Are Data-driven Explanations Robust against Out-of-distribution Data?

Tang Li   Fengchun Qiao   Mengmeng Ma   Xi Peng
University of Delaware
{tangli, fengchun, mengma, xipeng}@udel.edu

Abstract

As black-box models increasingly power high-stakes applications, a variety of data-driven explanation methods have been introduced. Meanwhile, machine learning models are constantly challenged by distributional shifts. A question naturally arises: Are data-driven explanations robust against out-of-distribution data? Our empirical results show that even though predict correctly, the model might still yield unreliable explanations under distributional shifts. How to develop robust explanations against out-of-distribution data? To address this problem, we propose an end-to-end model-agnostic learning framework Distributionally Robust Explanations (DRE). The key idea is, inspired by self-supervised learning, to fully utilizes the inter-distribution information to provide supervisory signals for the learning of explanations without human annotation. Can robust explanations benefit the model's generalization capability? We conduct extensive experiments on a wide range of tasks and data types, including classification and regression on image and scientific tabular data. Our results demonstrate that the proposed method significantly improves the model's performance in terms of explanation and prediction robustness against distributional shifts.

1. Introduction

There has been an increasing trend to apply *black-box* machine learning (ML) models for high-stakes applications. The lack of explainability of models can have severe consequences in healthcare [48], criminal justice [61], and other domains. Meanwhile, ML models are inevitably exposed to unseen distributions that lie outside their training space [28, 56]; a highly accurate model on average can fail catastrophically on *out-of-distribution* (OOD) data due to naturally-occurring variations, sub-populations, spurious correlations, and adversarial attacks. For example, a cancer detector would erroneously predict samples from hospitals having different data acquisition protocols or equipment
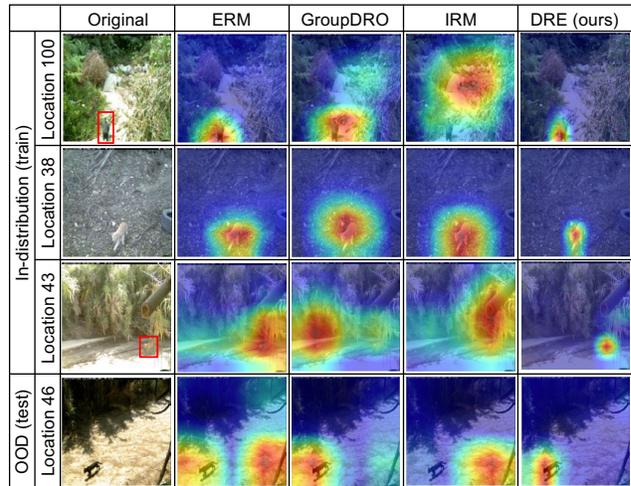

Figure 1. The explanations for *in-* and *out-of-distribution* data of *Terra Incognita* [4] dataset. Note that GroupDRO [50] and IRM [2] are explicitly designed methods that can predict accurately across distributions. Although with correct predictions, the explanations of models trained by such methods would also highlight distribution-specific associations (*e.g.*, tree branches) except the object. This leads to unreliable explanations on OOD data. On the contrary, our model consistently focuses on the most discriminative features shared across distributions.

manufacturers. Therefore, reliable explanations across distributions are crucial for the safe deployment of ML models. However, existing works focus on the reliability of data-driven explanation methods [1, 64] while ignoring the robustness of explanations against distributional shifts.

A question naturally arises: *Are data-driven explanations robust against out-of-distribution data?* We empirically investigate this problem across different methods. Results of the Grad-CAM [51] explanations are shown in Fig. 1. We find that *the distributional shifts would further obscure the decision-making process due to the black-box nature of ML models.* As shown, the explanations focus not only on the object but also spurious factors (*e.g.*, background pixels). Such distribution-specific associations would yield *inconsistent* explanations across distributions.

---

[1]The source code and pre-trained models are available at: https://github.com/tangli-udel/DRE.

Eventually, it leads to unreliable explanations (*e.g.*, tree branches) on OOD data. This contradicts with human prior that the most discriminative features ought to be invariant.

How to develop robust explanations against out-of-distribution data? Existing works on OOD generalization are limited to data augmentation [42, 52, 59], distribution alignment [3, 16, 31], Meta learning [12, 29, 40], or invariant learning [2, 26]. However, without constraints on explanations, the model would still recklessly absorb all associations found in the training data, including *spurious correlations*. To constrain the learning of explanations, existing methods rely on explanation annotations [44, 54] or one-to-one mapping between image transforms [8, 19, 39]. However, there is no such mapping in general *naturally-occurring* distributional shifts. Furthermore, obtaining ground truth explanation annotations is prohibitively expensive [60], or even impossible due to subjectivity in real-world tasks [46]. To address the aforementioned limitations, we propose an end-to-end model-agnostic training framework *Distributionally Robust Explanations (DRE)*. The key idea is, inspired by self-supervised learning, to fully utilize the inter-distribution information to provide supervisory signals for explanation learning.

Can robust explanations benefit the model's generalization capability? We evaluate the proposed methods on a wide range of tasks in Sec. 4, including the classification and regression tasks on image and scientific tabular data. Our empirical results demonstrate the robustness of our explanations. The explanations of the model trained via the proposed method outperform existing methods in terms of explanation consistency, fidelity, and scientific plausibility. The extensive comparisons and ablation studies prove that our robust explanations significantly improve the model's prediction accuracy on OOD data. As shown, the robust explanations would alleviate the model's excessive reliance on *spurious correlations*, which are unrelated to the causal correlations of interest [2]. Furthermore, the enhanced explainability can be generalized to a variety of data-driven explanation methods.

In summary, our main contributions:

- We comprehensively study the robustness of data-driven explanations against naturally-occurring distributional shifts.

