

Contrastive Domain Generalization via Logit Attribution Matching

Han Gao^{1,*} Kaican Li^{2,*} Yongxiang Huang¹ Lunling Wang¹ Caleb Chen Cao² Nevin L. Zhang^{2,*}

¹ Huawei Hong Kong AI Framework & Data Technologies Lab, Hong Kong SAR, China

²The Hong Kong University of Science and Technology, Hong Kong SAR, China

¹ {gaohan19, huang.yongxiang2, wangluning2}@huawei.com ² {klibf, lzheng}@cse.ust.hk, cao@ust.hk

Abstract

Domain Generalization (DG) is an important open problem in machine learning. Deep models are susceptible to domain shifts of even minute degrees, which severely compromises their reliability in real applications. To alleviate the issue, most existing methods enforce various invariant constraints across multiple training domains. However, such an approach provides little performance guarantee for novel test domains in general. In this paper, we investigate a different approach named Contrastive Domain Generalization (CDG), which exploits semantic invariance exhibited by strongly contrastive data pairs in lieu of multiple domains. We present a causal DG theory that shows the potential capability of CDG; together with a regularization technique, Logit Attribution Matching (LAM), for realizing CDG. We empirically show that LAM outperforms state-of-the-art DG methods with only a small portion of paired data and that LAM helps models better focus on semantic features which are crucial to DG.

1. Introduction

Deep learning models are successful under the i.i.d. assumption that test data are drawn from the same distribution as that of training data. However, models that generalize well in-distribution (ID) may be generalizing in unintended ways out-of-distribution (OOD) [60, 58, 20, 16]. Some image classifiers with great ID performance, in fact, rely on background cues to predict the class of foreground objects, leading to poor OOD performance [8, 69, 65]. Such reliance on spurious correlations is subject to various kinds of domain shift, affecting many real-world applications where the i.i.d. assumption cannot be guaranteed [44, 4, 31, 5].

Domain Generalization (DG) deals with the conundrum of generalizing under domain shift. It is difficult because actual test domains may vary from the training domains in numerous ways. Multi-source DG methods tackle this issue

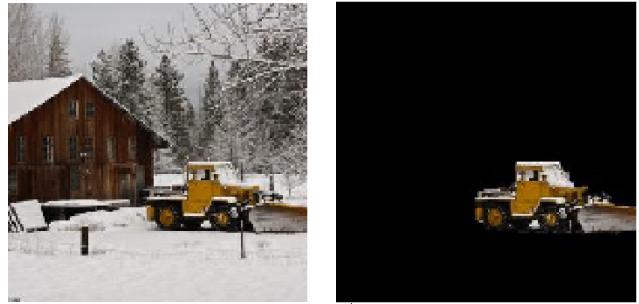


Figure 1. An example of *strongly contrastive* data pairs. The goal is to predict the class of the foreground object (vehicle) [65]. The background of the real example on the left is removed to produce the counterfactual example on the right, which is harder to predict since the context is lost, but it also forces us (and models) to focus on the foreground object and ignore the background cues.

by utilizing data from multiple training domains, however, as [21] has shown, these methods are not generally better than the classic Empirical Risk Minimization (ERM) [61]. The stagnation of algorithmic improvement under the multi-source setting suggests that we may need to rethink what kind of data to use and how to use them.

In this work, we consider *single-source* DG where there is only one training domain. On top of the training domain, we introduce a relatively small collection of *strongly contrastive* data pairs (SC pairs) that are tailored for different datasets and problem scenarios. An example of SC pairs against background shift [8] is shown in Fig. 1. To create such an SC pair, we simply put shift-inducing features in contrast, while preserving semantic features that causally explain the label. The pairwise contrast efficiently isolates the semantic features which are essential to DG, whereas multi-source DG usually requires a large amount of data collected from diverse environments to achieve a similar effect, which is sometimes prohibitively impractical or expensive.

We term our approach *Contrastive Domain Generalization* (CDG) for the utilization of SC pairs. Although bear-

* Equal contribution.

ing a similar name, CDG differs from Contrastive Learning (CL) [13] as CDG uses relatively few but highly informative data pairs for the task at hand in a supervised manner. In comparison, the positive data pair in CL is a form of *weakly contrastive* pair created via basic data augmentations that are generally task-agnostic, involving little to no domain knowledge. As for existing DG approaches, there are only several attempts to utilize data pairs that are either discovered from real data [24, 41] or generated by cross-domain image translation [52, 63]. These methods do not emphasize the proactive creation of SC pairs to enhance training signals in data. In a broader scope, CDG shares the view of data-centric AI [43] that moves from big data to good data. We believe that practitioners are willing to spend a fair amount of effort to create those SC pairs if they can significantly improve DG as we show in this paper. Moreover, we note that there are already many tools [14, 46, 70] that can be used to create high-quality SC pairs at a small cost.

To make full use of available SC pairs, we further propose a regularization technique called *Logit Attribution Matching* (LAM) for CDG. The key idea of LAM is that every SC pair should be classified using the same set of features in the latent space. Technically, this is achieved through two collaborative sub-objectives for a model consisting of a feature extractor f and a classifier g . First, LAM encourages g to put more weights on invariant features (output by f), *i.e.*, features with small pairwise difference. Second, for features that are heavily weighted by g , LAM tries to minimize the pairwise feature difference through updating f . Together, the two sub-objectives guide the model to rely on the semantically invariant features captured by the SC pairs for prediction.

Although existing pair-based DG methods also aim at a similar goal, they are not as effective as LAM due to some technical issues. For example, MatchDG [41] pushes the feature representations of a data pair $(x, \tilde{x}; y)$ close to each other regardless of the class label y . This uniform matching of $f(x)$ and $f(\tilde{x})$ may distort features that are important to other classes, harming both ID and OOD performance. In contrast, LAM adaptively matches features by their relevance to the prediction of y . Other pair-based methods propose to minimize the difference between the logits $g \circ f$ of a data pair [24, 52, 63]. These methods are free of the feature distortion issue but do not guarantee the use of semantic features. As we will elaborate in Sec. 4, the design of LAM combines their advantages.

The contribution of this paper is summarized as follows:

- We propose CDG, a generic data-centric approach to DG, which emphasizes the creation of SC data pairs tailored for different datasets and problems.
- We present a causal DG theory that formally demonstrates the potential capability of CDG.

	multi-source	single-source
domain	DANN [19], IRM [7], VREx [32], GDRO [56]	RSC [26], SD [48]
data pair	MBDG [52], RICE [63]	CoRe [24], MatchDG [41]

Table 1. Categorization of representative DG baselines. The baselines are categorized according to the expected number of training domains (“multi-source” and “single-source”), and the level where the regularization takes place (“domain” and “data pair”).

- We propose LAM, a regularization technique for CDG, which outperforms most existing DG methods including both multi-source and single-source (with and w/o additional pairs) methods in challenging OOD tasks.

2. Related Work

2.1. Domain Generalization

DG methods can be roughly divided into multi-source and single-source methods while multi-source methods are the majority. Despite this difference, almost all the methods propose a form of regularization that prevents models from overfitting the training domain(s). To get the big picture, we categorize representative DG baselines in Tab. 1. LAM, as a regularization technique, belongs to the single-source “data pair” category, but the general approach of CDG differs from the other ones in this category as we explicitly advocate the creation of SC pairs.

2.1.1 Multi-Source Domain Generalization

Good DG can be expected if the test domain is close to the training domains. This motivates the *Data Augmentation* approach, which aims at generating novel training data via methods such as input/feature mixup [42, 68, 67] and deep generative models [72, 59, 64, 36, 29]. However, the generated data usually lack of variations beyond the training domains and thus fail to bridge the domain gap.

Robust Optimization (RO) seeks to minimize the worst-case risk across all possible training domains [10]. Representative methods include Distributionally RO (DRO) with adversarial data augmentation [62], and GDRO [56] which deals with pre-defined groups (domains). It is shown that RO minimizes the risk on some mixture of the training domains, but there is no guarantee beyond the mixture [6].

Invariant Representation Learning (IRL) aims at mapping inputs from each training domain to an invariant distribution in the feature space from which the original domain is indistinguishable. IRL is motivated by a formal result that bounds the test error by the training error and the divergence between the domains [9]. A seminal work of IRL is Domain-Adversarial Neural Networks (DANN) [19] in which many related works take root [35, 3, 73, 37, 2, 71,

[40]. However, it is recently found for IRL that training-domain invariance alone does not always imply test-domain invariance and may even harm test performance [18].

Invariant Prediction (IP) is an emerging trend of research that seeks invariant causal predictors [47]. Such predictors are invariant to features other than causal features. To achieve this, Invariant Risk Minimization (IRM) [7] regularizes ERM with a penalty that enforces cross-domain optimality on the classifier. As alternatives, [32, 66, 27, 12] propose to penalize the variance or the difference of training-domain risks. These methods are usually effective when the domain shift is small [1] but have difficulties in more general situations as being unable to discover the causal features [54, 28].

