Causally motivated multi-shortcut identification & removal

Jiayun Zheng [1] **Maggie Makar** [1]

Abstract

We present an approach to discourage shortcut learning. Our approach has two steps: (1) efficiently identify relevant shortcuts, and (2) leverage the identified shortcuts to build robust and efficient models. We present theoretical and empirical arguments which show that our approach leads to robust and efficient estimators.

1. Introduction

Despite their immense success, predictors constructed from deep neural networks (DNNs) tend to have poor performance under distribution shift [7, 22, 5, 12]. One reason behind such brittleness is "shortcut learning": when a predictor relies on spurious correlations between the inputs and the target label that are easy to learn (i.e., shortcuts) and are predictive of the label in the training data [13]. If these spurious correlations no longer exist when the test distribution shifts, the accuracy of the predictor deteriorates. Here, we study the problem of learning a performant predictor whose risk is invariant to interventions that change the association between irrelevant factors (i.e., shortcuts) and the target label. Our work tackles two limitations in previous literature on preventing shortcut learning. First, previous work often assumes that the set of shortcuts are known in advance, or is easily identifiable using interpretability methods such as saliency maps. Second, much of the existing work assumes that there are a few (often one) shortcuts.

To tackle these limitations, we propose an approach to identify shortcuts, and build models that are invariant to possibly many shortcuts. Throughout, we will use the example of detecting the presence and severity of diabetic retionpathy (DR) using images taken using a funduscope. We focus on a setting where we are also given multiple auxiliary labels (e.g., the type of funduscope, patient age, sex and previous medical history) at training but not test time. A subset of these auxiliary data label factors of variation (i.e., shortcuts) that we want to be invariant to but the rest might be redundant for the purpose of shortcut removal.

Our contributions can be summarized as follows. (1) We leverage ideas from causality to show that robustness to a large set of distribution shifts is possible through ensuring invariance to a small set of shortcuts. (2) We develop a method for identifying these shortcuts, provide theoretical arguments about validity of our approach and show that it leads to more efficient predictors. (3) We extend previous work on single shortcut removal to a more general formulation that allows for high dimensional shortcuts of arbitrary types (4) We empirically validate our theoretical findings using a semi-simulated benchmark and a medical task, showing our approach has favorable in- and out-of-distribution generalization properties.

Related Work. Unlike previous work [35, 26, 33, 23, 30, 4], we do not assume that the relevant shortcuts are known *a priori* and we do not make any assumptions about the type or dimension of the auxiliary and target labels. A more extensive review of related work is in the appendix.

2. Preliminaries

Setup. We consider a supervised learning setup where the task is to construct a predictor $f(\mathbf{X})$ that predicts a label Y (e.g., presence and severity of DR) from an input $\mathbf{X}$ (e.g., image). We assume that at training time only, we have a set of auxiliary labels $\mathbf{V}^d$, with $d = \{0, \dots, D\}$. We use V^i to denote the i^{th} column of $\mathbf{V}^d$, and $\boldsymbol{V}^{d\setminus i}$ to denote all columns except the i^{th} column. We use $\mathcal{X}, \mathcal{Y}, \mathcal{V}^d$ to denote the domains of $\mathbf{X}, Y$, and $\mathbf{V}^d$ respectively. We make no assumptions about these domains: they can contain binary, categorical or continuous variables. We use the notation $Z \perp\!\!\!\perp_P Z'$ to denote that the two variables Z, Z' are independent under the distribution P. We use capital letters to denote variables, and small letters to denote their value. Our training data consist of tuples $\mathcal{D} = \{(\mathbf{x}_i, y_i, \mathbf{v}_i^d)\}_{i=1}^n$ drawn from a source training distribution P_s. We will consider predictors f of the form $f = h(\phi(\mathbf{x}))$, where ϕ is a representation mapping and h is the final predictor.

We assume that P_s follows an anti-causal structure, meaning that $\mathbf{X}$ is generated by the labels Y and $\mathbf{V}^p$, where $\mathbf{V}^p$ is a subset of $\mathbf{V}^d$. Importantly, we do not assume that such a subset is known *a priori*. We use $\mathbf{V}^c$ to denote the complement of $\mathbf{V}^p$, i.e., all the variables in $\mathbf{V}^d$ that do not di-

[1]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor. Correspondence to: Maggie Makar <mmakar@umich.edu>.

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

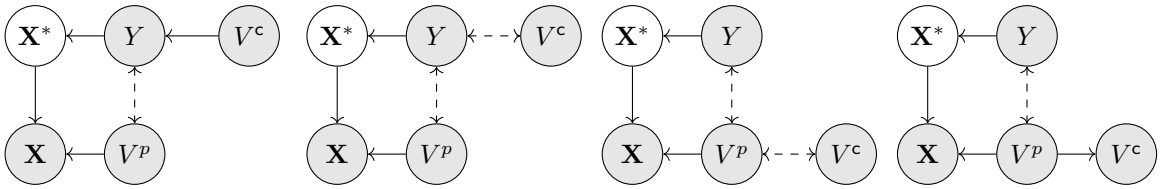

Figure 1. Examples of causal DAGs studied in this paper.

rectly affect $\mathbf{X}$. We assume that the labels Y and $\mathbf{V}^p$ are are correlated, but not causally related; that is, an intervention on $\mathbf{V}^p$ does not imply a change in the distribution of Y, and vice versa. Such correlation often arises through the influence of an unobserved third variable such as the environment from which the data is collected. We make no assumptions about the relationship between Y and $\mathbf{V}^c$ or $\mathbf{V}^c$ and $\mathbf{V}^p$: they can be causal or correlations. Figure 1 shows examples of the causal directed acyclic graphs (DAGs) that conform with our assumptions. Solid edges in the figure depict causal relationships, and dashed bidirectional arrows depict correlations.

We assume that there is a sufficient statistic $\mathbf{X}^*$ such that Y only affects $\mathbf{X}$ through $\mathbf{X}^*$, and $\mathbf{X}^*$ can be fully recovered from $\mathbf{X}$ via the function $\mathbf{X}^* := e(\mathbf{X})$, where $e(\mathbf{X})$ is unknown. We make an overlap assumption with respect to $\mathbf{V}^p$ on the source distribution, P_s: we assume that $P_s(\mathbf{V}^p)P_s(Y) \ll P_s(\mathbf{V}^p, Y)$. We also assume that $\mathbf{V}^p$ has a bounded variance. In the appendix, we give examples for the DAGs in figure 1.

