CINet: Causal Intervention Network for Cross-Component Few-Shot Fault Diagnosis

Jiahan Zhu, Juan Xu*, Qile Ren, Mingguang Dai, Xiaohui Yuan*

***Abstract*—Existing few-shot cross-component fault diagnosis methods primarily focus on the correlation between input data and fault classes, neglecting causal relationships. This limits the model's ability to separate and eliminate confounding factors, limiting the improvement of cross-component prediction accuracy. To address this issue, this paper proposes a Causal Intervention Network for Cross-Component Few-Shot Fault Diagnosis (CINet), which constructs a causal structure model to perform causal decomposition, extracting and decoupling the instrumental variable, confounding variable, and adjustment variable from vibration signals, thus enabling the modeling of cross-component causal relationships, which enhances diagnostic accuracy under few-shot conditions. Specifically, the CINet is composed of three main modules: a feature encoding module, a causal disentanglement module, and a relation metric module, jointly optimizing the fault diagnosis and relation metric selection loss functions through multi-task learning. Experimental results on multiple fault diagnosis datasets demonstrate that the CINet significantly outperforms existing methods, especially in handling cross-component fault diagnosis problems, particularly in few-shot scenarios, by better capturing causal relationships and improving prediction accuracy and model interpretability.**

***Index Terms*—Causal Disentanglement, Structural Causal Model, Few-Shot Learning, Fault Diagnosis.**

I. Introduction

In industrial systems, mechanical fault diagnosis plays a critical role in ensuring equipment reliability and operational safety [1, 2]. Traditional methods have been extensively studied in this field, such as the multivariate statistical process control (MSPC) method based on linear correlation [3–6]. Denkena et al. [7] studied Principal Component Analysis (PCA) for sensor fusion to monitor preload loss in single nut ball screws. Pandhare et al. [8] focused on cross-sensor domain adaptation to diagnose discrete variations in preload levels and backlash. Additionally, the $PCA - T^2$ method [9] has been applied to detect early-stage degradation and quantify fault severity. Such methods rely on a large amount of Low-frequency yet large-scale continuous multivariate data and cannot handle the problems of disjoint fault categories and scarce samples across components. The inherent diversity of mechanical components (such as bearings, gears, and motors) leads to completely different types of faults for different components. Moreover, the sample data from different components show significant differences in distribution [10]. Under such circumstances, the applicability of traditional methods is limited. To address the problem of scarce samples in industrial practice, few-shot learning (FSL) has received extensive attention in mechanical fault diagnosis. A valuable practice is to build high-performance fault diagnosis models using components with abundant labeled data, then transfer them to target different components with scarce samples for fault diagnosis. This scenario is known as few-shot cross-component fault diagnosis. This problem features disjoint classes and significant cross-domain distribution shifts between training and testing datasets, representing a critical research direction in fault diagnosis and a pressing challenge in real-world industrial applications.

Current studies on few-shot cross-component fault diagnosis approaches mostly adopt conventional FSL frameworks, e.g., metric-learning networks as backbone networks [11–13]. Existing few-shot learning approaches primarily exploit statistical correlations between input data and output fault classes, for example, by using a similarity matrix for graph embedding to reduce the adverse impact of noisy samples [14], by employing attention mechanisms to weight-fuse signal features [15, 16], and by aligning the joint distributions of source and target domains [17], while neglecting to disentangle causal representations governed by underlying causal mechanisms [18]. This oversight compromises model interpretability [19] and retains spurious correlations that hinder generalization capability in cross-component scenarios with scarce target samples.

Causal theory offers a promising tool for addressing these limitations by modeling the intrinsic causal mechanisms of faults. Methods based on causal features can effectively eliminate the interference of confounding factors such as noise, thereby enhancing the interpretability and robustness of fault diagnosis models. Uchida et al. [20] proposed a causal plot based on LiNGAM for fault diagnosis in MSPC. However, their method focuses on discovering causal structures among multiple process variables rather than disentangling representations from a single vibration signal, and it is mainly designed for continuous process fault cause localization, requiring suf-

This work was supported in part by the National Natural Science Foundation of China under Grant 52375089, Dreams Foundation of Jianghuai Advance Technology Center under Grant 2023-ZM01J003, Open Foundation of State Key Laboratory of High-end Compressor and System Technology under Grant SKL-YSJ202307, and Open Foundation of Henan Key Laboratory of High-performance Bearings under Grant ZYSKF202307.

JiaHan Zhu is with School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China (2023170636@mail.hfut.edu.cn).

Juan Xu is with School of Electrical Engineering and Automation, Anhui University, Hefei, 230061, China, and Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, China (xujuan@ahu.edu.cn).

Qile Ren is with State Key Laboratory of High-end Compressor and System Technology, Hefei, China (renqile@hgmri.com).

Mingguang Dai is with Jianghuai Advance Technology Center, Hefei, China (dmgjxdyq@foxmail.com).

Xiaohui Yuan is with College of Engineering, University of North Texas, USA (Xiaohui.Yuan@unt.edu).

* corresponding authors

ficient sample data. To the best of our knowledge, causal theory–guided FSL in discrete mechanical fault diagnosis remains largely unexplored, with only a few studies reported to date. Chang et al. proposed CIS2N for mechanical fault diagnosis [21], which decomposes causal/non-causal feature sets via sample intervention and optimizes intra-causal feature-pair distances to achieve shift sparsity. Notably, this study requires a large number of samples for computational causal feature estimation, and the proposed Structural Causal Model (SCM) violates the assumption of domain-factor independence. Therefore, it may be ineffective in few-shot cross-component scenarios.

Thereby, in our previous work, we proposed a causal intervention relation network for cross-component few-shot fault diagnosis via a backdoor adjustment strategy [22]. As shown in Fig. 1(a), we constructed the prior SCM for FSL-based fault diagnosis and then identified the meta-training knowledge as the confounding factors (C) and eliminated them using backdoor adjustment-based causal intervention methods. However, this method only considers confounding factors for SCM and fails to incorporate other variables such as instrumental variables (I) and adjustment variables (A), leading to the omission of certain causal paths.

Compared with CIRNet, which models pretraining knowledge as a confounder, the proposed SCM explicitly incorporates three causal variables (instrumental, confounding, and adjustment), resulting in a complete causal-variable specification and fewer omitted causal paths. To address these limitations, we develop a causal framework for FSL-based fault diagnosis, depicted in Fig. 1(b). This framework systematically incorporates additional variables through explicit causal modeling, contrasting fundamentally with existing approaches in both causal assumptions and variable dependency structures.

Under the meta-task learning framework, it is commonly assumed that the causal relationship between features and fault classes remains invariant across components, and that these features exhibit a degree of independence within the causal graph [23, 24].

Building upon this SCM, our methodology employs systematic causal decomposition within cross-component diagnostic networks. This enables precise isolation of instrumental variables, confounding factors, and adjustment parameters directly from vibrational data, while establishing explicit cross-component causal relationships to optimize diagnostic performance in the data-scarce scenario.

The contributions of this paper are as follows:

1) We propose a novel causal intervention network, termed CINet, for cross-component few-shot fault diagnosis. By modeling the underlying causal structure and decomposing the vibration signal into distinct causal representations, CINet separates cross-component causal factors, improves model interpretability, and alleviates data scarcity in target components.
2) Within the relational network framework, we develop three causality-constrained loss functions to enforce: (1) instrumental variable I-fault label Y independence, (2) the adjustment variable A -relation metric T independence, and (3) the elimination of the error influence of confounding variable C on relational scores. This integrated optimization strategy simultaneously optimizes fault classification and metric learning objectives while enforcing cross-component causal feature separation, significantly enhancing model generalizability.
3) We conduct extensive experiments on multiple cross-component fault diagnosis datasets, and demonstrate improved performance under different task settings. The experimental results show that CINet can better capture the causality and improve the diagnosis accuracy in few-shot scenarios.

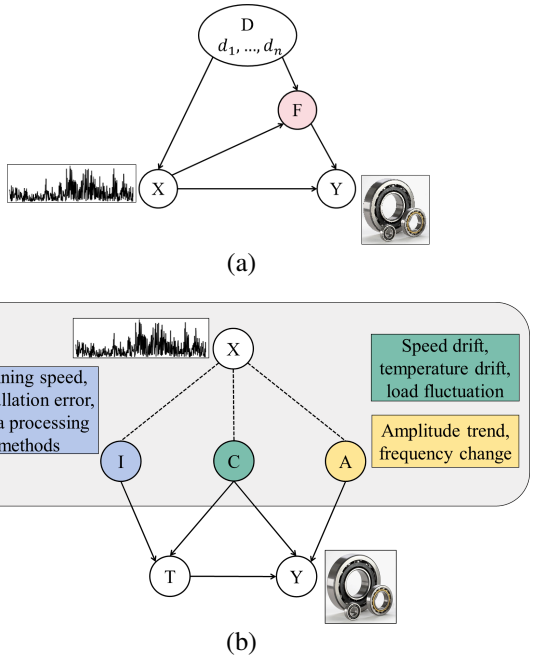


Fig. 1: Different SCMs for Cross-Component Few-Shot Fault Diagnosis. (a) The previous SCM, where X denotes the raw vibration signal, F represents fault-related causal features, D denotes meta training knowledge identified as the confounding factors, and Y indicates the fault class. (b)Enhanced SCM, where X is the vibration signal, Y is the fault class, I denotes the instrumental variables, C represents confounding variables, A is the adjustment variables, and T denotes the relation metric.

The remainder of this paper is organized as follows. Section II reviews related work on causal modeling and causal disentanglement. Section III defines the problem addressed by the proposed model and introduces the causal structural model and the proposed framework. Section IV presents and analyzes the experimental results, including comparisons with state-of-the-art methods. Section V concludes the paper with a summary.

II. Related Work

A. Causal Modeling

Traditional machine-learning pipelines (e.g., support vector machines or shallow neural networks) typically depend on sufficient labeled data to learn stable decision boundaries.

However, in few-shot fault diagnosis, the available training samples are extremely limited, making these correlation-based models prone to overfitting the small training set and thus generalizing poorly under distribution shifts (e.g., cross-component transfer) [25, 26]. In contrast, causal theory goes beyond correlation by incorporating intervention and counterfactual reasoning to systematically identify causal mechanisms [27, 28]. Specifically, if intervening on a variable X leads to a change in another variable Y, then X can be regarded as a cause of Y. These causal relationships are expected to be more invariant across environments, offering advantages over conventional statistical dependencies that typically rely on i.i.d. assumptions.

There are two common approaches to causal modeling: (1) the potential outcome framework [29], which studies average causal effects when the treatment and outcome variables are specified; and (2) SCMs, which use graphs to represent causal relationships while supporting effect estimation and, in some settings, structure learning. The potential outcome framework can recover causal effects via well-defined treatment manipulations, but it is less suited to explicitly representing causal pathways or visualizing causal networks [30]. In contrast, SCMs facilitate uncovering and reasoning about causal mechanisms among variables. In fault diagnosis, multiple factors interact in complex ways, and SCMs provide a principled way to capture, interpret, and quantify these causal dependencies.