2.1.2 Single-Source Domain Generalization

Single-source DG does not assume access to multiple training domains. One of the main approaches to single-source DG is to discover predictive features that are more sophisticated than simple cues spuriously correlated with labels. RSC [26] and SD [48] are two prominent methods in this direction. Another main direction of single-source DG utilizes additional data pairs. These methods include MatchDG [41] and CoRe [24]. Given a data pair $(x, \tilde{x}; y)$, they directly push either the feature representations ($f(x), f(\tilde{x})$) or the logits ($g \circ f(x), g \circ f(\tilde{x})$) close to each other. These constraints are too stringent sometimes as they do not take the label y of every pair into consideration, enforcing uniform invariance across all classes at all time. On a side note, although methods like MBDG [52] and RICE [63] also utilize data pairs, they are not single-source methods as they expect multiple training domains to generate the data pairs via cross-domain image translation.

2.2. Contrastive Learning

At first glimpse, CDG might appear similar to self-supervised contrastive learning (S²CL) [13, 11] and supervised contrastive learning (SCL) [30]. However, there are important differences. In S²CL, example pairs (groups) are created via basic data augmentations such as scaling, rotation, and random cropping. Such pairs are not informative enough for LAM, and they resulted in a performance drop in our experiments (see Sec. 5.6). In SCL, groups of examples are simply taken from the same class. They are not as informative as SC pairs in terms of domain knowledge.

3. A Causal Theory for Domain Generalization

We present a causal theory for DG in support of CDG. In single-source DG, there is only one source (or training) domain. We use $P_s(X, Y)$ to denote the generative model of the source domain, and $\hat{P}_s(Y | X)$ to denote the prediction model learned from $P_s(X, Y)$. Let $P_t(X, Y)$ be the generative model of an unseen target (or test) domain. For sim-

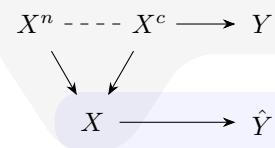


Figure 2. Causal graph of the data generation process P (gray-shaded) and the prediction process \hat{P} (blue-shaded). The input variable X is generated from two latent variables X^c and X^n , where only X^c contains semantic information for determining the target variable Y . The dashed line between X^c and X^n indicates potential correlation. In the prediction model \hat{P} , the output variable \hat{Y} is predicted from X without knowing X^c and X^n a priori. Similar models are also proposed in [38, 24, 41, 45], where X^c/X^n are referred to as semantic/variation factors, causal/non-causal factors, core/non-core factors, and content/style.

plicity, we use cross entropy as the loss function. The objective of DG is to minimize the following out-of-distribution (OOD) loss:

$$\mathbb{E}_{(X, Y) \sim P_t} [-\log \hat{P}_s(Y | X)]. \quad (1)$$

Note that the OOD loss depends on the generative models, P_s and P_t , of both domains. We show that the OOD loss is minimized under a set of general conditions involving how P_s and P_t are related, and how $\hat{P}_s(Y | X)$ should behave.

First of all, DG from P_s to arbitrary P_t is clearly impossible. To reasonably constrain them, we consider a class of *causally invariant* generative models that are consistent with the causal graph illustrated in Fig. 2. These generative models can also be formally stated as follows:

Definition 1 (Causally-Invariant Generative Models).

A collection \mathcal{P} of generative models for the input-output pair (X, Y) is causally invariant if and only if

1. for every $P \in \mathcal{P}$, the input X is generated from two latent factors, X^c and X^n , while the label Y is generated from X^c , implying that

$$P(Y | X) = P(Y | X^c); \quad (2)$$

2. for every pair $P, P' \in \mathcal{P}$,

$$P(Y | X^c) = P'(Y | X^c). \quad (3)$$

Based on \mathcal{P} that is causally invariant, we further define a class of prediction models that is *causally faithful* to \mathcal{P} .

Definition 2 (Causally-Faithful Prediction Model).

A prediction model $\hat{P}(Y | X)$ is causally faithful to a collection \mathcal{P} of causally-invariant generative models if and only if \hat{P} uses X^c to make predictions, i.e.,

$$\hat{P}(Y | X) = \hat{P}(Y | \tilde{X}) \quad \text{when} \quad X^c = \tilde{X}^c. \quad (4)$$

These causally-faithful prediction models are of particular interests because they can minimize the OOD loss as shown by the following theorem.

Theorem 1 (Generalization via Causal Invariance). Suppose \mathcal{P} is a collection of causally-invariant generative models. Let $\hat{P}_s(Y | X)$ be the optimal prediction model for any $P_s \in \mathcal{P}$, i.e., it minimizes the in-distribution (ID) loss:

$$\mathbb{E}_{(X,Y) \sim P_s} [-\log \hat{P}_s(Y | X)]. \quad (5)$$

If $\hat{P}_s(Y | X)$ is also causally faithful to \mathcal{P} , then it minimizes the OOD loss given by Eq. (1) for any $P_t \in \mathcal{P}$ such that $\text{supp}[P_t(X^c)] \subseteq \text{supp}[P_s(X^c)]$.

The proof of this theorem can be found in Appendix A. Theorem 1 suggests that we only need to find an ID-optimal and causally faithful predictor to solve a DG problem. The first part is relatively straightforward, but simultaneously ensuring the causal faithfulness of $\hat{P}_s(Y | X)$ is not trivial by all means. Fortunately, CDG with SC pairs can help.

How can CDG help? Under the generative models given by Definition 1, an SC pair can be seen as $(x, \tilde{x}; y)$ such that x and \tilde{x} are generated from the same x^c but different x^n . Suppose that for every example $(x; y)$ drawn from the source domain P_s , we can create additional examples $(\tilde{x}; y)$ that maintain *semantic invariance* (i.e. $\tilde{x}^c = x^c$), then we have a collection of SC pairs. For every $(\tilde{x}; y)$ from this collection, suppose that \tilde{x}^n is randomly drawn from $P_s(X^n)$, and denote the resulting distribution of \tilde{X} by $Q_X(\tilde{X})$. We show that the loss of \hat{P}_s on a training domain P_s with SC pairs is minimized only if \hat{P}_s is causally faithful.

Theorem 2 (Semantic Invariance \Rightarrow Causal Invariance). Suppose \mathcal{P} is a collection of causally-invariant generative models. For any $P_s \in \mathcal{P}$, if \hat{P}_s minimizes

$$\mathbb{E}_{(X,Y) \sim P_s} \mathbb{E}_{\tilde{X} \sim Q_X} [-\log \hat{P}_s(Y | \tilde{X})], \quad (6)$$

then \hat{P}_s is causally faithful to \mathcal{P} while also minimizing the ID loss given by Eq. (5). Moreover, such \hat{P}_s is unique.

The theorem points out a direction to learn an ID-optimal and causally faithful predictor. Recall that Theorem 1 suggests that if a predictor meets both of the conditions, then DG is realized. Combining it with Theorem 2, DG can now be achieved by minimizing the loss on the domain “expanded” by the SC pairs. In short, CDG converts DG problems into ID problems as Eq. (6).

A naive solution. The theorems suggest a naive solution to DG, which is minimizing the empirical version of Eq. (6). We term this approach *Naive CDG*, which seems plausible

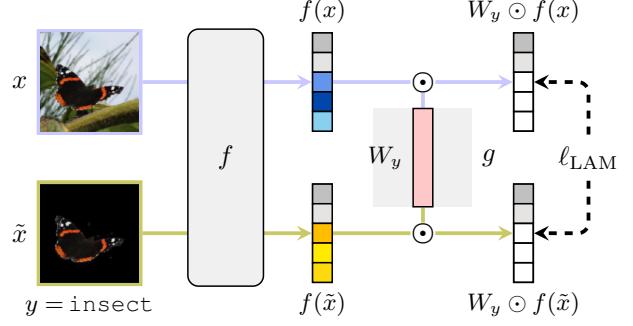


Figure 3. Illustration of how LAM works. Following the notation in the main text, $(x, \tilde{x}; y)$ is an SC pair, f is a feature extractor, W_y is the weight vector of a linear classifier g with respect to class y , and \odot is the Hadamard product. Different colors in the vector cells (except W_y) represent different feature values. By comparing x and \tilde{x} , ℓ_{LAM} preserves features that are invariant (the gray cells), while preventing the use of other features (the blue/yellow cells).

but is in fact still difficult for two reasons. First, it is sometimes hard to create \tilde{x} with arbitrary \tilde{x}^n while keeping x^c fixed. As a result, the diversity and quality of \tilde{x} may be somewhat limited in practice. This suggests that we can only minimize an estimate of Eq. (6) that is not completely free of spurious correlations between Y and X^n . Second, deep neural networks are notoriously prone to be trapped by local minima, especially by the highly-attractive ones induced by the spurious correlations [20, 58].

