

CGU-BAYES: Causal Graph Uncertainty-Guided Bayesian Inference for Domain Generalization

Naiyu Yin
 Lehigh University
 Bethlehem, Pennsylvania
 nay224@lehigh.edu

Hanjing Wang
 Rensselaer Polytechnic Institute
 Troy, New York
 wangh36@rpi.edu

Yue Yu
 Lehigh University
 Bethlehem, Pennsylvania
 yuy214@lehigh.edu

Tian Gao
 IBM Research
 Yorktown Heights, New York
 tgao@us.ibm.com

Amit Dhurandhar
 IBM Research
 Yorktown Heights, New York
 adhuran@us.ibm.com

Chung-Hao Lee
 University of California, Riverside
 Riverside, California
 chunghal@ucr.edu

Qiang Ji
 Rensselaer Polytechnic Institute
 Troy, New York
 jiq@rpi.edu

Abstract

Causal graphs play a crucial role in AI research as they reveal the data generation processes underlying real-world machine learning and computer vision tasks. Recent studies have leveraged causal graphs to develop more robust and interpretable models. However, limited or biased data often lead to inaccurate causal graph estimation, reducing a model's transferability to unseen domains. To address this challenge, we propose a novel framework that performs Bayesian inference over causal graphs to capture potential underlying causal relations and identify invariant causal features for DG prediction. The key advantage of our framework lies in its ability to quantify causal graph uncertainty in the context of prediction tasks and incorporate it into the prediction process. Our proposed uncertainty provides valuable insights into (i) the reliability of our method on specific datasets, (ii) the alignment between learned causal graphs and unseen test domains, and (iii) the confidence of our predictions. In particular, we go beyond merely quantifying uncertainty and leverage it as weighting factors in a weighted Bayesian inference scheme. Empirical results on multiple benchmark distribution-shift datasets show that our algorithm, Causal Graph Uncertainty-guided Bayesian Inference (CGU-BAYES), outperforms existing DG methods on challenging datasets and achieves state-of-the-art performance overall.

1. Introduction

Although modern deep learning models have demonstrated impressive performance across a spectrum of tasks with large-scale annotated datasets, they still face significant limitations, including poor generalization, lack of explainability, and fairness issues due to their reliance on spurious correlations. These correlations are easy to capture but prone to change across domains. To address this, a natural approach is to leverage the invariance of causality and model causal relations within data using structural causal models (SCMs). Guided by SCMs, multiple studies [7, 26, 35–37, 40] can identify and eliminate spurious representations that show domain-varying correlations with the label, distinguishing them from causal representations that remain stable across domains. These works successfully improve domain generalization (DG) prediction performance by constructing a predictive mechanism using domain-invariant causal representations.

Unfortunately, the true SCMs over the variables of interest (the latent representations, the prediction target, and the domain variable) are unknown. They are either constructed heuristically [7, 26, 37] using common sense and domain expertise, or learned statistically [35, 56]. Given the challenges of acquiring reliable domain knowledge in many real-world tasks, recent studies [34, 35, 44, 56] tend to learn the SCM rather than assume one for DG prediction. However, these methods intrinsically perform point estimation of the SCM and select a single set of causal features for

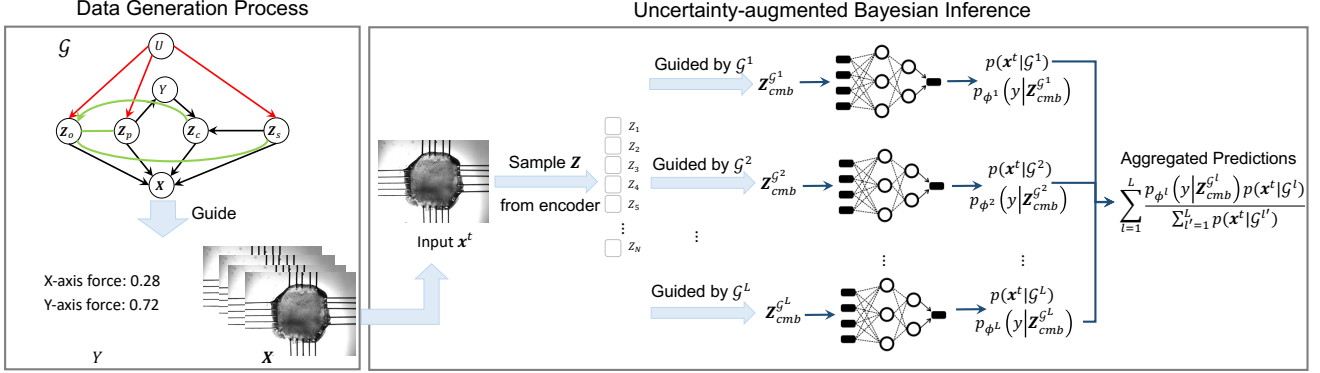


Figure 1. An illustration of using SCM to model the data generation process for vision data and our uncertainty-augmented Bayesian framework, CGU-BAYES. For the SCM causal graph, **black** edges denote known relationships, **green** edges indicate possible acyclic edges, and **red** edges highlight domain-dependent relations influenced by U . In CGU-BAYES, multiple CMB sets are selected across sampled G s, robust predictors are constructed, and their uncertainties are used as weights to aggregate predictions for the final output.

DG prediction. Their DG performance heavily depends on the accuracy of SCM learning and causal feature identification, which can degrade significantly under scarce or noisy data conditions.

To tackle this issue, several works [45, 54] employ Bayesian approaches to capture the posterior distribution of invariant causal features using Bayesian Neural Networks (BNNs). Bayesian inference is known to have several theoretical benefits. Averaging predictions over the posterior distribution reduces the likelihood of overfitting as it does not rely on a single, potentially biased parameter estimate [6]. More importantly, empirical studies [53] and theoretical results from PAC-Bayesian theory [15] support that Bayesian models generalize better than point estimate models. Inspired by these works, we propose integrating Bayesian inference into the causal framework to enhance DG performance. While traditional Bayesian approaches for DG prediction learn and sample from the posterior distribution of invariant features, our method goes a step further by learning and sampling from the posterior distribution of the SCM’s causal graph itself. The corresponding invariant causal features are then selected from the sampled graphs to perform DG prediction. One advantage of our approach is that, by learning the causal graph and selecting features causally related to the target, it enables identification of a feature set that is theoretically guaranteed to be invariant and maximally predictive.

Yet the key advantage of integrating the Bayesian framework is that it enables explicit quantification of uncertainty. Uncertainty has been empirically shown to be crucial for Out-of-Distribution (OOD) generalization [30, 41, 46], as it mitigates the issue of overconfident model predictions. We quantify causal graph uncertainty in the context of prediction tasks, providing insight into the reliability of our approach on specific datasets, the fitness of the learned graphs and selected features in modeling unseen domains, and the

confidence of the resulting predictions. More importantly, we go beyond merely quantifying uncertainty. Although leveraging uncertainty for weighted aggregation is a relatively common practice in Bayesian Deep Learning [2], to the best of our knowledge it has not previously been applied to DG prediction in computer vision tasks. Specifically, we use the quantified uncertainty, which reflects the alignment between the learned causal structures and unseen test domains, as weights in a weighted Bayesian inference scheme. Empirical results show that this uncertainty-augmented inference consistently outperforms the standard Bayesian approach, validating both the effectiveness of our uncertainty quantification and its role in guiding causal feature selection for improved DG performance.