- We propose an end-to-end model-agnostic learning framework Distributionally Robust Explanations (DRE). It fully utilizes inter-distribution information to provide supervisory signals for explanation learning without human annotations.

- Empirical results in a wide range of tasks including classification and regression on image and scientific tabular data demonstrate superior explanation and prediction robustness of our model against OOD data.

2. Related work

Explainable machine learning. A suite of techniques has been proposed to reveal the decision-making process of modern *black-box* ML models. One direction is intrinsic to the model design and training, rendering an explanation along with its output, *e.g.*, attention mechanisms [58] and joint training [7, 23]. A more popular way is to give insight into the learned associations of a model that are not readily interpretable by design, known as post-hoc methods. Such methods usually leverage backpropagation or local approximation to offer saliency maps as explanations, *e.g.*, Input Gradient [53], Grad-CAM [51], LIME [43], and SHAP [32]. Recent works have shed light on the downsides of post-hoc methods. The gradient-based explanations (*e.g.*, Input Gradient) are consistent with sample-based explanations (*e.g.*, LIME) with comparable fidelity but have much lower computational cost [47]. Moreover, only the Input Gradient and Grad-CAM methods passed the sanity checks in [1]. In our work, we incorporate gradient-based methods into optimization to calculate explanations efficiently.

Out-of-distribution generalization. To generalize machine learning models from training distributions to unseen distributions, existing methods on OOD generalization can be categorized into four branches: (1) Data augmentation. [33, 36, 52, 59] enhance the generalization performance by increasing the diversity of data through augmentation. (2) Distribution alignment. [3, 31, 41] align the features across source distributions in latent space to minimize the distribution gaps. (3) Meta learning. [12, 29, 34, 35] using meta-learning to facilitate fast-transferable model initialization. (4) Invariant learning. [2, 26, 37] learn invariant representations that are general and transferable to different distributions. However, recent works [18, 25] show that the classic empirical risk minimization (ERM) [57] method has comparable or even outperforms the aforementioned approaches. We argue that this is because the existing approaches barely have constraints on explanations, the model would still recklessly absorb any correlations identified in the training data.

Explanation-guided learning. Several recent works have attempted to incorporate explanations into model training to improve predictive performance. [44, 54] match explanations with human annotations based on domain knowledge, to alleviate the model's reliance on background pixels. [19, 39, 60] align explanations between spatial transformations to improve image classification and weakly-supervised segmentation. [8, 10, 20] synchronize the explanations of the perturbed and original samples to enhance the robustness of models. However, acquiring ground truth explanations is prohibitively labor-intensive [60] or even impossible due to subjectivity in real-world tasks [46]. Furthermore, image transformations are insufficient to address different data types and the general *naturally-occurring* dis-

tributional shifts, there is no one-to-one correlation between samples from different distributions to provide supervision.

3. Methods

3.1. Robustness against Out-of-distribution Data

Question: *Are data-driven explanations robust against out-of-distribution data?*

We empirically investigate this problem on an out-of-distribution generalization benchmark image dataset *Terra Incognita* [4] and a scientific tabular dataset *Urban Land* [17]. For image data, we leverage the Grad-CAM [51] method to generate explanations, and leverage explanation fidelity [38] as the metric to measure explanation quality. Specifically, we evaluate ERM [57] and two representative out-of-distribution generalization methods, Group-DRO [50] and IRM [2]. For scientific tabular data, we leverage the Input Gradient [53] method to generate explanations, and leverage scientific consistency (Sec. 4.2) as the metric to measure explanation quality. Qualitatively, the more explanation plausibility on OOD data, the better the explanation robustness. Quantitatively, the higher the explanation fidelity or scientific consistency on OOD data, the better the explanation robustness.

For qualitative evaluation, we observe that even with correct predictions, the explanations on out-of-distribution data might still focus on spurious correlations. Fig. 1 shows the Grad-CAM explanations of models trained via ERM, GroupDRO, and IRM. As shown, the explanations on in-distribution data would not only highlight the object, but also distribution-specific associations (*e.g.*, background pixels). This eventually leads to unreliable explanations on OOD data (*e.g.*, tree branches). We find that the out-of-distribution generalization methods perform even worse than the classic ERM in terms of explanation robustness. This finding verify the results in [18, 25] that ERM outperforms the out-of-distribution generalization methods from an explanation perspective. For quantitative evaluation, we empirically verify that the explanation fidelity and scientific consistency severely dropped on OOD data. Tab. 1 reports the evaluation results, as shown, on the average of each distribution as the testing set, the explanation fidelity dropped by 24.9%, 20.8%, and 15.0% for ERM, GroupDRO, and IRM on *Terra Incognita*; the scientific consistency dropped by 32.4% for ERM on *Urban Land*.

To conclude, *data-driven explanations are not robust again out-of-distribution data.* The distributional shifts further obscure the decision-making process due to the *black-box* nature of modern ML models.

3.2. Distributionally Robust Explanations

Question: *How to develop robust explanations against out-of-distribution data?*

Dataset	Method	Evaluation*		$\Delta \downarrow$
		ID	OOD	
Terra Incognita [4]	ERM [57]	0.778	0.584	**0.194**
	GroupDRO [50]	0.742	0.597	0.145
	IRM [2]	0.612	0.520	0.092
Urban Land [17]	ERM [57]	0.535	-0.113	**0.648**

Table 1. Evaluation of the explanation quality of *in-distribution* (ID) and *out-of-distribution* data. The results are on the average of each distribution as the testing set. Note that the explanation quality of both image and scientific data severely dropped on OOD data. *The reported evaluation scores are explanation fidelity [38] (*Terra Incogenita*) and scientific consistency (*Urban Land*).