To fully exploit available SC pairs and prevent a model from overfitting the spurious correlations, we further propose a regularization technique designed for CDG.

4. Logit Attribution Matching

In the previous section, we discussed how CDG works in principle and why we need extra regularization to fully exploit the potential of CDG. In this section, we introduce a regularization technique, called Logit Attribution Matching (LAM), to improve Naive CDG.

Formulation. Let f be the feature extractor of a model and g be the linear classifier on top of f . For every SC pair $(x, \tilde{x}; y)$, the objective of LAM is to minimize

$$\ell_{\text{LAM}}(x, \tilde{x}; y) = \|W_y \odot [f(x) - f(\tilde{x})]\|_2^2, \quad (7)$$

where W_y is the weight vector of g with respect to the class y , and \odot is the Hadamard product. Rewriting the term as $\|W_y \odot f(x) - W_y \odot f(\tilde{x})\|_2^2$, we see that $W_y \odot f(x)$ is the contribution of features $f(x)$ to the logit of class y in the case of x , and $W_y \odot f(\tilde{x})$ is that in the case of \tilde{x} . The whole term therefore aims to match the contributions of the features to the class y in the two cases (hence the name of LAM). See Fig. 3 for an illustration of how it works.

From the formulation, it should be easy to see the desiderata behind LAM—*some (but not necessarily all) components of the latent representations of x and \tilde{x} should be similar, and the classification should be based only on those components*. In connection with our DG theory, LAM guides f to represent data with the semantic factors X^c since they are the only invariant components shared by an SC pair. Furthermore, LAM also guides g to use only X^c to predict, making the model causally faithful. To see these concretely, we rewrite the right hand side of Eq. (7) as

$$\sum_{i=1}^d W_{iy}^2 [f_i(x) - f_i(\tilde{x})]^2, \quad (8)$$

where d is the dimension of the feature space, f_i is the u -th output dimension of f , and W_{iy} is the weight between f_i and the logit of class y . According to Eq. (8), LAM encourages feature components f_i with large W_{iy}^2 to have similar activations on x and \tilde{x} . This helps f extract X^c from the data. Meanwhile, LAM penalizes those W_{iy} of feature components with large $|f_i(x) - f_i(\tilde{x})|$, forcing g to use f_i with similar activations on x and \tilde{x} , which are essentially representations of X^c . In summary, f and g are optimized collaboratively towards a causally faithful model.

To realize DG, now we only need to ensure that the ID loss is also minimized. This can be done by incorporating the ERM loss. Given a dataset $\mathcal{D} = \{(x_j; y_j)\}_{j=1}^N$, out of which M examples are paired with an SC example (denoted by $\tilde{\mathcal{D}} = \{(x_j, \tilde{x}_j; y_j)\}_{j=1}^M$), the final objective function is

$$\mathcal{L}_{\text{ERM}}(\mathcal{D}) + \lambda \mathcal{L}_{\text{LAM}}(\tilde{\mathcal{D}}), \quad (9)$$

where $\mathcal{L}_{\text{ERM}}(\mathcal{D}) = \frac{1}{N} \sum_{j=1}^N \ell_{\text{CE}}(g \circ f(x_j), y_j)$ is regularized by $\mathcal{L}_{\text{LAM}}(\tilde{\mathcal{D}}) = \frac{1}{M} \sum_{j=1}^M \ell_{\text{LAM}}(x_j, \tilde{x}_j; y_j)$ with the coefficient λ being a hyperparameter.

Comparison with similar methods. Existing methods that are most related to LAM are CoRe [24], MatchDG [41], MBDG [52], and RICE [63]. These methods aim to align either the features or the logits of a data pair. MatchDG takes the former approach, while the rest take the latter.

More specifically, given an SC pair $(x, \tilde{x}; y)$, MatchDG tries to minimize $\|f(x) - f(\tilde{x})\|_2$. This uniform alignment is different from LAM as it does not consider the class label y , whereas from Eq. (8), we can see that LAM enforces the invariance separately for each f_i , depending on their contribution to the prediction of the class y . It is relevant because the uniform alignment may distort f_i that is important only to some other class, *i.e.*, f_i with small $|W_{iy}|$ but large $|W_{iy'}|$ for some $y' \neq y$, while LAM is more lenient with such f_i . The other three methods seek to minimize the distance $D(g \circ f(x), g \circ f(\tilde{x}))$ under some metric D . Although they are free of the feature distortion issue, they do not guarantee the use of invariant features.

Moreover, these methods share a common issue as they all aim at some sort of alignment across *all* classes. For such alignment to be possible, the entire X^c needs to be the same for every SC pair, which is sometimes too demanding and unrealistic. In comparison, LAM only requires the part of X^c that explains the class y to be the same, softening the requirement. As we will show in the experiments, LAM does outperform these similar methods when provided with SC pairs of reasonably fair quality.

5. Experiments

We evaluate LAM on three OOD image classification datasets with different domain shift: two with background shift and one with style shift, representing a good portion of domain shift encountered in real applications. On these datasets, we compare LAM against multiple strong DG methods which are listed in Tab. 1.

5.1. Datasets

ImageNet-9 [65] is a dataset with background shift across three domains: *original*, *only-fg*, and *mixed-rand*. The first domain consists of original images from ImageNet [15], while the other two domains are derived from the first domain, one with pure-black backgrounds (*only-fg*) and the other with random backgrounds (*mixed-rand*). In our experiments, we use the original domain as the training domain and *mixed-rand* as the test domain.

NICO [23] is a dataset with background shift across multiple domains. It contains 19 classes of animals and vehicles in various contexts such as “at home” and “on beach”. In total, NICO covers 288 class-context combinations. For each class, we randomly pick one context as the test domain, while the rest are used as the training domain.

PACS [34] is a dataset with style shift across four domains: *photo*, *art painting*, *cartoon*, and *sketch*. Each domain depicts seven classes of objects and persons in a different style. Commonly, three domains are used for training and the remaining one is used for testing in an alternating fashion [34, 21]. We adopt the same setting except that we always use the *photo* domain for training because we use the *photo* domain to generate SC pairs.

5.2. SC-Pair Creation

As illustrated in Fig. 1, an SC pair can be formed by a *real example* and a *counterfactual example*. Given a real example, we may need to use different tools to create the corresponding counterfactual example for different domain shift. We use two simple SC-pair creation methods in the case of background shift and style shift.

Background shift. We utilize GrabCut [55], an interactive segmentation tool, to segment foreground objects from backgrounds. With the segmentation masks, we remove the

	ImageNet-9 [65]		NICO [23]	
Method	ViT	RN-50	ViT	RN-50
ERM [61]	85.26	78.49	94.85	86.14
RSC [26]	85.95	79.86	95.45	80.00
SD [48]	86.88	81.31	95.64	84.06
DANN [19]	86.16	77.75	93.66	80.40
GDRO [56]	85.11	65.64	92.38	64.25
IRM [7]	86.38	75.29	93.47	72.48
VREx [32]	86.21	75.94	93.86	77.62
MBDG [52]	83.97	69.77	93.96	71.68
RICE [63]	84.28	74.05	91.88	70.99
CoRe [24]	77.85	68.63	94.55	77.22
MatchDG [41]	89.61	74.17	94.55	73.86
LAM (ours)	91.18	81.72	96.93	87.62

Table 2. OOD performance of LAM and baselines on ImageNet-9 and NICO with background shift. Experiments are conducted on both ViT and ResNet-50 (RN-50) architectures.

backgrounds by coloring them in black. In addition, we may also randomly replace the original background of an image with the background of another image. Using this method, 20 SC pairs are created for each class of NICO, taking up about 5% of the training data. It only takes us around three seconds on average to segment an image. As for ImageNet-9, the counterfactual examples are also created using GrabCut. The segmentation masks are already provided by [65]. If not stated otherwise, the proportion of SC pairs we use is also 5% for ImageNet-9, same as NICO.

Style shift. For style shift in PACS, we use StableDiffusion [53] to translate the photos of PACS into a different style. For each class of PACS, 100 SC pairs are created (less than 10% of the training data). The generated examples cover a wide range of styles, so there is no need for augmentations like randomizing background as in the case of background shift. More details on the creation of SC pairs and examples of the created SC pairs are provided in Appendix B.

5.3. Baselines

We compare LAM with ERM [61] and the ten DG baselines listed in Tab. 1. Two of them, *i.e.*, CoRe [24] and MatchDG [41], are closely related to LAM since they also based on data pairs. The baselines are selected for two main considerations: (i) they represent a diverse category of DG approaches, and (ii) they are the state of the art in their respective category.