Contributions. In this paper, we make three key contributions: 1) To improve OOD performance, particularly under insufficient or noisy data regimes, we propose a novel framework that performs Bayesian inference over causal graphs to capture underlying data relationships and identify causal features for prediction. 2) Our Bayesian formulation enables uncertainty quantification, providing insights into the reliability of the method on specific datasets, the alignment between learned causal graphs and unseen test domains, and the confidence of the predictions. 3) We further leverage the uncertainty that reflects the alignment between the learned causal graphs and unseen domains as weighting factors in a weighted Bayesian inference scheme. We evaluate our framework on benchmark distribution-shift datasets. Empirical results demonstrate that our method outperforms existing DG approaches and standard Bayesian baselines on challenging datasets while achieving state-of-the-art performance overall.

2. Related Work

Domain Generalization. Domain generalization aims to learn a universally applicable prediction model from one or

multiple observational training domains and leverage it for prediction in unseen test domains. The popular approaches include disentanglement representation learning [1, 8–10], data augmentation [24, 32, 33, 46], meta-learning [29, 47], adversarial training strategies [50, 51], spurious correlations mitigation methods [25, 42] and “mix-up” kinds of strategies [16, 57, 58]. Since we employ both causality and Bayesian inference in the context of DG prediction tasks, we discuss the related works from these two perspectives.

Causal Representation Learning for Domain Generalization. Recent causal representation learning approaches for domain generalization employ SCMs to capture causal relations within data and to select domain-invariant features for robust prediction. One major line of work [3–5, 7, 11, 12, 18–21, 27, 35, 38, 43] constructs SCMs using domain-specific knowledge and derives causal constraints from these models to regularize the learning of causal representations. The differences among these methods mainly lie in the specific causal constraints imposed. However, they typically rely on expert domain knowledge to construct SCMs and lack theoretical guarantees for the learned representations. Consequently, another line of research focuses on learning SCMs rather than assuming them *a priori*. Jointly learning high-level representations and the underlying SCM can be ill-posed and challenging. To address this, several recent studies [35, 56] first learn identifiable high-level representations leveraging identifiable variational autoencoder (iVAE) frameworks, then infer the partial SCM structure via independence tests to select causally related features of the target. While Lu et al. [35] focuses on parent and child variables, Yin et al. [56] extends this to include causal Markov blanket (CMB) features and proposes an efficient CMB discovery method. These approaches consider more general and flexible SCM structures, offering theoretical guarantees for the identified causal features, and demonstrating strong OOD prediction performance. Hence we adopt this line of works as the foundation of our framework. Nevertheless, they rely heavily on independence tests for revealing the SCM structure, which are unreliable under limited or noisy data. This compromises the accuracy of the estimated causal graph and, in turn, the DG performance of the identified causal features. Our method addresses this limitation by performing Bayesian inference over the SCM structure, identifying multiple plausible sets of causal features to improve robustness and DG performance under challenging data conditions.

Bayesian Approaches for Domain Generalization. Several studies have employed Bayesian approaches to address DG tasks under limited observational data. However, existing methods primarily focus on learning the posterior distribution of target representations or model parameters given the observational data. For example, Xiao et al. [54] estimates the distribution of invariant classifier param-

eters using BNNs, Liu et al. [31] introduces a variational Bayesian framework for aligning conditional distributions and marginal label shifts, and Shen et al. [45] infers the posterior distribution of domain-invariant features. In contrast, our approach estimates the posterior distribution of the causal structure given data. By integrating both Bayesian inference and causal discovery in DG tasks, our method enables more theoretically grounded identification of causal features for each Bayesian model, leading to improved DG performance under challenging data conditions. Existing Bayesian DG methods have not fully exploited Bayesian inference for uncertainty quantification. We close this gap by not only quantifying predictive uncertainty, but also clarifying its implications and showing how to effectively leverage it to further improve DG performance.

Bayesian Causal Discovery. Bayesian causal discovery (BCD), instead of performing point-estimation of causal graph from observational data, aims to estimate the posterior distribution of causal graph given data. In our framework, we directly employ existing BCD methods rather than proposing a new one. To this end, we comprehensively evaluate state-of-the-art approaches and select the method with optimal performance. A detailed discussion of related BCD work is provided in Appendix B.3.

3. Bayesian Framework for Domain Generalization Task

In this section, we first introduce the causal framework for DG prediction tasks in Section 3.1, followed by our Bayesian inference formulation in Section 3.2 and uncertainty quantification approach in Section 3.3.

3.1. Causal Framework for Prediction Tasks

The core of constructing a causal framework for prediction tasks lies in employing a SCM to capture the data-generating process among variables of interest. Existing approaches lack consensus on the specific form of the SCM, often adopting models that share general principles but differ in structural details. In this work, we follow the structural settings proposed by Yin et al. [56] due to their practicality and flexibility. An example of an SCM causal graph consistent with these settings is illustrated in Figure 1.

SCM. In Figure 1, $\mathbf{X} \in \mathbb{R}^d$ represents the high-dimensional input data (the images) and Y the target variable for prediction. Domain variable U encodes domain-specific information, which remains constant within a given domain. $\mathbf{Z} \in \mathbb{R}^N$ represents the latent, high-level variables responsible for generating $\mathbf{X}$. Based on their relationships to Y , $\mathbf{Z}$ can be divided into four types: parent variables $\mathbf{Z}_p \in \mathbb{R}^{n_p}$, child variables $\mathbf{Z}_c \in \mathbb{R}^{n_c}$, spouse variables $\mathbf{Z}_s \in \mathbb{R}^{n_s}$, and spurious variables $\mathbf{Z}_o \in \mathbb{R}^{n_o}$. $\mathbf{Z}_p$, $\mathbf{Z}_c$, and $\mathbf{Z}_s$ together constitute the CMB representations $\mathbf{Z}_{cmb}$. Conditioning on this set d-separates the prediction

target Y from other variables in the causal graph, particularly domain-specific variations (Z_o and U), thereby ensuring robustness across environments. Moreover, this set includes the direct causes, direct effects, and indirect causes (through direct effects) of Y , thus capturing maximal information about the target and being inherently predictive.

Multiple domains and their variations. Given the causal structure $\mathcal{G}$, the SCM defines a set of conditional distributions for each variable given its parents, parameterized by Θ , to model the data generation process. Existing studies commonly assume that data from different domains share the same $\mathcal{G}$ but differ in parameters. Specifically, conditional distributions such as $p(y|z_p)$, $p(z_c|y, z_s)$, and $p(x|z)$ are unaffected by the domain variable U and thus considered domain-invariant, while others, including $p(z_o|\pi_o)$, $p(z_p|\pi_p)$, and $p(z_s|\pi_s)$, vary across domains due to their dependence on U . These domain-dependent conditionals cause both the joint distribution over all variables and the marginal distribution $p(x, y)$ to shift across domains.