An SCM is often represented as a directed acyclic graph (DAG) whose vertices correspond to variables $X_1, \ldots, X_n$. Directed edges such as $A \to B$ indicate that A is a direct cause of B. The causal structure (i.e., the causal graph) is denoted by $G = (V, E)$, where $V = \{X_1, \ldots, X_n\}$ and E encodes assumptions about direct causal relations among variables [31]. In this paper, we employ SCMs to model the cross-component few-shot fault diagnosis problem.

B. Causal Disentanglement

Causal disentanglement is the process of decomposing observed data representations into distinct latent factors that correspond to different causal variables in the causal structure [32, 33]. By mapping observational data into a latent space structured by causal relationships, causal disentanglement explicitly models the underlying dependencies among variables, distinguishing features that correspond to true causal factors from those associated with confounding effects or spurious correlations [34, 35].

Existing causal disentanglement methods can be broadly divided into two categories: those based on *causal structural models* [36, 37] and those based on *causal discovery techniques* [38, 39].

Methods based on causal structural models typically construct causal graphs using prior knowledge and leverage SCMs to guide the disentanglement of causal features. For instance, the IFSL framework proposed by Yue et al. [40] employs backdoor adjustment to eliminate confounding effects introduced by pre-trained knowledge, allowing the model to focus on features with genuine causal effects on labels. By using SCMs to remove confounders and decouple causal structures, IFSL effectively enhances the model's generalization ability. Another representative method, CDN [41], explicitly models the causal relationship between domain and label information. It assumes that both domain-related and causal features influence the signal, but only causal features affect class labels. Based on this assumption, CDN guides the model to disentangle domain-invariant causal features from domain-specific ones, thereby improving cross-domain classification accuracy.

In contrast, methods based on causal discovery aim to learn causal structures directly from data without relying on strong prior assumptions. These approaches typically apply causal discovery algorithms to construct causal graphs, which are then used to guide causal feature disentanglement. For example, Liu et al. [42] proposed a method that integrates convolutional networks with graph attention mechanisms to derive causal graphs from fault-free data using Granger causality analysis. This framework not only detects faults but also identifies their root causes based on the learned causal structure.

III. Methodology

A. Notations and Problem Definition

In few-shot, cross-component fault diagnosis problems, there are typically two distinct mechanical components: Component A and Component B. Component B is the target component, while training samples are collected from Component A. Due to the limited sample size of the target component, our method employs metric-based meta-learning. Meta-task learning allows the model to generalize across tasks by learning from a distribution of related tasks, rather than individual samples. Each meta-task is constructed as a classification problem consisting of several classes and a few samples per class [43]. Consequently, the fault classification problem can be formulated as a c-way k-shot meta-task learning problem [44].

Given the few-shot scenario of the target component, we construct c-way k-shot tasks from the training dataset D_A, where $Y_A = \{y^1, y^2, \ldots, y^d\}$ denotes the class set. We randomly select c classes to form the support set $S = \{(x_i, y_i)\}_{i=1}^{c\times k}$, with x_i and y_i being the i-th sample and its label. From the remaining samples of these c classes, m samples per class are selected to construct the query set $Q = \{(x_j, y_j)\}_{j=1}^{c\times m}$.The support set S and query set Q together form a c-way k-shot meta-training task. For each query sample, we compute weighted relation scores between its extracted features and the average features (prototypes) of each class in the support set. As each of the $c \times m$ query samples is compared with all c prototypes, this yields a total of $c^2 \times m$ prototype-query feature pairs per task. The final label is assigned according to the highest relation score.

For the target component B, we denote its dataset as D_B with the class set $Y_B = \{\hat{y}^1, \hat{y}^2, \ldots, \hat{y}^c\}$. The support set consists of c classes, each containing k labeled samples: $\hat{S} = \{(\hat{x}_i, \hat{y}_i)\}_{i=1}^{c\times k}$, where $\hat{x}_i$ and $\hat{y}_i$ represent the i-th sample and its label, respectively, drawn from the limited labeled samples of the target component. The remaining unlabeled samples of the target component form the test set $\hat{T} = \{(\hat{x}_j)\}_{j=1}^{n}$, where

$\hat{x}_j$ represents the j-th sample selected from a larger pool of unlabeled instances. The support set $\hat{S}$ and test set $\hat{T}$ together form a meta-testing task. For each test sample, features are extracted, and a weighted relation score is computed using the average features of each class in the support set. The test sample is then classified into the class with the highest relation score:

$$\hat{y}_j = \arg\max_{\hat{y}_i} \hat{r}_{i,j}. \tag{1}$$

The notations used throughout this paper are summarized in Table I.

TABLE I: GLOSSARY OF NOTATIONS

Notations	Description
$\mathcal{S}$	Support set
$\mathcal{Q}$	Query set
c	Number of classes (ways)
k	Number of support samples per class (shots)
x_i, y_i	Input sample and corresponding label of support set
x_j, y_j	Input sample and corresponding label of query set
$I(\cdot), C(\cdot), A(\cdot)$	Representation networks for instrumental and confounding/adjustment variables, respectively
$\overline{C(x_i)}$, $\overline{A(x_i)}$	Class-i averaged confounding features and averaged adjustment features
$G_A(\cdot)$	Relation metric module conditioned on A
$\hat{t_j}$	Predicted hard type/index of relation metric function for sample j
$t_{j,l}$	Soft classification probability over the l-th relation metric function for sample j
p_i, p_j	Concatenated prototype features of class i and features of query sample j
$\mathcal{A}(\cdot)$	Attention module
$T_l(\cdot)$	l-th relation metric function
$r_{i,j}$	Relation score between class i and query sample j
$r_{i,j,l}$	Relation score computed by T_l from class i and query sample j
$\mathbb{1}(\cdot)$	Indicator function (1 if true, 0 otherwise)
$\|\cdot\|^2$	Squared Euclidean distance
$\mathrm{disc}(\cdot)$	Distributional distance under different $\hat{t}$
$\oplus$	Feature concatenation operator

B. Structural Causal Model

The causal framework for few-shot cross-component fault diagnosis is depicted in Fig. 1. We assume that the vibration signal X of a component can be approximately decomposed into three latent variables, $\{I, C, A\}$, representing the instrumental variable I, confounding factor C, and adjustment variable A.

Instrumental Variable (I): The instrumental variable solely influences the relation metric T, determining the appropriate function for the vibration signal data. It does not directly affect the fault class Y, but influences Y indirectly through T. Examples include running speed, installation errors, and data processing methods. Physically, these factors mainly modulate signal representation and acquisition quality (e.g., amplitude scaling, noise level, and spectral appearance), rather than directly causing structural damage.

Confounding Variable (C): The confounding factor C is a common cause of both the relation metric T and the fault class Y. It may affect the relationship between the instrumental variable I and T, as well as between the adjustment variable A and Y. Examples include speed drift, temperature drift, and load fluctuations. For instance, increased load may simultaneously intensify gear vibration ($C \to T$) and accelerate fault occurrence ($C \to Y$). Its primary role is environmental coupling to both observation and degradation progression, rather than direct generation of intrinsic fault signatures.

Adjustment Variable (A): The adjustment variable A solely affects the fault class Y. For instance, factors such as vibration spectral kurtosis, entropy, frequency band energy, and characteristic fault frequencies determine the fault class Y. These descriptors are linked to damage mechanisms and are expected to be relatively stable across different sensing settings and moderate variations in operating conditions.

In practical industrial settings, these assumptions are reasonable because shared physical laws govern fault mechanisms across components. At the same time, component-specific variations and operating conditions can be treated as separable factors captured by the causal decomposition. Importantly, CINet does not assume perfect decorrelation in real data; instead, it enforces soft independence through $\mathcal{L}_I$, $\mathcal{L}_A$, $\mathcal{L}_{C\_B}$, and $\mathcal{L}_O$. If the assumptions are partially violated, these regularizers can still reduce entanglement and improve transferability, though they cannot remove all bias. To better capture these causal relationships, we employ a meta-task learning model. By simultaneously predicting both the optimal relation metric and fault class, the model gains an understanding of the causal structure and extracts actionable insights from the input data, thereby improving prediction accuracy.

This method extracts the features of the three components (I, C, and A) from the input signal X in the cross-component few-shot fault diagnosis SCM decomposition (as shown in Fig. 1). In the proposed SCM, the confounding variable C forms a backdoor path, which causes bias in the estimation of the causal effect. To address this issue, our causal decoupling module explicitly separates C from X, thereby blocking these backdoor paths. At the same time, the instrumental variable I and the adjustment variable A are retained to capture the effective causal path.

$$P(Y|do(T=t)) = \sum_c P(Y|T=t, C=c)\, P(C=c). \tag{2}$$

Decomposing A from X: As shown in the structural diagram, the adjustment variable A only affects the fault class Y and is independent of the intervention variable T. According to the principle of causal independence, we can decompose the adjustment variable A and prevent it from being entangled with other variables by enhancing the predictive capability of the classification model $G_A(A(X))$ for Y and enforcing the independence $A \perp T$.

Decomposing I from X: The instrumental variable I can only influence the outcome Y through the intervention variable T. Therefore, after balancing the confounding variable to ensure their independence from T, I can be made independent of Y given T. Accordingly, we decompose I by minimizing the mutual information between I and Y while fixing T, using $G_I(I(X))$ to predict T, and calculating the relationship score with T as the weight to predict the final outcome Y.

Balancing the Confounding Factor C: By introducing an attention mechanism, the distribution of C for different values

of T is aligned, i.e., $\mathcal{A}[C(X)] \perp T$. Under this independence condition, the backdoor adjustment formula can be simplified as:

$$\begin{aligned} P(Y|do(T=t)) &= \sum_c P(Y|T=t, C=c)\, P(C=c) \\ &= \sum_c P(Y|T=t, C=c)\, P(C=c|T=t) \\ &= P(Y|T=t). \end{aligned} \tag{3}$$

By disentangling cause and effect, particularly by separating instrumental variables, such as running speed, installation errors, and data processing methods, from the latent fault variables, the model can avoid being influenced by environmental or operational conditions. This enables the model to focus more on the fault itself rather than on irrelevant environmental changes. By integrating the model's ability to separate these latent variables, its robustness improves, leading to more reliable fault detection and diagnosis across varying conditions. As a result, the model can maintain consistent performance across different devices, environments, and operating conditions, unaffected by domain variations.

