From Invariant Representations to Invariant Data: Provable Robustness to Spurious Correlations via Noisy Counterfactual Matching

Ruqi Bai [1] **Yao Ji** [2] **Zeyu Zhou** [1] **David I. Inouye** [1]

Abstract

Spurious correlations can cause model performance to degrade in new environments. Prior causality-inspired work aim to learn invariant representations (e.g., IRM) but typically underperform empirical risk minimization (ERM). Recent alternatives improve robustness by leveraging test-time data, but such data may be unavailable in practice. To address these issues, we take a data-centric approach by leveraging *invariant data pairs*—training samples with equal true predictive distributions, such as counterfactuals intervening only on non-ancestors of the target. We introduce *noisy counterfactual matching* (NCM) that adds a linear constraint to ERM based on counterfactuals that achieves provable robustness to spurious correlations—even if the counterfactuals are noisy. For linear causal models, we prove that the test domain error can be upper bounded by the in domain error and a term that depends on the counterfactuals' diversity and quality. Empirically, we validate on a synthetic dataset that only a few counterfactual pairs are needed and demonstrate on real-world benchmarks (ColoredMNIST, Waterbirds, and PACS) that linear probing on a pretrained ViT-B/32 CLIP backbone improves robustness.

1. Introduction

Spurious correlations are misleading patterns in the training data—relationships between features and the target that do not hold across environments. Models trained on such correlations may perform well on training data but fail to generalize when the environment changes, as these correlations reflect confounding or coincidental associations rather than true causal links. Addressing spurious correlations is critical for building models that are robust to distribution shifts, particularly in high-stakes domains like healthcare, finance, and public services, where robustness and reliability are critical.

One major approach to handling spurious correlations is based on invariant representation learning (Peters et al., 2016; Li et al., 2018b;a; Arjovsky et al., 2019) in various forms. Some works aim to learn representations whose marginal $p(h(\mathbf{x}))$ (Li et al., 2018b) and conditional distributions $p(h(\mathbf{x})|\mathbf{y})$ (Li et al., 2018b) are invariant across domains or environments. Others invariance works are inspired by causality such as Invariant Causal Prediction (ICP) (Peters et al., 2016) which uses conditional independence tests to find features that are invariant predictors across domains. Similarly, invariant risk minimization (IRM) (Arjovsky et al., 2019) seeks to learn representations such that the predictive distribution $p(\mathbf{y}|h(\mathbf{x}))$ is invariant across domains. While these causality-based works are more theoretically grounded, they often fail to beat empirical risk minimization (ERM) on real-world tasks (Gulrajani and Lopez-Paz, 2021; Koh et al., 2021; Bai et al., 2024), which may be due to the strong assumptions required for correctness that do not necessarily hold in practice.

Given the inherent challenges of spurious correlation, other prior work leverages additional data, usually from the test domains. For example, some work use additional unlabeled test data to enhance predictions in the test domains (Ben-David et al., 2010; Daumé III, 2009; Mansour et al., 2009). A related approach is test-time adaptation which allows the model to dynamically adapt to test domain data quickly, usually without finetuning or model retraining (Azimi et al., 2022; Wang et al., 2023; Sun et al., 2020). Another relaxation requires test domain metadata, which are descriptive features of the domains like geographic location, timestamp, etc. For example, D^3G (Zhang et al., 2023a) addresses distribution shifts by leveraging such metadata under the assumption that the test distribution can be represented as a linear combination of training domain distributions. However, collecting test domain samples or metadata can be expensive or impractical in many real-world scenarios. For example, in medical related applica-

[1]Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA [2]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA. Correspondence to: David I. Inouye <dinouye@purdue.edu>.

Preprint.

tions, due to privacy restriction, test domain data might not be available when new deployment occurs. Consequently, practitioners may face an important question: *Is there a type of data solely from training domains that could provably make models invariant to spurious correlations?*

To address this question, we propose a data-centric approach that leverages *invariant data pairs*—samples from the training domains that have equal predictions under the robust classifier. One example of an invariant pair is a sample of the same object in two different domains. For example, an image of the same dog running on grass versus on a paved road would form an invariant pair. To formalize this notion with causality, we focus on counterfactuals (CF), where the intervened variables are non-ancestors of the target variable. The factual and this counterfactual will form a *counterfactual pair* that inherently satisfies the invariance property. Intuitively, these counterfactuals answer the question: "What would this sample be like if it were generated from different domain?"

While it might at first seem that counterfactual pairs (or more generally invariant pairs) would be infeasible to collect, we suggest several scenarios where these pairs are possible to collect in practice. First, when the spurious correlations are artifacts of a measurement process (e.g., x-ray machine, microscope, staining methodology, etc.), then an invariant pair could be collected by measuring the same specimen under two different environments (e.g., send the same patient to two x-ray machines). While this costs twice as much per sample, it would provide a powerful signal for making the model robust to future measurement changes (e.g., x-ray machine replacements)—indeed, we show both theoretically and empirically that only a small number of pairs are needed for enhanced robustness. Note that this approach would not even require manual labels (e.g., a diagnosis) since it is assumed that the labels would be the same. Second, when a domain expert can identify spurious features, they can directly edit spurious features of the sample. An example of this would be using image editing software including AI-based image editing to change the background of an image while keeping the subject the same (e.g., putting a cow on a boat or a fish in a desert). Furthermore, we explicitly design our approach to handle approximate or noisy counterfactual pairs, and it only requires a small number of pairs to achieve robustness. Thus, we believe this scenario is both practical and reasonable in some cases.

At a higher level, invariant data pairs enable a way to *implicitly* specify knowledge of spurious correlations instead of requiring explicit specification (e.g., specifying a causal graphical model). Analogously, collecting invariant pairs is to spurious correlation as collecting class labels is to classification. In both cases, explicitly defining the target object (either spurious correlations or class) is very challenging but implicitly defining them through examples is easier. In particular, we target applications where only implicit knowledge of spurious correlations is known (see Table 1 for a summary and Section 2.3 for a more in-depth discussion). **We hypothesize that invariant data pairs could provide a data-driven way to specify spurious correlations and give evidence for this hypothesis both empirically and theoretically in this paper.**

Table 1. An illustrative taxonomy of scenarios from explicit knowledge to no knowledge of spurious correlations. We target the applications between level 2 to level 1.

	knowledge on spurious features	pair data acquisition
level 3	explicit knowledge	model constraint
level 2	soft expert knowledge	sample editing
level 1	implicit assumed	CF pair collection
level 0	no knowledge	-

Given a small set of counterfactual pairs that satisfy the invariant property, we propose a simple method called Noisy Counterfactual Matching that simply adds a constraint to ERM that the predictions on the two points in the invariant pairs must match. Given perfect counterfactuals, NCM is similar to the objective in MatchDG (Mahajan et al., 2021), but MatchDG does not explicitly handle noisy counterfactuals or theoretically analyze the approach. For linear models, this constraint ensures that the classifier is orthogonal to difference between the two points of the pair. Figure 1 illustrates our CF pair matching idea using a simple two-dimensional binary classification with a single counterfactual pair. While ERM would learn a non-robust classifier, our method using only a *single counterfactual pair* would force the classifier to be orthogonal to the pair difference (denoted by green line) and thus make it robust to the test domain which modified the spurious feature. Intuitively, CF pairs act as negative signal and tell the model which features to ignore. Analogously, this is the opposite of the target variable $\mathbf{y}$, which acts as a positive signal and tell the model which features are relevant for prediction. While NCM does not require causal modeling to be applied, to better understand our method, we analyze NCM under the latent causal modeling setting and prove that even with noisy counterfactuals, our approach can be provably robust to spurious correlations that change in the test domain.

We summarize our contributions as follows:

1. We introduce a data-centric approach to spurious correlations based on a small dataset of (noisy) counterfactual pairs that adds a simple constraint to ERM to increase robustness.