We started by providing the formulation of the problem, then introduce a new framework *Distributionally Robust Explanations* (DRE) for explanation learning.

3.2.1 Problem formulation

The objective of typical supervised learning is to learn a predictor $f \in \mathcal{F}$ such that $f(x) \to y$ for any $(x, y) \sim P(X, Y)$, where $P(X, Y)$ is an unknown joint probability distribution, and $\mathcal{F}$ is a function class that is model-agnostic for a prediction task. However, in the scenario of *out-of-distribution* generalization, one can not sample directly from $P(X, Y)$ due to distributional shifts. Instead, it is assumed that we can only measure (X, Y) under different environmental conditions e so that data is drawn from a set of groups $\mathcal{E}_{\text{all}}$ such that $(x, y) \sim P_e(X, Y)$. For example, in the cancer detection task, the environmental conditions denote the latent factors (*e.g.*, data acquisition protocols or equipment manufacturers) that underlie different hospitals.

Informally, this assumption specifies that there should exist a function G that relates the random variables X and X^e via $G(X, e) = X^e$. Let $\mathcal{E}_{\text{train}} \subsetneq \mathcal{E}_{\text{all}}$ be a finite subset of training groups (distributions), given the task-specific objective function ℓ and explanation method $g(\cdot)$, our *Distributionally Robust Explanations* is equivalent to the following constrained problem:

$$
\begin{aligned}
\min_{f \in \mathcal{F}} \quad & \mathcal{R}(f) := \mathbb{E}_{(x,y) \sim P_{\text{train}}}[\ell(f(x), y)] \\
\text{s.t.} \quad & g(x) = g(G(x, e)) \quad \forall e \in \mathcal{E}_{\text{all}}.
\end{aligned}
\tag{DRE}
$$

Intuitively, we encourage the model to have invariant explanations for a sample under different environmental conditions (distributions) after optimization.

The problem in DRE is challenging to solve, since we do not have access to the set of all distributions $\mathcal{E}_{\text{all}}$ or the underlying distribution $P(X, Y)$, and to the distribution transformation function G. The alternative solutions would be: (i) acquire the ground truth explanations for all samples; (ii) obtain the one-to-one mapping between samples from
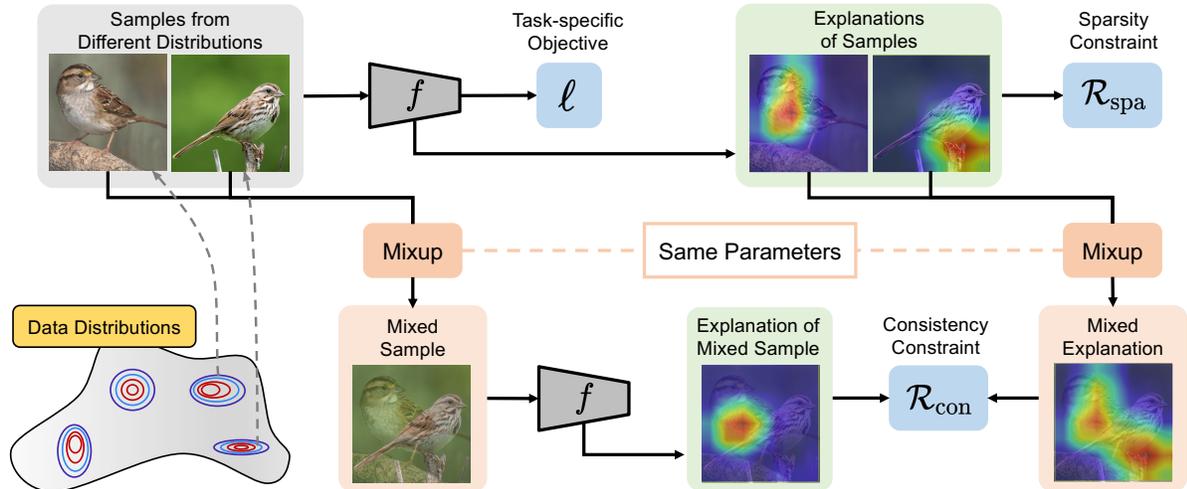
Figure 2. **Overview of the proposed Distributionally Robust Explanations (DRE) method.** Our method consists of distributional explanation consistency constraint and explanation sparsity constraint. Firstly, we load a batch of random samples from different distributions and feed them into the model to calculate the standard task-specific objective (*e.g.*, Cross-entropy loss for classification tasks). Then, we calculate the explanations of each sample *w.r.t.* their predictions and constrain the explanation sparsity. Next, we pair up the samples with the same prediction but from different distributions, mixing up both samples and explanations using the same parameters. We feed the mixed samples into the model and calculate the explanations *w.r.t.* their shared predictions. Finally, we constrain the consistency between the explanation of mixed samples and the mixed explanations. This figure is best viewed in color.

different distributions, such as original sample and its corresponding corrupted ones [22]. However, as our discussion in Sec. 2, the ground truth explanations and one-to-one mappings are practically unavailable in real-world tasks.