Risk invariance and shortcuts. We define the generalization risk of a function f on a distribution P as $R_P = \mathbb{E}_{X,Y\sim P}[\ell(f(X), Y)]$, where ℓ is an appropriate loss function. We focus on obtaining an optimal *risk invariant* predictor, whose risk is invariant across a family of target distributions $\mathcal{P}$ that can be obtained from P_s by interventions on the DAGs in Figure 1. Specifically, we consider interventions on any non-causal relationship that keep the marginal distribution of Y constant[1]. For example, each distribution in the target family of distributions described by the DAG in 1(a) can be obtained by replacing the source conditional distribution $P_s(\mathbf{V}^p \mid Y)$ with a target conditional distribution $P_t(\mathbf{V}^p \mid Y)$. In this case, the target set of distributions is: $\mathcal{P} = \{P_s(\mathbf{X} \mid \mathbf{X}^*, \mathbf{V}^p)P_s(\mathbf{X}^* \mid Y)P_s(Y \mid \mathbf{V}^c)P(\mathbf{V}^c)P_t(\mathbf{V}^p \mid Y)\}$ This family allows the marginal dependence between Y and $\mathbf{V}^p$ to change arbitrarily. We define the set of risk invariant predictors to be all predictors that have the same risk for all $P_t \in \mathcal{P}$, $\mathcal{F}_{\text{rinv}} = \{f : R_{P_t}(f) = R_{P'_t}(f) \quad \forall P_t, P'_t \in \mathcal{P}\}$ and an optimal risk-invariant predictor f_{rinv} to have the property $f_{\text{rinv}} \in \arg\min_{f\in\mathcal{F}_{\text{rinv}}} R_{P_t}(f) \quad \forall P_t \in \mathcal{P}$. The definition of $\mathcal{P}$ also allows us to define a set of shortcuts that we care to remove: these are the set of shortcuts that would lead to varying risk across different distributions in $\mathcal{P}$. We will refer to this set as $\mathcal{P}$-specific shortcuts, but drop such notation when it is implied from the text.

[1]Extending our analysis to settings where the marginal distribution of Y also changes is possible, but would introduce some notational overhead. It would require that a re-weighted risk be invariant across such a family.

The sufficiency of $\mathbf{V}^p$ for $\mathcal{P}$-shortcut removal. One of the insights of our work is that by taking into account the causal DAG that generates the data, we are able to identify a small subset of the auxiliary labels that are sufficient to induce robustness across $\mathcal{P}$. Specifically, for any DAG that satisfies the properties outlined above, we show that it is sufficient to remove shortcuts that are labeled by $\mathbf{V}^p$ to achieve robustness. We formally state this in the following proposition.

Proposition 1. *Let $T(P_s)$ be any transformation that renders $Y \perp\!\!\!\perp_{T(P_s)} \mathbf{V}^p$. Under such transformation, the Bayes optimal predictor is a function of $\mathbf{X}^*$ only and is asymptotically risk invariant.*

The proof follows from the fact that $\mathbf{X}^*$ d-separates $Y, \mathbf{X}$ when $Y \perp\!\!\!\perp_{T(P_s)} \mathbf{V}^p$. Since the full statement of the proof is identical that of proposition 1 in [26], it is omitted. The proposition states that any transformation that renders $\mathbf{X}$ independent of $\mathbf{V}^p$ is sufficient to give us risk invariance for DAGs that satisfy the assumptions outlined above. Meaning the only shortcuts that we care about are ones induced by $\mathbf{V}^p$. Transformations T include conditioning on $\mathbf{V}^p$ or reweighting the distribution. As shown in [26], conditioning might lead to poor estimators especially when training involves small batches. So we focus on reweighting schemes. We use P° to denote the outcome of such a reweighting transformation, i.e., $P^\circ = T(P_s)$, with $Y \perp\!\!\!\perp_{P^\circ} \mathbf{V}^p$. We refer to this P° as the ideal distribution. In the DR example, this distribution is one where we are equally likely to observe a man or a woman with DR.

3. Identifying a sufficient subset of shortcuts

Our training strategy follows two steps. First, we develop a novel approach to identify $\mathbf{V}^p$. Second, by extending

previous work on single shortcut removal, we suggest an approach which leverages the results from the first step to train predictors that are robust to arbitrary types and dimensionality of auxiliary labels and target labels.

Examining the DAGs described in figure 1 reveals that the variables $\mathbf{V}^p$ have two properties which can be exploited to differentiate them from $\mathbf{V}^{\mathsf{c}}$. We state those two properties in the following proposition.

Proposition 2. *For all $V^i \in \mathbf{V}^d$, the following two properties hold: (1) $Y \perp\!\!\!\perp_{P_s} V^i \mid \boldsymbol{V}^{d\setminus i} \Rightarrow V^i \notin \mathbf{V}^p$, and (2) $\mathbf{X} \not\perp\!\!\!\perp_{P_s} V^i \mid Y, \boldsymbol{V}^{d\setminus i} \Leftrightarrow V^i \in \mathbf{V}^p$*

Proposition 2 states that if any V^i is independent of Y conditional on the rest of the auxiliary variables, it is not in $\mathbf{V}^p$, and that for any V^i in $\mathbf{V}^p$, it must hold that $\mathbf{X}$ is not independent of such a variable conditional on all other auxiliary labels. These two properties provide us with two tests that enables us to identify which auxiliary labels mark shortcuts that are necessary to account for to induce robustness versus ones which are not. In principal, we can apply non-parametric conditional independence tests to each of the auxiliary labels to identify whether it satisfies the two properties. However, the power of non-parametric independence tests has been shown to decline as a function of of the dimension of the data [28, 29]. This dependence on the dimension of the data makes testing if $\mathbf{X} \not\perp\!\!\!\perp_{P_s} V^i \mid Y, \boldsymbol{V}^{d\setminus i}$ particularly difficult in situations where $\mathbf{X}$ is high dimensional, which is the case for high resolution images.

Instead, we seek out to find a low dimensional representation $s(\mathbf{X})$, with $s \in \mathcal{S}$ such that if and only if $\mathbf{X} \not\perp\!\!\!\perp_{P_s} V^i \mid Y, \boldsymbol{V}^{d\setminus i}$ then it also true that $s(\mathbf{X}) \not\perp\!\!\!\perp_{P_s} V^i \mid Y, \boldsymbol{V}^{d\setminus i}$. Intuitively, if $\mathbf{X}$ contains any information about a given $V^i \in \mathbf{V}^d$ in some source distribution P_s, $s(\mathbf{X})$ must retain such information. This intuition implies that taking $s(\mathbf{X})$ to be the empirical risk minimizing function that predicts $\mathbf{V}^d$ from $\mathbf{X}$, is a good reduction.

To prove the validity of this simple reduction, we assume that $\mathbf{V}^p$ is s-representable. Meaning there exists some $s \in \mathcal{S}$ that can perfectly predict $\mathbf{V}^p$. We do not require that such an s is identifiable using finite samples. We note that under the causal DAGs in figure 1, for an appropriately chosen $\mathcal{S}$, there should exist performant (albeit not perfect) predictors of $\mathbf{V}^p$ from $\mathbf{X}$ since $\mathbf{V}^p$ causes $\mathbf{X}$. In the appendix, we discuss cases where this assumption can be relaxed.

