Robustness to Spurious Correlation: A Comprehensive Review

Mohammadjavad Maheronnaghsh[1] and Taha Akbari Alvanagh[1]

Sharif University of Technology, Tehran, Iran
m.j.maheronnaghsh@gmail.com

Abstract. The persistence of spurious features in machine learning models remains a significant challenge. To address this issue, we identify several future directions that require attention. Firstly, we highlight the need for a new dataset that allows researchers to control the types and levels of spurious features, as this resource is currently lacking. Secondly, we emphasize the importance of addressing spurious features in natural language processing, where more attention is needed compared to vision-related tasks. We also stress the need for addressing spurious correlations at the core algorithmic level, rather than relying on complex, task-specific solutions that may not generalize well. Finally, we advocate for the development of weakly-supervised or unsupervised methods that reduce reliance on group labels, making the approaches more widely applicable. Our review aims to provide a comprehensive overview of existing work and guide future research in creating more robust machine learning models.

Keywords: Machine Learning · Deep Learning· Explainable Artificial Intelligence· Spurious Correlation· Spurious Features· Causal Features· Causality· Biases· Bias Detection and Identification· Bias Resolving· Feature Engineering

1 Introduction

The rise of machine learning has expanded its use across fields like computer vision, NLP, healthcare, and finance. However, this growth has highlighted a key challenge: spurious correlation. This occurs when a model mistakenly identifies a false correlation between variables, leading to unreliable predictions, especially in critical applications. While spurious correlation is not new, it has become more prominent with the increasing complexity of models. Traditional approaches like regularization and feature selection often fall short. As a result, new methods have been developed to enhance model robustness against spurious correlations. This paper provides a comprehensive survey of these methods, categorized into six main domains:

– **Mitigation Methods**: Papers that introduce a way to mitigate spurious correlation, such as regularization techniques or algorithmic modifications.

Method	1	2	3	4	5	6	7
DFR [23]	88.3	92.9	74.7	70.1	N/A	N/A	64.5
JTT [37]	81.1	86.7	72.6	69.3	74.5	N/A	64.2
CNC [81]	88.8	88.5	N/A	68.9	**77.4**	N/A	N/A
SCILL [6]	N/A	86.5	N/A	N/A	N/A	N/A	N/A
SMM [78]	N/A	90.5	N/A	N/A	N/A	N/A	N/A
PDE [12]	89.0	**94.5**	**75.5**	71.2	N/A	N/A	N/A
AFR [50]	**95.1**	84.7	N/A	67.1	N/A	N/A	**70.1**
SELF [66]	83.9	93.0	70.7	79.1	N/A	N/A	N/A
BAM [34]	83.5	89.2	71.2	**79.3**	N/A	N/A	N/A
BAM + ClassDiff [34]	80.1	89.1	70.8	**79.3**	N/A	N/A	N/A
EIIL [9]	N/A	N/A	N/A	N/A	72.8	N/A	N/A
LfF [45]	77.2	78.0	N/A	78.8	N/A	N/A	N/A
IRM [2]	N/A	N/A	N/A	N/A	N/A	**66.9**	N/A

Table 1: Worst Group Accuracies for various methods across datasets. Here is the list of datasets; 1: CelebA, 2: Waterbirds, 3: MultiNLI, 4: CivilComments, 5: ColorMnist, 6: CorruptedMnist, 7: Chest X-Ray

- **Benchmarks and Datasets**: Papers that introduce a new benchmark or dataset for evaluating spurious correlation, providing a common ground for comparing different methods.
- **Surveys and Reviews**: Papers that provide a comprehensive overview of the field, summarizing existing methods and techniques.
- **Evaluation Metrics**: Papers that introduce new evaluation metrics or refine existing ones to assess the robustness of machine learning models to spurious correlation.
- **Theoretical Hypotheses**: Papers that propose new theoretical hypotheses or frameworks for understanding the causes and consequences of spurious correlation.
- **Practical Hypotheses**: Papers that propose practical solutions or methods for mitigating spurious correlation in specific applications or domains.

Understanding these methods helps in designing more reliable machine learning models, which is crucial given the significant impact of accurate predictions.

The classical approach in machine learning consists of training a classifier on a training dataset $\mathcal{D}_{\text{tr}}$ with the same distribution of test dataset $\mathcal{D}_{\text{te}}$ which we call the ERM classifier which aims to minimize:

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{tr}}}[\ell(f(x),y)]$$

This may dramatically fail since training distribution may confront shifts. A spurious correlation is a special case of this distribution shift where an attribute may spuriously be correlated with the label in training distribution but this correlation may be absent or reversed in the testing distribution resulting in poor performance.

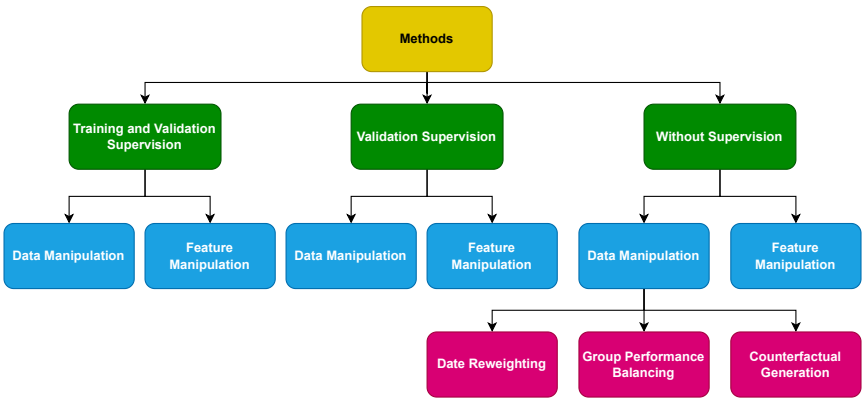

Fig. 1: Categorization of Methods

To formulate this problem we add this spurious attribute to the problem formulation i.e. The training distribution will consist of $\{x_i, y_i, a_i\}^n$ where a_i is the spurious attribute associated with the i'th attribute. To correctly formulate our goal we define the following objective function instead of the ERM objective function:

$$\max_{a',y'} \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{tr}}|a,y=a',y'}[\ell(f(x),y)]$$

The training data is split into **groups** based on values of (a, y), hence we are minimizing the **Worst group accuracy (WGA)**. Groups more apparent in the data are called majority groups on which we usually perform well. On the other hand groups less apparent in the data are called minority groups on which we usually perform poorly. Hence the general goal of WGA can be summarized as boosting the performance of minority groups while preserving the performance of majority groups.