3.2.2 Distributional Explanation Consistency

We address these challenges by explicitly designing the *distributional explanation consistency*. The key idea is, inspired by self-supervised learning, leveraging the mixed explanation to provide supervisory signals for the learning of explanations. Specifically, we first leverage the *distributional mixup* to achieve a simple but effective interdistributional transformation. The *mixup* [63] methods have been empirically shown to substantially improve performance and robustness to adversarial noise [40, 62]. In contrast to original *mixup* that mixes random pairs of samples and labels, we mix up samples with the same ground truth but from different distributions. Denote $(\mathbf{x}_e, \mathbf{x}_{e'})$ as random pairs of training samples with the same ground truth $\mathbf{y}$ but from different distributions, then our *mixup* operator can be defined as:

$$\mathcal{M}(\mathbf{x}_e, \mathbf{x}_{e'}) = \tau\mathbf{x}_e + (1-\tau)\mathbf{x}_{e'} \tag{1}$$

where $\tau \sim \text{Beta}(\alpha, \alpha)$ and the *mixup* hyper-parameter $\alpha \in (0, +\infty)$ controls the interpolation strength, in practice we use $\alpha = 0.2$. Secondly, we *mixup* the explanations of the samples using the same parameters, namely:

$$\mathcal{M}(g(\mathbf{x}_e), g(\mathbf{x}_{e'})) = \tau g(\mathbf{x}_e) + (1-\tau)g(\mathbf{x}_{e'}) \tag{2}$$

Thirdly, we feed the mixed samples into the model to calculate the explanations $g(\mathcal{M}(\mathbf{x}_e, \mathbf{x}_{e'}))$. Lastly, denote $\mathcal{D}$ as an arbitrary discrepancy metric, for example, ℓ_1 distance and KL-divergence [27], for all $e, e' \in \mathcal{E}_{\text{train}}$, we leverage the consistency between the mixed explanation and the explanation of the mixed sample as an alternative of DRE:

$$\min_{f \in \mathcal{F}} \mathcal{R}(f) \text{ s.t. } \mathcal{D}[g(\mathcal{M}(\mathbf{x}_e, \mathbf{x}_{e'})), \mathcal{M}(g(\mathbf{x}_e), g(\mathbf{x}_{e'}))] \leq \epsilon \tag{3}$$

Intuitively, the mixed explanation serves as the pseudo label to guide the learning of the explanation for the mixed sample. Note that $g(\cdot)$ is not restrictive, including any gradient-based explanation methods, such as Grad-CAM [51] and Input Gradient [53]. We leverage Karush–Kuhn–Tucker conditions [6] and introduce a Lagrange multiplier λ to convert the constrained problem in Eq. 3 into its unconstrained counterpart:

$$\min_{f \in \mathcal{F}} \{\mathcal{R}_{\text{con}}(f) := \mathbb{E}_{(x,y) \sim P_{\text{train}}}[\ell(f(x), y)]$$
$$+ \lambda\mathcal{D}[g(\mathcal{M}(\mathbf{x}_e, \mathbf{x}_{e'})), \mathcal{M}(g(\mathbf{x}_e), g(\mathbf{x}_{e'}))]\} \tag{4}$$

Explanation Regularization Our empirical results (Tab. 6) show that recklessly optimizing Eq. 4 could easily fall into a local minimum. For example, explanations that evenly attribute the prediction to all features in the sample would be a trivial solution to satisfy the constraint. To address this problem, we propose to further regularize the ℓ_1-norm of the explanations. Let γ be a dual variable, our

Algorithm 1: The proposed Distributionally Robust Explanations (DRE).

Input: Data of $\mathcal{E}_{\text{train}}$; Step size η
Output: Learned model parameters θ
1 **while** *not converged* **do**
2   Sample $(\mathbf{x}_e, \mathbf{y}) \sim P_e(X, Y) \ \forall e \in \mathcal{E}_{\text{train}}$
3   Sample $(\mathbf{x}_{e'}, \mathbf{y}) \sim P_{e'}(X, Y) \ \forall e' \in \mathcal{E}_{\text{train}}$
4   Calculate $\mathcal{R}(f)$ via Eq. 5
5   Update θ via $\theta^{t+1} = \theta^t - \eta^t \nabla \mathcal{R}(f)$
6 **end**

---

overall objective is formulated as follows:

$$\min_{f \in \mathcal{F}} \{ \mathcal{R}(f) := \mathcal{R}_{\text{con}}(f) + \gamma[||g(\mathbf{x}_e)||_1 + ||g(\mathbf{x}_{e'})||_1] \} \quad (5)$$

The entire training pipeline is summarized in Alg. 1. Our *Distributional Robust Explanations* has the following merits. First, in contrast to existing works that rely on expensive explanation annotations [44, 54] or one-to-one mapping between image transforms [19, 39, 60], the proposed method provides a supervisory signal for the explanation learning in general distributional shifts. Second, the proposed method fully utilizes the inter-distribution information which guarantees distributional invariance in a more continuous latent space. Third, the proposed method does not involve additional parameters, it can be considered as an add-on module for training a robust model without changing the model architecture. This enables DRE for robust explanation and prediction against distributional shifts as shown in Sec. 4.

3.3. Improve Model's Generalization Capability

Question: *Can robust explanations benefit the model's generalization capability?*

We empirically evaluate this problem on two out-of-distribution generalization benchmark image datasets (*Terra Incognita* [4] and *VLCS* [14]) and a scientific tabular dataset (*Urban Land* [17]). Tab. 2 shows the results of prediction performance on OOD data. For image datasets, on average of each distribution as the testing set, our method outperforms the ERM [57] results by 6.9% and 2.0% in terms of prediction accuracy. For scientific tabular data, on average of each continental region as the testing set, our method significantly outperforms the baseline results by 18.5% in terms of the prediction residual.

Therefore, *the robust explanations against OOD data can benefit the model's generalization capability*. Recklessly minimizing training error would leads machines to absorb all the correlations found in training data [2]. We argue that our robust explanations alleviate the excessive reliance of the model on *spurious correlations*. Our method constrains the model to rely on invariant causal correlations and leads to better generalization capability.

Dataset	Method	Evaluation*
Terra Incognita [4]	ERM	46.1%
	DRE (ours)	**53.0%**
VLCS [14]	ERM	77.5%
	DRE (ours)	**79.5%**
Urban Land [17]	ERM	9.70e-4
	DRE (ours)	**7.91e-4**

Table 2. Evaluation of predictive performance on *out-of-distribution* data. The results are on the average of each distribution as the testing set. Note that our method outperforms ERM on *Terra Incogenita*, *VLCS*, and *Urban Land*. *The reported evaluation scores are Accuracy (*Terra Incogenita*, *VLCS*) and Prediction Residual (*Urban Land*).