For **ERM** and single-source methods that do not exploit paired relations (**RSC** [26] and **SD** [48]), the counterfactual examples are not used. For the multi-source methods that

	PACS [34]			
Method	Art	Cartoon	Sketch	Avg.
ERM [61]	83.54	83.11	80.12	82.26
RSC [26]	82.86	85.41	83.35	83.87
SD [48]	85.35	87.80	82.62	85.26
DANN [19]	78.22	80.16	71.95	76.78
GDRO [56]	82.91	83.87	84.37	83.72
IRM [7]	80.62	86.43	84.37	83.81
VREx [32]	84.18	87.50	80.78	84.15
MBDG [52]	82.71	81.61	67.57	77.30
RICE [63]	89.26	73.29	84.70	82.42
CoRe [24]	86.23	83.40	82.94	84.19
MatchDG [41]	78.71	84.21	83.66	82.19
LAM (ours)	88.67	87.50	81.80	86.00

Table 3. OOD performance of LAM and baselines on PACS with style shift. Experiments are conducted on ResNet-50. The test domain is alternated within Art, Cartoon, and Sketch domain. Photo domain is excluded since SC pairs are generated from its images.

do not exploit paired relations (**DANN** [19], **GDRO** [56], **IRM** [7], and **VREx** [32]), the counterfactual examples of the SC pairs are used to form an additional training domain. **MBDG** [52] and **RICE** [63] generate data pairs by cross-domain image translation across multiple training domains, one of which is formed by the counterfactual examples. The two remaining baselines (**CoRe** and **MatchDG**) both exploit paired relations and are provided with the pairing information of the SC pairs just like LAM.

In some experiments, we also compare LAM with **Naive CDG**, which is basically ERM but with the counterfactual examples mixed with the real examples for training.

5.4. Network Architecture and Training Strategy

We conduct the experiments on two widely-used architectures, namely ResNet-50 [22] and ViT-B/16 [17], with CLIP pre-training weights [49]. We adopt the training strategy (LP-FT) proposed in [33] to more effectively reuse the learned representations of the pre-trained weights. In short, LP-FT consists of linear probing (LP) on the linear classifier while freezing the feature extractor, followed by fine-tuning (FT) the entire model. Hyperparameter choices and more implementation details are provided in Appendix C.

5.5. Results

LAM consistently outperforms the baselines. From Tab. 2, we see that LAM performs the best in all cases under background shift. In particular, the advantage of LAM is more significant on ViT than on ResNet-50, taking potential performance saturation into account. Overall, the most competitive baseline is SD, with an average accuracy

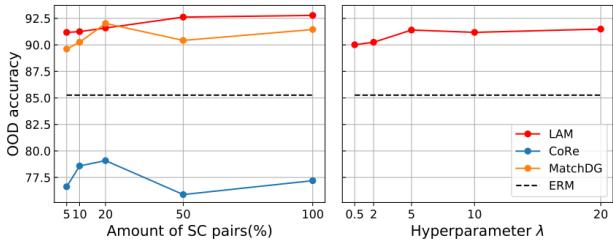


Figure 4. Left: OOD performance of LAM, CoRe, MatchDG, and ERM on ImageNet-9 as a function of the proportion of data provided with SC pairs. Right: OOD performance of LAM on ImageNet-9 as a function of the hyperparameter λ , the weight of LAM regularization, with 5% SC pairs.

	Naive	LAM		
Seg. method	GrabCut	box	auto	GrabCut
Acc. (5%-pair)	85.71	89.15	89.37	91.18
Acc. (100%-pair)	87.41	89.60	90.44	92.81

Table 4. OOD performance on ImageNet-9 with SC pairs produced from foreground masks of varying quality. We test LAM with bounding boxes as masks (box), masks automatically produced by a semantic segmentation model [39] (auto), and masks produced using the interactive segmentation tool, GrabCut. We include the performance of Naive CDG for reference. The accuracy in both cases of 5% and 100% paired training examples are provided.

of 86.97. LAM, which has an average accuracy of 89.36, improves the second-best result by a great margin.

As for style shift, Tab. 3 shows that LAM is also superior to the previous works. On average, the second-best method is still SD. LAM improves SD by some margin again, although less significantly than the previous results in Tab. 2. It is worth noting that the OOD performance on the sketch domain is relatively low for LAM. We suspect this is due to the fact that the images in the sketch domain are quite distinct from the other domains, sharing very few semantic features with the training domains.

From the results, we note that SD is a very strong baseline. Since SD is a single-source method that does not rely on any paired data, it is possible that LAM combined with SD can further boost the OOD performance.

LAM improves gracefully with more SC pairs. As with LAM, CoRe, and MatchDG also uses SC pairs. To further compare them, we increase the amount of SC pairs and see how the performance of these three methods changes. In Fig. 4, as the fraction of training examples with SC pairs increases from 5% to 100% on ImageNet-9, the performance of LAM improves steadily from 91.18 to 92.81. On the other hand, CoRe and MatchDG are unable to consistently improve, sometimes even adversely affected by more SC

Method	Contrast	LAM reg.	Rand. BG	Acc.
ERM	-	-	-	85.26
Naive CDG [†]	strong	-	-	85.71
Naive CDG	strong	-	✓	89.75
LAM [‡]	weak	✓	-	84.80
LAM [†]	strong	✓	-	88.12
LAM	strong	✓	✓	91.18

Table 5. Ablation study of LAM on ImageNet-9. **Contrast** indicates whether the training data are provided with contrastive data pairs and whether they are strongly or weakly contrastive pairs. **LAM reg.** indicates whether LAM regularization is active. **Rand. BG** indicates whether the background is randomized as opposed to pure-black background. [†]Background is not randomized for these methods. [‡]The data pairs for the method are weakly contrastive.

pairs. This observation is coherent with our analysis at the end of Sec. 4 where we have pointed out a fundamental difference between LAM and those two previous methods—CoRe and MatchDG have a high standard for the quality of SC pairs. When the quality does not meet the standard, merely increasing the quantity is not helpful.

LAM is robust to the quality of SC pairs. To see if LAM is robust to the quality of SC pairs, we test LAM with SC pairs of varying foreground-background segmentation quality. The results are shown in Tab. 4. Compared with Naive CDG, the improvement brought by LAM is still significant even when the segmentation masks reduce to bounding boxes whereas Naive CDG has access to GrabCut segmentation masks. When using segmentation masks produced by an old semantic segmentation model, FCN [39], LAM also significantly outperforms Naive CDG. Note that even with 100% GrabCut-produced SC pairs, Naive CDG is still unable to surpass LAM with only 5% bounding-box SC pairs. This again highlights the importance of LAM and its robustness to the quality of SC pairs. Examples of the SC pairs of varying quality are provided in Appendix B.

5.6. Further Analysis

In this section, we conduct additional experiments to corroborate our previous findings and provide some more empirical analysis as to why LAM works. We also provide visualizations of the attention maps of LAM in comparison with other baselines. All experiments in this section are conducted on ImageNet-9 (5% SC pairs) with the ViT model introduced in Sec. 5.4.

Ablation study. We analyze the impact of each major CDG component on the overall performance of LAM-regularized CDG (as opposed to Naive CDG) on ImageNet-9. The components are: (i) SC pairs (vs. no data pair and weakly contrastive pairs), (ii) LAM regularization term, and

(iii) whether random backgrounds are used to generate the SC pairs. The results are reported in Tab. 5.

In the table, we see a salient improvement from Naive CDG to LAM, highlighting the effectiveness of regularization such as LAM in CDG. Moreover, we note that LAM^\dagger is much better than LAM^\ddagger , showing that SC pairs are indeed much more informative and beneficial to CDG than weakly contrastive pairs generated via basic data augmentations like cropping, scaling, and flipping.

Last but not least, by comparing Naive CDG with Naive CDG^\dagger , and LAM with LAM^\dagger , we see that random background also plays an important role in CDG against background shift. This observation is consistent with our Theorem 2 which suggests that non-semantic factors X^n (the backgrounds in this case) should be diverse in order to achieve optimal CDG. If X^n is not diverse, the improvement is quite limited as we see from the comparison of Naive CDG^\dagger with ERM. But more importantly, even without random backgrounds, LAM still helps a lot in improving the performance (see LAM^\dagger vs. Naive CDG^\dagger).

Explanation of model predictions. GradCAM [57] is a commonly-used explanation method for image classification. It is often used to visualize the attention of a deep neural network. Fig. 5 shows the GradCAM heatmaps of ERM, CoRe, MatchDG, and LAM on examples from the test domain (mixed-rand¹) of ImageNet-9. In particular, CoRe and MatchDG use the same SC pairs as LAM during training.