1.1 Related Works

To our knowledge, there are only two survey and review papers in this field. We will discuss these papers, highlighting their advantages and limitations.

The paper [62] offers a basic overview of spurious correlations across medical imaging, NLP, and computer vision. It examines detection methods like adversarial training, representation learning, and interpretability techniques, and discusses challenges in these areas. However, it lacks depth, missing specific methods, problem definitions, datasets, evaluation metrics, and other crucial details.

The survey by [79] categorizes methods but does not account for the amount of group information needed, leading to incomplete or overlapping categories. It also provides limited insight into datasets, lacking clarity on their motiva-

tions. The paper briefly mentions methods within categories without a detailed investigation into their workings.

2 Methods

Methods for mitigating spurious correlations differ mainly in their supervision levels. Since labeled data is costly, methods with less supervision are preferred. Supervised learning needs large training sets, making group annotation-based methods expensive. Methods without group annotations can still suffer from spurious correlations. Most use group annotations only for hyperparameter tuning. They fall into three categories: no group annotations, annotations in validation only, and annotations in both training and validation. They also vary by data handling approach—augmenting to capture features or avoiding spurious attributes. Data manipulation methods for group fairness (WGA) include reweighting, accounting for group performance, and generating counterfactual data.

2.1 Training & Validation Supervision

Data Manipulation

Group Distributionally Robust Optimization (GDRO) [52] The paper makes three key contributions: it shows that DRO [10] struggles with overparameterized networks and suggests using stronger regularization; it introduces group-adjusted DRO, which reduces generalization gaps by scaling as $\frac{1}{\sqrt{n}}$; and it presents a training algorithm that alternates between adjusting group distributions and optimizing model parameters to minimize worst-case loss.

Deep Feature Reweighting (DFR) [23] The DFR method is widely used in the field. It involves training a network on standard data without group supervision and then retraining only the last layer on group-balanced data, keeping the earlier layers fixed. Some papers incorrectly claim DFR doesn't need group supervision during training [23] [84], but it does require it for the last layer. Essentially, the feature extractor retains both spurious and core features, while the classifier focuses solely on core features.

Progressive Data Expansion (PDE) [12] This method shows that models learn spurious features when they are easier to learn, especially if spurious correlations exceed 50%, causing slower learning of core features. To address this, the method suggests starting with a balanced dataset and gradually expanding to the full training data.

Feature Manipulation

Learning with explanatory interaction (LWEI) [29] This paper explores spurious correlations in concept drift, noting that models relying on these correlations may fail because spurious features often remain unchanged. To address this, the paper studies drift in explanations and encourages users to correct spurious correlations in these explanations.

Category	Index	Method	Venue	Keywords
TVD	1	GDRO [52]	ICLR (2020)	Distributional robustness, Generalization
	2	DFR [23]	arXiv (2022)	Last-layer retraining, Group-balanced
	3	PDE [12]	NeurIPS (2024)	Data expansion, Spurious features
TVF	4	LWEI [29]	arXiv (2024)	Concept drift, Explanations
VD	5	JTT [37]	PMLR (2021)	Misclassification, Upsampling
	6	SCILL [6]	NeurIPS (2022)	Group invariant, Conditional independence
	7	BPA [57]	CVPR (2022)	Clustering, Pseudo-groups
	8	AFR [50]	PMLR (2023)	Feature reweighting, Confidence-based
	9	SELF [28]	NeurIPS (2023)	Last-layer retraining, Class-balanced
	10	BAM [34]	TMLR (2024)	Bias amplification, Auxiliary variables
VF	11	IRM [2]	arXiv (2019)	Invariant risk, Domain generalization
	12	REX [27]	PMLR (2021)	Risk extrapolation, Robustness
	13	StableNet [82]	CVPR (2021)	Sample reweighting, Shifts
	14	SSA [47]	ICLR (2022)	Semi-supervised, Pseudo-Labeling
	15	CNC [81]	arXiv (2022)	Contrastive learning, Unsupervised
	16	CIU [68]	CL (2022)	Counterfactuals, Causal inference
	16	DivDis [33]	ICLR (2023)	Diversify, Disambiguation
	17	SIFER [66]	PMLR (2023)	Feature sieving, Forgetting loss
	18	SMM [78]	ICML (2023)	Multi-modal, Spurious features
WDD	19	LfF [45]	NeurIPS (2020)	Biased classifier, Debiased model
	20	Rebias [3]	PMLR (2020)	Biased representations, Independent
	21	LWBC [22]	NeurIPS (2022)	Biased ensemble, Reweighting
	22	SELF [28]	NeurIPS (2023)	Last-layer retraining, Class-balanced
	23	BAM [34]	TMLR (2024)	Bias amplification, Auxiliary variables
WDG	24	GEORGE [60]	NeurIPS (2020)	Clustering, Pseudo-labeling
	25	EIIL [9]	PMLR (2021)	Invariant learning, Environment inference
	26	FACTS [80]	CVPR (2023)	Bias discovery, Clustering
	27	DISC [75]	PMLR (2023)	Concept-aware, Pseudo-labels
WDC	28	AGC [73]	AAAI (2021)	Counterfactuals, Causal words
	29	GICL [41]	CVPR (2021)	Generative interventions, Causal learning
	30	CGKR [43]	ACM (2022)	Counterfactuals, Reinforcement
WF	31	IdMNLP [71]	arXiv (2021)	Cross-dataset, Semantic analysis
	32	NuRD [48]	arXiv (2021)	Nuisance-randomized, Independence
	33	CIM [65]	PMLR (2021)	Perceptual similarity, Contrastive
	34	Cobias [56]	AAAI (2022)	Bias measurement, Noise injection
	35	LBC [84]	arXiv (2024)	Spurious detection, Vision-language