3.2. Bayesian Inference for Domain Generation Prediction Tasks

Domain generalization tasks aim to leverage knowledge learned from training domains to make accurate predictions on unseen test domains. Let $\mathcal{D}$ denote the observational data from the training domains, and let x^t be an input sample drawn from an unseen test domain. The goal is to construct $p(y|x^t, \mathcal{D})$. In our framework, we recover the SCM causal graph $\mathcal{G}$ from $\mathcal{D}$ along with its associated parameters Θ to construct a universally applicable predictor. Specifically, as illustrated in Eq. (1), we perform Bayesian inference over $\mathcal{G}$ and build an ensemble of predictors corresponding to each sampled $\mathcal{G}$. The predictions are then aggregated to mitigate overfitting to the training domain, avoiding reliance on a single, potentially biased parameter estimate [6]. A detailed derivation is provided in Appendix A.1.

$$p(y|x^t, \mathcal{D}) \propto \mathbb{E}_{\mathcal{G} \sim p(\mathcal{G}|\mathcal{D})} [p(y|x^t, \mathcal{G})p(x^t|\mathcal{G})]. \quad (1)$$

Remark 1. Eq. (1) shows that estimating $p(y|x^t, \mathcal{D})$ requires the graph posterior $p(\mathcal{G}|\mathcal{D})$, the graph-based predictor $p(y|x^t, \mathcal{G})$, and the alignment term $p(x^t|\mathcal{G})$. In practice, we sample candidate graphs from $p(\mathcal{G}|\mathcal{D})$, build a robust predictor for each, and aggregate their predictions weighted by each graph’s compatibility with x^t .

In particular, to obtain $p(\mathcal{G}|\mathcal{D})$, we develop a learning framework that integrates iVAE and Bayesian causal discovery (BCD) using only observational data in vision tasks, as detailed in Section 3.3. In this section, we focus on the remaining two components in Eq. (1).

Robust Predictor $p(y|x^t, \mathcal{G})$. We seek a domain-invariant, transportable predictor for $p(y|x^t, \mathcal{G})$. As discussed in Section 3.1, the CMB representations Z_{cmb} are both robust and

highly predictive. Thus, we approximate $p(y|x^t, \mathcal{G})$ by extracting Z_{cmb} and using it to predict Y as in Eq. (2). Since the CMB set is determined by $\mathcal{G}$, we denote it by $Z_{cmb}^{\mathcal{G}}$. Please refer to Appendix A.2 for the detailed derivations.

$$p(y|x^t, \mathcal{G}) \approx p(y|(z_{cmb}^{\mathcal{G}})^*), (z_{cmb}^{\mathcal{G}})^* = \arg \max p(z_{cmb}^{\mathcal{G}}|x^t, \mathcal{G}). \quad (2)$$

Remark 2. Eq. (2) reduces estimating $p(y|x^t, \mathcal{G})$ to extracting the CMB representations from the input and learning $p(y|z_{cmb}^{\mathcal{G}})$.

Alignment Term $p(x^t|\mathcal{G})$ via Uncertainty. $p(x^t|\mathcal{G})$ measures how well an unseen input x^t aligns with a sampled causal graph $\mathcal{G}$. Directly estimating it via the full SCM joint and marginalizing non- X variables is infeasible, due to missing U and Y for x^t and unknown test-domain parameters. Instead, we approximate $p(x^t|\mathcal{G})$ using the **epistemic uncertainty of the predictor** $p(y|x^t, \mathcal{G})$, termed SINGLE-GRAPH PREDICTION UNCERTAINTY and denoted $\mathcal{U}_e(x^t|\mathcal{G})$. Epistemic (model) uncertainty arises from limited knowledge about the correct model due to insufficient data, and unlike aleatoric uncertainty, it can be reduced with more informative data. Prior work [14, 22, 23] emphasizes its inverse relationship with data density. However, there is no universal functional form linking $\mathcal{U}_e(x^t|\mathcal{G})$ and $p(x^t|\mathcal{G})$; this relationship depends on the specific model and uncertainty estimation method. In this work, we adopt a generic formulation¹ to estimate $p(x^t|\mathcal{G})$ from $\mathcal{U}_e(x^t|\mathcal{G})$, as shown in Eq. (3).

$$p(x^t|\mathcal{G}) \propto e^{-\alpha \mathcal{U}_e(x^t|\mathcal{G})}. \quad (3)$$

Here, α controls the dispersion of the weights: by tuning it, we avoid both nearly uniform weights (a naive ensemble) and overly peaked weights (effectively a single deterministic model), which would undermine the benefits of the Bayesian framework.

Remark 3. Low $\mathcal{U}_e(x^t|\mathcal{G})$ indicates strong alignment between x^t and $\mathcal{G}$, so the predictor trained under $\mathcal{G}$ generalizes well to x^t and contributes more to the aggregated prediction. Thus, our Bayesian approach naturally prioritizes predictors that best generalize to the domain of x^t .

3.3. Uncertainty Quantification

SINGLE-GRAPH PREDICTION UNCERTAINTY $\mathcal{U}_e(x^t|\mathcal{G})$. As noted above, we approximate $p(x^t|\mathcal{G})$ using the epistemic prediction uncertainty of $p(y|x^t, \mathcal{G})$. Eq. (4) first shows that $p(y|x^t, \mathcal{G})$ contains two sources of uncertainty,

$$p(y|x^t, \mathcal{G}) = \int \underbrace{p(y|z_{cmb}^{\mathcal{G}})}_{\text{aleatoric}} \underbrace{p(z_{cmb}^{\mathcal{G}}|x^t, \mathcal{G})}_{\text{epistemic}} dz_{cmb}^{\mathcal{G}}. \quad (4)$$

¹Using an exponential function with a tunable scaling factor is a standard practice in related literature. In practice, we normalize the estimated $p(x^t|\mathcal{G})$ values across different $\mathcal{G}$ s and use them as weights for the corresponding predictors.

We treat the predictor for $p(y|\mathbf{x}^t, \mathcal{G})$ as an epistemic neural network [39], where the distribution of $\mathbf{z}_{cmb}^{\mathcal{G}}$ induces model stochasticity. Following the entropy-based uncertainty quantification in Osband et al. [39], we obtain epistemic uncertainty $\mathcal{U}_e(\mathbf{x}^t|\mathcal{G})$ as the difference between total and aleatoric uncertainty, as shown in Eq. (5). Detailed derivations are provided in Appendix A.3.

$$\begin{aligned} \mathcal{U}_e(\mathbf{x}^t|\mathcal{G}) &= \underbrace{\mathcal{H}[p(y|\mathbf{x}^t, \mathcal{G})]}_{\text{Total uncertainty}} - \underbrace{\mathbb{E}_{p(\mathbf{z}_{cmb}^{\mathcal{G}}|\mathbf{x}^t, \mathcal{G})}[\mathcal{H}[p(y|\mathbf{z}_{cmb}^{\mathcal{G}})]]}_{\text{Aleatoric uncertainty}} \\ &\approx \mathcal{H}\left[\frac{1}{S} \sum_{s=1}^S p(y|(\mathbf{z}_{cmb}^{\mathcal{G}})^s)\right] - \frac{1}{S} \sum_{s=1}^S \mathcal{H}[p(y|(\mathbf{z}_{cmb}^{\mathcal{G}})^s)], \quad (5) \\ (\mathbf{z}_{cmb}^{\mathcal{G}})^s &\sim p(\mathbf{z}_{cmb}^{\mathcal{G}}|\mathbf{x}^t, \mathcal{G}). \end{aligned}$$

$\mathcal{H}[\cdot]$ denotes entropy. S^2 is the number of sampled CMB representations.