2. We theoretically prove that, in both linear and logistic

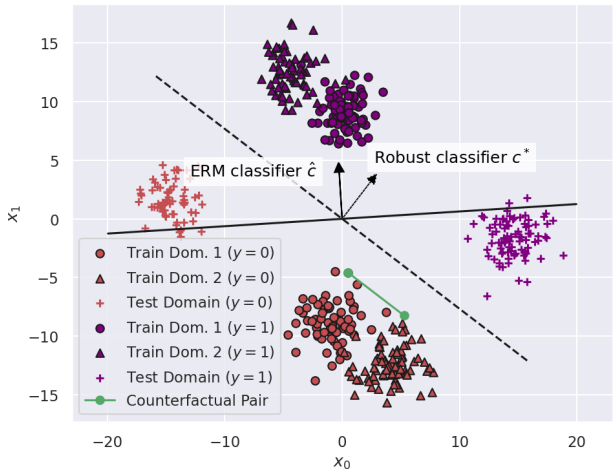


Figure 1. While ERM $\hat{c}$ on the training domains (circles and triangles) is not robust to the change in spurious feature in the unseen test domain (pluses), our NCM method with a *single* counterfactual pair (green dots) would find the robust classifier c^* by constraining the coefficients to be orthogonal to the difference in the counterfactual pair (green line). The color represents the binary label Y.

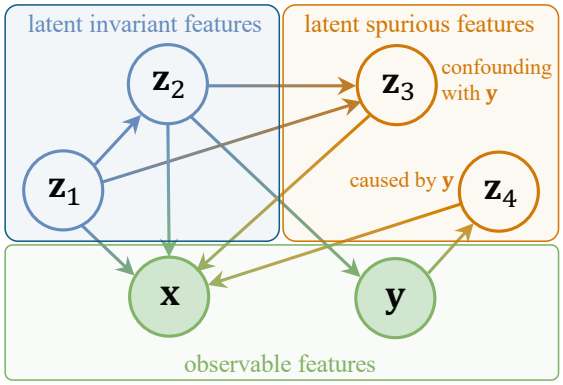


Figure 2. Illustration of the latent domain causal model. The ancestors of the target $\mathbf{y}$ are $\mathbf{z}_1, \mathbf{z}_2$, which are assumed to be invariant across domains (see Assumption 1). On the other hand, $\mathbf{z}_3, \mathbf{z}_4$ are spurious features because they are not ancestors of $\mathbf{y}$ but are descendants of $\mathbf{y}$ or its ancestors.

regression settings, NCM ensures robustness to spurious correlations, even with only small number of noisy or approximate counterfactuals.

3. We empirically validate our theory using a synthetic dataset as well as linear probing on a pretrained ViT-B/32 CLIP backbone across multiple datasets: ColoredMNIST (Arjovsky et al., 2019; Gulrajani and Lopez-Paz, 2021), Waterbirds dataset (Sagawa et al., 2019) and PACS dataset (Gulrajani and Lopez-Paz, 2021) (cf. Table 2, Table 3, Table 4).

2. Problem Setup

To formalize the goal of robustness to spurious correlation, we will consider a variant of the domain generalization (DG) problem for out-of-distribution robustness. Domain generalization considers a set of related domains or environments $\mathcal{E}$. The algorithm is given samples from a set of training environments $\mathcal{E}_{\text{train}} \subseteq \mathcal{E}$ but aims to perform well on the unseen test environments $\mathcal{E}_{\text{test}} \subset \mathcal{E}$. Formally, given samples from each training environment $\{\{(x_{e,i}, y_{e,i})\}_{i=1}^{n_e}\}_{e\in\mathcal{E}_{\text{train}}}$, the goal is to find model parameters θ that minimizes the worst case risk over all possible unseen distributions:

$$\min_{\theta} \max_{e \in \mathcal{E}_{\text{test}}} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathbb{P}_e}[\ell(h_\theta(\mathbf{x}), \mathbf{y})]\,, \tag{1}$$

where ℓ is a per-sample loss function such as mean squared or cross-entropy loss. The key challenge in domain generalization is the lack of data from the test environments. In particular, if there are spurious correlations in the training domains that do not hold in the test domains, then a model trained using ERM or similar will perform poorly on the test domains.

We consider a data-centric relaxation of this problem where we assume access to an additional small dataset of *invariant data pairs*. Intuitively, these invariant data pairs are ones that should have the same prediction under the robust model, e.g., if the task is medical diagnosis, x-rays from two different machines of the same patient should predict the same probabilities. However, to formally define invariant pairs, we need to first define optimally robust classifier given an arbitrary environment set (potentially infinite).

Definition 1 (Optimally Robust Classifier)**.** *Given a potentially infinite set of environments $\mathcal{E}$, the optimally robust classifier is defined as:*

$$h^*_{\mathcal{E}} := \arg\min_{h} \max_{e\in\mathcal{E}} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathbb{P}_e}[\ell(h(\mathbf{x}), \mathbf{y})]\,, \tag{2}$$

where the optimization is over all possible predictive functions h.

While this optimally robust classifier is not perfectly achievable like the Bayes optimal classifier, it provides a theoretically optimal classifier that will enable us to define invariant pairs. Specifically, invariant data pairs are formally defined as pairs that would give the same prediction values under

Definition 2 (Invariant Pair)**.** *Given a set of environments $\mathcal{E}$, a pair of distinct inputs $(\boldsymbol{x}, \boldsymbol{x}')$ with $\boldsymbol{x} \neq \boldsymbol{x}'$ is an* invariant data pair *if and only if the predictions under the*

optimally robust classifier are equal, i.e., $h^*_{\mathcal{E}}(\boldsymbol{x}) = h^*_{\mathcal{E}}(\boldsymbol{x}')$*, where* $h^*_{\mathcal{E}}$ *is defined as in Definition 1.*

While at first glance, invariant data pairs may seem impossible to collect, we believe the intuitive and formal notion of counterfactuals provides a good way to understand these invariant pairs. Intuitively, as described in the introduction, counterfactual data points where only spurious features (e.g., background of an image) are changed while the substantive or causal features (e.g., identity of the foreground object) are maintained. Formally, in the next sections, we will develop a causal perspective on invariant data pairs that formally defines (1) the set of environments $\mathcal{E}$ under the assumption of spurious correlation and (2) counterfactual pairs which will satisfy the invariant pairs property. Finally, we note that the idea of invariant data pairs could be considered in non-causal frameworks where the set of environments is defined in other ways (e.g., a Wasserstein-ball around the training distributions or total variation between the training distribution to the test distribution) (Albuquerque et al., 2019; Blanchard et al., 2021; Ben-David et al., 2010) and the invariant pairs would be defined differently, but we leave this to future work.

2.1. Causality Preliminaries

To formally define the set of spurious correlation environments and their corresponding invariant pairs, we first introduce some related concepts in causality. In summary, we consider that each domain (or environment[1]) corresponds to a distinct structural causal model (SCM) (Pearl, 2009, Definition 7.1.1), the differences of the SCMs are equivalent to interventions, and counterfactuals are based on applying two different SCMs to the same exogenous noise. First, we formally define an SCM.

Definition 3 (Structural Causal Model (Pearl, 2009, Definition 7.1.1))**.** *An SCM* $\mathcal{M}$ *is represented by a 3-tuple* $\langle \mathcal{U}, \mathcal{V}, \mathcal{F} \rangle$*, where* $\mathcal{U}$ *is the set of exogenous noise variables,* $\mathcal{V}$ *is a set of causal variables, and* $\mathcal{F} \coloneqq \{f_1, f_2, \ldots, f_m\}$ *denotes the set of causal mechanisms for each causal variable in* $\mathcal{Z}$ *given its corresponding exogenous noise and parents, i.e.,* $\mathbf{v}_i = f_i(\mathbf{u}_i, \mathbf{v}_{\mathrm{Pa}(i)})$.

Given this formulation, we consider two notions when comparing two different causal models: intervention set and counterfactuals.

Definition 4 (Intervention Set)**.** *Given two SCMs* $\mathcal{M}$ *and* $\mathcal{M}'$ *defined on the same set of exogenous noise and causal variables, the intervention set is defined only in terms of their causal mechanisms* $\mathcal{F}$ *and* $\mathcal{F}'$ *respectively:*

$$\mathcal{I}(\mathcal{F}, \mathcal{F}') = \{i : f_i \neq f'_i\}\,. \tag{3}$$

[1]We will use domain and environment interchangeably.

Note that this definition allows multiple types of intervention including soft, hard or do-style interventions. We now define counterfactual pairs as applying two SCMs to the same exogenous noise based on the original definition of counterfactuals in SCMs (Pearl, 2009, Definition 7.1.5).