Table 2: Methodologies categorized by combined taxonomy. Here are the abbreviations; TVD: Training Supervision & Validation Supervision & Data Manipulation, TVF: Training Supervision & Validation Supervision & Feature Manipulation, VD: Validation-only Supervision & Data Manipulation, VF: Validation-only Supervision & Feature Manipulation, WDD: Without Supervision & Data Manipulation (Data Reweighting), WDG: Without Supervision & Data Manipulation (Group Performance Balancing), WDC: Without Supervision & Data Manipulation (Counterfactual Generation), WF: Without Supervision & Feature Manipulation

2.2 Validation Only Supervision

Data Manipulation

Just Train Twice (JTT) [37] JTT aims to handle spurious correlations using group info from a small validation set. It trains an identifier network and treats misclassified examples as proxies for minority groups. A second network is then trained by upsampling these misclassified examples. Contrary to some sources [84], JTT does require group labels during validation.

Spurious-correlation-strata Invariant learning with Label-balance (SCILL) [6] This paper introduces a theoretical framework for group-IR methods and highlights the flaws in current approaches. It proposes SCILL, which splits labels into groups to make spurious attributes conditionally independent of labels. The method reweights each sample based on group label proportions, with group labels inferred by the proposed algorithm.

Bias Pseudo-Attribute (BPA) [57] This paper observes that for a model trained sufficiently, samples with similar spurious correlations fall into the same cluster. Hence it uses a clustering approach to make pseudo-groups. Based on the created group it up-weights groups with a small number of samples to make a group-balanced training.

Automatic Feature Reweighting (AFR) [50] AFR builds on JTT [37] and CNC [81] but differs in how it identifies minority and majority groups. It trains an ERM classifier until convergence, without regularization or early stopping. Then, it reweights samples, giving less weight to high-confidence examples, and retrains the last layer for improved robustness.

Selective Last-Layer finetuning (SELF) [28] First, the paper makes a practical observation that class-balanced datasets suffice for last-layer retraining for being robust to spurious correlations. It demonstrates that class-balanced datasets work well even if they are not group-balanced.

Bias Amplification (BAM) [34] This method introduces a trainable auxiliary variable for each sample, adding it to the logits. For hard samples, this variable learns the label, while for easy samples, the model's learning dominates. This amplifies errors in the majority samples group, up-weights them, and trains a new model to reduce reliance on spurious correlations. Hyperparameters, including up-weighting, are selected via WGA.

Spread Spurious Attribute (SSA) [47] The paper proposes a method to best use spurious attribute data by combining labeled and unlabeled data. A classifier is trained using cross-entropy loss on both, with predictions evaluated on remaining labeled data. High-confidence points are selected with different thresholds to balance the dataset, which is then used for final training.

Feature Manipulation

Invariant Risk minimization (IRM) [2] Invariant Risk Minimization aims to find a representation that is optimal for standard ERM [69] and across different environments. To achieve this, it minimizes the following loss function:

$$\min_{\phi:\mathcal{X}\to\mathcal{Y}} \sum_{e\in\varepsilon_{\text{tr}}} R^e(\phi) + \lambda||\nabla_{w|w=1.0} R^e(w.\phi)||^2 \tag{1}$$

The first term ensures optimal performance in the ERM setting [69], while the second term enforces optimality across all environments.

Risk Extrapolation (REX) [27] The paper's main idea is that a robust classifier must meet two criteria: it should have low risk across all domains and low risk in each domain. To address this, they propose two loss functions: Minmax-re and Variance-re, which account for these criteria.

StableNet [82] This paper discusses sample reweighting as a way to disassociate the spurious and core features. Since we do not have any supervision on the spurious correlation it disassociates on all features. With disassociated features, we achieve stable models under distribution shifts.

Correct-N-Contrast (CNC) [81] CNC learns core features without supervision, unlike GRDO [52], which requires group labels. It performs comparably to GDRO despite being unsupervised. The method involves training an ERM classifier [69] with cross-entropy loss and then applying a contrastive loss to remove spurious correlations while maintaining prediction accuracy with a weighted average of these losses, with weights tuned using validation data.

Counterfactual Inference Understanding (CIU) [68] NLU models [4] often rely on biased predictions and spurious shortcuts due to repetitive patterns and annotation artifacts. The paper defines causal relationships to assess these artifacts and uses counterfactual inference to reduce spurious correlations. This inference-focused approach is effective even for out-of-distribution data.

Diversify and Disambiguate (DivDis) [33] The Diversifying step trains a neural network with multiple heads, each minimizing loss on training data but producing different outputs on test data. The Disambiguate step chooses the most suitable head for each test data point based on its label.

Spuriousity Mitigator in Multi-modals (SMM) [78] SMM aims to reduce spurious features in multi-modal data. It uses GradCAM [55] to identify which parts of the feature map are attended to and then fine-tunes the model to focus more on core features. GradCAM is used only for qualitative evaluation, not directly in the method. SMM also employs CLIP and Vocabulary Open Detector to detect images with or without spurious features.

Sieving Features for Robust learning (SIFER) [66] This approach has two stages: first, a neural network is trained on standard data. Then, a classifier is built on the network's initial layers to focus on challenging features, using a forgetting loss to remove easy, spurious features. One version uses group annotations during validation, while the other does not.

2.3 Without Supervision

Data Manipulation: Data Reweighting

Learn from Failure (LfF) [45] The approach is similar to JTT [37]. It trains a biased classifier with Generalized Cross-Entropy loss [83] to focus on

easier examples. At the same time, a debiased model is trained, assigning weights to data points based on the loss difference between the biased and debiased models.

