Conditional Contrastive Domain Generalization for Fault Diagnosis

Mohamed Ragab, Zhenghua Chen, *Senior Member, IEEE*, Wenyu Zhang, Emadeldeen Eldele, Min Wu, *Senior Member, IEEE*, Chee-Keong Kwoh, and Xiaoli Li, *Senior Member, IEEE*

***Abstract*—Data-driven fault diagnosis plays a key role in stability and reliability of operations in modern industries. Recently, deep learning has achieved remarkable performance in fault classification tasks. However, in reality, the model can be deployed under highly varying working environments. As a result, the model trained under a certain working environment (i.e., certain distribution) can fail to generalize well on data from different working environments (i.e., different distributions). The naive approach of training a new model for each new working environment would be infeasible in practice. To address this issue, we propose a novel conditional contrastive domain generalization (CCDG) approach for fault diagnosis of rolling machinery, which is able to capture shareable class-information and learn environment-independent representation among data collected from different environments (also known as domains). Specifically, our CCDG attempts to maximize the mutual information of similar classes across different domains while minimizing mutual information among different classes, such that it can learn domain-independent class representation that can be transferable to new *unseen* domains. Our proposed approach significantly outperforms state-of-the-art methods on two real-world fault diagnosis datasets with an average improvement 7.75% and 2.60% respectively. The promising performance of our proposed CCDG on new unseen target domain contributes towards more practical data-driven approaches that can work under challenging real-world environments.**

***Index Terms*—Domain Generalization, Contrasting Learning, Intelligent Fault Diagnosis, Mutual Information**

I. Introduction

Fault diagnosis of rotating machinery plays a key role in reducing maintenance costs, improving reliability, and enhancing safety of operations in industries. The recent years have witnessed a remarkable success of deep learning in fault diagnosis for rotating machinery. However, deep learning may have limited generalization capability under dynamic conditions. For instance, fault diagnosis of rolling-element bearing can encounter highly varying working environments due to many factors such as loading torque, rotation speed, and radial force [1]. Such variability of working environments can yield different sensor readings even for the same fault type. As a result, the data generated from each working condition may have different statistics (i,e., distributions). Under those circumstances, a model that can correctly classify faults under one working environment may perform poorly when tested on data from different working environments, which is well-known as the domain shift problem. Such a problem can significantly hinder the model's generalization performance on fault diagnosis tasks [2], [3].

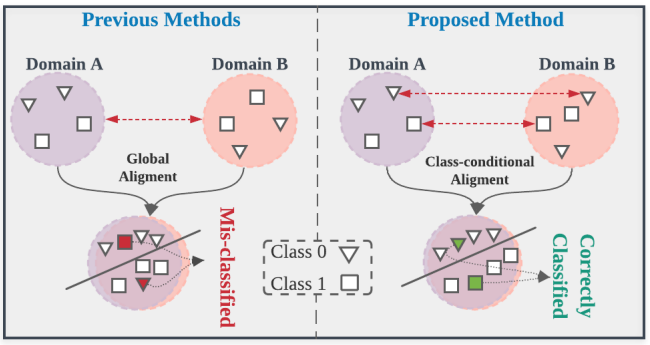


Fig. 1: (Best viewed in colors). Comparison between previous and proposed methods in distribution matching. **Left**: Aligning the marginal distribution without considering fine-grained class distributions tends to mis-classify samples among domains. **Right**: Our proposed approach with its class-conditional contrastive loss can successfully align the classes among different domains.

Domain adaptation (DA) aims to transfer knowledge from a labeled source domain to an unlabeled target domain, while tackling the domain shift problem. Recently, several attempts have been made to address the domain shift problem in data-driven fault diagnosis [1], [4]–[8]. For instance, Sohaib *et al.* consolidated a convolutional neural network (CNN) with bi-spectrum features to achieve more robust diagnostic performance [4]. Guo *et al.* integrated maximum mean discrepancy (MMD) with a domain confusion loss to adapt between different machines [7]. Yang *et al.* introduced polynomial kernel based MMD approach to reduce the computational complexity and improve the transferability of the learned features [8]. However, DA approaches may not be easily applicable for many practical scenarios due to the following reasons. First, DA approaches usually assume access to the target domain data during the training phase, which may not be attainable when encountering new target domain with no prior data. Additionally, for each new target domain, we need to retrain a new model independently, which can be time-consuming

Mohamed Ragab, Emadeldeen Eldele, and Chee-Keong Kwoh are with school of Computer Science and Engineering at Nanyang Technological University, Singapore (Email: mohamedr002@e.ntu.edu.sg, emad0002@ntu.edu.sg, asckkwoh@ntu.edu.sg).

Zhenghua Chen, Wenyu Zhang, Min Wu, and Xiaoli Li are with Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, Sinagpore 138632 (Email: chen0832@e.ntu.edu.sg, zhang_wenyu@i2r.a-star.edu.sg, wumin@i2r.a-star.edu.sg, xlli@i2r.a-star.edu.sg).

and not scalable solutions for dynamic working environments. Hence, there is an urgent need for a more realistic fault diagnosis model that can generalize to new unseen domains without prior data.

To tackle this issue, Domain Generalization (DG), a more practical yet challenging scenario, leverages data from multiple source domains to generalize well to new unseen domains. Few studies have investigated domain generalization for fault diagnosis [9]–[12]. A predominant approach is to leverage adversarial learning to minimize the discrepancy of the marginal distribution between multiple source domains. Yet, in fault classification tasks, the marginal distribution of the data can inherently encompass a multi-modal class structures. Thus, only aligning the marginal distribution without considering the fine-grained class distribution within each domain tends to fail in some challenging scenarios [13], as shown in Fig. 1. Recently, contrastive learning has achieved widely acclaimed performance in representation learning for visual applications. The key idea is to learn feature representations such that similar samples (i.e., positive pairs) are pulled together while dissimilar samples (i.e., negative pairs) are pushed away [14]. However, most of existing contrastive learning approaches heavily rely on domain-specific augmentations to construct the positive pairs, which can be laborious and require extensive expert knowledge.

In this work, we propose a novel conditional contrastive domain generalization approach (CCDG) for fault diagnosis. Particularly, instead of relying on the laborious domain-specific augmentations, we leverage the variability among multi-domains (i.e., working conditions) to define more representative and realistic positive and negative pairs. Additionally, to realize class-conditional invariance, we maximize the mutual information between the prediction scores of same classes among different domains while minimizing mutual information among different classes.

The main contributions of this paper are summarized as follows.

1) We propose a novel domain generalization approach to tackle a more realistic yet challenging task in real-world fault diagnosis. It does not require any data from the target domain during training.
2) We design a new conditional contrastive loss across realistic pairs from multi-source domains to find the domain invariant class representation, leading to robust and discriminative representations for domain generalization.
3) We provide a theoretical derivation of the mutual information lower bound of our conditional contrastive loss on multi-domain data.
4) We extensively evaluate our proposed CCDG approach on two rotating machinery datasets. The experimental results clearly show the efficacy of our approach in generalizing to new unseen domains with no prior data.

II. Related Work

A. Domain Adaptation and Generalization

Majority of domain adaptation approaches focus on finding domain invariant features between the source and target domains. To do so, some approaches aim to minimize the statistical distance between the source and target features among different moments of the distribution [15]–[18]. Another line of research, inspired by generative adversarial networks, leverages adversarial learning to find a target feature representation that can be indistinguishable from the source by a discriminator network [19]–[21].

Different from domain adaptation, domain generalization assumes no access for target domain data. To tackle domain generalization task, various approaches have been developed to tackle domain generalization task. One solution targeted to learn domain invariant features across the multiple source domains [22], [23]. Another line of research exploited various data augmentation strategies to improve the generalization performance [24], [25]. However, all these approaches are specifically designed for computer vision applications, and may not be extendable for fault diagnosis tasks.