Definition 5 (Counterfactual Pair)**.** *A pair of causal variable realizations* $(\boldsymbol{z}, \boldsymbol{z}')$ *is a* counterfactual pair *between two SCMs* $\mathcal{M}$ *and* $\mathcal{M}'$ *(with the same set of exogenous noise variables and causal variables) if and only if there exists a exogenous noise realization* $\boldsymbol{u}$ *such that* $\boldsymbol{z}$ *is the solution to* $\mathcal{M}$ *and* $\boldsymbol{z}'$ *is the solution to* $\mathcal{M}'$.

Note that this is different than *estimating* counterfactuals given some factual evidence, which would require the three steps of abduction, action, and prediction. Rather, here we simply define the theoretic notion of a counterfactual pair between two SCMs. However, in practice, we expect that perfect counterfactual pairs will not be feasible so we focus on providing theoretic analysis of noisy or approximate counterfactual pairs.

2.2. Latent Spurious Correlation

After introducing the causal preliminaries, we now formalize the problem latent spurious correlation by formalizing the collection of SCMs that define domains. This is follows many latent SCM multi-domain works (Liu et al., 2022; Zhang et al., 2023b; von Kügelgen et al., 2023; Zhou et al., 2024).

Definition 6 (Class of Latent Domain SCMs)**.** *Letting* $\mathcal{E}$ *denote the set of domains, a latent domain SCM class is a set of latent SCMs* $\mathcal{M}_{\mathcal{E}} = \{\mathcal{M}_e\}_{e \in \mathcal{E}}$ *such that:*

1. *The causal models share the same set of exogenous noise variables, causal variables, and exogenous noise distribution* $\mathbb{P}_{\mathcal{U}}$.
2. *The causal variables* $\mathcal{V}$ *are split into observed* $\mathcal{X} \cup \mathcal{Y}$ *and latent* $\mathcal{Z}$ *variables*
3. *The models share the same causal mechanisms for the observed variables, which we will denote by* $g_{\mathbf{x}}$ *and* $g_{\mathbf{y}}$.

The latent causal mechanisms for the i*-th variable in* $\mathcal{Z}$ *for the* e*-th domain will be denoted as* $f_{e,i}$*, and the induced distribution over the observed random variables for each domain will be denoted by* $\mathbb{P}_e(\mathbf{x}, \mathbf{y})$.

We now give our primary spurious correlation assumption that the domains in the class can only intervene on spurious latent variables with respect to the target variable $\mathbf{y}$, i.e., non-ancestors of $\mathbf{y}$.

Assumption 1 (Spurious Correlation Latent SCM Class)**.** *The intervened variables between any two domains in the*

latent SCM class $\mathcal{M}_{\mathcal{E}}$ must be non-ancestors of $\mathbf{y}$, i.e., $\mathcal{I}(\mathcal{F}_e, \mathcal{F}_{e'}) \cap \mathrm{Anc}(\mathbf{y}) = \emptyset, \forall e, e'$. Equivalently, all domains must share the mechanisms for ancestors of $\mathbf{y}$, i.e., $f_{e,i} = f_{e',i}, \forall i \in \mathrm{Anc}(\mathbf{y})$.

This assumption limits the types of shift that we could see at test time to only spurious features, i.e., non-ancestors of $\mathbf{y}$. However, this assumption does not limit the strength of these shifts, i.e., they could be arbitrarily strong in the spurious features. While this does not cover all possible distribution shifts, it nonetheless captures an interesting collection of shifts that could cause a model to perform poorly on unseen test domains.

2.3. Spurious Counterfactuals are Invariant Pairs

Given our causal model setup in the previous section, we can now prove that counterfactuals within a spurious correlation latent SCM class are invariant pairs w.r.t. the corresponding domain distributions.

Proposition 1. *Given a spurious correlation latent SCM class $\mathcal{M}_{\mathcal{E}}$, any counterfactual pair between models in this class will be an invariant pair w.r.t. the optimally robust classifier $h^*_{\mathcal{E}}$ induced by the domain distributions $\{\mathbb{P}_e\}_{e \in \mathcal{E}}$.*

See proof in Appendix B.1. This elegantly connects causal counterfactuals and invariant pairs though again we note that invariant pairs could be defined for other perspectives. The natural next question is: *Is it possible collect such pairs in reality?* We argue that while perfect counterfactual pairs are not possible, noisy or approximate counterfactual pairs could be reasonably simple to collect in different scenarios. We discuss some of these scenarios next.

Availability of such pairs: For certain applications, obtaining such CF pairs are both possible and effective. Table 1 from the introduction summarizes a range of cases where there could be enough implicit knowledge of spurious correlations to collect them. We outline these levels in more detail below.

Level 3 - Explicit knowledge : In some scientific settings, spurious correlation can be coded as an explicit and mathematical modeling constraint. For example, SchNet (Schütt et al., 2018) builds molecule symmetries and invariance directly into the model structure. This case is straightforward but does not hold in general, so we do not consider it in our work.

Level 2 - Domain expert "soft" knowledge of spurious features: In some applications, domain experts can articulate which features are irrelevant, even if they cannot encode this knowledge as model constraints. For example, an x-ray technician knows that certain medical equipments should not affect their diagnosis of cancer or not (Zech et al., 2018; Oakden-Rayner et al., 2020). In this case, CF pairs can be either manually curated (via image editing or generative models) or collected (e.g., by obtaining paired x-rays with and without fluid lines). Simple image augmentation techniques like rotations, flips or color distortions may also fall under this category as they implicitly encode spurious features that are assumed to not affect the downstream tasks (like ColoredMNIST and RotatedMNIST experiment (cf. Section 6)).

Level 1 - Implicit knowledge: At this level, the only differences between domains are assumed to be spurious features because of application-specific knowledge, but domain experts may not know the spurious features a priori. As one example, the differences between data coming from two similar microscopes can be assumed to be spurious since the measurement effects should not affect the underlying physical phenomena of interest. In this case, it is feasible to collect a small number of counterfactual pairs by measuring a small number of samples with both microscopes. Further, if the implicit knowledge could be verified after obtaining the pairs, it allows post-hoc validation of CF pairs.

Level 0 - No knowledge: Without any hints or assumptions about spurious features as in levels 1-3, making a model robust to spurious features is likely infeasible. To illustrate, consider a simple causal structure without any knowledge on (latent) spurious features: $\mathbf{z}_1 \to \mathbf{y} \to \mathbf{z}_2$ where only $\mathbf{z}_1$ is invariant. Without any knowledge, there is no information to distinguish between invariant feature $\mathbf{z}_1$ and spurious feature $\mathbf{z}_2$. Moreover, if $\mathbf{z}_2$ is more strongly correlated to $\mathbf{y}$ or related to $\mathbf{y}$ that is easier to extract from inputs $\mathbf{x}$, models are prone to shortcut learning (Hermann et al., 2024), the model prediction will rely heavily or nearly solely on $\mathbf{z}_2$.

We specifically target the hard and feasible levels 1 and 2, and suggest that in certain cases counterfactuals could feasibly be collected and in certain cases the counterfactuals could be created either via manual editing or using generative AI tools (Rombach et al., 2022; Betker et al., 2023). Noticing that this CF pairs acquisition are costly, so we ask that if we have k estimated CF pairs with noise ϵ, what kind of robustness guarantee can we get?

3. Noisy Counterfactual Matching (NCM)

Equipped with the above problem setup, we introduce the noisy counterfactual matching method that aims to identify and recover the spurious subspace by leveraging CF pairs. The key intuition is that by construction these CF pairs only differ in a spurious feature subspace and thus provide signal for identifying the spurious subspace. Once the correct spurious subspace is successfully recovered, the prediction can be made robust to this spurious subspace.