Rebias [3] This method trains a biased classifier with an ERM model [69] using regularization and early stopping. It then trains a second model to make independent predictions, hoping the first model captures spurious attributes while the second focuses on predicting the label.

Learning with biased committee (LWBC) [22] This paper designs an ensemble learning method to mitigate spurious correlations. To do so they train a biased ensemble so that the data of each classifier is a bootstrapped sample. Then this biased ensemble is used to weight the samples in the training dataset so that to train the actual classifier the minority group gets more weight.

Avoiding Spurious Correlations via Logit Correlation (ASCLC) [38] This approach uses a neural network with two heads the first head is used to train the model with generalized cross-entropy loss so that it becomes more dependent on the spurious correlation. The second head is then based on the first head to correct the logits and make a robust classifier.

Selective Last-Layer finetuning (SELF) [28] This method is similar to the SELF approach with validation supervision, but it selects a subset of training data, such as by misclassification, as a proxy for the majority group.

Bias Amplification (BAM) + ClassDiff [34] This method is like BAM but does not use validation data supervision. Instead, it uses the difference in class accuracy as a proxy for worst-group accuracy and aims to minimize this difference without group annotations.

Data Manipulation: Group Performance Balancing

Environment inference for invariant learning (EIIL) [9] Invariant learning methods typically require environment labels to find invariant features. This paper proposes a two-stage approach: first, it identifies partitions to maximize an invariant learning penalty with a soft, differentiable split, and then it trains a model using the invariant learning objective.

First Amplify Correlations and Then Slice to Discover Bias (FACTS) [80] This paper uses an ERM-based model [69] with regularization to extract biased representations, which are then fed into a Gaussian mixture model [51] as a proxy for group annotations.

Discover and Cure (DISC) [75] The paper proposes a multi-modal approach [69] to address spurious correlations by preprocessing data, creating a concept-sensitive environment, and balancing the dataset using a concept bank and cluster pseudo-labels. It calculates concept sensitivity for each cluster and generates 200 images per concept using a text-to-image model.

GEORGE [60] The paper proposes training an ERM classifier [69], clustering the resulting data to identify subclasses, and then using GDRO [52] to minimize worst-group accuracy in these clusters.

Data Manipulation: Counterfactual Generation

Automatically Generating Counterfactual (AGC) [73] The paper augments a review dataset with counterfactuals by identifying causal words, replacing them with antonyms, and flipping sentence labels.

Generative Interventions for Causal Learning (GICL) [41] In this paper generative models are exploited for generating interventions for data, they use BigGan to model the data and further make interventions on them. Then a discriminative model is fit to data with mixed coefficients of original data loss and intervented data loss.

Counterfactual Generator Knowledge-aware Recommender (CGKR) [43] The paper mitigates spurious correlation in recommendation systems using two counterfactual generators for positive and negative classes. These generators use reinforcement learning with a Markov reward process to create high-quality samples.

Feature Manipulation

Identify and Mitigate Spurious Correlations in NLP(IdMNLP) [71] This paper classifies important words as spurious or genuine using cross-dataset stability instead of labeled data. It extracts domain-specific synonyms and analyzes semantic differences in sentences.

Nuisance-Randomized Distillation(NuRD) [48] This algorithm aims to find a classifier where the label and spurious attribute are independent in datasets with balanced spurious correlations. It begins by creating a nuisance-randomized distribution, either by modeling $p(x|y,z)$ and varying $p(y)$ and $p(z)$, or by reweighting samples to match this distribution. Then, it identifies features $r(x)$ that make y and z independent given $r(x)$, while maximizing prediction accuracy. This is achieved by maximizing:

$$\max_{\theta,\gamma} \mathbb{E}_{\hat{p}_\perp(x,y,z)} \log p_\theta(y|r_\gamma(x)) - \lambda I_{\hat{p}_\perp}(y; z|r_\gamma(x)) \tag{2}$$

The objective is to find features that predict y without relying on the spurious attribute z.

Robust Representation Learning via Perceptual Similarity Metrics (CIM) [65] This paper uses the SSIM metric [18] to align representations with human perception. It applies a contrastive loss to the embeddings, combining this with cross-entropy loss to ensure high classification performance while reducing spurious features.

Bias Measurement with Conditional Mutual Information (Cobias) [56] The paper introduces a theoretical algorithm for measuring feature-level bias and proposes two debiasing frameworks: Bias Regularization, using a cobias loss, and Label Noise Injection, which adds synthetic label noise. It fits into "Spurious Correlation Mitigators" and "Evaluation Metrics," with a focus on debiasing.

Robust Classifier Learner (LBC) [84] LBC detects attributes using a Vision-Language model, selects relevant nouns and words, and creates an attribute set. It then calculates spurious features in each class, then clusters and

balances these features for training. This method automatically balances features without needing group labels during training or validation.

Research on methods requiring group annotations for both training and validation is nearly complete, with Group DRO [52] being a leading approach and benchmark. In contrast, methods that avoid group annotations during training typically have performance limited by those that use annotations only for validation. In such cases, annotations are usually for retraining the last layer or tuning hyperparameters, leading to some performance degradation. The DFR method [23] serves as a baseline for approaches using validation data for retraining, while AFR [50] is a baseline for those using it solely for hyperparameter tuning. Methods that completely avoid group supervision are relatively new, with BAM [34] and SELF [66] being notable examples in this emerging field.

