Information Theoretic Learning-Enhanced Dual-Generative Adversarial Networks With Causal Representation for Robust OOD Generalization

Xiaokang Zhou, *Member, IEEE*, Xuzhe Zheng, Tian Shu, Wei Liang, *Member, IEEE*, Kevin I-Kai Wang, *Member, IEEE*, Lianyong Qi, *Senior Member, IEEE*, Shohei Shimizu, and Qun Jin, *Senior Member, IEEE*

***Abstract*— Recently, machine/deep learning techniques are achieving remarkable success in a variety of intelligent control and management systems, promising to change the future of artificial intelligence (AI) scenarios. However, they still suffer from some intractable difficulty or limitations for model training, such as the out-of-distribution (OOD) issue, in modern smart manufacturing or intelligent transportation systems (ITSs). In this study, we newly design and introduce a deep generative model framework, which seamlessly incorporates the information theoretic learning (ITL) and causal representation learning (CRL) in a dual-generative adversarial network (Dual-GAN) architecture, aiming to enhance the robust OOD generalization in modern machine learning (ML) paradigms. In particular, an ITL- and CRL-enhanced Dual-GAN (ITCRL-DGAN) model is presented, which includes an autoencoder with CRL (AE-CRL) structure to aid the dual-adversarial training with causality-inspired feature representations and a Dual-GAN structure to improve the data augmentation in both feature and data levels. Following a newly designed feature separation strategy, a causal graph is built and improved based on the information theory, which can enhance the causally related factors among the separated core features and further enrich the feature representation with the counterfactual features via interventions based on the refined causal relationships. The ITL is incorporated to improve the extraction of low-dimensional feature representations and learn the optimized causal representations based on the idea of "information flow." A dual-adversarial training mechanism is then developed, which not only enables the generator to expand the boundary of feature distribution in accordance with the optimized feature representation from AE-CRL, but also allows the discriminator to further verify and improve the quality of the augmented data for OOD generalization. Experiment and evaluation results based on an open-source dataset demonstrate the outstanding learning efficiency and classification performance of our proposed model for robust OOD generalization in modern smart applications compared with three baseline methods.**

***Index Terms*— Autoencoder (AE), causal representation learning (CRL), deep learning, generative adversarial network (GAN), information theoretic learning (ITL), out-of-distribution (OOD).**

Manuscript received 20 September 2022; revised 15 April 2023 and 24 August 2023; accepted 1 November 2023. This work was supported in part by the Grants-in-Aid for Scientific Research (C) from Japan Society for the Promotion of Science (JSPS) under Grant 23K11064, in part by the National Natural Science Foundation of China under Grant 62072171 and Grant 72091515, and in part by the 2022 and 2023 Waseda University Grants for Special Research Projects under Grant 2022R-036 and Grant 2023C-216. *(Corresponding author: Wei Liang.)*

Xiaokang Zhou and Shohei Shimizu are with the Faculty of Data Science, Shiga University, Hikone 522-8522, Japan, and also with the RIKEN Center for Advanced Intelligence Project, RIKEN, Tokyo 103-0027, Japan (e-mail: zhou@biwako.shiga-u.ac.jp; shohei-shimizu@biwako.shiga-u.ac.jp).

Xuzhe Zheng is with the School of Frontier Crossover Studies, Hunan University of Technology and Business, Changsha 410205, China (e-mail: xuzhezheng245@gmail.com).

Tian Shu is with the Computer Science Institute, Hunan University of Technology and Business, Changsha 410205, China (e-mail: wenbenst@gmail.com).

Wei Liang is with the Changsha Social Laboratory of Artificial Intelligence, Hunan University of Technology and Business, Changsha 410205, China (e-mail: weiliang@csu.edu.cn).

Kevin I-Kai Wang is with the Department of Electrical, Computer, and Software Engineering, University of Auckland, Auckland 1010, New Zealand (e-mail: kevin.wang@auckland.ac.nz).

Lianyong Qi is with the College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China (e-mail: lianyongqi@upc.edu.cn).

Qun Jin is with the Faculty of Human Sciences, Waseda University, Tokorozawa 3591192, Japan (e-mail: jin@waseda.jp).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2023.3330864.

Digital Object Identifier 10.1109/TNNLS.2023.3330864

I. Introduction

In RECENT years, machine learning (ML) techniques have demonstrated excellent performance in developing various intelligent systems and applications, including smart manufacturing, smart healthcare, and intelligent transportation systems (ITSs). Classic ML models usually exhibit reliable classification or prediction capabilities, but mainly rely on a fixed and known distribution called in-distribution (ID) [1], [2], [3]. This is one of the most fundamental assumptions for ML, which sets the training and test data as independent and identically distributed (IID). It is noticed that many ML algorithms with empirical risk minimization (ERM) rely on the IID assumption. However, studies also revealed their vulnerability to different distributions of data [4]. In fact, ML may even face unknown test distributions in real scenario applications, which may complicate the original IID scenario. In particular, the test data may have the different distributions to the training data, and collecting data from different environments and rebuilding the training set are high cost or even impossible. It, thus, becomes important to consider and generate the out-of-distribution (OOD) data samples, so as to enhance the generalization of learning models in non-IID issues.

Basically, OOD generalization problem assumes that data are extracted from a set of available training domains, and, thus, can be generalized to a larger set, including all invisible domains [5]. Most ML models aim to learn the features with certain invariance in the training data domain, and these features need to be preserved in the invisible domain. As one of the important representation learning strategies, causal representation learning (CRL), which aims to embed prior knowledge into learning processes in an unsupervised or semisupervised way, would be able to realize more reliable predictions with the learned causal representations for domain adaptation in nonstatic environments, especially when facing the situation that distributions of training and test data may be related but different, due to the environmental variation, selection bias, or time shift in distribution [6]. Krueger et al. [7] pointed out the use of invariant causality prediction to achieve the generalizability in across domains, in which the causal mechanism during data generation process might be the same across different domains, while interventions can be varied by domains. Wang et al. [8] introduced CRL for OOD recommendations and proposed to transfer user features as intervention and object-oriented recommendations as postintervention inference of interaction probabilities. They further integrated the variational autoencoder (VAE) into causal modeling, which could leverage the encoder to infer the unobserved user features from historical interactions. Although the existing research works have shown advance in terms of alleviation of cross-domain OOD generalization problem, questions of which features need to be retained, and whether uncontrollable factors will be introduced again during the decoding of features for data generation, are still open issues, calling for new design to capture latent data correlations and resist distribution shifts with the newly learned feature representations.

Currently, researchers are paying more efforts to improve the robustness and generalization of ML models, e.g., by generating OOD samples that may appear near the ID boundary [9], and can be fed into the classification model to improve knowledge learning. Although several representative generative models, such as AE and generative adversarial network (GAN), have been widely used, they are prone to bring feature selection bias during the data generation process [10]. Generating data without restrictions will not only fail to make OOD samples cover the ID boundary, but also result in poor quality of the generated samples, which is not conducive to the generalization of learning models. Typically, addressing this issue can be considered from an information theoretic perspective, where the mutual information can be incorporated to effectively constrain the generated data. It is indicated that ML techniques are inseparable from the extensive exploration of information theory. Considering the bias problem in information theoretic learning (ITL) can essentially help finding the sources of bias [11], [12], which points out a significant way to find the optimal balance between accuracy and complexity using information theory, toward the robust OOD generalization.

In this study, both the ITL and CRL are integrated into a generative model framework to realize a dual-adversarial training for the robust OOD generalization. Specifically, an ITL- and CRL-enhanced Dual-GAN (ITCRL-DGAN) model is introduced, in which an AE with CRL (AE-CRL) structure is constructed to assist the dual-adversarial training in causality-inspired feature representation, while a Dual-GAN structure is built to improve the data augmentation in both feature and data levels. Based on a newly designed feature separation strategy, a causal graph is improved to enhance the causally related factors in core features, so as to generate the enriched feature representations according to the latent causal relationships with better learning efficiency. The ITL is incorporated to improve the extraction of low-dimensional feature representations and further learn the optimized causal representations among the separated core features using the idea of "information flow." A dual-adversarial training mechanism for intelligent data augmentation is finally developed to tackle the OOD problem. Main contributions of this study can be summarized as follows.

1) An integrated deep generative model framework is proposed, in which four loss functions, including an evidence lower bound loss to facilitate the modeling of posterior distributions of latent variables with better distribution description, a causal structure loss to refine relationships among causal representation variables based on maximum mean discrepancy (MMD), an information flow loss to improve the quality of causal representation by maximizing relations between labels and the core features, and a reconstruction loss to reduce the deviation between the generated data and real data, are defined and seamlessly work together to realize the causality-inspired feature representation in supporting the dual-adversarial training from both feature and data levels.
2) A CRL-enhanced AE structure is designed, in which a feature separation strategy is newly introduced to separate the so-called causal features that may be more relevant to the class labels; then, an information theory-enhanced causal graph is built to facilitate the finding of more causally related factors among core features and, thus, can enrich the feature representation with counterfactual features via interventions based on the refined causal relationships.
3) A Dual-GAN structure is devised, in which the first adversarial network is constructed to maximally retain the key information in the original data and optimize the latent features of GAN to expand the boundary of feature distribution, while the second adversarial network is constructed to further verify the quality of the augmented data and make it fit the real data distribution as much as possible.

The rest of this article is addressed as follows. The state-of-the-art techniques related to this study are reviewed in Section II. The basic model architecture is introduced in Section III. We explain the core mechanisms to realize the AE-CRL and Dual-GAN for robust OOD generalization in Section IV. Experiment and evaluation results are demonstrated and discussed in Section V. We conclude this study and give promising perspectives regarding future research in Section VI.

II. Related Works

In this section, the existing techniques and methods related to ITL, CRL, and GAN are reviewed and summarized, respectively.