Proposition 3. *For a loss function ℓ, and function space $\mathcal{S}$, let $s^*(\mathbf{X}) = \mathit{argmin}_{s\in\mathcal{S}} \mathbb{E}_{P_s}[\ell(s(\mathbf{X}), \mathbf{V}^d)]$. Then $s^*(\mathbf{X}) \not\perp\!\!\!\perp_{P_s} V^i \mid Y, \boldsymbol{V}^{d\setminus i} \Leftrightarrow \mathbf{X} \not\perp\!\!\!\perp_{P_s} V^i \mid Y, \boldsymbol{V}^{d\setminus i}$, for all $V^i \in \mathbf{V}^d$*

Propositions 2 and 3 give us a practical and efficient procedure to identify a subset of $\mathbf{V}^d$ that is sufficient for $\mathcal{P}$-shortcut removal. For each V^i, we propose first testing if $Y \perp\!\!\!\perp_{P_s} V^i \mid \boldsymbol{V}^{d\setminus i}$. We remove labels for which this relationship holds (consistent with condition 1 of proposition 2). We use $\underline{d}$ to denote the remaining set of auxiliary label indices. For the remaining labels in $\underline{d}$, we test if the second condition of proposition 2 holds as follows. We split the training data into two sub samples $\mathcal{D}_1$ and $\mathcal{D}_2$. We use $\mathcal{D}_1$ to train a model $s : \mathbf{X} \to \mathbf{V}^d$. We then proceed by predicting the value of $S = s(\mathbf{x}_i)$ for $i \in \mathcal{D}_2$, and testing if $S \perp\!\!\!\perp V^i \mid Y, \boldsymbol{V}^{\underline{d}/\,i}$ for all $i \in \underline{d}$.

To conduct the conditional independence tests, we use kernel-based conditional independence (KCIT) tests [37]. Such methods ascertain conditional independencies by analyzing the cross covariance operator. Intuitively, the cross-covariance operator can be thought of as an extension of the covariance matrix when the variables are infinite dimensional. We formally define it in the appendix.

In KCIT, the cross covariance operator is used to conduct a hypothesis test with the null hypothesis defined as $\mathbf{X} \perp\!\!\!\perp_{P_s} V^i \mid Y, \boldsymbol{V}^{d\setminus i}$, for example in our case. We use the Gamma approximation method suggested in [37] to approximate the null distribution and reject the null if the p-value corresponding to the independence test is less than a pre-specified significance level. To account for the multiple hypothesis tests, we set the significance level to be low (0.001), following the authors of KCIT. We use the radial basis function (RBF) to estimate the kernel matrices, and use the median heuristic described in [17] to set the kernel bandwidth. Finally, KCIT requires setting a parameter ϵ, which is a small regularization parameter. We set $\epsilon = 10^{-3}$ as suggested by the authors but we find that the tests are generally robust to this hyperparameter.

This procedure gives us a subset of $\widehat{\mathbf{V}}^p$, which is an estimate of $\mathbf{V}^p$ that is sufficient for shortcut removal. When characteristic kernels such as the RBF are used as the basis for the RKHS over which we measure the cross covariance operator, Zhang et al. [37] show that KCIT is asymptotically consistent, which in turns mean that $\widehat{\mathbf{V}}^p$ is an asymptotically consistent estimate of $\mathbf{V}^p$.

4. Building risk invariant predictors

Given the set $\widehat{\mathbf{V}}^p$, the challenge of building an invariant predictor reduces to an extension of Makar et al. [26]. In that work, the authors assume that (1) $\mathbf{V}^{\mathsf{c}} = \emptyset$, (2) the auxiliary and target labels are binary, and (3) there is a single, binary auxiliary label. We relax these assumptions.

Reweighting to recover P°. Makar et al. [26] show that by reweighting data sampled from an arbitrary P_s to generate a pseudo-sample from P°, the empirical risk minimizer f^* is asymptotically risk invariant across $\mathcal{P}$. However, their proposed reweighting scheme assumes that $\mathbf{V}^p$ and Y are binary. Instead we leverage permutation weighting [3] which allows for arbitrarily valued $\mathbf{V}^p$ and Y. Per-

mutation weighting proceeds by permuting Y in the training data to create $\mathcal{D}' = \{(\mathbf{x}_i, y_{\pi(i)}, \mathbf{v}_i^d)\}_{i=1}^n$, where π is a random permutation of the indices. The original $\mathcal{D}$ and $\mathcal{D}'$ are stacked and a label $C \in \{0, 1\}$ is given to examples in the observed and permuted data respectively. A classifier $\eta : \mathcal{Y} \times \mathcal{V}^p \to \{0, 1\}$ is trained to learn $P_s(C = 1 \mid Y, \mathbf{V}^p)$. The weights are then computed as:

$$u_i = \frac{\eta(\mathbf{v}_i^p, y_i)}{1 - \eta(\mathbf{v}_i^p, y_i)} = \frac{P_s(C = 1 \mid \mathbf{v}_i^p, y_i)}{P_s(C = 0 \mid \mathbf{v}_i^p, y_i)}. \quad (1)$$

We use $\tilde{u}_i$ to denote a normalized version of u_i such that $\sum_i \tilde{u}_i = 1$. Under this reweighting scheme, the empirical risk minimizer $f^* = \text{argmin}_f \sum_i \tilde{u}_i \ell(f(\mathbf{x}_i), y_i)$ is asymptotically risk invariant. The proof for this statement is identical to results by Makar et al. [26] and is hence omitted.

Causally-motivated regularization for lower variance. Reweighting leads to estimators that are inefficient in finite samples [10]. Similar to [26], we penalize models which encode a correlation between $\phi(\mathbf{X})$ and $\mathbf{V}^p$ to improve the efficiency of our approach. However, instead of using the maximum mean discrepancy, which assumes that the auxiliary label is a single binary label, to enforce independence between $\phi(\mathbf{X})$ and $\mathbf{V}^p$, we use the Hilbert Schmidt Independence Criterion (HSIC). The HSIC measures the dependence between two vectors [17]. When $P_s = P^\circ$, we can penalize $\text{HSIC}(\phi(\mathbf{X}), \mathbf{V}^p)$ to enforce the desired independencies. When $P_s \neq P^\circ$, we need to penalize a weighted version of the HSIC. This weighting in necessary since the independence property only holds under P°. We use the weighted HSIC estimator suggested by [21]. Putting all components of our approach together the final objective to minimize the following loss function

$$\mathcal{L} = \sum_i \tilde{u}_i \ell(h(\phi(\mathbf{x}_i)), y_i) + \alpha \cdot \widehat{\text{HSIC}}_\gamma^{\boldsymbol{u}}(\phi(\mathbf{X}), \widehat{\mathbf{V}}^p), \quad (2)$$

where $\alpha > 0$ is a hyperparameter that controls the cost of violating the HSIC penalty, $\widehat{\text{HSIC}}_\gamma^{\boldsymbol{u}}$ is the estimate of the HSIC, computed over samples weighted by u which is defined in equation (1) using a kernel with bandwidth γ. In the appendix, we show that our estimator inherits the finite sample efficiency guarantees of the methods described in [26]. Our cross-validation procedure (described in the appendix) is a modification of the one presented in [26].

