electronics

MDPI

Article

A Causality-Aware Perspective on Domain Generalization via Domain Intervention

Youjia Shao [1], Shaohui Wang [1] and Wencang Zhao [1,2,3,*]

[1] College of Automation and Electronic Engineering, Qingdao University of Science and Technology, Qingdao 266061, China; nicoleshao98@mails.qust.edu.cn (Y.S.); 4021040041@mails.qust.edu.cn (S.W.)
[2] Qingdao Institute of Intelligent Navigation and Control, Qingdao 266071, China
[3] Shandong Key Laboratory of Autonomous Landing for Deep Space Exploration, Qingdao 266061, China
* Correspondence: coinslab@qust.edu.cn

Abstract: Most mainstream statistical models will achieve poor performance in Out-Of-Distribution (OOD) generalization. This is because these models tend to learn the spurious correlation between data and will collapse when the domain shift exists. If we want artificial intelligence (AI) to make great strides in real life, the current focus needs to be shifted to the OOD problem of deep learning models to explore the generalization ability under unknown environments. Domain generalization (DG) focusing on OOD generalization is proposed, which is able to transfer the knowledge extracted from multiple source domains to the unseen target domain. We are inspired by intuitive thinking about human intelligence relying on causality. Unlike relying on plain probability correlations, we apply a novel causal perspective to DG, which can improve the OOD generalization ability of the trained model by mining the invariant causal mechanism. Firstly, we construct the inclusive causal graph for most DG tasks through stepwise causal analysis based on the data generation process in the natural environment and introduce the reasonable Structural Causal Model (SCM). Secondly, based on counterfactual inference, causal semantic representation learning with domain intervention (CSRDN) is proposed to train a robust model. In this regard, we generate counterfactual representations for different domain interventions, which can help the model learn causal semantics and develop generalization capacity. At the same time, we seek the Pareto optimal solution in the optimization process based on the loss function to obtain a more advanced training model. Extensive experimental results of Rotated MNIST and PACS as well as VLCS datasets verify the effectiveness of the proposed CSRDN. The proposed method can integrate causal inference into domain generalization by enhancing interpretability and applicability and brings a boost to challenging OOD generalization problems.

Keywords: domain generalization; causal inference; counterfactual representation; domain intervention

Citation: Shao, Y.; Wang, S.; Zhao, W. A Causality-Aware Perspective on Domain Generalization via Domain Intervention. *Electronics* **2024**, *13*, 1891. https://doi.org/10.3390/electronics13101891

Academic Editor: George A. Tsihrintzis

Received: 14 April 2024
Revised: 5 May 2024
Accepted: 8 May 2024
Published: 11 May 2024

1. Introduction

Today, deep learning (DL) has achieved remarkable success in various fields, showing superior performance [1,2]. However, this performance guarantee mostly relies on a general assumption that the distribution of training data (source domain) and test data (target domain) is independent and identical (Independent Identically Distribution, IID) [3,4]. IID assumptions mostly violate the real-life scene and affect the ability of models to generalize in unknown environments, which limits the sustainability of artificial intelligence (AI) in future fields. If DL models want to gain high-level machine intelligence targeting for simulating human cognition to make progress in landing various real-life applications, we must expand our horizons to the Out-Of-Distribution (OOD) problem, which means that the data distribution of the target domain is different from that of the source domains [5]. When traditional statistical models relying on IID data are applied to OOD generalization problems, catastrophic performance degradation occurs due to over-reliance on probabilistic correlations [6,7]. Based on the pursuit of breaking this restricted situation, the

motivation for our work is to design an effective training framework under the OOD distribution, which can reduce the over-reliance on probabilistic correlations and enable the model to stably perform OOD generalization, with superior generalization performance in the unseen target domain. Domain generalization (DG) is proposed, which is essentially a kind of OOD generalization problem, and can generalize the knowledge acquired from source domains to the unseen target domain [8–10]. In DG, the target domain is unavailable during the training phase, which fits the challenge of real-world situations [11]. In recent years, challenging DG has attracted the attention of many researchers. In this paper, our work focuses on the study of domain generalization and is committed to improving the stable generalization performance of the model.

We define domain shifts as the discrepancy between training and testing domains. In order to solve the problem caused by domain shifts and improve the generalization ability of DL models, the methods in the DG field are increasing, which can be roughly divided into domain alignment [9,12,13], meta-learning [14,15], data augmentation [16,17], etc. Although these have achieved some success, most of them are operated at the traditional probability hierarchy without explaining the intrinsically causal mechanism. This probabilistic correlation is insufficient for real-life tasks, such as in image classification, where the background of boats is often the lake, implying that boats and lakes are highly correlated. In this case, it is easier for the probabilistic model to learn the disruptive information of the lake as the discriminative feature for classifying boats. When the domain changes (e.g., the background is a desert), the model may not be able to recognize the label of the boat due to the absence of factors that contain lake information, resulting in poor generalization results. At this point, the cognitive level of the model is relatively low, and it is not able to achieve stable cross-domain generalization. Unlike naive machine models, humans can accurately identify the boat label in images without being affected by changing backgrounds and image styles. It means that people have the ability to adapt to complex and changing environments. Humans are capable of solving unknown problems with continuously accumulated knowledge based on their high-level human cognition, showing stable generalization ability. We argue that this intelligent generalization relies on human inference ability, which is centered on causality [18,19]. Inspired by the above, in order to enable deep learning models to enhance OOD generalization capabilities, we combine causal inference with domain generalization and utilize a novel causal perspective to view domain generalization. From a causal view, in this example, the characteristics of the boat, such as lines and shapes, are stable causal factors, while the background information is the unconcerned factor for the classification task. Causal models capture causal relationships between variables and allow us to predict how a system will behave under interventions or changes in distribution, which are more powerful than probability-dependent models [18,19]. Integrating causality into domain generalization is bound to have certain advantages and potential. With the deepening of causal inference in the field of artificial intelligence, some studies that combine causality and DG have appeared [20–22]. These include MatchDG [20], which regards samples of the same category from different fields as positive pairs and samples of different categories as negative pairs to pull in similar representations for label matching. In addition, Rojas-Carulla et al. [23] propose to utilize an effective subset of predictions in an adversarial environment, and Müller et al. [21] rely on the principle of independent causal mechanism to build a gradient-based learning framework. The key points of these methods differ, but they all contain causal commonalities. Based on the motivation of improving the OOD generalization ability, our work innovatively draws inspiration from human cognition, enhancing the interpretability of the combination of domain generalization and causal inference, and the purpose is to train the effective model that can improve domain generalization performance through an innovative causal perspective.

In this paper, we regard domain generalization from a novel causal perspective, and the task is dedicated to applying causal representation learning to domain generalization and exploring the causal invariant mechanism, thereby improving the OOD generaliza-

tion ability of the model. We propose the inclusive causal graph and leverage a novel domain intervention to learn stable causal semantic representations through counterfactual inference and ultimately seek the Pareto optimal solution in the optimization of the loss function to obtain more advanced results, which can provide a creative idea for improving the OOD generalization. Firstly, an inclusive causal Directed Acyclic Graph (DAG) is established, which can be applied to general DG tasks in the Computer Vision (CV) field. A proxy domain variable is introduced to enrich the causal graph based on the data generation process in the natural environment to visualize the domain shift and explain the generalization bias. In this regard, we point out that even though the domain changes, the semantics of the input would remain consistent as long as the object is unchanged. In generalization tasks, the model aims to learn high-level causal semantic representations that remain invariant under different disturbances. Inspired by the above, we further propose causal semantic representation learning with domain intervention (CSRDN), which learns causal representations by applying generative interventions, so as to train a robust model that can resist the domain changes caused by various disturbances. Based on the display of our causal graph and the concept of the causal do-calculus [19], we regard the change of domains as the interventional method to fit the domain shift. To remove confounding effects, we take perturbed control of the proxy domain variable and implement the cutoff for lifting the prejudicial restriction. Different from relatively plain image transformation about simple rotation and cropping, we utilize a novel domain intervention method to intervene on non-causal information to increase the randomness of the simulated natural world changes. We rely on the idea of counterfactual inference [24,25] to generate the counterfactual representations so as to simulate the different changes to the domain. With the help of this kind of intervention, stable causal representation learning is performed. The causal DAG gives us the hint, and the label of the image and random Gaussian noise serve as input to construct the training of our counterfactual representation generator. Due to the reverse generation from the latent distribution and the maximum preservation of object classification properties, the semantic traits of ideal counterfactual representations will remain unchanged while including the different domain changes. As the similarity between the distributions of the counterfactual representation and the original causal representation continues to approach, this model can learn the invariant causal mechanism across domains by seeking the Pareto optimal solution, thereby improving the OOD generalization ability. Our contributions are summarized as follows:

1. We view domain generalization through a causal lens derived from the intuitive core of human intelligence. Based on the data generation mechanism in natural environments for causal modeling, the Structural Causal Model (SCM) [18] is injected to construct the causal DAG that is inclusive of various DG tasks in CV.
2. A novel stable framework CSRDN for causal semantic representation learning is proposed. According to the presentation of our causal DAG, we conduct the domain intervention in the learning process based on counterfactual inference, which is achieved by the generation of counterfactual representations to change non-causal information. The model seeks the Pareto optimal solution based on the loss function in the optimization process. Our work is able to exploit causal invariance to improve OOD generalization.
3. Extensive experiments are conducted and sequentially analyzed on three widely used datasets, including the synthetic dataset Rotated MNIST [26], the dataset PACS [27] with significant differences in style, and the real-world dataset VLCS [28]. The effectiveness and superiority of our method are proved by the detailed experimental results.

This paper is organized as follows. Section 2 describes related work corresponding to the content of this paper. In Section 3, the analysis of the causal DAG and the methodological details of the CSRDN framework are presented. The experimental setup and the analysis and discussion of the experimental results are presented in Section 4. Finally, we summarize our work and put forward prospects for future research in Section 5.

2. Related Work

2.1. Domain Generalization

Domain generalization aims to generalize the knowledge learned from source domains to the unseen target domain [11]. Through nearly ten years of development, researchers have explored plenty of methods. Domain alignment [9,12,13] is applied to many DG tasks, and the core is to minimize the difference between source domains to learn domain invariant representations, improving the generalization ability. These include minimizing moments [29], minimizing contrastive loss [30], domain-adversarial learning [12,13], etc., which are used for distribution alignments. Data augmentation [16,17] increases the diversity of source domains at the data level to improve the robust generalization of the system. These include traditional image translation [16], the perturbative change of input images using adversarial gradients obtained by task classifiers [17], and synthesizing domain-agnostic images by using domain adversarial gradients [31]. Learnable augmentation networks [32] are also introduced to generate new data with augmentation neural networks to synthesize new domains. Meta-learning [14,15] is also popularly applied in DG as a fast-growing field of DL, which aims to learn from events sampled from related tasks to benefit future learning. The most relevant paper is MAML [33], which divides the training data into two sets for meta-training and meta-testing, and the model is trained on the meta-training data to improve the performance of meta-testing data. Our method applies a novel causality-based lens to the analysis of domain generalization, focusing on the learning of causal semantic representations to effectively improve the generalization ability.

2.2. Causal Representation Learning

Causal inference focuses on the mining of the causal mechanism, which is different from statistical correlation modeling. Statistical learning cannot be reliably applied to OOD generalization tasks, inaccurately responding to counterfactual input. Causal models have more important information in substance compared to statistical models [18,19]. Causality cannot be defined simply and directly in terms of Boolean logic language [34] or probabilistic inference [35]. It needs to take the concept of intervention into account [19], which can be regarded as a component of the chain of inference. The core of combining artificial intelligence and causal learning is causal representation learning, which can extract structured variables that can be used for causal inference from unstructured data [35]. Shalit et al. [36] propose generalization error bounds and corresponding algorithms for predicting Individual Treatment Effects (ITE), which can learn a balanced representation by making the distributions of treatment and control groups similar. Hassanpour and Greiner [37] propose to utilize context-aware importance sampling to balance the selection bias, thereby replacing fixed weights and learning reasonable representations. Kallus and Nathan [38] put forward a new approach based on adversarial training of weighted and discriminative networks. In cases where there are multiple covariates and complex relationships among them, this method achieves excellent covariate balance, enabling robust causal analysis. CausalVAE [39] adds a causal layer on the basis of the traditional variational autoencoder model. The causal layer converts independent exogenous factors into endogenous factors of the causal graph, and then the mask mechanism transmits the representation generated by the causal intervention and finally decodes the representation.

3. Method

An inclusive causal DAG is first created on the basis of the data generation process combined with SCM. This causal graph can cover most of the DG tasks in the CV field, having a certain degree of universality, and can show the essential causal mechanism. Secondly, we propose causal semantic representation learning with domain intervention (CSRDN) based on the structure of the proposed causal DAG by regarding domain changes as interventions. Counterfactual inference is utilized in CSRDN with pointing out three basic requirements that need to be met. We train counterfactual representation generators based on adversarial learning with Generative Adversarial Network (GAN) [40] architec-

ture. Finally, the robust model is continuously optimized by the Pareto optimal solution of the loss function from the goals and is able to obtain stable generalization from the source domain to the invisible target domain.

3.1. Causal Graph Modeling via SCM

In the DL field, it is assumed that there is a causal relationship between two variables, and if one is the cause, the other must be the effect (cause $\rightarrow$ effect). According to the data generation in a natural environment, we introduce SCM for causal modeling. Some methods [41,42] associate label Y directly with a subset of covariates of X, which we believe is conceptually unreasonable at a semantic level, as pixel-wise covariates of X cannot contain semantic information.

For input X ($X \in \mathbb{R}^d$) and output label Y ($Y \in \mathbb{R}$) from M ($M > 1$) domains, a joint space $X \times Y$ is generated. Observable input X is composed of two parts, causal semantic factors C and unconcerned factors U, $X \leftarrow (C, U)$, and only the former can effect the output label Y, $Y \leftarrow C$, as shown in Figure 1a. Causal semantic factors contain a series of information that determines the output. For instance, in object recognition, the shape information is the causal factor that contains object discrimination, while the photo background or shooting angle is the unconcerned factor, which is independent of the classification label. At the same time, we have an in-depth understanding of domain shifts and consider that they are attributed to two aspects.

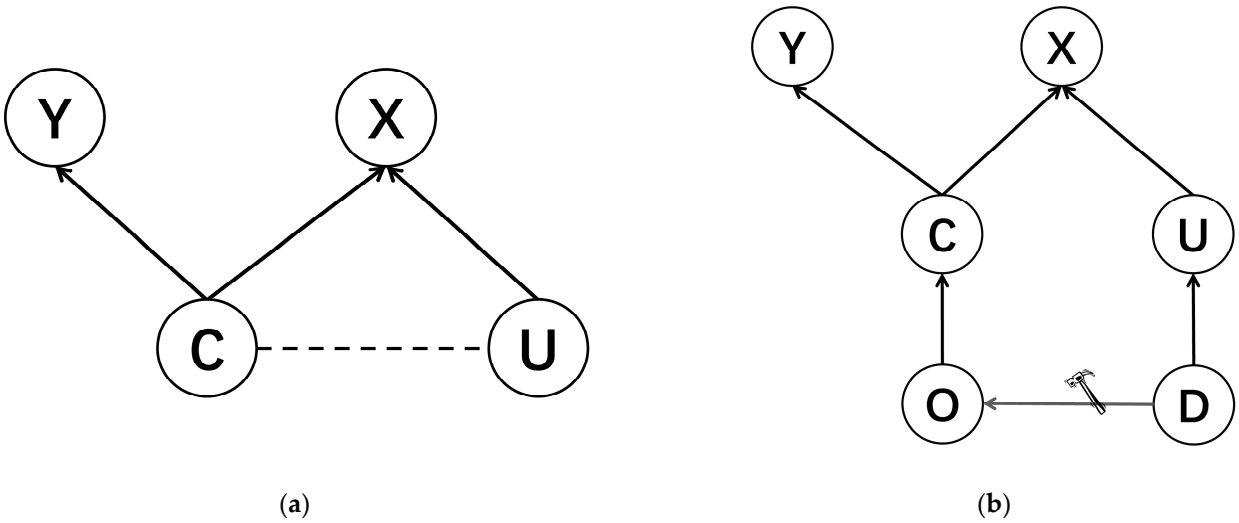

Figure 1. An inclusive causal graph for most DG tasks in the CV area. (**a**) Input X is composed of causal semantic factors C and unconcerned factors U, and only C points to output label Y through the causal solid line arrow. C and U are connected by a dashed line without an arrow, representing the spurious correlation. (**b**) The proxy domain variable D is introduced to enrich (**a**), pointing to both object variable O and U, replacing the dashed line between C and U. The hammer pattern represents the cutoff when controlling D as the intervention. Finally, a complete causal graph is presented.