4. Experiments

To best validate the performance, we conduct a series of experiments to compare our DRE method with existing methods. The experimental results prove that our method achieves superior explanation and prediction robustness against *out-of-distribution* data on a wide scope of tasks, including classification and regression tasks on image and scientific tabular data.

4.1. Datasets and Implementation Details

Datasets. We validate our method on two OOD generalization benchmark image datasets for classification and a scientific tabular dataset for regression. (1) *Terra Incognita* [4] ($\approx$ 11K images, 10 classes) consists of four sub-datasets: Location 100, Location 38, Location 43, and Location 46. Each sub-dataset indicates a camera trap location in the wild and can be viewed as a different distribution. Each image in these datasets contains one single animal category (*e.g.,* coyote) with different illumination, backgrounds, perspective, etc. (2) *VLCS* [14] ($\approx$ 25K images, 5 classes) consists of four sub-datasets: Caltech101 [15], LabelMe [49], SUN09 [9], and VOC2007 [13]. Each sub-dataset can be viewed as a different distribution. Each image in these datasets contains one single image category (*e.g.,* car) with different styles and backgrounds. (3) *Global National Total Amounts of Urban Land, SSP-Consistent Projections and Base Year, v1 (2000 - 2100)* [17] (hereinafter referred to as *Urban Land*). The dataset is used for urban land prediction, and the global land area has been divided into 997,022 grid cells. Each grid cell contains nine topographic, population, and historical urban fraction attributes. The task is to predict the urban fraction in the year 2010. The world is been divided into nine continental regions, each region can be viewed as a different distribution.

Implementation details. For all datasets, we alternately leave one distribution out as the testing set. We split the data

Metric	Method	Terra Incognita [4]					VLCS [14]				
		Loc.100	Loc.38	Loc.43	Loc.46	Avg.	Caltech101	LabelMe	SUN09	VOC2007	Avg.
DEC loss ↓	ERM [57]	1.014	0.971	0.998	1.016	1.000	0.991	0.902	0.987	1.120	1.000
	GroupDRO [50]	0.929	0.996	1.072	0.969	0.992	0.947	0.907	1.026	0.942	0.956
	IRM [2]	0.982	0.965	0.982	0.939	0.967	0.918	0.874	0.947	0.928	0.918
	Mixup [62]	0.998	0.926	1.032	0.970	0.982	0.942	1.014	1.072	0.947	0.995
	CGC [39]	0.382	0.412	0.396	0.356	0.387	1.000	0.915	0.992	0.918	0.956
	DRE (ours)	**0.195**	**0.250**	**0.213**	**0.226**	**0.221**	**0.061**	**0.431**	**0.421**	**0.229**	**0.286**
iAUC ↑	ERM [57]	0.517	0.644	0.614	0.560	0.584	0.964	0.787	**0.765**	0.758	0.819
	GroupDRO [50]	**0.678**	0.578	0.608	0.525	0.597	0.928	0.786	0.643	0.706	0.766
	IRM [2]	0.489	0.651	0.438	0.500	0.520	0.939	0.772	0.613	0.715	0.760
	Mixup [62]	0.535	0.471	0.488	0.450	0.486	0.924	0.767	0.607	0.641	0.735
	CGC [39]	0.238	0.250	0.380	0.322	0.298	0.804	0.581	0.524	0.618	0.632
	DRE (ours)	0.651	**0.685**	**0.633**	**0.605**	**0.644**	**0.973**	**0.837**	0.752	**0.763**	**0.840**
Acc (%) ↑	ERM [57]	49.8	42.1	56.9	35.7	46.1	97.7	64.3	73.4	74.6	77.5
	GroupDRO [50]	41.2	38.6	56.7	36.4	43.2	97.3	63.4	69.5	76.7	76.7
	IRM [2]	54.6	39.8	56.2	39.6	47.6	**98.6**	64.9	73.4	77.3	78.5
	Mixup [62]	59.6	42.2	55.9	33.9	47.9	98.3	64.8	72.1	74.3	77.4
	CGC [39]	51.8	44.6	54.9	39.8	47.8	97.1	63.2	73.6	70.6	76.1
	DRE (ours)	**64.1**	**48.1**	**57.1**	**42.8**	**53.0**	98.3	**65.5**	**73.8**	**80.2**	**79.5**

Table 3. Comparison of the *out-of-distribution* explanation and prediction performance on *Terra Incognita* [4] and *VLCS* [14] datasets. The models are tested on the specified distribution and trained on all other distributions. Our results are on the average of three trials of experiments. We highlight the **best results** and the second best results. Note that the Acc (%) numbers for ERM [57], GroupDRO [50], IRM [2], and Mixup [62] are from Gulrajani and Lopez-Paz [18].