Remark 4. Our uncertainty quantification in Eq. (5) is efficient because it only relies on samples of the CMB representations and the conditional distribution $p(y|\mathbf{z}_{cmb}^{\mathcal{G}})$, both of which are already computed when estimating $p(y|\mathbf{x}^t, \mathcal{G})$ in Eq. (2).

We also assess the reliability of the aggregated Bayesian predictions.

BAYESIAN INFERENCE UNCERTAINTY $\mathcal{U}_e(\mathbf{x}^t|\mathcal{D})$. In contrast to $\mathcal{U}_e(\mathbf{x}^t|\mathcal{G})$, $\mathcal{U}_e(\mathbf{x}^t|\mathcal{D})$ measures the epistemic uncertainty of the full Bayesian predictor $p(y|\mathbf{x}^t, \mathcal{D}) = \int_{\mathcal{G}} p(y|\mathbf{x}^t, \mathcal{G}) p(\mathcal{G}|\mathbf{x}^t, \mathcal{D}) d\mathcal{G}$. Here, $p(y|\mathbf{x}^t, \mathcal{G})$ encodes aleatoric uncertainty, while $p(\mathcal{G}|\mathbf{x}^t, \mathcal{D})$ captures epistemic uncertainty. Analogous to $\mathcal{U}_e(\mathbf{x}^t|\mathcal{G})$, we estimate $\mathcal{U}_e(\mathbf{x}^t|\mathcal{D})$ via an entropy-based decomposition, as the difference between total and aleatoric uncertainty (Eq. (6)).

$$\mathcal{U}_e(\mathbf{x}^t|\mathcal{D}) \approx \mathcal{H}\left[\frac{\sum_{l=1}^L p(y|\mathbf{x}^t, \mathcal{G}^l) p(\mathbf{x}^t|\mathcal{G}^l)}{\sum_{l=1}^L p(\mathbf{x}^t|\mathcal{G}^l)}\right] - \left[\frac{\sum_{l=1}^L \mathcal{H}[p(y|\mathbf{x}^t, \mathcal{G}^l)] p(\mathbf{x}^t|\mathcal{G}^l)}{\sum_{l=1}^L p(\mathbf{x}^t|\mathcal{G}^l)}\right], \quad (6)$$

where $\mathcal{G}^l \sim p(\mathcal{G}|\mathcal{D})$. The terms $p(y|\mathbf{x}^t, \mathcal{G}^l)$ and $p(\mathbf{x}^t|\mathcal{G}^l)$ are estimated using Eq. (2) and Eq. (3), respectively. The full derivation is provided in Appendix A.4. Intuitively, $\mathcal{U}_e(\mathbf{x}^t|\mathcal{D})$ characterizes the confidence of our Bayesian predictions. Moreover, given the inferred graph posterior $p(\mathcal{G}|\mathcal{D})$, we can assess the variability among sampled graphs. **This variability is a key indicator of whether the Bayesian approach is likely to outperform non-Bayesian methods on a given dataset.**

CAUSAL GRAPH UNCERTAINTY $\mathcal{U}(\mathcal{G})$. The effectiveness of Bayesian inference in Eq. (1) depends critically on the variability of the sampled causal graphs. If the sampled $\mathcal{G}$ s are identical or lie within the same Markov equivalence class, the resulting $\mathbf{z}_{cmb}^{\mathcal{G}}$ and predictors $p(y|\mathbf{z}_{cmb}^{\mathcal{G}})$ coincide, so the Bayesian procedure effectively reduces to a single

deterministic model, offering little benefit. This typically occurs when abundant, high-quality data are available [13], where deterministic methods already perform well. In contrast, under scarce or noisy data, existing BCD methods tend to produce more diverse causal graphs that are less likely to encode domain-specific noise or bias, thereby improving generalization to unseen domains; our Bayesian approach is particularly advantageous in this regime. Practically, we quantify $\mathcal{U}(\mathcal{G})$ by measuring variation in CMB selection across sampled graphs: for each variable Z_i , we record 1 if $Z_i \in \mathbf{Z}_{cmb}$ and 0 otherwise to form a $|\mathbf{Z}|$ -dimensional binary vector per graph, and then compute $\mathcal{U}(\mathcal{G})$ as the average pairwise distance between these vectors over all sampled graphs.

4. Proposed Algorithm: CGU-Bayes

Guided by the theoretical results in Section 3, we propose a novel approach for performing domain generalization prediction, denoted as Causal Graph Uncertainty-guided Bayesian Inference procedure (CGU-BAYES).

Algorithm 1 CGU-BAYES Training Procedures

- 1: **Input:** Training domain observational data for variables $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{U}$.
- 2: **Phase 1:** Learn the $\mathbf{Z} \in \mathbb{R}^N$ via NF-iVAE [35].
- 3: Train the NF-iVAE, with its encoder $q_{\psi}(\mathbf{Z}|\mathbf{X})$ and decoder $p_{\theta}(\mathbf{X}|\mathbf{Z})$, by minimizing the objective in Eq. (7).

$$\begin{aligned} \mathcal{L}_{\text{iVAE}} &= \underbrace{-\mathbb{E}_{q_{\psi}(\mathbf{z}|\mathbf{x})}[\log p_{\theta}(\mathbf{x}|\mathbf{z}) + \log p_{T,\lambda}(\mathbf{z}|\mathbf{y}, \mathbf{u}) - \log q_{\psi}(\mathbf{z}|\mathbf{x})]}_{\mathcal{L}_{\text{ELBO}}} \\ &\quad + \underbrace{\mathbb{E}_{q_{\psi}(\mathbf{z}|\mathbf{x})}[\|\nabla_{\mathbf{z}} q_{\psi}(\mathbf{z}|\mathbf{x}) - \nabla_{\mathbf{z}} p_{T,\lambda}(\mathbf{z}|\mathbf{y}, \mathbf{u})\|^2]}_{\mathcal{L}_{\text{SM}}} \end{aligned} \quad (7)$$

