Finding and Fixing Spurious Patterns with Explanations

Anonymous authors
Paper under double-blind review

Abstract

Machine learning models often use spurious patterns such as "relying on the presence of a person to detect a tennis racket," which do not generalize. In this work, we present an end-to-end pipeline for identifying and mitigating spurious patterns for image classifiers. We start by finding patterns such as "the model's prediction for tennis racket changes 63% of the time if we hide the people." Then, if a pattern is spurious, we mitigate it via a novel form of data augmentation. We demonstrate that this approach identifies a diverse set of spurious patterns and that it mitigates them by producing a model that is both more accurate on a distribution where the spurious pattern is not helpful and more robust to distribution shift.

1 Introduction

With the adoption of machine learning models in real-world applications, there is a growing concern about *Spurious Patterns* (SPs) – when models rely on patterns that do not align with domain knowledge and do not generalize (Ross et al., 2017; Shetty et al., 2019; Rieger et al., 2020; Teney et al., 2020; Singh et al., 2020). For example, a model trained to detect tennis rackets on the COCO dataset (Lin et al., 2014) learns to rely on the presence of a person, which leads to systemic errors: it is significantly less accurate at detecting tennis rackets for images without people than with people (41.2% vs 86.6%) and only ever has false positives on images with people. Relying on this SP works well on COCO, where the vast majority of images with tennis rackets also have people, but would not be as effective for other distributions. Further, relying on SPs may also lead to serious concerns when they relate to protected attributes such as race or gender (Buolamwini & Gebru, 2018).

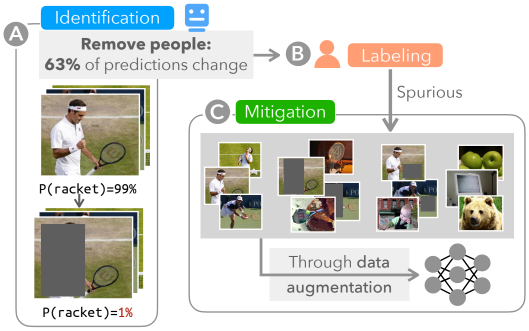

Figure 1: For the tennis racket example, SPIRE identifies this pattern by observing that, when we remove the people from images with both a tennis racket and a person, the model's prediction changes 63% of the time. Since this process does not remove the tennis racket itself, we label this pattern as spurious. Then, SPIRE carefully adds/removes tennis rackets/people from different images to create an augmented training set where tennis rackets and people are independent while minimizing any new correlations between the label and artifacts in the counterfactual images (e.g., grey boxes).

We focus on SPs where an image classification model is relying on a spurious object (e.g., using people to detect tennis rackets) and we propose Spurious Pattern Identification and REpair[1] (SPIRE) as an end-to-end solution for these SPs. As illustrated in Figure 1, SPIRE *identifies* which patterns the model is using by measuring how often it makes different predictions on the original and counterfactual versions of an image. Since it reduces a pattern to a single value that has a clear interpretation, it is easy for a user to (when needed) label that pattern as spurious or valid. Then, it *mitigates* SPs by retraining the model using a novel form of data augmentation that aims to shift the training distribution towards the *balanced distribution*, a distribution where the SP is no longer helpful, while minimizing any new correlations between the label and artifacts in the counterfactual images (e.g., the grey boxes from removing people).

[1] Code will be released at https://github.com/user/repo

In order to verify that the baseline model relies on a SP and quantify the impact of mitigation methods, we measure *gaps* in accuracy between images with and without the spurious object (e.g., there is a 45.4% accuracy drop between images of tennis rackets with and without people). Intuitively, the more a model relies on a SP, the larger these gaps will be and the less robust the model is to distribution shift. Additionally, we measure performance on the balanced distribution. Then, an effective mitigation method will decrease the gap metrics and improve performance on the balanced distribution. Empirically, we demonstrate SPIRE's effectiveness with three sets of experiments:

- *Benchmark Experiments.* We induce SPs with varying strengths by sub-sampling COCO in order to better understand how mitigation methods work in a controlled setting. Overall, we find that SPIRE is substantially more effective than prior methods. Interestingly, we also find that most prior methods are ineffective at mitigating negative SPs (i.e., SPs where the presence of the spurious object is negatively associated with the label).
- *Full Experiment.* We show that SPIRE is useful "in the wild" on the full COCO dataset. For identification, it finds a diverse set of SPs and is the first method to identify negative SPs (e.g., neck ties and cats), and, for mitigation, it is more effective than prior methods. Additionally, we show that it improves zero-shot generalization (i.e., evaluation without any re-training) to two challenging datasets: UnRel (Peyre et al., 2017) and SpatialSense (Yang et al., 2019). Collectively, these results are notable because most methods produce no improvements in terms of robustness to natural distribution shifts (Taori et al., 2020).
- *Generalization Experiments.* We illustrate how SPIRE generalizes beyond the setting from our prior experiments, where we considered the object-detection task and assumed that the dataset has annotations to use to create counterfactual images. Specifically, we explore three examples that consider a different task and/or do not make this assumption.