5. Experiments

We study the performance of our approach in two tasks: predicting diabetic retinopathy (presented in the appendix), and predicting bird type. The latter is based on semi-simulated task where the data generation process follows the DAG described in figure 1(a). Specifically, we generate a high dimensional set of auxiliary labels with a small subset that affects both Y and $\mathbf{X}$ while the rest only affect Y. We follow Sagawa et al. [30] by constructing a semi-synthetic waterbirds dataset where the task is to predict Y, the type of bird (land or water). In this setting $\mathbf{V}^p$ is 2 dimensional, with V^{p0} representing the image background (land or water) and V^{p1} camera artifacts (present or absent). To generate the background shortcut, we combine images of water and land birds extracted from the CUB dataset [36] with water and land background extracted from the Places dataset [38]. To generate the camera artifact shortcut, we add small black patches to the image if camera artifacts are present. In addition, we generate 10 auxiliary labels ($\mathbf{V}^\mathbf{c}$) that affect the outcome Y but not the image $\mathbf{X}$. Additional details are included in the appendix.

We generate the source distribution P_s such that $P_s(V^{p0} = 1 \mid Y = 1) = P_s(V^{p0} = 0 \mid Y = 0) \approx 0.75$, and $P_s(V^{p1} = 1 \mid Y = 1) = P_s(V^{p1} = 0 \mid Y = 0) \approx 0.65$. We also generate three test distributions: P_s, P_{Flip}, and P°. P_s is the same as the training distribution. P° is the ideal distribution, where $P^\circ(V^{p0} = 1 \mid Y = 1) = P^\circ(V^{p0} = 0 \mid Y = 0) = P^\circ(V^{p1} = 1 \mid Y = 1) = P^\circ(V^{p1} = 0 \mid Y = 0) = 0.5$. Finally, P_{Flip} is the most dissimilar to the training distribution, where the relationship between V^{p0}, V^{p1} and Y is flipped in that $P_{\text{Flip}}(V^{p0} = 1 \mid Y = 1) = P_{\text{Flip}}(V^{p0} = 0 \mid Y = 0) \approx 0.25$, and $P_{\text{Flip}}(V^{p1} = 1 \mid Y = 1) = P_s(V^{p1} = 0 \mid Y = 0) \approx 0.35$. We introduce noise by randomly flipping 1% of the labels. We present the results from 10 simulations. Additional training details are included in the appendix.

Baselines. We compare our approach to the following baselines: **L2** is the standard neural network trained to minimize the empirical risk, with an $L2$ penalty on the model weights. **W-L2-FullV** minimizes the weighted empirical risk, with the weights computed as defined in equation 1 but using the full set of 12 auxiliary variables, $\mathbf{V}^d$. **W-L2-S** is similar to W-L2-FullV but it follows the first step in our approach to identify a sufficient set of auxiliary labels to compute the sample weights. **W-L2-HDX** is similar to W-L2-S but instead of first reducing $\mathbf{X}$ to the low dimensional $s(\mathbf{X})$, it conducts the conditional independence tests on the raw input $\mathbf{X}$. **W-HSIC-FullV** and **W-HSIC-HDX** are similar to W-L2-FullV and W-L2-HDX respectively but instead of an L2 penalty, they penalizes the HSIC penalty. Note that as Sagawa et al. [30] show, the baselines W-L2-FullV, W-L2-S and W-L2-HDX are equivalent to distributionally robust optimization in some special cases.

Results. By reducing $\mathbf{X}$ to its low dimensional sufficient statistic, our approach is able to correctly identify the two true auxiliary labels in all 10 simulations. By contrast, utilizing the full $\mathbf{X}$ rather than $s(\mathbf{X})$ to conduct the conditional independence tests identifies the correct auxiliary labels in only 1 simulation. Figure 2 shows the AUROC

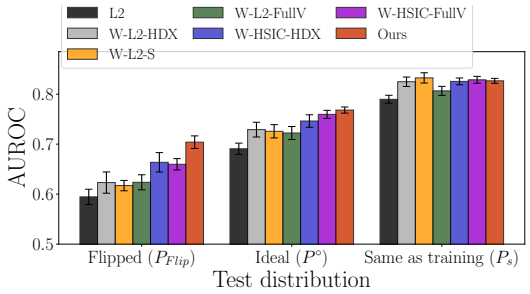

Figure 2. Results on the Waterbirds dataset: test distribution (x-axis) versus AUROC ($y-$axis). Our approach outperforms others in the most severe distribution shifts and performs comparably to others in-distribution. Note that our approach is equivalent to W-HSIC-S.

(y-axis), on the three different test distributions $P_{\text{Flip}}, P^{\circ}$, and P_s (x-axis). Our approach outperforms all others under distribution shift and performs comparably to the best models in-distribution. W-HSIC-HDX and W-HSIC-FullV are unable to achieve the same level of robustness as our approach highlighting the limitation of conducting the conditional independence tests on the full $\mathbf{X}$, and the importance of selecting a sufficient subset of shortcuts respectively.

6. Conclusion

We presented an approach to identify a sufficient set of shortcuts and leverage the identified shortcuts to build predictors that are invariant to distribution shifts. We analyzed the theoretical properties of our approach, showing that it is both consistent and efficient. Empirically, we showed that our approach outperforms others using a semi-simulated dataset and a medical dataset.

Acknowledgements

We thank the anonymous reviewers for their feedback. We also thank Alex D'Amour for his insightful comments. This work was partially supported by the National Science Foundation under Grant No. 2153083.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

[2] A. Alqaraawi, M. Schuessler, P. Weiß, E. Costanza, and N. Berthouze. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 275–285, 2020.

[3] D. Arbour, D. Dimmery, and A. Sondhi. Permutation weighting. In *International Conference on Machine Learning*, pages 331–341. PMLR, 2021.

[4] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[5] A. Azulay and Y. Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv preprint arXiv:1805.12177*, 2018.

[6] G. Balakrishnan, Y. Xiong, W. Xia, and P. Perona. Towards causal benchmarking of bias in face analysis algorithms. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII*, pages 547–563, 2020.

[7] S. Beery, G. Van Horn, and P. Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, 2018.

[8] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 79(1): 151–175, 2010.