Given a set of counterfactual pairs $\{(\boldsymbol{x}_{e_j}^{(j)}, \boldsymbol{x}_{e_j \to e'_j}^{(j)})\}_{j=1}^{k}$,

we define a simple CF matching method that merely adds a constraint to ERM that enforces the outputs of the pairs to match:

$$\min_{\theta} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathbb{P}_{\text{train}}}[\ell(h_\theta(\mathbf{x}),\mathbf{y})], \quad \text{s.t. } h_\theta(\boldsymbol{x}_{e_j}) = h_\theta(\boldsymbol{x}_{e_j\to e'_j}) \ \ \forall j, \tag{4}$$

where the loss $\ell(\cdot,\cdot)$ measures data fidelity by prediction function h_θ. If h_θ is linear, the constraint could simplify to $\theta^\top \delta_j = 0, \forall j$, where $\delta_j := \boldsymbol{x}_{e_j}^{(j)} - \boldsymbol{x}_{e_j\to e'_j}^{(j)}$. Intuitively, this constraint forces the classifier θ to be orthogonal to the CF differences, whose subspace only spans spurious features. The objectives in MatchDG (Mahajan et al., 2021) and DIRT (Nguyen et al., 2021) consider Lagrangian formulations of this problem. However, they lack theoretic analysis and do not explicitly identify the necessary properties that the data pairs must satisfy to achieve robustness. In particular, they do not consider how to be robust when the data pairs are non-ideal (e.g., approximate or noisy)—which is precisely the practical scenario.

To address non-ideal pairs for linear models, we propose a slightly modified objective based on finding a lower-dimensional subspace spanned by counterfactual pair differences. Concretely, let us define a (potentially noisy) counterfactual pair difference matrix as:

$$\tilde{\Delta}_x \coloneqq \left[\boldsymbol{x}_{e_1}^{(1)} - \boldsymbol{x}_{e'_1}^{(1)}, \ldots, \boldsymbol{x}_{e_k}^{(k)} - \boldsymbol{x}_{e'_k}^{(k)}\right] \in \mathbb{R}^{d\times k}.$$

Because the pairs are noisy, making the linear model orthogonal to the whole matrix $\tilde{\Delta}_x$ may remove many non-spurious features and degrade performance. Thus, we propose to make the linear classifier only orthogonal to the best r-dimensional subspace of $\tilde{\Delta}_x$ and propose the Noisy Counterfactual Matching problem for linear models:

$$\min_{\theta} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathbb{P}_{\text{train}}}[\ell(\theta;\mathbf{x},\mathbf{y})] \quad \text{s.t. } \theta^\top \tilde{U}^r = 0, \tag{5}$$

where $\tilde{U}^r \in \mathbb{R}^{d\times r}$ denotes the space of left singular vectors corresponding to the r-truncated SVD of $\tilde{\Delta}_x$. Notice that the constraint enforces the classifier θ to be orthogonal to the subspace $\tilde{U}^r$. With perfect counterfactuals, $\tilde{U}^r$ would correspond to only the spurious subspace and the classifier would be robust to changes in the spurious subspace. Because of noise, a much more careful analysis is required (which we do in the next section) to show that this approach improves robustness based on the diversity and quality of the pairs.

Algorithmically, to ensure the learned model orthogonal to be the spurious feature subspace, we consider preprocessing approach that projects the samples $\mathbf{x}$ onto the orthogonal complement of $\tilde{U}^r$ and then trains an unconstrained classifier on top (See Algorithm 1). This ensures that the effective classifier, i.e., the composition of the preprocessing and unconstrained classifier, is orthogonal to the subspace defined by $\tilde{U}^r$. Unlike projected gradient descent, this enables simple optimization without any special projection step every iteration. A projected gradient descent method could also be used here, and we expect it would have similar results, but we do not explore it further.

Algorithm 1 Noisy Counterfactual-Matching (NCM)

Input: Training data set $\mathcal{D}_{\text{train}}$; pair difference matrix $\tilde{\Delta}_x \in \mathbb{R}^{d\times k}$; truncated SVD dimension r; number of epochs T; step size η; batch size B.

// Phase I: Find projection matrix to remove estimated spurious subspace $\tilde{U}^r$.

$\tilde{U}_r, \tilde{\Sigma}_r, \tilde{V}_r^\top = \text{TruncatedSVD}(\tilde{\Delta}_x, r)$

$P = I - \tilde{U}_r \tilde{U}_r^\top$

// Phase II: Gradient descent with preprocessing.

for $t = 1, 2, \ldots, T$ **do**

  for sample mini-batch $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^B \subset \mathcal{D}_{\text{train}}$ **do**

    $\theta \leftarrow \theta - \eta \nabla \frac{1}{B} \sum_{i=1}^B \ell(h_\theta(P\boldsymbol{x}_i), \boldsymbol{y}_i)$,

  end for

end for

Output θ

4. Theoretic Guarantees of NCM for Linear Models

In this section, we provide theoretic guarantees of NCM (5) for both linear regression and logistic regression. Our study builds on the following steps. 1) We first show that the test domain error could be decomposed into the in-domain error and the misalignment of the estimated spurious subspace $\tilde{U}^r$ and the relevant test domain spurious subspace (cf. Theorem 1 and Theorem 3). 2) We then apply Wedin's $\sin\Theta$ theorem (Wedin, 1972) to decompose the spurious subspace misalignment into two components: (a) the discrepancy between the truncated singular subspaces of the noisy counterfactual pair matrix $\tilde{\Delta}_{\boldsymbol{x}}$ and the true matrix $\Delta_{\boldsymbol{x}}$, where the truncation is determined by the user-specified rank r; and (b) the model misspecification error arising from the unknown dimensionality of the spurious subspace $|\tilde{S}|$ (cf. Corollary 2 and Corollary 4).

4.1. Error Bounds of NCM for Linear Regression

We will assume the data generating process for linear regression with latent causal variables is as follows: $\mathbf{z}_e = A_e^* \mathbf{u}$, $\mathbf{x}_e = B^* \mathbf{z}_e$, $\mathbf{y} = A^* \mathbf{z}_e$, where $\mathbf{u} \sim \mathbb{P}_{\mathbf{u}}, \mathbf{z}_e \in \mathbb{R}^m$, $A_e \in \mathbb{R}^{m\times m}$ and $\mathbf{x}_e \in \mathbb{R}^d$, $B^* \in \mathbb{R}^{d\times m}, \theta^* \in \mathbb{R}^m$.

To quantify the deviation of the test domain $e^+ \in \mathcal{E}_{\text{test}}$ from the training domain $\mathcal{E}_{\text{train}}$, we introduce the second moment matrix $m_{e^+,e}$ of the difference between the test domain

e^+ and its induced training domain e which $(x_{e^+}, x_{e^+\to e})$ forms an oracle CF pair. $x_{e^+\to e}$ is only conceptual and not observed in the training data. Further, these conceptual samples are also realizations of the training distributions because their exogenous noise follows distribution $\mathbb{P}_{\mathcal{U}}$. (cf. Definition 6) We define its largest eigenvalue of the population second moment matrix $\lambda_1(e^+)$ as the largest eigenvalue of the following population matrix

$$\mathbb{E}_u[(x_{e^+} - x_{e^+\to e})(x_{e^+} - x_{e^+\to e})^\top]. \tag{6}$$

We adopt the conventional definition of the distance between any two subspaces U_1, U_2 of the same size $\mathbb{R}^{d\times j}, j \in [d]$ as (Chen et al., 2021) $\text{dist}(U_1, U_2) := \|U_1U_1^\top - U_2U_2^\top\|$. We have the following guarantee on the spurious correlation using NCM (5) for linear regression.

Theorem 1. *Suppose Assumption 1 and the number of estimated counterfactual pairs satisfies $k \geq \min\{r, |\mathcal{S}|\}$. Then, for any solution $\hat{\theta}$ of NCM (5) with mean square loss, for any test domain satisfying $\mathcal{M}_{e^+} \in \mathcal{M}_{\mathcal{E}}$, there holds*

$$\mathbb{E}_{p(\boldsymbol{x}_{e^+}, y_{e^+})}[\|\hat{\theta}^\top \boldsymbol{x}_{e^+} - y_{e^+}\|^2] \leq \underbrace{2\mathbb{E}_{p(e)p(\boldsymbol{x}_e,y)}[\|\hat{\theta}^\top \boldsymbol{x}_e - y\|^2]}_{\text{In-domain error}} + \underbrace{2\lambda_1(e^+)\|\hat{\theta}\|^2\|\tilde{U}_{r,\perp}^\top U_{\mathcal{S}}\|^2}_{\text{Spurious subspace misalignment}},$$

where $\lambda_1(e^+)$ is defined in (6), and r is the user defined dimension for the truncated SVD, and $|\mathcal{S}|$ is the dimension for the spurious space.