B. Cross-domain Fault Diagnosis

In realistic fault diagnosis environments, domain shift problem can be prominent due to the variability of working conditions. Several approaches aim to tackle domain shift problem for fault diagnosis. For instance, Song *et al.* augmented the adversarial adaptation with a pseudo-label retraining strategy for fault diagnosis task [26]. While Chen *et al.* proposed a domain adversarial transfer network to find domain invariant features for fault diagnosis of rotary machinery [3]. Jiao *et al.* adversarial trained shared feature extractor network against two task-specific classifiers to better align the source and target domains [6]. Zhang *et al.* jointly aligned the feature and class distributions to improve the cross-domain performance [27]. Ragab et al. proposed a more scalable adversarial adaptation approach that can generalize to multiple target domains concurrently [1]. Recently, some approaches have been developed

Yet, all these approaches require access to target data during training, which may not be attainable for many real-world scenarios. Besides, it requires to train a new model for each new domain. Therefore, there is a necessity to develop approaches that can generalize to new unseen domains.

Few works have realized domain generalization for fault diagnosis problems. For instance, Li *et al.* proposed a gradient reversal layer and a domain augmentation to improve generalization for machinery fault diagnosis [10]. While Liao *et al.* leveraged Wasserestien generative adversarial network with gradient penalty to learn the invariant features to generalize well to new working conditions [11]. Differently, Zhang *et al.* relied on saliency maps to remove superficial features and improve the performance of fault diagnosis [28]. Han *et al.* leveraged adversarial learning and triplet loss to improve generalization performance on unseen target domains [12]. Most of these approaches enforce invariance between the global distribution between domains to improve the generalization performance. However, in fault classification tasks, the global distribution of each working environment encloses multiple fine-grained distributions for each class. Furthermore, aligning the global distribution between different working environments does not obligate the alignment of these fine-grained class distribution within each domain. Consequently,

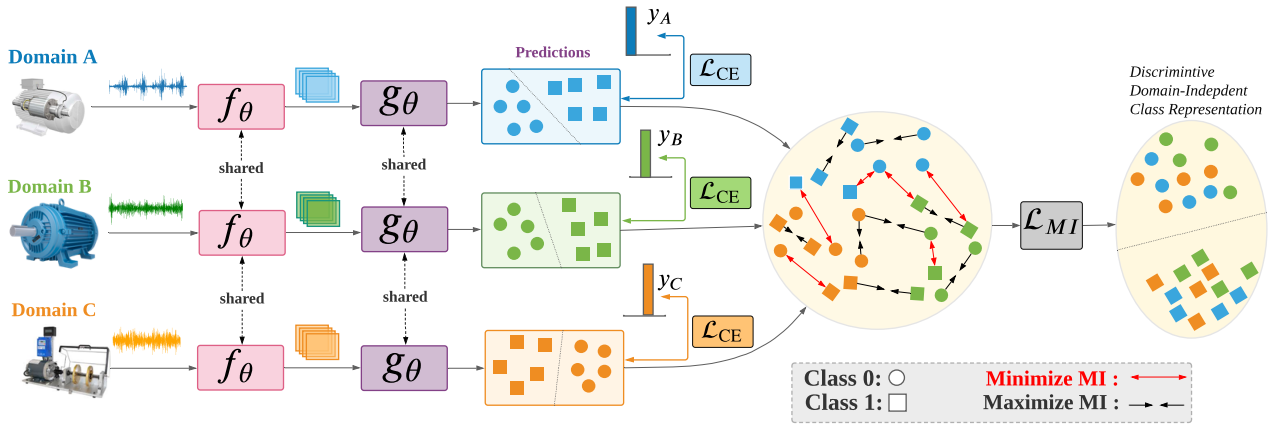


Fig. 2: Framework of the proposed conditional contrastive domain generalization (CCDG). First, we pass the data from different domains through a shared network f_θ (i.e., feature extractor) to find the feature representation of the data. Second, a shared prediction network g_θ (i.e., predictor) converts the features to prediction scores, and we calculate the task-specific cross-entropy loss $\mathcal{L}_{CE}$ between the predicted labels and true labels. Third, we minimize a novel conditional contrastive loss $\mathcal{L}_{MI}$ to maximize both the intra-class similarity and inter-class separability. We train both the feature extractor f_θ and the predictor g_θ by optimizing $\mathcal{L}_{CE}$ and $\mathcal{L}_{MI}$ together (Best viewed in colors).

the fine grained class-distributions can still be mis-aligned even if the global distributions are well matched, yielding poor cross-domain performance for the trained model. Differently, our CCDG approach considers the class-conditional distribution by maximizing mutual information among same class in different domains to obtain environment independent class representation.

III. Methodology

In this section, we first formulate the domain generalization problem. Then, an overview of our CCDG approach is presented. Subsequently, we describe each loss component of our CCDG approach in details. Last, we provide an information-theoretic perspective for our class-conditional contrastive loss.

A. Problem Formulation

In this subjection, we provide a mathematical formulation for the domain generalization and domain invariance problems.

1) Domain Generalization Problem: Let $\mathcal{X}$ denote the feature space and $\mathcal{Y}$ denote the label space, a domain is represented by the joint distribution P_{XY} defined on $\mathcal{X} \times \mathcal{Y}$. In the domain generalization problem, we are given multiple source domains $(\mathcal{D}^1, \mathcal{D}^2, \ldots, \mathcal{D}^M\}$, where M is the total number of source domains. Each domain $\mathcal{D}^i = \{\mathbf{x}_j^i, y_j^i\}_{j=1}^{N_i} \sim P_{XY}^i$ has N_i number of samples. The joint distributions vary among the multiple source domains, i.e., $P_{XY}^1 \neq P_{XY}^2 \ldots \neq P_{XY}^M$. The objective of the domain generalization is to learn a model $h : \mathcal{X} \rightarrow \mathcal{Y}$ based on all the source domains and generalize well to a new unseen target domain $D^{test} = \{\mathbf{x}_j^{test}, y_j^{test}\}_{j=1}^{N_{test}} \sim P_{XY}^{test}$, where $P_{XY}^{test} \neq P_{XY}^i \; for \; i \in \{1, \ldots, M\}$. It worth emphasizing that we do not have any access to D^{test} during the training procedure.

2) Domain Invariance Problem: Most of existing domain generalization approaches assume that only marginal distribution changes among domains, i.e., $P_1^S(X) \neq P_2^S(X) \ldots \neq P_M^S(X)$, while the conditional distribution $P(Y|X)$ remains stable. Therefore, most of the existing approaches address the domain generalization problem via finding domain invariant representation on the feature space such that $P_1^S(f_\theta(X)) = P_2^S(f_\theta(X)) \cdots = P_M^S(f_\theta(X))$, where $f_\theta : \mathcal{X} \rightarrow \mathcal{F}$ is a deep neural network that maps from the space of the raw input time series signals (i.e., $\mathcal{X}$) to the vectorized feature space (i.e., $\mathcal{F}$).

In reality, $P(Y|X)$ may also be unstable among the domains. Therefore, only finding domain invariant features can be insufficient condition to perfectly align the domains.

B. Overview of CCDG

In this work, we propose a novel conditional contrastive domain generalization (CCDG) approach that aligns the class-conditional distributions among multiple source domains to boost the generalization performance on unseen target domains. Fig. 2 shows the overall framework of our proposed CCDG method. First, we design a shared feature extractor f_θ for feature learning and a shared predictor g_θ for fault diagnosis. Second, we calculate the task-specific cross-entropy loss $\mathcal{L}_{CE}$ between the predicted labels and true labels. Third, we propose a novel conditional contrastive loss $\mathcal{L}_{MI}$ to maximize the mutual information among same classes from different domains while minimizing the mutual information among different classes. Eventually, we train both the feature extractor f_θ and the predictor g_θ by optimizing the task-specific loss $\mathcal{L}_{CE}$ and the conditional contrastive loss $\mathcal{L}_{MI}$ concurrently in an end-to-end manner. In the next subsections, we will introduce these two losses, i.e., $\mathcal{L}_{CE}$ and $\mathcal{L}_{MI}$, in details.