[9] K.-M. Chueh, Y.-T. Hsieh, and S.-L. Huang. Prediction of gender from macular optical coherence tomography using deep learning. *Investigative Ophthalmology & Visual Science*, 61(7):2042–2042, 2020.

[10] C. Cortes, Y. Mansour, and M. Mohri. Learning bounds for importance weighting. *Advances in neural information processing systems*, 23, 2010.

[11] J. Cuadros and G. Bresnick. Eyepacs: an adaptable telemedicine system for diabetic retinopathy screening. *Journal of diabetes science and technology*, 3 (3):509–516, 2009.

[12] R. Geirhos, C. R. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann. Generalisation in humans and deep neural networks. In *Advances in neural information processing systems*, pages 7538–7550, 2018.

[13] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*, 2020.

[14] N. Golowich, A. Rakhlin, and O. Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299. PMLR, 2018.

[15] M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, and B. Schölkopf. Domain adaptation with conditional transferable components. In *International conference on machine learning*, pages 2839–2848. PMLR, 2016.

[16] A. Gretton and L. Györfi. Consistent nonparametric tests of independence. *The Journal of Machine Learning Research*, 11:1391–1423, 2010.

[17] A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. *Advances in neural information processing systems*, 20, 2007.

[18] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.

[19] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

[20] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[21] R. Hu, D. Sejdinovic, and R. J. Evans. A kernel test for causal association via noise contrastive backdoor adjustment. *arXiv preprint arXiv:2111.13226*, 2021.

[22] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pages 125–136, 2019.

[23] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

[24] F. Li, Z. Liu, H. Chen, M. Jiang, X. Zhang, and Z. Wu. Automatic detection of diabetic retinopathy in retinal fundus photographs based on deep learning algorithm. *Translational vision science & technology*, 8 (6):4–4, 2019.

[25] Z. Lipton, Y.-X. Wang, and A. Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018.

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

[27] R. Okada, Y. Yasuda, K. Tsushita, K. Wakai, N. Hamajima, and S. Matsuo. Glomerular hyperfiltration in prediabetes and prehypertension. *Nephrology Dialysis Transplantation*, 27(5):1821–1825, 2012.

[28] A. Ramdas, S. J. Reddi, B. Póczos, A. Singh, and L. Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[29] S. Reddi, A. Ramdas, B. Poczos, A. Singh, and L. Wasserman. On the High Dimensional Power of a Linear-Time Two Sample Test under Mean-shift Alternatives. In G. Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 772–780, San Diego, California, USA, 09–12 May 2015. PMLR. URL https://proceedings.mlr.press/v38/reddi15.html.

[30] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[32] A. Subbaswamy and S. Saria. Counterfactual normalization: Proactively addressing dataset shift and improving reliability using causal mechanisms. *arXiv preprint arXiv:1808.03253*, 2018.

[33] A. Subbaswamy, P. Schulam, and S. Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3118–3127. PMLR, 2019.

[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. corr abs/1512.00567 (2015), 2015.

[35] V. Veitch, A. D'Amour, S. Yadlowsky, and J. Eisenstein. Counterfactual invariance to spurious correlations in text classification. *Advances in Neural Information Processing Systems*, 34, 2021.

[36] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[37] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 804–813, 2011.

[38] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

A. Related work

Existing work tackling out-of-distribution generalization tends to fall into two categories: those which assume access to some (usually unlabeled) examples from the target domain (e.g., [15, 18, 25, 8]) and those which do not (e.g., [33, 32, 26, 35, 30]). Our work falls into the latter category.

Robustness to known shortcuts. Similar to our work, a number of authors adapt causal ideas for the purpose of out-of-distribution generalization when samples from the target domain are unavailable. By contrast to our work, this line of work tends to assume that the sources of bias (or shortcuts) are known *a priori*. For example, Subbaswamy et al. [33] assume the availability of a "selection diagram" that specifies which variables have a unstable relationship with the target label, and hence could be shortcuts. Absent prior knowledge, the authors suggest constructing this selection diagrams using conditional independence tests. We show here that such tests are unreliable when the variables are high dimensional, and present an solution to this limitation. The assumption of known shortcuts is implicit in other work (e.g., [23, 30, 4]) where the authors aim to find the best predictor over a set of possible distributions. Here, defining such a set requires knowledge of the meaningful shortcuts. In the experiments section, we show that our approach, by identifying a subset of relevant shortcuts, is able to outperform approaches equivalent to [30].

Most similar to our work is [26], where the authors study an anti-causal prediction problem similar to ours. Unlike us, they assume that there is a single shortcut labeled by a binary auxiliary label. Our work can be viewed as a direct extension of this work to relax assumptions about the type and dimension of the auxiliary label as well as the prior knowledge about the shortcut.

Shortcut identification. One approach that has been suggested to identify possible shortcuts is by leveraging interpretability methods such as saliency maps [31] which visually highlight which parts of an image is most important for a prediction. However, user-based studies have found that saliency maps often have limited utility in explaining model features [2]. In addition, in domains such as healthcare, leveraging saliency maps to identify shortcuts might require expert knowledge. In [6], the authors suggest manipulating the observed examples by intervening on possible shortcuts and measuring the behviour of the model under such interventions. However, such work relies on being able to faithfully manipulate the observed data, which is not possible in most cases.

B. Additional preliminary information

Examples of DAGs. To establish the intuition underlying the DAGs in figure 1, we highlight some possible scenarios that these DAGs depict. In all DAGs, $\mathbf{V}^p$ can denote the quality of the funduscope, which is used to capture the image $\mathbf{X}$, or the sex of the patient which has been shown to affect the shape of the retina [9]. In figure 1(a), $\mathbf{V}^c$ can denote high sugar intake: it can cause diabetes and its complications such as DR but it likely does not directly affect the appearance of the retina ($\mathbf{X}$) independently of Y. In figure 1(b), $\mathbf{V}^c$ can denote conditions that tend to co-occur with DR such as kidney diseases [27] in figure 1(c), $\mathbf{V}^c$ could be socio-economic characteristics correlated with access to high quality funduscopes (or healthcare in general) while in figure 1(d) $\mathbf{V}^c$ could be sex-specific diseases such as cervical cancer.

Relaxing s-representability In cases where $\mathbf{V}^p$ is binary, the assumption of s-representability can be relaxed. In that case it is sufficient to assume that $\mathcal{S}$ contains some s with bounded δ error such that δ is less than the proportion of the smallest subgroup defined by $Y, \mathbf{V}^p$. Under such assumption, the following proposition establishes the validity of this simple reduction.