2 Related Work

We discuss prior work as it pertains to identifying and mitigating SPs for image classification models.

Identification. While several prior works measure the extent to which the model is relying on "context" in a general sense (e.g., the model is relying on *something* other than the tennis racket to detect the tennis racket) (Shetty et al., 2019; Agarwal et al., 2020; Xiao et al., 2021), SPIRE identifies specific SPs (e.g., the model is relying on *the person* to detect the tennis racket). For identifying specific SPs, the most common approach uses explainable machine learning (Simonyan et al., 2013; Ribeiro et al., 2016; Selvaraju et al., 2017; Ross et al., 2017; Singh et al., 2018; Dhurandhar et al., 2018; Goyal et al., 2019; Koh et al., 2020; Joo & Kärkkäinen, 2020; Rieger et al., 2020). For image datasets, these methods rely on local explanations, resulting in a slow process that requires the user to look at the explanation for each image, infer what that explanation is telling them, and then aggregate those inferences to assess whether or not they represent a consistent pattern (Figure 2). In addition to this procedural difficulty, there is uncertainty about the usefulness of some of these explanations for model debugging (Adebayo et al., 2018; 2020).

In contrast, SPIRE inherently avoids these challenges by measuring the aggregated effect that a specific counterfactual has on the model's predictions (e.g., the model's prediction changes 63% of the time when we remove the people from images with both a tennis racket and a person). Note that our proposed explanations fall into the broad definition of a global counterfactual explanation described in (Plumb et al., 2020); however, our technical approach is different. Singh et al. (2020) follow a similar principle and look for object pairs such that the presence of one object increases the prediction probability of the other object. This method relies on the existence of images of objects in isolation, which may be rare (e.g., the entire COCO dataset contains 34 images of tennis rackets without people), while SPIRE counterfactually generates such images.

Mitigation. While prior work has explored data augmentation for mitigation (Hendricks et al., 2018; Shetty et al., 2019; Teney et al., 2020; Chen et al., 2020a; Agarwal et al., 2020), it has done so with augmentation strategies that are agnostic to the training distribution (e.g., QCEC (Shetty et al., 2019) simply removes either the tennis rackets or the people, as applicable, uniformly at random for each image). In contrast, SPIRE aims to use counterfactual images to create a training distribution where the label is independent of the spurious object, while minimizing any new correlations between the label and artifacts in the counterfactual images.

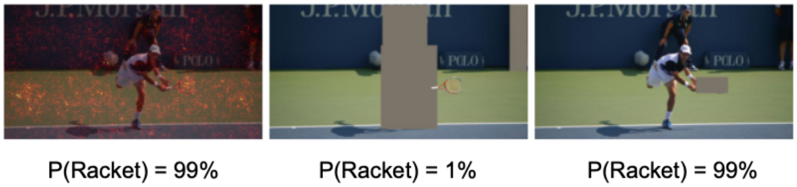

Figure 2: Based on the saliency map (Simonyan et al., 2013) (Left), one might mistakenly infer that the model is not relying on the person. However, the model fails to detect the racket after the person is removed (Center) and incorrectly detects a racket after it is removed (Right).

Another line of prior work adds regularization to the model training process (Ross et al., 2017; Hendricks et al., 2018; Wang et al., 2019; Rieger et al., 2020; Teney et al., 2020; Liang et al., 2020; Singh et al., 2020). Some of these methods specify which parts of the image should not be relevant to the model's prediction (Ross et al., 2017; Rieger et al., 2020). Other methods encourage the model's predictions to be consistent across counterfactual versions of the image (Hendricks et al., 2018; Teney et al., 2020; Liang et al., 2020). All of these methods could be used in conjunction with SPIRE.

Finally, there are two additional lines of work that make different assumptions. Making weaker assumptions, there are methods based on sub-sampling, re-weighting, or grouping the training set (Chawla et al., 2002; Sagawa* et al., 2020; Creager et al., 2020). These methods have been found to be less effective than methods that use data augmentation or regularization (Rieger et al., 2020; Neto, 2020; Singh et al., 2020; Goel et al., 2021). Making stronger assumptions, there are methods from domain adaptation, which assume access to several distinct training distributions (Wen et al., 2020; Chen et al., 2020b); we do not assume this.