- 4: Sample observations of $\mathbf{Z}$ using the trained encoder $q_{\psi}(\mathbf{Z}|\mathbf{X})$. Obtain observations of variable set $\mathcal{D}^c = \{\mathbf{Z}, \mathbf{Y}, \mathbf{U}\}$ for learning $\mathcal{G}$.
- 5: **Phase 2:** Estimate $p(\mathcal{G}|\mathcal{D}^c)$ via DAG-GFlowNet [13].
- 6: Sample $\mathcal{G}^l \sim p(\mathcal{G}|\mathcal{D}^c)$, $l = 1, 2, \dots, L$.
- 7: **Phase 3:** Construct the robust predictor for each $\mathcal{G}^l$.
- 8: **for** $l \leftarrow 1$ **to** L **do**
- 9: Identify the CMB representations $\mathbf{Z}_{cmb}^{\mathcal{G}^l}$ from $\mathcal{G}^l$.
- 10: Train the predictor $p_{\phi^l}(Y|\mathbf{Z}_{cmb}^{\mathcal{G}^l})$ by minimizing the objective in Eq. (8).

$$\hat{\phi}^l = \arg \min_{\phi^l} \mathbb{E}_{\mathbf{x}, \mathbf{y}} \mathcal{L}_{\text{pred}}(y, p_{\phi^l}(y|\mathbf{z}_{cmb}^{\mathcal{G}^l})), \mathbf{z}_{cmb}^{\mathcal{G}^l} \sim q_{\psi}(\mathbf{z}|\mathbf{x}) \quad (8)$$

- 11: **end for**
 - 12: **return** Encoder $q_{\hat{\psi}}(\mathbf{Z}|\mathbf{X})$, the CMB representations $\mathbf{Z}_{cmb}^{\mathcal{G}^l}$ in terms of their indices, and learned predictors $\{p_{\hat{\phi}^l}(Y|\mathbf{Z}_{cmb}^{\mathcal{G}^l})\}_{l=1}^L$ for inference.
-

²Since the expectation in Eq. (5) is generally intractable, we approximate it with a sample average.

Training Procedure. We outline our training procedures in Algorithm 1. We build on the NF-iVAE framework [35], as Yin et al. [56] have shown it can learn the fully identifiable representation set $\mathcal{Z}$ under our SCM. To match our inference procedure, we replace the original encoder $q(\mathcal{Z}|\mathbf{X}, Y, U)$ with $q(\mathcal{Z}|\mathbf{X})$. Detailed derivations of $\mathcal{L}_{\text{ELBO}}$ and the resulting training steps are provided in Appendix A.5. We also prove that the learned representation $\mathcal{Z}$ remains component-wise identifiable in theory (Appendix B.2) and exhibits a comparable level of identifiability in practice (Appendix C.5), which is crucial for reproducibility of $\mathcal{Z}$ and for transferring the identified CMB sets and encoder across domains. The quality of the learned $\mathcal{Z}$ is crucial: if $\mathcal{Z}$ is not fully disentangled, the BCD will yield an inaccurate posterior over $\mathcal{G}$, causing the selected CMB features to be contaminated with domain-variant information. For BCD, we comprehensively evaluate several state-of-the-art methods on vision datasets, and select the one with the best overall accuracy and reasonable runtime: DAG-GFlowNet [13]. Because our SCM imposes structural assumptions on the causal graphs (Assumption ?? in Appendix B.1), we discard sampled graphs that violate these assumptions³ and retain only those that satisfy them. For efficiency, we sample and use only the first $L = 5$ valid graphs. We train robust predictors with cross-entropy loss $\mathcal{L}_{\text{pred}}(\cdot)$ for classification, and with the Gaussian negative log-likelihood for regression, where the network predicts the mean and variance of $p(y|z_{\text{cmb}}^{\mathcal{G}^l})$.

Inference Procedure. For a test input $\mathbf{x}^t$, we first compute $\mathcal{U}_e(\mathbf{x}^t|\mathcal{G}^l)$ for each predictor via Eq. (5). Concretely, we sample $\mathcal{Z}$ from the trained encoder $q_{\hat{\phi}}(\mathbf{z}|\mathbf{x}^t)$, extract $\mathcal{Z}_{\text{cmb}}^{\mathcal{G}^l}$ using the CMB indices identified during training, and then compute the weight $p(\mathbf{x}^t|\mathcal{G}^l)$ using Eq. (3). We finally normalize $\{p(\mathbf{x}^t|\mathcal{G}^l)\}_{l=1}^L$ and combine them with $\{p_{\hat{\phi}^l}(y|\mathbf{z}_{\text{cmb}}^{\mathcal{G}^l})\}_{l=1}^L$ in Eq. (1) to obtain the final prediction.

Complexity Analysis. For the training procedure, the iVAE and BCD phases scale quadratically with respect to the latent representation dimension and the number of variables in $\mathcal{G}$, respectively. Specifically, the computational complexities are $\mathcal{O}(|\mathcal{Z}|^2)$ for Phase 1 and $\mathcal{O}((|\mathcal{Z}| + 2)^2)$ for Phase 2. The training of the L predictors in Phase 3 scales linearly with the size of the corresponding CMB sets, i.e., $\mathcal{O}(\sum_{l=1}^L |\mathcal{Z}_{\text{cmb}}^{\mathcal{G}^l}|)$. For the inference procedure, let M_e and M_p denote the computational costs of the encoder and the predictor per sample, respectively. During inference, we sample S latent representations from the encoder for each input and evaluate L graph-specific predictors conditioned on their corresponding CMB representations. The resulting complexity per test sample is roughly $\mathcal{O}(SM_e + LSM_p)$. We also report wall-clock training times and inference latencies in the Appendix C.7, showing that (1) overall infer-

ence costs are comparable, with only a marginal gap when parallel inference is used, and (2) the extra training time is a well-justified trade-off for improved accuracy, especially in challenging data regimes.

A key requirement of our approach is a set of plausible causal graphs over the latent representations, target, and domains. As clarified in Remark 4, this can be relaxed to identifying a set of plausible latent features that are causally related to the target. However, neither causal graphs nor causal representations are typically available, especially in real-world vision datasets. To address this, we incorporate iVAE and BCD modules into our training pipeline, which provide a practical and reliable way to obtain such sets while building on existing techniques. We also study an alternative variant that uses the “causal representations” and predictors learned by several state-of-the-art baselines, estimates their uncertainties, and aggregates their predictions; we denote this variant CGU-BAYES++ and evaluate both versions empirically.

5. Experiments

We investigate the effectiveness of our proposed algorithm CGU-BAYES, in terms of OOD prediction performance. In particular, we highlight our performance gain on the **challenging limited and noisy data settings**. We also show that our estimated uncertainties can provide intuition of our predictions and further improve the OOD predictions.

