Causal Structure-guided Distributionally Robust Optimization under Domain Shifts

Seonggyeom Kim*
Dept. of Artificial Intelligence
Hanyang University
Seoul, South Korea
canino@hanyang.ac.kr

Eunjung Choi*
Dept. of Artificial Intelligence
Hanyang University
Seoul, South Korea
sinegi@hanyang.ac.kr

Dong-Kyu Chae
Dept. of Computer Science
Hanyang University
Seoul, South Korea
dongkyu@hanyang.ac.kr

Abstract

Domain generalization (DG) aims to learn predictive models from multiple source domains that maintain performance on unseen target domains. *Distributionally robust optimization* (DRO) addresses distribution shift by minimizing the worst-case risk over an uncertainty set of plausible test distributions. However, if this uncertainty set is overly large, it may include unrealistic shifts, leading to low-confidence predictions. To address this issue, we propose Causal Structure-guided DRO (**CS-DRO**), which estimates a *directed acyclic graph* (DAG) that encodes the predictive relationships between representations and labels, serving as a proxy for causal structure shared across source domains. Using gradient-based signals, we quantify how well the candidate distributions explored during optimization maintain the estimated structure and impose a structure-preserving constraint on the uncertainty set. The resulting DRO objective is reformulated via Lagrangian relaxation into a tractable primal–dual learning problem. Experiments on standard DG benchmarks show that our method achieves competitive performance compared to state-of-the-art methods while improving robustness to unseen target domains. Our Github repository including code and supplementary material is here: https://github.com/gyeomo/CS-DRO.

CCS Concepts

• **Computing methodologies** → **Machine learning**; • **Theory of computation** → **Continuous optimization**.

Keywords

Domain generalization, Distribution shift, Distributionally robust optimization, causal structure learning

ACM Reference Format:
Seonggyeom Kim, Eunjung Choi, and Dong-Kyu Chae. 2026. Causal Structure-guided Distributionally Robust Optimization under Domain Shifts. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3770855.3818154

*Both authors contributed equally to this research.

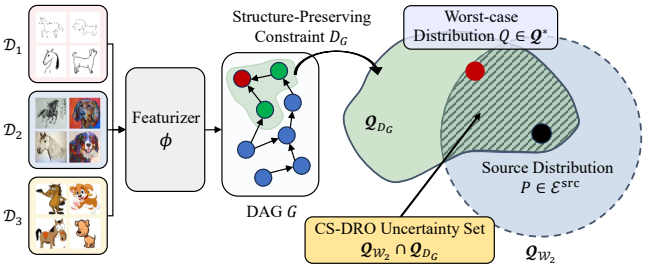


Figure 1: Conceptual illustration of CS-DRO. Q_{W_2} denotes the standard geometric (Wasserstein) uncertainty set, and Q_{D_G} denotes the set defined by the proposed structure-preserving constraint D_G. CS-DRO optimizes within the intersection region, thereby preventing unrealistic shifts.

1 Introduction

Deep Neural Networks demonstrate remarkable performance when training and target domains share similar distributions. In real-world scenarios, this assumption is often violated, leading to performance degradation under distribution shifts. Domain Generalization (DG) aims to learn predictive models from multiple sources that can generalize to out-of-distribution data from unseen domains.

In this context, Distributionally Robust Optimization (DRO) has emerged as a promising approach to handle distribution discrepancy by ensuring generalization performance against unknown distributional shifts [18, 32, 42]. DRO defines an uncertainty set centered on the observed empirical distribution that covers possible target data distributions, and attains robustness by minimizing the worst-case risk over this set. However, to capture potential target distributions, the uncertainty set often becomes overwhelmingly large, leading to low-confidence predictions [11, 17]. An overly large uncertainty set may include unrealistic scenarios that are semantically unrelated to the training data. Optimizing a model for such unrealistic shifts can lead it to a suboptimal solution that produces near-uniform predictions across classes, ultimately harming generalization performance [33].

Therefore, the uncertainty set should be guided so as to contain only potential distributional shifts that can realistically occur [6, 52]. Prior works attempt to achieve this by preserving only the stable correlations present in the source data [11, 33], or by imposing topology-based constraints [51]. However, such approaches can miss plausible shifts that share semantics with the source domains, and they may still fail to exclude certain unrealistic shifts. Another line of work tackles this issue by integrating additional constraints

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
KDD '26, Jeju Island, Republic of Korea
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3818154

reflecting causal relationships into the optimal transport (OT) formulation [9, 15]. These methods, however, either do not provide a practical computational solution [9] or rely on strong assumptions such as requiring pre-defined causal structures and access to the functional forms that formalize the underlying causal mechanisms [15]. Such requirements are infeasible in practice, limiting their empirical scope to low-dimensional numerical data with known causal structures; their applicability to high-dimensional visual tasks remains largely unexplored.

To bridge this gap without relying on external priors, we start from the property of DG that source domains share the same label space and that the label-generating mechanism remains invariant across domains. From this perspective, while domain-specific features may vary under domain shifts, the *causal mechanism* between the input and the label is assumed to remain invariant [36, 41, 60]. Therefore, the uncertainty set should be designed to include distributional information that preserves such underlying causal mechanism across domains so as to leverage the generalization ability of DRO. Herein, we propose Causal Structure-Guided Distributionally Robust Optimization (**CS-DRO**) to address domain shifts.

Specifically, we estimate a *directed acyclic graph* (DAG) in the latent representation space to encode the directed dependencies between variables, serving as a proxy for the common causal structure across source domains. During training, we utilize differentiable gradient-based signals to quantify the structural consistency of the explored distributions with the estimated DAG. Based on this measure, we impose a structure-preserving constraint on the uncertainty set accordingly. Consequently, our approach constrains variations in label-generating features while permitting greater flexibility in non-causal ones, thereby guiding the exploration of potential distributions that are semantically compatible with the source domains. We reformulate the constrained objective as a tractable primal–dual problem using Lagrangian relaxation that supports efficient end-to-end training.

Beyond its practical advantages, we provide a theoretical analysis demonstrating that our structure-preserving uncertainty set ensures target inclusion under mild assumptions. We demonstrate through extensive analyses that these theoretical results remain valid under practical, finite-sample settings. Experimental results on standard DG benchmarks show that our method achieves superior performance on unseen target domains compared to recent state-of-the-art baselines.

2 Preliminaries

Throughout the paper, several additional details and analyses are provided in our **supplementary material** (**link is here**: https://github.com/gyeomo/CS-DRO/blob/main/Supplementary_Material.pdf). Its overall structure is outlined in Appendix.

2.1 Notation and Data-Generating Process

Let $\mathcal{X}$ denote the input space and $\mathcal{Y} = \{1, \dots, C\}$ the label space. For each domain e, let (X^e, Y^e) be random variables taking values in $\mathcal{X}\times\mathcal{Y}$. We denote by P^e_{XY} the joint distribution of (X^e, Y^e) on $\mathcal{X}\times\mathcal{Y}$, and we will often abbreviate it as P^e when there is no ambiguity. Given K source domains, for each domain $e \in \mathcal{E}^{\text{src}}$, we assume that $(X^e_i, Y^e_i) \overset{\text{i.i.d.}}{\sim} P^e_{XY},\ i = 1, \dots, n_e$, and the observed labeled dataset is given by the realizations $\mathcal{D}^e = \{(x^e_i, y^e_i)\}_{i=1}^{n_e}$. The goal of DG is to learn a predictive model $f_\theta : \mathcal{X} \to \mathbb{R}^C$ parameterized by $\theta \in \Theta$ on given source domains such that it generalizes well to an unseen target domain $T \notin \mathcal{E}^{\text{src}}$. We decompose the model f_θ into a featurizer $\phi : \mathcal{X} \to \mathcal{Z}$ and a classifier $w : \mathcal{Z} \to \mathbb{R}^C$, such that $f_\theta = w \circ \phi$, where $\mathcal{Z} \subseteq \mathbb{R}^d$ represents the latent feature space.

Data-Generating Process. Following prior works [23, 35], we assume a data-generating process between the input and the label. To treat the label as a continuous variable, we introduce an auxiliary variable denoted by $\tilde{Y} \in \mathbb{R}$. For each domain $e \in \mathcal{E}^{\text{src}} \cup \{T\}$, the input $X^e \in \mathcal{X}$ is generated from a domain-invariant (causal) factor $X_c \in \mathcal{X}_c$ and a domain-specific (spurious) factor $X^e_s \in \mathcal{X}_s$ via a generating mechanism $\mathbf{G} : \mathcal{X}_c \times \mathcal{X}_s \to \mathcal{X}$, such that $X^e = \mathbf{G}(X_c, X^e_s)$. The variables are governed by the following relations:

$$\tilde{Y} = \mathbf{G}_{\tilde{Y}}(X_c) + \epsilon_{\tilde{Y}}, \qquad X^e_s = \mathbf{G}^e_s(\tilde{Y}) + \epsilon^e_s, \tag{1}$$

where $\epsilon_{\tilde{Y}}$ and ϵ^e_s denote mutually independent exogenous noises. As shown in the left term of Eq. (1), the label variable $\tilde{Y}$ is determined solely by X_c through the label-generating mechanism $\mathbf{G}_{\tilde{Y}}$. In contrast, the right term represents the spurious correlation where $\tilde{Y}$ induces domain-specific attributes X^e_s such as background or style during the data collection process. Following [11, 38, 41], we assume that while the dependency between $\tilde{Y}$ and X^e_s learned from the source domains may vary across different domains, the label-generating mechanism $X_c \to \tilde{Y}$ remains invariant. This invariance is intrinsic to the DG problem setting itself, under which a model trained on source domains can generalize to unseen targets.