Consequently, the methods designed for image classification that use data augmentation or regularization represent SPIRE's most direct competition. As a result, we compare against "Right for the Right Reasons" (RRR) (Ross et al., 2017), "Quantifying and Controlling the Effects of Context" (QCEC) (Shetty et al., 2019), "Contextual Decomposition Explanation Penalization" (CDEP) (Rieger et al., 2020), "Gradient Supervision" (GS) (Teney et al., 2020), and the "Feature Splitting" (FS) method from (Singh et al., 2020). Note that, with the exception of FS, all of these methods require dataset annotations for the location of the spurious object; we make the same assumption in Sections 5.1 and 5.2, but explore relaxing it in Section 5.3.

3 Spurious Pattern Identification and REpair

In this section, we explain SPIRE's approach for addressing SPs. We use the object detection task as a running example, where *Main* is the object being detected and *Spurious* is the other object in a SP.[2]

Preliminaries. We view a dataset as a probability distribution over a set of *image splits*, which we call *Both*, *Just Main*, *Just Spurious*, and *Neither*, depending on which of Main and/or Spurious they contain (e.g., Just Main is the set of images with tennis rackets but no people). Figure 3 (Left) shows these splits for the tennis racket example. Note that we can take an image from one split and create a counterfactual version of it in a different split by either adding or removing either Main or Spurious (e.g., removing the people from an image in Both moves it to Just Main) (see Appendix B.1).

Identification. SPIRE measures the degree to which the model relies on Spurious to detect Main by measuring the probability that, when we remove Spurious from a training image from Both, the model's prediction changes (e.g., the model's prediction for tennis racket changes 63% of the time when we remove the people from an image with both a tennis racket and a person). Intuitively, the higher this probability, the stronger this pattern is.

To identify the full set of patterns that the model is using, SPIRE measures this probability for all (Main, Spurious) pairs, where Main and Spurious are different, and then sorts this list to find the pairs that represent the strongest patterns. Recall that not all patterns are necessarily spurious and that the user may label patterns as spurious or valid as necessary before moving to the mitigation step.

[2] The same methodology directly applies to any binary classification with a binary "spurious feature."

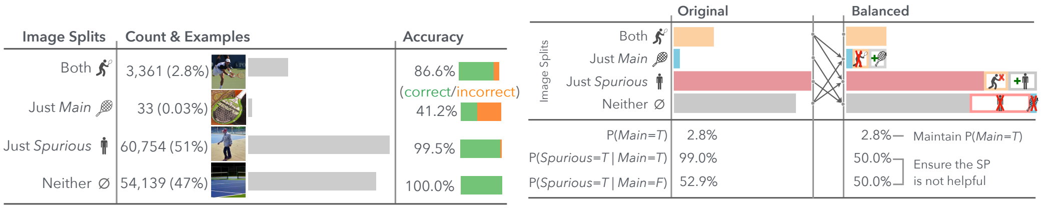

Figure 3: **Left.** The training image splits and the original training distribution for the tennis racket example. Because of the strong positive correlation between Main and Spurious, it is helpful for the model to rely on this SP. **Right.** The balanced distribution for the tennis racket example. This SP is no longer helpful because Main and Spurious are now independent and there are the same number of images in Both and Just Main.

Mitigation. It is often the case that there is a strong correlation between Main and Spurious in the original training distribution, which incentivizes the model to rely on this SP. As a result, we want to define a distribution, which we call the *balanced distribution*, where relying on this SP is neither inherently helpful nor harmful. This is a distribution, exemplified in Figure 3 (Right), that:

- *Preserves P(Main).* This value strongly influences the model's relative accuracy on {Both, Just Main} versus {Just Spurious, Neither} but does not incentivize the SP. As a result, we preserve it in order to maximize the similarity between the original and balanced distributions.
- *Sets P(Spurious | Main) = P(Spurious | not Main) = 0.5.* This makes Main and Spurious independent, which removes the statistical benefit of relying on the SP, and assigns equal importance to images with and without Spurious. However, this does not go so far as to invert the original correlation, which would directly punish reliance on the SP.