C. Task-specific Learning

The first step in our CCDG is to optimize the model performance on the corresponding task by leveraging the known class labels in the training data. The task-specific cross-entropy loss is formalized in Equation 1 below.

$$\mathcal{L}_{CE} = -\frac{1}{M}\sum_{i=1}^{M}\frac{1}{N_i}\sum_{j=1}^{N_i}\left[\ell(h_\theta(\mathbf{x}_j^i), y_j^i))\right], \tag{1}$$

where $\mathbf{x}_j^i$ refers to the j^{th} sample from the domain D^i, y_j^i is the class label of $\mathbf{x}_j^i$, and $h_\theta(\mathbf{x}_j^i) = g_\theta(f_\theta(\mathbf{x}_j^i))$ is a function that maps the raw inputs to the class predictions. Given C classes, both $\hat{y}_j^i$ and y_j^i are C-dimensional vectors, where $\hat{y}_j^i = h_\theta(\mathbf{x}_j^i)$ and y_j^i represent the predicted and true class labels, respectively. Then, the cross-entropy loss between the predicted labels and the true ones $\ell(\hat{y}_j^i, y_j^i)$ can be represented as follows:

$$\ell(\hat{y}_j^i, y_j^i) = \sum_{c=1}^{C} y(c)_j^i \times \log \hat{y}(c)_j^i. \tag{2}$$

Here, $y(c)_j^i$ is 1 if this sample $\mathbf{x}_j^i$ is from class c and 0 otherwise, and $\hat{y}(r)_j^i$ is the predicted score for the class c.

D. Class-conditional Contrastive Loss

The variability of working environments among the source domains can produce task irrelevant feature on each source domain. As a result, the class representation can be different from one domain to another. Such inconsistency of class representations among the multiple source domains can harm the generalization performance on new unseen target domains. Therefore, only optimizing the cross-entropy loss $\mathcal{L}_{CE}$ without considering the cross-domain relationship can have negative impact on the generalization performance. To tackle this issue, we jointly optimize our class-conditional contrastive loss $\mathcal{L}_{MI}$. Particularly, we consider the cross-domain relationship via maximizing the mutual information between similar classes across different domains while removing the task-irrelevant information that is caused by the change in working environments. By doing so, our $\mathcal{L}_{MI}$ can capture the shared class information and obtain environment independent class representation which can boost generalization performance on new unseen environments (i.e., domains).

Formally, given an anchor sample $\mathbf{x}_u \in B = \{b_1, b_2, \ldots, b_M\}$, where B is the set of batches from M source domains. We aim to maximize mutual information between $\mathbf{x}_u$ and the corresponding positive samples in B while minimizing its similarity with the corresponding negative pairs in B. In particular, the positive samples of $\mathbf{x}_u$ are represented by all the samples in B that belong to the same class $pos(u) = \{\mathbf{x}_v \in B : y_u = y_v\}$. While the negative samples are the remaining samples in B that belong to different classes $neg(u) = \{\mathbf{x}_k \in B : y_u \neq y_k\}$. Our proposed approach is different from traditional approaches in two aspects. First, traditional approaches only align the global feature representation between different domains. As a result, samples that belong to different classes (i.e., $\mathbf{x}_u$ and $\mathbf{x}_k$) can be pulled together, despite that the global distributions of different domains are well aligned. Differently, we consider a class-wise alignment where only samples that belong to the same class are pulled together. Second, these approaches usually applied on the feature space (i.e., $f_\theta(\mathbf{x}_j^i)$. While our CCDG approach is applied on the class prediction level (i.e., $g_\theta(f_\theta(\mathbf{x}_j^i))$) of the corresponding samples, which enables the model to find class-conditional invariant representation among multiple source domains.

The class-conditional contrastive loss of the sample $\mathbf{x}_u$, denoted as $\mathcal{L}_{MI}^u$. To formalize our class-conditional contrastive loss, we follow standard notations of contrastive learning in [29], which is represented as follows:

$$\begin{aligned}\mathcal{L}_{MI}^u &= \frac{-1}{|pos(u)|}\sum_{v\in pos(u)}\left(\log\frac{e^{(\sigma(\boldsymbol{h}_u,\boldsymbol{h}_v)/\tau)}}{\sum_{k\in neg(u)}e^{(\sigma(\boldsymbol{h}_u,\boldsymbol{h}_k)/\tau)}}\right)\\ &= \frac{-1}{|pos(u)|}\sum_{v\in pos(u)}\left(\log\underbrace{e^{(\sigma(\boldsymbol{h}_u,\boldsymbol{h}_v)/\tau)}}_{\text{positives}}\right.\\ &\qquad\left.-\log\sum_{k\in neg(u)}\underbrace{e^{(\sigma(\boldsymbol{h}_u,\boldsymbol{h}_k)/\tau)}}_{\text{negatives}}\right).\end{aligned} \tag{3}$$

Here, $|pos(u)|$ is cardinality of the positive sample set, $\sigma(\mathbf{a},\mathbf{b}) = \frac{\mathbf{a}^\intercal\mathbf{b}}{|\mathbf{a}||\mathbf{b}|}$ is the similarity scoring function for any given vectors $\mathbf{a},\mathbf{b}$, and τ is the temperature parameter to control the contrastive power. $\boldsymbol{h}_u = g_\theta \odot f_\theta(\mathbf{x}_u) \in \mathbb{R}^C$ are the logits of $\mathbf{x}_u$, $g_\theta \odot f_\theta$ represents the composition of the feature extractor and the predictor, and C is the dimension of the logits vector (i.e., the number of classes). The $\mathcal{L}_{MI}^u$ represents the contrastive loss for the anchor sample $\mathbf{x}_u$. Particularly, via optimizing $\mathcal{L}_{MI}^u$, we maximize the mutual information between the classifier prediction of the anchor sample $\boldsymbol{h}_u$ and all corresponding positive samples (i.e., samples that belong to the same class of the anchor sample $pos(u)$) across different domains, where $\boldsymbol{h}_v$ is sampled from the set of positive samples. Concurrently, we minimize the mutual information between the anchor sample and all the negative samples (i.e., all samples that belong to different class from the anchor sample $neg(u)$.

The overall contrastive loss for all the samples is calculated as follows:

$$\mathcal{L}_{MI} = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{N_i}\sum_{u=1}^{N_i}\mathcal{L}_{MI}^u. \tag{4}$$

By minimizing $\mathcal{L}_{MI}$, we can maximize the lower bound of the mutual information among the positive samples [30].

E. Connections to Mutual Information

Here, we provide the theoretical derivation of the lower bound of mutual information of our multi-domain class-conditional contrasting loss. Formally speaking, as in [30], the optimal similarity score is proportional to the density

ratio between the joint distribution and product of marginal distributions. This can be formulated as follows:

$$\phi(a,b) \propto [\frac{p\,(a,b)}{p\,(a)\,p(b)}], \tag{5}$$

where $p\,(a,b)$ is the joint distribution of (a,b), and $p(a)p(b)$ is the product of marginal distributions.

