Modeling the Data-Generating Process is Necessary for Out-of-Distribution Generalization

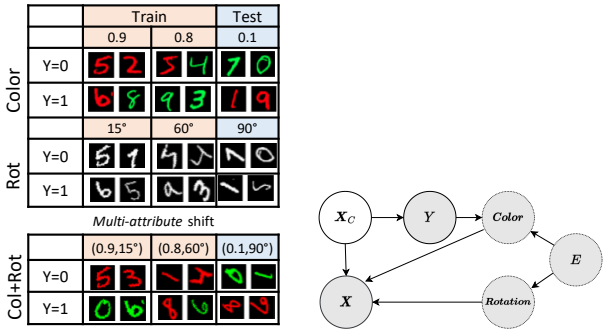
Jivat Neet Kaur [1] **Emre Kıcıman** [1] **Amit Sharma** [1]

Abstract

Real-world data collected from multiple domains can have multiple, distinct distribution shifts over multiple attributes. However, state-of-the art advances in domain generalization (DG) algorithms focus only on specific shifts over a single attribute. We introduce datasets with *multi*-attribute distribution shifts and find that existing DG algorithms fail to generalize. Using causal graphs to characterize the different types of shifts, we show that each multi-attribute causal graph entails different constraints over observed variables, and therefore any algorithm based on a single, fixed independence constraint cannot work well across all shifts. We present *Causally Adaptive Constraint Minimization (CACM)*, an algorithm for identifying the correct independence constraints for regularization. Experiments confirm our theoretical claim: correct independence constraints lead to the highest accuracy on unseen domains. Our results demonstrate the importance of modeling the causal relationships inherent in a data-generating process, without which it can be impossible to know the correct regularization constraints for a dataset.

1. Introduction

To perform reliably in real world settings, machine learning models must be robust to distribution shifts – where the training distribution differs from the test distribution. The *domain generalization (DG)* task (Wang et al., 2021; Zhou et al., 2021) encapsulates this challenge by evaluating accuracy on an unseen domain given data from multiple domains that share a common optimal predictor. Recent state-of-the-art advances in representation learning for DG (Li et al., 2018a; Arjovsky et al., 2019; Krueger et al., 2021; Mahajan et al., 2021; Sun & Saenko, 2016) focus on a limited setting where the domains exhibit a single kind of distribution shift over one attribute (where an attribute refers to a spurious high-level variable). Using MNIST as an example, domains are created either by adding new values of a spurious attribute like rotation (e.g., Rotated-MNIST dataset (Ghifary et al., 2015; Piratla et al., 2020)) or by changing the correlation between the class label and a spurious attribute like color (e.g., Colored-MNIST (Arjovsky et al., 2019)), but not both simultaneously. Recent work (Wiles et al., 2022; Ye et al., 2022) shows that the accuracy of state-of-the-art DG algorithms are not consistent over these different datasets, indicating the importance of the kind of shift in a dataset.



(a) (b)

Figure 1: (a) Our *multi-attribute* distribution shift dataset Col+Rot-MNIST. We combine Colored MNIST (Arjovsky et al., 2019) and Rotated MNIST (Ghifary et al., 2015) to introduce distinct shifts over *Color* and *Rotation* attributes. (b) The causal graph representing the data generating process for (a) – *Color* has a correlation with Y which changes across environments while *Rotation* varies independently.

In real-world data, however, different sources of distribution shift can *co-exist*. Differences across domains may involve multiple attributes with different kinds of shifts. For example, in our Col+Rot-MNIST dataset (see Figure 1), the color and rotation angle of digits can shift independently across data distributions. In satellite imagery (Koh et al., 2021), the appearance of land cover such as vegetation (trees and grasses) changes seasonally and independently of regional variations in vegetation. To capture such data, we provide a

[1]Microsoft Research. Correspondence to: Jivat Neet Kaur <t-kaurjivat@microsoft.com>.

Published at the ICML 2022 Workshop on Spurious Correlations, Invariance, and Stability. Baltimore, Maryland, USA.

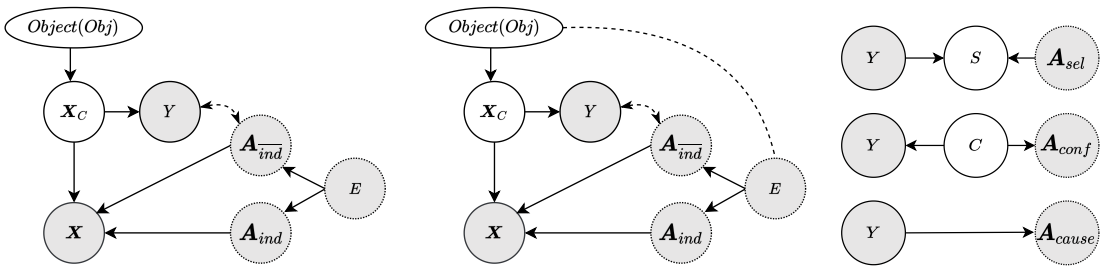


Figure 2: (a), (b) Causal graphs used for specifying *multi-attribute* distribution shifts. Shaded nodes denote observed variables; since not all attributes may be observed, we use dotted boundary. Dashed lines denote correlation. (c) represents different mechanisms for the Y $A_{\overline{ind}}$ relationship leading to *Causal*, *Confounded* and *Selected* shifts (bottom to top).

characterization of *multi-attribute* distribution shifts based on the relationship between each attribute and the class label. Using causal graphs and the principle of d-separation, we show that each type of shift leads to a different set of independence constraints on the observed variables. As a consequence, for datasets like Col+Rot-MNIST and a multi-attribute dataset derived from small NORB (LeCun et al., 2004), we find that existing DG algorithms that are often targeted for a specific shift fail to generalize. Specifically, applying an incorrect constraint leads to substantially lower accuracy than the correct constraint, highlighting the importance of identifying a suitable constraint for each dataset.

Beyond existing DG algorithms, our theoretical analysis shows that any representation learning algorithm based on a single, fixed independence constraint will fail to generalize under multi-attribute shifts. Therefore, we propose to leverage the information provided by multiple independent shifts across attributes, assuming structural knowledge of the shifts. Then, given a dataset and the canonical causal graph for multi-attribute shifts (Figure 2), our proposed algorithm, *Causally Adaptive Constraint Minimization (CACM)*, identifies the correct constraints and applies them as a regularizer in the learning algorithm's loss function.[1]

2. Generalization under multi-attribute shifts

We focus on representation learning-based (Wang et al., 2021) DG algorithms, typically characterized by a regularization term that constrains an ERM loss such as cross-entropy (see Table 5 in Suppl.). *Which regularization constraint is the correct one?* This question has attracted much discussion (Johansson et al., 2019; Mahajan et al., 2021; Wiles et al., 2022; Ye et al., 2022; Zhao et al., 2019) without resolution. Various works have discussed failure modes of unconditional independence (Akuzawa et al., 2019; Johansson et al., 2019; Zhao et al., 2019), conditional independence (Mahajan et al., 2021) and Invariant Risk Minimization (Rosenfeld et al., 2021). Moreover, recent empirical work (Wiles et al., 2022; Ye et al., 2022) shows that different algorithms perform better under different shifts, but none performs across all shifts. As a result, (Wiles et al., 2022) suggest that instead of a universal algorithm for any shift, adaptable algorithms that use auxiliary attribute information can be more useful. To explore this question beyond domains with a single distribution shift, we consider the generalization problem over a more realistic setup where each domain can have multiple shifts over different attributes.

[1]Full version of the paper is available at: https://arxiv.org/abs/2206.07837.

2.1. Risk-invariant predictor over a set of distributions

We consider the supervised learning setup from Wiles et al. (2022) where each row of train data $(x_i, a_i, y_i)_{i=1}^n$ contains input features x_i (e.g., X-ray pixels), a set of nuisance or spurious attributes a_i (e.g., hospital) and class label y_i (e.g., disease diagnosis). The attributes represent variables that are often recorded during data collection or can be inferred.

Since the nuisance attribute's distribution or its correlation with the label can change, we obtain different data distributions. Given a set of domains sampled from P, the train data is sampled from domains, $P_{Etr} = \{P_{E1}, P_{E2}, \ldots\} \subset P$ while the test data is assumed to be sampled from a single unseen domain, $P_{Ete} = \{P_{Ete}\} \subset P$. The goal is to learn a classifier $g(x)$ using train domains such that it generalizes and achieves a similar, small risk on test data from P_{Ete} as it achieves on P_{Etr}. Formally, given a set of distributions P, we define a risk-invariant predictor (Makar et al., 2022) as,

Definition 2.1. ***Optimal Risk Invariant Predictor for P (from (Makar et al., 2022))*** *Define the risk of predictor g on distribution $P \in P$ as $R_P(g) = \mathbb{E}_{x,y \sim P}\ell(g(x), y)$ where ℓ is cross-entropy or another classification loss. Then, the set of risk-invariant predictors obtain the same risk across all distributions $P \in P$, and set of the optimal risk-invariant predictors is defined as the risk-invariant predictors that obtain minimum risk on all distributions.*

2.2. Using causal graphs for multi-attribute shifts

To specify the set of distributions P to generalize over, using causal graphs, we characterize the different data-generating processes that can lead to a multi-attribute shift dataset. Figure 2 shows the corresponding causal directed acylic graph (DAG). Shaded nodes represent observed variables X, Y; and the sets of attributes $A_{\overline{ind}}$, A_{ind}, and E such that $A_{\overline{ind}} \cup A_{ind} \cup \{E\} = A$. $A_{\overline{ind}}$ represents the attributes correlated with label, A_{ind} the attributes that are independent of label, while E is a special attribute for the domain. All attributes, along with the stable/causal features X_c, determine the observed features X. And the stable features, X_c are the only features that cause Y. In the simplest case, we assume no label shift across environments i.e. marginal distribution of Y is constant across train domains and test, $P_{Etr}(y) = P_{Ete}(y)$ (see Figure 2(a)). More generally, different domains may have different distribution of objects and hence there may be a correlation between E and Obj, as represented by the right subfigure (Figure 2(b)).

We characterize different kinds of shifts based on the relationship between nuisance attributes A and the classification label Y. Specifically, A_{ind} has varying distribution across environments but is *Independent* of the class label. The dashed bidirectional arrow represents the correlation between $A_{\overline{ind}}$ and Y. There are different mechanisms which can introduce the dashed-line relationship (Figure 2(c)) – direct-causal relationship (Y causing $A_{\overline{ind}}$), confounding between Y and $A_{\overline{ind}}$ due to a common cause, or selection during the data-generating process. Thus, we define four kinds of shifts based on the causal graph: *Independent*, *Causal*, *Confounded*, and *Selected*. As we shall see, these shifts correspond to different independence constraints between observed variables. Thus, in addition to the dataset, to fully specify the problem of multi-attribute shift generalization for a learning algorithm, we require knowledge of the kind of shift for each observed attribute.