As shown in Figure 3 (Right), SPIRE's mitigation strategy uses counterfactual images to manipulate the training distribution. The specifics are described in Section 3.1, but they implement two goals:

- *Primary: Shift the training distribution towards the balanced distribution.* While the original training distribution often incentivizes the model to rely on the SP, the balanced distribution does not. However, adding too many counterfactual images may compromise the model's accuracy on natural images. As a result, we want to shift the training distribution towards, but not necessarily all the way to, the balanced distribution.
- *Secondary: Minimize the potential for new SPs.* While shifting towards the balanced distribution, we may inadvertently introduce new potential SPs between Main and artifacts in the counterfactual images. For example, augmenting the dataset with the same counterfactuals that SPIRE uses for identification (i.e., images from Both where Spurious has been covered with a grey box) introduces the potential for a new SP because P(Main | "Artifact") = 1.0 where, in this case, the "Artifact" is "grey boxes". Because the augmentation will be less effective if the model learns to rely on new SPs, we minimize their potential by trying to set P(Main | Artifact) = 0.5.

3.1 Specific Mitigation Strategies

While SPIRE's augmentation strategy follows the aforementioned goals, its specific details depend on the problem setting, which we characterize using two factors:

- *Can the counterfactuals change an image's label?* For tasks such as object detection, counterfactuals can change an image's label by removing or adding Main. However, for tasks such as scene identification, we may not have counterfactuals that can change an image's label. For example, we cannot turn a runway into a street or a street into a runway by manipulating a few objects. Fundamentally, this defines the space of counterfactuals an augmentation strategy can use.
- *Is the dataset class balanced?* While working with class balanced datasets drastically simplifies the problem and analysis, it is not an assumption that usually holds in practice.

These two factors define the three problem settings that we consider, which correspond to the experiments in Sections 5.1, 5.2, and 5.3 respectively. For each setting, we summarize what makes it interesting, define SPIRE's specific augmentation strategy for it, and then discuss how that strategy meets SPIRE's primary and secondary goals.

Table 1: Setting 1. For $p = 0.9$ and $p = 0.1$, we show the original size of each split for a dataset of size 200 as well as the size of each split after SPIRE's or QCEC's augmentation. Note that SPIRE produces the balanced distribution, while QCEC does not even make Main and Spurious independent.

	p = 0.9			p = 0.1		
Split	Original	SPIRE	QCEC	Original	SPIRE	QCEC
Both	90	90	90	10	90	10
Just Main	10	90	55	90	90	95
Just Spurious	10	90	55	90	90	95
Neither	90	90	110	10	90	190

Setting 1: Counterfactuals can change an image's label and the dataset is class-balanced. Because we have class balance, P(Main) = P(Spurious) = 0.5 and we can specify the training distribution by specifying p = P(Main | Spurious). If $p > 0.5$, SPIRE moves images from $\{$Both, Neither$\}$ to $\{$Just Main, Just Spurious$\}$ with probability $\frac{2p-1}{2p}$ for each of those four combinations. If $p < 0.5$, SPIRE moves images from $\{$Just Main, Just Spurious$\}$ to $\{$Both, Neither$\}$ with probability $\frac{p-0.5}{p-1}$.

Table 1 shows how SPIRE changes the training distributions for $p = 0.9$ and $p = 0.1$. For $p = 0.9$, it succeeds at both of its goals. For $p = 0.1$, it produces the balanced distribution, but does add the potential for new SPs because P(Main | Removed an object) = 0 and P(Main | Added an object) = 1. We contrast SPIRE to the most closely related method, QCEC (Shetty et al., 2019), which removes either Main or Spurious uniformly at random from each image. For both values of p, QCEC does not make Main and Spurious independent and adds the potential for new SPs. This example highlights the fact that, while prior work has used counterfactuals for data augmentation, SPIRE uses them in a fundamentally different way by considering the training distribution.

Setting 2: Counterfactuals can change an image's label, but the dataset has class imbalance. In the presence of significant class imbalance, two parts of the definition of the balanced distribution become problematic for an augmentation strategy:

- *Preserves P(Main).* When P(Main) is small, this constraint requires that we generate more counterfactual images without Main than with it, which can introduce new potential SPs.
- *Sets P(Spurious | not Main) = 0.5.* When P(Spurious) is also small, this constraint requires that most of the counterfactual images we generate belong to Just Spurious, which can lead to the counterfactual data outnumbering the original data by a factor of 100 or more for this split.