In our approach, we aim to obtain the optimal contrastive loss $\mathcal{L}_{\text{MI}}^{opt}$ for multi-domain setting. Particularly, for anchor $\boldsymbol{h}_u$, we contrast between the positive sample in the numerator, i.e., $(\boldsymbol{h}_u, \boldsymbol{h_v})$ for $v \in pos(u)$, and all samples in the denominator, i.e., $(\boldsymbol{h}_u, \boldsymbol{h_i})$ for $i \in \{0, \ldots, k\}$. It is assumed that the positive sample is at index $i = 0$, and the rest are negative samples, as in [31]. We substitute ϕ function from Eq. 5 into Eq. 6. Subsequently, we split the summation of the denominator into two parts: the term with $i = 0$, and the summation from $i = 1$ to k. Given that $\mathcal{L}_{\text{MI}}^{opt}$ is the optimal contrastive loss, the mutual information between the anchor sample and negative sample, i.e. $(\boldsymbol{h}_u, \boldsymbol{h_i})$ for $i > 0$, will be minimized. This means that the two samples can be considered as approximately independent from each other and the ratio is $\frac{p(\boldsymbol{h_u},\boldsymbol{h_i})}{p(\boldsymbol{h_u})p(\boldsymbol{h_i})} \approx 1$ for $i > 0$ [30]. Last, given that $log(k)$ is independent from the summation, we take it out while replacing the remaining summation term by the mutual information notation (i.e., I). The full derivation can be viewed as follows:

$$\begin{aligned}
\mathcal{L}_{\text{MI}}^{opt} &= -\sum_{v\in pos(u)} \log\left[\frac{\phi\,(\boldsymbol{h}_u,\boldsymbol{h}_v)}{\sum_{i=0}^{k}\phi\,(\boldsymbol{h}_u,\boldsymbol{h_i})}\right] \\
&= -\sum_{v\in pos(u)} \log\left[\frac{\frac{p(\boldsymbol{h}_u,\boldsymbol{h}_v)}{p(\boldsymbol{h}_u)p(\boldsymbol{h}_v)}}{\sum_{i=0}^{k}\frac{p(\boldsymbol{h}_u,\boldsymbol{h_i})}{p(\boldsymbol{h}_u)p(\boldsymbol{h_i})}}\right] \\
&= -\sum_{v\in pos(u)} \log\left[\frac{\frac{p(\boldsymbol{h}_u,\boldsymbol{h}_v)}{p(\boldsymbol{h}_u)p(\boldsymbol{h}_v)}}{\frac{p(\boldsymbol{h}_u,\boldsymbol{h_i})}{p(\boldsymbol{h}_u)p(\boldsymbol{h_i})}\Big|_{i=0} + \sum_{i=1}^{k}\frac{p(\boldsymbol{h}_u,\boldsymbol{h_i})}{p(\boldsymbol{h}_u)p(\boldsymbol{h_i})}}\right] \\
&= \sum_{v\in pos(u)} \log\left[\frac{\frac{p(\boldsymbol{h}_u,\boldsymbol{h}_v)}{p(\boldsymbol{h}_u)p(\boldsymbol{h}_v)}}{\frac{p(\boldsymbol{h}_u,\boldsymbol{h_i})}{p(\boldsymbol{h}_u)p(\boldsymbol{h_i})}\Big|_{i=0} + \sum_{i=1}^{k}\frac{p(\boldsymbol{h}_u,\boldsymbol{h_i})}{p(\boldsymbol{h}_u)p(\boldsymbol{h_i})}}\right]^{-1} \\
&= \sum_{v\in pos(u)} \log\left[\frac{\frac{p(\boldsymbol{h}_u,\boldsymbol{h_i})}{p(\boldsymbol{h}_u)p(\boldsymbol{h_i})}\Big|_{i=0} + \sum_{i=1}^{k}\frac{p(\boldsymbol{h}_u,\boldsymbol{h_i})}{p(\boldsymbol{h}_u)p(\boldsymbol{h_i})}}{\frac{p(\boldsymbol{h}_u,\boldsymbol{h}_v)}{p(\boldsymbol{h}_u)p(\boldsymbol{h}_v)}}\right] \\
&= \sum_{v\in pos(u)} \log\left[1 + \frac{p\,(\boldsymbol{h}_u)\,p\,(\boldsymbol{h}_v)}{p\,(\boldsymbol{h}_u,\boldsymbol{h}_v)}\sum_{i=1}^{k}\frac{p\,(\boldsymbol{h}_u,\boldsymbol{h_i})}{p\,(\boldsymbol{h}_u)\,p\,(\boldsymbol{h_i})}\right] \\
&\approx \sum_{v\in pos(u)} \log\left[1 + \frac{p\,(\boldsymbol{h}_u)\,p\,(\boldsymbol{h}_v)}{p\,(\boldsymbol{h}_u,\boldsymbol{h}_v)}k\right] \\
&\geq \log(k) - \sum_{v\in pos(u)} \log\left[\frac{p\,(\boldsymbol{h}_u,\boldsymbol{h}_v)}{p\,(\boldsymbol{h}_u)\,p\,(\boldsymbol{h}_v)}\right] \\
&\geq \log(k) - \sum_{v\in pos(u)} I\,(\boldsymbol{h}_u;\boldsymbol{h}_v)\,.
\end{aligned} \tag{6}$$

With all the available positive samples $\sum_{v\in pos(u)} I\,(\boldsymbol{h}_u;\boldsymbol{h}_v) \geq \log(k) - \mathcal{L}_{\text{MI}}^{opt}$, by minimizing $\mathcal{L}_{\text{MI}}^{opt}(\boldsymbol{h}_u,\boldsymbol{h}_v)$, we can maximize the lower bound on the mutual information $I\,(\boldsymbol{h}_u;\boldsymbol{h}_v)$. Notably, as the number of negative samples k increases, the approximation can be more accurate.

F. Overall Loss for Optimization

In our CCDG, both $\mathcal{L}_{MI}$ and $\mathcal{L}_{CE}$ are jointly optimized to improve the generalization performance on new unseen domains. Particularly, the objective of the cross-entropy loss is to improve the task specific performance within each source domain. While the objective of the contrastive loss $\mathcal{L}_{MI}$ is to consider cross-domain relationships via finding domain-independent class representation, which can boost the generalization performance on unseen domains. The overall objective is then represented by a convex combination between them in Equation 7.

$$\min_{f_\theta, g_\theta} \mathcal{L} = \alpha\mathcal{L}_{MI} + (1-\alpha)\mathcal{L}_{CE}, \tag{7}$$

where α is the weight between the two losses. We use Adam [32] as the optimizer to minimize the overall objective and learn the feature extractor f_θ and the predictor network g_θ.

IV. Experiments and Results

In this section, we first introduce the datasets and the setup of our experiments. Then, we present the evaluation results of our proposed CCDG method.

A. Datasets

1) CWRU Dataset: The Case Western Reserve University (CWRU) dataset is a widely adopted dataset for rolling bearing elements [33]. Fig 3 shows the test rig for CWRU dataset. Accelerometer sensors were deployed at both the drive-end and fan-end of the housing motors. Vibration signals were collected with 12 KHz sampling rate under eight different operating conditions. Particularly, we have four different operating conditions with different loading torques collected from the drive-end, denoted as domain A, B, C, and D. Similarly, we have other four operating conditions collected from the fan-end of the motor, denoted as domain F, G, H, and I. For each operating condition, there are one healthy state and three faulty states, namely, inner fault (IF), outer fault (OF), and bearing fault (BF). Each faulty state has three levels of severity with dimensions of 7, 14, 21 mil. In total, we have 10 classes with 1 healthy class and 9 faulty classes. Table Ishows the detailed description of the CWRU dataset. To prepare the data for our experiments, we partitioned the sensor readings into smaller samples using sliding windows with a fixed length of 4,096 and the shifting step of 290, which is widely used in previous studies [1], [34]. Overall, we can generate 4,000 samples for each domain.

TABLE I: Details of CWRU bearing dataset.