A. Information Theoretic Learning

Recently, in addition to optimizing more cost-effective ML schemes, connections among statistics, information theory, and ML are received remarkable attentions and are extensively explored, aiming to capture more informative but low-dimensional features from large-scale data. Honkela and Valpola [13] discussed the variational learning and bits-back coding from an information theoretic view. They reviewed the general idea of statistical Bayesian framework with information theoretic minimum-description-length principle and indicated the benefit by combining the views of Bayesian statistics and information theory for model selection and parameter optimization in learning tasks. Tan et al. [14] introduced an information theoretic framework to examine the information-based learning capacity using a so-called interactively and integratively connected deep recurrent neural network, which could theoretically analyze the model's learning behavior, and further prove the capacity of capturing the multiscale dependence of spatiotemporal data in predictive spatiotemporal analytics tasks. Goldfeld and Polyanskiy [15] focused on the information bottleneck theory as one of the specific information theoretic paradigms for deep learning analysis. They discussed operational interpretations, especially the Gaussian information bottleneck setup, and pointed out advances in deep neural network-based feature representation when adopting mutual information in information bottleneck framework. Li et al. [16] presented an ITL-based method for diffusion distributed estimation, in which they employed the error entropy criterion to improve the cost function based on two algorithms named diffusion minimum error entropy algorithm and diffusion entropy bound minimization algorithm. Boscolo et al. [17] introduced an information theoretic exploratory framework based on the concept of statistical coinformation, which incorporated a so-called moment-based approximation of coinformation measure, to estimate the high-dimensional multivariate probability density functions based on a set of conditioning variables. Considering the uncertainty issue in discriminative data representation tasks, Deng et al. [18] built a robust information theoretic framework for feature transformation, in which a discriminative classifier was developed to maximize the mutual information with a transformation function in the latent space. They also discussed the three different implementations in terms of linear subspace embedding, deep transformation, and structured sparse learning. Meyer et al. [19] designed an information theoretic feature selection criterion for microarray data filtering, which defined two properties to realize the maximization in both lower bound and upper bound of the mutual information of a subset. They further implemented it using a backward elimination combined with a sequential replacement strategy to deal with the densest subgraph problem. Zhang et al. [20] developed a reinforcement learning model with an asynchronous advantage actor–critic algorithm, which utilized the mutual information to measure the user-level and aggregate-level privacy leakage, respectively, aiming to optimize the centralized privacy-preserving aggregate mobility from the information theoretic perspective. To deal with the generally defined feature transformation learning problem, Özdenizci and Erdogmus [21] proposed a maximum mutual information linear transformation method based on ITL. They further built a graphical model-based hierarchical multiclass decoding framework to improve the multiclass classification performance in brain–computer interface tasks. Xu et al. [22] developed two algorithms using diffusion cooperative strategy, to deal with the linear and nonlinear multilabel classification, respectively. They introduced an information theoretic measure to optimize the cost function based on a label correlation term defined on some anchor data with a distributed matrix completion scheme.

B. Causal Representation Learning

In current years, the concept of CRL has drawn more and more attentions, which could help to learn more about structural knowledge or causal relations among variables, exploring the robustness of deep neural network architectures in modern ML research. Sun et al. [23] integrated the causal graph into the reinforcement learning framework and proposed a relation transfer method to infer the target domain model based on the summarized causal relations from source task variables using the prior causal knowledge. Wang et al. [24] constructed a causal graph to analyze causal relationships in terms of the unbiased user intents with specific semantics in recommendation tasks. They developed a causal intervention mechanism, to refine the semantic-aware representations by eliminating the confounding bias and disentangling users' true intents in specific item context. Guo et al. [25] focused on the causal contextual entropy prediction in image compression and built a causal context model to better utilize the so-called channelwise relationships, which might capture the highly informative adjacent contexts. They defined the separate entropy coding in a causal global prediction model with a group-separated attention module, to pursue more accurate predictions of undecoded points. Yang et al. [26] designed a causal AE with a causal structure learning module, to learn the causal representation in a single source domain. They employed the CRL to generate causal representations and task-irrelevant representations from low-dimensional representations, in order to solve the robust domain adaptation problem. Rao et al. [27] considered the causal model for decision-making in the medicine field. They built a transformer-based model to analyze causal associations from electronic health records, and their experiment results indicated the benefit of using causal inference for more accurate estimations. Zhang et al. [28] presented a domain adaptation model for weather condition discovery, in which a dynamic item extraction strategy was proposed to guarantee the representativeness of each prototype of object features in a confounder dictionary, and a causal intervention reasoning module was constructed to improve the invariant feature representation. Xiang and Truong [29] addressed an expressive causal interaction function for the nonimpeding noisy-AND

tree and formulated a concise structure representation based on two structural acquisition methods, which could be used in both elicitation-based and ML-based acquisitions in Bayesian networks. Wang et al. [30] introduced a so-called reinforced causal explainer based on reinforcement learning techniques, to improve the explainability in graph neural network models. Their method enhanced the importance of causal effects and dependencies on each edge, to build an explanatory subgraph and explain the prediction result as a sequential decision process. Sahoh et al. [31] built a graph-based structural causal model to improve the qualitative knowledge representation in internet of things (IoT) systems when detecting fault events from sensory signals, in which a causal discovery algorithm was developed to determine cause-and-effect in ML models toward human-like perception from observational data. Zhu et al. [32] applied causal inference into the deep neural networks, aiming to enhance the causal effect in a supervised learning scheme. They constructed a direct learning framework with a shared representation layer and a propensity prediction regularizer to improve the treatment effect estimation in multitask learning.

C. Generative Adversarial Networks

It is noticed that GAN has become a promising unsupervised learning technique in dealing with a variety of intelligent tasks, ranging from data augmentation, anomaly detection, to dialog generation, and image translation. Huang et al. [33] presented an attentive GAN in a self-supervised framework, to deal with the video anomaly detection in a unsupervised way. They designed a self-attentive predictor to extract long-term dependences, and a vanilla discriminator and a self-supervised discriminator to improve the true–false discrimination and self-supervised rotation detection. Cai et al. [34] introduced a dual-attentional GAN for the task of image synthesis from text description, in which a textual attention module was used to model the interaction between vision and text, while a visual attention module was used to learn the internal vision representation from channel and spatial axes. Wang et al. [35] proposed a conditional GAN with a two-step data augmentation strategy in aircraft design. The GAN-based neural network was employed to augment the original airfoil dataset with different properties and further improve the prediction of pressure coefficient curve. Choi et al. [36] addressed a GAN-based defense model in two transformation steps. They defined a joint loss function to optimize parameters in the generator, which could better learn the vulnerability of the target neural network model based on the generated adversarial examples. Zheng et al. [37] employed GAN into the attribute augmented network embedding task and constructed an attribute augmented network to extract the node attribute and structural feature by capturing the latent distribution of data in low-dimensional representations. They used the teacher forcing scheme with a pretraining algorithm to improve the training efficiency and stability during the generative model implementation process. Shao et al. [38] incorporated GAN into the capsule network for the better utilization of view angle invariance and rotation equivariance in image-to-image translation issues. They applied two capsule networks as the discriminators, in which the routing algorithm was optimized with the combination of margin loss and original adversarial loss in a multiagent competition mechanism. Jiang et al. [39] developed a weakly supervised discriminative learning scheme based on a spectral constrained GAN for hyperspectral anomaly detection, in which a so-called orthogonal projection divergence spectral constraint based on Kullback–Leibler (KL) divergence was proposed to enhance the discrimination capacity in an end-to-end architecture. To improve the training process of GAN, Franci and Grammatico [40] introduced a stochastic relaxed forward–backward algorithm and its variant with averaging, which only included one evaluation of the pseudogradient mapping for each iteration. They proved that the convergence could be guaranteed even only a few samples are available. Considering the searching process as a bilevel minimax problem, Tian et al. [41] tried to improve the learning performance in GAN based on an automated search framework. They trained GAN with a searched architecture from the training dataset, while the search process was guided by a differentiable evaluation metric to obtain the optimized network parameters of both generator and discriminator.

III. Framework of ITCRL-DGAN

In this section, we first address the problem definition in terms of the OOD issue and then introduce the network architecture of our proposed ITCRL-DGAN model, which basically consists of an AE-CRL structure and a Dual-GAN structure.

A. Problem Definition

Traditional deep learning models heavily depend on the IID assumption, which means the data distribution in model training is required to be the same as in model testing, $P_{\text{train}}(X, Y) = P_{\text{test}}(X, Y)$. However, it is noticed that the phenomenon of distribution shift may happen in many real application scenarios, which means distributions of the training data and the test data may become different, and the test data distribution may even become unknown due to the spatial or temporal data evaluation or sample selection bias in data collection [5]. To handle such OOD problem in reality, the existing studies tried to use data augmentation techniques to generate OOD samples for model training enhancement [42], [43], but it still needs to face the inevitable distribution bias when training the data generator.

Accordingly, both ITL and CRL are involved in our framework to constrain the distribution bias and further generate the augmented data to enhance the OOD generalization issue. Specifically, given a generative model $G(x, y, \mathcal{N}; \theta_G)$, where $\mathcal{N}$ is the random noise and θ_G is the parameter of the generation model, the generator is designed to reduce the distribution bias and be able to augment the original training data D, where $D = \{(x_i, y_i)|x_i \in X, y_i \in Y\}$, X is the set of original samples, Y is the corresponding set of the labels, and the distribution of D is $P(X, Y)$ for data training. Based on the augmented data $D_o = \{(x_j, y_j)|x_j \in X_o, y_j \in Y\}$, where X_o is the set of generated OOD samples with distribution $P_o(X_o, Y)$, the goal of the framework is to optimize the generator G and discriminator D during a dual-adversarial training process,

so as to enhance the data augmentation with generated OOD samples.