Consequently, we relax these constraints. If P(Spurious | Main) > P(Spurious), SPIRE creates an equal number of images to add to Just Main/Spurious by removing the appropriate object from an image from Both. Specifically, this number is the smallest positive solution to: $\frac{|\text{Both}|}{|\text{Both}|+|\text{Just Spurious}|+\delta} = \frac{|\text{Just Main}|+\delta}{|\text{Just Main}|+|\text{Neither}|+\delta}$. Otherwise, SPIRE creates an equal number of images to add to Both/Just Spurious by adding Spurious to Just Main/Neither. Specifically, this number solves: $\frac{|\text{Both}|+\delta}{|\text{Both}|+|\text{Just Spurious}|+2\delta} = \frac{|\text{Just Main}|}{|\text{Just Main}|+|\text{Neither}|}$. In both cases, we cap δ to be no larger than the smallest source split. SPIRE achieves its primary goal by making P(Main | Spurious) = P(Main | not Spurious) (i.e., Main and Spurious are now independent) and it achieves its secondary goal by adding an equal number of counterfactual images with and without Main (i.e., P(Main | Artifact) = 0.5).

Setting 3: Counterfactuals cannot change an image's label Because of this new constraint, the previously described augmentation strategies cannot be applied. As a result, SPIRE removes Spurious from every image with it and adds Spurious to every image without it. While this does achieve its primary goal, it does not achieve its secondary goal (e.g., the correlation between the label and grey boxes from removing Spurious is the same as the correlation between the label and Spurious).

4 Evaluation

Because relying on the SP is usually helpful on the original distribution, we cannot measure the effectiveness of a mitigation method using that distribution. Instead, we measure the model's performance on the balanced distribution using metrics such as accuracy and average precision. Intuitively, using the balanced distribution provides a fairer comparison because the SP is neither helpful nor harmful on it. However, like any performance metric that is aggregated over a distribution, these metrics hide potentially useful details and are dependent on the distribution itself.

We address these limitations by measuring the model's accuracy on each of the image splits.[3] These per split accuracies yield a more detailed analysis and, further, allow us to calculate two "gap metrics," which give us a distribution-independent form of evaluation. The *Recall Gap* is the difference in accuracy between Both and Just Main; the *Hallucination Gap* is the difference in accuracy between Neither and Just Spurious. Intuitively, a smaller recall gap means that the model is more robust to distribution shifts that move weight between Both and Just Main. The same is true for the hallucination gap and shifts between Neither and Just Spurious. As a concrete example of these metrics, consider the tennis racket example (Figure 3 Left), where we observe that the recall gap is 45.4% (i.e., the model is much more likely to detect a tennis racket when a person is present) and a hallucination gap of 0.5% (i.e., the model is more likely to hallucinate a tennis racket when a person is present; see Appendix C.2).

It is important to note that these per split accuracies are measured using only natural (i.e., not counterfactual) images, in order to prevent the model from "cheating" by learning to use artifacts in the counterfactual images. As a result, the gap metrics and the performance on the balanced distribution also only use natural images because they are estimated from these accuracies.

Class Balanced vs Imbalanced Evaluation. When the dataset is class balanced, we use the standard prediction threshold of 0.5 to measure a model's performance using accuracy on the balanced distribution (i.e., *balanced accuracy*) and its gap metrics. When there is class imbalance, Average Precision (AP), which is the area under the precision vs recall curve, is the standard performance metric. Analogous to AP, we can calculate the Average Recall Gap by finding the area under the "absolute value of the recall gap" vs recall curve; the Average Hallucination Gap is defined similarly. As a result, we measure a model's performance using AP on the balanced distribution (i.e., *balanced AP*) and the Average Recall/Hallucination Gaps.

5 Experiments

We divide our experiments into three groups:

- In Section 5.1, we induce SPs with varying strengths by sub-sampling COCO in order to better understand how mitigation methods work in a controlled setting. We show that SPIRE is more effective at mitigating these SPs than prior methods. We also use these results to identify the best prior method, which we use for comparisons for the remaining experiments.
- In Section 5.2, we find and fix SPs "in the wild" using all of COCO; this means finding multiple naturally occurring SPs and fixing them simultaneously. We show that SPIRE identifies a wider range of SPs than prior methods and that it is more effective at mitigating them. Additionally, we show that it improves zero-shot generalization to two challenging datasets (UnRel and SpatialSense).
- In Section 5.3, we show how SPIRE generalizes beyond the setting considered for COCO. Specifically, we consider tasks other than object-detection and/or not using dataset annotations to create counterfactual images.

For the baseline models (i.e., the normally trained models that contain the SPs that we are going to identify and mitigate), we fine-tune a pre-trained version of ResNet18 (He et al., 2016) (Appendix D). We compare SPIRE to RRR (Ross et al., 2017), QCEC (Shetty et al., 2019), CDEP (Rieger et al., 2020), GS (Teney et al., 2020), and FS (Singh et al., 2020). We use the evaluation described in Section 4 and, for any split that is too small to produce a reliable accuracy estimate, we acquire additional images (using Google Images) such that each split has at least 30 images to use for evaluation.

