang, Y. Lei, F. Jia, N. Li, and Z. Du, "A Polynomial Kernel Induced Distance Metric to Improve Deep Transfer Learning for Fault Diagnosis of Machines," *IEEE Transactions on Industrial Electronics,* vol. 67, no. 11, pp. 9747-9757, 2020.

[27] C. Cheng, B. T. Zhou, G. J. Ma, D. R. Wu, and Y. Yuan, "Wasserstein distance based deep adversarial transfer learning for intelligent fault diagnosis with unlabeled or insufficient labeled data," *Neurocomputing,* vol. 409, pp. 35-45, Oct 2020.

[28] Y. Liu, Y. Wang, T. W. S. Chow, and B. Li, "Deep Adversarial Subdomain Adaptation Network for Intelligent Fault Diagnosis," *IEEE Transactions on Industrial Informatics,* vol. 18, no. 9, pp. 6038-6046, 2022.

[29] J. Jiao, J. Lin, M. Zhao, K. Liang, and C. Ding, "Cycle-consistent Adversarial Adaptation Network and its application to machine fault diagnosis," *Neural Networks,* vol. 145, pp. 331-341, 2022.

[30] Y. Shi, A. Deng, X. Ding, S. Zhang, S. Xu, and J. Li, "Multisource domain factorization network for cross-domain fault diagnosis of rotating machinery: An unsupervised multisource domain adaptation method," *Mechanical Systems and Signal Processing,* vol. 164, 2022.

[31] Y. B. Li, Y. Song, L. Jia, S. Y. Gao, Q. Q. Li, and M. K. Qiu, "Intelligent Fault Diagnosis by Fusing Domain Adversarial Training and Maximum Mean Discrepancy via Ensemble Learning," *Ieee Transactions on Industrial Informatics,* vol. 17, no. 4, pp. 2833-2841, Apr 2021.

[32] Y. Xiao, H. Zhao, and T. Li, "Learning Class-Aligned and Generalized Domain-Invariant Representations for Speech Emotion Recognition," *IEEE Transactions on Emerging Topics in Computational Intelligence,* vol. 4, no. 4, pp. 480-489, 2020.

[33] Y. Chong, C. Peng, C. Zhang, Y. Wang, W. Feng, and S. Pan, "Learning domain invariant and specific representation for cross-domain person re-identification," *Applied Intelligence,* vol. 51, no. 8, pp. 5219-5232, 2021.

[34] J. Wang *et al.*, "Generalizing to Unseen Domains: A Survey on Domain Generalization," *IEEE Transactions on Knowledge and Data Engineering,* pp. 1-1, 2022.

[35] Q. Qian, J. Zhou, and Y. Qin, "Relationship Transfer Domain Generalization Network for Rotating Machinery Fault Diagnosis Under Different Working Conditions," *IEEE Transactions on Industrial Informatics,* vol. 19, no. 9, pp. 9898-9908, 2023.

[36] S. Angarano, M. Martini, F. Salvetti, V. Mazzia, and M. Chiaberge, "Back-to-Bones: Rediscovering the role of backbones in domain generalization," *Pattern Recognition,* vol. 156, 2024.

[37] Z. Zhu, G. Peng, Y. Chen, and H. Gao, "A convolutional neural network based on a capsule network with strong generalization for bearing fault diagnosis," *Neurocomputing,* vol. 323, pp. 62-75, 2019.

[38] Y. Yang, J. Yin, H. Zheng, Y. Li, M. Xu, and Y. Chen, "Learn Generalization Feature via Convolutional Neural Network: A Fault Diagnosis Scheme Toward Unseen Operating Conditions," *IEEE Access,* vol. 8, pp. 91103-91115, 2020.

[39] M. Ragab *et al.*, "Conditional Contrastive Domain Generalization for Fault Diagnosis," *IEEE Transactions on Instrumentation and Measurement,* vol. 71, pp. 1-12, 2022.

[40] P. Cui and S. Athey, "Stable learning establishes some common ground between causal inference and machine learning," *Nature Machine Intelligence,* vol. 4, no. 2, pp. 110-115, 2022.

[41] L. Jia, T. W. S. Chow, and Y. Yuan, "Causal Disentanglement Domain Generalization for time-series signal fault diagnosis," *Neural Networks,* vol. 172, 2024.