In particular, the OOD generalization problem based on data augmentation can be summarized as follows. Given a labeled dataset D with the corresponding distribution $P = (X, Y)$, our generative model $G(x, y, \mathcal{N}; \theta_G)$ aims to generate x_j in D_o using the knowledge learned from D, while the generated data keeps the same label in Y. It can effectively enhance the diversity and reality of available data in model training and finally improve the prediction of downstream classifier h, which can be formulated as follows:

$$\min_h E_{X,y}\big[\mathcal{L}(h(X), y)\big] + E_{X_o,y}\big[\mathcal{L}(h(X_o), y)\big]. \tag{1}$$

B. Basic Framework

In general, we design and propose a Dual-GAN structure, which incorporates ITL and CRL to solve the OOD generalization problem. As shown in Fig. 1, the basic architecture includes two important parts: an AE-CRL module and a Dual-GAN module. The AE-CRL is designed to assist the Dual-GAN in causality-inspired feature representation, in which an information theory-enhanced causal model is constructed and integrated into the AE structure to better learn the latent causal relationships among the separated core features; then, generate the enriched feature representations with counterfactual features. In the Dual-GAN, the first adversarial network is designed to realize the data generation in the feature level, which aims to maximally retain the key information in the original data, and optimize the latent features of GAN based on the intervention via an improved causal graph, making the generator be able to expand the boundary of feature distribution for the OOD generalization. The second adversarial network is devised to realize the data augmentation in the data level, which aims to make the generated data fit the real data distribution as much as possible, and further verify the quality of the augmented OOD data to be more realistic.

Specifically, data x are first input into the AE-CRL; then, the hidden variable Z is compressed and obtained after the encoder based on ITL, to learn a low-dimensional feature representation while avoiding the loss of key information. The variational inference is used to calculate the evidence lower bound loss L_{ELBO}, which could benefit the modeling of posterior distributions of latent variables and result in better data distribution. Following a feature separation strategy that newly separates the obtained Z into the so-called causal feature and confounder feature, an improved causal graph with the direct acyclic graph (DAG) structure is constructed, which is used to enhance the causally related factors among the separated core features, based on two causal losses. In particular, a causal structure loss L_{MMD} is defined to refine the causal relationships among causal representation variables based on the MMD method, and an information flow loss L_{IF} is defined to improve the causal representation by maximizing relations between labels and the core features. A optimized feature representation Z' is then generated based on counterfactual features via intervention. Differing from [26], which also introduced CRL into the AE structure, our design initializes the causal graph by the Markov blanket and adds multilayer perceptron (MLP) to better learn the nonlinear relationships among the core features. The idea of "information flow" is improved and newly applied to further enhance the quality of causal representations, rather than simply using the structural loss for feature optimization. In addition, we define a reconstruction loss L_{rec} to control the reduction of the deviation between the generated data x' and the real data x by the decoder. Finally, the optimized Z' is used as the input to train the first discriminator.

The design of the Dual-GAN structure basically consists of one generator network, one decoder, and two discriminator networks, to form a dual-adversarial training process. In the first adversarial network, we input the random noise $\mathcal{N}$ and label y_i, $y_i \in \{y_1, y_2, \ldots, y_{|Y|}\}$, into the generator G and generate the hidden variable Z_G in the GAN. The first adversarial training is then conducted between Z' and Z_G, which could make the generated features Z_G continuously approach the optimized Z'. It is noted that, inspired by Yang et al. [44], the intervention is employed and added during the CRL so as to enrich the feature representation in Z' with counterfactual features, aiming at expanding the boundary of feature distribution in Z_G. In the second adversarial network, the augmented data sample x'' is generated from Z_G via the decoder on the basis of the first adversarial training. Considering the generated x' usually in a relatively lower quality with less diversity, while the enhancement of x' could also benefit the improvement of Z' and further influence Z_G and x'' in a collaborative way, the second adversarial training is then conducted among x, x', and x'', so as to augment more reliable samples for the robust OOD generalization.

IV. ITL- and CRL-Enhanced Dual-Adversarial Training for OOD Generalization

In this section, following the introduction on the AE-CRL structure, we explain how to construct the improved causal graph based on a feature separation strategy, and discuss the implementation for the dual-adversarial training. An intelligent data augmentation algorithm based on the proposed ITCRL-DGAN for OOD generalization is developed finally.

A. AE With CRL

In general, we introduce the CRL into the AE structure, so that the causal graph can be trained based on the compressed feature representation Z; then, the learned causal relationships can be incorporated to generate the refined feature representation Z', which is also used as the input to the decoder to generate x'.

Basically, the AE-CRL first inputs x into the compression network to get the hidden variable Z, which can be simply expressed as follows:

$$Z = \text{Encoder}(x). \tag{2}$$

The variational inference approach is employed to model the distribution of Z, which approximates an incalculable posterior distribution $p_\theta(Z|x, y)$ by constructing a new distribution $q_\varphi(Z|x, y)$. The difference between the two distributions is

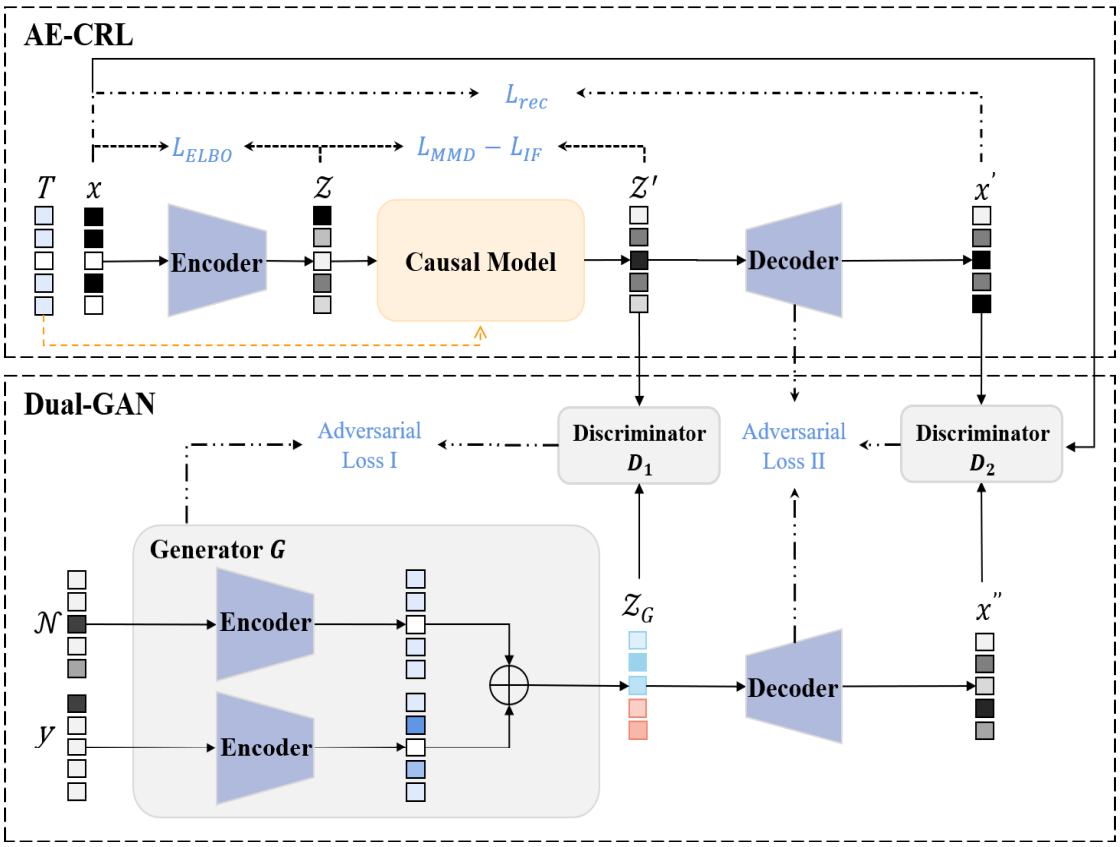


Fig. 1. Framework architecture of ITCRL-DGAN.

quantified by the KL divergence, and the specific calculation process can be formulated as follows:

$$\begin{aligned}&\mathrm{KL}[q_\varphi(Z|x,y)||p_\theta(Z|x,y)]\\&\quad=E_{z\sim q_\varphi}[\log q_\varphi(Z|x,y)-\log p_\theta(Z|x,y)]\end{aligned}\tag{3}$$

where $p_\theta(Z|x,y)]$ can be deconstructed by a Bayesian approach; thus, (1) can be updated as a new expression

$$\begin{aligned}&\mathrm{KL}[q_\varphi(Z|x,y)||p_\theta(Z|x,y)]\\&\quad=E_{z\sim q_\varphi}[\log q_\varphi(Z|x,y)-\log p_\theta(x|Z,y)\\&\qquad-\log p_\theta(Z|y)]+\log p_\theta(x|y)\end{aligned}\tag{4}$$

where $\log p_\theta(x|y)$ denotes the log likelihood of the sample and $E_{z\sim q_\varphi}[\log p_\theta(x|Z,y)]$ denotes the expectation of reconstructing Z into the value of likelihood function of x.

We hope that this expectation is as large as possible, so that the result of the hidden variable can become better. Thus, (4) can be expanded to further obtain a lower bound for the distribution, which can be fixed as follows:

$$\begin{aligned}&\mathrm{KL}[q_\varphi(Z|x,y)||p_\theta(Z|x,y)]\\&\quad=-E_{z\sim q_\varphi}[\log p_\theta(x|Z,y)]\\&\qquad+\mathrm{KL}[q_\varphi(Z|x,y)||p_\theta(Z|y)]+\log p_\theta(x|y)\end{aligned}\tag{5}$$

where $-E_{z\sim q_\varphi}[\log p_\theta(x|Z,y)]+\mathrm{KL}[q_\varphi(Z|x,y)||p_\theta(Z|y)]$ denotes the evidence lower bound and can be used to quantified the evidence lower bound loss L_{ELBO}.

Assuming that q_φ and p_θ are codistributed, then $\mathrm{KL}[q_\varphi(Z|x,y)||p_\theta(Z|x,y)]=0$; thus, it can be further improved as follows:

$$\begin{aligned}\log p_\theta(x|y)&=E_{z\sim q_\varphi}[\log p_\theta(x|Z,y)]\\&\quad-\mathrm{KL}[q_\varphi(Z|x,y)||p_\theta(Z|y)].\end{aligned}\tag{6}$$

Accordingly, the target of the variational inference is to maximize the likelihood function $p_\theta(x|y)$, which equals to maximize the evidence lower bound. It is noted that, following this way, the reconstruction process of unobservable variables can be efficiently improved.