5.1 Benchmark Experiments

We construct a set of benchmark tasks from COCO consisting of different SPs with varying strengths, by manipulating the model's training distribution, in order to better understand how mitigation methods work in a controlled setting. Appendix E has additional details.

Creating the benchmark. We start by finding each pair of objects that has at least 100 images in each split of the testing set (13 pairs). For each of those pairs, we create a series of controlled training

[3] In fact, we have to do this in order to estimate the model's performance on the balanced distribution because the only way we can estimate metrics like accuracy and average precision is by re-weighting the model's accuracy on each split. The same process allows us to estimate the model's performance on any distribution.

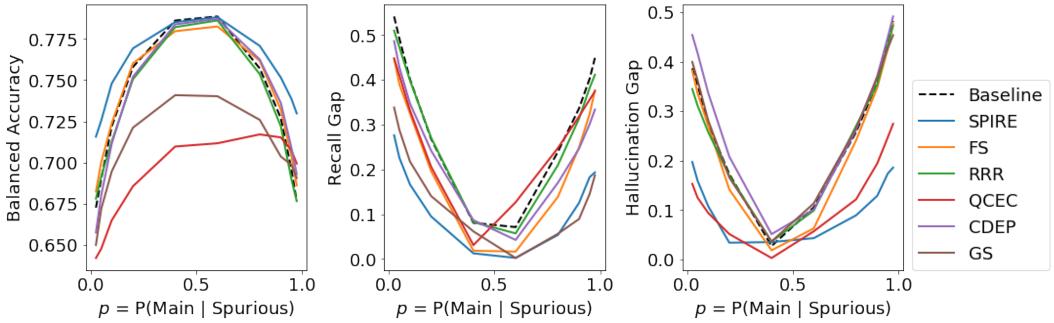

Figure 4: A comparison of the baseline model to various mitigation methods. The results shown are averaged across both the pairs accepted for our benchmark and across eight trials. **Left - Balanced Accuracy.** For $p \leq 0.2$ and $p \geq 0.8$, SPIRE produces the most accurate models. None of the methods have much of an impact for $p = 0.4$ or $p = 0.6$, likely because those create weak SPs. **Center/Right - Recall/Hallucination Gaps.** SPIRE generally shrinks the absolute value of both of the gap metrics by more than prior methods.

sets of size 2000 by sampling images from the full training set such that P(Main) = P(Spurious) = 0.5 and p = P(Main | Spurious) ranges between 0.025 and 0.975. Each controlled training set represents a binary task, where the goal is to predict the presence of Main.

While varying p allows us to control the strength of the correlation between Main and Spurious (i.e., p near 0 indicates a strong negative correlation while p near 1 indicates a strong positive correlation), it does not guarantee that the model actually relies on the intended SP. Indeed, when measure the models' balanced accuracy as p varies, we observe that 5 out of the 13 pairs show little to no loss in balanced accuracy as p approaches 1. Consequently, subsequent evaluation considers the other 8 pairs. For these pairs, the model's reliance on the SP increases as p approaches 0 or 1 as evidenced by the increasing loss of balanced accuracy and confirmed via counterfactual evaluation.

Results. Figure 4 (Left) presents the balanced accuracy results. We find that SPIRE consistently improves balanced accuracy and that it does so by significantly more than prior methods. Interestingly, while most prior methods are beneficial for strong positive SPs ($p \geq 0.9$), only FS is also (mildly) beneficial for negative SPs ($p < 0.5$).

Figure 4 (Center/Right) presents the gap metric results. We find that SPIRE is the most effective method at shrinking these metrics, which indicates that it produces a model that is more robust to distribution shift. Interestingly, QCEC and GS, which are the two prior methods that include some form of data augmentation, are the only prior methods that substantially shrink the gap metrics; however, they do so at the cost of balanced accuracy for $p < 0.9$.

Overall, this experiment shows that SPIRE is an effective mitigation method and that our evaluation framework enables us to easily understand how methods affect the behavior of a model. We use FS as the baseline for comparison for the remaining experiments because, of the prior methods, it had the best average balanced accuracy across p's range.

5.2 Full Experiment

We evaluate SPIRE "in the wild" by identifying and mitigating SPs learned by a multi-label binary object-detection model trained on the full COCO dataset. Appendix F has additional details.