Under this data-generating process, the featurizer ϕ induces a representation $Z^e = \phi(X^e)$, which we assume can be partitioned into a domain-invariant (causal) component Z_c and a domain-variant (spurious) component Z^e_s [1, 23, 38, 41, 57]. Our work aims to achieve generalization to unseen target domains by encouraging the model to rely on these stable relationships rather than on highly variable spurious signals.

2.2 Distributionally Robust Optimization.

In this work, we consider the *Wasserstein Distributionally Robust Optimization* (WDRO) framework, which defines the uncertainty set using the Wasserstein distance [27].

Let $\mathcal{V}_{XY} := \mathcal{X} \times \mathcal{Y}$, and for a sample (x, y) we denote $v_{xy} := (x, y) \in \mathcal{V}_{XY}$. For any two points $v_{xy}, v'_{xy} \in \mathcal{V}_{XY}$, the transportation cost function $c : \mathcal{V}_{XY} \times \mathcal{V}_{XY} \to [0, \infty]$ is defined as follows:

$$c(v_{xy}, v'_{xy}) = \|x - x'\|_2^2 + \infty \cdot \mathbb{I}(y \neq y'), \tag{2}$$

where the indicator $\mathbb{I}(y \neq y')$ restricts the transport between samples with different labels. The type-2 Wasserstein distance between any two distributions P and Q on $\mathcal{V}_{XY}$ is then defined as [56]:

$$W_2(P, Q) := \inf_{\pi \in \Pi(P,Q)} \left\{ \mathbb{E}_{(v_{xy}, v'_{xy}) \sim \pi} \left[c(v_{xy}, v'_{xy}) \right] \right\}^{\frac{1}{2}}, \tag{3}$$

where $\Pi(P, Q)$ denotes the set of all couplings on $\mathcal{V}_{XY}$ with marginals P and Q. The uncertainty set centered at P is defined to include all distributions within a specified radius $\rho > 0$:

$$\mathcal{Q}_{W_2}(P) = \{\, Q :\ W_2(P, Q) \leq \rho \,\}. \tag{4}$$

In our multi-source setting, we define the nominal distribution P as the mixture distribution over the source domains. Our objective

is to find the model parameters θ that minimize the worst-case risk over all distributions Q within the uncertainty set $\mathcal{Q}_{W_2}(P)$:

$$\min_{\theta\in\Theta}\Big\{R(\theta) := \sup_{Q\in\mathcal{Q}_{W_2}(P)} \mathbb{E}_{(x,y)\sim Q}\big[\ell(f_\theta(x),y)\big]\Big\}, \tag{5}$$

where ℓ denotes the cross-entropy loss $\ell : \mathbb{R}^C \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$. In practice, directly applying DRO in the original high-dimensional input space can be challenging, so following [40], we solve the above DRO problem in the learned representation space $\mathcal{Z}$. Note that the entire model f_θ (both ϕ and w) is trained end-to-end; however, for ease of exposition, we formulate the subsequent derivations using the feature vector $Z = \phi(X)$ instead of the raw input X. Accordingly, we redefine the product space as $\mathcal{V}_{ZY} := \mathcal{Z} \times \mathcal{Y}$ (and we will simply write $\mathcal{V}$ for brevity). Furthermore, P and Q will refer to joint distributions over features and labels (Z, Y), and c, W_2, and the uncertainty set $\mathcal{Q}_{W_2}(P)$ are likewise defined over $\mathcal{V}$.

3 Method

This section presents our Causal Structure-guided DRO (**CS-DRO**). Based on the graphical relationships between variables shared across source domains, we impose a constraint to ensure that candidate distributions explored within the uncertainty set maintain the underlying causal structure. The proposed structure-preserving constraint guides the model to optimize over distributions that preserve mechanisms invariant across domains, thereby making it robust to distribution shifts.

3.1 Domain-Invariant DAG

Under the assumption that the data-generating process remains consistent across multiple source domains, identifying the causal structure that reflects this mechanism provides an effective strategy for domain generalization. We use a DAG learned from the representations as a proxy for this domain-shared structure to guide robust generalization.

Let $\mathbf{U}^e := (Z^e, \tilde{Y}^e)$ denote the random vector in the representation space for each domain e. The causal structure is represented by a DAG $\mathcal{G} = (\mathcal{U}, E)$ consisting of a node set $\mathcal{U} = \{U_1, \ldots, U_{d+1}\}$ and a set of directed edges E. Each edge $(U_j \to U_i) \in E$ indicates that U_j is a direct cause of U_i. Under the Causal Markov Condition, $\mathcal{G}$ stipulates a mechanism where each node is generated by its parent nodes. Accordingly, the joint distribution $P^e(\mathbf{U}^e)$ of each domain e can be factorized as $P^e(\mathbf{U}^e) = \prod_{i=1}^{d+1} P^e\big(U_i^e \mid Pa_{\mathcal{G}}(U_i)\big)$, where $Pa_{\mathcal{G}}(U_i) := \{U_j \in \mathcal{U} : (U_j \to U_i) \in E\}$ denotes the parent set of node U_i. Following prior works [45, 47, 58], we assume these edges may have causal interpretations. We consider a DAG model formulated as linear structural equation models (SEMs). We identify $\mathcal{G}$ with a weighted adjacency matrix $G \in \mathbb{R}^{(d+1)\times(d+1)}$, where a non-zero entry $G_{ji} \neq 0$ signifies the existence of an edge $(U_j \to U_i)$ and $G_{ii} = 0$. Each variable follows the structural equation $U_i^e = G_{:i}^\top \mathbf{U}^e + \epsilon_i^e$, where ϵ_i^e represents the exogenous noise. From the perspective that data in each domain follows a consistent data-generating process, we assume the following structural consistency.

Assumption 3.1 (Existence of a source-shared DAG)**.** There exists a common structure $\mathcal{G}$ that can simultaneously explain the joint distributions $P^e(\mathbf{U}^e)$ of all source domains $e \in \mathcal{E}^{\text{src}}$.

Under this assumption, the DAG G we aim to estimate represents a common structure shared across all source domains. Under the Causal Markov condition, $\mathcal{G}$ encodes each node as being generated from its parents. Therefore, if the mechanism in Eq. (1) induces the directed relation $\tilde{Y} \to Z_s^e$, the reverse edge $Z_s^e \to \tilde{Y}$ is ruled out by the acyclicity of $\mathcal{G}$. As a result, domain-specific variables Z_s^e are excluded from the set of candidate parents for $\tilde{Y}$, while Z_c, which has consistent explanatory power for $\tilde{Y}$ across all domains, is selected as a parent variable. The detailed procedure for DAG learning is provided in Section 3.4.

Remark 3.2 (Proxy nature and limitation). The properties described above are based on the selective characteristics observed when learning structures from multi-domain data [23]. Given that DAG learning relies on observational data, it is difficult to guarantee the identification of every true causal edge or the complete removal of certain spurious edges that consistently appear across domains. We note that recovering a uniquely identifiable causal graph is not our objective; rather, jointly learning across source domains suffices to obtain a DAG that consistently explains y for our purpose. Therefore, in this work we regard the learned DAG G as a proxy for the causal structure, and use it to define the uncertainty set of our method as a guide to ensure that the structural relationships induced by G remain invariant.

3.2 Causal Structure-Preserving Uncertainty Set

3.2.1 Structural Signature via Gated Reconstruction. To ensure that a candidate distribution Q explored within the uncertainty set preserves the causal structure shared in the source domains, we use the fixed DAG G learned from the source data as a guide. We first define a differentiable reconstruction risk to quantify how well a candidate distribution Q preserves the structure encoded by G. We introduce a matrix variable $A \in \mathbb{R}^{(d+1)\times(d+1)}$ of the same shape as G to control the influence of individual edges of G. Using this, we define $\tilde{G} = G \odot A$, where $\odot$ denotes element-wise product. For a distribution Q over V, the reconstruction risk is defined as follows:

$$\mathcal{L}_{\text{rec}}(\tilde{G}; Q) = \mathbb{E}_{(z,y)\sim Q}\Big[\big\|\tilde{v} - \tilde{G}^\top \tilde{v}\big\|_2^2\Big], \quad \tilde{v} := (z, \tilde{y}) = (z, \Gamma(y)), \tag{6}$$

where $\Gamma : \mathcal{Y} \to \mathbb{R}$ is a fixed encoding function used to treat the label as a single scalar variable, denoted by $\tilde{Y} := \Gamma(Y) \in \mathbb{R}$. We employ the gradient of $\mathcal{L}_{\text{rec}}$ at $A = \mathbf{1}$ as the structural signature of distribution Q with respect to G. For linear least-squares errors, $\mathcal{L}_{\text{rec}}(\tilde{G}; Q)$ is convex in A. If $A = 1$ is a local minimum for Q, then by the first-order necessary condition the gradient at this point is zero [54, 66]. Consequently, the norm of the gradient at $A = \mathbf{1}$ provides a quantitative measure of the violation of the stationarity condition for Q when reconstructed based on the structure G.