Definition 2.2. ***Generalization under Multi-attribute shifts.*** *Given training data $(x_i, a_i, y_i)_{i=1}^n$ and the type of causal relationship of each attribute A with the label Y, construct a realized causal graph G based on the canonical graph in Figure 2 and define P_G as the set of all distributions obtained by changing the relationship between Y and each attribute while keeping the same graph (type of shift). The generalization goal is to learn an optimal risk-invariant predictor over P_G.*

Availability of multiple attributes. Unlike the full causal graph, type of relationship between label and an attribute is often known. Suppl. A contains real-world examples where these relationships as well as attribute values are known.

3. Correct regularizer for multi-attribute shifts

3.1. Deriving conditional independence constraints for a risk-invariant representation

We assume the predictor can be represented as $g(x) = g_1(\phi(x))$ where ϕ is the learnt representation. To derive the constraints that should be satisfied by a risk-invariant $g_1(\phi)$, we utilize a strategy from past work (Mahajan et al., 2021; Veitch et al., 2021). We identify the conditional independence constraints satisfied by X_c in the causal graph and enforce that ϕ should follow the same constraint.

Proposition 3.1. *Given a dataset $(x_i, a_i, y_i)_{i=1}^n$ and a causal DAG G over $\{X_c, X, A, Y\}$ such that X_c is the only variable (or set of variables) that causes Y and is not independent of X, then the conditional independence constraints satisfied by X_c are necessary for a risk-invariant predictor over P_G. That is, if a predictor does not satisfy any of these constraints, then there exists a data distribution $P' \in P_G$ such that predictor's risk will be higher than its risk in other distributions.*

We examine two common constraints on independence between ϕ and a nuisance attribute: either unconditional (Albuquerque et al., 2020; Ganin et al., 2016b; Muandet et al., 2013) or conditional on the label Y (Ghifary et al., 2016; Hu et al., 2019; Li et al., 2018c;d) (see Suppl. for details on these baseline methods). Under the canonical graph from Figure 2(b), none of these constraints are valid because there could be a correlation path between X_c and E (under the X-ray example, this can be because more women visit one hospital compared to the other). When we simplify the graph by removing the correlation between object and E (Figure 2(a)), the unconditional constraint is true when $A \perp\!\!\!\perp Y$ ($A \in A_{ind}$) but not always for $A_{\overline{ind}}$. For any attribute $A \in A_{\overline{ind}}$, if the relationship between Y and A is *Confounded*, then the unconditional constraint is correct; if it is *Causal* or *Selected*, then the conditional constraint is correct. Below we provide the set of valid constraints.

Theorem 3.1. *Given a causal DAG with the structure as shown in Figure 2(a), the correct constraint depends on the relationship of label Y with the nuisance attributes A. As shown, A can be split into $A_{\overline{ind}}$, A_{ind} and E, where $A_{\overline{ind}}$ can be further split into subsets that have a causal (A_{cause}), confounded (A_{conf}), selected (A_{sel}) relationship with Y ($A_{\overline{ind}} = A_{cause} \cup A_{conf} \cup A_{sel}$). Then, the (conditional) independence constraints that X_c should satisfy are,*

1. *Independent: $X_c \perp\!\!\!\perp A_{ind}$; $X_c \perp\!\!\!\perp E$; $X_c \perp\!\!\!\perp A_{ind} | Y$; $X_c \perp\!\!\!\perp A_{ind} | E$; $X_c \perp\!\!\!\perp A_{ind} | Y, E$*
2. *Causal: $X_c \perp\!\!\!\perp A_{cause} | Y$; $X_c \perp\!\!\!\perp E$; $X_c \perp\!\!\!\perp A_{cause} | Y, E$*
3. *Confounded: $X_c \perp\!\!\!\perp A_{conf}$; $X_c \perp\!\!\!\perp E$; $X_c \perp\!\!\!\perp A_{conf} | E$*

4. *Selected:* X_c ?? A_{sel} *jY;* X_c ?? A_{sel} *jY, E*

Corollary 3.1.1. *All the above derived constraints are valid for Graph 2(a). However, in the presence of a correlation between E and Obj (Graph 2(b)), only the constraints conditioned on E hold true.*

Hence, if information on Obj-E correlation is not available, it is advisable to use E-$conditioned$ constraints. While we list all constraints, if one of the attributes is unobserved (say $A_{\overline{ind}}$ or A_{ind} is not available), then we use the subset of constraints derived for the observed features.

3.2. An algorithm for generalizing under multi-attribution shifts

We now describe the proposed *CACM* algorithm.

Given a causal graph, the *CACM* algorithm first utilizes the steps highlighted above to identify the correct independence constraints. Then it applies those constraints as a regularizer to the standard ERM loss, g_1, ϕ = arg min$_{g_1,\phi}$; $\ell(g_1(\phi(X)), y)$ + λ $(RegPenalty)$, where λ is a hyperparameter and ℓ is cross-entropy loss. We design the regularizer such that it optimizes for valid constraints over all observed variables V 2 A. If there is a choice between multiple constraints, we choose the constraint that will be valid over both Figure 2(a) and 2(b). We describe implementation details, full algorithm and the regularization penalty for individual shifts in Section D.3.

3.3. A fixed conditional independence constraint cannot work for all datasets

Since the observed data distribution can be identical for all three types of relationship between Y and $A_{\overline{ind}}$, the type of relationship cannot be learned from observed data. Since the constraints are different for different relationship types, it implies that any algorithm relying on a single (conditional) independence constraint (Gretton et al., 2012; Arjovsky et al., 2019; Li et al., 2018b; Sun & Saenko, 2016) cannot work for all datasets.

Theorem 3.2. *Under the canonical causal graph in Figure 2, there exists no (conditional) independence constraint such that it is valid for all realizations of the graph as the type of multi-attribute shifts vary. Thus, for any predictor algorithm for Y that uses a single type of (conditional) independence constraint, there exists a realized graph G and a corresponding training dataset such that the learned predictor cannot be a risk-invariant predictor across distributions in* P_G .

Table 1: MNIST. Accuracy on unseen domain for singe-(color, rotation) and multi-attribute (col+rot) shifts.

Algo.	Accuracy					
	color (A_{cause})		rotation (A_{ind})		col+rot	
ERM	30.9	1.6	61.9	0.5	25.2	1.3
IRM	50.0	0.1	61.2	0.3	39.6	6.7
VREx	30.3	1.6	62.1	0.4	23.3	0.4
MMD	29.7	1.8	62.2	0.5	24.1	0.6
CORAL	28.5	0.8	**62.5**	**0.7**	23.5	1.1
DANN	20.7	0.8	61.9	0.7	32.0	7.8
C-MMD	29.4	0.2	62.3	0.4	32.2	7.0
CDANN	30.8	8.0	61.8	0.2	32.2	7.0
CACM	**70.4**	**0.5**	62.4	0.4	**54.1**	**1.3**

Table 2: small NORB. Accuracy on unseen domain for single- (lighting, azimuth) and multi-attribute ($l + azi$) shifts.

Algo.	Accuracy					
	lighting (A_{cause})		azimuth (A_{ind})		l + azi	
ERM	65.5	0.7	78.6	0.7	64.0	1.2
IRM	66.7	1.5	75.7	0.4	61.7	0.5
VREx	64.7	1.0	77.6	0.5	62.5	1.6
MMD	66.6	1.6	76.7	1.1	62.5	0.3
CORAL	64.7	0.5	77.2	0.7	62.9	0.3
DANN	64.6	1.4	78.6	0.7	60.8	0.7
C-MMD	65.8	0.8	76.9	1.0	61.0	0.9
CDANN	64.9	0.5	77.3	0.3	60.8	0.9
CACM	**85.4**	**0.5**	**80.5**	**0.6**	**69.6**	**1.6**

4. Empirical Evaluation

We perform experiments on semi-synthetic (MNIST) and natural (small NORB) datasets to demonstrate our main claims: *CACM* with the correct graph-based constraints significantly outperforms these algorithms, and incorrect constraints cannot match the above accuracy. We compare to baseline algorithms: IRM (Arjovsky et al., 2019), VREx (Krueger et al., 2021), MMD (Li et al., 2018b), CORAL (Sun & Saenko, 2016), DANN (Gretton et al., 2012), Conditional-MMD (C-MMD) (Li et al., 2018b), and conditional-DANN (CDANN) (Li et al., 2018d). Refer to Suppl. D for further experimental details.

MNIST. Colored (Arjovsky et al., 2019) and Rotated (Ghifary et al., 2015) MNIST present A_{cause} and A_{ind} distribution shifts, respectively. We combine these to obtain a multi-attribute dataset with A_{cause} = $f color g$ and A_{ind} = $f rotation g$.

small NORB. We use small NORB (LeCun et al., 2004; Wiles et al., 2022), an object recognition dataset, to create a challenging task with multi-valued classes and attributes over realistic 3D objects with varying lighting (l) and azimuths (azi). We create multi-attribution shifts,

Table 3: Comparing constraints $X_c \perp\!\!\!\perp A_{cause} | Y, E$ and $X_c \perp\!\!\!\perp A_{cause} | Y$ in MNIST and small NORB.

Constraint	MNIST Acc.		small NORB Acc.	
$X_c \perp\!\!\!\perp A_{cause} \mid Y$	69.7	0.2	79.7	0.9
$X_c \perp\!\!\!\perp A_{cause} \mid Y, E$	70.4	0.5	85.4	0.5

wherein there is a correlation between lighting condition $A_{cause} = flg$ and object category y; and $A_{ind} = fazig$ that varies independently across domains.

Further dataset details are in Suppl. D.2. To compare the effect of shifts over two attributes, for each dataset, we also create single-attribute shift datasets involving a change in only one of the attributes. Thus, we have three evaluation setups for each dataset: A_{cause}, A_{ind} and $A_{cause} \not\perp A_{ind}$.

4.1. Results

Correct constraint derived from the causal graph matters. Table 1 shows the accuracy on MNIST dataset. Comparing the three prediction tasks, for all algorithms, accuracy is lowest under two-attribute shift ($A_{ind} \not\perp A_{cause}$), reflecting the difficulty of a distribution shift over multiple attributes. On the two-attribute shift task, all DG algorithms obtain less than 40% accuracy whereas *CACM* obtains a 14.5% absolute improvement. Results on the small NORB dataset (Table 2) are similar – *CACM* obtains 69.6% accuracy on the two-attribute task while the nearest baseline is ERM at 64%. *CACM* also obtains highest accuracy on the A_{cause} task for both datasets. On MNIST, we find that *CACM* achieves a substantially higher accuracy (70%) than IRM and VREx, just 5 units lower than the optimal 75%. While the A_{ind} task is relatively easier, algorithms optimizing for the correct constraint achieve highest accuracy. Note that MMD, CORAL, DANN, and *CACM* are based on the same independence constraint (see Table 5 in Suppl.). These results indicate the importance of regularization based on data-specific correct constraints for generalization.