Compared with the baselines, it is clear from the heatmaps that the attention of LAM is better focused on the semantic area (which explains the the label) of an image, while ERM and CoRe completely failed to do so. MatchDG performs much better than ERM and CoRe but still slightly worse than LAM, mistaking bird for fish due to the marine-like background. Fig. 5 helps explain why LAM has superior OOD performance in Tab. 2 as the semantic areas contain all the information of the semantic factors X^c (in the case of background shift), which are crucial to DG according to our theory in Sec. 3.

Sensitivity to the hyperparameter λ . The weight λ of the regularization term is the only additional hyperparameter of LAM on top of ERM. Here, we analyze if the performance of LAM is sensitive to the choice of λ . From the result shown in the right sub-figure of Fig. 4, as λ varies across a wide range (from 0.5 to 20), the OOD accuracy of LAM is stably above 90% (whereas ERM is around 85%). Furthermore, when λ is between 5 and 20, the OOD accuracy is consistently around 91.5%. All these indicate that LAM is not sensitive to the choice of λ , reflecting the le-

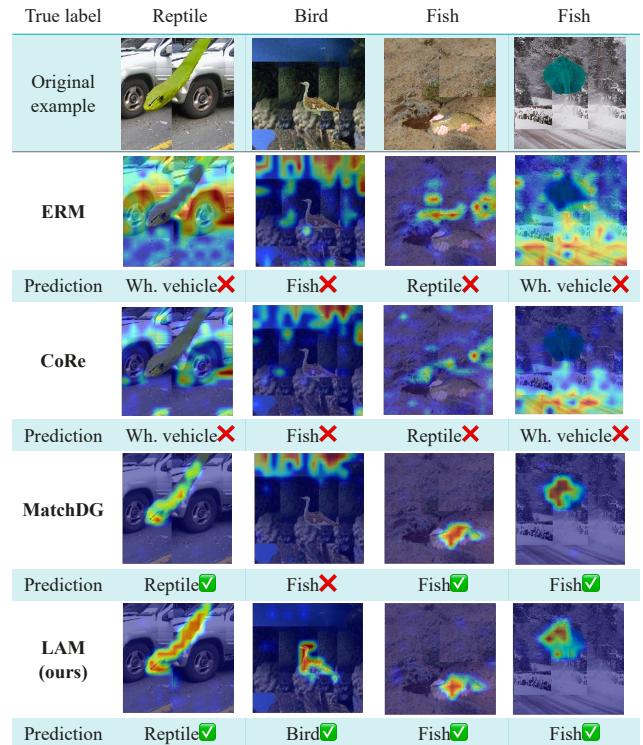


Figure 5. GradCAM heatmaps of LAM compared with ERM and baselines that also utilize data pairs. The examples are from the test domain of ImageNet-9. Images in this domain has random backgrounds from other images. The heatmaps show that LAM helps the model focus on the semantic area of an image, *i.e.* the area of the foreground object corresponding to the true label. “Wh. vehicle” stands for “Wheeled vehicle”. See Appendix D for more GradCAM heatmaps.

nience of LAM as it hardly interferes with the ERM objective even when λ is large.

6. Conclusion

In this paper, we have proposed and investigated CDG as a generic data-centric approach to DG through theoretical analysis and extensive experiments. Our method, LAM, as an effective regularization technique for CDG, demonstrates superior DG performance against background shift and style shift in comparison with various state-of-the-art DG methods. We have also shown that LAM is robust to varying quantity and quality of SC pairs as well as a wide range of hyperparameter choices. Attention visualizations further highlight the strength of LAM in capturing semantic features which are essential to DG. From the results, we believe LAM (and CDG in general) opens up a promising new avenue for future DG research.

¹Images in the mixed-rand domain of ImageNet-9 has random backgrounds from other images.

References

- [1] Kartik Ahuja, Jun Wang, Amit Dhurandhar, Karthikeyan Shanmugam, and Kush R. Varshney. Empirical or invariant risk minimization? a sample complexity perspective. In *ICLR*, 2021. 3
- [2] Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo. Adversarial invariant feature learning with accuracy constraint for domain generalization. In *ECML-PKDD*, 2019. 3
- [3] Isabela Albuquerque, João Monteiro, Mohammad Darvishi, Tiago H Falk, and Ioannis Mitliagkas. Generalizing to unseen domains via distribution matching. *arXiv:1911.00804*, 2019. 3
- [4] Michael A Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *CVPR*, 2019. 1
- [5] Sharib Ali, Noha Ghatwary, Debesh Jha, Ece Isik-Polat, Gorkem Polat, Chen Yang, Wuyang Li, Adrian Galdran, Miguel-Ángel González Ballester, Vajira Thambawita, et al. Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge. *arXiv preprint arXiv:2202.12031*, 2022. 1
- [6] Martin Arjovsky. *Out of distribution generalization in machine learning*. PhD thesis, New York University, 2020. 2
- [7] Martin Arjovsky, Léon Bottou, Ishaaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv:1907.02893*, 2019. 2, 3, 6
- [8] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *ECCV*, 2018. 1
- [9] 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(1):151–175, 2010. 2
- [10] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton university press, 2009. 2
- [11] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 3
- [12] Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. Invariant rationalization. In *International Conference on Machine Learning*, pages 1448–1458. PMLR, 2020. 3
- [13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 3
- [14] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 2
- [15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5
- [16] Lauro Langosco Di Langosco, Jack Koch, Lee D Sharkey, Jacob Pfau, and David Krueger. Goal misgeneralization in deep reinforcement learning. In *International Conference on Machine Learning*, pages 12004–12019. PMLR, 2022. 1
- [17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 6
- [18] Tigran Galstyan, Hrayr Harutyunyan, Hrant Khachatrian, Greg Ver Steeg, and Aram Galstyan. Failure modes of domain generalization algorithms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19077–19086, 2022. 3
- [19] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 2016. 2, 6
- [20] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020. 1, 4
- [21] Ishaaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *ICLR*, 2021. 1, 5, 18
- [22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6
- [23] Yue He, Zheyuan Shen, and Peng Cui. Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, 2020. 5, 6
- [24] Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *Machine Learning*, 110(2):303–348, 2021. 2, 3, 5, 6
- [25] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018. 17
- [26] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 124–140. Springer, 2020. 2, 3, 6
- [27] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Domain extrapolation via regret minimization. *arXiv preprint arXiv:2006.03908*, 2020. 3
- [28] Prithish Kamath, Akilesh Tangella, Danica Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? In *International Conference on Artificial Intelligence and Statistics*, pages 4069–4077. PMLR, 2021. 3
- [29] Juwon Kang, Sohyun Lee, Namyup Kim, and Suha Kwak. Style neophile: Constantly seeking novel styles for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7130–7140, 2022. 2
- [30] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 3

- [31] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, et al. Wilds: A benchmark of in-the-wild distribution shifts. *arXiv:2012.07421*, 2020. 1
- [32] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). *arXiv:2003.00688*, 2020. 2, 3, 6
- [33] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022. 6, 13
- [34] D. Li, Y. Yang, Y. Song, and T. M. Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017. 5, 6
- [35] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, 2018. 3
- [36] Lei Li, Ke Gao, Juan Cao, Ziyao Huang, Yepeng Weng, Xiaoyue Mi, Zhengze Yu, Xiaoya Li, and Boyang Xia. Progressive domain expansion network for single domain generalization. In *CVPR*, 2021. 2
- [37] 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, 2018. 3
- [38] Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. Learning causal semantic representation for out-of-distribution prediction. In *NeurIPS*, 2021. 3
- [39] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 7, 13
- [40] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018. 3
- [41] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International Conference on Machine Learning*, pages 7313–7324. PMLR, 2021. 2, 3, 5, 6
- [42] Massimiliano Mancini, Zeynep Akata, Elisa Ricci, and Barbara Caputo. Towards recognizing unseen categories in unseen domains. In *ECCV*, 2020. 2
- [43] Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Douwe Kiela, David Jurado, et al. Dataperf: Benchmarks for data-centric ai development. *arXiv preprint arXiv:2207.10062*, 2022. 2
- [44] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv:1907.07484*, 2019. 1
- [45] Jovana Mitrovic, Brian McWilliams, Jacob C Walker, Lars Holger Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. In *ICLR*, 2021. 3
- [46] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 2
- [47] 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. 3
- [48] Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34:1256–1272, 2021. 2, 3, 6
- [49] 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. 6
- [50] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 13
- [51] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 13
- [52] Alexander Robey, George Pappas, and Hamed Hassani. Model-based domain generalization. In *NeurIPS*, 2021. 2, 3, 5, 6
- [53] 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. 6, 13
- [54] Elan Rosenfeld, Pradeep Kumar Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *ICLR*, 2021. 3
- [55] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. “grabcut” interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004. 5, 13
- [56] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *ICLR*, 2020. 2, 6
- [57] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv 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. 8
- [58] Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias

- in neural networks. *Advances in Neural Information Processing Systems*, 33:9573–9585, 2020. 1, 4
- [59] Nathan Somavarapu, Chih-Yao Ma, and Zsolt Kira. Frustratingly simple domain generalization via image stylization. *arXiv preprint arXiv:2006.11207*, 2020. 2
- [60] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1
- [61] Vladimir Vapnik. *Statistical Learning Theory*. Wiley, 1998. 1, 6
- [62] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *NeurIPS*, 2018. 2
- [63] Ruoyu Wang, Mingyang Yi, Zhitang Chen, and Shengyu Zhu. Out-of-distribution generalization with causal invariant transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 375–385, 2022. 2, 3, 5, 6
- [64] Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotagh. Learning to diversify for single domain generalization. In *ICCV*, 2021. 2
- [65] Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020. 1, 5, 6
- [66] Chuanlong Xie, Haotian Ye, Fei Chen, Yue Liu, Rui Sun, and Zhenguo Li. Risk variance penalization. *arXiv preprint arXiv:2006.07544*, 2020. 3
- [67] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *AAAI*, 2020. 2
- [68] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv:2001.00677*, 2020. 2
- [69] 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. 1
- [70] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 2
- [71] Shanshan Zhao, Mingming Gong, Tongliang Liu, Huan Fu, and Dacheng Tao. Domain generalization via entropy regularization. In *NeurIPS*, 2020. 3
- [72] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *ECCV*, 2020. 2
- [73] Wei Zhu, Le Lu, Jing Xiao, Mei Han, Jiebo Luo, and Adam P Harrison. Localized adversarial domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7108–7118, 2022. 3

Supplementary Materials

A. Proofs

To prepare the proof of the theorems, we define a new joint distribution:

$$\tilde{P}_s(X, X^c, Y) := \hat{P}_s(Y | X) P_s(X, X^c), \quad (\text{A.1})$$

which combines the prediction model $\hat{P}_s(Y | X)$ and the joint distribution of X and X^c in the generative model P_s . As we will show in the proof of Theorem 1, Eq. (A.1) bridges the ID loss in Eq. (5) and the OOD loss in Eq. (1). First, we have the following lemma:

Lemma 1. *If $\hat{P}_s(Y | X)$ is causally faithful, then it makes prediction based on X^c , i.e., $\hat{P}_s(Y | X) = \tilde{P}_s(Y | X^c)$.*

Proof. Since $\hat{P}_s(Y | X)$ is causally faithful, we have

$$\begin{aligned} \tilde{P}_s(Y | X^c) &= \int \tilde{P}_s(\tilde{X}, Y | X^c) d\tilde{X} \\ &= \int \frac{\tilde{P}_s(\tilde{X}, X^c, Y)}{\tilde{P}_s(X^c)} d\tilde{X} \\ &= \int \frac{\hat{P}_s(Y | \tilde{X}) P_s(\tilde{X}, X^c)}{P_s(X^c)} d\tilde{X} \\ &= \int \hat{P}_s(Y | \tilde{X}) P_s(\tilde{X} | X^c) d\tilde{X} \\ &= \hat{P}_s(Y | X) \int P_s(\tilde{X} | X^c) d\tilde{X} \\ &= \hat{P}_s(Y | X), \end{aligned}$$

where the third equality follows from Eq. (A.1) and the fifth equality follows from

$$\begin{aligned} \hat{P}_s(Y | \tilde{X}) &= \hat{P}_s(Y | X) && \text{if } \tilde{X}^c = X^c; \\ P_s(\tilde{X} | X^c) &= 0 && \text{if } \tilde{X}^c \neq X^c. \end{aligned} \quad \square$$

A.1. Proof of Theorem 1

Consider the ID loss

$$\begin{aligned} &\mathbb{E}_{(X, Y) \sim P_s} [-\log \hat{P}_s(Y | X)] \\ &= - \int P_s(X, Y) \log \hat{P}_s(Y | X) dXdY \\ &= - \int P_s(X, X^c, Y) \log \hat{P}_s(Y | X) dXdX^c dY \\ &= - \int P_s(X, X^c, Y) \log \tilde{P}_s(Y | X^c) dXdX^c dY \\ &= - \int P_s(X^c, Y) \log \tilde{P}_s(Y | X^c) dX^c dY \\ &= - \int P_s(X^c) [P_s(Y | X^c) \log \tilde{P}_s(Y | X^c) dY] dX^c, \end{aligned} \quad (\text{A.2})$$

where the third equality uses $\hat{P}_s(Y | X) = \tilde{P}_s(Y | X^c)$ from Lemma 1. By Gibbs' inequality, the quantity on the last line, and hence the ID loss, is minimized if and only if

$$P_s(Y | X^c) = \tilde{P}_s(Y | X^c)$$

for all possible values of Y and X^c such that $P_s(X^c) > 0$.²

For the OOD loss, we can carry out the same derivation as above in Eq. (A.2) and obtain:

$$\begin{aligned} &\mathbb{E}_{(X, Y) \sim P_t} [-\log \hat{P}_s(Y | X)] \\ &= - \int P_t(X^c) [P_t(Y | X^c) \log \tilde{P}_s(Y | X^c) dY] dX^c. \end{aligned}$$

Since all generative models in \mathcal{P} , including both P_t and P_s , are causally invariant, we have $P_t(Y | X^c) = P_s(Y | X^c)$. Consequently, under an additional technical condition that $\text{supp}[P_t(X^c)] \subseteq \text{supp}[P_s(X^c)]$, we have

$$\begin{aligned} &\mathbb{E}_{(X, Y) \sim P_t} [-\log \hat{P}_s(Y | X)] \\ &= - \int P_t(X^c) [P_s(Y | X^c) \log \tilde{P}_s(Y | X^c) dY] dX^c. \end{aligned}$$

Because the quantity on the last line of Eq. (A.2) is minimized, the OOD loss above is also minimized. \square

A.2. Proof of Theorem 2

Note that

$$\begin{aligned} &\mathbb{E}_{(X, Y) \sim P_s} \mathbb{E}_{\tilde{X} \sim Q_X} [-\log \hat{P}_s(Y | \tilde{X})] \\ &= - \int P_s(X, Y) Q_X(\tilde{X}) \log \hat{P}_s(Y | \tilde{X}) dXd\tilde{X} dY \quad (\text{A.3}) \\ &= - \int P_s(X) Q_X(\tilde{X}) h(X, \tilde{X}) dXd\tilde{X}, \end{aligned}$$

where

$$h(X, \tilde{X}) := \int P_s(Y | X) \log \hat{P}_s(Y | \tilde{X}) dY.$$

By Gibbs' inequality, Eq. (A.3) is minimized if and only if $\hat{P}_s(Y | \tilde{X}) = P_s(Y | X)$ for all possible values of X , Y , and \tilde{X} such that $\hat{P}_s(\tilde{X}) > 0$ and $P_s(X) > 0$.³ It follows that (i) $\hat{P}_s(Y | \tilde{X}) = \hat{P}_s(Y | X)$, i.e. $\hat{P}_s(Y | X)$ is causally faithful and thus Eq. (A.2) holds; and (ii) the ID loss is also uniquely minimized by $\hat{P}_s(Y | X)$. \square

B. Strongly-Contrastive (SC) Data Pairs

Background shift. An SC pair can be formed by a real example and a counterfactual example. In practice, for creating counterfactual examples under background shift, we

²Strictly speaking, this requires the probability density functions of $P_s(Y | X^c)$ and $\tilde{P}_s(Y | X^c)$ to be continuous.

³Again, we assume that $\hat{P}_s(Y | X)$ and $P_s(Y | X)$ are continuous.

may utilize object detection models [51, 50] to generate bounding boxes for target objects in large quantities. Then, combining tools like GrabCut [55], we can obtain coarse object masks automatically. After that we can further manually filter and refine the masks, which can be done within an acceptable amount of human labor. Besides, to further save human labor, we can also make use of segmentation models like FCN [39] to generate segmentation masks automatically, although there is a trade-off: more automation leads to less accurate masks. See Fig. B.3 for some examples.

With the segmented foreground object, optionally, we can further create more diverse SC pairs by randomly augmenting (*e.g.*, cropping and scaling) the original background or replacing it with a background from another image. Fig. B.1 shows the overall pipeline. Here we set a hyperparameter $p \in (0, 1)$, the probability of transforming the original background, otherwise it is replaced by another background. Examples of SC pairs created through this pipeline are shown in Fig. B.2.