Shifted unconcerned factors. Unconcerned features are not independent of the domain, which includes the style of the image, background, perspective, etc. They vary across domains, resulting in a changing distribution. Statistical models may pay too much attention to these factors due to probability, which is far from human cognitive classification ability. People can keenly discover the object related to the label in multiple images and automatically ignore irrelevant background elements. We hope that the model can learn the generalization ability like humans to the greatest extent.

Shifted C–U spurious correlation effect. There is an undeniable spurious correlation between causal factors and unconcerned factors, which is not true causality. In most cases, data sampling preference leads to the existence of confounding effects between the two factors. For example, since the camera prefers to capture images of boats sailing

in water, this sampling preference leads to a high correlation between discriminating lines of boats and water, which is often detrimental to model predictions when domain shifting. When shifting, non-causal objects in the background may disappear, or this unstable spurious correlation may change, and a series of decisions based on probabilistic correlation may break into different domains.

Due to the above aspects, C and U are connected by a dashed line without an arrow, which is different from the causal solid line with an arrow, as shown in Figure 1a. Taking into account the data generalization in the natural environment, we introduce the proxy domain variable D for explaining the limitations of generalization performance, adding $D \to O \to C$ and $D \to U$ to replace the correlation linkage between C and U, as shown in Figure 1b. An example is given to illustrate this enhanced causal graph. For example, in nature pictures, unique animals survive in specific environments, and we can infer from an Arctic environment D that object O is the Arctic fox adapted to the ice in the subdivision of fox classification, and thus $D \to O$. It is a precarious one-way line of reasoning. The object provides a range of causal information for predicting the label, with $O \to C$. Object O can be considered an interpretable bridge. However, the presence of Arctic foxes alone cannot infer that the current environment is the North Pole. D simultaneously provides a set of unconcerned factors that would interfere in predicting Y; $D \to U$. We do not simply define domains as distributions of variation [22,41] but present it as a separate proxy variable, which can simply and clearly show the impact of domain shift. This enhanced causal DAG can make it easier for people to understand the mechanism of DG, which has a certain degree of intuition.

In our causal DAG, SCM is introduced. The SCM views a set of variables $X_1, \ldots, X_n$ associated with the vertices of a DAG. It is assumed that each variable can be represented by a deterministic function that depends on X_i's parents in the graph (denoted by PA_i) and an unexplained random variable V_i.

$$X_i = f_i(PA_i, V_i), \ (i = 1, 2, \ldots, n) \tag{1}$$

where random noise variables $V_1, \ldots, V_n$ are joint independent. In SCM, the intervention can be seen as an operation to modify the formula, such as changing the random noise V_i or changing the form of the function f_i. According to Equation (1) and Figure 1, we can formalize general DG tasks as follows:

$$\begin{aligned} X &= f(C, U, V_1) \\ Y &= h(C, V_2) = h(\Phi(X), V_2) \end{aligned} \tag{2}$$

where f, h, and Φ are unknown structure functions. In light of the causal invariant mechanism, if C is known, for any distribution $P(X, Y)$, we can train the optimal predictor according to the naive Empirical Risk Minimization (ERM).

$$h^* = \text{argmin}_h \mathrm{E}_{P(X,Y)}[L(Y, h(C))] = \text{argmin}_h \mathrm{E}_{P(X,Y)}[L(Y, h(\Phi(X)))] \tag{3}$$

where $L(,)$ denotes the cross entropy loss. However, it is an overly shallow idea. In actual situations, we cannot accurately obtain unobservable causal factors C while all we have is the observable input X. Meanwhile, although a large number of priori hypotheses can be placed again, it is difficult to directly construct ambiguous causal factors from the input. Therefore, from the perspective of causal semantic stability, we simulate the mining of causal factors to the greatest extent through causal semantic representation learning. In the process, we rely on the idea of counterfactual inference to achieve domain intervention to learn the robust predictive model. We expand the details of our proposed method in the following section.

3.2. Causal Semantic Representation Learning with Domain Intervention

In this section, we introduce CSRDN in detail. We consider the change of domain as an intervention that occurs more commonly in real life than in image transformation. In the causal graph in the previous subsection, D is a confounder for learning causal effects from inputs to outputs and imposes limitations on prediction tasks, so we need to control domain variable D to remove confounding effects. At the same time, $D \rightarrow O$ is pruned due to unstable bias, blocking the information flow, as shown in Figure 1b. By implementing the domain intervention, $do(D)$, theoretically, $P(Y|C, do(D))$ is invariant across different domains. In the process of exploring causal effects, the model can better learn stable causal semantic representations. We start from the stability of causal semantics and randomization of interventions and propose a novel domain intervention method by generating counterfactual representations. The causal direction $O \rightarrow C \rightarrow Y$ gives us the hint, the inputs of the counterfactual representation generator are label Y and random Gaussian noise, and the output is the counterfactual representation whose semantic characteristics are consistent with the original input. It is worth noting that we perform inference operations at the representation level, focusing on generating counterfactual representations rather than counterfactual images in order to directly pull in the similarity with the original causal representation. Inspired by counterfactual inference [24], we need to achieve the following three goals as much as possible and make effective trade-offs between each other in CSRDN: (1) using original causal representations to make accurate predictions for estimating good facts, (2) achieving low-error generation of counterfactual representations for effectively estimating good counter-facts, and (3) balancing the distributional similarity of representations under different interventions.

Based on the above three basic goals, we design the overall framework, as shown in Figure 2. It mainly includes three modules, namely, representation extractor r, predictor h, and counterfactual representation generator cg. For the input data from source domains, we need to train the representation extractor r, inputting raw X from source domains to extract the original causal representation containing causal information. As the domain intervention, corresponding ground-truth label Y and random Gaussian noise are input to cg for extracting counterfactual representations. We take the original causal representation and the counterfactual representation as input to jointly train the predictor h. During the training process of these three modules, it is necessary to constantly weigh the above three goals and perform loss approximation. In other words, the similarity between the counterfactual representation and the original causal representation is continuously pulled in, which is also the process of mining causal invariance. Learning the underlying causal invariance mechanism can help the model achieve successful generalization in different environments.