Incorrect constraints hurt generalization. We now directly compare the effect of using correct versus *incorrect* (but commonly used) constraints for a dataset. To isolate the effect of a single constraint, we consider the single-attribute shift on A_{cause}. Comparing small NORB and MNIST (Table 3) reveals the importance of making the right structural assumptions. Typically, DG algorithms assume that distribution of causal features X_c does not change across domains. Then, both $X_c \perp\!\!\!\perp A_{cause} | Y, E$ and $X_c \perp\!\!\!\perp A_{cause} | Y$ should be correct constraints. However, conditioning on both Y and E provides a 5% point gain over conditioning on Y in NORB while the accuracy is comparable for MNIST. Auxiliary information about the data-generating process explains the result: Different domains in MNIST include samples from the same distribution whereas small NORB domains are sampled from a different set of toy objects, thus creating a correlation between Obj and E (Figure 2(b)). Without such auxiliary information, such gains will be difficult.

More ablations are in Suppl. E. To investigate the differences in shifts over *Independent*, *Causal* and *Confounded*, results of evaluation on synthetic data are in Suppl. E.2.

5. Discussion

We introduced *CACM*, an adaptive OoD generalization algorithm to characterize *multi-attribute* shifts and apply the correct independence constraints. Through empirical experiments and theoretical analysis, we show the importance of modeling the causal relationships in the data-generating process. The main limitation is that *CACM* does not address data sparsity – applying the constraints might be statistically inefficient if an attribute value is undersampled compared to others. Future work includes statistical improvements in the regularization penalty (e.g., multiple regularization coefficients λ).

References

Akuzawa, K., Iwasawa, Y., and Matsuo, Y. Adversarial invariant feature learning with accuracy constraint for domain generalization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 315–331. Springer, 2019.

Albuquerque, I., Monteiro, J., Darvishi, M., Falk, T. H., and Mitliagkas, I. Generalizing to unseen domains via distribution matching, 2020.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization, 2019. URL https://arxiv.org/abs/1907.02893.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. Randaugment: Practical automated data augmentation with a reduced search space. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18613–18624. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/d85b63ef0ccb114d0a3bb7b7d808028f-Paper.pdf.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., and Lempitsky, V. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016a. URL http://jmlr.org/papers/v17/15-239.html.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016b.

Ghifary, M., Kleijn, W., Zhang, M., and Balduzzi, D. Domain generalization for object recognition with multi-task autoencoders. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2551–2559, Los Alamitos, CA, USA, dec 2015. IEEE Computer Society. doi: 10.1109/ICCV.2015.293. URL https://doi.ieeecomputersociety.org/10.1109/ICCV.2015.293.

Ghifary, M., Balduzzi, D., Kleijn, W. B., and Zhang, M. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1414–1430, 2016.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. *ArXiv*, abs/2007.01434, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

Hu, S., Zhang, K., Chen, Z., and Chan, L. Domain generalization via multidomain discriminant analysis. In *Uncertainty in artificial intelligence: proceedings of the... conference. Conference on Uncertainty in Artificial Intelligence*, volume 35. NIH Public Access, 2019.

Johansson, F. D., Sontag, D., and Ranganath, R. Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 527–536, 2019.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Beery, S., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. Wilds: A benchmark of in-the-wild distribution shifts. In *ICML*, 2021.

Komura, D. and Ishikawa, S. Machine learning methods for histopathological image analysis. *Computational and Structural Biotechnology Journal*, 16:34–42, 2018. ISSN 2001-0370. doi: https://doi.org/10.1016/j.csbj.2018.01.001. URL https://www.sciencedirect.com/science/article/pii/S2001037017300867.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Priol, R. L., and Courville, A. Out-of-distribution generalization via risk extrapolation (rex). In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5815–5826. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/krueger21a.html.

LeCun, Y., Huang, F. J., and Bottou, L. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pp. II–104 Vol.2, 2004. doi: 10.1109/CVPR.2004.1315150.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. Learning to generalize: Meta-learning for domain generalization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018a. URL https://ojs.aaai.org/index.php/AAAI/article/view/11596.

Li, H., Pan, S. J., Wang, S., and Kot, A. C. Domain generalization with adversarial feature learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5400–5409, 2018b. doi: 10.1109/CVPR.2018.00566.

Li, Y., Gong, M., Tian, X., Liu, T., and Tao, D. Domain generalization via conditional invariant representations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018c.

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

Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., and Tschannen, M. Weakly-supervised disentanglement without compromises. ICML'20. JMLR.org, 2020.

Mahajan, D., Tople, S., and Sharma, A. Domain generalization using causal matching. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7313–7324. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/mahajan21b.html.

Makar, M., Packer, B., Moldovan, D., Blalock, D., Halpern, Y., and D'Amour, A. Causally motivated shortcut removal using auxiliary labels. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 739–766. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/makar22a.html.

Muandet, K., Balduzzi, D., and Schölkopf, B. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pp. 10–18, 2013.

Piratla, V., Netrapalli, P., and Sarawagi, S. Efficient domain generalization via common-specific low-rank decomposition. *Proceedings of the International Conference of Machine Learning (ICML) 2020*, 2020.

Rosenfeld, E., Ravikumar, P. K., and Risteski, A. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=BbNIbVPJ-42.

Sagawa*, S., Koh*, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 109 (5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9363924.

Shah, H., Tamuly, K., Raghunathan, A., Jain, P., and Netrapalli, P. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33, 2020.

Sun, B. and Saenko, K. Deep coral: Correlation alignment for deep domain adaptation. In Hua, G. and Jégou, H. (eds.), *Computer Vision – ECCV 2016 Workshops*, pp. 443–450, Cham, 2016. Springer International Publishing. ISBN 978-3-319-49409-8.

Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.-M., Ciompi, F., and van der Laak, J. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, 58:101544, 2019. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2019.101544. URL https://www.sciencedirect.com/science/article/pii/S1361841519300799.

Veitch, V., D'Amour, A., Yadlowsky, S., and Eisenstein, J. Counterfactual invariance to spurious correlations in text classification. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=BdKxQp0iBi8.

Wang, J., Lan, C., Liu, C., Ouyang, Y., Zeng, W., and Qin, T. Generalizing to unseen domains: A survey on domain generalization. *arXiv preprint arXiv:2103.03097*, 2021.

Wiles, O., Gowal, S., Stimberg, F., Rebuffi, S.-A., Ktena, I., Dvijotham, K. D., and Cemgil, A. T. A fine-grained analysis on distribution shift. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Dl4LetuLdyK.

Ye, N., Li, K., Bai, H., Yu, R., Hong, L., Zhou, F., Li, Z., and Zhu, J. Ood-bench: Quantifying and understanding two dimensions of out-of-distribution generalization. In *CVPR*, 2022.

Zhao, H., Combes, R. T. d., Zhang, K., and Gordon, G. J. On learning invariant representation for domain adaptation. *arXiv preprint arXiv:1901.09453*, 2019.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., and Change Loy, C. Domain generalization: A survey. *arXiv e-prints*, pp. arXiv–2103, 2021.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. Unpaired image-to-image translation using cycle-consistent adversarial networks. pp. 2242–2251, 10 2017. doi: 10.1109/ICCV.2017.244.

A. Presence of auxiliary attribute information in datasets

Unlike the full causal graph, attribute values as well as the relationships between class labels and attributes is often known. *CACM* assumes access to attribute labels A only during training time, which are collected as part of the data collection process (e.g., as metadata with training data (Makar et al., 2022)). We start by discussing the availability of attributes in WILDS (Koh et al., 2021), a set of real-world datasets adapted for the domain generalization setting. Attribute labels available in the datasets include, the *time (year)* and *region* associated with satellite images in FMoW dataset (Christie et al. 2018) for predicting land use category, *hospital* from where the tissue patch was collected for tumor detection in Camelyon17 dataset (Bandi et al., 2018) and the *demographic* information for CivilComments dataset (Borkan et al., 2019). (Koh et al., 2021) create different domains in WILDS using this metadata, consistent with our definition of E 2 A as a special domain attribute.

In addition, CACM requires the type of relationship between label Y and attributes. This is often known, either based on how the dataset was collected or inferred based on domain knowledge or observation. While the distinction between A_{ind} and $A_{\overline{ind}}$ can be established using a statistical test of independence on a given dataset, the distinction between A_{cause} , A_{sel} and A_{conf} within $A_{\overline{ind}}$ must be provided by the user. In the above datasets, for FMoW, *time* can be considered an *Independent* attribute (A_{ind}) since it reflects the time at which images are captured which is not correlated with Y; whereas *region* is a *Confounded* attribute since certain regions associated with certain Y labels are over-represented due to ease of data collection. Note that region cannot lead to *Causal* shift since the decision to take images in a region was not determined by the final label nor *Selected* for the same reason that the decision was not taken based on values of Y. Similarly, for the Camelyon17 dataset, it is known that differences in slide staining or image acquisition leads to variation in tissue slides across *hospitals*, thus implying that *hospital* is an *Independent* attribute (A_{ind}) (Koh et al., 2021; Komura & Ishikawa, 2018; Tellez et al., 2019); As another example from healthcare, a study in MIT Technology Review[2] discusses biased data where a person's *position* (A_{conf}) was spuriously correlated with disease prediction as patients lying down were more likely to be ill. As another example, (Sagawa* et al., 2020) adapt MultiNLI dataset for OoD generalization due to the presence of spurious correlation between *negation words* (attribute) and the contradiction label between "premise" and "hypothesis" inputs. Here, negation words are a result of the contradiction label (*Causal* shift), however this relationship between negation words and label may not always hold. Finally, for the CivilComments dataset, we expect the *demographic* features to be *Confounded* attributes as there could be biases which result in spurious correlation between comment toxicity and demographic information.

To provide examples showing the availability of attributes and their type of relationship with the label, Table 4 lists some popular datasets used for DG and the associated auxiliary information present as metadata. In addition to above discussed datasets, we include the popularly used Waterbirds dataset (Sagawa* et al., 2020) where the type of *background* (land/water) is assigned to bird images based on bird label; hence, being a *Causal* attribute.

Table 4: Commonly used DG datasets include auxiliary information.

Dataset	**Attribute(s)**	Y	A **relationship**
FMoW-WILDS (Koh et al., 2021)	time		A_{ind}
	region		A_{conf}
Camelyon17-WILDS (Koh et al., 2021)	hospital		A_{ind}
Waterbirds (Sagawa* et al., 2020)	background (land/water)		A_{cause}
MultiNLI (Sagawa* et al., 2020)	negation word		A_{cause}
CivilComments-WILDS (Koh et al., 2021)	demographic		A_{conf}

Datasets cited in this section

G. Christie, N. Fendley, J. Wilson, and R. Mukherjee. Functional map of the world. In Computer Vision and Pattern Recognition (CVPR), 2018.