See Appendix B.3 for proofs. Observe that the generalization error can be controlled by the in-domain error and spurious subspace misalignment. It sheds light on the impact of the problem parameters on the spurious correlation. Specifically, the following comments are in order.

(i) Few shot counterfactual pairs requirement: Notice that the number of counterfactual pairs only need to satisfy $k \geq \min\{r, |\mathcal{S}|\}$. It is easily satisfied as the spurious feature space is in general low dimensional due to the sparse mechanism shift hypothesis (Schölkopf et al., 2021).

(ii) Difficulty of the test domain: $\lambda_1(e^+)$ denotes the difficulty of the test domain. As $\lambda_1(e^+)$ increases, the test domain becomes more distant from the training domain. A special case arises when e^+ chosen to be sampled from mixture of training domains. In this case, the spurious subspace misalignment vanishes as the test domain is already seen, thus $\boldsymbol{x}_{e^+} = \boldsymbol{x}_{e^+\to e}$ i.e., $\lambda_1(e^+) = 0$, NCM (5) reduces to empirical risk minimization (ERM).

(iii) Accuracy trade-off induced by r: Observe that the constraint $\theta^\top \tilde{U}_r = 0$ in the NCM objective (5) implies that the classifier θ must lie within $\tilde{U}_{r,\perp}$. Therefore, choosing a smaller r increases the feasible region $U_{r,\perp}$, allowing for a smaller in-domain error. An extreme case is $r = 0$, in which case (5) reduces to minimizing only the in-domain error, i.e., ERM. However, a smaller r leads to a greater spurious subspace misalignment, as $\tilde{U}_r \in \mathbb{R}^{d\times r}$ fails to accurately recover the true spurious subspace which is of dimension $\mathbb{R}^{d\times|\tilde{S}|}$. Figure 4 presents empirical results demonstrating how model performance varies with different values of r, supporting our theoretical analysis. To achieve optimal performance in practice, it is important to carefully select r to balance those error terms. This can be accomplished by tuning r on a held-out validation domain.

It becomes critical to quantify the second part: spurious subspace misalignment. Notice that $\tilde{U}_r \in \mathbb{R}^{d\times r}$ and $U_{|\mathcal{S}|} \in \mathbb{R}^{d\times|\mathcal{S}|}$ are of different sizes, thus $\text{dist}\left(\tilde{U}_r, U_{|\mathcal{S}|}\right)$ is not defined. However, it is straightforward to verify that

$$\|\tilde{U}_{r,\perp}^\top U_{\mathcal{S}}\| \leq \text{dist}\left(\tilde{U}_{\min\{r,|\mathcal{S}|\}}, U_{\min\{r,|\mathcal{S}|\}}\right) + \mathbf{1}_{\{r\neq|\mathcal{S}|\}}.$$

Observe that **1)** the first term corresponds to the misalignment between the $\min\{r, |\mathcal{S}|\}$ truncated singular subspace of the noisy counterfactual pair matrix $\tilde{\Delta}_{\boldsymbol{x}}$, and that of the true counterfactual pair matrix $\Delta_{\boldsymbol{x}}$. That is, the distance between $\tilde{U}_{\min\{r,|\mathcal{S}|\}}$ and $U_{\min\{r,|\mathcal{S}|\}}$. The misalignment can further be characterized by the Wedin's $\sin\Theta$ theorem (Wedin, 1972), which measures how perturbations in a matrix, $\Delta_{\boldsymbol{x}} \to \tilde{\Delta}_{\boldsymbol{x}}$ influence the alignment of its singular subspaces, $U_{\min\{r,|\mathcal{S}|\}} \to \tilde{U}_{\min\{r,|\mathcal{S}|\}}$. **2)** The second term relates to the model misspecification error due to the unknown size of the spurious subspace $|\tilde{S}|$. This term vanishes only when we have prior knowledge of $|\tilde{S}|$ and set r, the size of the truncated SVD, to match it. In general, $|\mathcal{S}|$ is unknown; thus, one may overestimate or underestimate it using r. Our theory addresses both cases and quantifies the impact of such model misspecification in the bound. To do so, we define $\sigma_1 \geq \cdots \geq \sigma_k$ as the singular values of the perfect counterfactual matrix $\Delta_{\boldsymbol{x}}$ in a descending order, and define the singular value gap of $\Delta_{\boldsymbol{x}}$ as follows.

$$\rho := \sigma_{\min\{r,|\mathcal{S}|\}} - \sigma_{\min\{r,|\mathcal{S}|\}+1}. \tag{7}$$

We have the following guarantee.

Corollary 2. *Instate the setting in Theorem 1. Further, suppose the noise $\Delta_{\boldsymbol{x}} - \tilde{\Delta}_{\boldsymbol{x}}$ satisfies*

$$\sigma_{\max}(\Delta_{\boldsymbol{x}} - \tilde{\Delta}_{\boldsymbol{x}}) \leq (1 - 1/\sqrt{2})\rho. \tag{8}$$

Then, for any solution $\hat{\theta}$ of NCM (5) with mean square loss, for any test domain satisfying $\mathcal{M}_{e^+} \in \mathcal{M}_{\mathcal{E}}$, there holds

$$\mathbb{E}_{p(\boldsymbol{x}_e^+, y)}[\|\hat{\theta}^\top \boldsymbol{x}_e^+ - y\|^2] \leq 2\mathbb{E}_{p(e)p(\boldsymbol{x}_e,y)}[\|\hat{\theta}^\top \boldsymbol{x}_e - y\|^2] + 8\lambda_1(e^+)\|\hat{\theta}\|^2\frac{\|\Delta_{\boldsymbol{x}} - \tilde{\Delta}_{\boldsymbol{x}}\|^2}{\rho^2} + 4\lambda_1(e^+)\|\hat{\theta}\|^2\mathbf{1}_{\{r\neq|\mathcal{S}|\}},$$

where $\lambda_1(e^+)$ and ρ are defined in (6) and (7).

If the pairs are oracle (a.k.a. noiseless), without explicit choosing r, $\tilde{\Delta}_x = \Delta_x$ is a rank $|\mathcal{S}|$ matrix, equivalent to choosing $r = |\mathcal{S}|$. In this case, both second term and third term vanish, and the test domain error converges to 0 as the training risk converges to 0.

Remark 1 (Oracle counterfactual matching)**.** *When the estimated counterfactual is error-free, i.e., $\tilde{\Delta}_{\boldsymbol{x}} = \Delta_{\boldsymbol{x}}$. Suppose Assumption 1 and the number of counterfactual pairs in $\Delta_{\boldsymbol{x}}$ satisfy $k \geq |\mathcal{S}|$. Then, for any solution $\hat{\theta}$ of NCM (4) with mean square loss, the following holds:*

$$\mathbb{E}_{p(\boldsymbol{x}_e^+, y)}[\|\hat{\theta}^\top \boldsymbol{x}_e^+ - y\|^2] \leq 2\mathbb{E}_{p(e)p(\boldsymbol{x}_e, y)}[\|\hat{\theta}^\top \boldsymbol{x}_e - y\|^2]$$

4.2. Error bounds of logistic regression NCM (5)

In this section, we further study the guarantee of NCM (5) for logistic regression. We focus on the analysis on linear observation function in this work, where the data generating process is specified as follows: $\mathbf{z}_e = A_e^* \mathbf{u}$, $\mathbf{x}_e = B^* \mathbf{z}_e$, $\mathbf{y} = \text{sign}(A^* \mathbf{z}_e)$, where $\mathbf{u} \sim \mathbb{P}_{\mathbf{u}}, \mathbf{z}_e \in \mathbb{R}^m$, $A_e \in \mathbb{R}^{m \times m}$ and $\mathbf{x}_e \in \mathbb{R}^d$, $B^* \in \mathbb{R}^{d \times m}, \theta^* \in \mathbb{R}^m$. Here ℓ takes logistic loss $\ell(\mathbf{x}) = \log(1 + \exp(-\mathbf{y}\theta^\top \mathbf{x}))$.