C. CINet Model

Building upon the proposed causal structural model, we designed a causal intervention network for cross-component few-shot fault diagnosis. Our network comprises two modules: the causal feature disentanglement module and the relation metric module. The network architecture is illustrated in Fig. 2. The meta-training task set provides sample sets and query sets across various categories. The causal disentanglement module separates the input data into instrumental variable I(X), confounding variable C(X), and adjustment variable A(X), each of which is optimized through a specific loss function. I(X) is fed into the metric function prediction network to determine the optimal relation metric T, which is optimized using the loss function L_I. C(X) and A(X) are input into the relation metric module, where the disentanglement of A(X) is optimized by L_A, and both fault type prediction and metric function selection are refined with L_R. The relation metric module outputs a relation score between the feature of the query sample and the mean feature representation of each fault category. The model parameters and structure are shown in Table II.

The three representation networks are denoted as $I(\cdot)$, $C(\cdot)$, and $A(\cdot)$, which extract the representations of the instrumental variable I, the confounding variable C, and the adjustment variable A, respectively. These representations are expressed as: $\hat{I} = I(x), \quad \hat{C} = C(x), \quad \hat{A} = A(x)$ We use $\hat{I}$ to predict T and use $\hat{A}$ to predict Y. The prediction of T employs a classification network, while the prediction of Y is achieved by computing relation scores, which match the most similar fault class. Fig. 2 presents the structure of the CINet model.

1) Causal Disentanglement Module: The causal structural model shows that the adjustment variable should be independent of the treatment variable, i.e., $A(x) \perp T$. Hence, we disentangle the representation of $A(X)$. Since the relation metric function selection is binary, we decompose the part of the representation unrelated to T by minimizing the distributional discrepancy of $A(X)$ when $\hat{t} = G_I[I(x)] \arg\max$ differs. Here, T is predicted by the regression network G_I using the features of $I(X)$.Simultaneously, to ensure that the representation of the adjustment variable is fully disentangled without residual information in other parts, we use only $A(X)$ to predict the fault class Y. The objective function for the adjustment variable representation network is as follows:

$$\begin{aligned} \mathcal{L}_A = \operatorname{disc}\left(\{A(x_j)\}_{j:\hat{t_j}=0}, \{A(x_j)\}_{j:\hat{t_j}=1}\right) \\ + \sum_j l\left[y_j, G_A(A(x_j))\right]. \end{aligned} \tag{4}$$

where $\{A(x_j)\}_{j:\hat{t_j}=0}$ represents the distribution of the adjustment variable representation $A(x)$ when $\hat{t} = 0$, disc$(\cdot)$ is a function that computes the distributional distance of $A(x)$ under different values of $\hat{t}$, and $\hat{t} = G_I[I(x)] \arg\max$. The term $\sum_j l\left[y_j, G_A(A(x_j))\right]$ measures the distance between the predicted fault class $G_A(A(x_j))$ and the true label y_j, where cross-entropy is used in this paper.Here, G_A denotes the relational metric module (see the section on the relation metric module for details).

For the instrumental variable $I(X)$, after ensuring that the attention-weighted confounding variable $\mathcal{A}[C(X)]$ is independent of T through the learning of an attention mechanism (i.e., under the assumption $\mathcal{A}[C(X)] \perp T$), we have $I(X) \perp Y|T$. Therefore, we disentangle I from X by minimizing the mutual information between $I(X)$ and Y.

As illustrated in the causal diagram in Fig. 1, the effect of variable T on Y is confounded by variable C. Based on the backdoor adjustment principle, by employing the attention mechanism to balance the influence of C and ensuring the independence condition $C \perp T$, the interventional distribution can be estimated from the observational conditional distribution. This enables the model to infer the true causal effect of T on Y without the bias introduced by C, allowing the downstream prediction to focus on the causal contribution of T rather than spurious correlations. The objective function for balancing the confounding variable is:

$$\mathcal{L}_{C\_B} = \operatorname{disc}\left(\{\mathcal{A}[C(x_j)]\}_{j:\hat{t_j}=0}, \{\mathcal{A}[C(x_j)]\}_{j:\hat{t}_j=1}\right), \tag{5}$$

where $\{\mathcal{A}[C(x_j)]\}_{j:\hat{t_j}=0}$ represents the distribution of the confounding variable $C(x)$ after balancing via the attention mechanism $\mathcal{A}$, and $\hat{t}_j = 0$ indicates that $\hat{t} = G_I[I(X)] \arg\max$ for this sample is 0. The function disc$(\cdot)$measures the distributional distance of the attention-weighted and balanced confounding variable $\mathcal{A}[C(x)]$ under different values of $\hat{t}$.

The objective function for the instrumental variable representation network is:

$$\mathcal{L}_I = \sum_{l=\{0,1\}} \mathrm{MI}\left(I(x_j), y_j\right)_{j:\hat{t_j}=l}, \tag{6}$$

where $\mathrm{MI}(I(x), y)$ denotes the mutual information between the instrumental variable $I(x)$ and the fault class y.

To ensure that the three representation extraction networks $\{I(\cdot), C(\cdot), A(\cdot)\}$ effectively disentangle the variables I, C, and A, we introduce an orthogonality constraint to reduce

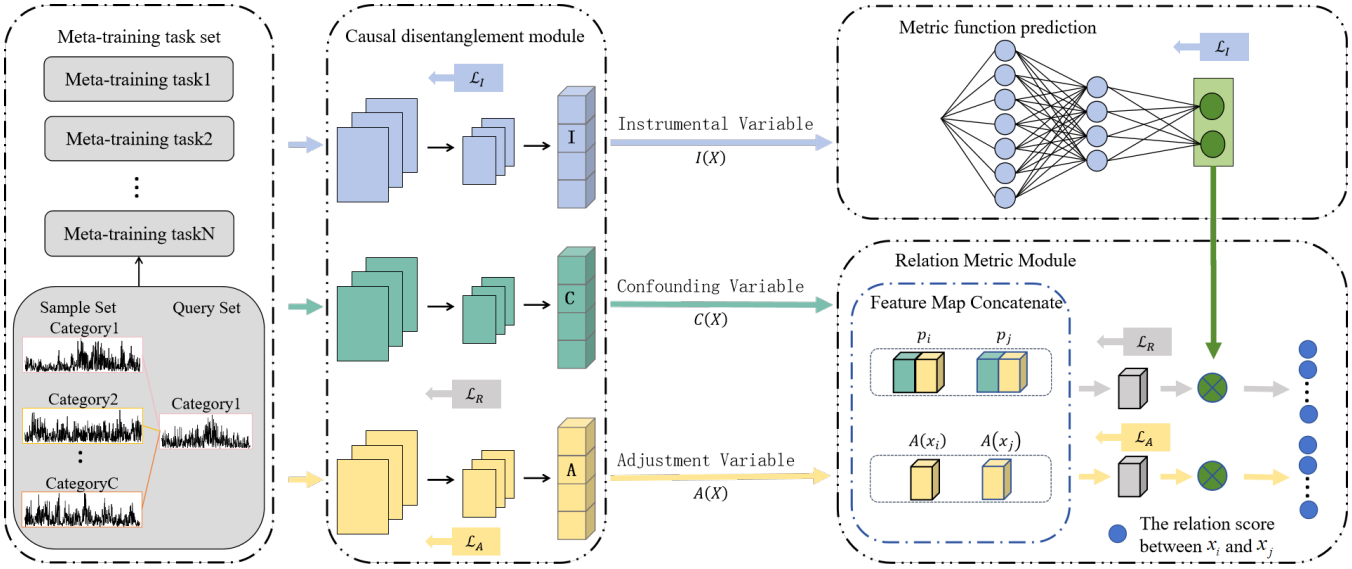


Fig. 2: The network architecture of CINet.

TABLE II: Model parameters and structure

Module	Layer	Operator	Output Shape
Representation Networks	Layer 1	Conv1d, BatchNorm1d, ReLU, MaxPool1d	(64, 20)
	Layer 2	Conv1d, BatchNorm1d, ReLU, MaxPool1d	(64, 9)
	Layer 3	Conv1d, BatchNorm1d, ReLU	(64, 9)
	Layer 4	Conv1d, BatchNorm1d, ReLU	(1024, 9)
Relation Network	FC1	Linear	(128)
	FC2	Linear	(64)
	FC3	Linear	(2)
SE	AvgPool	AdaptiveAvgPool2d	(1, 1)
	FC1	Linear	(1)
	FC2	Linear	(25)

information overlap among them. The objective function for enforcing the orthogonality constraint is

$$\mathcal{L}_O = \overline{W}_I^T \overline{W}_C + \overline{W}_C^T \overline{W}_A + \overline{W}_A^T \overline{W}_I, \tag{7}$$

where the contribution of each variable in X to $I(X)$ is denoted as W_I, and $\overline{W}_I$ represents the row-wise average of W_I. Since the three representation networks share the same structure, their average weight vectors have the same dimensionality. Thus, we impose pairwise orthogonality constraints to achieve hard disentanglement.

2) Relation Metric Module: Based on the proposed causal structural model, after performing causal disentanglement on the input samples to obtain $I(X)$, $C(X)$, and $A(X)$, we use the instrumental variable $I(X)$ to select the corresponding relation metric function T_l. The relation metric function calculates the relation score between the test sample features $\mathcal{A}[C(X), A(X)]$ and the average features of samples from each fault class in the query set, and weighs the relation scores using t:

$$\bar{C}(x_i) = \frac{1}{k} \sum_{i=1}^{k} C(x_i^c), \tag{8}$$

$$\bar{A}(x_i) = \frac{1}{k} \sum_{i=1}^{k} A(x_i^c). \tag{9}$$

To integrate complementary information from different feature extractors, we perform concatenation over the representations:

$$p_i = \bar{C}(X_i) \oplus \bar{A}(X_i), \tag{10}$$

$$p_j = C(x_j) \oplus A(x_j), \tag{11}$$

where $\oplus$ denotes vector concatenation.

$$r_{i,j,l} = T_l \left[\mathcal{A}(p_i); \mathcal{A}(p_j) \right], \tag{12}$$

$$r_{i,j} = \sum_{l=0,1} t_{j,l} \; r_{i,j,l}. \tag{13}$$

The relation score $r_{i,j,l}$ is computed with the relation metric function T_l using the concatenated prototype p_i of class i from the support set and the concatenated feature p_j of a query sample $x_j \in X_c$. The relation scores $r_{i,j,l}$ are weighted and summed according to the weights $t_{j,l}$. To facilitate the optimization of the objective function via gradient descent, and $t_{j,l} = G_I[I(x_j)]\,\mathrm{softmax}$.

The test sample is ultimately classified into the fault class of the query set sample with the highest relation score. The objective function for predicting the fault class is:

$$\mathcal{L}_R = -\frac{1}{mc} \sum_{j=1}^{mc} \log \left(\frac{e^{r_{y_j,j}}}{\sum_{i=1}^{c} e^{r_{i,j}}} \right), \tag{14}$$

where y_i is the true label of the test sample, and $r_{i,j}$ is the weighted relation score between x_i and the average features of samples from class j in the query set.