Dataset. We evaluate on one synthetic dataset (Colored MNIST, CMNIST), one challenging real-world dataset (Biaxial Loading Test, BLT), and three standard distribution shift benchmarks (PACS, VLCS, OfficeHome). **CMNIST** [7] is constructed by coloring grayscale digits so that color is spuriously correlated with the label; the task is to classify whether a digit is less than 5. Each of the two training domains has 30,000 images, and the test domain has 10,000 images, with substantially different color-label correlations inducing distribution shift. We further create more challenging insufficient and noisy data variants, described below and in Appendix C.2. **BLT** is a biaxial loading dataset collected from a tricuspid valve anterior leaflet (TVAL) specimen of a porcine heart. Seven biaxial stress-ratio protocols ($P_{11} : P_{22} = 1:1, 1:0.66, 1:0.33, 0.66:1, 0.33:1, 0.05:1, 1:0.1$) yield 3,539–4,175 samples per protocol and 26,524 samples in total, with each protocol treated as a domain. The task is to predict x - and y -axis stresses from images, under shifts induced by changing stress ratios, limited samples, and gradual microscale tissue damage that introduces non-uniform, heteroskedastic noise (see Appendix C.3 for examples). **PACS** [28] contains 9,991 images from 4 domains (Photo, Art painting, Cartoon, Sketch) and 7 classes. **VLCS** [48] has 10,729 images from 4 domains (PASCAL VOC 2007, LabelMe, Caltech, Sun) and 5 classes. **OfficeHome** [49] includes 15,588 images from

³For example, the graphs with the presence of link $U \rightarrow Y$.

Table 1. Empirical results on BLT dataset in terms of OOD relative MSE (%) against state-of-the-art baselines. We report CGU-BAYES with and without applying uncertainty as weights, and the causal graph uncertainty $\mathcal{U}(\mathcal{G})$. We set $|\mathcal{Z}| = 50$.

Algorithms	Out-of-distribution relative MSE (%) $\downarrow$							Average
	1:1	1:0.66	1:0.33	0.66:1	0.33:1	0.05:1	1:0.1	
ERM	68.3	21.5	40.4	32.6	46.5	111.0	214.5	76.4
CORAL	67.7	21.4	34.1	34.5	36.1	90.5	193.0	68.2
IRM	65.3	21.1	38.8	35.1	45.8	99.0	200.1	72.2
CAUSALREP	63.5	20.9	36.5	31.8	39.5	98.1	198.0	69.8
CTrans	60.9	20.5	35.0	31.0	30.8	85.1	185.6	64.1
iCaRL	60.7	22.1	38.5	36.1	35.2	89.2	185.3	66.7
CMBRL	58.3	20.1	36.5	33.2	28.2	85.0	179.2	62.9
BITEBAYES	65.3	21.1	35.8	32.1	26.8	89.0	182.1	64.6
PTG	63.2	20.8	39.4	31.7	35.4	98.5	187.5	68.1
CGU-BAYES (w/o UG)	50.7	21.4	32.7	33.3	33.5	80.6	166.0	59.7
CGU-BAYES	41.4	18.4	29.6	30.6	25.7	73.2	150.2	54.2
$\mathcal{U}(\mathcal{G})$	12.3	3.7	6.2	5.3	6.7	25.1	37.6	-

4 domains (Artistic, Clipart, Product, Real World) and 65 classes. For BLT, PACS, VLCS, and OfficeHome, we follow the standard leave-one-domain-out protocol: one domain is held out for testing and the remaining domains are used for training. Implementation details are provided in Appendix C.1, and ablation studies on hyperparameters are discussed in Appendix C.6.

5.1. Results on Challenging Synthetic Data

We aim to demonstrate that CGU-BAYES remains robust under challenging scenarios where existing baselines fail. To this end, we design two protocols on synthetic CMNIST: 1) an insufficient data setting and 2) a noisy data setting. In the insufficient data setting, we limit the number of training images and evaluate on the full test domain, sampling $M \in \{500, 1000, 2000, 5000\}$ images from each training domain. In the noisy data setting, we corrupt the training images by adding Gaussian noise to each pixel, where the noise is drawn from $\mathcal{N}(0, \sigma^2)$ with $\sigma^2 \in \{0.01, 0.1, 1, 10\}$. Figure 2 presents the empirical results of our CGU-BAYES

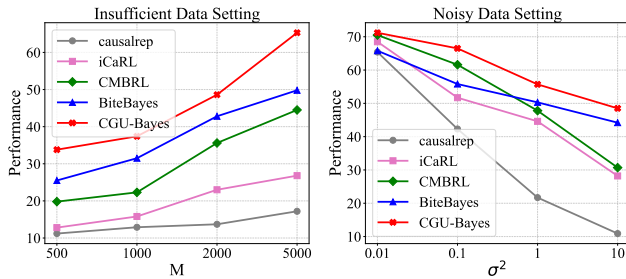


Figure 2. Evaluations on challenging data scenarios. The left figure shows performance under insufficient data, while the right figure shows performance under noisy data.

on these two challenging data settings. We compare it with several state-of-the-art causal domain generalization baselines, iCaRL [35], CMBRL [56], and CAUSALREP [52], as well as the Bayesian baseline BITEBAYES [54]. Empirical results show that under a limited number of training domain images and high noise levels, existing causal domain generalization methods iCaRL, CMBRL, and CAUSALREP fail to match the performance of Bayesian approaches. Our

Table 2. Empirical results on CMNIST in terms of prediction accuracy (%) against state-of-the-art baselines. We set $|\mathcal{Z}| = 32$.

Algorithms	Prediction Accuracy (%)	
	In-distribution ($\uparrow$)	OOD ($\uparrow$)
ERM	85.2 $\pm$ 0.5	10.8 $\pm$ 0.2
ROBUST MIN MAX	84.3 $\pm$ 0.4	10.9 $\pm$ 0.5
F-IRM GAME	63.4 $\pm$ 1.1	60.0 $\pm$ 2.7
V-IRM GAME	64.0 $\pm$ 1.0	49.2 $\pm$ 3.4
RSC	76.3 $\pm$ 0.3	20.5 $\pm$ 1.0
IRM	59.3 $\pm$ 4.4	62.8 $\pm$ 9.6
CAUSALREP	70.1 $\pm$ 1.5	68.6 $\pm$ 5.5
CTrans	76.9 $\pm$ 0.8	72.5 $\pm$ 1.1
iCaRL	70.6 $\pm$ 0.8	68.8 $\pm$ 1.5
CMBRL	76.4 $\pm$ 0.8	74.1 $\pm$ 1.2
CGU-BAYES	76.0 $\pm$ 1.2	74.1 $\pm$ 1.3
CGU-BAYES++	85.2 $\pm$ 0.3	74.2 $\pm$ 1.3

CGU-BAYES consistently achieves the best results, outperforming the existing Bayesian baseline by 8.3% and 5.9% with 500 and 1000 images, respectively, and by 4.3% and 5.4% under severe noise settings with $\sigma^2 = 10$ and 1.

Although the results in Figure 2 are promising, one may question whether these findings generalize beyond CMNIST, given that the tasks are relatively simple and the data is synthetic. To address this concern, we introduce the BLT dataset, a real-world, complex and naturally collected datasets designed to evaluate domain generalization under more complex and realistic conditions. Using BLT, we can test the robustness and applicability of our CGU-BAYES approach in practical scenarios.