P. Bandi, O. Geessink, Q. Manson, M. V. Dijk, M. Balkenhol, M. Hermsen, B. E. Bejnordi, B. Lee, K. Paeng, A. Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. IEEE Transactions on Medical Imaging, 38(2):550–560, 2018.

[2] https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagno

D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In WWW, pages 491–500, 2019.

B. Related Work

Improving the robustness of models in the face of distribution shifts is a key challenge. Several works have attempted to tackle the domain generalization problem (Wang et al., 2021; Zhou et al., 2021) using different approaches – data augmentation (Cubuk et al., 2020; He et al., 2016; Zhu et al., 2017), and representation learning (Arjovsky et al., 2019; Deng et al., 2009; Higgins et al., 2017) being popular ones. Trying to gauge the progress made by these approaches, Gulrajani and Lopez-Paz (Gulrajani & Lopez-Paz, 2021) find that existing state-of-the-art DG algorithms do not improve over ERM. More recent work (Wiles et al., 2022; Ye et al., 2022) empirically shows that different algorithms perform well over different distribution shifts, but no single algorithm performs consistently across all. While they evaluate on single-attribute shift datasets, (Wiles et al., 2022) discuss the importance of having auxiliary knowledge of and evaluating methods under different underlying shifts. To this end, we provide (1) multi-attribute shift benchmark datasets; (2) a causal interpretation of different kinds of shifts; and (3) an adaptive algorithm to identify the correct regularizer.

Causally-motivated learning. There has been recent work focused on *causal representation learning* (Arjovsky et al., 2019; Krueger et al., 2021; Locatello et al., 2020; Schölkopf et al., 2021) for OoD generalization. While these works attempt to learn the constraints for causal features from input features, we show that it is necessary to model the data-generating process and have access to auxiliary attributes to obtain a risk-invariant predictor, especially in *multi-attribute* distribution shift setups. Recent research has shown how causal graphs can be used to characterize and analyze the different kinds of distribution shifts that occur in real-world settings (Makar et al., 2022; Veitch et al., 2021). Our approach is similar in motivation but we extend from single-domain, single-attribute setups in past work to formally introduce *multi-attribute* distribution shifts in more complex and real-world settings. Additionally, we do not restrict ourselves to binary-valued classes and attributes.

C. Proofs

C.1. Proof of Proposition 3.1

Proposition 3.1. *Given a dataset* $(x_i, a_i, y_i)_{i=1}^n$ *and a causal DAG* G *over* $\{X_c, X, A, Y\}$ *such that* X_c *is the only variable (or set of variables) that causes* Y *and is not independent of* X*, then the conditional independence constraints satisfied by* X_c *are necessary for a risk-invariant predictor over* P_G*. That is, if a predictor does not satisfy any of these constraints, then there exists a data distribution* $P^0 \in P_G$ *such that predictor's risk will be higher than its risk in other distributions.*

Proof. Let X, Y, X_c be random variables where X_c causes Y. X_c also causes the observed features X but X may be additionally affected by the attributes A. Let $\hat{Y} = g(X)$ be a candidate predictor. Then $g(X)$ represents a random vector based on a deterministic function g of X. Suppose there is an independence constraint ψ that is satisfied by X_c but not $g(X)$. [3] Since A refers to the set of all other variables (attributes) that also cause X, A cannot be empty otherwise X is only caused by X_c and hence would satisfy all independence constraints that X_c satisfies. Below we show that such a predictor g is not risk-invariant: there exist two data distributions generated according to Definition 2.2 such that the risk of g is different for them.

Without loss of generality, we can write $g(x)$ as,

$$g(x) = (g(x)/h(x_c)) \quad h(x_c) = g^0(x, x_c)h(x_c) \qquad \forall x \quad P(X) \tag{1}$$

where h is an arbitrary, non-zero, deterministic function of the random variable X_c. Since X_c satisfies the (conditional) independence constraint ψ and h is a deterministic function, $h(X_c)$ also satisfies ψ. Also since the predictor $g(X)$ does not satisfy the constraint ψ, it implies that the random vector $g^0(X, X_c)$ cannot satisfy the constraint ψ. Thus, $g^0(X, X_c)$ cannot be a function of X_c only. Since X has two parents, X_c and A, this implies that $g^0(X, X_c)$ and A are not independent.

Now, let us construct two data distributions P_1 and P_2 such that $P(X_c, Y)$ stays invariant, i.e., $P_1(X_c, Y) = P_2(X_c, Y)$. But $P(A)$ can change or $P(A|Y)$ can change. Since A causes X, the conditional distribution $P(Y|X)$ will also

[3] In practice, the constraint may be evaluated on an intermediate representation of g, such that g can be written as, $g(X) = g_1(\phi(X))$ where ϕ denotes the representation function. However, for simplicity, we assume it is applied on $g(X)$.

change. Further, since $g^0(X, X_c)$ and A are not independent, $P(Y|g^0(X, X_c))$ will change, i.e., $P_1(Y|g^0(X, X_c)) \neq P_2(Y|g^0(X, X_c))$.

The risk over any distribution P can be written as (using the cross-entropy loss),

$$\begin{aligned} R_P(g) &= \mathbb{E}_P[\ell(Y, g^0(X, X_c)h(X_c))] \\ &= \mathbb{E}_P[\sum_y y \log g^0(X, X_c)h(X_c)] \\ &= \mathbb{E}_P[\sum_y y \log g^0(X, X_c)] + \mathbb{E}_P[\sum_y y \log h(X_c)] \end{aligned} \tag{2}$$

The risk difference is,

$$\begin{aligned} & R_{P_2}(g) - R_{P_1}(g) \\ &= \mathbb{E}_{P_1}[\sum_y y \log g^0(X, X_c)] - \mathbb{E}_{P_2}[\sum_y y \log g^0(X, X_c)] + \mathbb{E}_{P_1}[\sum_y y \log h(X_c)] - \mathbb{E}_{P_2}[\sum_y y \log h(X_c)] \\ &= \mathbb{E}_{P_1}[\sum_y y \log g^0(X, X_c)] - \mathbb{E}_{P_2}[\sum_y y \log g^0(X, X_c)] \end{aligned}$$

where the second equality is because $P_1(X_c, Y) = P_2(X_c, Y)$. The risk of $h(X_c))$ would be the same across P_1 and P_2 but not for g^0 since $g^0(X, X_c)$ changes across the two distributions. Thus the absolute risk difference is non-zero,

$$|R_{P_2}(g) - R_{P_1}(g)| > 0 \tag{3}$$

and g is not a risk-invariant predictor. Hence, satisfying conditional independencies that X_c satisfies is necessary for a risk-invariant predictor. □

C.2. Proof of Theorem 3.1

Theorem 3.1. *Given a causal DAG with the structure as shown in Figure 2(a), the correct constraint depends on the relationship of label Y with the nuisance attributes A. As shown, A can be split into $A_{\overline{ind}}$, A_{ind} and E, where $A_{\overline{ind}}$ can be further split into subsets that have a causal (A_{cause}), confounded (A_{conf}), selected (A_{sel}) relationship with Y ($A_{\overline{ind}} = A_{cause} \cup A_{conf} \cup A_{sel}$). Then, the (conditional) independence constraints that X_c should satisfy are,*

1. *Independent: $X_c \perp\!\!\!\perp A_{ind}$; $X_c \perp\!\!\!\perp E$; $X_c \perp\!\!\!\perp A_{ind}|Y$; $X_c \perp\!\!\!\perp A_{ind}|E$; $X_c \perp\!\!\!\perp A_{ind}|Y, E$*
2. *Causal: $X_c \perp\!\!\!\perp A_{cause}|Y$; $X_c \perp\!\!\!\perp E$; $X_c \perp\!\!\!\perp A_{cause}|Y, E$*
3. *Confounded: $X_c \perp\!\!\!\perp A_{conf}$; $X_c \perp\!\!\!\perp E$; $X_c \perp\!\!\!\perp A_{conf}|E$*
4. *Selected: $X_c \perp\!\!\!\perp A_{sel}|Y$; $X_c \perp\!\!\!\perp A_{sel}|Y, E$*

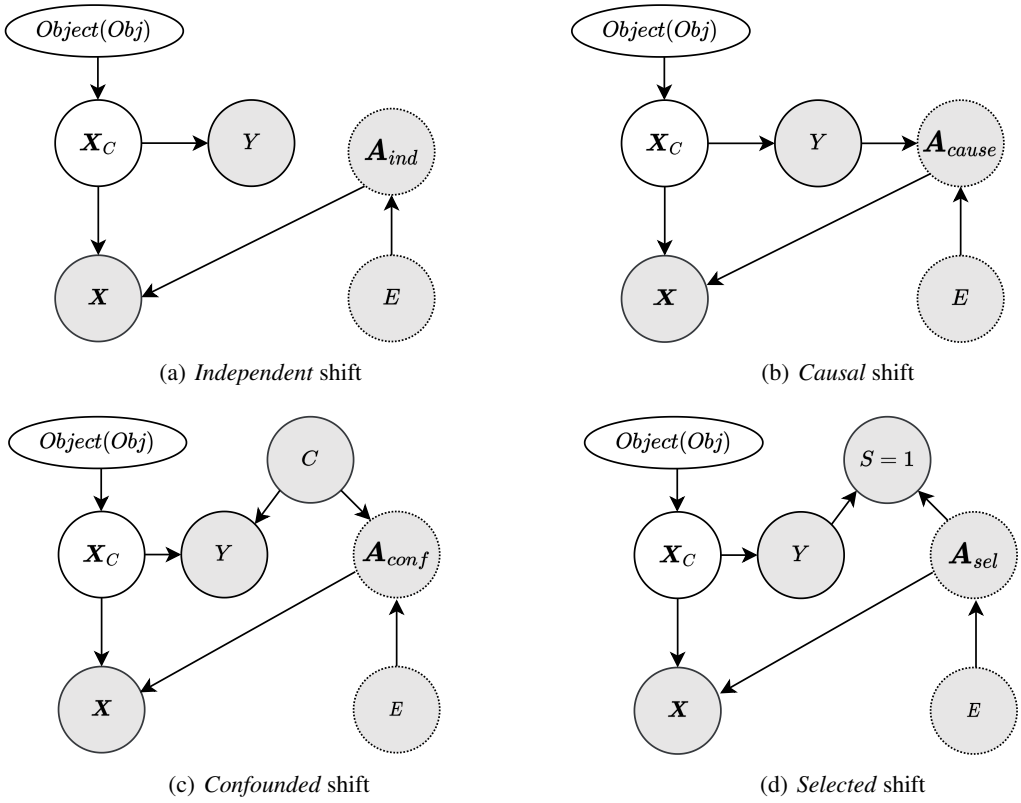
Proof. The proof follows from d-separation (Pearl, 2009) on the causal graphs realized from Figure 2(a). For each condition, *Independent*, *Causal*, *Confounded* and *Selected*, we provide the realized causal graphs below and derive the constraints.