3) Contrastive Loss: For sample feature vectors that are close to the average feature vectors of other classes, we introduce a contrastive loss to penalize cases where the distance between feature vectors of different classes is smaller than the margin δ:

$$\mathcal{L}_C = \frac{1}{mc^2} \sum_{i=1}^{c} \sum_{j=1}^{mc} \Big[\mathbb{1}(y_i = y_j)(c-1)\|p_i - p_j\|^2 + \big[1 - \mathbb{1}(y_i = y_j)\big] \big(\delta - \|p_i - p_j\|\big)^2 \Big], \tag{15}$$

where $\mathbb{1}(y_i = y_j)$ indicates whether samples x_i and x_j belong to the same class (1 if true, 0 otherwise). $\|p_i - p_j\|^2$ denotes the squared Euclidean distance between the two vectors, where δ is the minimum allowed distance between feature vectors of different classes.

4) Loss Function: Therefore, the CINet algorithm optimizes the following objective function:

$$\mathcal{L} = \mathcal{L}_R + \alpha\mathcal{L}_A + \beta\mathcal{L}_I + \mu\mathcal{L}_O + \gamma\mathcal{L}_{C\_B} + \epsilon\mathcal{L}_C, \tag{16}$$

where α, β, μ, γ, and ϵ are hyperparameters that control the contribution of each loss term. These losses are complementary: $\mathcal{L}_R$ drives fault classification and metric learning, $\mathcal{L}_A$ enforces the causal independence of adjustment features, $\mathcal{L}_I$ guides instrumental-variable disentanglement, and $\mathcal{L}_{C\_B}$ and $\mathcal{L}_C$ mitigate confounding effects through balancing and regularization, while $\mathcal{L}_O$ stabilizes the overall optimization. Together, they align discriminative performance with causal separation, improving generalization under cross-component shifts. For a detailed description of the training process, please refer to the pseudocode in the Appendix. By separating cause and effect, especially by separating "adjustment variables" such as temperature drift and load fluctuations from the fault signals, it can help the model avoid being affected by environmental or operational conditions. For instance, speed drift, temperature changes, and load fluctuations are some external factors that may alter the way the fault signals are observed. By separating these factors from the fault signals, the model can focus more on the fault itself rather than the irrelevant environmental changes. Therefore, the model can maintain consistent performance across different devices, environments, or working conditions without being disturbed by these domain variations.

IV. Experimental Results

A. Dataset Description

We experimentally validate the generalization performance of the CINet model using the following datasets: the bearing dataset from Case Western Reserve University (CWRU) [45], the gear dataset from the University of Connecticut (UConn) [46], and a lab-built bearing dataset from our laboratory test rig.

UConn Gear Dataset: The test rig consists of a motor, a two-stage reducer with replaceable gears, an electromagnetic brake, an acceleration sensor, and a tachometer. Vibration signals from the gears are measured by the acceleration sensor at a sampling frequency of 20 kHz. The gear unit includes nine fault classes: crack, missing, health, spall, and chip (with five severity levels from 1a to 5a under the chip class). For our experiments, we selected five classes: crack, missing, health, spall, and chip5a.

CWRU Bearing Dataset: The CWRU bearing test rig comprises a 2-horsepower motor, a torque sensor/encoder, a dynamometer, and control electronics. The sampling frequencies are 12 kHz and 48 kHz for the drive end and 12 kHz for the fan end. The bearing faults vary in diameter from 0.007 inches to 0.040 inches, with fault locations including the inner race, rolling element (ball), and outer race. The motor load ranges from 0 to 3 horsepower, and the motor speed ranges from 1797 to 1720 RPM. Our experiments utilize vibration signals from four fault diameters, six classes (including the normal condition), two load levels (0 and 3 hp), and two sampling locations, resulting in 80 data classes. The sampling locations include the fan end (FE) and drive end (DE). The fault classes are N (normal), IF (inner race fault), BF (ball fault), OF@3, OF@6, and OF@12, where "@n" indicates the fault is located at the n o'clock position relative to the load zone.

Lab-Built Bearing Dataset: The laboratory test rig is shown in Fig. 3. The bearing speed is controlled by an AC motor via a flexible coupling, and vibration signals are collected using an acceleration sensor at a sampling frequency of 51,200 Hz. The dataset includes five bearing classes: four fault classes and one normal condition - rolling: ball fault (BF), inner race fault (IF), outer race fault (OF), compound fault of rolling element and outer race (BOF), and normal signal (N).

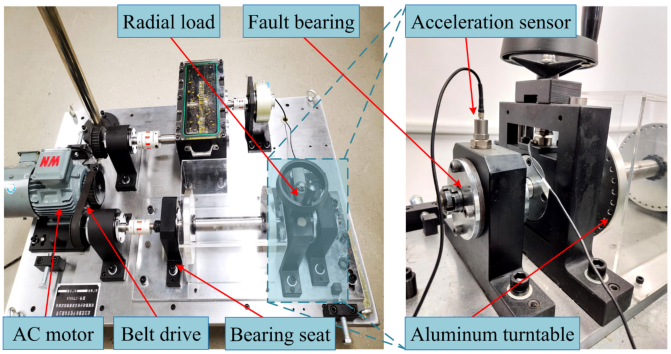


Fig. 3: Physical view of the laboratory-built test bench

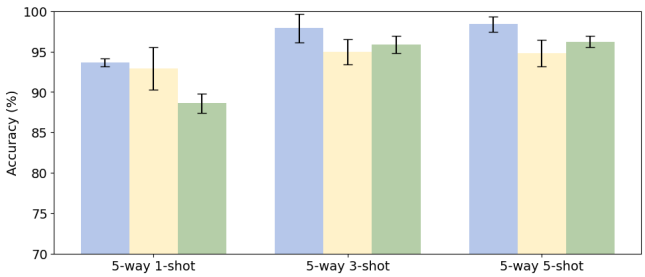
For our few-shot classification experiments, training samples are selected from the source component dataset in a meta-task fashion. In a 5-way 1-shot meta-task, the support set S contains 5 classes with 1 labeled sample per class. The query set Q, which the model needs to classify during training, also consists of 5 classes but randomly samples 20 labeled examples per class, resulting in 100 samples per batch.

To evaluate the model's generalization capability, test samples are selected from the target component dataset, also organized as meta-tasks. Thus, "cross-component" refers to training tasks sampled from the source dataset and testing tasks sampled from a different target dataset, rather than a single task mixing datasets. The source and target datasets

TABLE III: Description of fault diagnosis task

Diagnosis Tasks	Diagnosis Experiments Train→Test	Fault Class Train	Fault Class Test
Task A1 (5-way 1-shot) Task A2 (5-way 3-shot) Task A3 (5-way 5-shot)	CWRU bearing → UConn gearing	80 classes	Spall,Crack,Health, Missing,Chip5a
Task B1 (5-way 1-shot) Task B2 (5-way 3-shot) Task B3 (5-way 5-shot)	UConn gearing → CWRU bearing	Spall,Crack,Health, Missing,Chip5a	80 classes
Task C1 (5-way 1-shot) Task C2 (5-way 3-shot) Task C3 (5-way 5-shot)	UConn gearing → Lab-built bearing	Spall,Crack,Health, Missing,Chip5a	N,IF,OF, BF,BOF
Task D1 (5-way 1-shot) Task D2 (5-way 3-shot) Task D3 (5-way 5-shot)	CWRU bearing → Lab-built bearing	80 classes	N,IF,OF, BF,BOF

used to construct meta-training and meta-testing tasks for Task A/B/C/D are listed in Table III. For a 5-way 1-shot test meta-task, analogous to the training phase, we select 5 classes from the target dataset with 1 labeled sample per class (5 total samples) as the support set, and the remaining unlabeled samples from the test set. The model uses a learning rate of 0.0003, with loss function weights set as $\epsilon = 0.02$, $\alpha = 0.2$, $\beta = 1$, $\gamma = 0.03$, $\mu = 1$, and $\delta = 5$. Section V.B will detail the hyperparameter selection experiments.



Task	5-way 1-shot	5-way 3-shot	5-way 5-shot
taskA	93.64±0.51	97.90±1.78	98.40±0.96
taskB	92.89±2.65	94.94±1.57	94.83±1.63
taskC	88.62±1.20	95.88±1.08	96.23±0.72

Fig. 4: Average accuracy for each task

Fig. 4 shows the average classification accuracy over 10 runs for the nine tasks. Under the few-shot setting, all cross-component fault diagnosis tasks achieved strong accuracy, with Task A (5-way 5-shot) reaching the highest accuracy of 98.4% and Task C (5-way 1-shot) the lowest at 88.62%.

Fig. 5 shows the confusion matrices for task A 5-way 1-shot, 5-way 3-shot, and 5-way 5-shot, where the vertical axis of the matrix corresponds to the actual fault label, the horizontal axis represents the fault label predicted by the model, and the classification accuracy is shown in each cell. The accuracy of each classification in 5-way 1-shot is relatively close, in which the classification accuracy of fault missing is slightly lower (93%), the classification accuracy of healthy is 95%, and the classification accuracy of other missing, crack, and peeling are all 94%. The accuracy of each fault class in the tasks of 5-way 3-shot and 5-way 5-shot is 98%. For Task C, the 5-way 1-shot, 5-way 3-shot, and 5-way 5-shot results are relatively less balanced: IF, OF, and BallOut achieve higher classification accuracy, while Normal has lower accuracy in Task C1, and BF reaches the lowest accuracy (69%) in Task C1 but improves substantially in Task C2 and Task C3.

Fig. 6 illustrates the variation of accuracy and loss for task A over 400 iterations. The accuracy on tasks A1, A2, and A3 improves significantly after about 50 iterations, and, at the 272nd, 107th, and 379th iterations, the model achieves 97.5%, 99.0%, and 99.3% highest test accuracy on tasks A1, A2, and A3, respectively. Then, as the number of iterations increases, the loss function curve becomes stable, leveling off after about 150 iterations.

B. Hyperparameter Analysis

For hyperparameter tuning, we adopt a coordinate-wise (one-factor-at-a-time) procedure on Task A1 across predefined ranges, covering the learning rate, the margin for negative sample pairs, and the weights associated with different loss terms. Starting from a default setting, at each step we update one hyperparameter based on the current best-performing configuration while keeping all others fixed. The best-performing value is then fixed to update the current best-performing configuration before tuning the next hyperparameter. Each candidate's value is trained on the meta-training task set and evaluated on the test set; five independent runs are conducted, and the average classification accuracy is reported. The best-performing value is then fixed before tuning the next hyperparameter. This approach is a computationally efficient tuning strategy and does not guarantee a globally optimal joint combination. Fig. 7 illustrates the box plots of our method under different hyperparameter values.