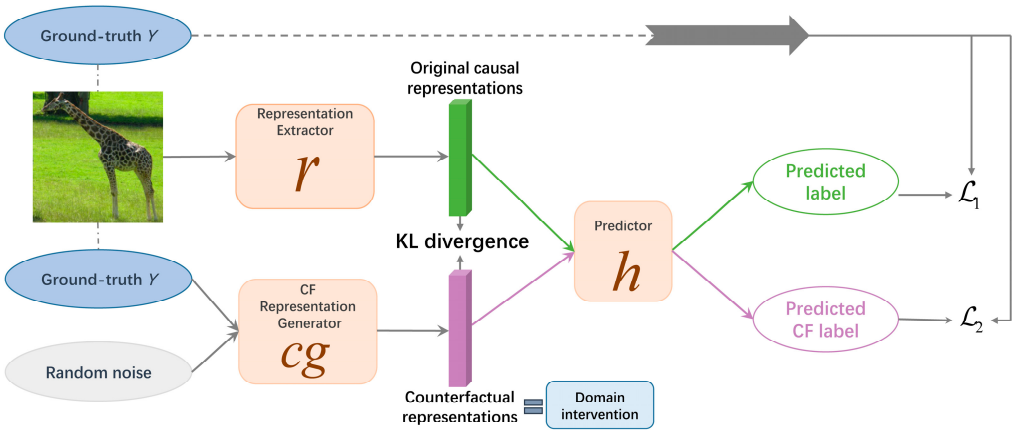

Figure 2. The overall learning framework of CSRDN. It mainly includes three modules, namely, the representation extractor r, the counterfactual representation generator cg, and the predictor h.

3.2.1. Counterfactual Representation Generator

In this subsection, we specifically introduce the training process of the counterfactual representation generator cg. In the learning process of CSRDN, an important part is the generation of counterfactual representations. We need to complete the operation in the latent space. First, we encode X classified as the intervention into the latent space to obtain the corresponding distribution. Secondly, to generate higher-quality counterfactual representations, we make the latent-variable distributions close to each other through an adversarial game between the counterfactual representation generator and the discriminator. We introduce an encoder E, a decoder D, and a discriminator S to assist the training of cg. The training framework is shown in Figure 3.

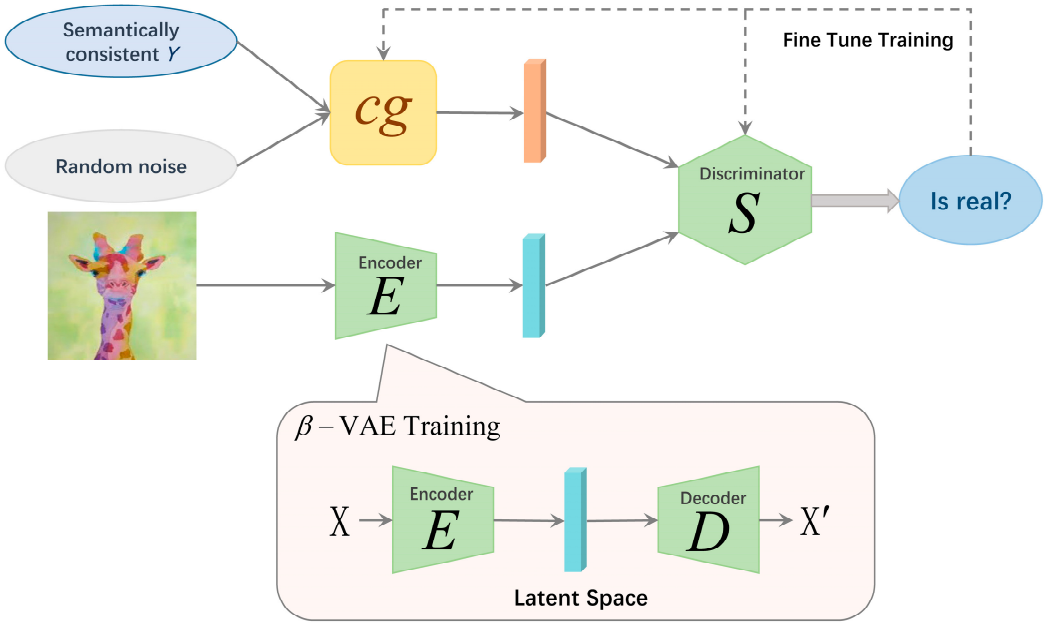

Figure 3. The training framework of counterfactual representation generator cg. It mainly contains three modules, namely, encoder E, decoder D, and discriminator S.

Since the target domain is not visible in the domain generalization task, we cannot obtain samples of the target domain. In order to simplify the calculation process, when there are existing M groups of source domains, we randomly fix one of the domains and treat the remaining M-1 sets of domains as interventions. In actual training, M-1 counterfactual representation generators are created. Specifically, the training of each cg requires the involvement of sets of encoder E and the corresponding decoder D, as well as discriminator S.

Step 1: We generate counterfactual representations via cg, which needs to be manipulated in the latent space. By using the technology of $\beta - VAE$ [43] (Variational Autoencoder, VAE), input X from M-1 source domains as the intervention is sequentially encoded into the latent space, and the corresponding distributions are obtained, respectively. The latent space needs to contain high-level semantic information and be able to reconstruct the original input through a matching decoder, providing prior guarantees for the formal training of cg. Below, the training process from the perspective of one counterfactual representation generator is introduced. $\beta - VAE$ is utilized to approximate the posterior distribution $p(z|x)$, which is a certain extension of VAE, having training stability and stronger decoupling performance. After the latent variable $z = E(x) \sim q(z|x)$ is obtained through E with the latent space, D reconstructs the image to obtain $x' = D(E(x)) = D(z) \sim p(x|z)$. E and D are trained by maximizing the Evidence Lower Bound (ELBO) loss with added coefficient β ($\beta > 1$).

$$\max_{E,D} \mathrm{ELBO} = \mathrm{E}_{q(z|x)}[\log p(x|z)] - \beta \mathrm{KL}[q(z|x)||p(z)] \tag{4}$$

where KL() denotes Kullback–Leibler (KL) divergence [44]. By training the set of $E - D$ well, we can optimize the latent space with superior decoding operations.

Step 2: We choose the encoder E that was trained through Step 1 and fix it, and additionally introduce a discriminator S. The decoder D is muted. At this point, the training of cg is formally started. The input to E is the image of the current source domain treated as the intervention, which can be well encoded into the latent space. Relying on the intervention guarantee of causal semantic consistency, the distribution of semantically consistent Y from the previous fixed single domain that corresponds to the label consistency of the current input of E and random Gaussian noise are simultaneously input to cg, which is utilized to obtain counterfactual representations. Since it is desirable that the distribution of the output of E and the corresponding output of cg can be as close as possible in the latent space to obtain more realistic and comprehensive counterfactual representations, the discriminator S is introduced to conduct an adversarial game (min–max) with cg. Unlike receiving the original image and the generated image, our discriminator discriminates representations according to the distribution. Finally, after the adversarial game, S will be in a chaotic state and cannot effectively distinguish. The technology of Wasserstein GAN (WGAN) [40] is utilized, which uses Wasserstein distance to calculate the difference between generated data and real data, effectively solving the problem of unstable training. We achieve the goal of stably training the optimal generator by min–maximizing the loss.

$$\min_{cg}\max_{S,\|S\|_L\leq 1} L_{\mathrm{WGAN}} = \mathrm{E}_{x\sim p(x)}[S(E(x))] - \mathrm{E}_{(y,u)\sim(p(y)\sim p(u))}[S(cg(y,u))] \quad (5)$$

where $p(x)$ and $p(y)$ represent the distribution of input X from this intervention and semantically consistent label Y from the fixed single source domain, respectively, and $p(u)$ represents the Gaussian distribution of noise. WGAN explicitly adds the Lipschitz constraint to the discriminator and requires the Lipschitz constant not to exceed 1. Through this min–max game, a delicate balance is obtained between cg and S. cg is effectively trained to generate counterfactual representations that are sufficiently realistic and comprehensive.

Based on the above steps, M-1 counterfactual representation generators are trained step by step according to M-1 interventions, capable of generating counterfactual representations directly from semantic labels and random noise to simulate interventions in the domain, thus contributing to the learning of the causal invariant mechanism. The trained cg will be involved in the training of the representation extractor r and the predictor h to implement causal semantic representation learning by counterfactual inference, promoting stable generalization of the model.