Definition 3.3 (Structural signature)**.** Given a fixed DAG G, the structural signature under the distribution Q is defined as:

$$M_G(Q) \;:=\; \nabla_A\, \mathcal{L}_{\text{rec}}(\tilde{G}; Q)\Big|_{A=\mathbf{1}} \;\in\; \mathbb{R}^{(d+1)\times(d+1)}. \tag{7}$$

Specifically, as $A = \mathbf{1}$ corresponds to the case where G is applied without modification, $M_G(Q)$ measures how far the G-induced reconstruction is from stationarity under Q. When G explains Q well, the first-order necessary condition is satisfied and $M_G(Q)$ vanishes; otherwise its magnitude reflects the edge-wise extent

to which Q departs from the structure encoded by G. We provide a detailed interpretation of this definition in Section B.3 of our supplementary material.

Since G is estimated from the source distribution P, we can quantify the structural shift by comparing the differences in M_G values between P and any candidate distribution Q.

Definition 3.4 (Structure-preserving discrepancy)**.** Given two distributions P, Q over V, and a fixed DAG G, the structure-preserving discrepancy is defined as:

$$D_G(P, Q) := \|M_G(P) - M_G(Q)\|_F. \tag{8}$$

Given a structural budget $\kappa > 0$, we say that Q is κ-structure-preserving with respect to (P, G) if $D_G(P, Q) \le \kappa$.

$D_G(P, Q)$ directly compares the G-induced signatures of the two distributions, thereby quantifying changes in the G-induced causal structure that a purely geometric distribution discrepancy might fail to capture. For example, even if Q is close to P in terms of W_2, a large $D_G(P, Q)$ would be interpreted as a shift that alters the structure from the perspective of G. In contrast, a small $D_G(P, Q)$ implies that the first-order necessary condition of the G-induced structure is stably maintained across two distributions. We provide a justification in Section B.4 of our supplementary material.

3.2.2 Anisotropic Geometry for Uncertainty Set. To preserve the domain-invariant $Z_c \to Y$ relationship, we introduce an anisotropic cost that aligns the geometry of the uncertainty set with the $Z \to Y$ dependencies encoded in G. We define a mask vector $G_Y \in [0, 1]^d$ from the label-node column of G as $G_Y := \mathrm{sigmoid}(G_{1:d,\, d+1})$. Using this mask, we define a cost function that assigns a higher penalty on shifts in dimensions of Z that correspond to the parents of Y while assigning a relatively lower cost to shifts along other dimensions:

$$c_G\big((v, v')\big) = \|G_Y \odot (z - z')\|_2^2 + \infty \cdot \mathbb{I}(y \neq y'). \tag{9}$$

Let W_G denote the Wasserstein distance induced by c_G. The uncertainty set based on W_G forms an anisotropic geometry that suppresses expansion along directions corresponding to the G-induced $Z \to Y$ structure, thereby preserving that causal structure. We now construct the uncertainty set by combining the structure-preserving discrepancy in Eq. (8) with the cost function in Eq. (9).

Definition 3.5 (Causal structure-preserving uncertainty set)**.** Given a fixed DAG G and source distribution P, the causal structure-preserving uncertainty set with radius ρ and structural budget κ is defined as follows:

$$\boldsymbol{Q}_{W_G, D_G}(P) := \Big\{Q : W_G(P, Q) \le \rho,\ D_G(P, Q) \le \kappa\Big\}. \tag{10}$$

Here, causal structure-preservation means that the DAG G learned from P remains near a local optimum under the candidate distribution Q, with structural discrepancy within κ.

As a result, the proposed uncertainty set allows distributional shifts only in an anisotropic manner and only within a range that does not significantly damage the structural signature induced by G. Based on this, we formulate the robust optimization problem.

3.3 Causal Structure-guided DRO

We now introduce the learning objective of minimizing the worst-case risk within the uncertainty set $Q_{W_G, D_G}(P)$ defined around the source distribution P. The training objective for the model f_θ can be expressed in the following constrained inf–sup formulation:

$$\begin{aligned} \inf_{\theta \in \Theta} \sup_{Q}\ & \mathbb{E}_{(z,y)\sim Q}\big[\ell\big(w(z), y\big)\big] \\ \text{s.t. } & W_G(P, Q) \le \rho, \quad D_G(P, Q) \le \kappa. \end{aligned} \tag{11}$$

In the inner supremum problem , the search for the candidate distribution Q is confined to a region where (i) the representation transport cost from P is at most ρ and (ii) the structural discrepancy is at most κ. By minimizing the worst-case risk over this restricted uncertainty set, the model is trained to be robust to realistic shifts that preserve the DAG G.

However, the inner supremum problem is intractable as it involves optimization over distributions subject to constraints. We therefore reformulate the problem into a tractable form via Lagrangian relaxation. We introduce dual variables $\gamma \ge 0$ and $\beta \ge 0$ corresponding to the two constraints. For a fixed θ and G, by imposing penalties proportional to the constraint violations, $W_G(P, Q) - \rho$ and $D_G(P, Q) - \kappa$, we define the Lagrangian risk for Q as follows:

$$\begin{aligned} R(P, Q, \gamma, \beta; \theta) := \ & \mathbb{E}_{(z,y)\sim Q}[\ell(w(z), y)] \\ & - \gamma\Big(W_G(P, Q)^2 - \rho^2\Big) - \beta\Big(D_G(P, Q) - \kappa\Big), \end{aligned} \tag{12}$$

where γ and β serve as penalty coefficients that regulate the enforcement of the respective constraints. In this way, instead of directly handling the hard constraints of the primal problem, we optimize a single objective augmented with these penalty terms. This formulation leads to a primal–dual approach in which (γ, β, θ) are updated jointly, and the model is guided to train on a worst-case distribution that (approximately) satisfies the constraints.

By maximizing the Lagrangian $R(P, Q, \gamma, \beta; \theta)$ with respect to Q, we indirectly address the inner supremum problem. Then, minimizing with respect to (γ, β, θ) yields the following reformulated primal–dual problem:

$$\begin{aligned} \min_{\substack{\theta \in \Theta, \\ \gamma \ge 0,\ \beta \ge 0}} \Big\{ & \gamma\rho^2 + \beta\kappa + \max_{Q} \big[\mathbb{E}_{(z,y)\sim Q}[\ell(w(z), y)] \\ & - \gamma W_G(P, Q)^2 - \beta D_G(P, Q)\big]\Big\}. \end{aligned} \tag{13}$$

3.4 Training Procedure

Our training procedure is divided into a warm-up stage (to stabilize the model's initial representation and estimate G) and a CS-DRO stage (to fix the obtained G and use it as a guide for robust training).

3.4.1 DAG Learning. Our approach adopts differentiable causal discovery methods to estimate a causal structure in the form of a DAG from observational data [22, 69, 70]. DAG learning is formulated as an optimization problem that maximizes the likelihood of the given data under constraints such as acyclicity and sparsity. Following these methods, we describe a DAG learning procedure tailored to our framework.

As noted in Section 3.1, we perform DAG learning on the representation Z induced by the featurizer ϕ [23]. Since the latent

representation changes during training, we take two specific measures during the warm-up phase to provide a stable environment for DAG estimation: (i) alongside the classification loss ℓ, we apply a Hilbert–Schmidt Independence Criterion (HSIC) [19] based regularization to reduce dependencies between representation dimensions, encouraging them to behave as independent causal variables; (ii) we construct samples based on cumulative statistics so that the DAG is learned from consistent information in the representations. Specifically, for an input mini-batch $B = (x_i, y_i)_{i=1}^{n_B}$, we compute cumulative statistics $\mathcal{N}(\mu_{e,c}, \sigma_{e,c})$ for each domain-label pair (e, c), and sample from these statistics to obtain $\{\bar{v}_i\}_{i=1}^{m} = \{(\bar{z}_i, \bar{y}_i)\}_{i=1}^{m}$, which are used for DAG learning. Since each $\bar{z}_i$ concentrates around a class–domain representative, the linear SEM in Section 3.1 operates as a local approximation around these representatives. Based on these samples, $(G^\top v_i)_{1:d} \in \mathbb{R}^d$ is trained to reconstruct $\bar{z}_i$ while a linear classifier g forms the parent connections for y via G_Y. Integrating the differentiable acyclicity constraint $h(G)$ and an ℓ_1 sparsity constraint, the DAG learning objective is formulated as:

$$\min_{G,g} \frac{1}{m} \sum_{i=1}^{m} \Big(\big\| \bar{z}_i - (G^\top \bar{v}_i)_{1:d} \big\|_2^2 + \ell\big(g(G_Y \odot \bar{z}_i), \bar{y}_i\big) \Big) + \lambda_a h(G) + \lambda_{l_1} \|G\|_1, \tag{14}$$

where $h(G) = tr\big(e^{G \odot G}\big) - (d+1)$, and $h(G) = 0$ indicates that G satisfies acyclicity (see Zheng et al. [69] for a detailed derivation).