Identification. Out of all possible (Main, Spurious) pairs, we consider those which have at least 25 training images in Both ($\approx$ 2700). From these, SPIRE identifies 29 where the model's prediction changes at least 40% of the time when we remove Spurious; we verified that the model is relying on these SPs by checking that it has large recall and hallucination gaps. Table 2 shows a few of the identified SPs; overall, they are quite diverse: the spurious object ranges from common (e.g., person) to rare (e.g., sheep); the SPs range from objects that are commonly co-located (e.g., tie-person) to usually separate (e.g., dog-sheep); and a few Main objects (e.g., tie and frisbee) have more than one associated SP. Notably, SPIRE identifies negative SPs (e.g., tie-cat) while prior work (Shetty et al., 2019; Singh et al., 2020; Teney et al., 2020) has only found positive SPs (e.g., frisbee-person).

Table 2: A few examples of the SPs identified by SPIRE for the Full Experiment. For each pair, we report several basic dataset statistics including *bias*, $\frac{\text{P(Spurious | Main) - P(Spurious)}}{\text{P(Spurious)}}$, which captures how far Main and Spurious are away from being independent as well as the sign of their correlation.

Main	Spurious	P(M)	P(S)	P(S | M)	bias
tie	cat	0.03	0.04	0.01	-0.66
toothbrush	person	0.01	0.54	0.54	-0.01
bird	sheep	0.03	0.01	0.01	0.00
frisbee	person	0.02	0.54	0.83	0.54
tie	person	0.03	0.54	0.95	0.76
tennis racket	person	0.03	0.54	0.99	0.83
dog	sheep	0.04	0.01	0.03	1.05
frisbee	dog	0.02	0.04	0.24	5.44
fork	dining table	0.03	0.10	0.76	6.56

Table 3: Mitigation results for the Full Experiment. Balanced AP is averaged across the SPs identified by SPIRE. Similarly, the gap metrics are reported as the "mean (median)" change from the baseline model, aggregated across those SPs.

	Original MAP	Balanced AP	%Δ Avg. Recall Gap	%Δ Avg. Hallucination Gap
Baseline	**64.1**	46.2	—	—
SPIRE	63.7	**47.3**	**-14.2 (-14.5)**	**-28.1 (-27.3)**
FS	62.5	44.7	9.7 (-5.9)	25.7 (-6.9)

Table 4: The MAP results of a zero-shot evaluation on the classes that are in the UnRel/SpatialSense datasets that SPIRE also identified as being Main in a SP.

	UnRel	SpatialSense
Baseline	38.9	20.3
SPIRE	**41.3**	**20.7**
FS	39.6	18.6

Mitigation. Unlike the Benchmark Experiments, this experiment requires mitigating many SPs simultaneously. We do this by re-training the slice of the model's final layer that corresponds to Main's class on an augmented dataset that combines SPIRE's augmentation for each SP associated with Main. All results shown (Tables 3 and 4) are averaged across eight trials.

We conclude that SPIRE significantly reduces the model's reliance on these SPs based on two main observations. First, it increases balanced AP by 1.1% and shrinks the average recall/hallucination gaps by a factor of 14.2/28.1%, relative to the baseline model, on COCO. As expected, this does slightly decrease Mean Average Precision (MAP) by 0.4% on the original (biased) distribution. Second, it increases MAP on the UnRel (Peyre et al., 2017) and SpatialSense (Yang et al., 2019) datasets. Because this evaluation was done in a zero-shot manner and these datasets are designed to have objects in unusual contexts, this is further evidence that SPIRE improves performance and distributional robustness. In contrast, FS decreases the model's performance, has inconsistent effects on the gap metrics, and has mixed results on the zero-shot evaluation.

SPIRE and Distributional Robustness. Noting that robustness to specific distribution shifts is one of the consequences of mitigating SPs, we can contextualize the impact of SPIRE by considering an extensive meta-analysis of methods that aim to provide general robustness (Taori et al., 2020). This analysis finds that the only methods that consistently work are those that re-train the baseline model on several orders of magnitude more data. It also describes two necessary conditions for a method to work. Notably, SPIRE satisfies both of those conditions: first, it improves performance on the shifted distributions (i.e., the balanced distributions, UnRel, and SpatialSense) and, second, this improvement cannot be explained by increased performance on the original distribution. Consequently, SPIRE 's results are significant because they show improved robustness without using orders of magnitude more training data. We hypothesize that SPIRE is successful because it targets specific SPs that the baseline model relies on rather than using a less targeted approach.