3.2.2. Joint Learning Procedure

As shown in Figure 2, we summarize the overall training process following the CSRDN framework. Firstly, we need to train the counterfactual representation generator cg. Through the two steps mentioned above, we train a total of M-1 cg according to the number of source domains. Secondly, trained M-1 cg participates in the training of the representation extractor r and the predictive classifier h in the main module. The input of r is all raw X in source domains, and the output is the original causal representation. The input of the current M-1 counterfactual representation generator is the corresponding ground-truth label Y from the input of the current r and random Gaussian noise, which can generate M-1 sets of counterfactual representations successively. These counterfactual representations are semantically consistent with the input but have different interventions on the domain to mitigate confounding effects, which can provide assistance for stable causal semantic representation learning in OOD situations. The input of h is all the original causal representations and counterfactual representations and the output is the predicted labels and predicted counterfactual labels, respectively. Through utilizing the technology of counterfactual inference [24], we always pursue the above three goals during the training process and are able to propose the final loss function for the training of r and h to optimize the model. The first goal corresponds to the original prediction loss L_1, and the second goal corresponds to the counterfactual prediction loss L_2. The third goal corresponds to

the distributional discrepancy distance loss L_3. For training the representation extractor r and predictor h, the overall optimization objective of our proposed CSRDN is summarized as follows:

$$\begin{aligned} \min_{r,h} L_{all} = \mathrm{E}_{(x,y,u)\sim(p(x),p(y),p(u))} L_1(y, h(r(x))) + \lambda_1 \tfrac{1}{M-1} \textstyle\sum_m^{M-1} L_2(y, h(cg_m(y,u))) \\ + \lambda_2 \tfrac{1}{M-1} \textstyle\sum_m^{M-1} L_3(r(x), cg_m(y,u)) \end{aligned} \tag{6}$$

where λ_1 and λ_2 denote weighting parameters that are utilized to balance the overall loss. We choose L_1 and L_2 as the cross-entropy loss, and L_3 as the KL divergence. During training, CSRDN is able to learn the invariant causal mechanism that enables stable generalization from source to target domains.

4. Experiments

This section describes the experimental part in detail. We conduct experiments and analyze several benchmark datasets to verify the feasibility and superiority of the method and compare CSRDN with a set of other DG methods. Specifically, we set the weighting parameters of the loss function to be dynamic and learnable, which means that we seek Pareto optimality in the training process of the model. We also conduct the experiment to show that adding interventions can improve the learning effect of CSRDN. Meanwhile, we discuss the improvement effect of each goal of CSRDN in the overall learning through ablation experiments and exhibit the results of the visualization.

4.1. Datasets

Rotated MNIST [26] is a synthetic dataset constructed from the MNIST handwritten digit dataset, which is a variant of it. Rotated MNIST is originated from grayscale MNIST handwritten digits performed with six different rotation angles, individually: 0°, 15°, 30°, 45°, 60°, and 75°. These six different rotation angles can also be regarded as six different domains. The images of Rotated MNIST are artificially rotated to achieve domain transformation.

PACS [27] is a widely used dataset of DG, with a total of 9991 images. These images are drawn from four different fields, namely, painting, cartoon, photo, and sketch, and contain seven categories, namely, dog, elephant, giraffe, guitar, house, horse, and person. PACS is considered to have a significant dataset shift, with large differences in image style.

VLCS [28] is a classic dataset of DG, with a total of 10,729 images. These images are drawn from four datasets, namely, PASCAL VOC2007 (V), LabelMe (L), Caltech(C), and SUN09 (S), which can also be regarded as four different domains. These images contain five classes, namely, bird, car, chair, dog, and person. It is worth noting that the images of VLCS are all derived from the real world and have a more realistic domain shift, which is also more challenging in DG tasks.

4.2. Implement Details

In the experimental setup, we follow the commonly used leave-one-domain-out protocol, designating one of the domains as the invisible target domain, while the rest are used as source domains for model training. For instance, for the PACS dataset, we regard cartoon-style images as the unavailable target domain and conduct training on painting, photo, and sketch images. Finally, the trained model is applied in the cartoon domain to obtain prediction results. Briefly, our experiments are performed on three widely used datasets, Rotated MNIST, PACS, and VLCS. By treating each domain of each dataset as an unseen target domain, in turn, the model is trained and tested, and the results are compared with the ERM baseline and a series of DG methods, which are presented in Section 4.3 below. The related analysis experiment on the number of domain interventions is performed on the Rotated MNIST dataset, which is presented in Section 4.4 below. The ablation experiment and visualization results of the PACS dataset are introduced in

Sections 4.6 and 4.7, respectively. Our basic settings follow DomainBed [45]. For the PACS and VLCS datasets, we use ResNet50 [1] as the backbone and MNIST ConvNet [45] based on the smaller Convolutional Neural Networks (CNN) architecture for Rotated MNIST. We resize PACS and VLCS datasets to 224 × 224 pixels and Rotated MNIST to 28 × 28 pixels. CSRDN is implemented by Python 3.6 and PyTorch 1.10.0 and we use 4 NVIDIA GEFORCE GTX 1080Ti 11G graphics cards (manufacturer: ASUS, location: Shanghai, China) for training and testing on Ubuntu 20.04. The model is trained by using an SGD optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4}. The batch size is 64 for the Rotated MNIST dataset and 32 for the PACS and VLCS datasets, and the training epoch is set to 50 with a 0.001 initial learning rate. Accuracy is the main evaluation indicator being compared for this part of the experiment. We run experiments on each dataset three times, and all results are represented according to the average accuracy of three runs. Specifically, the higher prediction accuracy obtained when testing in the target domain means that the model trained only from the source domains can better resist the challenges brought by domain shift, representing better OOD generalization performance of the model.

4.3. Experimental Results

We compare CSRDN with a large number of other methods for DG, including the baseline Empirical Risk Minimization (ERM) [46], Invariant Risk Minimization (IRM) [47], Inter-domain Mixup (Mixup) [48], Maximum Mean Discrepancy (MMD) [9], Marginal Transfer Learning (MTL) [49], Meta-Learning Domain Generalization (MLDG) [14], Domain Adversarial Neural Network(DANN) [50], Deep CORAL (CORAL) [51], Group Distributionally Robust Optimization (GroupDRO) [52], the causality-related method (MatchDG) [20], the regularization-related method for invariant gradient variances (Fishr) [53], Representation Self-Challenging (RSC) [54], Style Agnostic Networks (SagNet) [55], and Exact Feature Distribution Mixing (EFDMix) [56].

4.3.1. Results of the Rotated MNIST Dataset

The results of the Rotated MNIST dataset are presented in Table 1. As shown in Table 1, the average accuracy of our proposed method is 98.4%, beating all other methods, and is 0.3% higher than the second place and 0.4% higher than baseline. At the same time, we also achieve optimal generalization performance at 45°, 60°, and 75° test domains, which are 0.2%, 0.2%, and 0.6% higher than the baseline, respectively. Since Rotated MNIST is a synthetic dataset, the domain shift is caused by artificial rotation, so there is no significant style difference in domain changes. In this case, no matter which domain is chosen as the target domain, the average accuracy of all methods is higher than 95%. CSRDN pursues stable causal representation learning, which can tap into the core causal mechanism as much as possible and resist artificially brought rotational changes.

Table 1. Leave-one-domain-out results of the Rotated MNIST dataset from 0° to 75° (accuracy in %). Each column name indicates the target domain. The best results are expressed in bold.