Domain	Torque	Location	Fault Type	Fault Size (inches)
A	0 hp	Drive End	Normal, IF, OF, BF	0, 007, 0.014, 0.021
B	1 hp	Drive End	Normal, IF, OF, BF	0, 007, 0.014, 0.021
C	2 hp	Drive End	Normal, IF, OF, BF	0, 007, 0.014, 0.021
D	3 hp	Drive End	Normal, IF, OF, BF	0, 007, 0.014, 0.021
E	0 hp	Fan End	Normal, IF, OF, BF	0, 007, 0.014, 0.021
F	1 hp	Fan End	Normal, IF, OF, BF	0, 007, 0.014, 0.021
G	2 hp	Fan End	Normal, IF, OF, BF	0, 007, 0.014, 0.021
H	3 hp	Fan End	Normal, IF, OF, BF	0, 007, 0.014, 0.021

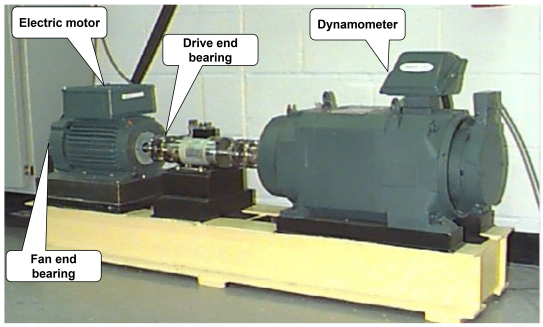


Fig. 3: Experimental equipment of CWRU dataset [33].

2) Paderborn Dataset: The second dataset was generated by the KAT data center in Paderborn University with the sampling rate of 64 KHz [35]. The test rig of Paderborn dataset is shown in Fig 4. The damages were generated using both artificial and natural ways. More specifically, an electric discharge machine (EDM), a drilling, and an electric engraving were used to manually produce the artificial faults. While the natural damages were caused by using accelerated run-to-failure tests. The data collection process for both types of damages, i.e., artificial and real, was exposed under working conditions with different operating parameters such as loading torque, rotational speed and radial force. In total, Paderborn dataset were collect under 6 different operating conditions including 3 conditions with artificial damages (denoted as domain I, J and K) and 3 conditions with real damages (denoted as domain L, M and N). Table II demonstrates the detailed specifications of each working condition. For example, the loading torque varies from 0.1 to 0.7 Nm and the radial force varies from 400 to 1000 N, while the rotational speed is fixed at 1500 RPM. Each operating condition (i.e., domain) contains three classes, namely, healthy class, inner fault (IF) class, and outer fault (OF) class. To prepare the data samples for Paderborn dataset, we adopted sliding windows with the fixed length of 5,120 and the shifting size of 4,096 [1]. As such, we generated 12,340 for each artificial domain (i.e., I, J and K) and 13,640 samples for each real domain (i.e., L, M and N) respectively.

B. Experimental Setup

1) Baseline Methods: To show the efficacy of our proposed CCDG method, we first adapted state-of-the-art methods on visual domain generalization. Besides, we re-implemented domain generalization methods proposed for fault diagnosis tasks. To ensure fair evaluation, same backbone architecture has been used for both our approach and the baseline methods. The compared baselines are shown as follows:

- Empirical Risk Minimization (**ERM**) [36]: it minimizes the sum of the empirical risk among the samples from all the domains.
- Maximum Mean Discrepancy (**MMD**) [37]: it minimizes the MMD distance between each pair of domains to improve the generalization performance.
- Deep Correlation Alignment (**Deep CORAL**) [38]: it aligns the co-variance matrices for each pair of domains.
- Conditional Domain Adversarial Networks (**CDANN**) [39]: it applies a separate domain classifier for each class to further improve the alignment performance.
- Beyond empirical risk minimization **Mixup** [40]: it generates new data samples via linear interpolation of samples and labels among random pair of domains.
- Self-supervised Contrastive Regularization for Domain Generalization (**Self-Reg**) [41]: it applies stochastic weight averaging with inter-domain curriculum learning with contrastive regularization.
- A Hybrid Generalization Network for Intelligent Fault Diagnosis (**IEDGNet**) [12]: it leverages both extrinsic and intrinsic generalization objectives to regularize the discriminating performance of the deep neural network on fault diagnosis tasks.
- Domain Generalization for Rotating machinery (**DGRM**) [10]: it adversarially trains a feature extractor against a multi-class domain discriminator network to to improve the generalization performance of rotating machinery.

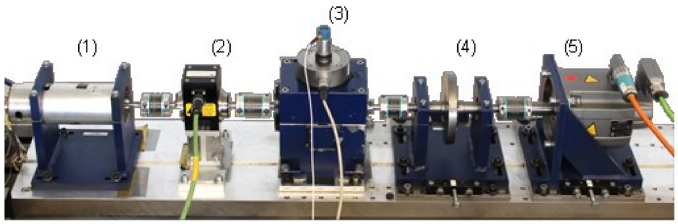


Fig. 4: Test rig for Paderborn dataset [35]. The rig is composed of (1) Electric motor; (2) Torque measurement; (3) Bearing test modular; (4) Fly wheel; (5) Load motor.

TABLE II: Details for Paderborn dataset.

Domain	Damage Type	Load Torque [Nm]	Radial Force [N]
I	Artificial	0.1	1000
J	Artificial	0.7	400
K	Artificial	0.7	1000
L	Real	0.1	1000
M	Real	0.7	400
N	Real	0.7	1000

2) Implementation Details: In our experiments, during the training phase, we split the data from multiple source domains

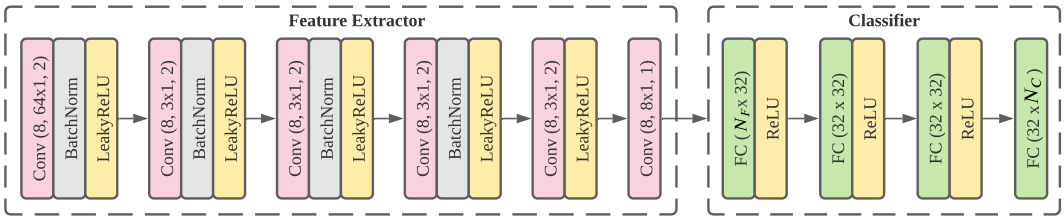


Fig. 5: Architecture of backbone model with a 6-layer 1-D CNN as feature extractor and a 4-layer FCN as classifier, where N_F represents the number of output features from the CNN and N_C is the number of output classes.

TABLE III: Details of hyper-parameter tuning setup.

Method	Hyperparameter	Range
MMD	Regularization λ	10^{-3} to 10^{-1}
Deep CORAL	Regularization λ	10^{-3} to 10^{-1}
DGRM, CDANN	Discriminator lr	10^{-5} to $10^{-3.5}$
	Discriminator weight decay	10^{-6} to 10^{-2}
	Discriminator Adam β_1	{0, 0.5}
	Discriminator steps	2^0 to 2^3
	Discriminator GP	10^{-2} to 10^{1}
	Adversarial regularization λ	10^{-2} to 10^{2}
Mixup	Beta shape parameter α	10^{-1} to 10^{1}
IEDGNet	adversarial regularization λ	10^{-2} to 10^{2}
	triplet loss weight β	10^{-2} to 10^{2}
CCDG	Temperature τ	10^{-1} to 10^{0}
	Contrastive weight α	10^{-1} to 10^{0}

into 80% for training and 20% for validation to monitor the model convergence on the source domains during the training phase. To test the generalization performance, we iteratively leave one whole domain out for testing while using all the other domains for training, following the standard protocol of domain generalization [42]. It worth highlighting that we neither use the testing data for the training nor for the model selection. Instead, we select the model based on the its performance on the validation sets of the source domains, which can be more practical and aligns with domain generalization assumption that no test data are available during the training process. The accuracy score (i.e., the number of correctly classified test samples divided by the total number of test samples) is utilized to measure the diagnosis performance. All the experiments were conducted using PyTorch 1.7 on NVIDIA GeForce RTX 2080 Ti GPU. We implemented a 6-layer 1-D CNN to extract features from the raw input signals followed by a 4-layer fully connected network (FC) for fault classification, as shown in Fig. 5. We used LeakyReLU as non-linear activation layer and batch normalization to speed up the training process. To promote fair evaluation, same network architecture is adopted for all the baseline methods.