Learning Rate (lr): As shown in Fig. 7(a), we evaluate the influence of model learning rate on the overall optimization of the model by comparing the optimization effect between different values of learning rate. In the experiment, we set lr to several possible values (lr=0.0001, 0.0003, 0.0005, 0.0007, 0.001), where lr=0.0001 has the lowest average accuracy of

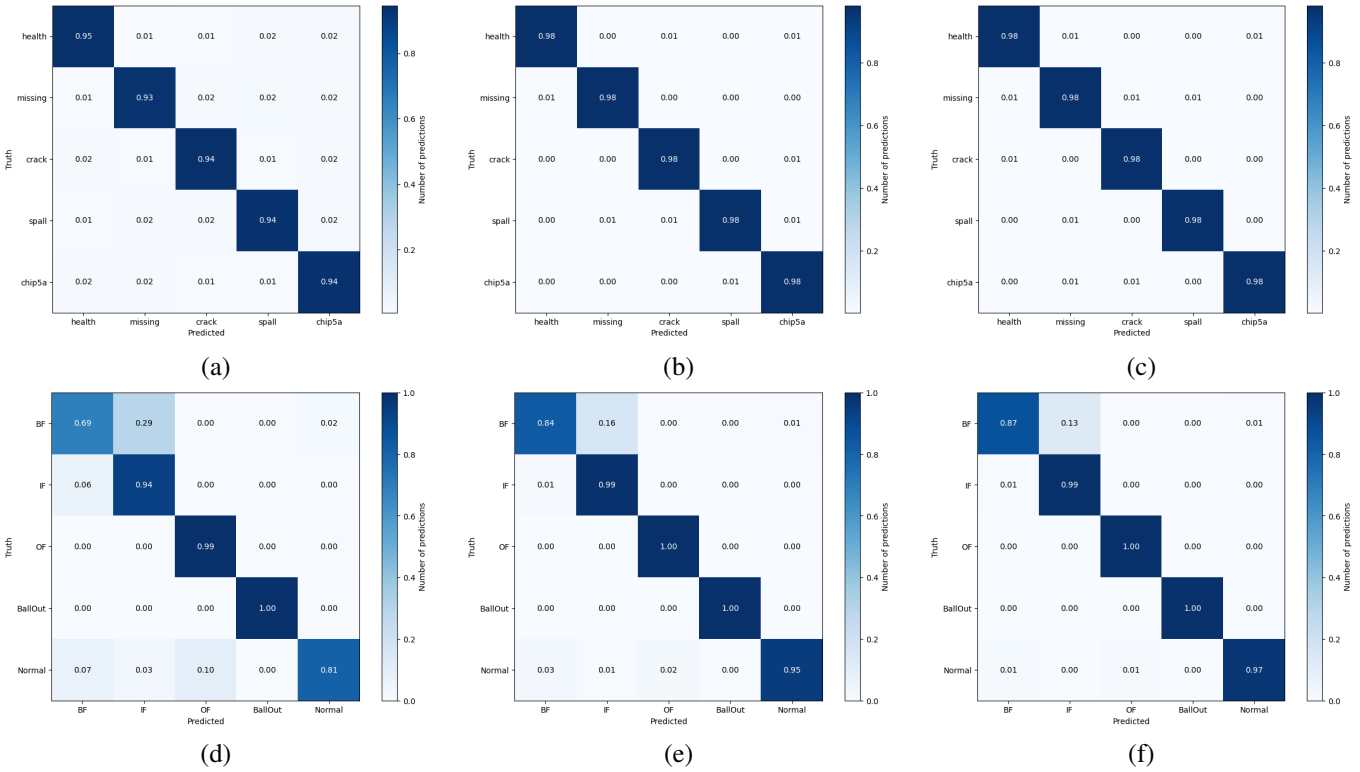


Fig. 5: Confusion matrix of Task A and Task C

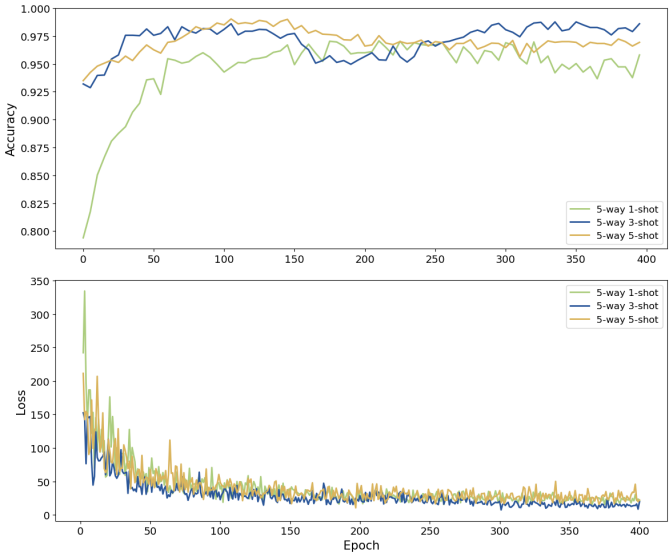


Fig. 6: Accuracy curve and loss curve of Task A

91.63%, and lr=0.0005 has the highest average accuracy of 93.79%.

Weight Parameter for $\mathcal{L}_A$ (α): As shown in Fig. 7(b), we study the weight parameter α of $\mathcal{L}_A$ in Formula (10), and evaluate the influence of this partial loss function on the overall optimization of the model through the comparison of optimization effects between different values. In the experiment, we set α to several possible values (α=0.02, 0.03, 0.05, 0.1, 0.2) for grid search, and we can see that the average classification accuracy of the model is the highest when α is 0.2, which indicates that causal disentanglement can improve the classification effect of the model.

Weight Parameter for $\mathcal{L}_I$ (β): As shown in Fig. 7(c), for the weight parameter β of $\mathcal{L}_I$, we also compare the average classification accuracy of the model under different values. Similarly, β was set to several possible values (β=0.3, 0.5, 1, 2, 3) for experiments. When β was set to 1, the average classification accuracy of the model was the highest, reaching 93.57%.

Weight Parameter for $\mathcal{L}_O$ (μ): As shown in Fig. 7(d), we evaluate the impact of the weight parameter μ of the orthogonal constraint loss term $\mathcal{L}_O$ on the model performance. A grid search is performed on μ (μ=0.2, 0.3, 0.5, 0.7, 1) while other hyperparameters are fixed. The average classification accuracy of the model is highest when μ is set to 1 and relatively low for other values, which indicates that moderate orthogonality constraints have a clear positive effect on causal decomposition.

Weight Parameter for $\mathcal{L}_{C_B}$ (γ): As shown in Fig. 7(e), we evaluate the impact of the weight parameter γ of the orthogonal constraint loss term γ on the model performance. Grid search was performed on γ (γ) with other hyperparameters fixed. The average classification accuracy of the model is the highest when γ is set to 0.03, and the accuracy is relatively low when γ is set to other values, which indicates that the introduction of $\mathcal{L}_{C_B}$ can effectively eliminate the confounding variable.

Weight Parameter for $\mathcal{L}_C$ (ϵ): As shown in Fig. 7(f), we evaluate the impact of the weight parameter ϵ of the orthogonal

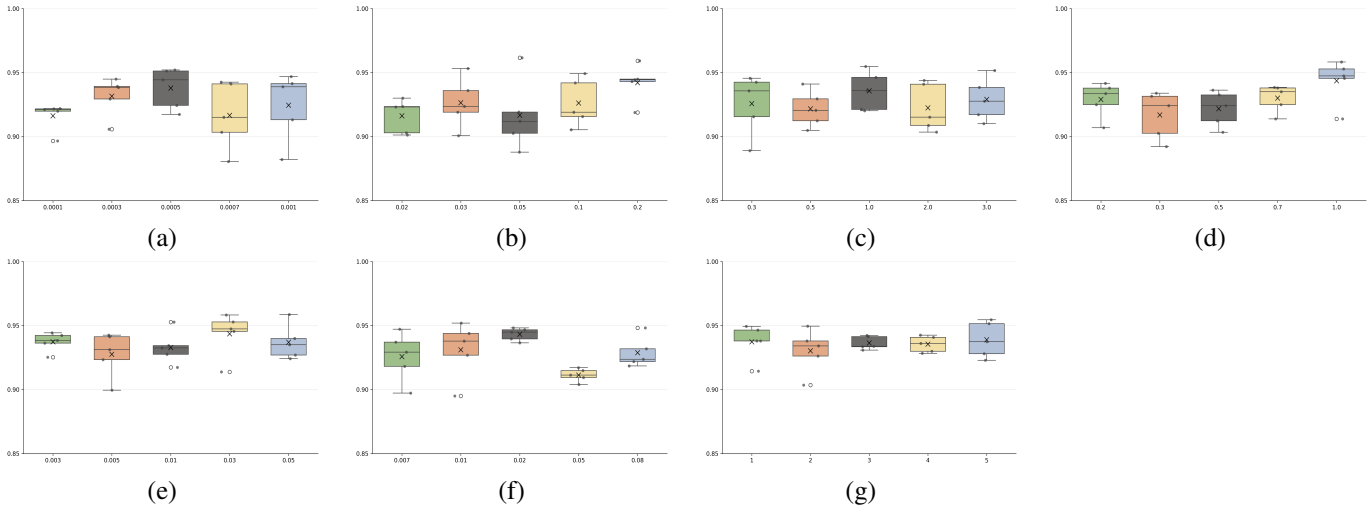


Fig. 7: Hyperparameter contrast line chart

constraint loss term $\mathcal{L}_C$ on the model performance. Grid search for ϵ (ϵ=0.007, 0.01, 0.02, 0.05, 0.08) with the other hyperparameters fixed. The results show that when ϵ=0.02, the average classification accuracy of the model is the highest, which is better than other values. Further analysis shows that the moderate introduction of contrastive loss can effectively distinguish difficult sample pairs.

Margin for Negative Sample Pairs (δ): As shown in Fig. 7(g), we investigate the minimum relative distance margin of negative sample pairs in Eq. (15) to assess its impact on the robustness of the model. We set the minimum relative distance between negative pairs to different values (δ=1, 2, 3, 4, 5). When the δ is 5, the model has the highest average classification accuracy in task A1.

C. Performance Analysis of Causal Disentanglement

To analyze the impact of causal disentanglement on model performance, we conduct ablation experiments and use t-SNE and PCA to visualize the feature distributions before and after causal disentanglement. As shown in Fig. 8, the feature space becomes more structured after introducing causal disentanglement, with smaller intra-cluster distances indicating more compact clusters. Specifically, Fig. 8(b) shows the feature distribution of Task A1 after causal disentanglement, where the three classes (missing, chip5a, and health) are more clearly separated than in the baseline. Fig. 8(d) shows larger inter-class distances and more compact intra-class clusters than Fig. 8(c), which corresponds to the model without causal disentanglement. In the PCA view, Fig. 8(e) shows the feature distribution of Task A1 before causal disentanglement, where the spall class is relatively dispersed, and the separation among the health, missing, and chip5a classes is limited; these issues are substantially alleviated after causal disentanglement, as shown in Fig. 8(f). Similarly, Fig. 8(g) presents the PCA visualization of Task C1 before causal disentanglement, where different classes show small margins and considerable overlap, while Fig. 8(h) demonstrates a clear improvement after causal disentanglement. These results indicate that the proposed method enhances both the discriminability and interpretability of the extracted features.