Methods	**0°**	**15°**	**30°**	**45°**	**60°**	**75°**	**Avg**
ERM	95.6	**99.0**	98.9	99.1	99.0	96.7	98.0
IRM	95.9	98.9	99.0	98.9	98.9	95.6	97.9
Mixup	96.1	99.1	98.9	99.0	99.0	96.6	98.1
MMD	**96.6**	98.9	98.9	99.1	99.0	96.2	98.1
MLDG	95.9	98.9	99.0	99.1	99.0	96.0	98.0
DANN	95.6	98.9	98.9	99.0	98.9	95.9	97.9
CORAL	95.7	**99.0**	99.1	99.1	99.0	96.7	98.1
GroupDRO	95.9	98.9	99.0	99.0	99.0	96.9	98.1
MatchDG	95.9	98.4	98.6	98.9	98.7	95.1	97.6
Fishr	95.0	98.5	**99.2**	98.9	98.9	96.5	97.8
CSRDN (ours)	96.5	98.8	99.1	**99.3**	**99.2**	**97.3**	**98.4**

4.3.2. Results of the PACS Dataset

The results of the PACS dataset are presented in Table 2. As shown in Table 2, the average accuracy of CSRDN is 88.8%, which outperforms all other methods and is 0.9% higher than suboptimal RSC, far superior to the ERM baseline by 3.3%. Not only does it achieve the best result in overall accuracy, but CSRDN also achieves the highest prediction accuracy when testing the cartoon and sketch domains. Sketch's style is very different from the other three datasets, which makes it more of a challenge to test, yet CSRDN leads the way with 84.9% accuracy, well ahead of the baseline of 5.6%. The difference between domains of PACS is mainly presented as the style transfer, which means that the domain shift is more significant than that of the synthetic dataset, and the accuracy of all methods is relatively low. At this challenging moment, the core superiority of our method is revealed compared to others. CSRDN has superior stable generalization capabilities based on learned invariant causal mechanisms that can be resistant to different stylistic variations.

Table 2. Leave-one-domain-out results of the PACS dataset (accuracy in %). Each column name indicates the target domain. The best results are expressed in bold.

Methods	P	A	C	S	Avg
ERM	97.2	84.7	80.8	79.3	85.5
IRM	96.7	84.8	76.4	76.1	83.5
Mixup	97.6	86.1	78.9	75.8	84.6
Fishr	97.0	88.4	78.7	77.8	85.5
SagNet	97.1	87.4	80.7	80.0	86.3
RSC	97.9	87.8	82.1	83.8	87.9
MMD	96.6	86.1	79.4	76.5	84.6
MLDG	97.4	85.5	80.1	76.6	84.9
DANN	97.3	86.4	77.4	73.5	83.6
CORAL	97.5	88.3	80.0	78.8	86.2
GroupDRO	96.7	83.5	79.1	78.3	84.4
EFDMix	**98.1**	**90.6**	82.5	76.4	86.9
MatchDG	97.9	85.6	82.1	78.8	86.1
CSRDN (ours)	97.5	88.3	**84.5**	**84.9**	**88.8**

4.3.3. Results of the VLCS Dataset

The results of the VLCS dataset are presented in Table 3. As shown in Table 3, CSRDN achieves an average accuracy of 79.3%, which shows the best performance and far exceeds the ERM baseline by 1.8%. Especially, when LabelMe is tested as the unseen domain, the prediction accuracy of all methods does not exceed 70%, but CSRDN still outperforms the second place CORAL by 0.9%. It fully demonstrates the superiority of CSRDN in the face of real and diverse domain changes. Since our method can create generative domain interventions through counterfactual inference, it can excavate stable causal representations and adapt to complex environmental disturbances.

Table 3. Leave-one-domain-out results of the VLCS dataset (accuracy in %). Each column name indicates the target domain. The best results are expressed in bold.

Methods	V	L	C	S	Avg
ERM	74.6	64.3	97.7	73.4	77.5
IRM	77.3	64.9	98.6	73.4	78.5
RSC	75.6	62.5	97.9	72.3	77.1
SagNet	**77.5**	64.5	97.9	71.4	77.8
Fishr	76.8	64.0	98.9	71.5	77.8
Mixup	74.3	64.8	98.3	72.1	77.4
MMD	75.3	64.0	97.7	72.8	77.5
MLDG	75.3	65.2	97.4	71.0	77.2
DANN	77.2	65.1	**99.0**	73.1	78.6

Table 3. *Cont.*

Methods	V	L	C	S	Avg
CORAL	**77.5**	66.1	98.3	73.4	78.8
GroupDRO	76.7	63.4	97.3	69.5	76.7
CSRDN (ours)	77.1	**67.0**	98.8	**74.2**	**79.3**

4.4. Results of the Number of Interventions

On the basis of the above, we conduct another interesting experiment. When the number of interventions increases, the training effect of the model is better, and the results are shown in Table 4. We fix the 75° domain of Rotated MNIST as the unseen target domain to test the generalization ability, while the 0° and 15° domains are the initial source domains and sequentially increase the number of source domains for model training. For the intuitiveness of the experiment, we define the 0° domain as the fixed domain for training counterfactual representation generators. Increasing the number of source domains can be seen as a means of increasing interventions. According to Table 4, every time an intervention is added, the average accuracy in the target domain is relatively improved, and the OOD generalization ability of the model is enhanced. When there is only one intervention, we train only one counterfactual representation generator, CSRDN achieves an average accuracy of 95.9%, but as the intervention increases to 4, the accuracy improves to 97.3%. It demonstrates the importance of source domain diversity. The more source domains, the more interventions we can implement, and the model can better focus on stable causal semantic representations with stronger learning ability.

Table 4. Results of the Rotated MNIST dataset for the number of interventions (accuracy in %). "√" indicates that the current domain is the available source domain. The 75° domain is the target domain (TD). The best result is expressed in bold.

Methods	0°	15°	30°	45°	60°	75° (TD)
Variant 1	√	√				95.9
Variant 2	√	√	√			96.5
Variant 3	√	√	√	√		97.1
Variant 4	√	√	√	√	√	**97.3**

4.5. Pareto Optimality

Our total loss function consists of three loss functions, and during the training process of the model, it is necessary to seek parameter optimization by continuously minimizing the total loss function. Inspired by Ref. [57], for CSRDN, we view it as a multi-objective optimization problem, not fixing the weighting parameters λ_1 and λ_2 but viewing them as dynamic and learnable. During the training process of the model, there is a high chance that the three loss functions will conflict and have a competitive relationship. Instead of choosing a weighted linear combination of the loss functions, we pursue Pareto optimization with the help of convex optimization to maximize the optimization performance of the activated model. Our approach utilizes MGDA-UB to reduce the expensive computational cost and seeks the upper bound of MGDA for model optimization during training. With Pareto optimization, the performance of our model is further improved, and optimal results are achieved on all three datasets.

4.6. Ablation Experiment

Based on the above experimental results, we further explored the effect of each loss of CSRDN on the entire training process. The results of CSRDN ablation experiments on the PACS dataset are shown in Figure 4. When the trained model only utilizes the original prediction loss L_1, the generative domain interventions do not participate in the overall optimization, which is not different from the common ERM method, and the average accuracy is 85.5%, the lowest among all variants. When loss L_1 and loss L_3 are involved

in model training at the same time, as shown in Variant 3, the average accuracy is 88.0%, which is 1.5% higher than the combination of L_1 and L_2, as shown in Variant 2. This suggests that only guaranteed generation of counterfactual representations on the basis of L_1 is insufficient to compete with Variant 3, which includes mining its intrinsic causal semantic invariance. L_3 focuses on balancing the distribution similarity of representations under different interventions, which is the core guarantee link for OOD generalization in the process of model training. When L_2 is combined with L_3 as seen in Variant 4, the average accuracy is 0.7% lower than that of Variant 3, which means that including L_3 in the overall training can guarantee a certain performance improvement under the premise that the counterfactual representation generator is well trained. Importantly, when L_1, L_2, and L_3 all participate in model training, CSRDN achieves the best performance, with an accuracy of 88.8%. It can be seen that every part of the loss for each counterfactual inference goal is crucial to the performance improvement of the model, and all three are indispensable. These three modules of the loss complement and promote each other, jointly improving the performance.