To quantify the deviation of the test domain from the training domain, for all $i \in [k]$, we introduce k second-moment random matrices of the difference between the test domain e_i^+ and its counterfactual in the training domain $\boldsymbol{x}_{e_i^+ \to e_i}$ (similar to (6), which may not be observed in the training set, but shares the same distribution), denoted as $m_{e_i^+, e_i}$ and define the empirical largest eigenvalue as $\lambda_1(e_i^+)$.

$$\begin{aligned} m_i &\coloneqq (x_{e_i^+} - x_{e_i^+ \to e_i})(x_{e_i^+} - x_{e_i^+ \to e_i})^\top \in \mathbb{R}^{d \times d} \\ \lambda_1(e_i^+) &\coloneqq \lambda_1(m_{e_i^+, e_i}) \in \mathbb{R}_+. \end{aligned} \tag{9}$$

We have the following guarantee on the spurious correlation using NCM for logistic regression.

Theorem 3. *Suppose Assumption 1 and the number of estimated counterfactual pairs satisfies $k \geq \min\{r, |\mathcal{S}|\}$. Then, for any solution $\hat{\theta}$ of NCM (5) with negative log likelihood, there holds*

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{x}_{e^+}, y_{e^+}} \left[\log\left(1 + \exp(-y_{e^+} \theta^\top \boldsymbol{x}_{e^+})\right)\right] \\ &\leq \mathbb{E}_{p(e)p(\boldsymbol{x}_e, y)} \log \left\{ 1 + \frac{1}{k} \sum_{i=1}^k \exp\left(-y_{e_i} \theta^\top \boldsymbol{x}_{e_i}\right) \right. \\ &\quad \left. + \frac{1}{k} \sum_{i=1}^k \exp\left(\sqrt{\lambda_1(e_i^+)} \|\theta\| \|\tilde{U}_{r,\perp}^\top U_{\mathcal{S}}\|\right) - 1 \right\} \\ &\quad + \sqrt{\mathbb{E}_u[\lambda_1(e^+)]} \|\theta\| \|\tilde{U}_{r,\perp}^\top U_{\mathcal{S}}\|. \end{aligned} \tag{10}$$

Notice that, similar to the linear regression case in Theorem 1, (10) implies that the test domain error in the test domain consists of two main components: the first term relates to the in domain error, while the second and third terms capture the misalignment of the spurious subspace due to noisy counterfactuals. Further, the number of counterfactual pairs only has to satisfy $k \geq \min\{r, |\mathcal{S}|\}$, which is the same condition as in Theorem 1. There' s also accuracy trade-offs induced by the truncated SVD choice r. Notice that, unlike the linear regression case Theorem 1, where the hardness of the test domain is characterized by the largest eigenvalue of the population matrix $\mathbb{E}_u[m_{e,e^+}]$ as defined in (6), the hardness of the test domain is now bounded by the exponential moment of the largest eigenvalue of the sample covariance matrix m_i defined in (9). This distinction arises from the difference in the loss structures between the ℓ_2 and logistic losses and do not scale significantly differently.

We refer to Appendix B.6 for the results using Wedin's $\sin\Theta$ theorem to bound the spurious misalignment of the subspaces (cf. Corollary 4), and we also provide guarantees when perfect counterfactual pairs are available (cf. Remark 2).

To recover the correct invariant subspace, the difference in invariant features must be smaller than the difference in spurious features. However, random pairing does not ensure this condition.

Our theoretical results explain why CF pairs can enhance robustness. The differences between these pairs enable effective separation of invariant features from spurious ones when there is a substantial gap between the largest eigenvalue of the invariant features and the smallest eigenvalue of the spurious features. Both random pairs from same target and different domain and the pairs from nearest neighbor matching do not satisfy the property of CF pairs, which leads to poor performance (cf. Section 6).

5. Related Works

In this section, we discuss some related approaches.

Data augmentation and generation: The data augmentation approach is closely related to our proposed method, particularly in the context of domain generalization. LISA (Yao et al., 2024) is a Mixup-inspired augmentation strategy that learns domain-invariant predictors through two types of interpolation: intra-label mixing, which combines samples sharing the same class label but originating from different domains to enforce prediction consistency across domains; and intra-domain mixing, which combines samples from the same domain but with different class labels to encourage the model to respect class boundaries within each domain. Other methods, such as those proposed by Honarvar Nazari and Kovashka (2020); Shorten and Khoshgoftaar (2019), use standard augmentation functions including rotation, scaling, and noise addition. An-

other line of research seeks to inflate the training data to improve domain generalization. For example, Rahman et al. (2019) use ComboGAN (Anoosheh et al., 2018) to generate new data, while DIRT (Nguyen et al., 2021) suggests using StarGAN (Choi et al., 2018) to generate counterfactual samples.

Distribution or Sample Matching in Addressing Spurious Features: Distribution matching methods, such as Invariant Risk Minimization (IRM) (Arjovsky et al., 2019) and Risk Extrapolation (REx) (Krueger et al., 2021), aim to mitigate spurious correlations by learning domain-invariant representations. Despite their theoretical appeal, IRM-based approaches often underperform in practice, prompting several works to analyze and refine them (Rosenfeld et al., 2020; Ahuja et al., 2022).

Beyond distribution matching, MatchDG (Mahajan et al., 2021) introduces an iterative sample-level matching objective that aligns representations across domains in latent space. Our method similarly employs sample-wise matching but departs from prior group-level matching techniques. Crucially, we provide a theoretical guarantee that, for domain generalization, only $O(r)$ samples suffice—where r denotes the dimension of spurious features in the latent structural causal model (SCM)—in contrast to the $n_d \to \infty$ requirement in earlier methods.

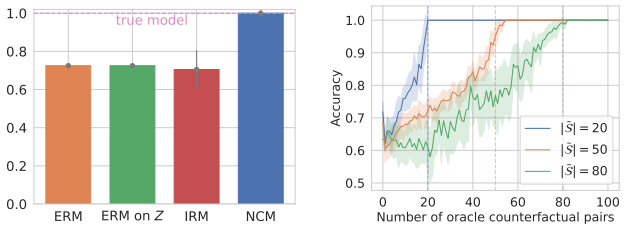
Multi-view or multi-modal data: A more recent line of research integrates the text modality into visual domain generalization tasks, as demonstrated by models like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021). These methods focus on aligning text and image modalities during training, leveraging large-scale datasets. This alignment enhances the model's ability to generalize to out-of-distribution samples, improving performance on unseen data beyond the training set.

6. Empirical Evaluation

In this section, we provide some experiments on synthetic and real-world datasets. **(i)** The synthetic data is instrumental in validating our theoretical findings. Specifically, we validate the theory of robustness performance using oracle CF pairs (Corollary 4). We also validate the linear dependency between the number of counterfactual pairs and the spurious feature dimensions. **(ii)** Beyond the synthetic data, we thoroughly evaluate NCM (5) by a linear probing of a preprocessed CLIP model (Radford et al., 2021) on multiple real-world datasets. While the CLIP model already demonstrates superior zero-shot transfer capabilities (Radford et al., 2021) , our NCM (5) further enhances robustness against spurious correlation. **(iii)** Moreover, we investigate three types of matching methods: random matching, one nearest neighbor matching (Mahajan et al., 2021), and noisy CF matching, and demonstrate the superiority of CF matching over learned matching approaches. In all experiments, we include in-domain validation and oracle validation for evaluation (Gulrajani and Lopez-Paz, 2021). See detailed validation mechanism as well as the hyperparameter tuning process in Appendix C.2. **(iv)** Lastly, we perform several ablation studies to explore how the performance of NCM (5) changes when key assumptions about the data or model are relaxed. Specifically, we examine the impact of varying the truncated SVD parameter r, and testing scenarios with insufficient counterfactual data. We also assess the model's performance in the context of deep neural network pretraining, analyzing its robustness and generalization across these different settings.

6.1. Synthetic Experiments

We model both the structural causal models (SCMs) and the observation function as linear functions adapting the data generation approach from Rosenfeld et al. (2020).



(a) A comparison of ERM, ERM on $\mathcal{Z}$, IRM, and NCM shows that NCM achieves oracle accuracy, while IRM performs similarly to ERM but with higher variance. ERM on $\mathcal{Z}$ means model access the true observation function.

(b) Accuracy vs. number of oracle counterfactual pairs for three different intervention set sizes. The solid lines represent the mean accuracy over 5 repeated runs, and the shaded regions indicate the standard deviation.