The extracted hidden variable Z is then input into the causal graph structure, aiming to obtain the optimized feature representation Z' based on the refined causal relationships among core features. Finally, our AE-CRL puts Z' into the decoder and obtains the reconstructed sample x'. It is noted that the dimension of x' should be the same as that of x. The general expression of x' can be simply described as follows:

$$x'=\mathrm{Decoder}(Z').\tag{7}$$

The mean square error is used to quantified the reconstruction loss L_{rec} between input x and output x', which can be defined as follows:

$$L_{\mathrm{rec}}=||x-x'||^2.\tag{8}$$

B. Causal Model Constructing

As shown in Fig. 2, to enhance the learning effect in terms of finding more causally related factors among the extracted features when constructing causal graph, we design to only input those core features that may be more relevant to the class

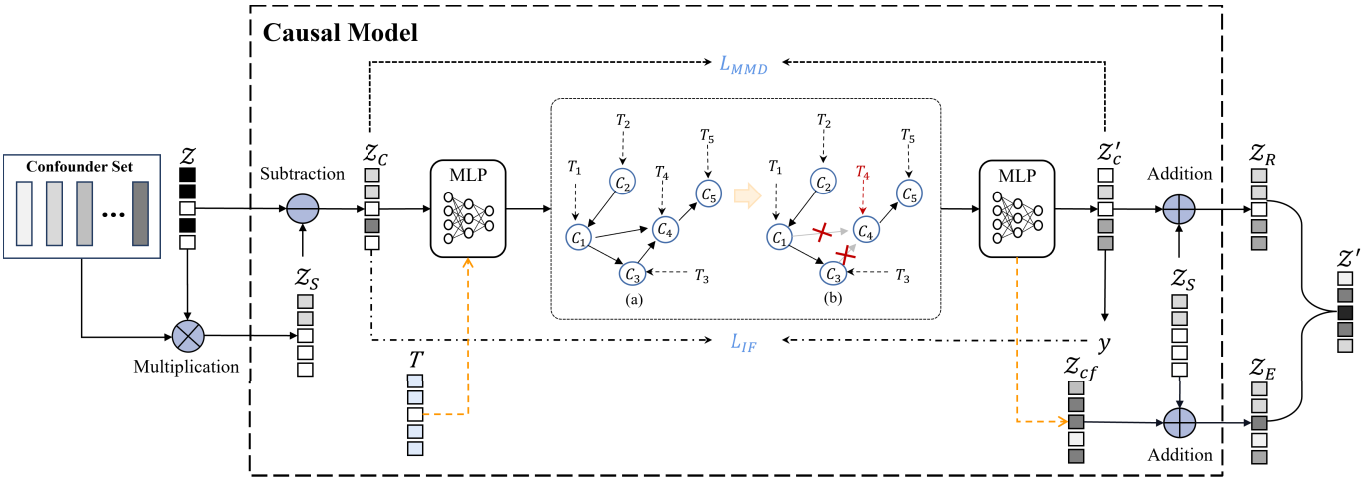


Fig. 2. Causal model structure based on feature separation.

labels into the causal graph model, which means the extracted Z needs to be defined and separated into two categories: the causal feature Z_C and the confounder feature Z_S.

Given S as a confounder set $\{s_i\}_{i=1}^{|Y|}$, where $|Y|$ is the number of labels in dataset, s_i is defined as a vector to describe the confounder features in terms of y_i in feature space. For example, considering a image classification task, the confounder features can be recognized as the background features, which may not very related but confuse the identification of the core object in a given picture.

Similar to [45], considering both the causal features and confounder features can be represented by manifolds, Z_S can be modeled as follows:

$$Z_S = f(Z, S) = \sum_{i=1}^{|Y|} P(s_i|Z)s_i \tag{9}$$

where $P(s_i|Z)$ indicates the probability that Z belongs to the confounder features of s_i.

Inspired by Hinton et al. [46], which considers a classifier based on (9) as the distilled knowledge for feature extraction, we design to separate the confounder features from Z by a subtraction operation. The separated causal feature Z_C can then be obtained and described as follows:

$$Z_C = Z - Z_S \tag{10}$$

where "$-$" denotes the operation of matrix subtraction.

We go further to discuss the construction of the causal graph, which may causally describe a hidden relationship between the parent and child nodes, using a DAG structure. In particular, we employ the Markov blanket to accelerate the convergence during the initialization process, and incorporate the MLP to enhance the learning of nonlinear relationships among features, which can be expressed as follows:

$$C = \text{MLP}(Z_C) \tag{11}$$

where $C = \{C_1, C_2, \ldots, C_n\}$ denotes the new nonlinear representations of features and MLP$(*)$ indicates a two-layer neural network.

Interventions are then involved to refine the more exact causal relationships among the input causal features. $T = \{T_1, T_2, \ldots, T_n\}$ represents a collection of interventions that work on different feature nodes, respectively. For example, as shown in Fig. 2, when an intervention T_4 is added to C_4, it will cut off all edges pointed to C_4 (i.e., C_1 and C_3); thus, C_4 is not affected by other nodes and can be viewed as a constant. Following this way, the intervention can break down the spurious correlation by controlling C_4 as a constant and then analyze the causality among other features.

Since the causal graph is implemented based on DAG, it can be described as an adjacency matrix A. Then, the improved nonlinear representation C' based on the refined causal relationships can be obtained and described as follows:

$$C' = A^T C + \epsilon \tag{12}$$

where $C' = \{C'_1, \ldots, C'_4, \ldots, C'_n\}$ and ϵ denotes the additive noise.

The obtained C' is then used as the input to another MLP to generate the reconstructed feature representation Z'_C, which can be expressed as follows:

$$Z'_C = \text{MLP}(C'). \tag{13}$$

Accordingly, the optimized causal features Z'_C from the improved causal graph can be fused again with the separated confounder features to generate the reconstructed feature representation Z_R, which can be expressed as follows:

$$Z_R = Z_S + Z'_C \tag{14}$$

where "$+$" denotes the operation of matrix addition.

To better learn the casual structure, we define a causal structure loss L_{MMD} to calculate the error between Z_C and Z'_C based on MMD, which can be quantified as follows:

$$L_{\text{MMD}} = \text{MMD}_k\left(Z_C, Z'_C\right) + \lambda|A| \tag{15}$$

where $|A|$ is defined to denote the number of edges in A. $\text{MMD}_k = (1/n^2)\sum_{i,j=1}^{n} k(x_i, x_j) + (1/n^2)\sum_{i,j=1}^{n} k(x'_i, x'_j) - (2/n^2)\sum_{i,j=1}^{n} k(x_i, x'_j)$, and kernel k is usually taken as the Gaussian kernel, $k(x, x') = \exp(-\gamma\|x - x'\|_2^2)$.

Furthermore, to more effectively mine the causal relationships and explore how the causal variables affect the predicted labels, the information theory, especially the idea of "information flow" [47], is improved and applied to measure the causal information among those core features, which can

be viewed as the causal counterpart of the mutual information $I(Z_C; y)$, and is defined as $I(Z_C \rightarrow y)$. It is noted that $I(Z_C \rightarrow y)$ is not equal to $I(Z_C; y)$ and can be formulated as follows:

$$I(Z_C \rightarrow y) = \int P(Z_C) \sum_y P(y|\text{do}(Z_C)) \log \frac{P(y|\text{do}(Z_C))}{\int P(y|\text{do}(Z_C)) dZ_C} dZ_C \tag{16}$$

where $\text{do}(Z_C)$ indicates the intervention and the classic backdoor adjustment is used to calculate $P(y|\text{do}(Z_C))$

$$P(y|\text{do}(Z_C)) = \sum_Z P(y|Z_C, Z) P(Z). \tag{17}$$

Based on the above calculations, we may keep Z_C fixed and estimate the causal effect of Z_C on y, so as to maximally retain the causal representation related to the label information and refine the causal associations among the extracted core features.

Finally, by using $I(Z_C \rightarrow y)$ to quantify the information flow loss L_{IF}, a total causal loss L_c based on the constructed causal graph can be expressed as follows:

$$L_C = L_{\text{MMD}} - L_{\text{IF}}. \tag{18}$$

We then employ the counterfactual estimation to generate the counterfactual features Z_{cf} based on the "abduction-action-prediction" process [48], toward the alleviation of OOD issue. Following the "abduction" step implemented by the improved causal graph, the counterfactual interventions, which will make a certain causal variables as the constant, are used as actions in the "action" step, so as to generate the counterfactual features. Then, the generated counterfactual features will be utilized as the input in the "prediction" step. If the predicted result is consistent with the original label, it indicates that this result falls within the feature space of that label; thus, this generated counterfactual feature can be kept, while those features whose predicted labels are changed will be discarded. The enriched feature representation with the generated counterfactual features can be expressed as follows:

$$Z_E = Z_S + Z_{\text{cf}} \tag{19}$$

where "+" denotes the operation of matrix addition.

Finally, the optimized feature representation based on AE-CRL can be obtained and described as follows:

$$Z' = Z_R \cup Z_E. \tag{20}$$

C. Dual-Adversarial Training

As shown in Fig. 1, our proposed Dual-GAN structure is mainly composed of a redesigned generator with two discriminators. Similar to a conventional conditional GAN, the input includes a random noise $\mathcal{N}$ with the label y. The generator consists of two encoders, and we encode noise $\mathcal{N}$ by one encoder, which can be simply expressed as follows:

$$\mathcal{N}_e = \text{Encoder}(\mathcal{N}). \tag{21}$$

Label y needs to be re-encoded to obtain the label representation, so that it can be matrix-summed with the updated random noise, which is simply expressed as follows:

$$\begin{aligned} y_e &= \text{Encoder}(y) \\ Z_G &= f(\mathcal{N}_e + y_e). \end{aligned} \tag{22}$$

During the first and second adversarial training, the generator $G(*)$ needs to make the generated feature or data as realistic as possible to fool the discriminator, for example, making the gap between $G_1(\mathcal{N}, y)$ and Z' as small as possible. The discriminator $D(*)$ outputs a scalar that discriminates whether the input is the real data or from the generated data. The $G(*)$ and $D(*)$ work together to learn the features, and maximize the probability of $D(*)$ in making mistakes that determine the input is from the original data rather than $G(*)$, according to the following adversarial game:

$$\min_G \max_D E[D(*)] + E[1 - D(G(*))] \tag{23}$$

where $E(*)$ is the expectation. On the one hand, we expect $E[1 - D(G(*))]$ to be as small as possible; on the other hand, we expect $E[D(*)]$ to be as large as possible.