Metric	Method	Africa	E. Asia	Europe	N. Africa	N. America	Oceania	Russia	S. America	S. Asia	Avg.
DEC loss ↓	ERM [57]	3.69e-1	2.25e-1	9.98e-1	1.43e+0	2.14e+1	2.27e+0	5.56e-1	5.58e-1	4.54e-1	1.00e+0
	DRE (ours)	**2.52e-1**	**2.11e-1**	**9.20e-2**	**1.12e-3**	**4.86e-6**	**7.25e-1**	**5.17e-6**	**6.38e-6**	**1.10e-1**	**1.55e-1**
SC ↑	ERM [57]	0.810	**0.779**	0.100	-0.707	-0.653	0.855	-0.856	-0.837	-0.504	-0.113
	DRE (ours)	**0.967**	**0.779**	**0.760**	**0.135**	**0.189**	**0.894**	**0.130**	**0.058**	**0.324**	**0.471**
Prediction Residual ↓	ERM [57]	7.45e-4	1.46e-3	2.43e-3	7.10e-4	5.40e-4	**1.53e-4**	8.04e-5	3.55e-4	2.26e-3	9.70e-4
	DRE (ours)	**6.58e-4**	**1.16e-3**	**1.95e-3**	**5.74e-4**	**4.14e-4**	1.56e-4	**6.34e-5**	**2.94e-4**	**1.85e-3**	**7.91e-4**

Table 4. Comparison of the *out-of-distribution* explanation and prediction performance for short-term urbanization estimation (2000–2010) on the *Urban Land* [17] dataset. Our results are on the average of three trials of experiments. Note that a residual of 0.01 indicates a one-percentage point difference between the estimated and observed built-up land fractions.

from each training distribution into 80% and 20% splits, and use the larger splits for training, the smaller splits for validation and model selection. All models are trained using Adam [24] optimizer for 5,000 steps. (1) For both of the image datasets, following the setup in [18], we use a ResNet50 [21] model pretrained on ImageNet [11] and fine-tune. We freeze all batch normalization layers before fine-tuning and insert a dropout layer before the final linear layer. We crop the images of random size and aspect ratio, resizing to 224 × 224 pixels, random horizontal flips, random color jitter, grayscaling the image with 10% probability, and normalization using the ImageNet channel statistics. We use learning rate=5e-5 and batch size=16. (2) For scientific data, following the setup in [30], we transform the tabular data into image-like data with each grid cell as a pixel. With each land pixel as the center pixel, we densely sampled 997,022 images with size 16 × 16. We use a standard U-Net [45] model training from scratch. We augment images by random vertical and horizontal flips, rotation by 90°, 180°, 270°. We use learning rate=1e-4 and batch size=256.

4.2. Evaluation Metrics

Distributional explanation consistency (DEC). We evaluate the explanations using the same explanation consistency we used for our model training. Note that although we optimized the in-distribution explanation consistency, the explanation consistency on *out-of-distribution* data is still an important metric to evaluate our goal. We mix up the OOD samples and their explanations with in-distribution samples and their explanations to calculate the DEC loss. For better comparison, we set the average DEC loss for the model trained via standard ERM [57] as 1.0 and rescale others correspondingly. We expect the DEC loss to be lower for the robust explanations against *out-of-distribution* data.
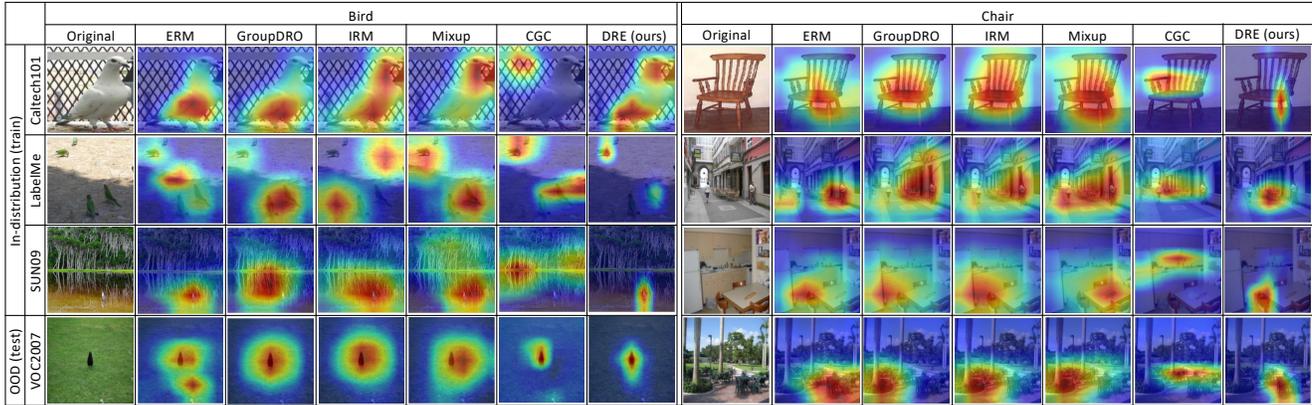
Figure 3. Grad-CAM explanations for images from *Bird* (left) and *Chair* (right) classes in *VLCS* dataset. The model trained via existing methods, such as ERM [57], Mixup [62], and CGC [39], not only focuses on the objects, but also distribution-specific associations, it getting even severe on OOD data. On the contrary, our model alleviates the reliance on *spurious correlations* (*e.g.*, background pixels), and makes consistent explanations on OOD data. This figure is best viewed in color.

Explanation fidelity (iAUC). This metric measures an increase in the probability of the predictive score as more and more pixels are introduced in a descending order of importance, where the importance is obtained from the explanation [5, 38]. We measure the area under the insertion curve (iAUC), which accumulated the probability increase while gradually inserting the features from the original input into a reference input. The reference input we used is an initially blurred canvas, since the blurring takes away most of the finer details of an image without exposing it to sharp edges which might introduce spurious correlations. To compare between models, we normalize the logit of the correct class to 1. We expect the iAUC score to be higher for a robust explanation against OOD data.

Scientific consistency (SC). We introduce this metric for the explanations of scientific data. Scientific consistency means the explanations obtained are plausible and consistent with existing scientific principles [46]. We leverage explanations of domain expert verified models as the ground truth domain knowledge. We use it as a posteriori by cosine similarity between feature importance obtained via explanations and domain knowledge. We expect the SC score to be higher for a robust explanation against distributional shifts.