Figure 3. Result on the synthetic dataset with oracle counterfactual pairs with $m = 100$. In two training domains, $\sigma_{p,1} = 0.1, \sigma_{p,2} = 0.2$ and in the test domain it is $\sigma_{p,3} = 10$. The total number of training samples are 200 with 100 for each domain.

Data generation: Invariant features are sampled from a standard normal distribution, i.e., $\mathbf{z}_{\text{inv}} \sim \mathcal{N}(0, I)$, The observation function g_y is linear, with parameter $\theta_y \sim \mathcal{N}(0, \sigma I)$, and the label $\boldsymbol{y} = \text{sign}(\mathbf{z}_{\text{inv}}\theta_y)$. The spurious features is correlated to the label $\boldsymbol{y}$, i.e., $\mathbf{z}_{\text{spu}} \sim \mathcal{N}\left(\frac{\boldsymbol{y}}{|\mathcal{S}|}, \sigma_s I\right)$ where σ_s varies across domains. The observation function g_x is a random orthonormal matrix. The dimension of $\mathbf{z}$ and $\mathbf{x}$ are both 100, i.e., $m = d = 100$.

We compare four methods on a binary logistic regression within this framework: ERM, ERM with known observation function g_x (ERM on $\mathcal{Z}$), IRM, and NCM . **(i)** Effectiveness of NCM (5): the result shows that NCM (5)

Table 2. Main Results on ColoredMNIST

	in-domain validation		oracle validation	
	in acc	test acc	in acc	test acc
ERM (CLIP)	0.852	0.093	0.753	0.253
IRM	0.799	0.118	0.724	0.469
REx	0.797	0.121	0.691	0.664
GroupDRO	0.798	0.127	0.786	0.201
Fish	0.798	0.118	0.495	0.486
SWAD	0.800	0.113	0.501	0.505
LISA	0.705	**0.693**	0.705	0.693
MatchDG w. random	0.799	0.120	0.511	0.512
MatchDG w. 1NN	0.789	0.217	0.728	0.662
MatchDG w. clean	0.793	0.181	0.742	0.672
NCM w. random	0.794	0.176	0.680	0.706
NCM w. 1NN	0.736	0.649	0.711	0.707
NCM w. clean	0.740	**0.693**	0.727	**0.714**
random guess	0.500	0.500	0.500	0.500
ERM oracle	0.735	0.730	0.735	0.730
theory oracle	0.750	0.750	0.750	0.750

achieves 100% accuracy as if we have data from the test domain while other methods achieves around 70% test accuracy because these methods rely on the spurious features (cf. Figure 3a). **(ii)** Linear number of counterfactual pairs: Figure 3b shows that for linear casual model, we only need approximately $|\mathcal{S}|$ counterfactual pairs to find spurious space correctly . When the number of counterfactual pairs matches or exceeds the dimension of intervened set, i.e, $k \geq |\mathcal{S}|$, NCM (5) achieves optimal test domain accuracy, therefore validating few shot of counterfactual requirement (cf. Remark 1).

6.2. ColoredMNIST

ColoredMNIST is first introduced in DomainNet (Arjovsky et al., 2019). Although its imagery is deceptively simple, it is indeed one of the hardest domain generalization benchmark as shown in Salaudeen et al. (2024) due to strong accuracy on the inverse line effect. Indeed, poor performance on ColoredMNIST has been widely reported for baseline methods including ERM, IRM, DRO, Mixup, MLDG, CORAL, MMD, ADA, and CondADA (Gulrajani and Lopez-Paz, 2021, Section B.1).

We report the accuracy of results on ColoredMNIST (Table 2) on a linear-probing pretrained ViT-B/32 CLIP model (Radford et al., 2021). NCM (5) performs well on both in-domain and oracle validation, achieving test domain accuracies of 69.3% and 71.4%, respectively, nearly matching the ERM oracle accuracy of 73%, demonstrating the effectiveness of NCM (5). The performance difference between two validation methods are only 2%, indicating that NCM (5) is less sensitive to hyperparameter tuning. This stands in sharp contrast to other algorithms such as ERM, IRM, GroupDRO, Fish, and REx, which only achieve around 10% accuracy with in-domain validation and 20%-66% with oracle validation except for LISA which achieves 69.3% on both validation methods.

We further observe that both random pairing and 1NN pairing are effective on ColoredMNIST. We suspect this is due to the inherent similarity of invariant features across samples. As a result, any random pair sharing the same target label y but from different domains d is a reasonable approximation of CF pairs.

6.3. PACS

PACS (Li et al., 2017) contains total 9991 images from four domains: Photos (P), Art painting (A), Cartoon (C) and Sketch (S). The task is to classify objects over 7 classes. We report NCM as well as the baselines on the pretrained-CLIP model in Table 4. From the result, **(i)** notice that ERM linear-probing CLIP outperforms all previous methods shown in the previous benchmark (Gulrajani and Lopez-Paz, 2021). **(ii)** Our method NCM (5) could further improve test domain accuracy. Using in-domain validation, we achieves 95.3% outperforms ERM by 1.6%. Similarly, with oracle validation, we also achieves 1.6% performance boost against ERM and other methods. **(iii)** The quality of the pairing matters on this dataset. Unlike ColoredMNIST, random pairing samples with same labels across different domains dramatically affects the model's

Table 3. Main Results on Waterbirds-CF

	In-domain Validation		Oracle Validation	
	in acc	wg acc	in acc	wg acc
ERM (CLIP)	0.885	0.781	0.882	0.800
ERM+UW	0.889	0.795	0.882	0.829
IRM	0.838	0.707	0.820	0.767
REx	0.891	0.617	0.878	0.729
GroupDRO	0.906	0.684	0.896	0.827
Fish	0.900	0.744	0.869	0.805
LISA	0.904	0.722	0.876	0.812
MatchDG w. random	0.793	0.009	0.785	0.149
MatchDG w. 1NN	0.886	0.411	0.886	0.411
MatchDG w. estimated CF	0.906	0.536	0.896	0.651
NCM w. random	0.804	0.269	0.804	0.269
NCM w. 1NN	0.892	0.521	0.882	0.560
NCM w. estimated CF	0.864	**0.812**	0.854	**0.860**

performance. We suspect that random pairing on PACS misleads the model due to significant differences in the invariant features across samples. When invariant features vary greatly, they can be conflated with spurious features during truncated SVD decomposition. As a result, the model may utilize spurious features with smaller variance while neglecting invariant features that exhibit larger differences.

6.4. Waterbirds-cf

We consider Waterbirds-cf in this section, similar to the Waterbirds (Sagawa et al., 2019) in the context of spurious correlation. In the original Waterbirds dataset, the background (water or land) is highly correlated with the bird species (waterbirds or landbirds) in the training set, but this correlation changes in the test set. The dataset contains "landbirds on land", "waterbirds on water", "landbirds on water" and "waterbird on land" as training domains with $3498, 1057, 184, 56$ training samples respectively. In this 4795 training samples, background is highly correlated to the objects, and only 240 out of 4795 samples are from the minority groups ("landbird on water", "waterbird on land"). The background here is the spurious features. To evaluate our NCM (5), we modified all the 240 minor group samples in the original dataset to estimated CF samples corresponding to 240 random samples from majority groups , estimated by samples in the majority groups. (See Appendix C.1 for details of data construction.)

(i) The results in Table 3 show that our method with estimated CF pairs achieves 86.0% accuracy, significantly outperforming ERM (raw CLIP) and other baseline methods with CLIP on oracle validation. It also achieves 81.2% accuracy using validation from the in-domain test set, continually outperforming CLIP, the best among all baselines, by 3.1% **(ii)** Further, note that the quality of the pairs significantly affects the model's performance on this dataset. When using random matching or 1-nearest-neighbor matching, the model's performance drops significantly compared to the CLIP baseline, showing the pairs are damaging instead of helping the model performance. . Our theory suggests that the model can only find the correct invariant feature space when the spurious feature difference gap is larger that of the invariant features. In the Waterbirds dataset, which contains versatile invariant features, random pairing or nearest-neighbor pairing results in a large difference in invariant features. This contrasts with ColoredMNIST, where the invariant features are inherently similar, leading to similar performance across different pairing methods.