In the first adversarial training, we employ Z' generated by AE-CRL to train the first discriminator D_1, and input the generated Z_G to it. Through the backpropagation, the discriminator can be trained adversarially with the generator. The loss function L_G of the generator can be expressed as follows:

$$L_G = \text{Entropy}(D_1(G(\mathcal{N}, y), 1)). \tag{24}$$

While training the AE-CRL, Z' and Z_G are input into D_1 for the first adversarial training. The loss function L_{D_1} of D_1 can be expressed as follows:

$$L_{D_1} = \frac{1}{2}(\text{Entropy}(D_1(Z'), 1) + \text{Entropy}(D_1(G(\mathcal{N}, y)), 0)) \tag{25}$$

where $\text{Entropy}(*)$ is the binary cross entropy.

It is noted that by adding interventions through the constructed causal graph, the enriched causal representation can be added to the original compressed hidden variable Z based on the refined causal relationships, as the new representation Z'. Thus, it can be used to achieve the goal of data generation in the feature level during the first adversarial training, enabling the generator to generate data with more causality-enhanced features against the OOD issue.

In second adversarial training, Z_G generated in the first adversarial training is used to obtained the augmented data x'' through the decoder, which can be simply expressed as follows:

$$x'' = \text{Decoder}(Z_G). \tag{26}$$

The augmented x'' based on Z_G and x' from AE-CRL is then adversarially trained with the original x. The loss functions, L_{Dec_1} of the decoder in the AE-CRL structure and

Algorithm 1 Data Augmentation for OOD Generalization

Input: An original dataset D
Output: A generated dataset D_o

1: Initialize the model
2: Initialize iteration T, batch size b, steps s, intervention I
3: **for** t in T **do**
4: **for** s steps **do**
5: Sample mini-batch of b original data $\{x_1^s, x_2^s, \ldots, x_b^s\}$
6: Sample mini-batch of b noise data $\{n_1^s, n_2^s, \ldots, n_b^s\}$
7: Sample mini-batch of b class label $\{y_1^s, y_2^s, \ldots, y_b^s\}$
8: Construct $\{z_1^s, z_2^s, \ldots, z_b^s\}$ as the input of the causal model by Eq. (2)
9: Generate $\{z_1^{s\prime}, z_2^{s\prime}, \ldots, z_b^{s\prime}\}$ by Eq. (10)-(14)
10: Generate $\{z_{G_1}^s, z_{G_2}^s, \ldots, z_{G_b}^s\}$ by Eq. (21)-(22)
11: Generate $\{x_1^{s\prime}, x_2^{s\prime}, \ldots, x_b^{s\prime}\}$ by Eq. (7)
12: Generate $\{x_1^{s\prime\prime}, x_2^{s\prime\prime}, \ldots, x_b^{s\prime\prime}\}$ by Eq. (26)
13: Update AE-CRL by Eq. (18) and Eq. (27)
14: Update generator by Eq. (24)
15: Update decoder in Dual-GAN by Eq. (28)
16: Update discriminator D_1 by Eq. (25)
17: Update discriminator D_2 by Eq. (29)
18: **end for**
19: **end for**
20: Sample m noise data $\{n_1, n_2, \ldots, n_m\}$
21: Sample m class label $\{c_1, c_2, \ldots, c_m\}$
22: Generate $\{z_{G_1}, z_{G_2}, \ldots, z_{G_m}\}$ by Eq. (21)-(22)
23: Generate the augmented data $X'' = \{x_1'', x_2'', \ldots, x_m''\}$ by Eq. (26)
24: $D_o \longleftarrow X''$
25: **return** D_o

L_{Dec_2} of the decoder in the Dual-GAN structure, for the second adversarial training, can be described, respectively, as follows:

$$L_{\text{Dec}_1} = \text{Entropy}(D_2(\text{Decoder}(Z'), 1)) \tag{27}$$

$$L_{\text{Dec}_2} = \text{Entropy}(D_2(\text{Decoder}(Z_G), 1)). \tag{28}$$

The generator and AE-CRL aim to make the generated data fool the discriminator as much as possible, which means the generated x' and x'' need to be as closer as possible to the original sample x.

After putting x, x', and x'' into the second discriminator D_2 for training, the loss function L_{D_2} of D_2 can be expressed as follows:

$$L_{D_2} = \frac{1}{3}(\alpha\text{Entropy}(x, 1) + \text{Entropy}(x', 0) + \text{Entropy}(x'', 0)) \tag{29}$$

where α denotes the hyperparameter.

Following this way, based on our designed Dual-GAN structure, the first adversarial training is performed between Z_G and Z' in the feature level, which allows the generator to expand the boundary of feature distribution in accordance with the optimized feature representation from AE-CRL. The second adversarial training is conducted among x, x', and x'' in the data level, which not only enables the augmented data to be as realistic as possible, but also can effectively alleviate the OOD issue.

D. Data Augmentation for OOD Generalization

Based on the discussion above, the loss function of AE-CRL can be optimized according to the evidence lower bound loss L_{ELBO}, causal loss L_C, and reconstruction loss L_{rec}, which can be expressed as follows:

$$L_{\text{AE-CRL}} = -L_{\text{ELBO}} + \beta_1 L_C + \beta_2 L_{\text{rec}} \tag{30}$$

where β_1 and β_2 are the hyperparameters.

Considering those adversarial losses discussed in Dual-GAN, finally, the overall loss function for our ITCRL-DGAN model can be expressed as follows:

$$L_{\text{ITCRL-DGAN}} = L_{\text{AE-CRL}} + \lambda_1 L_G + \lambda_2 L_{D_1} + \lambda_3 L_{\text{Dec}_1} + \lambda_4 L_{\text{Dec}_2} + \lambda_5 L_{D_2} \tag{31}$$

where λ_1–λ_5 are the hyperparameters.

The detailed algorithm to realize the intelligent data augmentation for OOD generalization in a dual-adversarial training process is shown in Algorithm 1. It basically consists of two parts: one is the training part of AE-CRL and Dual-GAN (from Lines 3 to 19), where $\{z_1^{s\prime}, z_2^{s\prime}, \ldots, z_b^{s\prime}\}$ generated from the causal model in AE-CRL and $\{z_{G_1}^s, z_{G_2}^s, \ldots, z_{G_b}^s\}$ generated from the generator in Dual-GAN are used to adversarially train discriminator D_1, while the original data $\{x_1^s, x_2^s, \ldots, x_b^s\}$ and $\{x_1^{s\prime}, x_2^{s\prime}, \ldots, x_b^{s\prime}\}$ generated from decoder in AE-CRL and $\{x_1^{s\prime\prime}, x_2^{s\prime\prime}, \ldots, x_b^{s\prime\prime}\}$ generated from decoder in Dual-GAN are used to adversarially train discriminator D_2. Another one is the generation part based on the trained generative model, where $X'' = \{x_1'', x_2'', \ldots, x_m''\}$ will be used as the augmented data to facilitate the robust OOD generalization problem.

V. Experiment and Analysis

In this section, the experiment design with the used public dataset is first introduced, followed by a series of evaluations conducted to compare and demonstrate the usefulness and effectiveness of the proposed ITCRL-DGAN model in data augmentation for OOD generalization.

A. Experimental Design

The well-known Canadian Institute for Advanced Research, 10 classes (CIFAR-10) dataset is employed to conduct the experiment using our proposed model. The dataset consists of 60 000 of 32×32 color images from ten categories, namely, airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. We also use the CIFAR-10-C dataset to evaluate the extent to which our ITCRL-DGAN model can generalize to OOD issues. CIFAR-10-C contains variants of CIFAR-10 test images altered by various corruptions (e.g., Gaussian noise and motion blur). When computing sufficient input subsets on CIFAR-10-C images, we use a uniform random sample of 2000 images across the entire CIFAR-10-C set.

To verify the validity of the proposed model, several classical or newly proposed methods, including DAG-graph neural network (DAG-GNN) [49], causal autoencoder (CAE) [26], and regularized training with invariance on causal essential set (RICE) [50], are employed to conduct the comparison evaluations, which are summarized as follows.