5.2. Results on Complex Realistic Data

In Table 1, we compare against ERM, state-of-the-art causal approaches (IRM [7], CAUSALREP, CTRANS [37], iCaRL, and CMBRL), and two Bayesian approaches (BITEBAYES and PTG [45]). Since the BLT dataset involves a regression task, we evaluate OOD performance using relative MSE. The results indicate that the BLT domains differ significantly in their levels of distribution shift. Some domains, such as 1:0.66 and 0.66:1, are easier to generalize and show relatively low ERM errors. Others, including 1:1, 0.05:1, and 1:0.1, exhibit stronger shifts and are more difficult to generalize.

Overall, CGU-BAYES performs consistently well across all domains. On domains with stronger shifts and more difficult generalization, the performance gains are substantially larger. For example, on 1:1, 0.05:1, and 1:0.1, our improvements over the best baselines are 9.3%, 7.4%, and 15.8%. On domains with mild shifts such as 1:0.66 and 0.66:1, where most methods already perform reasonably well, the gains remain positive but smaller.

The results also demonstrate the effectiveness of our quantified uncertainties. First, incorporating uncertainty-guided weights improves performance compared with the version without uncertainty guidance. On challenging domains the gains are substantial, for example on 0.05:1 where the error decreases from 80.6 to 73.2. Second, the causal

graph uncertainty $\mathcal{U}(\mathcal{G})$ shows a clear pattern: uncertainties are small on easy domains and large on difficult ones, which aligns with the intuition that easy domains yield similar CMB sets across sampled graphs. This correspondence between the magnitude of the distribution shift and the level of graph uncertainty indicates that the uncertainty can signal when our method is likely to perform well. This supports our claim in Section 3.3.

5.3. Results on Benchmark Domain Shifts Data

We then compare CGU-BAYES with a broad set of DG methods on four standard distribution shift benchmarks: CMNIST, PACS, VLCS, and OfficeHome. Table 2 reports results on the clean and complete CMNIST dataset, evaluated against baselines including IRM and its variants F-IRM and V-IRM [3], the ROBUST MIN MAX method [3], RSC [17], CAUSALREP, CTRANS, and ICARL. As shown in Table 2, CGU-BAYES performs on par with these baselines, while the alternative version CGU-BAYES++ achieves the best accuracy, although by a modest margin. A likely reason is that CMNIST provides sufficient and noiseless images for training, which leads to low variability in the CMB set. Consequently, our method relies on nearly the same CMB features as deterministic approaches such as CMBRL, resulting in similar OOD performance. For CGU-BAYES++, we aggregate predictions from selected feature sets and predictors produced by CMBRL, ICARL, CAUSALREP, and ERM. Inputs from different domains tend to favor different models, allowing CGU-BAYES++ to adaptively rely on the most suitable model, achieving strong performance across both training and testing domains.

We report the full evaluation of our approaches on PACS, VLCS, and OfficeHome in Table 3, comparing against a broad suite of DG baselines. In addition to the causal and Bayesian methods discussed earlier, we include widely used DG approaches such as GroupDRO [44] and Mixup [55].

Overall, CGU-BAYES performs on par with the strongest baselines, while CGU-BAYES++ achieves the best results, with only a modest margin on PACS. Together with the CMNIST findings, this suggests that standalone CGU-BAYES does not consistently yield gains on all benchmarks, but its boosted variant built on multiple state-of-the-art methods reliably matches or surpasses them. Comparing CGU-BAYES++ to a uniform ensemble of its base models (iCaRL, CMBRL, CAUSALREP, ERM), the uniform average attains 55.7, 80.5, and 87.2 accuracy on CMNIST, VLCS, and PACS, far below our weighted scheme (74.2, 82.9, 90.5). This highlights the importance of our weighting mechanism, which adaptively emphasizes the most suitable base models: unlike uniform averaging, which is dragged down by weak OOD methods such as ERM, our approach downweights ERM under shift while still exploiting its strength in-distribution.

Table 3. Empirical results on VLCS, PACS, and OfficeHome datasets in terms of OOD prediction accuracy (%). We report the average performance over all domains. Please refer to Tables 5 and 6 in the Appendix C.4 for detailed results.

Algorithms	VLCS ($\uparrow$)	PACS ($\uparrow$)	OfficeHome ($\uparrow$)
ERM	77.2	83.5	66.5
GROUPDRO	77.9	84.4	66.0
RSC	77.5	85.2	65.5
MIXUP	77.7	84.6	68.4
FACT	78.3	88.2	66.6
IRM	78.5	85.5	64.3
SAGNET	77.5	86.3	68.1
ICARL	81.8	88.7	-
CASN	79.1	87.3	68.1
CIRL	78.5	90.1	67.1
CMBRL	82.3	89.1	67.1
BITEBAYES	79.1	85.5	66.4
PTG	76.1	83.7	61.6
CGU-BAYES	82.5	89.5	69.5
CGU-BAYES++	82.9	90.5	69.5

5.4. Evaluation of Uncertainty

We have demonstrated the effectiveness of single graph uncertainty $\mathcal{U}_e(\mathbf{x}^t|\mathcal{G})$ in improving OOD predictions and of causal graph uncertainty $\mathcal{U}(\mathcal{G})$ in indicating when our approach is likely to perform well on a specific dataset. In this section, we evaluate Bayesian inference uncertainty $\mathcal{U}_e(\mathbf{x}^t|\mathcal{D})$ by using it to filter out low-confidence predictions in the test domain. Figure 3 reports the OOD prediction accuracy after discarding the top 5%, 10%, and 15% of predictions with the highest uncertainty. The results indicate that removing predictions with high uncertainty consistently improves OOD accuracy.

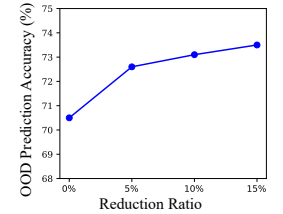


Figure 3. The evaluation of Bayesian inference uncertainty $\mathcal{U}_e(\mathbf{x}^t|\mathcal{D})$.

6. Conclusion

We propose CGU-BAYES, a framework that performs Bayesian inference over causal graphs to perform domain generalization prediction. Through extensive experiments on a wide-range of synthetic and real datasets, we show that our approach consistently improves OOD generalization, particularly in **challenging scenarios with limited or noisy data**. We demonstrate that quantified uncertainties at both the single graph and graph ensemble levels provide valuable guidance for prediction confidence and model reliability. Furthermore, our boosted variant, CGU-BAYES++, effectively aggregates predictions across state-of-the-art models to achieve optimal performance across multiple domains. Overall, our results highlight the importance of combining causal structure learning with Bayesian uncertainty modeling for robust domain generalization, offering a practical and interpretable framework applicable to both synthetic and complex real-world datasets.

Acknowledgment

This work was supported in part by the National Science Foundation award IIS 2236026. N. Yin was partially supported by the AFOSR grant FA9550-22-1-0197. Y. Yu was funded by the NSF grant DMS 2436624. C-H. Lee would like to acknowledge the support from the National Institute of Health under award 1R01GM157589-01. Portions of this research were conducted on Lehigh University’s Research Computing infrastructure partially supported by NSF Award 2019035.