Cross covariance operator definition The formal definition of the cross covariance operator is stated below

Definition 1. *Let Z, Z' be a pair of random variables defined on $\mathcal{Z} \times \mathcal{Z}'$ and let $\Omega_{\mathcal{Z}}$ and $\Omega_{\mathcal{Z}'}$ be two Reproducing Kernel Hilbert Spaces (RKHSs) defined on $\mathcal{Z}$ and $\mathcal{Z}'$. Define the cross-covariance operator of Z, Z, $C_{zz'} : \Omega_{\mathcal{Z}} \to \Omega_{\mathcal{Z}'}$ such that $\langle f, C_{zz'} g \rangle = Cov[g(Z), g'(Z')],\ \forall g \in \Omega_{\mathcal{Z}}, g' \in \Omega_{\mathcal{Z}'}$*

C. Proofs for section 3

C.1. Proof for proposition2

The proof relies on examining the d-separation properties implied by the DAGs. We will assume that V^p and V^c are single dimensional for simplicity.

First property. Note that for all DAGs, the two paths $Y \leftarrow U \rightarrow V^i$ is unblocked by any other $\boldsymbol{V}^{d\backslash i}$, which means that for all $V^i \in \mathbf{V}^p$, $Y \not\perp\!\!\!\perp V^i \mid \boldsymbol{V}^{d\backslash i}$. Unfortunately, the same property holds for $V^i \in \mathbf{V}^c$ in all DAGs except the DAG in 1(d). To see that note that in DAG 1(a) the path $Y \leftarrow \mathbf{V}^p$ is unblocked by any other variables in $\mathbf{V}^d$, in DAG 1(b), $Y \leftarrow U_2 \rightarrow \mathbf{V}^c$ is unblocked by any other variables in $\mathbf{V}^d$, in DAG 1(c) the path $Y \leftarrow U_1 \rightarrow V^p \rightarrow U_2 \leftarrow V^c$ is unblocked by conditioning on V^p, which is a collider. So in all these DAGs, $V^i \in \mathbf{V}^p$, $Y \not\perp\!\!\!\perp V^i \mid \boldsymbol{V}^{d\backslash i}$. However, in DAG 1(d) the path $Y \leftarrow U \rightarrow V^p \rightarrow V^c$ is blocked by V^p so $Y \perp\!\!\!\perp V^i \mid \boldsymbol{V}^{d\backslash i}$ for $V^i = V^c$ only.

Second property. We start by proving the direction: $\mathbf{X} \not\perp\!\!\!\perp_{P_s} V^i \mid Y, \boldsymbol{V}^{d/\,i} \Rightarrow V^i \in \mathbf{V}^p$. Suppose that there exists some i such that $\mathbf{X} \perp\!\!\!\perp_{P_s} V^i \mid Y, \boldsymbol{V}^{d/\,i}$ but $V^i \in \mathbf{V}^p$. This means that all paths between V^i and $\mathbf{X}$ are blocked by $\boldsymbol{V}^{d/\,i}$. However in all DAGs for all $V^i \in \mathbf{V}^p$, the the path $\mathbf{X} \leftarrow V^p$ cannot be blocked via conditioning on any other variables, which represents a contradiction.

We next prove the direction $V^i \in \mathbf{V}^p \Rightarrow \mathbf{X} \not\perp\!\!\!\perp_{P_s} V^i \mid Y, \boldsymbol{V}^{d/\,i}$. Suppose there exists some i such that $V^i \notin \mathbf{V}^p$ but $\mathbf{X} \not\perp\!\!\!\perp_{P_s} V^i \mid Y, \boldsymbol{V}^{d/\,i}$. Then there exists an unblocked path between V^i and $\mathbf{X}$. This is a contradiction because:

1. In DAG 1(a) the only path between V^c and $\mathbf{X}$ is $\mathbf{X} \leftarrow V^p \leftarrow U \rightarrow Y \leftarrow V^c$. This path is blocked by conditioning on V^p.
2. In DAG 1(b) the only path between V^c and $\mathbf{X}$ is $\mathbf{X} \leftarrow V^p \leftarrow U \rightarrow Y \leftarrow U_2 \rightarrow V^c$. This path is blocked by conditioning on V^p.
3. In DAG 1(c) the only path between V^c and $\mathbf{X}$ is $\mathbf{X} \leftarrow V^p \leftarrow U_2 \rightarrow V^c$, which is blocked by V^p.
4. In DAG 1(d) the path between V^c and $\mathbf{X}$ is $\mathbf{X} \leftarrow V^p \rightarrow V^c$ which is blocked by V^p.

C.2. Proof for proposition 3

The direction $\mathbf{X} \not\perp\!\!\!\perp_{P_s} V^i \mid Y, \boldsymbol{V}^{d/\,i} \Rightarrow s^*(\mathbf{X}) \not\perp\!\!\!\perp_{P_s} V^i \mid Y, \boldsymbol{V}^{d/\,i}$ is easy to prove as follows. Suppose that $\mathbf{X} \perp\!\!\!\perp_{P_s} V^i \mid Y, \boldsymbol{V}^{d/\,i}$ but $s^*(\mathbf{X}) \perp\!\!\!\perp_{P_s} V^i \mid Y, \boldsymbol{V}^{d/\,i}$. This statement presents an immediate contradiction since any functions of independent random variables must be independent so such an s^* cannot exist.

Next, the direction $s^*(\mathbf{X}) \not\perp\!\!\!\perp_{P_s} V^i \mid Y, \boldsymbol{V}^{d/\,i} \Rightarrow \mathbf{X} \not\perp\!\!\!\perp_{P_s} V^i \mid Y, \boldsymbol{V}^{d/\,i}$. Suppose that $s^*(\mathbf{X}) \perp\!\!\!\perp_{P_s} V^i \mid Y, \boldsymbol{V}^{d/\,i}$ but $\mathbf{X} \not\perp\!\!\!\perp_{P_s} V^i \mid Y, \boldsymbol{V}^{d/\,i}$. By proposition 2, $\mathbf{X} \not\perp\!\!\!\perp_{P_s} V^i \mid Y, \boldsymbol{V}^{d/\,i}$ implies that $V^i \in \mathbf{V}^p$. And by the assumption that $\mathbf{V}^p$ is s-representable, we have that there exists some s such that $V^i = s(\mathbf{X})$. Such an s is an empirical risk minimizer achieving the minimum possible risk of 0. By definition for such an s, $s(\mathbf{X}) \not\perp\!\!\!\perp_{P_s} V^i \mid Y, \boldsymbol{V}^{d/\,i}$ and $\mathbb{E}_{P_s}[s(\mathbf{X})|V^i = v^i, Y = y, \boldsymbol{V}^{d/\,i} = \boldsymbol{v}^{d/\,i}] \neq \mathbb{E}_{P_s}[s(\mathbf{X})|V^i = \tilde{v}^i, Y = y, \boldsymbol{V}^{d/\,i} = \boldsymbol{v}^{d/\,i}]$ for all $v^i \neq \tilde{v}^i$ and all z, y. However, for $s^*(\mathbf{X}) \perp\!\!\!\perp_{P_s} V^i \mid Y, \boldsymbol{V}^{d/\,i}$ to hold, it must be true that $\mathbb{E}_{P_s}[s(\mathbf{X})|V^i = v^i, Y = y, \boldsymbol{V}^{d/\,i} = \boldsymbol{v}^{d/\,i}] = \mathbb{E}_{P_s}[s(\mathbf{X})|V^i = \tilde{v}^i, Y = y, \boldsymbol{V}^{d/\,i} = \boldsymbol{v}^{d/\,i}]$ for all $v^i \neq \tilde{v}^i$ and all z, y. This means that s^* must have an empirical risk greater than 0, i.e., it is not an empirical risk minimizer which is a contradiction.