3.4.2 Structural Regularization. Although the DAG G learned in the warm-up stage is fixed thereafter, the featurizer ϕ continues to be updated, potentially causing the distribution of the representations to drift away from the structure captured by G. Thus, we regularize the model so that the representation does not deviate excessively from G. For a mini-batch B, let $\hat{P}_B$ be its empirical distribution. We overload notation and write $M_G(B) := M_G(\hat{P}_B)$ and $D_G(B', B) := D_G(\hat{P}_{B'}, \hat{P}_B)$. We define the structural regularization loss for the mini-batch B as:

$$\mathcal{L}_G(B) = \big\| M_G(B) \big\|_F^2. \tag{15}$$

Here, minimizing L_G encourages the feature outputs z_i of featurizer ϕ to remain well-explained by the fixed G. This term is incorporated into the overall objective for f_θ as a weighted sum with the coefficient λ_G.

3.4.3 Stochastic Optimization of the Empirical CS-DRO. For tractability, we replace the population distribution P in CS-DRO objective with the empirical distribution $\hat{P}_n$ of n training samples, and estimate gradients stochastically using mini-batches. Specifically, following the literature [34, 40], we adopt a batch-based stochastic primal–dual optimization. Given a mini-batch $B = \{(z_i, y_i)\}_{i=1}^{n_B} \sim \hat{P}_n$, we approximate the distribution-level inner maximization by optimizing perturbed representations $\{z_i'\}$ within the mini-batch. This batch-level approach is a practical technique that has been widely validated in WDRO frameworks [34, 40]. Concretely, for each mini-batch B, we obtain a perturbed mini-batch $B' = \{(z_i', y_i)\}_{i=1}^{n_B}$ by approximately solving the maximization of the Lagrangian objective in Eq. (13) as follows:

$$B' \approx \arg\max_{B'} \Big[\frac{1}{n_B} \sum_{i=1}^{n_B} \ell\big(w(G_Y \odot z_i'), y_i\big) - \frac{\gamma}{n_B} \sum_{i=1}^{n_B} \big\| G_Y \odot (z_i - z_i') \big\|_2^2 - \beta\, D_G(B', B) \Big]. \tag{16}$$

Following Bui et al. [7], Liu et al. [34], in implementation we approximate the solution of Eq. (16) by performing J steps of gradient ascent on $\{z_i'\}$, yielding an adversarial batch B_J'. Importantly, this procedure is not a pointwise perturbation of individual samples; rather, it implements a mini-batch-level adversary (a restricted class of Q) operating on the batch support set.

In the outer minimization, directly minimizing only the worst-case risk can lead to excessively conservative model updates [68]. To balance the trade-off between robustness and source accuracy, we introduce a surrogate loss $\tilde{\ell}$. This surrogate designed to reduce the classification error on the perturbed representations z_i' while simultaneously preventing excessive prediction drift compared to the original representations z_i. We define the surrogate loss as:

$$\tilde{\ell}(z_i, z_i', y_i) = \ell\big(w(G_Y \odot z_i'), y_i\big) + \lambda_r\, \mathcal{L}_{\text{reg}}(z_i, z_i'), \tag{17}$$

where the first term represents the worst-case classification loss estimated over the batch and the second term $\mathcal{L}_{\text{reg}}$ is a regularization that encourages prediction consistency between the original and perturbed inputs. Let p_i and q_i be the predicted probability distributions derived from z_i and z_i', respectively. We define the regularization term using a bidirectional KL divergence after smoothing the distributions with a temperature parameter $\tau > 1$:

$$\mathcal{L}_{\text{reg}}(z_i, z_i') = D_{\text{KL}}\big(p_i^{(1)} \big\| \lfloor q_i^{(\tau)} \rfloor\big) + D_{\text{KL}}\big(q_i^{(1)} \big\| \lfloor p_i^{(\tau)} \rfloor\big), \tag{18}$$

where $p_i^{(\tau)}$ is the probability distribution obtained by applying temperature τ to the logits of classifier w. The notation $\lfloor \cdot \rfloor$ denotes the stop-gradient operation, which mitigates excessive sharpening of predictions by targeting smoothed distributions and stabilizes the overall training process.

Finally, given a mini-batch $B = \{(z_i, y_i)\}_{i=1}^{n_B}$ and its perturbed version B' obtained via Eq. (16) under the current values of (θ, γ, β), the outer update minimizes the following sample-level objective:

$$\min_{\substack{\theta \\ \gamma \geq 0,\ \beta \geq 0}} \frac{1}{n_B} \sum_{i=1}^{n_B} \tilde{\ell}(z_i, z_i', y_i) + \lambda_G\, \mathcal{L}_G(B) + \gamma \rho^2 + \beta \kappa, \tag{19}$$

where (γ, β) are the dual variables that are constrained to remain non-negative after each update in our implementation. All predictions are performed by applying the $w(G_Y \odot \cdot)$ operation to the representations z_i and z_i'. Consistent with prior work [40] showing that performing DRO in the representation space achieves generalization performance on par with input-space, we adopt this representation-space approach. The training procedure of CS-DRO is summarized in Algorithm 1, and further implementation details are provided in Section G.2 of our supplementary material.

Algorithm 1 Causal Structure-guided DRO (CS-DRO)

Require: Multi-source datasets $\{\mathcal{D}^e\}_{e\in\mathcal{E}^{\text{src}}}$, radius ρ, structural budget κ, step sizes η' (inner), η (outer), η_ρ and η_κ, weights λ_G.
Ensure: Trained predictor $f_\theta = w \circ \phi$.

Warm-up: train (ϕ, w) on $\cup_e \mathcal{D}^e$ by ERM with HSIC.
Proxy DAG: train $\widehat{G}$ on samples from $\mathcal{N}(\mu_{e,c}, \sigma_{e,c})$.
Fix $(\widehat{G}, \widehat{G}_Y)$.
while not converged **do**
  Sample a batch $B = \{(x_i, y_i)\}_{i=1}^{n_B}$ from $\cup_e \mathcal{D}^e$.
  Compute $z_i = \phi(x_i)$ and $\tilde{z}_i = \widehat{G}_Y \odot z_i$ for all i.
  Initialize adversarial representations $z'_i \leftarrow z_i$ + noise.
  repeat ▷ inner $\max_{B'}$ (gradient ascent)
    Compute $\tilde{z}'_i = \widehat{G}_Y \odot z'_i$ for all i and let $B' = \{(z'_i, y_i)\}_{i=1}^{n_B}$.
    Update z' by ascent:

$$z' \leftarrow z' + \eta' \nabla_{z'} \Big(\frac{1}{n_B} \sum_{i=1}^{n_B} \ell(w(\tilde{z}'_i), y_i) - \gamma \big\| \widehat{G}_Y \odot (z - z') \big\|_2^2 - \beta D_{\widehat{G}}(B', B) \Big).$$

  until a fixed number of steps (J) is reached
  Update (θ, γ, β) by gradient descent: ▷ outer $\min_{\theta,\gamma,\beta}$

$$\theta \leftarrow \theta - \eta \nabla_\theta \Big(\frac{1}{n_B} \sum_{i=1}^{n_B} \tilde{\ell}(z_i, z'_i, y_i) + \lambda_G \, \mathcal{L}_G(B) \Big).$$

  $\gamma \leftarrow \gamma - \eta_\rho \big(\rho^2 - \widehat{\mathbb{E}}_B[c(z, z')] \big)_+, \;\; \beta \leftarrow \beta - \eta_\kappa \big(\kappa - D_{\widehat{G}}(B', B) \big)_+.$
end while

4 Theoretical Analysis

In this section, we reformulate the distribution-level constrained objective of Eq. (11) into the tractable primal–dual training objective of Eq. (13) via Lagrangian relaxation, and provide theoretical certification. We also derive the probability that an unseen target domain is included within the uncertainty set and establish the resulting generalization risk bounds.

4.1 Tractable Reformulation via Compactness and Strong Duality

To establish the well-posedness of the objective in Eq. (11) and secure the topological and convex-analytic properties of the uncertainty set in Definition 3.5, we introduce the following assumptions regarding the representation space, loss function, and constraint functionals.