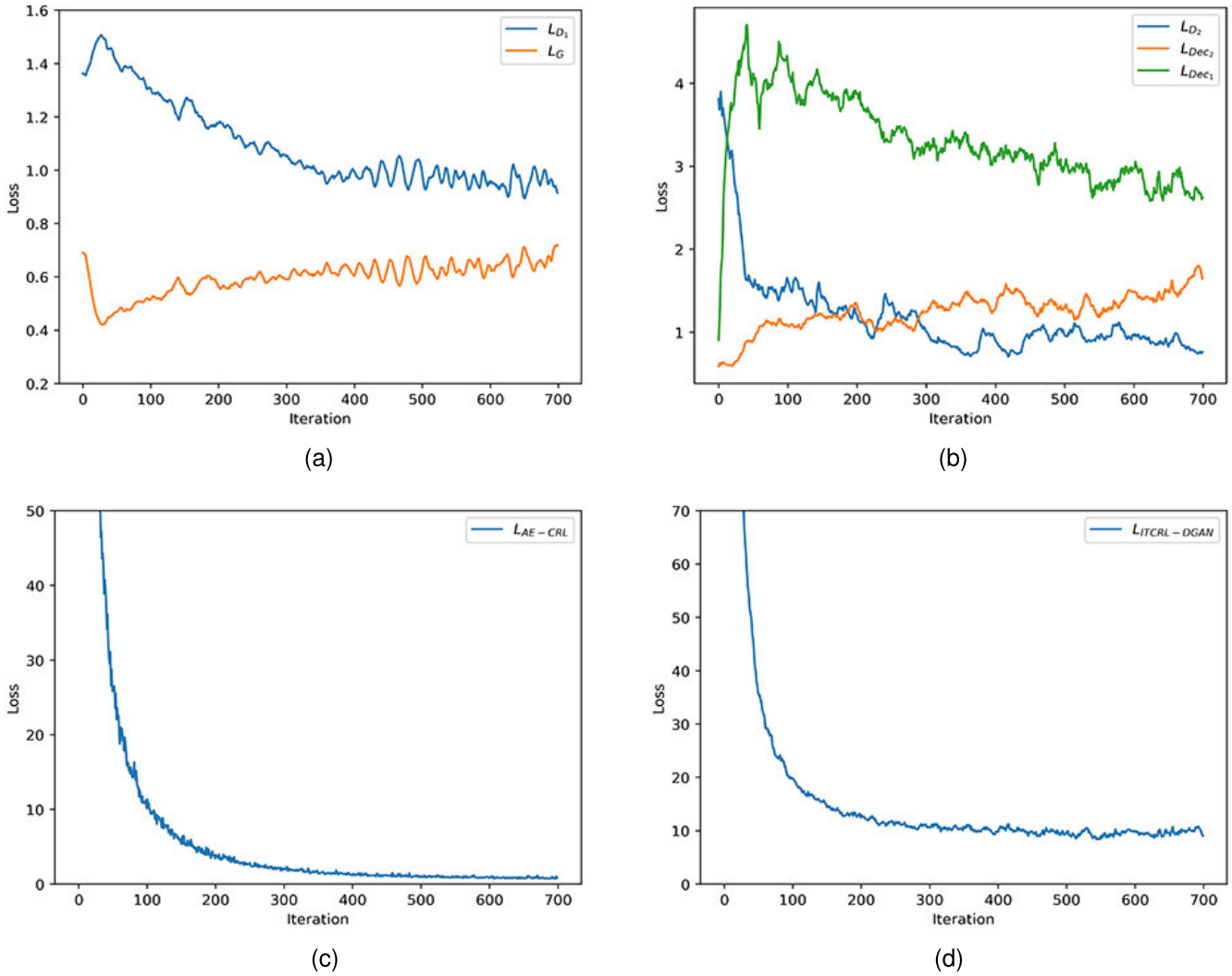


Fig. 3. Learning performance analysis. (a) First adversarial training. (b) Second adversarial training. (c) AE-CRL. (d) ITCRL-DGAN.

1) *DAG-GNN:* A deep generative model based on VAE, which introduced a GNN architecture with MLP to improve the data generation.
2) *CAE:* A causal structure learning method based on VAE, which combined a causal graph with an AE to capture causal representations for robust domain adaptation problem.
3) *RICE:* A regularized training algorithm based on Cycle-GAN, which employed the so-called causal invariant transformations to modify the noncausal feature for OOD generalization.

In addition, to demonstrate performances among our ITCRL-DGAN and the three baseline methods, evaluation metrics that include precision, recall, $F1$ score, and accuracy are considered. Experiments are conducted on an Ubuntu server, GTX 3080Ti, E5-2683 Core, 32G RAM, and Python 3.8.

B. Evaluation on Training Efficiency

We first compare and demonstrate the learning performance of our proposed model in the first adversarial training, the second adversarial training, AE-CRL, and the complete ITCRL-DGAN, respectively. The results are shown in Fig. 3.

As shown in Fig. 3(a) and (b), the loss values of the generator and discriminator are trending toward each other and eventually converge to their respective optimal losses, which indicates that our proposed model can be effectively trained to generate high-quality data. In particular, during the first adversarial training, the loss values of L_G and L_{D_1} converge to 0.66 and 0.93 around 500 iterations, respectively. This indicates the effectiveness of our model that is able to capture the latent features and enrich the feature representation in AE-CRL. As for the second adversarial training, a similar trend can be observed, where L_{Dec_2} and L_{D_2} are converging toward each other, reaching the values of 1.5 and 0.8, respectively. However, due to the complexity of the three-way adversarial process, three points are worth noting during the second adversarial training. First, during the first ten iterations, a rapid increase in L_{Dec_1} and a rapid decrease in L_{D_2} can be observed. This is due to the fact that the AE-CRL module shows a rapid drop in $L_{\text{AE-CRL}}$ loss in Fig. 3(c), which may cause the model instability. This affected the quality of the data generated by the model for a brief period of time. While after the initial instability, the quality of the generated data could recover and improve rapidly. Second, unlike in Fig. 3(a), where L_G and L_{D_1} have the tendency to converge, the loss of

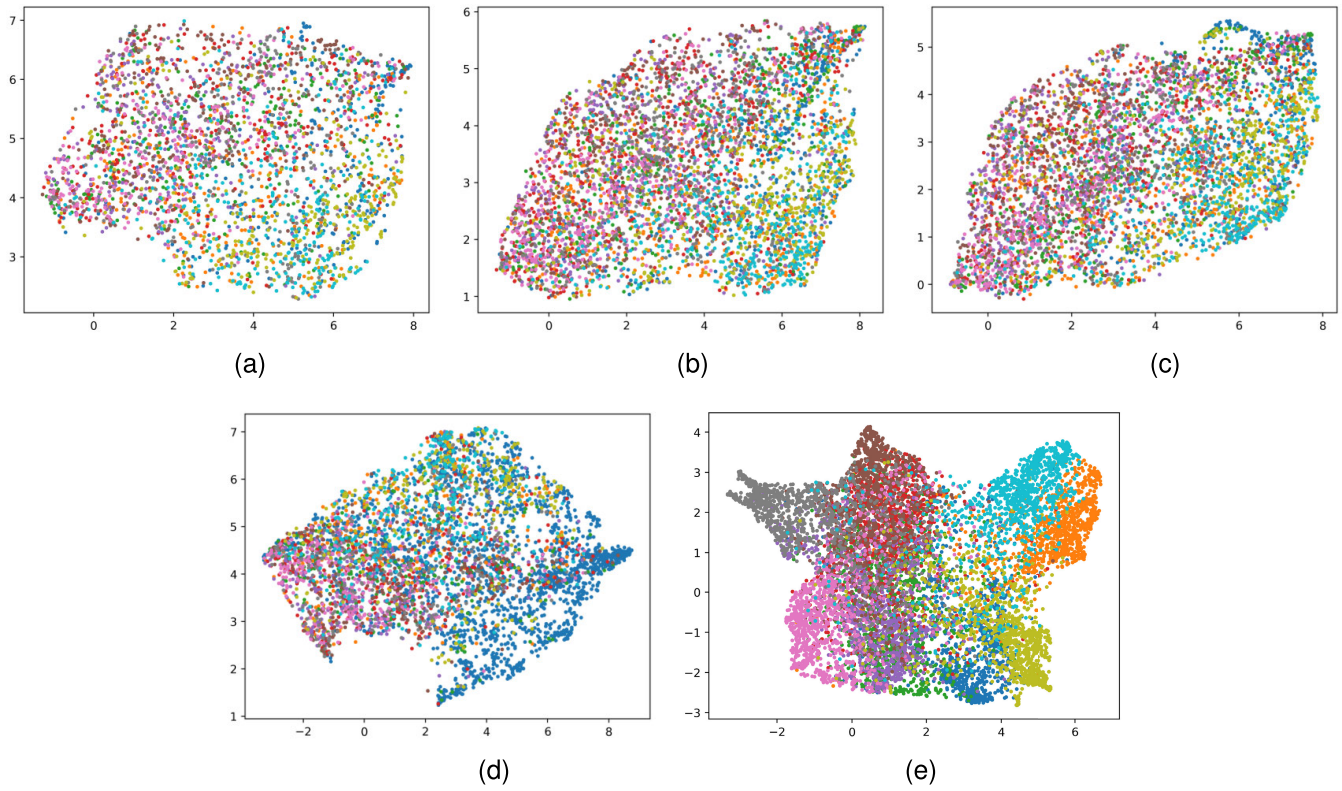


Fig. 4. Comparison on feature visualization based on augmented data using UMAP. (a) Original. (b) DAG-GNN. (c) CVAE. (d) RICE. (e) ITCRL-DGAN.

L_{Dec_2} crosses and exceeds the loss of L_{D_2}. This is primarily due to the complex interactions in the three-way adversarial training among the decoder in Dual-GAN, decoder in AE-CRL, and the second discriminator. In the second adversarial training, the rapid increase of L_{Dec_1} in the early stage leads to rapid decrease of L_{D_2}. Then, both L_{Dec_1} and L_{D_2} need to reduce and eventually force L_{Dec_2} to rise faster than that of common cases. In addition, it is found that all of them are able to continuously optimize their modules and eventually converge to the optimal loss values. This result demonstrates that our method can efficiently learn the causality-inspired feature representation and train the generative model in the dual-adversarial training process.

Moreover, according to Fig. 3(c) and (d), it can be observed that $L_{\text{AE-CRL}}$ and $L_{\text{ITCRL-DGAN}}$ decrease rapidly in the first 20–30 iterations and then become relatively stable. These results demonstrate the strong ability of our causal model in learning the optimized feature representation among the refined causal relationships and adaptability of the ITCRL-DGAN model in dealing with the OOD problem.

C. Evaluation on Data Augmentation

Then, we demonstrate the feature visualization of the generated data using uniform manifold approximation and projection (UMAP), which is a data downscaling and visualization tool with faster computation and better representation of high-dimensional data in comparison with t-distributed stochastic neighbor embedding (t-SNE) and principal component analysis (PCA). The comparison results among our proposed model and three baseline methods are illustrated in Fig. 4.