D. Proofs for section 4

D.1. Reducing sample complexity

To explain how the HSIC penalty leads to a reduction in the sample complexity and hence the variance of the estimator, we follow the same strategy as Makar et al. [26] in studying a simple setting where we focus on a linear function class and analyze how the suggested HSIC penalty compares to a standard $L2$-regularized function class. Our analysis is extendable to more complex neural networks e.g., through approaches studied in [14].

For some $A > 0, \tau \geq 0$, define the two function classes:

$$\mathcal{F}_{L_2} := \{f : \mathbf{x} \mapsto \sigma(\mathbf{w}^\top \mathbf{x}), \|\mathbf{w}\|_2 \leq A\}, \tag{3}$$

$$\mathcal{F}_{L_2,\text{HSIC}} := \{f : \mathbf{x} \mapsto \sigma(\mathbf{w}^\top \mathbf{x}), \|\mathbf{w}\|_2 \leq A, \text{HSIC} \leq \tau\}. \tag{4}$$

In this simple function class, the HSIC constraint restricts the projection of the weights $\mathbf{w}$ onto $\Delta := \text{Cov}_{P^\circ}(\mathbf{X}, \mathbf{V}^p)$. To simplify notation, we assume that the variance of $\mathbf{V}^p$ is the vector of ones. However, if that is not true, $\mathbf{V}^p$ can be rescaled

such that the variance is the vector of ones, which is possible because of the assumption of bounded variance. The matrix Δ is the average change in $\mathbf{X}$ caused by intervening to change each dimension in $\mathbf{V}^p$ under P°. Define the projection matrix $\Pi := \Delta(\Delta^\top\Delta)^{-1}\Delta^\top = \|\Delta\|_2^{-2}\Delta\Delta^\top$, which projects any vector onto Δ, and $\mathbf{w}_\perp := \Pi\mathbf{w}$ as the projection of $\mathbf{w}$ onto Δ, which can be thought of as the "irrelevant" dimension of $\mathbf{X}$. To directly compare our results with Makar et al. [26], we consider the case where $\mathbf{V}^p$ is one dimensional, i.e., Δ is a vector.

Proposition A1. *Let $f(\mathbf{x}) = \sigma(\phi(\mathbf{x})) = \sigma(\mathbf{w}^\top\mathbf{x})$ be a function contained in $\mathcal{F}_{L_2,\text{HSIC}}$. Then, $\|\mathbf{w}_\perp\| \leq \frac{\tau}{\|\Delta\|}$. and*

$$\mathfrak{R}(\mathcal{F}_{L_2}) \leq \frac{A\sqrt{B_\parallel^2 + B_\perp^2}}{\sqrt{n}}, \quad \textit{and} \quad \mathfrak{R}(\mathcal{F}_{L_2,\text{HSIC}}) \leq \frac{A \cdot B_\parallel + \tau\frac{B_\perp}{\|\Delta\|}}{\sqrt{n}}.$$

Proof. By Gretton and Györfi [16] we have that:

$$\text{HSIC}(\phi(\mathbf{X}), \mathbf{V}^p, \Omega, \Psi) \geq \sup_{\omega\in\Omega,\psi\in\Psi} \|\text{Cov}(\omega(\phi(\mathbf{X})), \psi(\mathbf{V}^p))\|_{HS},$$

where Ω and Ψ are two RKHS spaces defined over $\phi(\mathbf{X})$ and $\mathbf{V}^p$ respectively.

Taking ω, and ψ to be the identity functions, and substituting ϕ for $\mathbf{w}^\top\mathbf{x}$, we have that:

$$\begin{aligned}
\tau &\geq \text{HSIC}(\phi(\mathbf{X}), \mathbf{V}^p, \Omega, \Psi) \\
&\geq \|\text{Cov}(\mathbf{w}^\top\mathbf{X}, \mathbf{V}^p)\|_F \\
&= \|\mathbf{w}\text{Cov}(\mathbf{X}, \mathbf{V}^p)\|_2 \\
&= \|\mathbf{w}\Delta\|_2 \\
&= |\mathbf{w}\Delta|
\end{aligned}$$

where $\|.\|_F$ is the Frobenius norm, and the equalities follow from the fact that $\mathbf{w}\text{Cov}(\mathbf{X}, \mathbf{V}^p)$ is a scalar. Note that $\|\mathbf{w}_\perp\| = \frac{|\mathbf{w}\Delta|}{\|\Delta\|}$, which completes our proof for the first part of the statemenet (bound on $\|\mathbf{w}_\perp\|$). The rest of the proof follows identically to Makar et al. [26]. □

The generalization bound can also be obtained identically to Makar et al. [26], and is hence omitted. We note that if $\mathbf{V}^d$ is used instead of $\mathbf{V}^p$, the term C_P in proposition A8 of [26] is larger, leading to a larger (less favorable) generalization error bound.

E. Cross-validation

Cross-validation. The objective function in (2) depends on two hyperparameters: the cost of the HSIC penalty α, and the penalty's kernel bandwidth γ. Unlike many regularizers, the HSIC penalty depends on the distribution of the data, and is vulnerable to overfitting, such that the estimated $\widehat{\text{HSIC}}$ on the training data underestimates the population HSIC. For this reason, we follow a two-step cross-validation procedure. Letting $\mathcal{D}_{\text{valid}}$ denote a held out validation set, ϕ_{valid} denote $\{\phi(\mathbf{x}_i)\}_{i\in\mathcal{D}_{\text{valid}}}$, and similarly define $\mathbf{V}^p_{\text{valid}}$, our cross validation procedure proceeds as follows. In the first step, for a given $\alpha = \alpha_0, \gamma = \gamma_0$, we first check if the corresponding ϕ_{valid} is independent of $\mathbf{V}^p_{\text{valid}}$. We do so using the permutation test suggested by Gretton et al. [17]. This test entails creating 100 permutations of the validation set, with the k^{th} permutation defined as $\mathcal{D}' = \{\mathbf{x}_i, y_i, \mathbf{v}^p_{\pi^k(i)}\}$, and $\pi^k(i)$ is a permutation of the indices. We compute a vector of HSIC values for each of the permuted datasets, and the corresponding $1-\beta^{\text{th}}$ quantile of that vector. β is a pre-specified significance level that we use to accept or reject the null hypothesis that the estimated $\phi(\mathbf{X}), \mathbf{V}^p$ are independent. Similar to before, we set that to be 0.001 as a heuristic to account for the multiple tests. We reject α_0, γ_0 as valid hyperparameters if $\widehat{\text{HSIC}}$ as calculated on the unpermutated validation set is larger than the value corresponding to the $1-\beta^{\text{th}}$ quantile. Repeating this process for all α, γ candidates gives us a subset of the that set that encode the desired invariances. In the second step, we pick the best performing model out of this subset of candidate functions.