Assumption 4.1 (Compact support in representation space)**.** Representation space $\mathcal{Z}$ is compact. Furthermore, as Y is a finite set, the joint space $V = Z \times Y$ is a compact metric space.

Assumption 4.2 (Regularity of objective and constraints)**.** We assume that for any fixed $\theta \in \Theta$, the following hold:

(1) **Loss regularity.** The loss function $\ell(\cdot)$ is Borel-measurable, and real-valued upper semi-continuous, and lower bounded on V. Furthermore, $\sup_{(z,y)\in\mathcal{V}} \ell(w_\theta(z), y) < \infty$.
(2) **Transport-cost regularity.** The cost function c_G is lower semi-continuous on $\mathcal{V}^2$ in the extended-real sense and satisfies $c_G(v, v) = 0$. Therefore, $W_G(P, Q)$ is lower semi-continuous with respect to weak convergence, and by continuity and monotonicity, $W_G(P, Q)^2$ is also lower semi-continuous under weak convergence.
(3) **Structural discrepancy regularity.** The discrepancy D_G is convex with respect to Q and is lower semi-continuous on $\mathcal{P}(\mathcal{V})$ with respect to weak convergence.

Assumption 4.3 (Slater condition)**.** For radius $\rho > 0$ and structural budget $\kappa > 0$, since $Q = P$ is feasible for the inner problem of Eq. (11), the Slater condition holds.

We follow the conditions established in prior literature [40, 55] for Assumption 4.1. This assumption can be practically satisfied by applying normalization to the latent features z. Assumption 4.2 (ii)–(iii) provide the necessary regularity for the sets in Definition 3.5 to be closed in the weak topology. Furthermore, convexity and lower semi-continuity in (iii) can be satisfied because G is linear in D_G and we use a squared-error formulation for D_G. We provide a detailed discussion in Section B.2 of our supplementary material.

We first establish that the uncertainty set in Definition 3.5 is compact under the weak topology, which in turn implies that the inner supremum in Eq. (11) is attained.

Proposition 4.4 (Weak compactness and attainment)**.** *Under Assumptions 4.1–4.2, the uncertainty set in Definition 3.5 is nonempty, convex, and weakly compact in the space of probability measures $\mathcal{P}(\mathcal{V})$ (endowed with the weak topology on $\mathcal{V}$). Furthermore, for any fixed $\theta \in \Theta$ and G, the worst-case risk in Eq.* (11) *is attained. In other words, there exists some $Q^\star(\theta) \in \mathcal{Q}_{W_G, D_G}(P)$ such that:*

$$\mathbb{E}_{Q^\star(\theta)}[\ell(w(z), y)] = \sup_{Q \in \mathcal{Q}_{W_G, D_G}(P)} \mathbb{E}_Q[\ell(w(z), y)]. \tag{20}$$

A detailed proof is provided in Section C.1 of our supplementary material. Proposition 4.4 guarantees that within the proposed structure-preserving uncertainty set, a distribution-level adversary Q^* exists, ensuring the problem is well-posed. This result serves as the foundation for validating the Lagrangian relaxation through strong duality.

Next, we demonstrate that the Lagrangian relaxation from Eq. (11) to Eq. (12)–(13) is valid at the distribution level.

Proposition 4.5 (Strong duality for the inner problem)**.** *Under Assumptions 4.1–4.3, for any fixed $\theta \in \Theta$ and G, the inner supremum of Eq.* (11) *satisfies a zero duality gap:*

$$\sup_{Q \in \mathcal{Q}_{W_G, D_G}(P)} \mathbb{E}_Q[\ell(w(z), y)] = \inf_{\gamma \geq 0, \; \beta \geq 0} \Big\{ \gamma \rho^2 + \beta \kappa + \sup_Q \Big(\mathbb{E}_Q[\ell(w(z), y)] - \gamma W_G(P, Q)^2 - \beta D_G(P, Q) \Big) \Big\}. \tag{21}$$

Therefore, the penalized objective in Eq. (12) *can indeed be viewed as a reformulation of the constrained inner problem (Eq.* (11)*), and Eq.* (13) *is justified as a primal–dual objective for distribution-level robustness.*

A detailed proof is given in Section C.2 of our supplementary material.

4.2 Certified Generalization

We now verify that $\mathcal{Q}_{W_G, D_G}$ includes the unseen target distribution P^T with high probability. This result is not intended to cover all forms of target shift; rather, the goal is to provide a certificate

within the bounded shift range consistent with the standard DG setting. We assume that the target shift is bounded with respect to the causal mechanism. This assumption means that only shifts at a controllable level along the two axes of geometry and structure are allowed.

Assumption 4.6 (Bounded Causal Shift)**.** Let $\mathcal{P}_G$ be the family of distributions generated by the invariant causal structure. Denote P as the population source distribution on $V = Z \times Y$ and $P^T \in \mathcal{P}_G$ as the unseen target domain. We assume P^T satisfies the data-generating process with respect to P:

$$W_G(P, P^T) \le \rho_0, \quad D_G(P, P^T) \le \kappa_0, \tag{22}$$

where $\rho_0, \kappa_0 \ge 0$ denote the intrinsic population-level radius and budget for geometric shift and structural discrepancy, respectively.

Since the population distribution P is not observable in practice, we construct an empirical distribution $\hat{P}_n$ from n i.i.d. samples. We then verify whether P^T is included under a radius ρ and structural budget κ defined based on the geometry and structure of $\hat{P}_n$.

Theorem 4.7 (Target inclusion certificate)**.** *If Assumptions 4.1–4.6 hold, let $\hat{P}_n$ be the empirical distribution of n i.i.d. samples drawn from P. For a fixed $\rho > \rho_0$ and $\kappa > \kappa_0$, we define the margins as $\Delta_\rho := (\rho - \rho_0)/2$ and $\Delta_\kappa := (\kappa - \kappa_0)/2$. Then, the target inclusion event holds with probability as follows:*

$$\mathbb{P}\left(P^T \in \mathcal{Q}_{W_G,D_G}(\hat{P}_n)\right) \ge 1 - \underbrace{\mathrm{C}(\Delta_\rho) e^{\left(-n\,\mathrm{a}(\Delta_\rho)\right)}}_{\text{geometric term}} - \underbrace{e^{\left(-n\,\mathrm{b}(\Delta_\kappa)\right)}}_{\text{structural term}}. \tag{23}$$

Here, $\mathrm{C}(\Delta_\rho)$ is a complexity term of the covering-number type induced by the c_G-geometry, which is primarily governed by the G_Y-aligned subspace. The functions $\mathrm{a}(\cdot)$ and $\mathrm{b}(\cdot)$ are positive and determined by the boundedness and diameter parameters. In particular, for any fixed (ρ, κ) satisfying $\rho > \rho_0$ and $\kappa > \kappa_0$, the lower bound in Eq. (23) converges to 1 at an exponential rate with respect to n.

We provide a detailed proof in Section C.3 of our supplementary material.

If the above inclusion event holds, then the worst-case risk for $\mathcal{Q}_{W_G,D_G}(\hat{P}_n)$ can provide an upper bound on the target risk. For the empirical distribution $\hat{P}_n$, define the following worst-case risk:

$$R(\hat{P}_n; \theta) = \sup_{Q \in \mathcal{Q}_{W_G,D_G}(\hat{P}_n)} \mathbb{E}_{(z,y)\sim Q}\left[\ell(w(z), y)\right]. \tag{24}$$

Then, on the event $P^T \in \mathcal{Q}_{W_G,D_G}(\hat{P}_n)$, the following holds:

$$R(P^T; \theta) := \mathbb{E}_{P^T}\left[\ell(w(z), y)\right] \le R(\hat{P}_n; \theta). \tag{25}$$

Therefore, combining with Eq. (23), we obtain:

$$\mathbb{P}\Big(R(P^T; \theta) \le R(\hat{P}_n; \theta)\Big) \;\ge\; \mathbb{P}\Big(P^T \in \mathcal{Q}_{W_G,D_G}(\hat{P}_n)\Big). \tag{26}$$

In other words, the probability that the worst-case risk defined on the uncertainty set centered at $\hat{P}_n$ serves as an upper bound for the target risk is bounded below by the inclusion probability from Theorem 4.7. However, this guarantee is provided only within the shift range specified by $\mathcal{Q}_{W_G,D_G}(\hat{P}_n)$, and should not be interpreted as covering any distribution shifts beyond that range.

Remark 4.8 (Practical optimization). Theorem 4.7 provides a theoretical guarantee that, under the worst-case risk defined on the empirical distribution $\hat{P}_n$, the target distribution is included in the uncertainty set. However, in the actual implementation of Algorithm 1, for efficiency, the inner problem is solved approximately via J-step gradient ascent on mini-batches. In this process, an objective suboptimality ε_{opt} arises between the global optimum of the inner optimization and the approximate solution obtained by stochastic optimization. Additionally, relying on individual mini-batch sampling instead of the entire dataset incurs a stochastic estimation error $\varepsilon_{\text{stoch}}$. Analyses of these approximation errors are provided in Section D of our supplementary material. Furthermore, potential errors arising from a sub-optimal estimation of G may further tighten the conditions for target inclusion, the analysis of which is included in Section E.