Independent: As we can see in Figure 3(a), we have a collider X on the path from X_c to A_{ind} and X_c to E. Since there is a single path here, we obtain the independence constraints $X_c \perp\!\!\!\perp A_{ind}$ and $X_c \perp\!\!\!\perp E$. Additionally, we see that conditioning on Y or E would not block the path from X_c to A_{ind}, which results in the remaining constraints: $X_c \perp\!\!\!\perp A_{ind}|Y$; $X_c \perp\!\!\!\perp A_{ind}|E$ and $X_c \perp\!\!\!\perp A_{ind}|Y, E$. Hence, we obtain,

$$X_c \perp\!\!\!\perp A_{ind}; X_c \perp\!\!\!\perp E; X_c \perp\!\!\!\perp A_{ind}|Y; X_c \perp\!\!\!\perp A_{ind}|E; X_c \perp\!\!\!\perp A_{ind}|Y, E$$

Causal: From Figure 3(b), we see that while the path $X_c \to X \to A_{cause}$ from X_c to A_{cause} contains a collider X, $X_c \not\perp\!\!\!\perp A_{cause}$ due to the presence of node Y as a chain. By the d-separation criteria, X_c and A_{cause} are conditionally independent given $Y \implies X_c \perp\!\!\!\perp A_{cause}|Y$. Additionally, conditioning on E is valid since E does not appear as a collider on any paths between X_c and $A_{cause} \implies X_c \perp\!\!\!\perp A_{cause}|Y, E$. We get the constraint $X_c \perp\!\!\!\perp E$ since all paths connecting X_c to E contain a collider (collider X in $X_c \to X \to A_{cause} \leftarrow E$, collider A_{cause} in $X_c \to Y \to A_{cause} \leftarrow E$). Hence, we obtain,

$$X_c \perp\!\!\!\perp A_{cause}|Y; X_c \perp\!\!\!\perp E; X_c \perp\!\!\!\perp A_{cause}|Y, E$$



(a) *Independent* shift (b) *Causal* shift

(c) *Confounded* shift (d) *Selected* shift

Figure 3: Causal graphs for distinct distribution shifts based on Y A relationship.

Confounded: From Figure 3(c), we see that all paths connecting X_c and A_{conf} contain a collider (collider X in X_c ! X ! A_{conf}, collider Y in X_c ! Y ! C ! A_{conf}). Hence, X_c ?? A_{conf}. Additionally, conditioning on E is valid since E does not appear as a collider on any paths between X_c and A_{conf} =) X_c ?? A_{conf} jE. We get the constraint X_c ?? E since all paths connecting X_c and E also contain a collider (collider X in X_c ! X ! A_{conf} ! E, collider Y in X_c ! Y ! C ! A_{conf} ! E). Hence, we obtain,

$$X_c \perp\!\!\!\perp A_{conf}\ ;\ X_c \not\perp\!\!\!\perp E;\ X_c \perp\!\!\!\perp A_{conf} \mid E$$

Selected: For the observed data, the selection variable is always conditioned on, with $S = 1$ indicating inclusion of sample in data. The selection variable S is a collider in Figure 3(d) and we condition on it. Hence, $X_c \not\perp\!\!\!\perp A_{sel}$. Conditioning on Y breaks the edge X_c ! Y, and hence all paths between X_c and A_{sel} now contain a collider (collider X in X_c ! X ! A_{sel}) =) X_c ?? A_{sel} jY. Additionally, conditioning on E is valid since E does not appear as a collider on any paths between X_c and A_{sel} =) X_c ?? A_{sel} jY, E. Hence, we obtain,

$$X_c \perp\!\!\!\perp A_{sel} \mid Y;\ X_c \perp\!\!\!\perp A_{sel} \mid Y, E$$

C.2.1. Proof of Corollary 3.1.1

Corollary 3.1.1. *All the above derived constraints are valid for Graph 2(a). However, in the presence of a correlation between E and Obj (Graph 2(b)), only the constraints conditioned on E hold true.*

If there is a correlation between Obj and E, $X_c \not\perp\!\!\!\perp E$. We can see from Figure 3 that in the presence of Obj E correlation, $X_c \not\perp\!\!\!\perp A_{ind}$; $X_c \not\perp\!\!\!\perp A_{ind} \mid Y$ (3(a)), $X_c \not\perp\!\!\!\perp A_{cause} \mid Y$ (3(b)), $X_c \not\perp\!\!\!\perp A_{conf}$ (3(c)) and $X_c \not\perp\!\!\!\perp A_{sel} \mid Y$ (3(d)). Hence, conditioning on environment E is required for the valid independence constraints.

□

C.3. Proof of Theorem 3.2

Theorem 3.2. *Under the canonical causal graph in Figure 2, there exists no (conditional) independence constraint such that it is valid for all realizations of the graph as the type of multi-attribute shifts vary. Thus, for any predictor algorithm for Y that uses a single type of (conditional) independence constraint, there exists a realized graph G and a corresponding training dataset such that the learned predictor cannot be a risk-invariant predictor across distributions in P_G .*

Proof. The proof follows from an application of Theorem 3.1 and Proposition 3.1. Under the canonical graph from Figure 2(a or b), the four types of attribute shifts possible are *Independent*, *Causal*, *Confounded* and *Selected*. From the constraints provided for these four types of attribute shifts in Theorem 3.1, it is easy to observe that there is no single constraint that is satisfied across all four shifts. Thus, given a data distribution (and hence, dataset) with specific types of multi-attribute shifts such that X_c satisfies certain (conditional) independence constraints, it is always possible to change the type of at least one of the those shifts to create a new data distribution (dataset) where the same constraints will not hold.

To prove the second claim, suppose that there exists a predictor for Y based on a single type of conditional independence constraint. Since the same constraint is not valid across all attribute shifts, we can always construct a data distribution (corresponding to a realized graph G) where X_c would not satisfy the same constraint, by changing the type of at least one attribute shift. From Proposition 3.1, all conditional independence constraints satisfied by X_c under G are necessary to be satisfied for a risk-invariant predictor. Hence, for the class of distributions P_G, a single constraint-based predictor cannot be a risk-invariant predictor. □

D. Experimental Details

All experiments are performed in PyTorch 1.10 with NVIDIA Tesla P40 and P100 GPUs. We build upon the code from DomainBed (Gulrajani & Lopez-Paz, 2021) and OoD-Bench (Ye et al., 2022). Regularizing on $g_1(\phi(x))$ provided better accuracy than $\phi(x)$; hence we adopt it for all our experiments.

D.1. Additional details about baseline methods

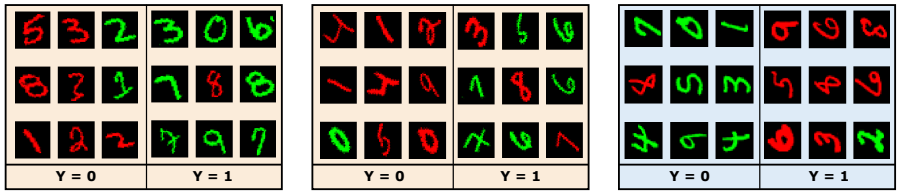
Table 5 lists the baseline methods we compare to, the independence constraints imposed and the statistics matched/optimized by each method across environments E.

Table 5: Statistic matched/optimized by different DG algorithms. match operation matches the statistic value across E. h is a learnable domain classifier on top of shared representation ϕ. ℓ represents the main classifier loss while ℓ_d is domain classifier loss.

Constraint	Statistic	DG Algorithm
$\phi \perp\!\!\!\perp E$	match $\mathbb{E}[\phi(x)\|E] \;\forall\; E$	MMD (Gretton et al., 2012)
	$\max_E \mathbb{E}[\ell_d(h(\phi(x)), E)]$	DANN (Ganin et al., 2016a)
	match $\mathrm{Cov}[\phi(x)\|E] \;\forall\; E$	CORAL (Sun & Saenko, 2016)
$Y \perp\!\!\!\perp E\|\phi$	match $\mathbb{E}[Y\|\phi(x), E] \;\forall\; E$	IRM (Arjovsky et al., 2019)
	match $\mathrm{Var}[\ell(f(x), y)\|E] \;\forall\; E$	VREx (Krueger et al., 2021)
$\phi \perp\!\!\!\perp E\|Y$	match $\mathbb{E}[\phi(x)\|E, Y = y] \;\forall\; E$	C-MMD (Li et al., 2018b)
	$\max_E \mathbb{E}[\ell_d(h(\phi(x)), E)\|Y = y)]$	CDANN (Li et al., 2018d)

D.2. Datasets

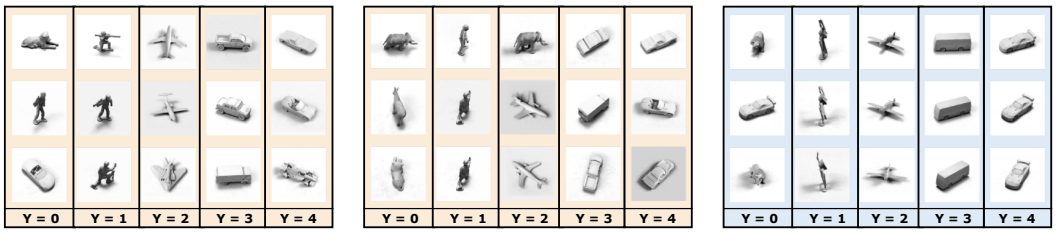
MNIST. Rotated (Ghifary et al., 2015) and Colored MNIST (Arjovsky et al., 2019) present distinct distribution shifts. While Rotated MNIST only has A_{ind} wrt. rotation attribute (R), Colored MNIST only has A_{cause} wrt. color attribute (C). We combine these datasets to obtain a multi-attribute dataset with $A_{cause} = \{C\}$ and $A_{ind} = \{R\}$. Each domain E_i has a specific rotation angle r_i and a specific correlation $corr_i$ between color C and label Y. Our setup consists of 3 domains: $E_1, E_2 \in E_{tr}$ (training), $E_3 \in E_{te}$ (test). We define $corr_i = P(Y = 1|C = 1) = P(Y = 0|C = 0)$ in E_i. In our setup, $r_1 = 15$, $r_2 = 60$, $r_3 = 90$ and $corr_1 = 0.9$, $corr_2 = 0.8$, $corr_3 = 0.1$. All environments have 25% label noise, as in (Arjovsky et al., 2019). For all experiments on MNIST, we use a two-layer perceptron consistent with previous works (Arjovsky et al., 2019; Krueger et al., 2021).



(a) (b) (c)

Figure 4: (a), (b) Train and (c) Test domains for MNIST.

small NORB. Moving beyond simple binary classification, we use small NORB (LeCun et al., 2004), an object recognition dataset, to create a challenging setup with multi-valued classes and attributes over realistic 3D objects. It consists of images of toys of five categories with varying lighting (l), elevation (ele) and azimuths (azi). The objective is to classify unseen samples of the five categories. (Wiles et al., 2022) introduced single-attribute shifts for this dataset. We combine them to yield $A_{cause} = flg$ wherein there is a correlation between lighting condition l and toy category y; and $A_{ind} = fazig$ that varies independently across domains. Training domains have 0.9 and 0.95 spurious correlation with l whereas there is no correlation in test domain. We add 5% label noise in all environments. We use ResNet-18 (pre-trained on ImageNet) for all settings and fine tune for our task.