To further examine the contributions of different latent variables, we conduct module-level ablation studies on the causal components in Tasks A1, B1, and C1 by setting the corresponding loss weights to zero. Each ablated variant is trained and evaluated over ten runs, and the mean accuracies are summarized in Table VI. The results show that removing causal components (NoCausal) yields the lowest accuracy, while removing A, C, or I also degrades performance, with a larger drop observed when excluding A, C than when excluding I. The significant accuracy decline demonstrates that disentangling these three types of variables is critical to the model's high performance, and further supports that such variable-level disentanglement is both feasible and practically meaningful for engineering applications.

D. Comparison with State-of-the-Art Methods

We compare the proposed CINet model with six advanced methods, including four FSL methods (Relation Net, Prototypical Net, TRNet, and CFDM) and one causal intervention-based method (CIRNet). For each task, all reported metrics are averaged over 10 runs and presented as mean $\pm$ standard deviation. The average diagnostic accuracy is summarized in Table IV. The corresponding Macro-F1 results are reported in Table V and exhibit the same overall performance trend. Macro-F1 is the macro-averaged harmonic mean of precision and recall across classes, and the results indicate that the models maintain relatively balanced class-level performance. In addition, we report MCC (Matthews correlation coefficient) in a manner consistent with Macro-F1 in Table V.

Among the compared methods, the three meta-learning-based methods—Prototypical Net, Relation Net, and TRNet—achieved accuracies of 79.09%, 84.03%, and 84.21% in Task A1, respectively, with corresponding Macro-F1 scores of 88.01%, 80.99%, and 88.03%. These results confirm the generalization advantage of meta-learning models in few-shot settings, as they effectively learn cross-domain invariant

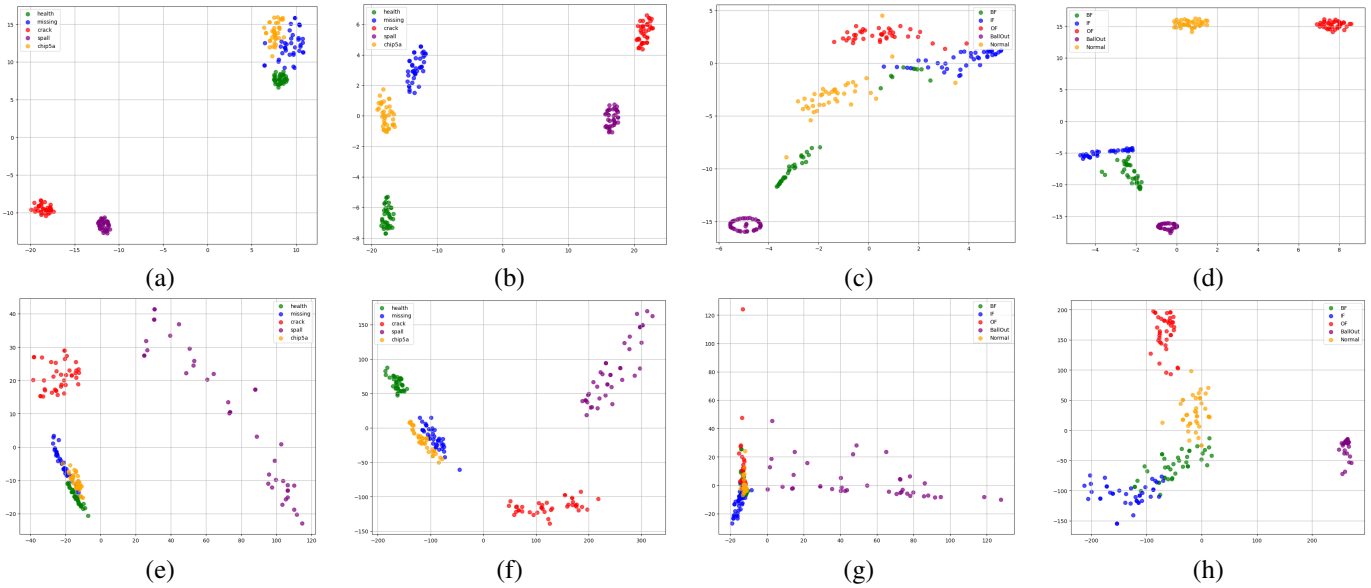


Fig. 8: Feature distribution visualization of representative tasks. (a)–(d) show the t-SNE projections before and after introducing causal disentanglement for Tasks A1 and C1, respectively. (e)–(h) show the visualization of projects using PCA for Tasks A1 and C1 before and after introducing causal disentanglement. In both t-SNE and PCA views, CINet produces more compact intra-class clusters and larger inter-class separation, indicating that causal disentanglement effectively suppresses confounding factors and improves feature discriminability for cross-component few-shot diagnosis.

TABLE IV: Average accuracy (%, mean ± std)

Method	Task A1	Task B1	Task C1	Task D1
Relation Net [11]	79.09 ± 0.95	82.11 ± 1.94	78.38 ± 1.28	83.36 ± 1.77
Prototypical Net [13]	84.03 ± 0.92	78.31 ± 1.52	79.60 ± 1.08	85.96 ± 2.29
TRNet [44]	84.21 ± 2.01	84.08 ± 1.16	79.29 ± 1.67	86.93 ± 1.61
CFDM [12]	84.99 ± 2.80	69.11 ± 1.50	69.95 ± 2.94	85.86 ± 1.01
CIRNet [22]	91.52 ± 1.02	87.25 ± 1.25	85.82 ± 1.90	93.13 ± 1.47
CINet	$\mathbf{93.64 \pm 0.51}$	$\mathbf{92.89 \pm 2.65}$	$\mathbf{88.62 \pm 3.56}$	$\mathbf{93.78 \pm 1.20}$

TABLE V: Average Macro-F1 and MCC (%, mean ± std)

Metric	Method	Task A1	Task B1	Task C1	Task D1
Macro-F1	Relation Net [11]	80.99 ± 1.26	80.24 ± 2.55	82.41 ± 0.91	85.20 ± 1.80
	Prototypical Net [13]	88.01 ± 0.47	77.79 ± 0.42	88.10 ± 0.68	89.05 ± 0.69
	TRNet [44]	88.03 ± 0.86	80.22 ± 1.73	86.28 ± 1.85	87.96 ± 0.79
	CFDM [12]	79.35 ± 3.12	72.17 ± 0.93	70.33 ± 1.89	88.42 ± 0.89
	CIRNet [22]	90.32 ± 0.84	81.99 ± 2.56	85.66 ± 2.41	93.96 ± 1.03
	CINet	$\mathbf{93.45 \pm 0.54}$	$\mathbf{92.51 \pm 3.97}$	$\mathbf{86.36 \pm 0.26}$	$\mathbf{93.40 \pm 2.11}$
MCC	Relation Net [11]	79.02 ± 2.10	77.23 ± 3.07	75.18 ± 2.60	82.98 ± 1.80
	Prototypical Net [13]	85.23 ± 0.75	77.10 ± 4.07	84.22 ± 1.90	85.08 ± 1.55
	TRNet [44]	86.28 ± 1.70	81.67 ± 1.94	85.29 ± 1.11	86.44 ± 2.61
	CFDM [12]	79.93 ± 2.58	72.11 ± 1.25	69.56 ± 1.06	82.82 ± 3.63
	CIRNet [22]	86.09 ± 3.81	79.67 ± 2.15	85.77 ± 4.24	88.62 ± 1.66
	CINet	$\mathbf{91.33 \pm 3.21}$	$\mathbf{89.03 \pm 3.99}$	$\mathbf{86.56 \pm 1.12}$	$\mathbf{91.79 \pm 2.51}$

TABLE VI: Ablation Results on Causal Variables

Setting	Task A1	Task B1	Task C1
CINet (full)	93.64 ± 0.51	92.89 ± 2.65	88.62 ± 3.56
NoCausal	77.44 ± 0.78	75.62 ± 0.80	72.90 ± 1.76
$without A, C$	83.14 ± 0.33	84.92 ± 0.34	80.66 ± 2.21
$without I$	90.86 ± 0.87	89.65 ± 0.90	86.40 ± 3.09

representations such as class-level metric spaces. However, since these methods do not explicitly consider the underlying causal structures of the data, their performance remains limited under complex distribution shifts.

For the metric-based method CFDM, the classification accuracies on Tasks A1, B1, C1, and D1 were 84.99%, 69.11%, 69.95%, and 85.86%, respectively. The corresponding Macro-F1 scores were 79.35%, 72.17%, 70.33%, and 88.42%. Overall, it achieved relatively high performance, which can be attributed to its ability to construct a metric space through an

encoding function, enabling the model to measure similarities between sample pairs from different domains rather than relying solely on within-domain classification. However, the large performance gap across tasks indicates that CFDM is significantly affected by the transferability between different components and remains vulnerable to confounding factors. CIRNet achieved accuracies of 91.52%, 86.65%, 85.82%, and 93.13% on tasks A1, B1, C1, and D1, respectively, with Macro-F1 scores of 90.32%, 81.99%, 85.66%, and 93.96%. Due to the introduction of a causal structural model, CIRNet shows a significant improvement in accuracy compared to previous models. However, its SCM construction is not comprehensive, as it only considers pretraining knowledge as a confounding factor, which still poses certain limitations.

As shown in Table IV, our method achieved average accuracies of 93.64%, 92.89%, 88.62%, and 93.78% in tasks A1, B1, C1, and D1, respectively, which demonstrated a significant improvement w.r.t. state-of-the-art methods. This is confirmed by the Macro-F1 scores: 93.45%, 92.51%, 86.36%, and 93.40%. The MCC values in Table V were 91.33%, 89.03%, 86.56%, and 91.79%, which are also the best among all compared methods and consistent with the trends of accuracy and Macro-F1. Task C1 yields the lowest Macro-F1 and MCC, suggesting that its class recognition is relatively less balanced and more challenging than the other tasks; this is also evident in the confusion matrix in Fig. 5. Overall, the Macro-F1 and MCC results suggest that none of the models exhibit severe class-imbalance failures, while CINet remains consistently superior to the competing methods. The standard deviations remain comparable to those of competing methods, underscoring the stability of our approach.

For the cross-component tasks (A1, B1, and C1) compared to the non-cross-component task (D1), the average accuracies of the comparison methods on the cross-component tasks are as follows: Relation Net at 79.86% (3.50% lower than Task D1), Prototypical Net at 80.65% (5.31% lower than Task D1), TRNet at 82.53% (4.40% lower than Task D1), CFDM at 74.68% (11.18% lower than Task D1), and CIRNet at 88.20% (4.93% lower than Task D1). In contrast, CINet achieved 91.72% average accuracy on the cross-component tasks, which is 2.06% lower than its performance on the non-cross-component task (93.78%). These results highlight a strong cross-component generalization capability of CINet, with performance minimally affected by component changes. Unlike conventional metric-based methods that heavily rely on sharing data distributions between training and testing components, CINet utilizes causal decoupling to extract invariant mechanisms related to faults. This makes knowledge transfer between different components more reliable, thereby achieving higher diagnostic accuracy in cross-component scenarios.