Style shift. For creating SC pairs under style shift, we use StableDiffusion v2 [53] to translate images from the photo domain of PACS into a different style. Given an example x , we add a mild level of Gaussian noise to the latent representation of x , and then remove the noise under the guidance of a text prompt. The prompt we use is “a minimalist drawing of a `class_name`, outline only, no texture” where `class_name` is the name of the class label of x . We choose this prompt because it produces the best visual quality among what we have explored. Finally, we decode the generated noise-free latent representation, producing the corresponding counterfactual example \tilde{x} . See Fig. B.4 for some examples.

C. Additional Implementation Details

C.1. Implementation Details of LAM

In the case of background shift, as shown in Algorithm 1, we use LAM to fine-tune a pretrained model ϕ which consists of a feature extractor f and a classification head g . As for training data, we start with a basic training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ with N examples. Following the method introduced in Appendix B, we generate a set of SC pairs $\tilde{\mathcal{D}}_0 = \{(x_i, \tilde{x}_i; y_i)\}_{i=1}^K$. Then we optionally replace the original backgrounds with random backgrounds to generate a more diverse set of SC pairs $\tilde{\mathcal{D}} = \{(x_i, \tilde{x}_i; y_i)\}_{i=1}^M$. With the SC pairs, we conduct linear-probing (LP) and then fine-tuning (FT) with the objective function given by Eq. (9) in Sec. 4.

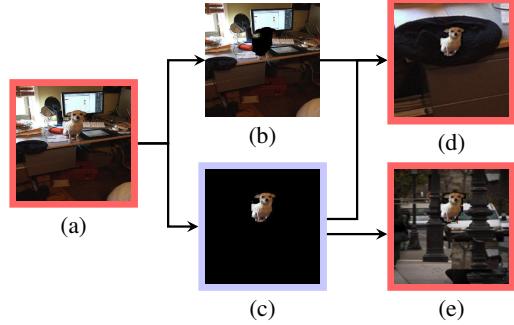


Figure B.1. An example of SC-pair creation against background shift. For every real example (a), we use a segmentation tool (*e.g.*, GrabCut [55]) to obtain its background (b) and foreground (c). (a) and (c) forms an SC pair. Optionally, we can either crop (b) and place (c) over the cropped (b) to obtain (d), or place (c) over random backgrounds from other images to obtain (e). In summary, (c) can form an SC pair with each red-boxed example.

C.2. Implementation Details of Baselines

In section 5.3, we have listed selected baselines to compare with LAM. Due to the diverse characteristics of the methods, the experimental settings are slightly different for each method. Tab C.1 and Tab C.2 summarize the settings.

C.3. Hyperparameter Setting and Other Details

Each experiment is conducted on a single Nvidia V100 GPU. Following [33], during LP, we set the learning rate to 0.003 for all experiments. During FT, we set the learning rate to 3e-5 for ViT on all datasets as well as ResNet50 on ImageNet-9 and PACS, and 3e-4 for the rest. All the experiments use cosine learning rate scheduling.

For ImageNet-9 and NICO datasets, the training batch size B is 128/64 during LP/FT respectively for all methods except the pair-based methods. For pair-based methods, the batch size is increased by B' for $B'/2$ SC pairs. Specifically, $B' = 256/32$ for both CoRe and MatchDG, while $B' = 256/64$ for LAM. For methods that require multiple training domains, the counterfactual examples are used to form another domain besides the original domain, while the batch size B is distributed according to the size of the two domains.

For PACS dataset, the training batch size B is 48 for all methods except the pair-based methods. For pair-based methods, the batch size is increased by B' for $B'/2$ SC pairs. Specifically, $B' = 32$. For methods that require multiple training domains, a training batch consists of even number of samples from 3 training domains. So the total batchsize is $16 \times 3 = 48$. Counterfactual examples are regarded as samples in Photo domain.

During LP, the model is trained for 10 epochs for all methods. During FT, the number is 20 epochs for ViT on

Algorithm 1: LAM under background shift

Input: pretrained model $\phi = g \circ f$, training set \mathcal{D} , number of LP iterations T_1 , number of FT iterations T_2 , probability p for SC-pair augmentation, hyperparameter λ of ℓ_{LAM} , batch size B for original examples, batch size B' for SC pairs;

Output: trained model ϕ' ;

$\tilde{\mathcal{D}}_0 \leftarrow \text{Create SC pairs ;}$

freeze f for LP;

for $t \leftarrow 1, \dots, T_1$ **do**

sampling B training examples from \mathcal{D} and $B'/2$

basic SC pairs from $\tilde{\mathcal{D}}_0$;

for every basic SC pair $(x_i, \tilde{x}_i; y_i)$ **do**

$z \leftarrow$ a sample from binomial $\mathcal{B}(1, p)$;

if $z = 1$ **then**

$b_i \leftarrow$ the background of x_i ;

$\tilde{b}_i \leftarrow b_i$ after transformation;

$\tilde{x}'_i \leftarrow \tilde{x}_i$ with new background \tilde{b}_i ;

else

$b_j \leftarrow$ a randomly sampled background;

$\tilde{x}'_i \leftarrow \tilde{x}_i$ with new background b_j ;

end

obtain SC pair $(\tilde{x}'_i, \tilde{x}_i; y_i)$;

end

train ϕ with the original examples and the SC pairs according to Eq. (9);

end

unfreeze f for FT;

for $t \leftarrow 1, \dots, T_2$ **do**

| repeat the same procedure as before during LP;

end

ImageNet-9 and NICO, and ResNet50 on ImageNet-9. For FT ResNet-50 on NICO, the number is 60 epochs. For FT ResNet-50 on PACS, the number is 40 epochs. All experiments are conducted with the random seed 3. The full hyperparameter setting is provided in Tab. C.3.

Background contexts in NICO. For NICO dataset, we only use contexts that are backgrounds (see Tab. C.4).

D. Additional Examples of Saliency Maps

We provide more examples of saliency maps visualized by GradCAM. The examples in Fig. D.1 and Fig. D.2 are all randomly sampled from *mixed-rand* of ImageNet-9, which is our OOD test set for ImageNet-9.

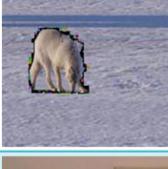
True label	Real example	Counterfactual example w/o background	Background	Counterfactual example with cropped original background	Counterfactual example with random background
Bird					
Wheeled vehicle					
Reptile					
Carnivore					
Musical instrument					

Figure B.2. More examples of SC pairs for ImageNet-9. The counterfactual examples are created by removing background or synthesizing cropped original background or randomly replaced with another background.

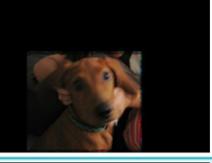
	Real example	Counterfactual example		
True label		Bounding box	Semantic segmentation	GrabCut
Dog				
Wheeled vehicle				

Figure B.3. Counterfactual examples created via different SC pair creation methods for ImageNet-9 under background shift. The bounding boxes are manually annotated, which are then used to produce the GrabCut results. The semantic segmentation masks are generated automatically. Note that there is a trade-off: more automation leads to less accurate masks.

True label	Real example	Counterfactual example	True label	Real example	Counterfactual example
Dog			Horse		
House			Giraffe		

Figure B.4. SC pairs created via StableDiffusion that generates counterfactual example from the real examples of the Photo domain in PACS dataset. The prompt we use is “a minimalist drawing of a `class_name`, outline only, no texture” where `class_name` is the name of the true class label.

Category	Method	Training data	Remark
Single-source	ERM		
	RSC	real examples	-
	SD		
Single-source pair-based	LAM	real examples	† Random backgrounds (in addition to the pure-black background) are used to generate the c.f. examples. Only 5% of the real examples are paired with a c.f. example.
	CoRe	+ c.f. examples [†]	
Multi-source pair-based	MatchDG		‡ An MUNIT [25] is learned to translate between real examples and the c.f. examples we provide. The c.f. examples generated by the MUNIT are then used instead. Every real example is paired with a c.f. example.
	MBDG	real examples	
Multi-source	RICE	+ c.f. examples (re-generated) [‡]	
	DANN	domain #1:	
	GDRO	real examples	-
	IRM	domain #2:	
	VREx	c.f. examples	

Table C.1. Implementation details of LAM and the baselines on ImageNet-9 and NICO under background shift. The category of the methods is explained in Tab. 1 of our paper. According to their categories, the training data are organized differently to meet the assumption of each individual method. On the other hand, the test data is the same for all methods: mixed-rand for ImageNet-9, and one unseen context per class for NICO. We use “c.f.” as a shorthand for “counterfactual”.