3 Datasets

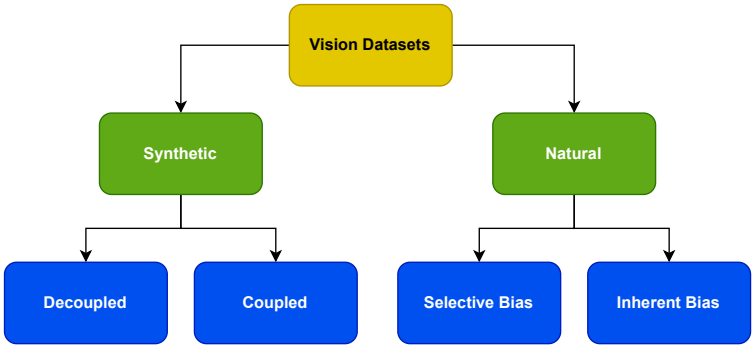

Fig. 2: Categorization of vision datasets

3.1 Vision Datasets

We classify vision task datasets into two main types: natural images and synthetic images. Synthetic datasets are further divided into coupled and decoupled types. Coupled datasets link core and spurious features, like a digit and its color, while decoupled datasets separate them, such as the foreground and background. For natural images, we distinguish between inherently biased datasets, where bias is intrinsic (e.g., swimming usually happens in pools), and intentionally biased datasets, where bias is deliberately introduced (e.g., cats with a grass background).

Taxonomy			Index	Dataset	Size	Year
Vision	**S**	**CF**	1	**CMNIST [32]**	60,000	2021
			2	**Corrupted CIFAR10 [32]**	60,000	2021
			3	**SpuCoMnist [20]**	118,100	2023
		DF	4	**Waterbirds [52]**	11,788	2019
			5	**Imagenet-9** [77]	5,495	2020
			6	**Spawrious** [40]	152,064	2023
	N	**SB**	8	**Imagenet-A** [11]	7,500	2009
			9	**ISIC** [8]	3,694	2019
			10	**NICO** [17]	24,214	2021
			11	**Meta Shift** [35]	12,868	2022
		IB	12	**Fmow**	1,047,691	2017
			13	**CelebA [52]**	202,599	2019
			14	**BAR** [46]	2,595	2020
NLP	-	-	16	**MultiNLI [52]**	206,175	2019
			17	**CivilComments [5]**	1,999,514	2019

Table 3: Overview of Datasets. Here are the abbreviations; S: Synthetic, N: Natural, CF: Coupled Features, DF: Decoupled Features, SB: Selective Bias, IB: Inherent Bias, ISIC: Skin lesions classification

Synthetic datasets: Coupled features

- Colored MNIST (CMNIST) [32]: The Colored MNIST dataset consists of MNIST digits [31] colored spuriously with the labels. So the model will predict the digit using colors instead of the actual digit.
- Corrupted CIFAR10 [32]: The Corrupted CIFAR10 consists of images of the CIFAR10 [26] dataset corrupted so that there is a spurious correlation between the label and the kind of corruption.
- SpuCoMnistT [20]: This dataset consists of MNIST dataset images [31] where the color is spuriously correlated with the digit.

Synthetic datasets: Decoupled Features

- Waterbirds [52]: Waterbirds is a dataset consisting of waterbirds and landbirds on various backgrounds. It was artificially generated by bird images from the CUB dataset [70] and background from the Places dataset [85]. In this case, the background is a spurious correlation for predicting whether the bird is a landbird or a waterbird.
- Imagenet-9 [77]: This dataset consists of a subset of the Imagenet [11] dataset with 9 labels such that the background is spuriously correlated with the foreground.
- Spawrious [40]: This dataset consists of images of dogs to classify into different categories where the background is spuriously correlated with the type of dog. This dataset consists of 152,064 images. This dataset brings a new paradigm for spurious correlation which is many to many spurious correlations.

Natural datasets: Selective Bias

- Imagenet-A [11]: This dataset consists of a subset of the Imagenet dataset that ResNet models [16] misclassify.
- ISIC [8] involves classifying skin lesions as benign or melanoma. It includes multiple train-test splits, each designed to highlight a specific type of spurious correlation.
- NICO [17] features object images in various contexts, allowing spurious correlations to be created by adjusting the proportions of contexts and objects.
- Meta Shift [35]: The Meta Shift dataset includes images of various objects in different contexts, using the Visual Genome dataset's metadata [25] to introduce spurious correlations. It clusters images based on context and provides annotations on the types of shifts caused by this metadata.

Natural datasets: Inherent Bias

- Fmow [7]: The Fmow dataset consists of satellite images comprised of different geographical locations that contribute to potential spurious correlations for predicting the functional purpose of the building.
- CelebA [52]: CelebA dataset consists of images of celebrities and the task is to predict attributes of faces. In this case, gender has a spurious correlation with hair color prediction.
- BAR [46] is designed for action recognition, where backgrounds are spuriously correlated with actions. For example, swimming is often associated with water backgrounds, creating a spurious correlation between the water background and the swimming action.

NLP Datasets

- MultiNLI [52]: This dataset consists of pairs of sentences labeled as either "entailed", "neutral" or "contradictory". The spurious features are "No negation" or "negation" where the existence of the negation word spuriously correlates with the sentences being contradictory
- CivilComments [5]: This dataset consists of comments and different attributes for predictions. It exhibits demographic information like gender, race, etc. as spurious correlations.

Datasets with spurious correlations include a core feature essential for the task and a spurious feature that, while correlated with the label, is not causally related. This correlation is strong in training data but weaker in validation and test data. Synthetic datasets like Domino, C-MNIST [32], and Waterbirds [52] allow control over spurious correlations, showing that increased spurious correlation typically reduces model performance. The spurious feature is usually simpler than the core feature, leading models to rely on it due to simplicity bias.

Current datasets lack controls for feature complexity, which is a research gap. Future work should focus on designing datasets with adjustable spurious feature complexity and explore scenarios with multiple simultaneous spurious correlations, particularly in vision tasks. Most datasets are vision-based, with synthetic or biased samples used to introduce spurious correlations.

4 Benchmarks

The following benchmarks have been proposed to evaluate the robustness of models against spurious correlations:

[20] introduces the SPUCO benchmark, which includes a dataset and a method to identify minority groups. The paper also highlights the importance of considering spurious features in text data.