4.3. Evaluation on Terra Incognita

We compare our model with models trained via four representative OOD generalization methods (*i.e.*, ERM [57], GroupDRO [50], IRM [2], and Mixup [62]), and a state-of-the-art explanation-guided learning method (*i.e.*, CGC [39]). We use Grad-CAM [51] to generate explanations after model training. Quantitatively, in Tab. 3 we report the results of explanation quality and predictive performance on OOD data. Our model outperforms the four OOD generalization methods and CGC [39] with significant im-

provement on all three metrics. On average of each distribution as the testing set, our method outperforms the second-best results by 42.9%, 7.9%, and 5.1% in terms of DEC loss, explanation fidelity, and prediction accuracy. Qualitatively, in Fig. 1 we visualize the Grad-CAM explanations of ERM, GroupDRO, IRM, and our model using Location 46 as the testing distribution. Our explanations corrected the wrongly focused explanations generated by the models of existing methods. Rather than backgrounds, our explanations are more concentrated on the most discriminative object. The results demonstrate the superiority of our model on explanation and prediction robustness against OOD data.

4.4. Evaluation on VLCS

Following the same settings in Sec. 4.3, we compare our model with ERM [57], GroupDRO [50], IRM [2], Mixup [62]), and CGC [39] models. Quantitatively, as shown in Tab. 3, our model outperforms existing methods with significant improvement on all three metrics. On average of each distribution as the testing set, our method outperforms the second-best results by 68.8%, 2.6%, and 1.0% in terms of DEC loss, explanation fidelity, and prediction accuracy. Qualitatively, in Fig. 3 we visualize the Grad-CAM explanation of five existing methods and our method. Our explanations are more concentrated on the most discriminative features of the object, and significantly alleviate the focus of background pixels on OOD data. Furthermore, we evaluate the efficiency of our method shown in Tab. 5.

4.5. Evaluation on Urban Land

We compare our model with models trained via ERM [57]. We leverage the Input Gradient [53] to generate explanations after model training, because of its fine-grained resolution and advanced explanation performance

Method	# of params.	Training Time	Acc (%)
ERM [57]	25.6M	18.4min	74.6
CGC [39]	25.6M	28.8min	70.6
DRE (ours)	25.6M	33.1min	**80.2**

Table 5. Efficiency comparison on *VLCS* [14] dataset using VOC2007 [13] as testing distribution. Our model significantly outperforms existing methods on prediction accuracy with a marginal increase in train time than the CGC [39] method, no additional parameters were introduced.

	DEC loss ↓	iAUC ↑	Acc (%) ↑
ERM [57]	1.000±0.02	0.758±0.01	74.6±1.3
DRE w/o reg.	0.909±0.01	0.756±0.01	78.1±0.9
DRE w/o consist.	**0.364**±0.03	0.698±0.02	71.2±2.1
DRE (full)	0.773±0.01	**0.772**±0.01	**80.2**±0.4

Table 6. Ablation study on *VLCS* [14] dataset using VOC2007 [13] as testing distribution. Although the variant (w/o explanation regularization) increases the prediction accuracy compared to ERM, it makes limited improvement in DEC loss and slightly dropped on iAUC. The variant (w/o explanation consistency) has the most significant DEC loss decrease, but it also leads to a severe drop in iAUC and prediction accuracy.

on truly continuous inputs [47]. As shown in Tab. 4, our model outperforms the ERM model with significant improvement on all three metrics. On average of each continental region (distribution) as the testing set, our method outperforms the ERM method by 84.5%, 29.2%, and 18.5% in terms of DEC loss, scientific consistency, and prediction residual. Note that the higher scientific consistency would further achieve promising and valuable scientific outcomes in the downstream tasks [46].

4.6. Ablation Study

In this section, we perform ablation studies to investigate key components of our method proposed in Eq. 4 and Eq. 5. The empirical results are shown in Tab. 6

Ablation on explanation regularization. The variant (w/o explanation regularization) increases the predictive accuracy by 3.5% compared to ERM, however, the it makes limited improvement in DEC loss and slightly dropped on iAUC. This indicates limited enhancement of explanation robustness against OOD data, the model might falls into a local minimum to satisfy the *distributional explanation consistency* constraint. For example, an explanation that uniformly attributes the prediction to all features.

Ablation on distributional explanation consistency. The variant (w/o explanation consistency) significantly improved the DEC loss by 63.6% over the ERM model. However, the explanation fidelity and prediction accuracy are significantly decreased by 6.0% and 3.4%. This indicates
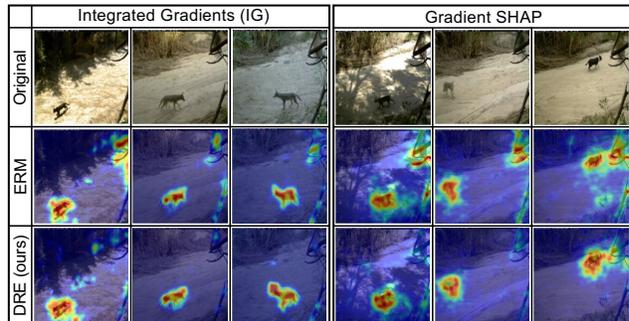

Figure 4. Integrated Gradients (IG) [55] and Gradient SHAP [32] saliency maps for OOD data from *Terra Incognita* dataset. Using location 46 as the testing set, for ERM model the saliency maps of both Integrated Gradients (IG) [55] and Gradient SHAP [32] methods are excessively focused on background pixels, such as branch and ground.

that blindly encouraging the explanation sparsity would hurt the explanation and predictive performance.