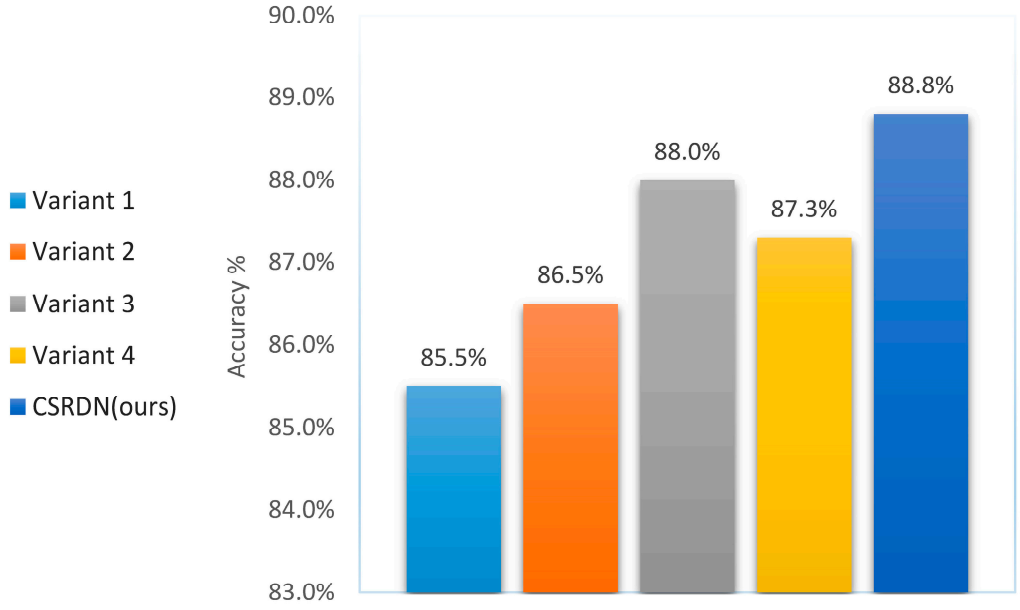

Figure 4. Ablation experiment of CSRDN on the PACS dataset. Variant 1: L_1; Variant 2: $L_1 + L_2$; Variant 3: $L_1 + L_3$; Variant 4: $L_2 + L_3$; CSRDN: $L_1 + L_2 + L_3$.

4.7. Visualization for Class Activation Map

To further visualize the superiority of CSRDN, we use the visualization technique of Gradient-weighted Class Activation Mapping (GradCAM) [58] to generate a set of attention maps for the ERM baseline and CSRDN, respectively. We select the cartoon domain on the PACS dataset as the test dataset for visualization generation. The results of the visualization are shown in Figure 5. The attention range of the causal semantics of the image is used as our evaluation indicator. It can be clearly seen that our method implements stable causal representation learning through different causal interventions based on counterfactual inference, and is able to seek true semantics in generalization tasks. For instance, in the classification of "person", the CSRDN's attention is focused on the faces of people with causal semantics, which provides stable discriminative information for the classification task. The ERM baseline, on the other hand, focuses on irrelevant factors, such as the background and texture of the image, which adversely affects the classification task. It suggests that CSRDN can help the model learn the core causal mechanism and extract stable causal semantic representations, giving it superior OOD generalization capabilities against perturbations brought by different domains.

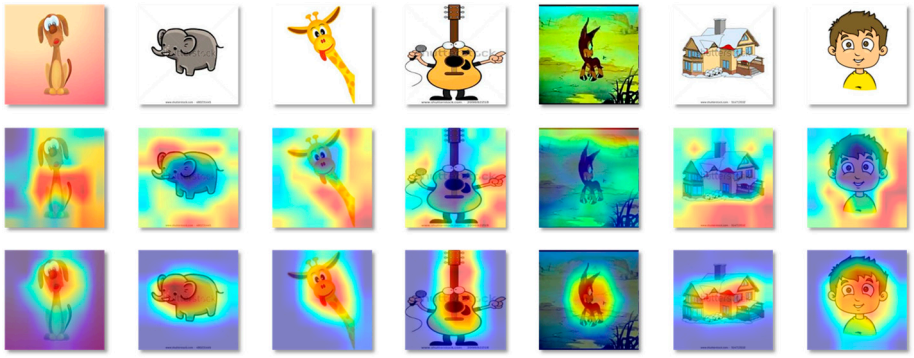

Figure 5. Visualization of attention maps using GradCAM in the PACS dataset. The first row represents original images with true labels dog, elephant, giraffe, guitar, horse, house and person (from left to right), while the second row corresponds to the baseline and the third row corresponds to our CSRDN.

5. Conclusions

In this paper, we shed light on the shortcomings of statistical models relying on spurious correlations in dealing with OOD problems and present a novel causal perspective on domain generalization, and the purpose and task are to improve the stable generalization ability of the DG model by implementing causal semantic representation learning through domain intervention. Based on the data generation process in natural environments, we construct the inclusive causal graph via SCM, which can be adapted to a variety of DG tasks. We point out that causal semantics are invariant across domains, and the core lies in mining the intrinsic causal invariance mechanism. A novel framework of CSRDN is proposed, utilizing generated counterfactual representations for different domain interventions, which can help the model learn cross-domain causal relationships and achieve robust generalization. Comprehensive experiments demonstrate the effectiveness and superiority of CSRDN. The proposed method can inject the prospective mind of causal learning into domain generalization and break the deadlock of the insufficient generalization ability of statistical modeling. Our CSRDN focuses on the standard domain generalization problem, that is, the multi-source domain setting. The generation of counterfactual representations of our method benefits from the diversity of data in source domains, which enables the corresponding interventions to be defined. In future work, we will consider the special single-source domain setting, that is, only one source domain can be obtained during model training. In this setting, the training of the counterfactual representation generator in the current method will be limited. Based on the pursuit of future development, we will consider extending the data distribution of a single source domain through a series of effective data augmentation methods that can safely preserve semantics to leverage counterfactual representations as domain interventions for counterfactual inference. Based on the proposed inclusive causal graph, under this special setting and with a positive outlook, the model will still perform causal semantic representation learning and pursue stable causal invariance.

Author Contributions: Conceptualization, Y.S.; methodology, Y.S.; software, Y.S. and S.W.; validation, Y.S.; formal analysis, Y.S. and S.W.; investigation, Y.S.; resources, W.Z.; data curation, Y.S.; writing—original draft preparation, Y.S.; writing—review and editing, Y.S., W.Z. and S.W.; visualization, Y.S.; supervision, W.Z.; project administration, W.Z.; funding acquisition, W.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Natural Science Foundation of Shandong Province of China (No. ZR2022MF320) and the National Defense Science and Technology 163 Program Project of China (No. 20-163-***-***).