To train the model, we fix the learning rate as 0.001, weight decay as $5e^{-5}$, and batch size as 32 through the whole experiments. To select the hyper-parameters, we ran 20 trials of hyper-parameter sweep for our approach and all baseline methods to ensure a fair evaluation. We sample the hyper-parameter values from a *uniform* distribution within a specific ranges [42]. Detailed ranges for our approach and baseline methods can be found in Table III. Note that the ERM method does not contain any tunable hyper-parameters in this setup. For more practical model selection, we validate our model only on the held-out subsets of the source domains without accessing to any data from the target domain.

C. Experimental Results

1) Results on CWRU Dataset: Table IV shows the results of different approaches on the CWRU dataset. In particular, column A in Table IV means that we use domain A as the unseen target domain and the other 7 domains as source domains. Last column shows the average performance for various methods over 8 target domains. Note that the best values on each target domain are highlighted in bold, while the second best values are underlined. It can be found that our CCDG approach significantly outperforms all the baseline approaches. On average, the CCDG outperforms the second best baseline (i.e., Mixup) by 7.75%.

Domains A and E have zero loading torque, while other domains have non-zero loading torque. Therefore, fault features in these two domains could be quite different from those in other domains, and it would be challenging to map other domains to A and E. This explains why various methods achieve unsatisfactory performance on A and E. However, our CCDG approach, with its class-conditional invariance, can still perform the best on these two challenging domains. It outperforms ERM (second best) on domain A by 12.62%, and outperforms Self-Reg (second best) on domain E by 6.54%, respectively.

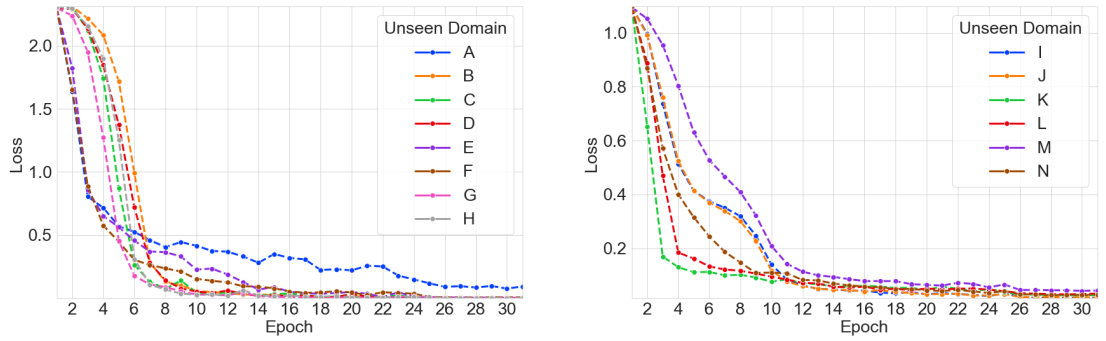
2) Results on Paderborn Dataset: To further evaluate the superiority of our CCDG method, we have conducted additional experiments on the Paderborn dataset. Table V shows the performance of various methods on the Paderborn dataset. It can be clearly observed that our CCDG approach outperforms the state-of-the-art methods on 5 out of 6 domains with an average accuracy of 88.52%. Moreover, it outperforms the second best baseline (i.e., Self-Reg) by 2.60%. Notably, generalization to domains with small radial force (i.e., J and M) can be a very challenging task. However, our CCDG method with its class-conditional contrasting can still be the best on these two domains, demonstrating the robustness of our method against the large domain shifts. Additionally, unlike Self-Reg that only focuses on positive pairs, our CCDG approach aims to push away negative pairs that belong to different classes, which can improve the generalization performance. Moreover, methods that only align the marginal distribution like Deep CORAL and MMD perform poorly on the tough domains, i.e., J and M.

TABLE IV: Domain generalization results on CWRU bearing dataset (Accuracy %).

	Unseen Target Domains								
Method	A	B	C	D	E	F	G	H	Average
ERM	64.17 ± 9.2	94.58 ± 1.6	96.38 ± 2.9	70.58 ± 4.4	68.33 ± 7.7	83.46 ± 3.1	94.38 ± 4.8	86.38 ± 4.6	82.28
MMD	47.83 ± 9.2	57.17 ± 13.4	75.67 ± 3.7	66.67 ± 1.2	61.54 ± 16.0	73.17 ± 8.5	71.75 ± 10.8	54.17 ± 17.7	63.50
Deep CORAL	63.29 ± 15.1	70.96 ± 8.8	87.25 ± 4.4	72.13 ± 4.0	66.67 ± 3.5	71.71 ± 6.9	68.71 ± 3.8	76.71 ± 5.0	72.18
CDANN	64.08 ± 11.6	82.67 ± 4.9	90.13 ± 6.7	77.50 ± 5.4	72.13 ± 8.5	74.75 ± 3.6	89.21 ± 7.3	74.29 ± 11.3	78.09
Mixup	57.63 ± 7.9	83.92 ± 5.9	93.96 ± 4.7	75.79 ± 7.8	81.88 ± 8.3	**95.38 ± 2.7**	**98.13 ± 0.5**	88.96 ± 3.6	84.45
Self-Reg	62.83 ± 1.7	95.71 ± 3.0	93.54 ± 4.2	76.29 ± 0.6	82.71 ± 0.3	79.54 ± 20.3	89.42 ± 0.9	93.79 ± 2.4	84.23
IEDGNet	45.54 ± 11.5	68.88 ± 5.4	81 ± 6.3	54.67 ± 1.1	54.71 ± 9.2	67.17 ± 1.6	65.25 ± 5.3	73.58 ± 7.7	63.85
DGRM	55.29 ± 7.1	78.42 ± 15.4	91.63 ± 4.5	81.79 ± 3.4	67.58 ± 3.4	77.29 ± 1.1	85.08 ± 4.1	85.25 ± 4.2	77.79
CCDG	**76.79 ± 2.9**	**97.25 ± 1.2**	**99.83 ± 0.1**	**93.71 ± 1.5**	**89.25 ± 1.4**	92.04 ± 3.5	94.42 ± 3.5	**94.33 ± 1.4**	**92.20**

TABLE V: Domain generalization results on Paderborn dataset (Accuracy %).

	Unseen Target Domains						
Method	I	J	K	L	M	N	Average
ERM	83.23 ± 2.0	67.50 ± 0.2	88.36 ± 0.9	99.25 ± 0.2	74.78 ± 7.2	99.40 ± 0.4	85.42
MMD	77.69 ± 3.2	69.18 ± 0.4	81.04 ± 1.4	69.53 ± 17.5	65.74 ± 5.8	82.38 ± 8.0	74.26
Deep CORAL	80.58 ± 0.9	71.30 ± 2.4	79.27 ± 1.3	81.33 ± 0.4	67.23 ± 2.0	80.58 ± 0.2	76.71
CDANN	78.77 ± 1.7	71.89 ± 5.2	86.90 ± 0.1	92.16 ± 7.8	74.16 ± 13.0	94.90 ± 3.0	83.13
Mixup	80.96 ± 4.3	71.66 ± 5.3	78.26 ± 1.5	72.52 ± 1.1	55.22 ± 3.7	74.24 ± 2.1	72.14
Self-Reg	**85.41 ± 1.5**	73.01 ± 2.5	91.79 ± 1.7	97.58 ± 0.6	68.76 ± 0.3	98.99 ± 0.5	85.92
IEDGNet	64.17 ± 8.3	63.38 ± 4.7	80.23 ± 7.5	91.83 ± 0.3	66.14 ± 3.8	92.14 ± 3.9	76.31
DGRM	76.20 ± 1.8	73.28 ± 3.7	86.56 ± 0.5	88.60 ± 10.4	76.37 ± 11.1	88.16 ± 7.5	81.53
CCDG	80.46 ± 1.7	**73.29 ± 4.2**	**94.61 ± 1.3**	**99.87 ± 0.1**	**83.08 ± 4.2**	**99.88 ± 0.1**	**88.52**



(a) CWRU Dataset (b) Paderborn Dataset

Fig. 6: Convergence of training loss for different Unseen Target Domains on both CWRU and Paderborn Datasets.