5 Related Work

5.1 Domain Generalization

Existing DG studies can be divided into three streams [64]: (i) Data manipulation involves transforming input data to learn general representations, which includes both data augmentation [29] and methods for generating diverse data [39, 62, 72]. (ii) Representation learning focuses on extracting domain-invariant features through techniques such as adversarial training [12], feature alignment across domains [67, 71], or invariant risk minimization [2, 26]. This category also encompasses research into the disentanglement of domain-shared and domain-specific features to foster better generalization [23, 43]. (iii) General learning strategies include ensemble learning [3] and methods that achieve generalization by optimizing the geometry of the loss function [8, 28, 65].

Adversarial training [12, 31, 37] can be understood as an approach that improves robustness to local perturbations by training the model to minimize the worst-case risk within a bounded perturbation set around each sample. In contrast, WDRO addresses distribution-level shifts by minimizing the worst-case risk over an uncertainty set defined by transport costs [7]. On the basis of WDRO, we restrict the uncertainty set in a structure-preserving manner with Wasserstein distance.

5.2 Distributionally Robust Optimization

DRO methods can be categorized by the choice of distribution discrepancy used to define the uncertainty set, such as $f-$divergence [5, 10, 44], Wasserstein distance [40, 56], and maximum mean discrepancy [59]. DRO has been utilized for various objectives, including ensuring robustness for minority groups [53], improving generalization under distribution shifts [24, 40, 49], mitigating robust overfitting [34], and addressing individual fairness [14]. However, their application to DG is often limited because capturing unknown target distributions necessitates an overwhelmingly large uncertainty set; such expansion can include unrealistic shifts, thereby degrading generalization performance [11, 17, 33].

To alleviate these issues, prior works have introduced additional constraints on the uncertainty set. For instance, SAL [33] and MODE [11] attempt to preserve stable correlations across domains, while TRO [51] employs topological constraints to restrict the search space to the vicinity of the source domains. Recent research into

Table 1: Performance comparison of the proposed method with existing methods. Results are reported as accuracy (%), averaged over three runs with different random seeds. Except for our method and the baselines denoted by †, all results are taken from their original papers.

Method	PACS	VLCS	OfficeH.	TerraInc.	Avg.
General DG Method					
Mixstyle [72]	85.2	77.9	60.4	44.0	66.9
SWAD [8]	88.1	79.1	70.6	50.0	72.0
SAGM [65]	86.6	80.0	70.1	48.8	71.4
iDAG [23]	88.8	76.9	71.8	46.1	70.9
GMDG [61]	85.6	79.2	70.7	50.1	71.4
MADG [12]	86.5	78.7	71.3	53.7	72.6
SFT [28]	88.3	79.8	70.9	50.7	72.4
GUIDE [62]	87.1	78.5	68.6	51.3	71.4
I^3C [71]	87.1	79.6	70.2	49.6	71.6
DSBH [46]	86.3	79.3	71.1	52.4	72.3
DgCD [25]	88.5	79.9	70.6	52.9	73.0
DRO-based Method					
GroupDRO [53]	84.4	76.7	66.0	43.2	67.6
SAL† [33]	86.1	78.0	71.4	52.2	71.9
DR-DG † [40]	87.5	78.9	70.7	50.4	71.9
GLOT-DR† [49]	86.6	78.1	71.9	50.2	71.7
TRO† [51]	86.8	77.4	71.5	51.3	71.8
SRM [50]	87.1	79.9	71.0	49.5	71.9
MEDRO † [24]	87.0	79.6	70.0	49.2	71.5
SR-WDRO† [34]	86.4	79.1	71.1	50.7	71.8
CS-DRO	**88.9**	**80.8**	**73.4**	**54.1**	**74.3**

$\mathcal{G}$−causal OT [9] and structural causal OT [15] proposes variants of OT to ensure the uncertainty set contains causally plausible distributions. However, these methodologies often fail to provide efficient computational methods [9] or rely on assumptions such as the availability of pre-defined causal structures and access to the functional forms of the causal mechanisms, which are difficult to satisfy in deep learning [15]. We address these limitations by estimating causal structures from source domains and formulating a differentiable structure-preserving constraint, enabling a tractable search for the worst-case distribution while maintaining causal consistency (see Section F of our supplementary material).

6 Experiments

Following the standard protocol in DomainBed [20], we evaluate our method on four widely used benchmarks: PACS [30], VLCS [16], OfficeHome [63], and TerraIncognita [4]. To ensure a fair comparison, we adopt DomainBed's dataset splitting method and evaluation protocol. Specifically, we split the source domain data into training and validation sets with an 8:2 ratio. For evaluation, we use a leave-one-domain-out protocol: one domain is held out as the unseen target domain and the remaining domains are used for training. The model selection and hyperparameter tuning are performed on a validation set created from the training domains. We rotate the held-out domain so that each domain is used as the target in turn, and report the image classification performance in each case. All experiments are repeated three times.

We use ResNet-50 [21] pre-trained on ImageNet [13] as the backbone. The model is fine-tuned using the Adam optimizer with a 5e-5 learning rate and 1e-6 weight decay. During training, each mini-batch includes samples from all source domains, with 32 samples per domain. The model is trained for 5,000 steps on each dataset. λ_G and κ were tuned on the validation set, using a log-scaled search space from 10^{-3} to 10^2 following [51]. We fixed the regularization hyperparameters τ and λ_r at 10 and 1, respectively. Further implementation details are provided in Section G.3 of our supplementary material.

We first show the results of our comparative analysis with the baselines, and then discuss the impact of our causal structure-preserving uncertainty set design on generalization. Due to the space limitation, additional experimental results, including the validation of DAG learning efficiency, the effect of the regularization loss, and a computational cost analysis, are provided in Section H of our supplementary material.

6.1 Comparative Results

We compare our CS-DRO with the DG baselines listed in Table 1. For a fair comparison, we employ the baselines following the same evaluation protocol (DomainBed) and backbone (ResNet-50) with our method. Meanwhile, the (reproduced) DRO-based methods were run in the same training and evaluation environment as our method. Although most DRO methods were not originally designed for the DG task, we include their performance in the DG setting as baselines to demonstrate the effectiveness of our CS-DRO. Note that our method without the structure-preserving constraint and regularization terms is equivalent to DR-DG [40].

Table 1 reports the result. CS-DRO achieves the highest mean performance (74.3%) among all methods. Even compared to recent DG methods like GUIDE [62] and I^3C [71], it maintains superior performance on most datasets, and especially in TerraIncognita (which involves large domain shifts), it shows a clear performance advantage. In addition, compared to the DRO-based methods, our CS-DRO provides consistent improvements in average accuracy and achieves the highest performance among DRO-based approaches. These results indicate that constraining the uncertainty set to preserve the causal structure between representations and labels enables CS-DRO to generalize better than existing DRO methods.

6.2 Ablation Study

We conduct an ablation study to analyze how each component contributes to the overall performance of the proposed method. We independently evaluate (i) the effect of additional regularization terms used during training, (ii) the impact of the structure-preserving constraint within the uncertainty set $\boldsymbol{Q}$, and (iii) the influence of acyclicity and sparsity constraints during DAG learning. All variants are compared under identical training and evaluation settings.

As shown in Table 2, removing the regularization terms leads to a general decline in performance. Specifically, the most significant drop is observed when $\mathcal{L}_G$ (Eq. (15)), which encourages the representation Z to align with the fixed DAG G, is removed. This suggests that a mismatch between the learned distribution of G and the current distribution of Z introduces errors in D_G, subsequently

Table 2: Ablation study of the proposed method on the DomainBed benchmark. We report average test accuracy (%) on four datasets when removing different regularization terms and components of CS-DRO.

Variant	PACS	VLCS	OfficeH.	TerraInc.	Avg.
Effect of additional regularization terms					
w/o $\mathcal{L}_G$	87.1	78.9	71.6	51.1	72.2
w/o $\mathcal{L}_{\text{reg}}$	87.8	79.9	72.4	53.2	73.3
Effect of components for uncertainty sets					
w/o G_Y	88.0	80.0	72.8	52.3	73.3
w/o D_G	87.5	79.5	72.3	51.6	72.7
w/o D_G & G_Y	87.3	79.1	71.4	51.0	72.2
Effect of components for DAG learning in Eq. (14)					
w/o $\mathcal{L}_1$	88.2	79.3	72.1	52.4	73.0
w/o h	87.5	79.7	72.3	51.9	72.9
CS-DRO	**88.9**	**80.8**	**73.4**	**54.1**	**74.3**

leading to an uncertainty set $\mathcal{Q}$ that favors non-causal directions. Regarding the uncertainty set, removing the constraint based on the structural discrepancy D_G results in a larger performance degradation compared to using an isotropic cost without G_Y. The most severe drop occurs when both are removed. This indicates that the synergy between the structure-preserving constraint and the anisotropic cost is crucial for forming an appropriate uncertainty set for unseen domains. Additionally, omitting constraints in DAG learning results in overly dense graphs that fail to capture the underlying causal structure, thereby weakening the effectiveness of structure preservation. In summary, each component plays a vital role in ensuring that the uncertainty set preserves the proper causal structure, ultimately contributing to enhanced generalization under domain shifts.