(a) (b) (c)

Figure 5: (a), (b) Train and (c) Test domains for MNIST.

D.3. Implementation details

All methods are trained using Adam optimizer. MNIST dataset is trained for 5000 steps (default in DomainBed (Gulrajani & Lopez-Paz, 2021)) while small NORB is trained for 2000 steps. Consistent with the default value in DomainBed, we use a batch size 64 for MNIST and 128 for small NORB.

Regularization Penalty. Since A includes multiple attributes, the regularizer penalty depends on the type of distribution shift for each attribute. For instance, for $A \; 2 \; A_{ind}$ (*Independent*), to enforce $\phi(x) \; ?? \; A$, we aim to minimize the distributional discrepancy between $P(g(x)jA = a_i)$ and $P(g(x)jA = a_j)$, for all i, j values of A. However, the same constraint is applicable on E. So if domain variable E is available, it is statistically efficient to apply the constraint on E since there would typically be multiple closely related values of A in a domain (e.g., slide stains collected from one hospital may be spread over similar colors, but not exactly the same). Hence, we apply the constraint on distributions $P(g_1(\phi(x))jE = E_i)$ and $P(g_1(\phi(x))jE = E_j)$ if E is observed (and A may/may not be unobserved), otherwise we apply the constraint over A.

$$RegPenalty_{A_{ind}} = \sum_{i=1}^{jEj} \sum_{j>i} \mathsf{MMD}(P(g_1(\phi(x))jE = E_i), P(g_1(\phi(x))jE = E_j)) \tag{4}$$

For $A \; 2 \; A_{cause}$ (*Causal*), following Theorem 3.1, we consider distributions $P(g_1(\phi(x))jA = a_i, Y = y)$ and $P(g_1(phi(x))jA = a_j, Y = y)$. We additionally condition on domain E as there may be a correlation between E

and Obj (Figure 2(b)), which renders other constraints incorrect (Corollary 3.1.1).

$$RegPenalty_{A_{cause}} = \sum_{jEj} \sum_{y2Y} \sum_{i=1}^{jA_{cause}j} \sum_{j>i} \mathsf{MMD}(P(g_1(\phi(x))ja_{i,cause}, y), P(g_1(\phi(x))ja_{j,cause}, y)) \tag{5}$$

The final $RegPenalty$ is a sum of penalties over all attributes, $RegPenalty = \sum_{A2A} Penalty_A$. We choose the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) metric to implement our penalty (although, in principle, any estimable metric for enforcing conditional independence would work). Unlike prior work (Makar et al., 2022; Veitch et al., 2021), we do not restrict ourselves to binary-valued attributes and classes.

We provide the *CACM* algorithm for a general graph G below.

Algorithm 1 *CACM*

Input: Dataset $(x_i, a_i, y_i)_{i=1}^n$, causal DAG G
Output: Function $g(x) = g_1(\phi(x)) : X \; ! \; Y$
A set of observed variables in G except Y, E (special domain attribute)
C fg $f$mapping of A to $A_s g$
Phase I: Derive correct independence constraints
for A 2 A **do**
  if (X_c, A) are d-separated **then**
    X_c ?? A is a valid independence constraint
  else if (X_c, A) are d-separated conditioned on any subset A_s of the remaining observed variables in $A \; n \; fAg$ **then**
    X_c ?? AjA_s is a valid independence constraint
    $C[A] = A_s$
  end if
end for
Phase II: Apply regularization penalty using constraints derived
for A 2 A **do**
  if X_c ?? A **then**
    $RegPenalty_A = \sum_{jEj} \sum_{i=1}^{jAj} \sum_{j>i} \mathsf{MMD}(P(g_1(\phi(x))jA_i), P(g_1(\phi(x))jA_j))$
  else if A is in C **then**
    $A_s = C[A]$
    $RegPenalty_A = \sum_{jEj} \sum_{a2A_s} \sum_{i=1}^{jAj} \sum_{j>i} \mathsf{MMD}(P(g_1(\phi(x))jA_i, a), P(g_1(\phi(x))jA_j, a))$
  end if
end for
$RegPenalty = \sum_{A2A} RegPenalty_A$
$g_1, \phi = \arg\min_{g_1,\phi}; \quad \ell(g_1(\phi(x)), y) + \lambda \; (RegPenalty)$

Remark. If E is observed, we always condition on E because of Corollary 3.1.1.

We provide the regularization penalty ($RegPenalty$) for *Independent*, *Causal*, *Confounded* and *Selected* shifts for our causal graph in Figure 2.

$$RegPenalty_{A_{ind}} = \sum_{i=1}^{jEj} \sum_{j>i} \mathsf{MMD}(P(g_1(\phi(x))jE = E_i), P(g_1(\phi(x))jE = E_j))$$

$$RegPenalty_{A_{cause}} = \sum_{jEj} \sum_{y2Y} \sum_{i=1}^{jA_{cause}j} \sum_{j>i} \mathsf{MMD}(P(g_1(\phi(x))ja_{i,cause}, y), P(g_1(\phi(x))ja_{j,cause}, y))$$

$$RegPenalty_{A_{conf}} = \sum_{jEj} \sum_{i=1}^{jA_{conf}j} \sum_{j>i} \text{MMD}(P(g_1(\phi(x))ja_{i,conf}), P(g_1(\phi(x))ja_{j,conf}))$$

$$RegPenalty_{A_{sel}} = \sum_{jEj} \sum_{y2Y} \sum_{i=1}^{jA_{sel}j} \sum_{j>i} \text{MMD}(P(g_1(\phi(x))ja_{i,sel}, y), P(g_1(\phi(x))ja_{j,sel}, y))$$

We want to emphasize that the constraints from the *CACM* algorithm are necessary but not sufficient. While regularizers like *CACM* restrict the set of possible solutions to a smaller subset that contains X_c (Mahajan et al., 2021), they are not guaranteed to return X_c. Formally, X_c is not identified under the current graph.

Model Selection. We create 90% and 10% splits from each domain to be used for training and model selection (as needed) respectively. For our main results, we use a validation set that follows the test domain distribution consistent with previous work on these datasets (Arjovsky et al., 2019; Ye et al., 2022; Wiles et al., 2022). Specifically, we adopt the *test-domain validation* from DomainBed where early stopping is not allowed and all models are trained for the same fixed number of steps to limit test domain access. We additionally report results using *test-domain validation* with early stopping as well as *train-domain validation* in Suppl. E. *Train-domain validation* uses a validation set that follows the distribution of the training domains.

D.4. Hyperparameter search

Following DomainBed (Gulrajani & Lopez-Paz, 2021), we perform a random search 20 times over the hyperparameter distribution and this process is repeated for total 3 seeds. The best models are obtained across the three seeds over which we compute the mean and standard error. The hyperparameter search space for all datasets and algorithms is given in Table 16.

E. Results

Table 6: Small NORB *Causal* shift. Comparing X_c ?? A_{cause} jY, E with possible incorrect constraints.

Constraint	Accuracy	
X_c ?? A_{cause}	72.7	1.1
X_c ?? A_{cause} jE	76.2	0.9
X_c ?? A_{cause} jY	79.7	0.9
X_c ?? A_{cause} jY, E	**85.4**	**0.5**

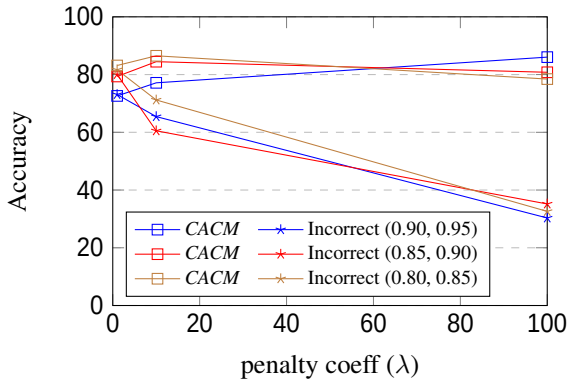


Figure 6: Accuracy of *CACM* and incorrect constraint on small NORB *Causal* shift with varying λ $f1$, 10, $100g$ and spurious correlation in train envs (in parantheses in legend).

E.1. Incorrect constraints hurt generalization.

Here, we present additional experiments to support the claim in Section 4.1 that incorrect constraints hurt model generalization.

Theorem 3.1 provides the correct constraint for A_{cause} : X_c ?? A_{cause} jY, E. In addition, using d-separation on Figure 2, we see the following invalid constraints, X_c ?? A_{cause} jE, X_c ?? A_{cause} . Without knowing that the shift corresponds to

a *Causal* shift, one may apply these constraints that do not condition on the class. Results on small NORB (Table 6) show that using the incorrect constraint has an adverse effect on model performance. The correct constraint yields 85% accuracy while the best incorrect constraint achieves 79.7%. Moreover, application of the incorrect constraint is sensitive to the λ (regularization weight) parameter (Figure 6): as λ increases, accuracy drops to less than 40%. However, accuracy with the correct constraint stays invariant across different values of λ.

Finally, we consider the *multi-attribute* shift setting for small NORB to demonstrate the significance of applying the correct constraints from the causal graph. In addition to applying the correct *CACM* constraints, we consider a case where we interchange the variables before inputting to *CACM* algorithm (A_{ind} gets used as A_{cause} and vice-versa) and then apply the resultant (incorrect) constraints. Accuracy with interchanged variables (65.1 1.6) is lower than that of correct *CACM* (69.6 1.6).

E.2. Synthetic Dataset

We create a synthetic dataset to investigate the differences in constraints over *Independent*, *Causal* and *Confounded* shifts. We also study the existing DG algorithms in a fairer synthetic setting and show their inability to close the performance gap with *CACM* even in such setting (Section E.3).

Dataset description. Our synthetic dataset is constructed based on the data-generating processes of the slab dataset (Mahajan et al., 2021; Shah et al., 2020). The original slab dataset was introduced by (Shah et al., 2020) to demonstrate the simiplicity bias in neural networks as they learn the linear feature which is easier to learn in comparison to the slab feature. Our extended slab dataset, adds to the setting from (Mahajan et al., 2021) by using non-binary attributes and class labels to create a more challenging task and allows us to study DG algorithms in the presence of linear spurious features.