[42] X. Zhang, P. Cui, R. Xu, L. Zhou, Y. He, and Z. Shen, "Deep stable learning for out-of-distribution generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5372-5382.

[43] B. Scholkopf *et al.*, "Toward Causal Representation Learning," *Proceedings of the IEEE,* vol. 109, no. 5, pp. 612-634, 2021.

[44] J. Pearl, *Causality*. Cambridge university press, 2009.

[45] X. Sun *et al.*, "Recovering latent causal factor for generalization to distributional shifts," *Advances in Neural Information Processing Systems,* vol. 34, pp. 16846-16859, 2021.

[46] T. Wang, J. Huang, H. Zhang, and Q. Sun, "Visual commonsense representation learning via causal inference," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 378-379.

[47] A. Ma, J. Li, K. Lu, L. Zhu, and H. T. Shen, "Adversarial Entropy Optimization for Unsupervised Domain Adaptation," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 33, no. 11, pp. 6263-6274, 2022.

[48] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:.03748,* 2018.

[49] P. Ping, C. Huang, W. Ding, Y. Liu, M. Chiyomi, and T. Kazuya, "Distracted driving detection based on the fusion of deep learning and causal reasoning," *Information Fusion,* vol. 89, pp. 121-142, 2023.

[50] W. A. Smith and R. B. Randall, "Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study," *Mechanical Systems and Signal Processing,* vol. 64-65, pp. 100-131, 2015.

[51] E. Bechhoefer, "A quick introduction to bearing envelope analysis," *Green Power Monit. Syst,* 2016.

[52] K. Li, X. Ping, H. Wang, P. Chen, and Y. Cao, "Sequential Fuzzy Diagnosis Method for Motor Roller Bearing in Variable Operating Conditions Based on Vibration Analysis," *Sensors,* vol. 13, no. 6, pp. 8013-8041, 2013.

[53] H. Huang and N. Baddour, "Bearing vibration data collected under time-varying rotational speed conditions," *Data in Brief,* vol. 21, pp. 1745-1749, 2018.

[54] C. Lessmeier, J. Kimotho, D. Zimmer, and W. Sextro, "KAt-DataCenter, chair of design and drive technology, paderborn university," ed, 2019.

[55] H. Zheng, Y. Yang, J. Yin, Y. Li, R. Wang, and M. Xu, "Deep Domain Generalization Combining A Priori Diagnosis Knowledge Toward Cross-Domain Fault Diagnosis of Rolling Bearing," *IEEE Transactions on Instrumentation and Measurement,* vol. 70, pp. 1-11, 2021.

[56] X. Li, W. Zhang, H. Ma, Z. Luo, and X. Li, "Domain generalization in rotating machinery fault diagnostics using deep neural networks," *Neurocomputing,* vol. 403, pp. 409-420, 2020.

[57] X. Li, H. Jiang, R. Wang, and M. Niu, "Rolling bearing fault diagnosis using optimal ensemble deep transfer network," *Knowledge-Based Systems,* vol. 213, p. 106695, 2021.

Chaochao Guo was born in Henan Province, China in 1990. He is currently a PhD candidate of college of civil aviation, Nanjing University of Aeronautics and Astronautics, Nanjing(NUAA), China. His research interests include aircarft engine fault diagnosis and prognostics, deep learning, and health management.

Youchao Sun was born in Henan Province, China in 1964. He is now a professor of college of civil aviation, NUAA, Nanjing, China. He is the author of seven books, more than 170 articles, and more than 20 patents. His research interests focus on reliability engineering, risk assessment and airworthiness technology of civil aircraft.

Rourou Yu was born in Jiangsu Province, China in 1996. She is currently a PhD candidate of college of civil aviation, Nanjing University of Aeronautics and Astronautics, Nanjing, China. Her research interests include aircraft safety monitoring and early warning, fault diagnosis, etc.

Xinxin Ren was born in Jiangsu Province, China in 1995. He is currently a PhD candidate of college of civil aviation, Nanjing University of Aeronautics and Astronautics, Nanjing, China. His research interests include reliability engineering, operational safety analysis and risk assessment of civil aircraft.