6.3 Structural Control Analysis.

To further investigate the impact of the structure-preserving constraint within the uncertainty set $\mathcal{Q}$, we evaluate performance variations with respect to κ and simultaneously analyze the effect of $\mathcal{L}_G$. In this setting, while ρ is dynamically adjusted during the training process, κ remains a fixed hyperparameter that controls the allowable range of the structural shift. To this end, we designate six candidates for each of κ and λ_G on a logarithmic scale and perform a 6×6 grid search.

Figure 2 illustrates the results. Overall, a relatively wide plateau is formed within the ranges of $\kappa \in \{10^{-2}, 10^{-1}, 10^{0}\}$ and $\lambda_G \in \{10^{0}, 10^{1}\}$. We observe that the performance remains stable when (κ, λ_G) is selected from these intervals. These findings suggest that within the range where λ_G is effectively applied, the representation Z maintains the DAG structure G well. Consequently, this enables the formation of appropriate worst-case distributions within $\mathcal{Q}$ where the structure is preserved according to the influence of D_G. Therefore, high performance can be maintained across a relatively broad range of κ. In contrast, when λ_G is small, performance remains relatively low despite variations in κ. This is interpreted as further evidence that the latent representation Z must faithfully preserve G. Similar experiments for other datasets are provided in Section H.3 of our supplementary material.

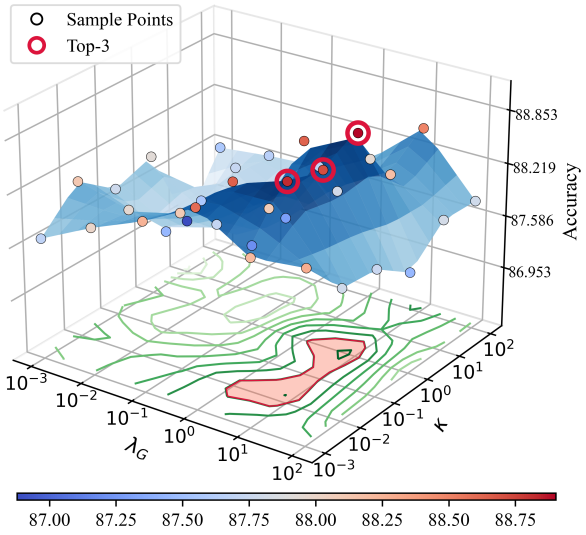


Figure 2: Performance comparison on the PACS dataset with respect to variations in κ and λ_G.

7 Conclusion

In this paper, we propose Causal Structure-guided Distributionally Robust Optimization (CS-DRO) to achieve generalization under domain shifts. We estimate the causal structure between representations and labels shared across multiple source domains as a DAG. Based on this, we design the uncertainty set to be restricted to preserving this structure. We formulate the DRO problem, which combines a structure-preserving constraint with an anisotropic Wasserstein cost, into a tractable primal-dual learning problem. Furthermore, we theoretically discuss that the target distribution can be located within the uncertainty set under certain conditions and provide an analysis in the form of a generalization upper bound based on this finding. Experimental results demonstrate that CS-DRO achieves the highest average performance across standard DG benchmarks compared to existing DG and DRO methods. We believe that aligning robustness with causal structure, rather than defining it solely based on geometric proximity, is an effective direction for DG. Future research could involve introducing more sophisticated structure estimation and expanding to various types.

Acknowledgement

This work was supported by (1) **Samsung Electronics** Co., Ltd., (2) the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(**RS-2025-24535423**), and the Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)((3) **RS-2025-25422680**, Metacognitive AGI Framework and its Applications, and (4) **RS-2020-II201373**, Artificial Intelligence Graduate School Program (Hanyang University)).

References

[1] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. 2021. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems* 34 (2021), 3438–3450.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893* (2019).

[3] Devansh Arpit, Huan Wang, Yingbo Zhou, and Caiming Xiong. 2022. Ensemble of averages: Improving model selection and boosting performance in domain generalization. *Advances in Neural Information Processing Systems* 35 (2022), 8265–8277.

[4] Sara Beery, Grant Van Horn, and Pietro Perona. 2018. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*. 456–473.

[5] Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. 2013. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* 59, 2 (2013), 341–357.

[6] Jose Blanchet, Karthyek Murthy, and Fan Zhang. 2022. Optimal transport-based distributionally robust optimization: Structural properties and iterative schemes. *Mathematics of Operations Research* 47, 2 (2022), 1500–1529.

[7] Anh Tuan Bui, Trung Le, Quan Hung Tran, He Zhao, and Dinh Phung. 2022. A Unified Wasserstein Distributional Robustness Framework for Adversarial Training. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Dzpe9C1mpiv

[8] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. 2021. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems* 34 (2021), 22405–22418.

[9] Patrick Cheridito and Stephan Eckstein. 2025. Optimal transport and Wasserstein distances for causal models. *Bernoulli* 31, 2 (2025), 1351–1376.

[10] Imre Csiszár. 1967. On information-type measure of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.* 2 (1967), 299–318.

[11] Rui Dai, Yonggang Zhang, Zhen Fang, Bo Han, and Xinmei Tian. 2023. Moderately Distributional Exploration for Domain Generalization. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 6786–6817. https://proceedings.mlr.press/v202/dai23d.html

[12] Aveen Dayal, Vimal KB, Linga Reddy Cenkeramaddi, C Mohan, Abhinav Kumar, and Vineeth N Balasubramanian. 2023. MADG: margin-based adversarial learning for domain generalization. *Advances in Neural Information Processing Systems* 36 (2023), 58938–58952.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[14] Ahmad Reza Ehyaei, Golnoosh Farnadi, and Samira Samadi. 2024. Wasserstein Distributionally Robust Optimization through the Lens of Structural Causal Models and Individual Fairness. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=piOzFx9whU

[15] Ahmad Reza Ehyaei, Golnoosh Farnadi, and Samira Samadi. 2025. Designing Ambiguity Sets for Distributionally Robust Optimization Using Structural Causal Optimal Transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 16462–16470.

[16] Chen Fang, Ye Xu, and Daniel N Rockmore. 2013. 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*. 1657–1664.

[17] Charlie Frogner, Sebastian Claici, Edward Chien, and Justin Solomon. 2019. Incorporating unlabeled data into distributionally robust learning. *arXiv preprint arXiv:1912.07729* (2019).

[18] Rui Gao and Anton Kleywegt. 2023. Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research* 48, 2 (2023), 603–655.

[19] Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. 2007. A kernel statistical test of independence. *Advances in neural information processing systems* 20 (2007).

[20] Ishaan Gulrajani and David Lopez-Paz. 2020. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434* (2020).

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[22] Yue He, Peng Cui, Zheyan Shen, Renzhe Xu, Furui Liu, and Yong Jiang. 2021. Daring: Differentiable causal discovery with residual independence. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. 596–605.

[23] Zenan Huang, Haobo Wang, Junbo Zhao, and Nenggan Zheng. 2023. idag: Invariant dag searching for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 19169–19179.

[24] Jinyong Jeong, Hyungu Kahng, and Seoung Bum Kim. 2025. Multi-Expert Distributionally Robust Optimization for Out-of-Distribution Generalization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=Lz5BUjArK4

[25] Seonggyeom Kim, Byeongtae Park, Harim Lee, and Dong-Kyu Chae. 2024. Discrepancy-guided channel dropout for domain generalization. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 1110–1120.

[26] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*. PMLR, 5815–5826.

[27] Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. 2019. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*. Informs, 130–166.

[28] Aodi Li, Liansheng Zhuang, Xiao Long, Minghong Yao, and Shafei Wang. 2025. Seeking consistent flat minima for better domain generalization via refining loss landscapes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 15349–15359.

[29] Chenming Li, Daoan Zhang, Wenjian Huang, and Jianguo Zhang. 2023. Cross contrasting feature perturbation for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1327–1337.

[30] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. 2017. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*. 5542–5550.

[31] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. 2018. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5400–5409.

[32] Fengming Lin, Xiaolei Fang, and Zheming Gao. 2022. Distributionally robust optimization: A review on theory and applications. *Numerical Algebra, Control and Optimization* 12, 1 (2022), 159–212.