Fig. 4(a) and (b) shows the visualization results of the original data and the data processed by DAG-GNN, respectively. It can be observed that the distribution of similar data (the same label) is more dispersed, along with no obvious boundary between different categories of data. This means it is difficult for the subsequent classification models to learn the class-to-class boundaries and classify the samples correctly, leading to the classification results that are not particularly satisfied, especially when facing the OOD issue. Fig. 4(c) and (d) shows the visualization results of the data generated by CVAE and RICE, respectively, which introduced the causal models to perform data augmentation on the samples. However, it can observed that although the distribution of some classes is more concentrated, they still have the problem of unclear boundaries between classes. The data augmented by the proposed ITCRL-DGAN model are visualized in Fig. 4(e). In comparison with the other methods, our model results in a more uniform distribution and more obvious boundaries between classes, which indicates the significance of our AE-CRL in learning the enriched feature representation, and the effectiveness of our Dual-GAN in data augmentation from both the feature and data levels.

D. Evaluation on Classification Performance

We go further to compare the overall classification performance of our ITCRL-DGAN against other three baseline methods when facing the OOD problem. The classification performance and the ablation analysis results are shown in Tables I and II, respectively.

As shown in Table I, when facing the OOD problem, the common trend is that the classification performance will decrease significantly in the test dataset in comparison with

TABLE I
COMPARISON ON CLASSIFICATION PERFORMANCE BASED ON DIFFERENT METRICS

Method	Validation Dataset				Test Dataset			
	Precision	Recall	F1_score	ACC	Precision	Recall	F1_score	ACC
DAG-GNN	92.6	91.4	92.0	91.0	73.8	72.5	72.7	71.0
CVAE	94.2	93.4	93.8	93.0	77.5	75.4	76.4	74.0
RICE	95.2	94.3	94.7	94.0	80.7	77.8	79.2	76.0
ITCRL-DGAN	98.5	97.2	97.6	97.0	86.6	82.2	84.2	80.0

TABLE II
ABLATION ANALYSIS BASED ON DIFFERENT METRICS

Method	Validation Dataset				Test Dataset			
	Precision	Recall	F1_score	ACC	Precision	Recall	F1_score	ACC
ITCRL-DGAN	98.5	97.2	97.6	97.0	86.6	82.2	84.2	80.0
/CRL	96.3	95.4	95.8	94.9	83.8	77.4	80.6	77.2
/First-GAN	96.4	95.2	95.5	95.1	80.5	77.0	79.4	76.1
/AE	95.6	94.8	95.1	94.4	77.7	74.8	76.2	74.0

the training dataset. This general trend can be observed in all four methods in Table I. Referring to Table I, all four methods are able to achieve reasonably high training accuracy and $F1$ score of higher than 90%. However, the DAG-GNN only achieved 71% accuracy and 72.7% $F1$ score in test dataset, which is the lowest among all four methods, demonstrating its inability to handle the OOD problem. In comparison, the proposed ITCRL-DGAN achieved 80% accuracy (4%, 6%, and 9% improvement against RICE, CVAE, and DAG-GNN, respectively) and 84.2% $F1$ score (5%, 7.8%, and 11.5% improvement against RICE, CVAE, and DAG-GNN, respectively). These results indicate that the proposed model can effectively address the OOD problem in classification and still maintain a reasonable performance in the test dataset. In comparison with its best opponent (i.e., RICE), our ITCRL-DGAN can improve the GAN model with a dual-adversarial structure and an information theoretically enhanced causal AE structure, to generate better quality data that address the OOD problem. This eventually leads to a significant improvement in the accuracy of the subsequent classification models.

In addition to the comparison against the other three state-of-the-art methods, an ablation analysis is also conducted to demonstrate the usefulness of each training steps. The performance results are summarized in Table II. The /CRL indicates the removal of the causal model in ITCRL-DGAN, the /First-GAN is the case where the first discriminator is removed on the basis of /CRL, and the /AE is when the AE is removed on the basis of /First-GAN. The results clearly show that the performance improves with each individual steps, and removing any of these steps will result in a negative impact on the classification performance. Overall, the complete ITCRL-DGAN offers the best performance in addressing the OOD problem and achieves the best result in the test dataset because of the consideration of finding more causally related factors among core features and using ITL- and CRL-enhanced feature representations to aid a dual-adversarial training for the robust generalization.

VI. CONCLUSION

In this article, we designed and proposed a generative model called ITCRL-DGAN, which incorporated ITL and CRL into a dual-adversarial training process, to better learn the causality-inspired feature representation and improve the data augmentation in both feature and data levels, toward the robust OOD generalization in modern smart application development.

Specifically, an AE-CRL structure was designed to aid the dual-adversarial training based on the optimized feature representation, while a Dual-GAN structure was devised to enhance the data augmentation with the OOD samples. A feature separation strategy was newly proposed, which separated the compressed hidden variable into the causal feature and confounder feature, and only input the causal feature that could be more relevant to the class labels into the causal graph, to improve the effect of causal learning. An improved causal graph was then built, which could find more causally related factors among the separated core features and further enrich the feature representation with counterfactual features via interventions based on the refined causal relationships. In addition, the ITL was involved in the generative model training, which could not only facilitate the extraction of low-dimensional feature representations, but also help learn the improved causal representations based on the idea of "information flow." In the Dual-GAN structure, the first adversarial network was constructed to maximally retain the key information in the original data and expand the boundary of feature distribution in the feature level, and the second adversarial network was constructed to further improve the quality of the augmented data to approach to the real data distribution as much as possible. A dual-adversarial training mechanism was finally developed to realize the intelligent data augmentation for the robust OOD generalization. Compared with three other similar data augmentation methods, experiment and evaluation results based on an open-source dataset demonstrated the usefulness and effectiveness of our proposed generative model in terms of the outstanding learning efficiency and classification performance for the improvement of OOD generalization in smart application scenarios.

In future works, we will further investigate more deep learning-based data augmentation schemes to enhance the OOD generalization. More evaluations in different smart application scenarios will be conducted to improve the performance of our generative model with more efficient algorithms.

REFERENCES

[1] V. Sandfort, K. Yan, P. J. Pickhardt, and R. M. Summers, "Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks," *Sci. Rep.*, vol. 9, no. 1, Nov. 2019, Art. no. 16884.

[2] J. Liu, Z. Hu, P. Cui, B. Li, and Z. Shen, "Heterogeneous risk minimization," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 6804–6814.

[3] Z. Shen, P. Cui, T. Zhang, and K. Kunag, "Stable learning via sample reweighting," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 4, 2020, pp. 5692–5699.

[4] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do ImageNet classifiers generalize to ImageNet?" in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5389–5400.

[5] H. Ye, C. Xie, T. Cai, R. Li, Z. Li, and L. Wang, "Towards a theoretical framework of out-of-distribution generalization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 23519–23531.

[6] B. Li et al., "Invariant information bottleneck for domain generalization," in *Proc. 36th AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 7399–7407.

[7] D. Krueger et al., "Out-of-distribution generalization via risk extrapolation," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 5815–5826.

[8] W. Wang, X. Lin, F. Feng, X. He, M. Lin, and T.-S. Chua, "Causal representation learning for out-of-distribution recommendation," in *Proc. ACM Web Conf.*, Apr. 2022, pp. 3562–3571.

[9] S. Vernekar, A. Gaurav, V. Abdelzad, T. Denouden, R. Salay, and K. Czarnecki, "Out-of-distribution detection in classifiers via generation," 2019, *arXiv:1910.04241*.

[10] K. Sricharan and A. Srivastava, "Building robust classifiers through generation of confident out of distribution examples," in *Proc. 3rd Workshop Bayesian Deep Learn. (NeurIPS)*, 2018, pp. 1–5.

[11] K. Kuang, P. Cui, B. Li, M. Jiang, S. Yang, and F. Wang, "Treatment effect estimation with data-driven variable decomposition," in *Proc. 31st AAAI Conf. Artif. Intell.*, vol. 31, 2017, pp. 140–146.

[12] D. Liu et al., "Mitigating confounding bias in recommendation via information bottleneck," in *Proc. 15th ACM Conf. Recommender Syst.*, Sep. 2021, pp. 351–360.

[13] A. Honkela and H. Valpola, "Variational learning and bits-back coding: An information-theoretic view to Bayesian learning," *IEEE Trans. Neural Netw.*, vol. 15, no. 4, pp. 800–810, Jul. 2004.

[14] Q. Tan, Y. Liu, and J. Liu, "Demystifying deep learning in predictive spatiotemporal analytics: An information-theoretic framework," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 8, pp. 3538–3552, Aug. 2021.

[15] Z. Goldfeld and Y. Polyanskiy, "The information bottleneck problem and its applications in machine learning," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 19–38, May 2020.

[16] C. Li, P. Shen, Y. Liu, and Z. Zhang, "Diffusion information theoretic learning for distributed estimation over network," *IEEE Trans. Signal Process.*, vol. 61, no. 16, pp. 4011–4024, Aug. 2013.

[17] R. Boscolo, J. Liao, and V. Roychowdhury, "An information theoretic exploratory method for learning patterns of conditional gene coexpression from microarray data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 5, no. 1, pp. 15–24, Mar. 2008.

[18] Y. Deng, F. Bao, X. Deng, R. Wang, Y. Kong, and Q. Dai, "Deep and structured robust information theoretic learning for image analysis," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4209–4221, Sep. 2016.

[19] P. E. Meyer, C. Schretter, and G. Bontempi, "Information-theoretic feature selection in microarray data using variable complementarity," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 3, pp. 261–274, Jun. 2008.

[20] W. Zhang, B. Jiang, M. Li, and X. Lin, "Privacy-preserving aggregate mobility data release: An information-theoretic deep reinforcement learning approach," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 849–864, 2022.

[21] O. Özdenizci and D. Erdogmus, "Information theoretic feature transformation learning for brain interfaces," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 1, pp. 69–78, Jan. 2020.