[15] proposes a benchmark to evaluate the robustness of visual transformers against spurious correlations. The benchmark consists of three datasets and evaluates the impact of fine-tuning and pretraining on mitigating spurious correlations.

[40] introduces a new benchmark, Spawrious, which includes a dataset of 152K images and evaluates the performance of models in various scenarios. The paper also highlights the low accuracy of state-of-the-art models on the benchmark.

[67] proposes a benchmark to evaluate the robustness of large language models against spurious correlations. The study shows that pre-trained datasets increase robust accuracy, but are inconsistent across benchmarks and datasets.

[76] introduces two datasets, DSNLI and DMNLI, which are created using a method to generate data for training. The paper also proposes a filtering algorithm to remove spurious examples.

[30] proposes a simple benchmark to evaluate the effectiveness of Deep Feature Reweighting (DFR) [23] against spurious correlations. The paper evaluates DFR on a realistic medical dataset and investigates why it works.

5 Evaluation Metrics

The effectiveness of post-hoc explanations for spurious correlation has been evaluated in several papers. For example, [1] found that post-hoc explanations, such as feature attribution, concept activation, and training point ranking, may not be as effective as expected in detecting unknown spurious correlations. They considered three types of post-hoc explanations: feature attribution (Input-Gradient [58], SmoothGrad [59], Integrated Gradients [49], Guided Backprop [61]), concept activation (TCAV) [21], and training point ranking (Influence Functions) [24].

Another approach to enhancing model robustness and fairness to spurious correlations is through regularization. [74] proposed a regularization approach that extracts a set of features with high importance and labels them as spurious or genuine. The model is then regularized using different weights for spurious and genuine features to ensure that spurious features are given lower weights.

6 Theoretical and Empirical Hypothesis

The relationship between spurious features and out-of-distribution detection has been studied in several papers. For example, [42] found that the existence of

spurious features can make OOD detection difficult. [53] showed that overparameterization can exacerbate spurious correlations, and that regularization can help mitigate this issue.

Other papers have explored the importance of regularization for worst-case generalization [52]. [13] proposed a robust reinforcement learning algorithm that can eliminate the effect of spurious correlations.

The importance of human annotations in mitigating spurious correlations has also been explored [63]. The paper introduced a new optimization objective, UV-DRO, which uses multiple annotations to reduce noise and improve accuracy.

Other papers have discussed the limitations of dataset balancing [54] and the importance of understanding the failure modes of out-of-distribution generalization [44]. The paper also discussed the importance of mechanistic mode connectivity [39].

The paper on informativeness and invariance [14] explores two perspectives on spurious correlations in NLP. [36] demonstrates that invariant learning without environment partition is infeasible due to different generative models producing identical distributions with varied causal features. [19] argues that spurious correlations should not be treated uniformly, as some methods may fail with specific datasets. Lastly, [64] assesses interpretable ML methods for handling spurious correlations, and [72] focuses on extracting keywords for text classification based on classifier weights.

7 Future Directions

To address the issue of spurious features in machine learning, we propose the following future directions:

- Develop a comprehensive dataset that enables researchers to control the types and levels of spurious features, bridging the current gap in this area.
- Shift the focus towards natural language processing tasks, where spurious feature mitigation has received less attention compared to vision-related tasks.
- Concentrate on core algorithmic solutions that can generalize across tasks, rather than relying on complex, task-specific solutions that may not generalize well.
- Explore weakly-supervised or unsupervised approaches that reduce reliance on group labels, making the approaches more widely applicable and robust.

8 Conclusion

In this comprehensive review, we have examined the current state of spurious features in machine learning, including both published and unpublished works. Our analysis highlights the importance of addressing this issue, which can have different types and origins. By identifying key areas for future research, we aim to provide a foundation for developing more robust machine learning models that can accurately generalize across various tasks.

Acknowledgements

We extend our gratitude to Professor Mohammad Hossein Rohban for his valuable advice on revising the paper.

References

1. Adebayo, J., Muelly, M., Abelson, H., Kim, B.: Post hoc explanations may be ineffective for detecting unknown spurious correlation. In: International conference on learning representations (2022)
2. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. ArXiv **abs/1907.02893** (2019), https://api.semanticscholar.org/CorpusID:195820364
3. Bahng, H., Chun, S., Yun, S., Choo, J., Oh, S.J.: Learning de-biased representations with biased representations. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 528–539. PMLR (13–18 Jul 2020), https://proceedings.mlr.press/v119/bahng20a.html
4. Baud, R., Lovis, C., Alpay, L., Rassinoux, A.M., Scherrer, J., Nowlan, A., Rector, A.: Modelling for natural language understanding. In: Proceedings of the Annual Symposium on Computer Application in Medical Care. p. 289. American Medical Informatics Association (1993)
5. Borkan, D., Dixon, L., Sorensen, J., Thain, N., Vasserman, L.: Nuanced metrics for measuring unintended bias with real data for text classification. In: Companion proceedings of the 2019 world wide web conference. pp. 491–500 (2019)
6. Chen, Y., Xiong, R., Ma, Z.M., Lan, Y.: When does group invariant learning survive spurious correlations? Advances in Neural Information Processing Systems **35**, 7038–7051 (2022)
7. Christie, G., Fendley, N., Wilson, J., Mukherjee, R.: Functional map of the world. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6172–6180 (2018)
8. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)
9. Creager, E., Jacobsen, J.H., Zemel, R.: Environment inference for invariant learning. In: International Conference on Machine Learning. pp. 2189–2200. PMLR (2021)
10. Delage, E., Ye, Y.: Distributionally robust optimization under moment uncertainty with application to data-driven problems. Operations research **58**(3), 595–612 (2010)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
12. Deng, Y., Yang, Y., Mirzasoleiman, B., Gu, Q.: Robust learning with progressive data expansion against spurious correlation. Advances in neural information processing systems **36** (2024)
13. Ding, W., Shi, L., Chi, Y., Zhao, D.: Seeing is not believing: Robust reinforcement learning against spurious correlation. Advances in Neural Information Processing Systems **36** (2024)