References

- [1] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980, 2018. 3
- [2] Idan Achituve, Idit Diamant, Arnon Netzer, Gal Chechik, and Ethan Fetaya. Bayesian uncertainty for gradient aggregation in multi-task learning. *arXiv preprint arXiv:2402.04005*, 2024. 2
- [3] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization game. In *International Conference on Machine Learning*, 2020. 3, 8
- [4] Kartik Ahuja, Ethan Caballero, Dinghui Zhang, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. In *Advances in Neural Inf. Proc. Systems*, 2021.
- [5] Kartik Ahuja, Jun Wang, Amit Dhurandhar, Karthikeyan Shanmugam, and Kush R. Varshney. Empirical or invariant risk minimization? a sample complexity perspective. In *International Conference on Learning Representations*, 2021. 3
- [6] Mark Andrews and Thom Baguley. Bayesian data analysis. *The Cambridge encyclopedia of child development*, pages 165–169, 2017. 2, 4
- [7] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 1, 3, 6, 7
- [8] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. 3
- [9] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β -vae. *arXiv preprint arXiv:1804.03599*, 2018.
- [10] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018. 3
- [11] Peng Cui and Susan Athey. Stable learning establishes some common ground between causal inference and machine learning. *Nature Machine Intelligence*, 4(2):110–115, 2022. 3
- [12] Peng Cui, Zheyang Shen, Sheng Li, Liuyi Yao, Yaliang Li, Zhixuan Chu, and Jing Gao. Causal inference meets machine learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3527–3528, 2020. 3
- [13] Tristan Deleu, António Góis, Chris Emezue, Mansi Rankawat, Simon Lacoste-Julien, Stefan Bauer, and Yoshua Bengio. Bayesian structure learning with generative flow networks. In *Uncertainty in Artificial Intelligence*, pages 518–528. PMLR, 2022. 5, 6
- [14] Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International conference on machine learning*, pages 1184–1193. PMLR, 2018. 4
- [15] Pascal Germain, Francis Bach, Alexandre Lacoste, and Simon Lacoste-Julien. Pac-bayesian theory meets bayesian inference. *Advances in Neural Information Processing Systems*, 29, 2016. 2
- [16] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019. 3
- [17] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 124–140. Springer, 2020. 8
- [18] Kasra Jalaldoust and Elias Bareinboim. Transportable representations for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12790–12800, 2024. 3
- [19] Kasra Jalaldoust, Alexis Bellot, and Elias Bareinboim. Partial transportability for domain generalization. *Advances in Neural Information Processing Systems*, 37:137768–137805, 2024.
- [20] Dominik Janzing. Causal regularization. *Advances in Neural Information Processing Systems*, 32, 2019.
- [21] Yibo Jiang and Victor Veitch. Invariant and transportable representations for anti-causal domain shifts. *arXiv preprint arXiv:2207.01603*, 2022. 3
- [22] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 4
- [23] Salman Khan, Munawar Hayat, Syed Waqas Zamir, Jianbing Shen, and Ling Shao. Striking the right balance with uncertainty. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 103–112, 2019. 4
- [24] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020. 3
- [25] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022. 3

- [26] Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, and Kun Zhang. Partial disentanglement for domain adaptation. In *International Conference on Machine Learning*, pages 11455–11472. PMLR, 2022. 1
- [27] Masanori Koyama and Shoichiro Yamaguchi. When is invariance useful in an out-of-distribution generalization problem? *arXiv preprint arXiv:2008.01883*, 2020. 3
- [28] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. 6
- [29] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 3
- [30] Xiaotong Li, Yongxing Dai, Yixiao Ge, Jun Liu, Ying Shan, and Ling-Yu Duan. Uncertainty modeling for out-of-distribution generalization. *arXiv preprint arXiv:2202.03958*, 2022. 2
- [31] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1013–1023, 2021. 3
- [32] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019. 3
- [33] Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pages 6348–6359. PMLR, 2020. 3
- [34] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Schölkopf, and Léon Bottou. Discovering causal signals in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6979–6987, 2017. 1
- [35] Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*, 2021. 1, 3, 5, 6, 7
- [36] Chengzhi Mao, Augustine Cha, Amogh Gupta, Hao Wang, Junfeng Yang, and Carl Vondrick. Generative interventions for causal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3947–3956, 2021.
- [37] Chengzhi Mao, Kevin Xia, James Wang, Hao Wang, Junfeng Yang, Elias Bareinboim, and Carl Vondrick. Causal transportability for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7521–7531, 2022. 1, 7
- [38] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8690–8699, 2021. 3
- [39] Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Morteza Ibrahimi, Xiuyuan Lu, and Benjamin Van Roy. Epistemic neural networks. *Advances in Neural Information Processing Systems*, 36, 2024. 5
- [40] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016. 1
- [41] Fengchun Qiao and Xi Peng. Uncertainty-guided model generalization to unseen domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6790–6800, 2021. 2
- [42] Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. In *International Conference on Machine Learning*, pages 28448–28467. PMLR, 2023. 3
- [43] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021. 3
- [44] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 1, 8
- [45] Shiyu Shen, Bin Pan, Tianyang Shi, Tao Li, and Zhenwei Shi. Bayesian domain invariant learning via posterior generalization of parameter distributions. *arXiv preprint arXiv:2310.16277*, 2023. 2, 3, 7
- [46] Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Weakly supervised disentangled generative causal representation learning. *The Journal of Machine Learning Research*, 23(1):10994–11048, 2022. 2, 3
- [47] Yang Shu, Zhangjie Cao, Chenyu Wang, Jianmin Wang, and Mingsheng Long. Open domain generalization with domain-augmented meta-learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9624–9633, 2021. 3
- [48] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011. 6
- [49] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017. 6
- [50] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018. 3

- [51] Haotao Wang, Chaowei Xiao, Jean Kossaifi, Zhiding Yu, Anima Anandkumar, and Zhangyang Wang. Augmax: Adversarial composition of random augmentations for robust training. *Advances in neural information processing systems*, 34: 237–250, 2021. [3](#)
- [52] Yixin Wang and Michael I Jordan. Desiderata for representation learning: A causal perspective. *arXiv preprint arXiv:2109.03795*, 2021. [7](#)
- [53] Zifan Wang, Nan Ding, Tomer Levinboim, Xi Chen, and Radu Soricut. Improving robust generalization by direct pac-bayesian bound minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16458–16468, 2023. [2](#)
- [54] Zehao Xiao, Jiayi Shen, Xiantong Zhen, Ling Shao, and Cees Snoek. A bit more bayesian: Domain-invariant learning with uncertainty. In *International Conference on Machine Learning*, pages 11351–11361. PMLR, 2021. [2](#), [3](#), [7](#)
- [55] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020. [8](#)
- [56] Naiyu Yin, Hanjing Wang, Yue Yu, Tian Gao, Amit Dhurandhar, and Qiang Ji. Integrating markov blanket discovery into causal representation learning for domain generalization. In *European Conference on Computer Vision*, 2024. [1](#), [3](#), [6](#), [7](#)
- [57] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. [3](#)
- [58] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. [3](#)