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Nevada, USA, 26–30 June 2016; pp. 770–778.
2. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
3. Li, S.; Xie, B.; Lin, Q.; Liu, C.H.; Huang, G.; Wang, G. Generalized domain conditioned adaptation network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 4093–4109. [CrossRef] [PubMed]
4. Lv, F.; Liang, J.; Gong, K.; Li, S.; Liu, C.H.; Li, H.; Liu, D.; Wang, G. Pareto domain adaptation. In Proceedings of the Advances in Neural Information Processing Systems, Virtual-only, 6–14 December 2021; pp. 12917–12929.
5. Wang, M.; Deng, W. Deep visual domain adaptation: A survey. *Neurocomputing* **2018**, *312*, 135–153. [CrossRef]
6. Recht, B.; Roelofs, R.; Schmidt, L.; Shankar, V. Do imagenet classifiers generalize to imagenet? In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 5389–5400.
7. Taori, R.; Dave, A.; Shankar, V.; Carlini, N.; Recht, B.; Schmidt, L. Measuring robustness to natural distribution shifts in image classification. In Proceedings of the Advances in Neural Information Processing Systems, Virtual-only, 6–12 December 2020; pp. 18583–18599.
8. Zhang, X.; Cui, P.; Xu, R.; Zhou, L.; He, Y.; Shen, Z. Deep stable learning for out-of-distribution generalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 5372–5382.
9. Li, H.; Pan, S.J.; Wang, S.; Kot, A.C. Domain generalization with adversarial feature learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5400–5409.
10. Chen, Y.; Wang, Y.; Pan, Y.; Yao, T.; Tian, X.; Mei, T. A style and semantic memory mechanism for domain generalization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 9164–9173.
11. Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; Loy, C.C. Domain generalization: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 4396–4415. [CrossRef] [PubMed]
12. Li, Y.; Tian, X.; Gong, M.; Liu, Y.; Liu, T.; Zhang, K.; Tao, D. Deep domain generalization via conditional invariant adversarial networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 624–639.
13. Shao, R.; Lan, X.; Li, J.; Yuen, P.C. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 10023–10031.
14. Li, D.; Yang, Y.; Song, Y.Z.; Hospedales, T. Learning to generalize: Meta-learning for domain generalization. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LI, USA, 2–7 February 2018.
15. Dou, Q.; Coelho de Castro, D.; Kamnitsas, K.; Glocker, B. Domain generalization via model-agnostic learning of semantic features. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019.
16. Shi, Y.; Yu, X.; Sohn, K.; Chandraker, M.; Jain, A.K. Towards universal representation learning for deep face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 6817–6826.
17. Volpi, R.; Namkoong, H.; Sener, O.; Duchi, J.C.; Murino, V.; Savarese, S. Generalizing to unseen domains via adversarial data augmentation. In Proceedings of the Advances in Neural Information Processing Systems, Montréal, QC, Canada, 3–8 December 2018.
18. Peters, J.; Janzing, D.; Schölkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms*; The MIT Press: Cambridge, MA, USA, 2017; ISBN 978-0-262-03731-0.
19. Pearl, J. *Causality: Models, Reasoning and Inference*, 2nd ed.; Cambridge University Press: New York, NY, USA, 2009; ISBN 978-0-521-89560-6.
20. Mahajan, D.; Tople, S.; Sharma, A. Domain generalization using causal matching. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 18–24 July 2021; pp. 7313–7324.
21. Müller, J.; Schmier, R.; Ardizzone, L.; Rother, C.; Köthe, U. Learning robust models using the principle of independent causal mechanisms. In Proceedings of the DAGM German Conference on Pattern Recognition, Bonn, Germany, 28 September–1 October 2021; pp. 79–110.
22. Christiansen, R.; Pfister, N.; Jakobsen, M.E.; Gnecco, N.; Peters, J. A causal framework for distribution generalization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 6614–6630. [CrossRef] [PubMed]
23. Rojas-Carulla, M.; Schölkopf, B.; Turner, R.; Peters, J. Invariant models for causal transfer learning. *J. Mach. Learn. Res.* **2018**, *19*, 1–34.
24. Johansson, F.; Shalit, U.; Sontag, D. Learning representations for counterfactual inference. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 3020–3029.
25. Bottou, L.; Peters, J.; Quiñonero-Candela, J.; Charles, D.X.; Chickering, D.M.; Portugaly, E.; Ray, D.; Simard, P.; Snelson, E. Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising. *J. Mach. Learn. Res.* **2013**, *14*.

26. Ghifary, M.; Kleijn, W.B.; Zhang, M.; Balduzzi, D. Domain generalization for object recognition with multi-task autoencoders. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2551–2559.
27. Li, D.; Yang, Y.; Song, Y.Z.; Hospedales, T.M. Deeper, broader and artier domain generalization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5542–5550.
28. Fang, C.; Xu, Y.; Rockmore, D.N. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 1657–1664.
29. Ghifary, M.; Balduzzi, D.; Kleijn, W.B.; Zhang, M. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1414–1430. [CrossRef] [PubMed]
30. Motiian, S.; Piccirilli, M.; Adjeroh, D.A.; Doretto, G. Unified deep supervised domain adaptation and generalization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5715–5725.
31. Shankar, S.; Piratla, V.; Chakrabarti, S.; Chaudhuri, S.; Jyothi, P.; Sarawagi, S. Generalizing Across Domains via Cross-Gradient Training. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
32. Carlucci, F.M.; Russo, P.; Tommasi, T.; Caputo, B. Hallucinating agnostic images to generalize across domains. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27–28 October 2019; pp. 3227–3234.
33. Finn, C.; Abbeel, P.; Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1126–1135.
34. Lewis, D. Causation. *J. Philos.* **1974**, *70*, 556–567. [CrossRef]
35. Schölkopf, B.; Locatello, F.; Bauer, S.; Ke, N.R.; Kalchbrenner, N.; Goyal, A.; Bengio, Y. Toward Causal Representation Learning. *Proc. IEEE* **2021**, *109*, 612–634. [CrossRef]
36. Shalit, U.; Johansson, F.D.; Sontag, D. Estimating individual treatment effect: Generalization bounds and algorithms. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 3076–3085.
37. Hassanpour, N.; Greiner, R. Counterfactual Regression with Importance Sampling Weights. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Macao, China, 10–16 August 2019; pp. 5880–5887.
38. Kallus, N. Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. In Proceedings of the International Conference on Machine Learning, Vienna, Austria, 12–18 July 2020; pp. 5067–5077.
39. Yang, M.; Liu, F.; Chen, Z.; Shen, X.; Hao, J.; Wang, J. Causalvae: Disentangled representation learning via neural structural causal models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9593–9602.
40. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein generative adversarial networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 214–223.
41. Peters, J.; Bühlmann, P.; Meinshausen, N. Causal inference by using invariant prediction: Identification and confidence intervals. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2016**, *78*, 947–1012. [CrossRef]
42. Magliacane, S.; Van Ommen, T.; Claassen, T.; Bongers, S.; Versteeg, P.; Mooij, J.M. Domain adaptation by using causal inference to predict invariant conditional distributions. In Proceedings of the Advances in Neural Information Processing Systems, Montréal, QC, Canada, 2–8 December 2018.
43. Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.P.; Glorot, X.; Botvinick, M.M.; Mohamed, S.; Lerchner, A. Beta-vae: Learning basic visual concepts with a constrained variational framework. In Proceedings of the International Conference on Learning Representations (Poster), Toulon, France, 24–26 April 2017.
44. Johnson, D.H.; Sinanovic, S. Symmetrizing the kullback-leibler distance. *IEEE Trans. Inf. Theory* **2001**, *1*, 1–10.
45. Gulrajani, I.; Lopez-Paz, D. In search of lost domain generalization. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
46. Vapnik, V. *The nature of Statistical Learning Theory*; Springer Science & Business Media: Heidelberg, Germany, 2013; ISBN 978–0-387-94559-0.
47. Arjovsky, M.; Bottou, L.; Gulrajani, I.; Lopez-Paz, D. Invariant risk minimization. *arXiv* **2019**, arXiv:1907.02893.
48. Yan, S.; Song, H.; Li, N.; Zou, L.; Ren, L. Improve unsupervised domain adaptation with mixup training. *arXiv* **2020**, arXiv:2001.00677.
49. Blanchard, G.; Deshmukh, A.A.; Dogan, U.; Lee, G.; Scott, C. Domain generalization by marginal transfer learning. *J. Mach. Learn. Res.* **2021**, *22*, 1–55.
50. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; Lempitsky, V. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **2016**, *17*, 1–35.
51. Sun, B.; Saenko, K. Deep coral: Correlation alignment for deep domain adaptation. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 443–450.
52. Sagawa, S.; Koh, P.W.; Hashimoto, T.B.; Liang, P. Distributionally robust neural networks. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
53. Rame, A.; Dancette, C.; Cord, M. Fishr: Invariant gradient variances for out-of-distribution generalization. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 18347–18377.

54. Huang, Z.; Wang, H.; Xing, E.P.; Huang, D. Self-Challenging improves cross-domain generalization. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 124–140.
55. Nam, H.; Lee, H.; Park, J.; Yoon, W.; Yoo, D. Reducing domain gap by reducing style bias. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, WA, USA, 20–25 June 2021; pp. 8690–8699.
56. Zhang, Y.; Li, M.; Li, R.; Jia, K.; Zhang, L. Exact Feature distribution matching for arbitrary style transfer and domain generalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–23 June 2022; pp. 8035–8045.
57. Sener, O.; Koltun, V. Multi-task learning as multi-objective optimization. In Proceedings of the Advances in neural information processing systems, Montréal, QC, Canada, 2–8 December 2018.
58. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-Cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.