14. Eisenstein, J.: Informativeness and invariance: Two perspectives on spurious correlations in natural language. arXiv preprint arXiv:2204.04487 (2022)
15. Ghosal, S.S., Li, Y.: Are vision transformers robust to spurious correlations? International Journal of Computer Vision **132**(3), 689–709 (2024)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
17. He, Y., Shen, Z., Cui, P.: Towards non-iid image classification: A dataset and baselines. Pattern Recognition **110**, 107383 (2021)
18. Hore, A., Ziou, D.: Image quality metrics: Psnr vs. ssim. In: 2010 20th international conference on pattern recognition. pp. 2366–2369. IEEE (2010)
19. Joshi, N., Pan, X., He, H.: Are all spurious features in natural language alike? an analysis through a causal lens. arXiv preprint arXiv:2210.14011 (2022)
20. Joshi, S., Yang, Y., Xue, Y., Yang, W., Mirzasoleiman, B.: Towards mitigating spurious correlations in the wild: A benchmark & a more realistic dataset. arXiv preprint arXiv:2306.11957 (2023)
21. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., sayres, R.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 2668–2677. PMLR (10–15 Jul 2018), https://proceedings.mlr.press/v80/kim18d.html
22. Kim, N., Hwang, S., Ahn, S., Park, J., Kwak, S.: Learning debiased classifier with biased committee. Advances in Neural Information Processing Systems **35**, 18403–18415 (2022)
23. Kirichenko, P., Izmailov, P., Wilson, A.G.: Last layer re-training is sufficient for robustness to spurious correlations. arXiv preprint arXiv:2204.02937 (2022)
24. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 1885–1894. PMLR (06–11 Aug 2017), https://proceedings.mlr.press/v70/koh17a.html
25. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**, 32–73 (2017)
26. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
27. Krueger, D., Caballero, E., Jacobsen, J.H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., Courville, A.: Out-of-distribution generalization via risk extrapolation (rex). In: International conference on machine learning. pp. 5815–5826. PMLR (2021)
28. LaBonte, T., Muthukumar, V., Kumar, A.: Towards last-layer retraining for group robustness with fewer annotations. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 11552–11579. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/265bee74aee86df77e8e36d25e786ab5-Paper-Conference.pdf
29. Lalletti, C., Teso, S.: Spurious correlations in concept drift: Can explanatory interaction help? arXiv preprint arXiv:2407.16515 (2024)

30. Le, P.Q., Schlötterer, J., Seifert, C.: Is last layer re-training truly sufficient for robustness to spurious correlations? arXiv preprint arXiv:2308.00473 (2023)
31. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
32. Lee, J., Kim, E., Lee, J., Lee, J., Choo, J.: Learning debiased representation via disentangled feature augmentation. Advances in Neural Information Processing Systems **34**, 25123–25133 (2021)
33. Lee, Y., Yao, H., Finn, C.: Diversify and disambiguate: Learning from underspecified data. arXiv preprint arXiv:2202.03418 (2022)
34. Li, G., Liu, J., Hu, W.: Bias amplification enhances minority group performance. arXiv preprint arXiv:2309.06717 (2023)
35. Liang, W., Zou, J.: Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. arXiv preprint arXiv:2202.06523 (2022)
36. Lin, Y., Zhu, S., Tan, L., Cui, P.: Zin: When and how to learn invariance without environment partition? Advances in Neural Information Processing Systems **35**, 24529–24542 (2022)
37. Liu, E.Z., Haghgoo, B., Chen, A.S., Raghunathan, A., Koh, P.W., Sagawa, S., Liang, P., Finn, C.: Just train twice: Improving group robustness without training group information. In: International Conference on Machine Learning. pp. 6781–6792. PMLR (2021)
38. Liu, S., Zhang, X., Sekhar, N., Wu, Y., Singhal, P., Fernandez-Granda, C.: Avoiding spurious correlations via logit correction. arXiv preprint arXiv:2212.01433 (2022)
39. Lubana, E.S., Bigelow, E.J., Dick, R.P., Krueger, D., Tanaka, H.: Mechanistic mode connectivity. In: International Conference on Machine Learning. pp. 22965–23004. PMLR (2023)
40. Lynch, A., Dovonon, G.J., Kaddour, J., Silva, R.: Spawrious: A benchmark for fine control of spurious correlation biases. arXiv preprint arXiv:2303.05470 (2023)
41. Mao, C., Cha, A., Gupta, A., Wang, H., Yang, J., Vondrick, C.: Generative interventions for causal learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3947–3956 (2021)
42. Ming, Y., Yin, H., Li, Y.: On the impact of spurious correlation for out-of-distribution detection. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 10051–10059 (2022)
43. Mu, S., Li, Y., Zhao, W.X., Wang, J., Ding, B., Wen, J.R.: Alleviating spurious correlations in knowledge-aware recommendations through counterfactual generator. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1401–1411 (2022)
44. Nagarajan, V., Andreassen, A., Neyshabur, B.: Understanding the failure modes of out-of-distribution generalization. arXiv preprint arXiv:2010.15775 (2020)
45. Nam, J., Cha, H., Ahn, S., Lee, J., Shin, J.: Learning from failure: Training debiased classifier from biased classifier, 2020. URL https://arxiv. org/abs (2007)
46. Nam, J., Cha, H., Ahn, S., Lee, J., Shin, J.: Learning from failure: De-biasing classifier from biased classifier. Advances in Neural Information Processing Systems **33**, 20673–20684 (2020)
47. Nam, J., Kim, J., Lee, J., Shin, J.: Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. arXiv preprint arXiv:2204.02070 (2022)
48. Puli, A., Zhang, L.H., Oermann, E.K., Ranganath, R.: Out-of-distribution generalization in the presence of nuisance-induced spurious correlations. arXiv preprint arXiv:2107.00520 (2021)