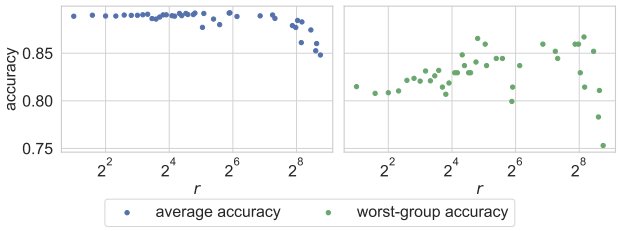


Figure 4. In-domain test and worst-group accuracy with changing hyperparameter r. on CLIP pretrained model. In-domain accuracy remains stable for small values of r, but drops when $r \geq 128$ approximately. In contrast, worst-group accuracy initially increases and then decreases as r grows.

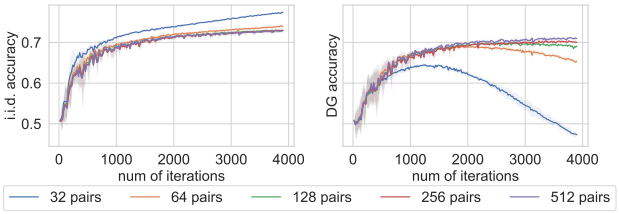


Figure 5. The number of counterfactuals vs. DG accuracy on ColoredMNIST using the CLIP + Linear model.

6.5. Ablation Study

Sensitivity on truncated SVD parameter r. We empirically evaluate the trade-off effect of the hyperparameter r on model performance during linear probing on the Waterbirds-CF dataset (cf. Figure 4), thus validating Theorem 3 comment **(iii)**: accuracy trade-off induced by r. This pattern reflects the model's shifting reliance from spurious to invariant features: when r is too small, spurious correlations dominate, resulting in high in-domain but low worst-group performance. As r increases and suppresses these spurious features, worst-group accuracy improves. However, beyond a certain point, further increases in r begin to remove invariant features as well, leading to a decline in both metrics.

Sensitivity on the number of CF Pairs. We evaluate the number of counterfactual pairs needed on ColoredMNIST dataset. The results show that with 32 counterfactual pairs, the number of pairs is insufficient for the model to eliminate spurious features, leading to spurious correlation (as indicated by an in-domain accuracy over 75%, meaning the classification relies on spurious features). However, when using 128 or 256 counterfactual pairs, the performance increases significantly and remains stable compared to the 32 counterfactual pairs.

Table 4. Main Results on PACS

	In-domain Validation					Oracle Validation				
	A	C	P	S	Avg	A	C	P	S	Avg
ERM (CLIP)	0.924	0.968	0.996	0.859	0.937	0.924	0.968	0.996	0.859	0.937
IRM	0.938	0.976	0.996	0.840	0.938	0.941	0.976	0.996	0.845	0.940
REx	0.953	0.963	0.993	0.836	0.936	0.953	0.975	0.996	0.845	0.942
GroupDRO	0.903	0.963	0.996	0.873	0.934	0.941	0.975	0.996	0.843	0.939
Fish	0.936	0.973	0.996	0.837	0.936	0.936	0.973	0.996	0.837	0.936
SWAD	0.941	0.976	0.996	0.838	0.938	0.941	0.977	0.996	0.838	0.938
LISA	0.926	**0.978**	0.997	0.848	0.937	0.940	**0.983**	0.997	0.864	0.946
MatchDG w. rand.	0.412	0.509	0.316	0.749	0.497	0.454	0.509	0.358	0.749	0.518
MatchDG w. 1NN.	**0.964**	0.971	0.995	0.880	**0.953**	**0.964**	0.973	0.996	**0.887**	**0.955**
NCM w. rand.	0.591	0.609	0.577	0.833	0.653	0.592	0.625	0.583	0.843	0.661
NCM w. 1NN.	0.957	0.974	**0.998**	**0.882**	**0.953**	**0.964**	0.974	**0.998**	0.885	**0.955**

7. Conclusion and Discussion

In this work, we tackle spurious correlation from a data-centric perspective and show that introducing (noisy) counterfactual pairs during training can enhance model robustness. This aligns with long-standing machine learning practices, where supervised learning uses labels to encode target concepts without formal definitions to help the model focus on the important features. Similarly, counterfactuals help capture spurious correlations implicitly without formal definition to help the model avoid spurious features.

One challenge of our method is obtaining counterfactual pairs. While straightforward in tasks like object classification (e.g., using image editing for spurious features, as shown in the Waterbirds dataset), it is more complex in fields like medical imaging, requiring expert involvement. However, experts can now help by creating or validating a few high-quality counterfactuals to improve robustness suggested by our findings.

References

Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization, 2022.

Isabela Albuquerque, João Monteiro, Mohammad Darvishi, Tiago H Falk, and Ioannis Mitliagkas. Generalizing to unseen domains via distribution matching. *arXiv preprint arXiv:1911.00804*, 2019.

Asha Anoosheh, Eirikur Agustsson, Radu Timofte, and Luc Van Gool. Combogan: Unrestrained scalability for image domain translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 783–790, 2018.

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

Fatemeh Azimi, Sebastian Palacio, Federico Raue, Jörn Hees, Luca Bertinetto, and Andreas Dengel. Self-supervised test-time adaptation on video data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3439–3448, 2022.

Ruqi Bai, Saurabh Bagchi, and David I. Inouye. Benchmarking algorithms for federated domain generalization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=wprSv7ichW.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *Journal of machine learning research*, 22(2):1–55, 2021.

Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, et al. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14 (5):566–806, 2021.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

Hal Daumé III. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

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

Katherine Hermann, Hossein Mobahi, Thomas FEL, and Michael Curtis Mozer. On the foundations of shortcut learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Tj3xLVuE9f.

Narges Honarvar Nazari and Adriana Kovashka. Domain generalization using shape representation. In *European Conference on Computer Vision*, pages 666–670. Springer, 2020.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

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

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

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

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 624–639, 2018b.

Yuhang Liu, Zhen Zhang, Dong Gong, Mingming Gong, Biwei Huang, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Identifying weight-variant latent causal models. *arXiv preprint arXiv:2208.14153*, 2022.

Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International Conference on Machine Learning*, pages 7313–7324. PMLR, 2021.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

A Tuan Nguyen, Toan Tran, Yarin Gal, and Atilim Gunes Baydin. Domain invariant representation learning with domain density transformations. *Advances in Neural Information Processing Systems*, 34:5264–5275, 2021.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pages 151–159, 2020.

J Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78 (5):947–1012, 2016.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Mohammad Mahfujur Rahman, Clinton Fookes, Mahsa Baktashmotlagh, and Sridha Sridharan. Multi-component image translation for deep domain generalization. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 579–588. IEEE, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Olawale Elijah Salaudeen, Nicole Chiou, and Sanmi Koyejo. On domain generalization datasets as proxy benchmarks for causal representation learning. In *NeurIPS 2024 Causal Representation Learning Workshop*, 2024.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24), 2018.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020.

Julius von Kügelgen, Michel Besserve, Wendong Liang, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=V87gZeSOL4.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. *The Caltech-UCSD Birds-200-2011 Dataset*. California Institute of Technology, Jul 2011.

Renhao Wang, Yu Sun, Yossi Gandelsman, Xinlei Chen, Alexei A Efros, and Xiaolong Wang. Test-time training on video streams. *arXiv preprint arXiv:2307.05014*, 2023.

Per-rAke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12:99–111, 1972.

Huaxiu Yao, Xinyu Yang, Xinyi Pan, Shengchao Liu, Pang Wei Koh, and Chelsea Finn. Improving domain generalization with domain relations. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Dc4rXq3HIA.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

Daoan Zhang, Mingkai Chen, Chenming Li, Lingyun Huang, and Jianguo Zhang. Aggregation of disentanglement: Reconsidering domain variations in domain generalization. *arXiv preprint arXiv:2302.02350*, 2023a.

Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b. URL https://openreview.net/forum?id=o16sYKHk3S.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Zeyu Zhou, Ruqi Bai, Sean Kulinski, Murat Kocaoglu, and David I. Inouye. Towards characterizing domain counterfactuals for invertible latent causal models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v1VvCWJAL8.