4.7. Generalize to Different Explanation Methods

In this experiment, we compare the saliency maps of ours and the ERM model. As shown in Fig. 4, the improved explainability of our model trained by Grad-CAM can be generalized to a variety of data-driven explanation methods. Using Location 46 in *Terra Incognita* as the testing set, for the ERM model, the saliency maps of both Integrated Gradients (IG) [55] and Gradient SHAP [32] methods are excessively focused on background pixels, such as branch and ground. On the contrary, the saliency maps of our model alleviate the reliance on background pixels, and clearly depicts the contour of the object.

5. Conclusion

In this paper, we present comprehensive study to show that the data-driven explanations are not robust against out-of-distribution data. To address this problem, we propose a an end-to-end model-agnostic learning framework Distributionally Robust Explanation (DRE). The proposed method fully utilizes inter-distribution information to provide supervisory signals for explanation learning without human annotation. We conduct extensive experiments on wide range of tasks, including the classification and regression tasks on image and scientific tabular data. Our results demonstrate the superior of our method in terms of explanation and prediction robustness against *out-of-distribution* data.

Acknowledgement

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018. 1, 2

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 1, 2, 3, 5, 6, 7

[3] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, pages 528–539. PMLR, 2020. 2

[4] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018. 1, 3, 5, 6

[5] Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and aggregating feature-based model explanations. In *Proceedings of IJCAI*, 2021. 7

[6] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004. 4

[7] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019. 2

[8] Jiefeng Chen, Xi Wu, Vaibhav Rastogi, Yingyu Liang, and Somesh Jha. Robust attribution regularization. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[9] Myung Jin Choi, Joseph J Lim, Antonio Torralba, and Alan S Willsky. Exploiting hierarchical context on a large database of object categories. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 129–136. IEEE, 2010. 5

[10] Ilke Cugu, Massimiliano Mancini, Yanbei Chen, and Zeynep Akata. Attention consistency on visual corruptions for single-source domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4165–4174, 2022. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[12] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 5, 8

[14] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1657–1664, 2013. 5, 6, 8

[15] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Pattern Recognition Workshop*, 2004. 5

[16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 2

[17] Jing Gao and Brian C O'Neill. Mapping global urban land for the 21st century with data-driven simulations and shared socioeconomic pathways. *Nature communications*, 11(1):1–12, 2020. 3, 5, 6

[18] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2020. 2, 3, 6

[19] Hao Guo, Kang Zheng, Xiaochuan Fan, Hongkai Yu, and Song Wang. Visual attention consistency under image transforms for multi-label image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 729–739, 2019. 2, 5

[20] Tao Han, Wei-Wei Tu, and Yu-Feng Li. Explanation consistency training: Facilitating consistency-based semi-supervised learning with interpretability. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 7639–7646, 2021. 2

[21] 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. 6

[22] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018. 4

[23] Michael Hind, Dennis Wei, Murray Campbell, Noel CF Codella, Amit Dhurandhar, Aleksandra Mojsilović, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Ted: Teaching ai to explain its decisions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 123–129, 2019. 2

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[25] 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. 2, 3

[26] 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. 2

[27] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951. 4

[28] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017. 1

[29] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 2

[30] Tang Li, Jing Gao, and Xi Peng. Deep learning for spatiotemporal modeling of urbanization. *Advances in Neural Information Processing Systems Workshops*, 2021. 6

[31] Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 2

[32] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017. 2, 8

[33] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18177–18186, 2022. 2

[34] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2302–2310, 2021. 2

[35] Xi Peng, Fengchun Qiao, and Long Zhao. Out-of-domain generalization from a single source: An uncertainty quantification approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[36] Xi Peng, Zhiqiang Tang, Fei Yang, Rogerio S Feris, and Dimitris Metaxas. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2226–2234, 2018. 2

[37] Xi Peng, Xiang Yu, Kihyuk Sohn, Dimitris N Metaxas, and Manmohan Chandraker. Reconstruction-based disentanglement for pose-invariant face recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1623–1632, 2017. 2

[38] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 3, 7

[39] Vipin Pillai, Soroush Abbasi Koohpayegani, Ashley Ouligian, Dennis Fong, and Hamed Pirsiavash. Consistent explanations by contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10213–10222, 2022. 2, 5, 6, 7, 8

[40] Fengchun Qiao and Xi Peng. Uncertainty-guided model generalization to unseen domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6790–6800, 2021. 2, 4

[41] Fengchun Qiao and Xi Peng. Topology-aware robust optimization for out-of-distribution generalization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. 2

[42] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12556–12565, 2020. 2

[43] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016. 2

[44] Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International conference on machine learning*, pages 8116–8126. PMLR, 2020. 2, 5

[45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 6

[46] Ribana Roscher, Bastian Bohn, Marco F Duarte, and Jochen Garcke. Explainable machine learning for scientific insights and discoveries. *Ieee Access*, 8:42200–42216, 2020. 2, 7, 8

[47] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: training differentiable models by constraining their explanations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2662–2670, 2017. 2, 8

[48] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. 1

[49] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1):157–173, 2008. 5

[50] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019. 1, 3, 6, 7

[51] 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. 1, 2, 3, 4, 7

[52] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. In *International Conference on Learning Representations*, 2018. 2

[53] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image

classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 2, 3, 4, 7

[54] Wolfgang Stammer, Patrick Schramowski, and Kristian Kersting. Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3619–3629, 2021. 2, 5

[55] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 8

[56] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011. 1

[57] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999. 2, 3, 5, 6, 7, 8

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[59] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018. 2

[60] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020. 2, 5

[61] Rebecca Wexler. When a computer program keeps you in jail. *The New York Times*, 13, 2017. 1

[62] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6502–6509, 2020. 4, 6, 7

[63] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 4

[64] Yilun Zhou, Serena Booth, Marco Tulio Ribeiro, and Julie Shah. Do feature attribution methods correctly attribute features? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9623–9633, 2022. 1