49. Qi, Z., Khorram, S., Li, F.: Visualizing deep networks by optimizing with integrated gradients. In: CVPR workshops. vol. 2, pp. 1–4 (2019)
50. Qiu, S., Potapczynski, A., Izmailov, P., Wilson, A.G.: Simple and fast group robustness by automatic feature reweighting. In: International Conference on Machine Learning. pp. 28448–28467. PMLR (2023)
51. Reynolds, D.A., et al.: Gaussian mixture models. Encyclopedia of biometrics **741**(659-663) (2009)
52. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. arXiv preprint arXiv:1911.08731 (2019)
53. Sagawa, S., Raghunathan, A., Koh, P.W., Liang, P.: An investigation of why overparameterization exacerbates spurious correlations. In: International Conference on Machine Learning. pp. 8346–8356. PMLR (2020)
54. Schwartz, R., Stanovsky, G.: On the limitations of dataset balancing: The lost battle against spurious correlations. arXiv preprint arXiv:2204.12708 (2022)
55. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. International journal of computer vision **128**, 336–359 (2020)
56. Seo, S., Lee, J.Y., Han, B.: Information-theoretic bias reduction via causal view of spurious correlation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2180–2188 (2022)
57. Seo, S., Lee, J.Y., Han, B.: Unsupervised learning of debiased representations with pseudo-attributes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16742–16751 (2022)
58. Singh, M., Kumari, N., Mangla, P., Sinha, A., Balasubramanian, V.N., Krishnamurthy, B.: Attributional robustness training using input-gradient spatial alignment. In: European Conference on Computer Vision. pp. 515–533. Springer (2020)
59. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017)
60. Sohoni, N., Dunnmon, J., Angus, G., Gu, A., Ré, C.: No subclass left behind: Fine-grained robustness in coarse-grained classification problems. Advances in Neural Information Processing Systems **33**, 19339–19352 (2020)
61. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806 (2014)
62. Srivastava, M.: Addressing spurious correlations in machine learning models: A comprehensive review. OSF Prepr (2023)
63. Srivastava, M., Hashimoto, T., Liang, P.: Robustness to spurious correlations via human annotations. In: International Conference on Machine Learning. pp. 9109–9119. PMLR (2020)
64. Sun, S., Koch, L.M., Baumgartner, C.F.: Right for the wrong reason: Can interpretable ml techniques detect spurious correlations? In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 425–434. Springer (2023)
65. Taghanaki, S.A., Choi, K., Khasahmadi, A.H., Goyal, A.: Robust representation learning via perceptual similarity metrics. In: International Conference on Machine Learning. pp. 10043–10053. PMLR (2021)
66. Tiwari, R., Shenoy, P.: Overcoming simplicity bias in deep networks using a feature sieve. In: International Conference on Machine Learning. pp. 34330–34343. PMLR (2023)

67. Tu, L., Lalwani, G., Gella, S., He, H.: An empirical study on robustness to spurious correlations using pre-trained language models. Transactions of the Association for Computational Linguistics **8**, 621–633 (2020)
68. Udomcharoenchaikit, C., Ponwitayarat, W., Payoungkhamdee, P., Masuk, K., Buaphet, W., Chuangsuwanich, E., Nutanong, S.: Mitigating spurious correlation in natural language understanding with counterfactual inference. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 11308–11321 (2022)
69. Vapnik, V.: The nature of statistical learning theory. Springer science & business media (2013)
70. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
71. Wang, T., Sridhar, R., Yang, D., Wang, X.: Identifying and mitigating spurious correlations for improving robustness in nlp models. arXiv preprint arXiv:2110.07736 (2021)
72. Wang, Z., Culotta, A.: Identifying spurious correlations for robust text classification. arXiv preprint arXiv:2010.02458 (2020)
73. Wang, Z., Culotta, A.: Robustness to spurious correlations in text classification via automatically generated counterfactuals. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 14024–14031 (2021)
74. Wang, Z., Shu, K., Culotta, A.: Enhancing model robustness and fairness with causality: A regularization approach. arXiv preprint arXiv:2110.00911 (2021)
75. Wu, S., Yuksekgonul, M., Zhang, L., Zou, J.: Discover and cure: Concept-aware mitigation of spurious correlation. In: International Conference on Machine Learning. pp. 37765–37786. PMLR (2023)
76. Wu, Y., Gardner, M., Stenetorp, P., Dasigi, P.: Generating data to mitigate spurious correlations in natural language inference datasets. arXiv preprint arXiv:2203.12942 (2022)
77. Xiao, K., Engstrom, L., Ilyas, A., Madry, A.: Noise or signal: The role of image backgrounds in object recognition. arXiv preprint arXiv:2006.09994 (2020)
78. Yang, Y., Nushi, B., Palangi, H., Mirzasoleiman, B.: Mitigating spurious correlations in multi-modal models during fine-tuning. In: International Conference on Machine Learning. pp. 39365–39379. PMLR (2023)
79. Ye, W., Zheng, G., Cao, X., Ma, Y., Hu, X., Zhang, A.: Spurious correlations in machine learning: A survey. arXiv preprint arXiv:2402.12715 (2024)
80. Yenamandra, S., Ramesh, P., Prabhu, V., Hoffman, J.: Facts: First amplify correlations and then slice to discover bias. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4794–4804 (October 2023)
81. Zhang, M., Sohoni, N.S., Zhang, H.R., Finn, C., Ré, C.: Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. arXiv preprint arXiv:2203.01517 (2022)
82. Zhang, X., Cui, P., Xu, R., Zhou, L., He, Y., Shen, Z.: Deep stable learning for out-of-distribution generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5372–5382 (2021)
83. Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in neural information processing systems **31** (2018)
84. Zheng, G., Ye, W., Zhang, A.: Learning robust classifiers with self-guided spurious correlation mitigation. arXiv preprint arXiv:2405.03649 (2024)

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