[33] Jiashuo Liu, Zheyan Shen, Peng Cui, Linjun Zhou, Kun Kuang, Bo Li, and Yishi Lin. 2021. Stable adversarial learning under distributional shifts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 8662–8670.

[34] Shuang Liu, Yihan Wang, Yifan Zhu, Yibo Miao, and Xiao-Shan Gao. 2025. Provable Robust Overfitting Mitigation in Wasserstein Distributionally Robust Optimization. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=sq5LLWk5SN

[35] Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. 2021. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*.

[36] Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. 2022. Causality inspired representation learning for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8046–8056.

[37] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*. https://openreview.net/forum?id=rJzIBfZAb

[38] Divyat Mahajan, Shruti Tople, and Amit Sharma. 2021. Domain generalization using causal matching. In *International conference on machine learning*. PMLR, 7313–7324.

[39] Noaman Mehmood and Kenneth Barner. 2025. Disentangled Embedding through Style and Mutual Information for Domain Generalization. *Transactions on Machine Learning Research* (2025). https://openreview.net/forum?id=552tedTByb

[40] Akshay Mehra, Bhavya Kailkhura, Pin-Yu Chen, and Jihun Hamm. 2022. Do domain generalization methods generalize well?. In *NeurIPS ML Safety Workshop*.

[41] Jovana Mitrovic, Brian McWilliams, Jacob C Walker, Lars Holger Buesing, and Charles Blundell. 2021. Representation Learning via Invariant Causal Mechanisms. In *International Conference on Learning Representations*. https://openreview.net/forum?id=9p2ekP904Rs

[42] Peyman Mohajerin Esfahani and Daniel Kuhn. 2018. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming* 171, 1 (2018), 115–166.

[43] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. 2021. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8690–8699.

[44] Hongseok Namkoong and John C Duchi. 2016. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in neural information processing systems* 29 (2016).

[45] Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. 2020. On the role of sparsity and dag constraints for learning linear dags. *Advances in Neural Information Processing Systems* 33 (2020), 17943–17954.

[46] Byeongtae Park, Harim Lee, Sungmoon Kim, and Dong-Kyu Chae. 2025. Dynamic Spuriosity Bias Harmonizer: Combatting Spurious Features for Domain Generalization. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 44–56.

[47] Judea Pearl. 2009. *Causality*. Cambridge university press.

[48] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. 2019. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1406–1415.

[49] Hoang Phan, Trung Le, Trung Phung, Anh Tuan Bui, Nhat Ho, and Dinh Phung. 2023. Global-local regularization via distributional robustness. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 7644–7664.
[50] Fengchun Qiao, Yanlin Chen, and Xi Peng. 2025. Structure-informed Risk Minimization for Robust Ensemble Learning. In *Forty-second International Conference on Machine Learning*. https://openreview.net/forum?id=GvV9MJs268
[51] Fengchun Qiao and Xi Peng. 2023. Topology-aware Robust Optimization for Out-of-Distribution Generalization. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=ylMq8MBnAp
[52] Hamed Rahimian and Sanjay Mehrotra. 2022. Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization* 3 (2022), 1–85.
[53] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally Robust Neural Networks. In *International Conference on Learning Representations*. https://openreview.net/forum?id=ryxGuJrFvS
[54] Xinwei Shen, Peter Bühlmann, and Armeen Taeb. 2025. Causality-oriented robustness: exploiting general noise interventions in linear structural causal models. *J. Amer. Statist. Assoc.* just-accepted (2025), 1–20.
[55] Aman Sinha, Hongseok Namkoong, and John Duchi. 2018. Certifiable Distributional Robustness with Principled Adversarial Training. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Hk6kPgZA-
[56] Aman Sinha, Hongseok Namkoong, and John Duchi. 2018. Certifying Some Distributional Robustness with Principled Adversarial Training. In *International Conference on Learning Representations*.
[57] Zeen Song, Siyu Zhao, Xingyu Zhang, Jiangmeng Li, Changwen Zheng, and Wenwen Qiang. 2025. Learning Invariant Causal Mechanism from Vision-Language Models. In *Forty-second International Conference on Machine Learning*. https://openreview.net/forum?id=GB9XiKIwfp
[58] Peter Spirtes, Clark N Glymour, and Richard Scheines. 2000. *Causation, prediction, and search*. MIT press.
[59] Matthew Staib and Stefanie Jegelka. 2019. Distributionally robust optimization and generalization in kernel methods. *Advances in Neural Information Processing Systems* 32 (2019).
[60] Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. 2019. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *International Conference on Machine Learning*. PMLR, 6056–6065.
[61] Zhaorui Tan, Xi Yang, and Kaizhu Huang. 2024. Rethinking multi-domain generalization with a general learning objective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23512–23522.
[62] Xavier Thomas and Deepti Ghadiyaram. 2025. What's in a Latent? Leveraging Diffusion Latent Space for Domain Generalization. In *Second Workshop on Visual Concepts*. https://openreview.net/forum?id=ay9jEQazmG
[63] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. 2017. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5018–5027.
[64] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S Yu. 2022. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering* 35, 8 (2022), 8052–8072.
[65] Pengfei Wang, Zhaoxiang Zhang, Zhen Lei, and Lei Zhang. 2023. Sharpness-aware gradient matching for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3769–3778.
[66] Yu Wang, An Zhang, Xiang Wang, Yancheng Yuan, Xiangnan He, and Tat-Seng Chua. 2022. Differentiable invariant causal discovery. *arXiv preprint arXiv:2205.15638* (2022).
[67] Xufeng Yao, Yang Bai, Xinyun Zhang, Yuechen Zhang, Qi Sun, Ran Chen, Ruiyu Li, and Bei Yu. 2022. Pcl: Proxy-based contrastive learning for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7097–7107.
[68] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. 2019. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*. PMLR, 7472–7482.
[69] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. 2018. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems* 31 (2018).
[70] Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. 2020. Learning sparse nonparametric dags. In *International conference on artificial intelligence and statistics*. Pmlr, 3414–3425.
[71] Chaoyang Zhou, Zengmao Wang, and Bo Du. 2025. Learning intrinsic invariance within intra-class for domain generalization. *IEEE Transactions on Multimedia* (2025).
[72] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. 2024. Mixstyle neural networks for domain generalization and adaptation. *International Journal of Computer Vision* 132, 3 (2024), 822–836.

Appendix

To support reproducibility, we publicly release our source code, datasets, and scripts used in the main experiments. In addition, all detailed derivations, proofs, additional experiments, and extended analyses are provided in a separate Supplementary Material:

- **Code repository:** `https://github.com/gyeomo/CS-DRO`
- **Supplementary Material:** `https://github.com/gyeomo/CS-DRO/blob/main/Supplementary_Material.pdf`

To help readers locate specific material of interest, we summarize the contents of the Supplementary Material below.

Section A — Notation. We summarize the mathematical notations used throughout the paper.

Section B — Structure-Preserving Discrepancy via Invariant Gradient Signatures. We provide the detailed definition of $M_G(Q)$ and $D_G(P, Q)$ introduced in Section 3.2, and prove the convexity and weak lower semi-continuity of D_G on $\mathcal{P}(\mathcal{V})$. We further justify the structure-preserving constraint via first-order optimality and local improvement bounds.

Section C — Proofs of Theoretical Results. We provide the detailed proofs of the results in Section 4, including Proposition 4.4, Proposition 4.5, and Theorem 4.7.

Section D — Approximation Gaps between Theory and Implementation. We formally define the two approximation errors arising from the mini-batch-based J-step stochastic optimization in Algorithm 1: the inner-solver suboptimality ε_{opt} and the mini-batch sampling variability $\varepsilon_{\text{stoch}}$, and empirically analyze them on PACS.

Section E — Impact of Structure Estimation Error. We analyze how the error $\epsilon_G := \|\hat{G} - G^\star\|_F$ between the estimated proxy DAG $\hat{G}$ and the optimal structure $G^\star$ affects the target inclusion certificate of Theorem 4.7, deriving a stability inequality that quantifies how ϵ_G inflates the radius ρ and budget κ.

Section F — Differences to Structural Causal OT.. We compare CS-DRO with the $\mathcal{G}$-causal OT of Cheridito and Eckstein [9] and the structural causal OT of Ehyaei et al. [15], and clarify why these approaches are difficult to apply to high-dimensional DG.

Section G — Implementation Details. We describe the model architecture and the full training procedure, and report the complete list of hyperparameters used for each dataset.

Section H — Additional Experiments. We report additional empirical results, including: (i) performance on the large-scale DomainNet [48] dataset; (ii) validity verification of the proxy DAG learning; (iii) additional structural control analysis on VLCS, OfficeHome, and TerraIncognita w.r.t. κ and λ_G; (iv) hyperparameter and calibration analysis of $\mathcal{L}_{\text{reg}}$; and (v) a computational cost comparison with existing DRO-based methods.

Section I — Full Results of Table 1. We report the full domain-wise accuracy results corresponding to Table 1 of the main paper under the leave-one-domain-out protocol.