[22] Z. Xu, Y. Liu, and C. Li, "Distributed information-theoretic semisupervised learning for multilabel classification," *IEEE Trans. Cybern.*, vol. 52, no. 2, pp. 821–835, Feb. 2022.

[23] Y. Sun, K. Zhang, and C. Sun, "Model-based transfer reinforcement learning based on graphical model representations," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 2, pp. 1035–1048, Feb. 2023.

[24] X. Wang, Q. Li, D. Yu, P. Cui, Z. Wang, and G. Xu, "Causal disentanglement for semantic-aware intent learning in recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 10, pp. 9836–9849, Oct. 2023.

[25] Z. Guo, Z. Zhang, R. Feng, and Z. Chen, "Causal contextual prediction for learned image compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2329–2341, Apr. 2022.

[26] S. Yang, K. Yu, F. Cao, L. Liu, H. Wang, and J. Li, "Learning causal representations for robust domain adaptation," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 3, pp. 2750–2764, Mar. 2023.

[27] S. Rao et al., "Targeted-BEHRT: Deep learning for observational causal inference on longitudinal electronic health records," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 23, 2022, doi: 10.1109/TNNLS.2022.3183864.

[28] H. Zhang, L. Xiao, X. Cao, and H. Foroosh, "Multiple adverse weather conditions adaptation for object detection via causal intervention," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 12, 2022, doi: 10.1109/TPAMI.2022.3166765.

[29] Y. Xiang and M. Truong, "Acquisition of causal models for local distributions in Bayesian networks," *IEEE Trans. Cybern.*, vol. 44, no. 9, pp. 1591–1604, Sep. 2014.

[30] X. Wang, Y. Wu, A. Zhang, F. Feng, X. He, and T.-S. Chua, "Reinforced causal explainer for graph neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2297–2309, Feb. 2023.

[31] B. Sahoh, C. Kaewrat, K. Yeranee, N. Kittiphattanabawon, and M. Kliangkhlao, "Causal AI-powered event interpretation: A cause-and-effect discovery for indoor thermal comfort measurements," *IEEE Internet Things J.*, vol. 9, no. 22, pp. 23188–23200, Nov. 2022.

[32] F. Zhu, J. Lu, A. Lin, J. Xuan, and G. Zhang, "Direct learning with multi-task neural networks for treatment effect estimation," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 3, pp. 2457–2470, Mar. 2023.

[33] C. Huang et al., "Self-supervised attentive generative adversarial networks for video anomaly detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 9389–9403, Nov. 2023.

[34] Y. Cai et al., "Dualattn-GAN: Text to image synthesis with dual attentional generative adversarial network," *IEEE Access*, vol. 7, pp. 183706–183716, 2019.

[35] Y. Wang et al., "An intelligent method for predicting the pressure coefficient curve of airfoil-based conditional generative adversarial networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 7, pp. 3538–3552, Jul. 2023.

[36] S.-H. Choi, J.-M. Shin, P. Liu, and Y.-H. Choi, "ARGAN: Adversarially robust generative adversarial networks for deep neural networks against adversarial examples," *IEEE Access*, vol. 10, pp. 33602–33615, 2022.

[37] C. Zheng, L. Pan, and P. Wu, "Attribute augmented network embedding based on generative adversarial nets," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 7, pp. 3473–3487, Jul. 2023.

[38] G. Shao, M. Huang, F. Gao, T. Liu, and L. Li, "DuCaGAN: Unified dual capsule generative adversarial network for unsupervised image-to-image translation," *IEEE Access*, vol. 8, pp. 154691–154707, 2020.

[39] T. Jiang, W. Xie, Y. Li, J. Lei, and Q. Du, "Weakly supervised discriminative learning with spectral constrained generative adversarial network for hyperspectral anomaly detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 11, pp. 6504–6517, Nov. 2022.

[40] B. Franci and S. Grammatico, "Training generative adversarial networks via stochastic Nash games," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 3, pp. 1319–1328, Mar. 2023.

[41] Y. Tian, L. Shen, L. Shen, G. Su, Z. Li, and W. Liu, "AlphaGAN: Fully differentiable architecture search for generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6752–6766, Oct. 2022.

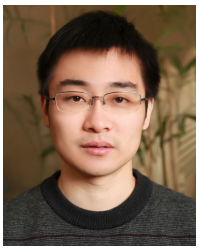
[42] K. Lee, H. Lee, K. Lee, and J. Shin, "Training confidence-calibrated classifiers for detecting out-of-distribution samples," in *Proc. ICLR*, 2018, pp. 1–16.

[43] D. Mandal et al., "Out-Of-distribution detection for generalized zero-shot action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9977–9985.

[44] M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, and J. Wang, "CausalVAE: Disentangled representation learning via neural structural causal models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9588–9597.

[45] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[46] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
[47] N. Ay and D. Polani, "Information flows in causal networks," *Adv. Complex Syst.*, vol. 11, no. 1, pp. 17–41, Feb. 2008.
[48] J. Pearl, M. Glymour, and N. P. Jewell, *Causal Inference in Statistics: A Primer*. Hoboken, NJ, USA: Wiley, 2016.
[49] Y. Yu, J. Chen, T. Gao, and M. Yu, "DAG-GNN: DAG structure learning with graph neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7154–7163.
[50] R. Wang, M. Yi, Z. Chen, and S. Zhu, "Out-of-distribution generalization with causal invariant transformations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 375–385.



Xiaokang Zhou (Member, IEEE) received the Ph.D. degree in human sciences from Waseda University, Tokorozawa, Japan, in 2014.

From 2012 to 2015, he was a Research Associate with the Faculty of Human Sciences, Waseda University. Since 2017, he has been working as a Visiting Researcher with the RIKEN Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo, Japan. He is currently an Associate Professor with the Faculty of Data Science, Shiga University, Hikone, Japan. He has been engaged in interdisciplinary research works in the fields of computer science and engineering, information systems, and social and human informatics. His recent research interests include ubiquitous computing, big data, machine learning, behavior and cognitive informatics, cyber–physical–social systems, and cyber intelligence and security.

Dr. Zhou is a member of the IEEE Computer Society (CS); Association for Computing Machinery (ACM), USA; Information Processing Society of Japan (IPSJ) and The Japanese Society for Artificial Intelligence (JSAI), Japan; and China Computer Federation (CCF), China.

Xuzhe Zheng received the bachelor's degree in management from the Hunan University of Technology and Business, Changsha, China, in 2020, where he is currently pursuing the master's degree in management science and engineering.

His main research interests include causal inference research and data-driven decision-making research.

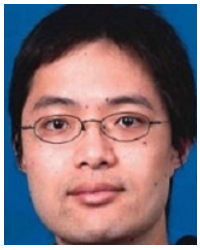
Tian Shu received the bachelor's degree in computer science, majoring in the Internet of Things, from Xiangnan University, Chenzhou, China, in 2020. She is currently pursuing the master's degree in electronic information with the Hunan University of Technology and Business, Changsha, China.

Her recent research interests include deep learning, medical big data, and intelligent transportation.

Wei Liang (Member, IEEE) received the M.S. and Ph.D. degrees in computer science from Central South University, Changsha, China, in 2005 and 2016, respectively.

From 2014 to 2015, he was a Researcher with the Department of Human Informatics and Cognitive Sciences, Waseda University, Tokorozawa, Japan. He is currently working at the Xiangjiang Laboratory and the Changsha Social Laboratory of Artificial Intelligence, Hunan University of Technology and Business, Changsha. He has published more than 20 papers at various conferences and journals. His research interests include information retrieval, data mining, and artificial intelligence.

Dr. Liang is a member of the IEEE Computer Society (CS) and China Computer Federation (CCF), China.



Kevin I-Kai Wang (Member, IEEE) received the B.E. degree (Hons.) in computer systems engineering and the Ph.D. degree in electrical and electronics engineering from the Department of Electrical and Computer Engineering, University of Auckland, Auckland, New Zealand, in 2004 and 2009, respectively.

He was a research engineer designing commercial home automation systems and traffic sensing systems from 2009 to 2011. He is currently a Senior Lecturer with the Department of Electrical, Computer, and Software Engineering, University of Auckland. His current research interests include wireless sensor network-based ambient intelligence, pervasive healthcare systems, human activity recognition, behavior data analytics, and biocybernetic systems.

Lianyong Qi (Senior Member, IEEE) received the Ph.D. degree from the Department of Computer Science and Technology, Nanjing University, Nanjing, China, in 2011.

In 2010, he visited the Department of Information and Communication Technology, Swinburne University of Technology, Melbourne, VIC, Australia. He is currently a Professor with the College of Computer Science and Technology, China University of Petroleum (East China), Qingdao, China. His research interests include big data and recommender systems.

Shohei Shimizu received the Ph.D. degree in engineering (statistical science) from Osaka University, Osaka, Japan, in 2006.

He is currently a Professor with the Faculty of Data Science, Shiga University, Hikone, Japan, where he leads the Causal Inference Team, RIKEN Center for Advanced Intelligence Project. His research interests include statistical methodologies for learning data generating processes, such as structural equation modeling and independent component analysis and their application to causal inference.

Dr. Shimizu received the Hayashi Chikio Award (Excellence Award) from the Behaviormetric Society in 2016. He has been a Coordinating Editor of *Behaviormetrika* (Springer) since 2016.

Qun Jin (Senior Member, IEEE) received the Ph.D. degree from Nihon University, Tokyo, Japan, in March 1992.

He is currently a Professor with the Department of Human Informatics and Cognitive Sciences, Faculty of Human Sciences, Waseda University, Tokorozawa, Japan. He has been extensively engaged in research works in the fields of computer science, information systems, and human informatics, with a focus on understanding and supporting humans through convergent research. His recent research interests cover computing for human well-being, behavior and cognitive informatics, health informatics, big data, personal analytics and individual modeling, digital twin, smart energy and behavioral data analytics for carbon neutrality, cyber security, blockchain, metaverse, artificial intelligence and machine learning, and applications in healthcare and learning support.

Dr. Jin is a Foreign Fellow of the Engineering Academy of Japan (EAJ).