TABLE VI: Results on Paderborn dataset for single source domain generalization.

Scenario	ERM	Self-Reg	CCDG
J→I	72.54 ± 0.7	**76.46 ± 2.8**	75.20 ± 2.1
I→J	65.45 ± 5.4	66.98 ± 1.1	**67.91 ± 3.1**
K→L	42.90 ± 12	34.72 ± 5.8	**44.49 ± 11**
L→K	50.50 ± 3.2	49.23 ± 3.0	**51.81 ± 1.8**
N→M	76.58 ± 2.6	77.13 ± 3.2	**78.18 ± 1.9**
M→N	87.16 ± 5.8	91.81 ± 1.2	**93.33 ± 6.6**
Average	65.85	66.06	**68.49**

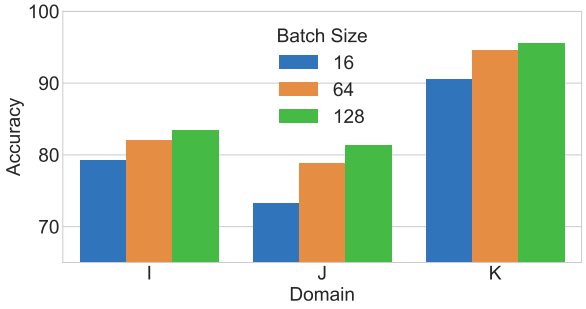
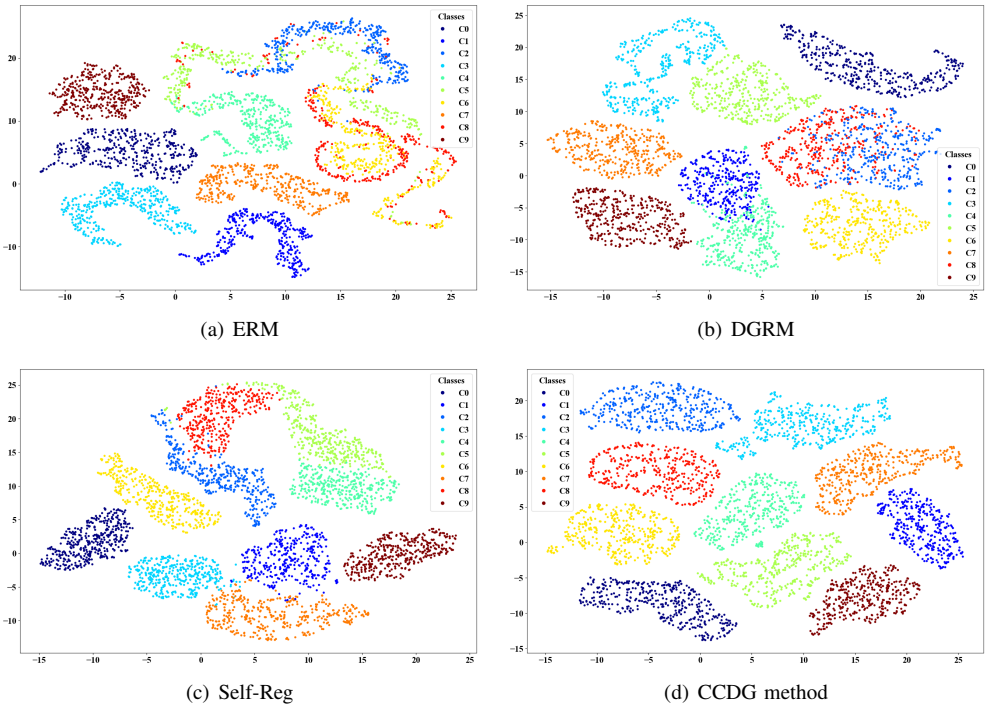


Fig. 7: Effect of batch size on CCDG performance.

3) Single Source Domain Generalization: In this experiment, we measure efficacy of our approach under extreme cases where only samples from single source domain are available for training. Table VI shows the results of the proposed approach versus the second best performing methods (i.e., ERM and Self-Reg) on randomly selected scenarios from



(a) ERM (b) DGRM

(c) Self-Reg (d) CCDG method

Fig. 8: UMAP visualization results on unseen target domain D of the CWRU dataset for different methods.

the Paderborn dataset. Overall, our CCDG performs best on 5 out of 6 cross-domain scenarios with an average improvement of 2.5%. While Self-Reg performs better than ERM, it can still under-preforms our CCDG approach. This is because that Self-Reg is mainly relying on image-specific augmentations, which may not be effective for fault diagnosis dataset. Besides Self-Reg can only align positive pairs (i.e., samples from the same class) without considering the negative pairs. Differently, our CCDG approach improves the performance via pushing away samples that belong to different classes within the same domain to have clear decision boundaries. Besides, the lower performance of our single-domain CCDG compared to multi-domain CDDG reflects the importance of having realistic representation of variability among multiple domains.

4) Generalization to multiple unseen domains: In this experiment, we measure the generalization ability of our approach on multiple unseen target domains. Particularly, we have evaluated our approach on 6 randomly selected scenarios with multiple unseen target domains. Besides, to further show the efficacy of our approach, we compared against SelfReg and ERM baselines (second and third best methods in this dataset). In our comparison, we reported the average performance on the multiple unseen domains. Clearly, our superior generalization performance persists even with more challenging scenarios of multiple unseen domains. Moreover, our CCDG approach significantly outperforms both ERM and SelfReg baselines (strongest baselines) on all the scenarios.

D. Analysis

TABLE VII: Generalization Performance of our CCDG approach on multiple unseen target domains on Paderborn Dataset (Accuracy %).

Scenario	Source Domains	Unseen Domains	ERM	SelfReg	CCDG
S1	K, L, M, N	I, J	75.68	77.95	**81.33**
S2	I, J, M, N	K, L	94.47	94.55	**97.65**
S3	I, j, K, L	M, N	85.49	82.61	**87.45**
S4	J, K, M	I, L, N	78.25	87.39	**81.75**
S5	K, L, M	I, J, N	84.05	77.64	**86.40**
S6	I, J, N	K, L, M	85.23	82.38	**91.48**

1) Training Convergence: To show the convergence of our proposed loss function, we have plotted the training convergence for each unseen target domain on both CWRU and Paderborn datasets, as shown in Fig. 6. For CWRU dataset, the convergence is smooth for all the unseen domains while the highest rate of convergence happens with unseen domain G, as shown in Fig. 6a. Similarly, for Paderborn dataset, our training loss converges smoothly while the highest convergence rate is corresponding with unseen domain K, as shown in Fig. 6b.

2) Effect of batch size: We investigate the dependency of our CCDG on the number negative samples within the mini-batch. To do so, we report the performance of three different batch sizes on three unseen domains of Paderborn datasets, as shown in Fig. 7. It can be clearly seen that our CCDG benefits from larger batch size consistently among the three different domains. Larger batch size can help in providing more negative examples per batch, which can improve the model convergence.

3) Parameter Sensitivity Analysis: In our CCDG, we have two main parameters, namely, the contrastive loss weight α and the contrasting temperature τ. To measure the model

sensitivity to each of them, we fix one parameter while changing the other and vice versa. We plot the average performance over the 8 domains of the CWRU dataset, as shown in Fig. 9. First, we fixed the temperature value τ as 0.03 while varying the weighting factor α from 0.01 to 0.9. Notably, better generalization performance can be achieved with higher weight for the contrastive loss against the cross-entropy loss, while the best performance is achieved with α equals to 0.7. However, the performance tend to degrade with α values larger than 0.7. This means that reducing the cross-entropy loss weight to lower values than 0.3 can have a negative impact on the performance. To sum up, despite that the contrastive loss can be more important than the cross entropy loss, the contribution of both losses is necessary for the best performance. Second, to measure the model dependency on the temperature parameter of the contrastive loss, we fixed α as 0.7 while varying the contrasting temperature τ from 0.01 to 0.9. It clearly shows that varying the temperature weight can change the performance from 82% at higher temperature values (i.e., 0.9) to 89% at lower temperature values (i.e., 0.03). In a nutshell, both α and τ can be of great importance to the model performance.