Our dataset consists of label Y ($jYj = 5$) and 3-dimensional input X consisting of features X_c, A_{ind} and $A_{\overline{ind}}$. This is consistent with the graph in Figure 2 where attributes and causal features together determine observed features X; we concatenate X_c, A_{ind} and $A_{\overline{ind}}$ to generate X in our synthetic setup. Causal feature X_c has a non-linear "slab" relationship with Y while $A_{\overline{ind}}$ has a linear, *Causal* relationship with Y. A_{ind} is independent of Y and has varying uniform distribution p_{ind} across environments. We have three environments, E_1, E_2 2 E_{tr} (training) and E_3 2 E_{te} (test). X_c has a uniform distribution Uniform$[0, 1]$ across all environments.

$$y = \begin{cases} 0 & \text{if } X_c \text{ 2 } [0, 0.2) \\ 1 & \text{if } X_c \text{ 2 } [0.2, 0.4) \\ 2 & \text{if } X_c \text{ 2 } [0.4, 0.6) \\ 3 & \text{if } X_c \text{ 2 } [0.6, 0.8) \\ 4 & \text{if } X_c \text{ 2 } [0.8, 1.0] \end{cases}$$

$$A_{cause} = \begin{cases} y & \text{with prob. } = p \\ abs(y \quad 1) & \text{with prob. } = 1 \quad p \end{cases}$$

$$p_{ind}(A_{ind} jE_i) = \begin{cases} \text{Uniform}[\quad 0.4, 0.4] & \text{if i = 1} \\ \text{Uniform}[\quad 0.5, 0.5] & \text{if i = 2} \\ \text{Uniform}[\quad 0.8, 0.8] & \text{if i = 3} \end{cases}$$

Hence, we have a five-way classification setup with multi-valued attributes and *multi-attribute* distribution shifts. Following (Mahajan et al., 2021), the two training domains have p as 0.9 and 1.0, and the test domain has $p = 0.0$. We add 10% noise to Y in all environments. We use the default 3-layer MLP architecture from DomainBed and use mean difference (L2) instead of MMD as the regularization penalty given the simplicity of the data.

Experiments. We run all baselines and *CACM* on the synthetic dataset similar to experiments in Section 4 (Table 7). We can see that *CACM* significantly outperforms all algorithms for *Causal* and *multi-attribute* shifts. As discussed previously, A_{ind} is a relatively easier task; however, algorithms optimizing for the correct constraint achieve highest accuracy. Note that MMD, CORAL, DANN, and *CACM* are based on the same independence constraint in the presence of A_{ind} (see Table 5 in Suppl., Theorem 3.1).

Table 7: Synthetic dataset. Accuracy on unseen domain for singe-attribute (A_{cause}, A_{ind}), and multi-attribute (A_{cause} [A_{ind}) distribution shifts.

Algo.	Accuracy					
	A_{cause}		A_{ind}		A_{cause} [A_{ind}	
ERM	32.2	2.9	86.3	0.7	26.4	1.3
IRM	68.4	3.4	84.7	1.0	51.0	3.9
VREx	66.0	2.2	84.1	1.4	62.4	5.6
MMD	23.3	1.7	86.0	1.0	23.8	2.1
CORAL	28.6	3.0	**87.6**	**0.4**	21.7	1.1
DANN	44.6	3.6	84.0	0.6	46.4	4.3
C-MMD	36.7	4.1	85.3	1.3	27.6	1.8
CDANN	40.0	7.2	84.9	1.1	40.5	2.1
CACM	**94.1**	**0.5**	86.4	0.7	**84.3**	**3.5**

Table 8: Synthetic dataset. Accuracy on unseen domain for *Causal* distribution shift when A_{cause} is provided in input (column 2) and when A_{cause} is additionally used to create domains (column 3).

Algo.	Accuracy			
	A_{cause} (input)		A_{cause} (input+domains)	
ERM	32.2	2.9	29.1	4.6
IRM	68.4	3.4	36.4	1.7
VREx	66.0	2.2	24.9	1.2
MMD	23.3	1.7	39.7	7.3
CORAL	28.6	3.0	37.7	4.8
DANN	44.6	3.6	58.0	11.6
C-MMD	36.7	4.1	33.9	5.6
CDANN	40.0	7.2	49.8	5.0
CACM		**94.1**	**0.5**	

Table 9: Comparison of constraints $X_c \perp\!\!\!\perp A | Y, E$ and $X_c \perp\!\!\!\perp A | E$ in *Causal* and *Confounded* shifts. $X_c \perp\!\!\!\perp A | Y, E$ is a correct constraint for *Causal* shift but invalid for *Confounded* shift; $X_c \perp\!\!\!\perp A | E$ is a correct constraint for *Confounded* shift but invalid for *Causal* shift.

Constraint	Accuracy			
	A_{cause}		A_{conf}	
$X_c \perp\!\!\!\perp A \mid E$	29.7	3.8	**62.4**	**1.9**
$X_c \perp\!\!\!\perp A \mid Y, E$	**94.1**	**0.5**	56.0	1.0

E.3. Providing attribute information to DG algorithms for a fairer comparison

CACM leverages attribute labels to apply the correct independence constraints derived from the causal graph. However, existing DG algorithms only use the input features X and the domain attribute. Here we provide this attribute information to existing DG algorithms to create a more favorable setting for their application. We show that even in a relatively fairer setup, these algorithms are not able to close the performance gap with *CACM*, showing the importance of the causal information through graphs.

We consider our Synthetic dataset with *Causal* distribution shift where our observed features $X = (X_c, A_{cause})$. Note that by construction of X, since one of our input dimensions already consists of A_{cause}, we explicitly make A_{cause} available to all DG algorithms for applying their respective constraints. Thus, in the synthetic setup, all baselines do receive information about A_{cause} in addition to the domain attribute E.

As a more informative way of providing the attribute information (A_{cause}) for existing DG algorithms, we run a separate experiment where the attribute is provided as the domain indicator. Using the same underlying data distribution, we group the data (i.e., create environments/domains) based on A_{cause} i.e, each environment E has samples with same value of A_{cause}. In this setup (Table 8, third column), we see MMD, CORAL, DANN and CDANN show significant improvement in accuracy but the best performance is still 36% lower than *CACM* while showing high estimate variance. This reinforces our motivation to use the causal graph of the data-generating process to derive the constraint, as the attribute values alone are not sufficient. We also see IRM and VREx perform much worse than earlier, highlighting the sensitivity of DG algorithms to domain definition. In contrast, *CACM* uses the causal graph to study the structural relationships and derive the regularization penalty, which remains the same in this new dataset too.

E.4. Comparing constraints for *Confounded* vs *Causal* shift

Here, we extend our experiments from the main paper to consider a *Confounded* shift setting. In Theorem 3.1, we see that for the *Causal* shift, $X_c \perp\!\!\!\perp A_{cause} | Y$; $X_c \not\perp\!\!\!\perp A_{cause}$ (also, $X_c \perp\!\!\!\perp A_{cause} | Y, E$; $X_c \not\perp\!\!\!\perp A_{cause} | E$) whereas for the *Confounded* shift, $X_c \perp\!\!\!\perp A_{conf}$; $X_c \not\perp\!\!\!\perp A_{conf} | Y$ (also, $X_c \perp\!\!\!\perp A_{conf} | E$; $X_c \not\perp\!\!\!\perp A_{conf} | Y, E$). We construct a synthetic setup with *Confounded* shift to demonstrate the importance of using the valid independence constraints for different kinds of shifts.

We have three environments, $E_1, E_2 \in E_{tr}$ (training) and $E_3 \in E_{te}$ (test). X_c has a uniform distribution $\mathsf{Uniform}[0, 1]$ across all environments. Our confounding variable c has different functional relationships with Y and A_{conf} which vary across environments. Our observed input X is 2-dimensional and formed by concatenating X_c and A_{conf}.

$$c_{E_1,E_2} = \begin{cases} 1 & \text{with prob. } = 0.25 \\ 0 & \text{with prob. } = 0.75 \end{cases} \qquad c_{E_3} = \begin{cases} 1 & \text{with prob. } = 0.75 \\ 0 & \text{with prob. } = 0.25 \end{cases}$$

$$y_{true} = \begin{cases} 0 & \text{if } X_c \in [0, 0.25) \\ 1 & \text{if } X_c \in [0.25, 0.5) \\ 2 & \text{if } X_c \in [0.5, 0.75) \\ 3 & \text{if } X_c \in [0.75, 1.0] \end{cases}$$

$$y_{E_1,E_2} = \begin{cases} y_{true} + c & \text{with prob. } = 0.9 \\ y_{true} & \text{with prob. } = 0.1 \end{cases} \qquad y_{E_3} = y_{true}$$

$$A_{conf} = \begin{cases} 2 \quad c & \text{with prob. } = p \\ 0 & \text{with prob. } = 1 \quad p \end{cases} \qquad ; p_{E_1} = 1.0, p_{E_2} = 0.9, p_{E_3} = 0.8$$

Table 9 compares the performance of these constraints in synthetic *Confounded* and *Causal* setups (Section E.2). We can see that the valid constraints according to the graph significantly outperform the incorrect constraints in both shifts. Hence, the information on the specific relationship between Y and A is necessary for obtaining an optimal predictor.

E.5. Complete results

We provide complete results here for experiments in Section 4.

Tables 10, 11 and 12 show results on *Causal* (A_{cause}), *Independent* (A_{ind}) and *multi-attribute* ($A_{cause} \times A_{ind}$) shifts respectively for MNIST. Tables 13, 14 and 15 show results on *Causal* (A_{cause}), *Independent* (A_{ind}) and *multi-attribute* ($A_{cause} \times A_{ind}$) shifts respectively for small NORB.

While we report results using *test-domain validation* without early stopping in Section 4.1, we present additional results here using early stopping. Overall, early stopping improves accuracy across datasets and shifts for all methods. *CACM* outperforms all methods using model selection with as well as without early stopping, with the exception of Table 11. Table 11 shows results for the *Independent* shift which is a relatively easier task and hence all methods perform similarly. For *Independent* shift in MNIST (Table 11), CORAL achieves the highest accuracy. It is important to note that CORAL uses the same valid independence constraint derived by *CACM* for *Independent* shift (Theorem 3.1).

For completeness, we also include results using *train-domain validation*. However, as noted by previous work (Ye et al., 2022), using a validation set based on training domain distribution may not be suitable in the presence of spurious correlations as achieving high accuracy in training domains often leads to low accuracy in sufficiently different, novel test domains.

F. Anti-Causal Graph

Figure 7 shows causal graphs used for specifying *multi-attribute* distribution shifts in an anti-causal setting. These graphs are identical to Figure 2, with the exception of change in direction of causal arrow from $X_c \to Y$ to $Y \to X_c$.

We derive the (conditional) independence constraints for the anti-causal DAG for *Independent*, *Causal*, *Confounded* and *Selected* shifts.