E. Interpretability Analysis

To validate the interpretability of CINet, we employ Grad-CAM (Gradient-weighted Class Activation Mapping) to visualize which regions of the input contribute most to model predictions. By computing gradients of the target-class score with respect to convolutional feature maps and using these gradients as importance weights, Grad-CAM generates a heatmap that highlights the frequency bands emphasized during decision making. Specifically, Grad-CAM computes the gradient of the target class score y^c with respect to the k-th feature map A^k, i.e., $\frac{\partial y^c}{\partial A^k}$. Channel-wise weights are computed using global average pooling:

$$\alpha_k^c = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \frac{\partial y^c}{\partial A_{ij}^k}. \tag{17}$$

The class activation maps are achieved using a ReLU activation to retain positively contributing regions:

$$\mathrm{ReLU}\left(\sum_k \alpha_k^c A^k \right). \tag{18}$$

After normalization and interpolation, the Grad-CAM heatmap is aligned with the input signal to support subsequent visual analysis.

Fig. 9 presents frequency-domain attention visualizations for three representative samples under the baseline model and CINet. In the baseline setting, the attention distribution is relatively diffuse, and part of the attention is allocated to low-energy or noise-dominated bands. In contrast, after introducing the causal intervention mechanism, the attention becomes more concentrated on bands aligned with dominant spectral peaks and fault-related harmonics, which have clear physical meaning. These results indicate that CINet captures more fault-relevant representations while suppressing spurious responses.

V. Conclusion

To address the challenge of few-shot cross-component fault diagnosis, this paper proposes CINet, a model built from a causal perspective. Within the meta-learning framework, CINet constructs a causal structural model to guide causal decomposition and uncover the intrinsic causal relationships between vibration signals and fault classes. By leveraging multi-task collaborative optimization, CINet improves both relation metric function prediction and fault class prediction, enabling rapid adaptation to new target-component tasks and enhancing overall diagnostic efficiency.

Experimental results across three datasets, including two public datasets and one laboratory-collected dataset, demonstrate that CINet consistently outperforms state-of-the-art methods in classification accuracy and robustness, even in the few-shot setting. CINet achieved an average accuracy of 92.52% on cross-component tasks, which is only 1.28% lower than its performance on the non-cross-component task. This demonstrates CINet's substantially improved cross-component generalization. To assess the impact of causal disentanglement on model performance, we conduct ablation experiments and use t-SNE and PCA to visualize feature distributions before and after causal disentanglement. The results show that the proposed method effectively enhances both the discriminability and interpretability of the extracted features.

Although the proposed CINet framework demonstrates strong cross-component generalization and provides an effective solution for the targeted few-shot fault diagnosis setting,

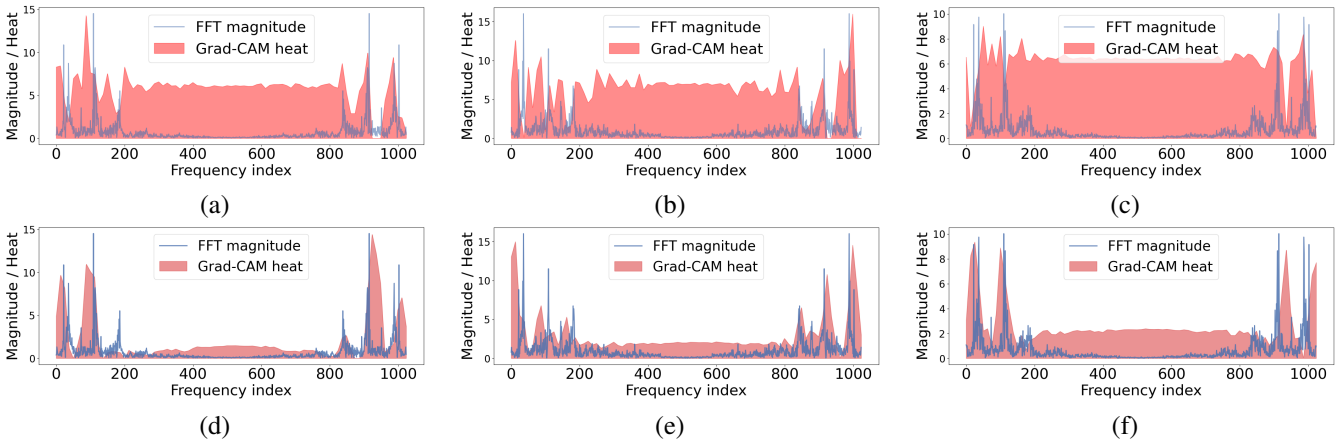


Fig. 9: Frequency-domain attention comparison between the baseline model and CINet for three representative samples. (a)–(c) Task A1/B1/C1 results of the baseline model. (d)–(f) Task A1/B1/C1 results of CINet, visualizing the extracted adjustment features $A(X)$.

its broader applicability to diverse industrial scenarios still requires further exploration. In real deployments, variations in sensor layout, operating regimes, and background disturbances may introduce partially entangled factors, so a small portion of confounding information may not be fully disentangled. Nevertheless, current results indicate that CINet's causal modeling enables relatively stable extraction of causal factors even when operating scenarios change, and the model remains robust and effective under representative cross-component and cross-operating-condition settings.

Future work will further investigate the scalability and transferability of CINet across heterogeneous industrial environments, while improving robustness through domain-specific calibration and developing efficiency-aware hyperparameter optimization strategies (e.g., Bayesian optimization and reinforcement learning) to reduce manual intervention.

ACKNOWLEDGEMENT

This work was supported in part by the National Key R&D Program of China (NO.2024YFB3311600) and the National Natural Science Foundation of China (52375089)

REFERENCES

[1] E. Zio, "Prognostics and Health Management (PHM): Where are we and where do we (need to) go in theory and practice," *Reliab. Eng. Syst. Saf.*, vol. 218, Art. no. 108119, Feb. 2022, doi: 10.1016/j.ress.2021.108119.

[2] K. Feng, J. C. Ji, Q. Ni, and M. Beer, "A review of vibration-based gear wear monitoring and prediction techniques," *Mech. Syst. Signal Process.*, vol. 182, Art. no. 109605, 2023, doi: 10.1016/j.ymssp.2022.109605.

[3] E. Iunusova, M. K. Gonzalez, K. Szipka, and A. Archenti, "Early fault diagnosis in rolling element bearings: comparative analysis of a knowledge-based and a data-driven approach," *J. Intell. Manuf.*, vol. 35, no. 5, pp. 2327–2347, Jun. 2024, doi: 10.1007/s10845-023-02151-y.

[4] V. Jorry, Z.-S. Duma, T. Sihvonen, S.-P. Reinikainen, and L. Roininen, "Statistical batch-based bearing fault detection," *J. Math. Ind.*, vol. 15, no. 1, p. 4, Feb. 2025, doi: 10.1186/s13362-025-00169-w.

[5] M. A. Khan, B. Asad, K. Kudelina, T. Vaimann, and A. Kallaste, "The Bearing Faults Detection Methods for Electrical Machines—The State of the Art," *Energies*, vol. 16, no. 1, p. 296, Jan. 2023, doi: 10.3390/en16010296.

[6] T. Kourti and J. F. MacGregor, "Multivariate SPC methods for process and product monitoring," *J. Qual. Technol.*, vol. 28, no. 4, pp. 409–428, Oct. 1996, doi: 10.1080/00224065.1996.11979699.

[7] B. Denkena, B. Bergmann, and A. Schmidt, "Preload monitoring of single nut ball screws based on sensor fusion," *CIRP J. Manuf. Sci. Technol.*, vol. 33, pp. 63–70, 2021, doi: 10.1016/j.cirpj.2021.02.006.

[8] V. Pandhare, X. Li, M. Miller, X. Jia, and J. Lee, "Intelligent diagnostics for ball screw fault through indirect sensing using deep domain adaptation," *IEEE Trans. Instrum. Meas.*, vol. 70, 2020, Art. no. 2504211, doi: 10.1109/TIM.2020.3043512.

[9] V. Pandhare, M. Miller, G. W. Vogl, and J. Lee, "Ball Screw Health Monitoring With Inertial Sensors," *IEEE Trans. Ind. Informat.*, vol. 19, no. 6, pp. 7323–7334, Jun. 2023, doi: 10.1109/TII.2022.3210999.

[10] Y. Zhang, J. C. Ji, Z. Ren, Q. Ni, F. Gu, K. Feng, K. Yu, J. Ge, Z. Lei, and Z. Liu, "Digital twin-driven partial domain adaptation network for intelligent fault diagnosis of rolling bearing," *Reliab. Eng. Syst. Saf.*, vol. 234, p. 109186, 2023, doi: 10.1016/j.ress.2023.109186.

[11] N. Lu, H. Hu, T. Yin, Y. Lei, and S. Wang, "Transfer relation network for fault diagnosis of rotating machinery with small data," *IEEE Trans. Cybern.*, vol. 52, no. 11, pp. 11927–11941, Nov. 2022, doi: 10.1109/TCYB.2021.3085476.

[12] S. Shao, R. Yan, Y. Lu, P. Wang, and R. X. Gao, "DCNN-based multi-signal induction motor fault diag-

nosis," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 6, pp. 2658–2669, Jun. 2020, doi: 10.1109/TIM.2019.2925247.

[13] J. Luo, H. Shao, J. Lin, and B. Liu, "Meta-learning with elastic prototypical network for fault transfer diagnosis of bearings under unstable speeds," *Reliab. Eng. Syst. Saf.*, vol. 245, p. 110001, May 2024, doi: 10.1016/j.ress.2024.110001.

[14] H. Pan, H. Xu, J. Cheng, J. Zheng, and J. Tong, "A Multiclass Graph Embedding Matrix Classification Method for Roller Bearing State Identification Under Limited Sample," *IEEE Trans. Reliab.*, vol. 74, no. 3, pp. 3824–3832, Sep. 2025, doi: 10.1109/TR.2025.3530441.

[15] H. Peng, W. Wang, J. Gao, Y. Wang, and J. Du, "A Lightweight Triple-Stream Network With Multisensor Fusion for Enhanced Few-Shot Learning Fault Diagnosis," *IEEE Trans. Reliab.*, vol. 74, no. 3, pp. 4062–4075, Sep. 2025, doi: 10.1109/TR.2025.3540500.

[16] L. Xue, A. Jiang, X. Zheng, Y. Qi, L. He, and Y. Wang, "Few-Shot Fault Diagnosis Based on an Attention-Weighted Relation Network," *Entropy*, vol. 26, no. 1, p. 22, Jan. 2024, doi: 10.3390/e26010022.