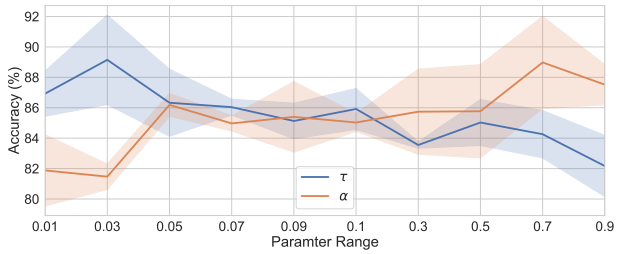


Fig. 9: Sensitivity analysis for the CCDG parameters.

4) Feature Visualization: To clearly show the effectiveness of the proposed CCDG, we visualize the learned features of our approach against three different baselines on unseen domain D of the CWRU dataset. We utilize the uniform manifold approximation and projection (UMAP) [43] to map the high dimensional features to a lower dimension as shown in Fig. 8. Fig. 8a shows the learned features of ERM, which can be hardly separable, specifically for classes C2, C5, C6, and C8. Similarly, the visualization of DGRM and Self-Reg can still have less-separable classes such as C2, C5, and C8, as shown in Fig. 8b, 8c. Differently, our CCDG approach, can separate different classes with clear boundaries, as illustrated in Fig. 8d. The visualization results of CCDG suggest that it is able to learn domain-independent class representation that can be transferable to new unseen domains.

V. Conclusion

In this work, we proposed a novel conditional contrastive domain generalization approach called CCDG that addresses a more challenging yet practical domain generalization problem for real-world fault diagnosis of rolling machinery. Specifically, we leveraged data from multiple source domains to generalize to a new unseen target domain (i.e., an unseen working condition for fault diagnosis). By evaluating on two datasets and comparing with various baseline methods, we showed that achieving the conditional invariance across the class-predictions of different source domains can significantly improve the generalization performance on unseen domains. The promising performance of our proposed CCDG method pushes towards more practical data-driven approaches that can work under challenging real-world environments.

References

[1] M. Ragab, Z. Chen, M. Wu, H. Li, C.-K. Kwoh, R. Yan, and X. Li, "Adversarial multiple-target domain adaptation for fault classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, 2021.

[2] Z. Shi, J. Chen, Y. Zi, and Z. Zhou, "A novel multitask adversarial network via redundant lifting for multicomponent intelligent fault detection under sharp speed variation," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–10, 2021.

[3] Z. Chen, G. He, J. Li, Y. Liao, K. Gryllias, and W. Li, "Domain adversarial transfer network for cross-domain fault diagnosis of rotary machinery," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 11, pp. 8702–8712, 2020.

[4] X. Li, W. Zhang, and Q. Ding, "Cross-domain fault diagnosis of rolling element bearings using deep generative neural networks," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 7, pp. 5525–5534, 2019.

[5] Y. Song, Y. Li, L. Jia, and M. Qiu, "Retraining Strategy based Domain Adaption Network for Intelligent Fault Diagnosis," *IEEE Transactions on Industrial Informatics*, vol. 3203, no. c, pp. 1–1, 2019.

[6] J. Jiao, M. Zhao, and J. Lin, "Unsupervised Adversarial Adaptation Network for Intelligent Fault Diagnosis," *IEEE Transactions on Industrial Electronics*, vol. 0046, no. c, pp. 1–1, 2019.

[7] 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, 2018.

[8] 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. 0046, no. c, pp. 1–1, 2019.

[9] H. Zheng, R. Wang, Y. Yang, Y. Li, and M. Xu, "Intelligent Fault Identification Based on Multisource Domain Generalization Towards Actual Diagnosis Scenario," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 2, pp. 1293–1304, 2020.

[10] 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.

[11] Y. Liao, R. Huang, J. Li, Z. Chen, and W. Li, "Deep Semi-supervised Domain Generalization Network for Rotary Machinery Fault Diagnosis under Variable Speed," *IEEE Transactions on Instrumentation and Measurement*, vol. 9456, no. c, pp. 1–1, 2020.

[12] T. Han, Y.-F. 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*, 2021.

[13] X. Yu, Z. Zhao, X. Zhang, C. Sun, B. Gong, R. Yan, and X. Chen, "Conditional adversarial domain adaptation with discrimination embedding for locomotive fault diagnosis," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2021.

[14] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[15] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *European Conference on Computer Vision*, 2016, pp. 443–450.

[16] D. Wang, P. Cui, and W. Zhu, "Deep asymmetric transfer network for unbalanced domain adaptation." in *AAAI*, 2018, pp. 443–450.

[17] S. Li, C. H. Liu, Q. Lin, B. Xie, Z. Ding, G. Huang, and J. Tang, "Domain conditioned adaptation network." in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 11 386–11 393.

[18] C. Chen, Z. Fu, Z. Chen, S. Jin, Z. Cheng, X. Jin, and X. sheng Hua, "Homm: Higher-order moment matching for unsupervised domain adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 3422–3429.

[19] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[20] M. Chen, S. Zhao, H. Liu, and D. Cai, "Adversarial-learned loss for domain adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 3521–3528.

[21] M. Xu, J. Zhang, B. Ni, T. Li, C. Wang, Q. Tian, and W. Zhang, "Adversarial domain adaptation with domain mixup," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 6502–6509.

[22] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in *International Conference on Machine Learning*. PMLR, 2013, pp. 10–18.

[23] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5400–5409.

[24] R. Volpi, H. Namkoong, O. Sener, J. Duchi, V. Murino, and S. Savarese, "Generalizing to unseen domains via adversarial data augmentation," *arXiv preprint arXiv:1805.12018*, 2018.

[25] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2229–2238.

[26] Y. Song, Y. Li, L. Jia, and M. Qiu, "Retraining Strategy based Domain Adaption Network for Intelligent Fault Diagnosis," *IEEE Transactions on Industrial Informatics*, vol. 3203, no. c, pp. 1–1, 2019.

[27] Y. Zhang, K. Yu, Z. Ren, and S. Zhou, "Joint domain alignment and class alignment method for cross-domain fault diagnosis of rotating machinery," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2021.

[28] W. Zhang, M. Ragab, and R. Sagarna, "Robust domain-free domain generalization with class-aware alignment," *arXiv preprint arXiv:2102.08897*, 2021.

[29] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[42] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," in *ICLR 2021: The Ninth International Conference on Learning Representations*, 2021.

[30] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[31] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 2020, pp. 776–794.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference for Learning Representations*, 2015.

[33] 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, pp. 100–131, 2015.

[34] W. Zhang, G. Peng, C. Li, Y. Chen, and Z. Zhang, "A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals," *Sensors*, vol. 17, no. 2, p. 425, 2017.

[35] C. Lessmeier, J. K. Kimotho, D. Zimmer, and W. Sextro, "Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification," in *Proceedings of the European conference of the prognostics and health management society*, 2016, pp. 05–08.

[36] V. N. Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 988–999, 1999.

[37] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5400–5409.

[38] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *European conference on computer vision*. Springer, 2016, pp. 443–450.

[39] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao, "Deep domain generalization via conditional invariant adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 624–639.

[40] L. Huang, C. Zhang, and H. Zhang, "Self-adaptive training: beyond empirical risk minimization," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[41] D. Kim, S. Park, J. Kim, and J. Lee, "Selfreg: Self-supervised contrastive regularization for domain generalization," *Proceedings of the IEEE international conference on computer vision [Accepted]*, 2021.

[43] L. McInnes, J. Healy, and J. Melville, "Umap: uniform manifold approximation and projection for dimension reduction," 2020.