Figure 3. Examples of the generated waterbirds images. Left: water bird on water background. Middle: water bird on land background. Right: water bird on land background with camera artifacts.

F. Waterbirds experiments details

F.1. Data generation

We use the subset of the places images provided by Makar et al. [26] in https://github.com/mymakar/causally_motivated_shortcut_removal. We generate the data as follows. $V^{p0} \sim \text{Binomial}(0.5)$, V^{p1} is generated such that it has a 70% correlation with V^{p0}. $\mathbf{V^c}$ is drawn from Binomial(0.01). We generate the outcome $Y = \sigma(\theta_0 + \theta_{p0} V^{p0} + \theta_{p1} V^{p1} + \boldsymbol{\theta}_c^\top \mathbf{V^c} + \varepsilon)$, where $\sigma(.)$ is the sigmoid function, and $\varepsilon \sim \mathcal{N}(0, 0.5)$.

For P_s: $\theta_0 = -0.84, \theta_{p0} = 0.84, \theta_{p1} = 0.4$ and $\boldsymbol{\theta}_c \sim \mathcal{N}(0, 1)$. For P_{Flip}: $\theta_0 = 0.45, \theta_{p0} = -0.84, \theta_{p1} = -0.4$ and $\boldsymbol{\theta}_c \sim \mathcal{N}(0, 1)$. For P°: $\theta_0 = -0.15, \theta_{p0} = 0, \theta_{p1} = 0$ and $\boldsymbol{\theta}_c \sim \mathcal{N}(0, 1)$.

Examples of the generated images are in figure 3.

F.2. Training details

We split the data into 70% training and validation and 30% is a held out test set. The training and validation data is further split into 75% training and 25% validation. We resize the images to a resolution of 128×128, and train for 500 epochs.

We use ResNet-50 [20], pretrained on ImageNet. All models in this paper are implemented in TensorFlow [1]. In each of the 10 simulations, we generate different train/test splits, different draws of auxiliary labels and different bird-background-camera artifact combinations.

For all HSIC based models, we cross validate over bandwidth values $= [1.0, 10.0, 100.0, 1000.0]$, and α values $= [1e3, 1e5, 1e7, 1e9]$. We picked this set of bandwidths to cross validate over using the following heuristic: for each HSIC model, we train its corresponding unpenalized (i.e., $\alpha = 0$) model. We evaluate the HSIC of the unpenalized model at various bandwidth levels, and pick the set that has non-zero HSIC as the reasonable set to cross validate over.

For all $L2$ models we cross validate over $L2$ penalty $= [0, 0.0001]$. We use Adam optimizer, with the default learning rate 0.001 and default $\epsilon = 1e-07$.

Each model takes roughly 50 minutes to train, with of 56 models per simulation and a total of 560 models, the total compute time is roughly 470 hours on a Tesla T4 GPU.

G. Additional experiments: Diabetic Retinopathy

Setup. In this setting, we examine the validity of our approach when the outcome is non-binary. We use a publicly available dataset made available by EyePACS, LLC [11]. Approval for the use of this data set for the purpose of research was obtained via correspondence with the data curators. Here, we predict the presence and severity of diabetic retinopathy (DR) using fundus images, with $Y \in \{0, \ldots, 4\}$. To focus the analysis on the challenges pertaining to categorical outcomes, we generate a single binary auxiliary label, V^p, reflecting the presence or absence of funduscope artifacts. Similar to before, we add small black patches to the image if funduscope artifacts are present. We simulate the training distribution P_s with $P_s(V^p = 1 \mid Y = 0) = P_s(V^p = 0 \mid Y > 0) = 0.9$. We introduce noise by randomly permuting 1% of the labels.

Here, we compare two baselines to our approach: L2 is defined similar to before, W-L2 is a weighted version of L2, using

Model	AUROC (STE) Flipped (P_{Flip})	Ideal (P°)	Same (P_s)
L2	0.69 (0.009)	0.82 (0.003)	**0.92 (0.001)**
W-L2	0.68 (0.015)	0.82 (0.005)	**0.92 (0.001)**
Ours	**0.72 (0.026)**	**0.83 (0.007)**	0.91 (0.007)

Table 1. Diabetic retinopathy results: AUROCs averaged over 10 simulations and standard deviations across 3 test distributions. Our approach outperforms others especially when the distribution shift is most severe, and performs comparably to others in-distribution

weights defined with respect to V^p. We follow Li et al. [24] in using an Inception-V3 architecture [34] to train all models. We present the results from 10 simulations. In each simulation, we generate different train/test splits and different draws of auxiliary labels.

We split the data into 70% training and validation and 30% is a held out test set. The training and validation data is further split into 75% training and 25% validation. We follow [19] in preprocessing the images such that they are macula-centered, and resize them to be 299 $\times$ 299. We train each model for 2 epochs (which is sufficient since the DR data is larger than the waterbirds data). We use Adam as our optimizer, and follow tensorflow guidance in setting $\epsilon = 0.1$. We also find that a slower learning rate leads to better results for all models, so set it to 0.0001.

For the HSIC based model, we consider bandwidths = $[0.1, 1.0]$, which were picked using the same heuristic described in the waterbirds experiment, and $\alpha = [1e3, 1e5, 1e7]$. Each model takes roughly 30 minutes. With 10 models per simulation, we have 100 models which take a total of 5 hours to train on a Tesla T4 GPU.

Results. Similar to the waterbirds setting, we measure the performance of the three models on three distributions P_s, P_{Flip}, and P°, where P_{Flip} has $P_{\text{Flip}}(V^p = 1 \mid Y = 0) = P_{\text{Flip}}(V^p = 0 \mid Y > 0) = 0.1$ and P° is the ideal distribution. Table 1 shows the AUROCs averaged over 10 simulations and their corresponding standard errors. The results show that our approach vastly outperforms others in the most severe distribution shifts, and performs relatively on par with the other models in-distribution. The slight drop in accuracy in-distribution is attributable to the fact that the baselines exploit the shortcut whereas our approach does not. The results confirm that our approach extends to setting where the target label is non-binary.