Theorem F.1. *Given a causal DAG with the structure as shown in Figure 7(a), the correct constraint depends on the relationship of label Y with the nuisance attributes A. As shown, A can be split into $A_{\overline{ind}}$, A_{ind} and E, where $A_{\overline{ind}}$ can be further split into subsets that have a causal (A_{cause}), confounded (A_{conf}), selected (A_{sel}) relationship with Y ($A_{\overline{ind}} = A_{cause} \cup A_{conf} \cup A_{sel}$). Then, the (conditional) independence constraints that X_c should satisfy are,*

1. *Independent:* $X_c \perp\!\!\!\perp A_{ind}$; $X_c \perp\!\!\!\perp E$; $X_c \perp\!\!\!\perp A_{ind} | Y$; $X_c \perp\!\!\!\perp A_{ind} | E$; $X_c \perp\!\!\!\perp A_{ind} | Y, E$
2. *Causal:* $X_c \perp\!\!\!\perp A_{cause} | Y$; $X_c \perp\!\!\!\perp E$; $X_c \perp\!\!\!\perp A_{cause} | Y, E$
3. *Confounded:* $X_c \perp\!\!\!\perp A_{conf} | Y$; $X_c \perp\!\!\!\perp E$; $X_c \perp\!\!\!\perp A_{conf} | Y, E$
4. *Selected:* $X_c \perp\!\!\!\perp A_{sel} | Y$; $X_c \perp\!\!\!\perp A_{sel} | Y, E$

Proof. The proof follows from d-separation using the same logic as earlier proof in Section C.2. We observe that for all attributes $A \in A_{\overline{ind}}$ (A_{cause}, A_{conf}, A_{sel}), it is required to condition on Y to obtain valid constraints as Y node appears as a chain or fork in the causal graph but never as a collider due to the $Y \rightarrow X_c$ causal arrow. □

Corollary F.1.1. *All the above derived constraints are valid for Graph 7(a). However, in the presence of a correlation between E and Obj (Graph 7(b)), only the constraints conditioned on E hold true.*

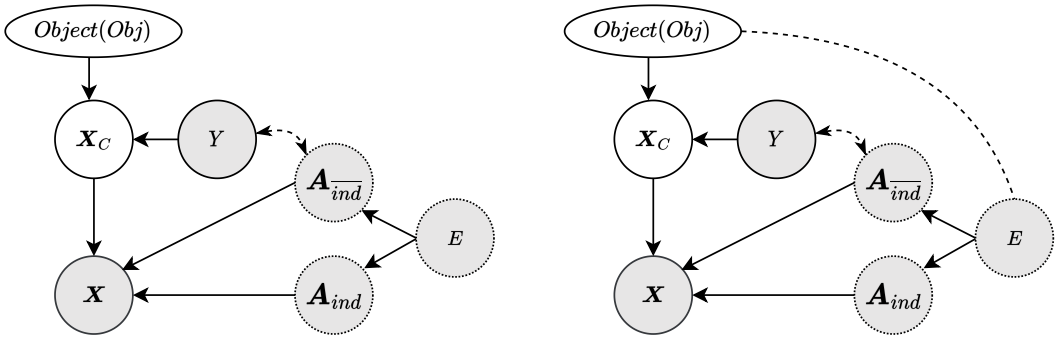


Figure 7: Corresponding anti-causal graphs for Figure 2. Note the graphs are identical to Figure 2 with the exception of the causal arrow pointing from $Y \rightarrow X_c$ instead of from $X_c \rightarrow Y$.

G. Broader impact and ethical considerations

Our work on modeling the data-generating process for improved out-of-distribution generalization is an important advance in building robust predictors for practical settings. Such prediction algorithms, including methods building on representation learning, are increasingly a key element of decision-support and decision-making systems. We expect our approach to creating a robust predictor to be particularly valuable in real world setups where *spurious* attributes and real-world multi-attribute settings lead to biases in data. While not the focus of this paper, *CACM* may be applied to mitigate social biases (e.g., in language and vision datasets) whose structures can be approximated by the graphs in Figure 2. Risks of using methods such as *CACM*, include excessive reliance or a false sense of confidence. While methods such as *CACM* ease the process of building robust models, there remain many ways that an application may still fail (e.g., incorrect structural assumptions). AI applications must still be designed appropriately with support of all stakeholders and potentially affected parties, tested in a variety of settings, etc.

Table 10: Colored + Rotated MNIST. Complete results for *Causal* (A_{cause}) shift.

Algorithm	Test-domain validation				Train-domain validation	
	no early stopping		early stopping			
ERM	30.9	1.6	63.2	2.7	10.1	0.1
IRM	50.0	0.1	66.1	1.5	**10.9**	**0.7**
VREx	30.3	1.6	62.1	2.6	10.2	0.4
MMD	29.7	1.8	57.8	4.5	10.1	0.1
CORAL	28.5	0.8	63.3	4.8	10.2	0.1
DANN	20.7	0.8	64.1	2.4	9.6	0.0
C-MMD	29.4	0.2	68.3	1.3	10.1	0.4
CDANN	30.3	9.1	63.3	3.4	10.2	0.2
CACM	**70.4**	**0.5**	**71.7**	**0.7**	10.1	0.2

Table 11: Colored + Rotated MNIST. Complete results for *Independent* (A_{ind}) shift.

Algorithm	Test-domain validation				Train-domain validation	
	no early stopping		early stopping			
ERM	61.9	0.5	63.4	0.8	61.1	0.4
IRM	61.2	0.3	63.1	1.0	60.5	0.6
VREx	62.1	0.4	62.2	0.5	61.5	0.2
MMD	62.2	0.5	61.6	0.2	60.7	0.6
CORAL	**62.5**	**0.7**	62.0	0.4	60.3	0.6
DANN	61.9	0.7	62.8	0.5	61.7	0.7
C-MMD	62.3	0.4	62.4	0.3	62.3	0.1
CDANN	61.8	0.2	**63.5**	**0.5**	**62.6**	**0.4**
CACM	62.4	0.4	63.0	0.1	61.6	0.3

Table 12: Colored + Rotated MNIST. Complete results for *multi-attribute* (A_{cause} [A_{ind}) shift.

Algorithm	Test-domain validation				Train-domain validation	
	no early stopping		early stopping			
ERM	25.2	1.3	64.2	5.3	10.3	0.1
IRM	39.6	6.7	66.2	3.1	10.5	0.0
VREx	23.3	0.4	65.2	4.4	10.0	0.1
MMD	24.1	0.6	62.6	3.4	**10.6**	**0.3**
CORAL	23.5	1.1	65.9	5.5	10.2	0.3
DANN	32.0	7.8	62.1	2.4	10.9	0.5
C-MMD	32.2	7.0	60.0	2.4	10.4	0.4
CDANN	30.8	8.0	67.6	2.8	10.3	0.2
CACM	**54.1**	**1.3**	**69.7**	**2.6**	10.2	0.1

Table 13: Small NORB. Complete results for *Causal* (A_{cause}) shift.

Algorithm	Test-domain validation				Train-domain validation	
	no early stopping		early stopping			
ERM	65.5	0.7	67.6	1.3	60.0	1.4
IRM	66.7	1.5	68.4	1.2	62.3	2.1
VREx	64.7	1.0	67.5	0.3	58.1	0.9
MMD	66.6	1.6	67.5	1.2	60.7	0.1
CORAL	64.7	0.5	67.4	0.2	61.5	1.7
DANN	64.6	1.4	69.6	0.5	61.5	1.1
C-MMD	65.8	0.8	68.5	0.1	62.1	2.4
CDANN	64.9	0.5	70.9	1.1	64.6	1.2
CACM	**85.4**	**0.5**	**87.2**	**0.4**	**75.7**	**4.7**

Table 14: Small NORB. Complete results for *Independent* (A_{ind}) shift.

Algorithm	Test-domain validation				Train-domain validation	
	no early stopping		early stopping			
ERM	78.6	0.7	79.2	1.1	74.2	1.5
IRM	75.7	0.4	79.4	0.4	72.0	0.9
VREx	77.6	0.5	79.6	0.1	75.2	0.7
MMD	76.7	1.1	79.9	0.7	74.7	0.9
CORAL	77.2	0.7	79.5	0.9	75.3	0.8
DANN	78.6	0.7	80.0	0.3	74.4	0.8
C-MMD	76.9	1.0	79.4	0.3	75.5	1.5
CDANN	77.3	0.3	78.6	0.9	72.5	1.4
CACM	**80.5**	**0.6**	**81.3**	**0.7**	**77.4**	**1.5**

Table 15: Small NORB. Complete results for *multi-attribute* (A_{cause} [A_{ind}) shift.

Algorithm	Test-domain validation				Train-domain validation	
	no early stopping		early stopping			
ERM	64.0	1.2	64.2	1.1	55.6	0.7
IRM	61.7	0.5	64.1	1.3	57.4	1.0
VREx	62.5	1.6	63.1	1.5	48.1	6.7
MMD	62.5	0.3	63.1	0.2	60.1	1.9
CORAL	62.9	0.3	63.9	1.6	42.4	5.0
DANN	60.8	0.7	65.1	1.0	57.9	1.4
C-MMD	61.0	0.9	62.9	1.2	58.7	3.0
CDANN	60.8	0.9	65.6	1.1	**60.5**	**1.8**
CACM	**69.6**	**1.6**	**69.5**	**1.6**	55.4	6.5

Table 16: Search space for random hyperparameter sweeps.

Method	Sweeps
MLP	learning rate: [1e-2, 1e-3, 1e-4, 1e-5] dropout: 0
ResNet	learning rate: [1e-2, 1e-3, 1e-4, 1e-5] dropout: [0, 0.1, 0.5]
MNIST	weight decay: 0 generator weight decay: 0
not MNIST	weight decay: $10^{\text{Uniform}(-6, -2)}$ generator weight decay: $10^{\text{Uniform}(-6, -2)}$
IRM	learning rate: [1e-2, 1e-3, 1e-4, 1e-5] λ: [0.01, 0.1, 1, 10, 100] iterations annealing: [10, 100, 1000]
VREx	learning rate: [1e-2, 1e-3, 1e-4, 1e-5] λ: [0.01, 0.1, 1, 10, 100] iterations annealing: [10, 100, 1000]
MMD	learning rate: [1e-2, 1e-3, 1e-4, 1e-5] λ: [0.1, 1, 10, 100] γ: [0.01, 0.0001, 0.000001
CORAL	learning rate: [1e-2, 1e-3, 1e-4, 1e-5] λ: [0.1, 1, 10, 100]
DANN, CDANN	generator learning rate: [1e-2, 1e-3, 1e-4, 1e-5] discriminator learning rate: [1e-2, 1e-3, 1e-4, 1e-5] discriminator weight decay: $10^{\text{Uniform}(-6, -2)}$ λ: [0.1, 1, 10, 100] discriminator steps: [1, 2, 4, 8] gradient penalty: [0.01, 0.1, 1, 10] adam β_1: [0, 0.5]
C-MMD	learning rate: [1e-2, 1e-3, 1e-4, 1e-5] λ: [0.1, 1, 10, 100] γ: [0.01, 0.0001, 0.000001
CACM	learning rate: [1e-2, 1e-3, 1e-4, 1e-5] λ: [0.1, 1, 10, 100] γ: [0.01, 0.0001, 0.000001