[17] J. Li, W. Deng, X. Dang, and H. Zhao, "Cross-Domain Adaptation Fault Diagnosis With Maximum Classifier Discrepancy and Deep Feature Alignment Under Variable Working Conditions," *IEEE Trans. Reliab.*, vol. 74, no. 3, pp. 4106–4115, Sep. 2025, doi: 10.1109/TR.2025.3551155.

[18] X. Liang, M. Zhang, G. Feng, Y. Yu, D. Zhen, and F. Gu, "A Novel Deep Model with Meta-learning for Rolling Bearing Few-shot Fault Diagnosis," *J. Dyn. Monit. Diagn.*, pp. 102–114, Apr. 2023, doi: 10.37965/jdmd.2023.164.

[19] D. Gunning and D. Aha, "DARPA's explainable artificial intelligence (XAI) program," *AI Mag.*, vol. 40, no. 2, pp. 44–58, 2019.

[20] Y. Uchida, K. Fujiwara, T. Saito, and T. Osaka, "Causal plot: Causal-based fault diagnosis method based on causal analysis," *Processes*, vol. 10, no. 11, p. 2269, Nov. 2022, doi: 10.3390/pr10112269.

[21] C. Guo, Z. Shang, J. Ren, Z. Zhao, B. Ding, S. Wang, and X. Chen, "CIS2N: Causal independence and sparse shift network for rotating machinery fault diagnosis in unseen domains," *Reliab. Eng. Syst. Saf.*, vol. 251, p. 110381, Nov. 2024, doi: 10.1016/j.ress.2024.110381.

[22] X. Ding, J. Ying, G. Chen, and J. Xu, "CIRNet: An interpretable cross-component few-shot mechanical fault diagnosis," *IEEE Trans. Reliab.*, pp. 1–15, 2024, doi: 10.1109/TR.2024.3432970.

[23] C. Liu, X. Sun, J. Wang, H. Tang, T. Li, T. Qin, W. Chen, and T.-Y. Liu, "Learning causal semantic representation for out-of-distribution prediction," in *Adv. Neural Inf. Process. Syst.* (NeurIPS), vol. 34, 2021.

[24] Y. Bengio, T. Deleu, N. Rahaman, N. R. Ke, S. Lachapelle, O. Bilaniuk, A. Goyal, and C. Pal, "A meta-transfer objective for learning to disentangle causal mechanisms," in *Proc. Int. Conf. Learn. Representations* (ICLR), 2020.

[25] L. B. Jack and A. K. Nandi, "Fault detection using support vector machines and artificial neural networks, augmented by genetic algorithms," *Mech. Syst. Signal Process.*, vol. 16, no. 2–3, pp. 373–390, 2002, doi: 10.1006/mssp.2001.1454.

[26] B. Samanta, K. R. Al-Balushi, and S. A. Al-Araimi, "Artificial neural networks and support vector machines with genetic algorithm for bearing fault detection," *Eng. Appl. Artif. Intell.*, vol. 16, no. 7–8, pp. 657–665, 2003, doi: 10.1016/j.engappai.2003.09.006.

[27] J. Pearl, "Theoretical impediments to machine learning with seven sparks from the causal revolution," *Proc. 11th ACM Int. Conf. Web Search Data Mining*, Marina Del Rey, CA, USA, 2018, p. 3.

[28] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," *Proc. IEEE*, vol. 109, no. 5, pp. 612–634, May 2021, doi: 10.1109/JPROC.2021.3058954.

[29] J. Splawa-Neyman, D. M. Dabrowska, and T. P. Speed, "On the application of probability theory to agricultural experiments," *Statist. Sci.*, vol. 5, no. 4, pp. 465–472, 1990.

[30] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, "A Survey on Causal Inference," *ACM Trans. Knowl. Discov. Data*, vol. 15, no. 5, pp. 1–46, Oct. 2021, doi: 10.1145/3444944.

[31] J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*. New York, NY, USA: Basic Books, 2018.

[32] M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, and J. Wang, "CausalVAE: Disentangled representation learning via neural structural causal models," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9588–9597, doi: 10.1109/CVPR46437.2021.00947.

[33] W. Yao, G. Chen, and K. Zhang, "Temporally disentangled representation learning," *Proc. 36th Int. Conf. Neural Inf. Process. Syst.*, New Orleans, LA, USA, 2022, Art. 1921, pp. 1–12.

[34] G. Qi and H. Yu, "CMVAE: Causal Meta VAE for unsupervised meta-learning," *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 8, pp. 9485–9493, 2023.

[35] J. Li, Y. Zhang, W. Qiang, L. Si, C. Jiao, X. Hu, C. Zheng, and F. Sun, "Disentangle and remerge: interventional knowledge distillation for few-shot object detection from a conditional causal perspective," *Proc. 37th AAAI Conf. Artif. Intell.*, 2023, Art. 147, pp. 1–11, doi: 10.1609/aaai.v37i1.25216.

[36] A. Wu, J. Yuan, K. Kuang, B. Li, R. Wu, Q. Zhu, Y. T. Zhuang, and F. Wu, "Learning decomposed representations for treatment effect estimation," *IEEE Trans. Knowl. Data Eng.*, pp. 1–1, 2022, doi: 10.1109/TKDE.2022.3150807.

[37] S. Wang, X. Chen, Q. Z. Sheng, Y. Zhang, and L. Yao, "Causal disentangled variational auto-encoder for preference understanding in recommendation," in *Proc. 46th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval (SIGIR '23)*, New York, NY, USA: ACM, Jul. 2023, pp. 1874–1878.

[38] A. Nazaret, J. Hong, E. Azizi, and D. Blei, "Stable

differentiable causal discovery," *Proc. 41st Int. Conf. Mach. Learn. (ICML)*, Vienna, Austria, 2024, Art. no. 1520, pp. 37413–37445.

[39] M. Liu, X. Sun, Y. Qiao, and Y. Wang, "Causal discovery via conditional independence testing with proxy variables," *Proc. 41st Int. Conf. Mach. Learn. (ICML)*, vol. 235, Vienna, Austria: JMLR.org, Jul. 2024, pp. 31866–31889.

[40] Z. Yue, H. Zhang, Q. Sun, and X. Hua, "Interventional few-shot learning," *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2020, Art. 230, pp. 1–13.

[41] J. Li, Y. Wang, Y. Zi, H. Zhang, and Z. Wan, "Causal disentanglement: A generalized bearing fault diagnostic framework in continuous degradation mode," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 6250–6262, Sep. 2023, doi: 10.1109/TNNLS.2021.3135036.

[42] Y. Liu and B. Jafarpour, "Graph attention network with Granger causality map for fault detection and root cause diagnosis," *Comput. Chem. Eng.*, vol. 180, p. 108453, Jan. 2024, doi: 10.1016/j.compchemeng.2023.108453.

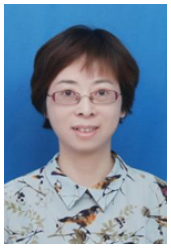
[43] M. Rostami, A. Faysal, H. Wang, A. Sahoo, and R. Antle, "Meta-Task: A Method-Agnostic Framework for Learning to Regularize in Few-Shot Learning," Feb. 2024. Accessed: Oct. 24, 2025. [Online]. Available: https://www.semanticscholar.org/paper/b231c326a733549d3d47df4d05cb75ef74951e6a.

[44] M. Luo, J. Xu, Y. Fan, and J. Zhang, "TRNet: A cross-component few-shot mechanical fault diagnosis," *IEEE Trans. Ind. Informat.*, vol. 19, pp. 6883–6894, 2023.

[45] B. Yang, Y. Lei, F. Jia, and S. Xing, "An intelligent fault diagnosis approach based on transfer learning from laboratory bearings to locomotive bearings," *Mech. Syst. Signal Process.*, vol. 122, pp. 692–706, 2019.

[46] P. Cao, S. Zhang, and J. Tang, "Preprocessing-free gear fault diagnosis using small datasets with deep convolutional neural network-based transfer learning," *IEEE Access*, vol. 6, pp. 26241–26253, 2018.

Jiahan Zhu received her Bachelor's degree from the School of Information and Electronic Engineering at Zhejiang University of Science and Technology. She is currently working toward the M.Sc. degree in the School of Computer and Information Sciences at Hefei University of Technology, Hefei, China. Her current research interests include deep learning and few-shot learning methods for intelligent fault diagnosis.



Juan Xu (Senior Member, IEEE) received the Ph.D. degree from the School of Computer and Information, Hefei University of Technology, Hefei, China, in 2012. She is currently a Professor at Anhui University. Her research interests include Industrial IoT and intelligent fault diagnosis, prognostics and health management, and predictive maintenance.

Qile Ren received the Master degree from the School of Mechanical Engineering , Hefei University of Technology, Hefei, China, in 2007. He works with the Machinery Research Institute Co. His current research interests include intelligent fault diagnosis, prognostics and health management, and predictive maintenance.

Mingguang Dai received the Ph.D. degree in Electrical Engineering from Northwestern Polytechnical University, Xi'an, China, in 2022. He is currently an electrical engineer at Jianghuai Advance Technology Center, Hefei, China. His research interests include high performance servo control, drive motor fault diagnosis, measurement, and detection system development.

Xiaohui Yuan (Senior Member, IEEE) is an Associate Professor at the University of North Texas, Denton, TX, USA, and the Director of the Computer Vision and Intelligent Systems Laboratory. His research interests include artificial intelligence, computer vision, and machine learning. Dr. Yuan was a recipient of the Ralph E. Powe Professor Award in 2008 and the U.S. Air Force Visiting Professor Award in 2011, 2012, and 2013. He serves as an associate editor, an editorial board member, and a guest editor for several journals, and an organizing member for many international conferences.

APPENDIX

PSEUDOCODE OF THE CINET TRAINING PROCEDURE

The following pseudocode describes the training process of the proposed Causal Intervention Network (CINet) under the meta-task learning framework.

Algorithm 1 Training Procedure of Causal Intervention Network (CINet)

Input: Support set $S = \{(x_i, y_i)\}_{i=1}^{c\times k}$ and query set $Q = \{(x_j, y_j)\}_{j=1}^{c\times m}$ in training component
Hyperparameters: α, β, μ, γ, ϵ, margin
Output: Predicted label $\hat{y}_i$
Components: Representation extraction networks $I(\cdot)$, $C(\cdot)$, $A(\cdot)$, regression networks $G_I(\cdot)$ to enforce $I(\cdot)$ to predict Treatment, attention mechanism balancing heterogeneous elements $Att(\cdot)$.
for epoch = 1 **to** epochs **do**
    Combine the data S and Q into the form of meta-task
    Calculate the features of S and Q using $\{I(\cdot), C(\cdot), A(\cdot)\}$
    for each sample feature in task **do**
        Calculate the causality disentanglement loss using Eq. (4)-(7).
        Calculate the probability of belonging to each class using Eq. (8)-(13).
        Calculate the classification loss and contrastive loss using Eq. (14)-(15).
        Update model parameters according to Eq. (16)
    end for
end for