Category	Method	Training data	Remark
Single-source	ERM	mixed real examples from 3	
	RSC	domains (PCS, PAS, or PAC)	-
	SD	then mixed with c.f. examples	
Single-source pair-based	LAM	mixed real examples from 3	† Only 100 real examples per class in the Photo (P) domain are paired with a c.f. example.
	CoRe	domains (PCS, PAS, or PAC)	
Multi-source pair-based	MatchDG	+ c.f. examples [†]	‡ An MUNIT [25] is learned to translate across the 3 domains. The c.f. examples generated by the MUNIT are used instead. Every real example is paired with a c.f. example.
	MBDG	separate real examples from 3	
Multi-source	RICE	domains (PCS, PAS, or PAC)	
	DANN	+ c.f. examples (re-generated) [‡]	
	GDRO	separate real examples from 3	
	IRM	domains (PCS, PAS, or PAC)	-
	VREx	with c.f. examples mixed with	
		the Photo (P) domain	

Table C.2. Implementation details of LAM and the baselines on PACS under style shift. The category of the methods is explained in Tab. 1 of our paper. According to their categories, the training data are organized differently to meet the assumption of each individual method. Since PACS has four domains, we evaluate with three training-domain settings, namely PCS, PAS, and PAC; and see how the models generalize to the remaining domain in each case. We use “c.f.” as a shorthand for “counterfactual”.

Architecture	ViT		ResNet-50				
Dataset	ImageNet-9	NICO	ImageNet-9	NICO	PACS		
LAM	LP learning rate: 0.003 LP epochs: 10						
	FT learning rate: 3e-5 LP training batchsize: 128 LP sc-pair batchsize: 256 FT training batchsize: 64 FT sc-pair batchsize: 64	FT learning rate: 3e-4 LP training batchsize: 64 LP sc-pair batchsize: 64 FT training batchsize: 64 FT sc-pair batchsize: 64	FT learning rate: 3e-5 LP training batchsize: 48 LP sc-pair batchsize: 32 FT training batchsize: 48 FT sc-pair batchsize: 32				
	FT epochs: 20		FT epochs: 60	FT epochs: 40			
	p=0.9	p=0.7	p=0.9	N/A			
	lambda=10	lambda=0.5		lambda=0.2			
	sc-pair transform: RandomHorizontalFlip background transform: RandomCrop	sc-pair transform: RandomHorizontalFlip	sc-pair transform: RandomHorizontalFlip background transform: RandomCrop				
	LP learning rate: 0.003 LP epochs: 10						
CoRe/ MatchDG	FT learning rate: 3e-5 LP training batchsize: 128 LP sc-pair batchsize: 256 FT training batchsize: 32 FT sc-pair batchsize: 32		FT learning rate: 3e-4 LP training batchsize: 64 LP sc-pair batchsize: 64 FT training batchsize: 64 FT sc-pair batchsize: 64	FT learning rate: 3e-5 LP training batchsize: 48 LP sc-pair batchsize: 32 FT training batchsize: 48 FT sc-pair batchsize: 32			
	LP learning rate: 0.003 LP epochs: 10 Training batchsize: 128						
Other methods	FT learning rate: 3e-5	FT learning rate: 3e-4	FT learning rate: 3e-5				

Table C.3. Hyperparameter setting for all the main experiments. Background transformation refers to the transformation applied to background when generating counterfactual examples. SC-pair transformation refers to the transformation applied to both real example and counterfactual example of an SC pair. Regular data augmentations (RandomCrop, RandomHorizontalFlip) are applied to all unpaired examples for all methods. If not specified otherwise, for CoRe and MatchDG, the method-specific hyperparameters follow the default setting in the corresponding papers. Pair related hyperparameters follow setting of LAM. For other methods, method-specific hyperparameters follow the default setting in DomainBed [21].

Class	Contexts
monkey	in forest, on snow, on grass, in cage, on beach, in water
sheep	on snow, on road, on grass, in water, in forest, at sunset, aside people
rat	on snow, on grass, in water, in hole, in forest, in cage, at home
horse	on snow, on grass, on beach, in street, in river, in forest, at home, aside people
elephant	on snow, on grass, in zoo, in street, in river, in forest, in circus
dog	on snow, on grass, on beach, in water, in street, in cage, at home
cow	on snow, on grass, in river, in forest, at home, aside people
cat	on tree, on snow, on grass, in water, in river, in street, in cage, at home
bird	on shoulder, on ground, on grass, on branch, in water, in hand, in cage
bear	on tree, on snow, on ground, in water, in forest, eating grass
truck	on snow, on road, on grass, on bridge, on beach, in sunset, in race, in forest, in city, aside mountain
train	subway, on snow, on bridge, on beach, in sunset, in forest, cross tunnel, at station, aside mountain
motorcycle	with people, on track, on snow, on road, on grass, on beach, in sunset, in street, in garage, in city
helicopter	with people, on snow, on sea, on grass, on beach, in sunset, in forest, in city, at heliport, aside mountain
car	with people, on track, on snow, on road, on bridge, on booth, on beach, in sunset, in city, at park
bus	with people, on snow, on bridge, in city, at yard, at station, aside tree, aside traffic light
boat	with people, on beach, in sunset, in river, in city, cross bridge, at wharf
bicycle	with people, velodrome, shard, on snow, on road, on grass, on beach, in sunset, in street, in garage

Table C.4. NICO contexts used in our experiments.

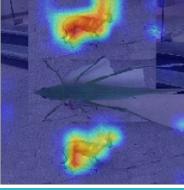
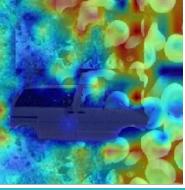
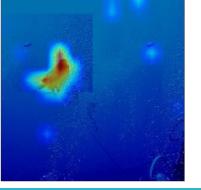
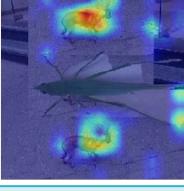
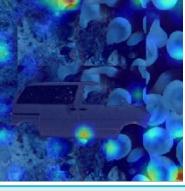
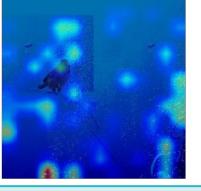
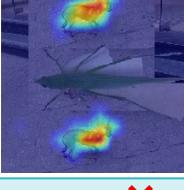
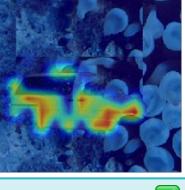
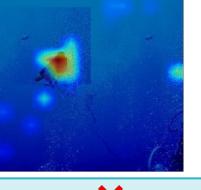
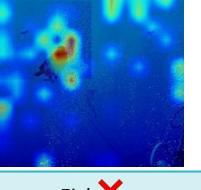
True label	Insect	Dog	Wheeled vehicle	Bird
Original example				
ERM				
prediction	Carnivore 	Dog 	Fish 	Carnivore 
CoRe				
prediction	Carnivore 	Dog 	Fish 	Fish 
MatchDG				
prediction	Carnivore 	Dog 	Wheeled vehicle 	Fish 
LAM(ours)				
prediction	Insect 	Bird 	Wheeled vehicle 	Fish 

Figure D.1. Non-cherry-picked GradCAM heatmaps of LAM compared with those of ERM, CoRe and MatchDG.

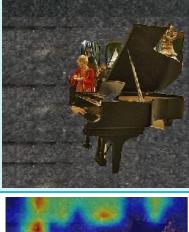
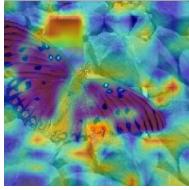
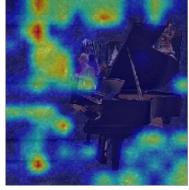
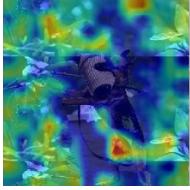
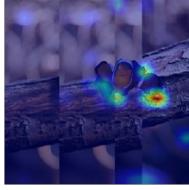
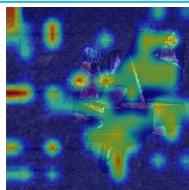
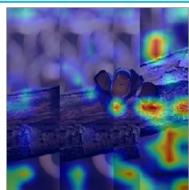
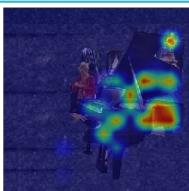
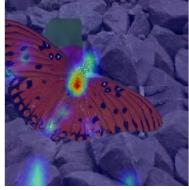
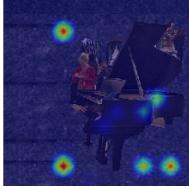
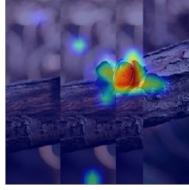
True label	Insect	Musical Instrument	Dog	Fish
Original example				
ERM				
prediction	Insect 	Musical Instrument 	Primate 	Insect 
CoRe				
prediction	Insect 	Musical Instrument 	Reptile 	Reptile 
MatchDG				
prediction	Insect 	Musical Instrument 	Carnivore 	Fish 
LAM(ours)				
prediction	Insect 	Musical Instrument 	Dog 	Fish 

Figure D.2. Non-cherry-picked GradCAM heatmaps of LAM compared with those of ERM, CoRe and MatchDG.