5.3 Generalization Experiments

We illustrate how SPIRE generalizes beyond the setting from our prior experiments, where we considered the object-detection task and assumed that the dataset has annotations to use to create counterfactual images. Specifically, we explore three examples that consider a different task (*Generalization 1*) and/or do not make this assumption (*Generalization 2*).

Scene Identification Experiment (Generalization 1). In this experiment, we construct a scene identification task using the image captions from COCO and show that SPIRE can identify and mitigate a naturally occurring SP. To do this, we define two classes: one where the word "runway" (the part of an airport where airplanes land) appears in the caption (1,134 training images) and another where "street" appears (12,543 training images); images with both or without either are discarded. For identification, we observe that removing all of the airplanes from an image of a runway changes the model's prediction 50.7% of the time and label this pattern as a SP.

Table 5: Results for the Scene Identification Experiment (averaged across sixteen trials).

	Original AP	Balanced AP	%Δ Avg. Recall Gap	%Δ Avg. Hallucination Gap
Baseline	**95.0**	48.9	—	—
SPIRE	92.8	**83.2**	**-82.1**	**-75.5**
FS	93.7	47.8	-11.5	3.7

Table 5 shows the results. SPIRE reduces the model's reliance on this SP because it substantially increases balanced AP and it reduces the average recall and hallucination gaps by factors of 82.1% and 75.5%. In contrast, FS is not effective at mitigating this SP.

No Object Annotation Experiment (Generalization 2). In this experiment, we mitigate the SP from the tennis racket example without assuming that we have pixel-wise object-annotations to create counterfactuals. Instead, we train a linear (in the model's representation space) classifier to predict whether or not an image contains a person (similar to Kim et al. (2018)). Then, we project across this linear classifier to essentially add or remove a person from an image's *representation* (so we call this method SPIRE-R).

Table 6: Results for the No Object Annotation Experiment for the tennis racket example (averaged across eight trials).

	Original AP	Balanced AP	%Δ Avg. Recall Gap	%Δ Avg. Hallucination Gap
Baseline	93.9	79.9	—	—
SPIRE	92.9	80.5	**-31.3**	-27.3
SPIRE-R	**94.0**	**81.0**	-9.1	**-44.9**
FS	92.9	80.7	-10.4	-22.3

Table 7: Results for the ISIC Experiment (averaged across eight trials).

	Original AP	Balanced AP	%Δ Avg. Recall Gap	%Δ Avg. Hallucination Gap
Baseline	78.3	71.0	—	—
SPIRE-EM	**78.8**	**76.4**	-20.5	-39.0
FS	70.7	68.0	**-31.3**	**-61.0**

Table 6 shows the results. There are two main results to note. First, that SPIRE provides a small increase in Balanced AP while providing the largest average decrease in the gap metrics. Second, that SPIRE-R is preferable to FS because it produces a larger reduction in the hallucination gap while being comparable otherwise.

ISIC Experiment (Generalizations 1 & 2). In this experiment, we imitate the setup from (Rieger et al., 2020) for the ISIC dataset (Codella et al., 2019). Specifically: the task is to predict whether an image of a skin lesion is malignant or benign; the model learns to use a SP where it relies on a "brightly colored sticker" that is spuriously correlated with the label; and the dataset does not have annotations for those stickers to use create counterfactual images. For this experiment, we illustrate another approach for working without annotations: using *external models* (so we call this method SPIRE-EM) to produce counterfactual images. The external model could be an off-the-shelf model (e.g., a model that locates text in an image or a GAN) or a simple pipeline such as the super-pixel clustering one we use (see Appendix G.1).

Table 7 shows the results. We can see that SPIRE-EM is effective at mitigating this SP because it generally improves performance while also shrinking the gap metrics.[4] In contrast, FS does not seem to be beneficial because it substantially reduces performance on both the original and balanced distributions (which outweighs shrinking the gap metrics).

6 Conclusion

In this work, we introduced SPIRE as an end-to-end solution for addressing Spurious Patterns for image classification models that are relying on spurious objects to make predictions. SPIRE identifies potential SPs by measuring how often the model's prediction changes when we remove the Spurious object from an image with a positive label and mitigates SPs by shifting the training distribution towards the balanced distribution while minimizing any correlations between the label and artifacts in the counterfactual images. We demonstrated that SPIRE is able identify and, at least partially, mitigate a diverse set of SPs by improving the model's performance on the balanced distribution and by making it more robust to specific distribution shifts. We found that these improvements lead to improved zero-shot generalization to challenging datasets. Finally, we showed that SPIRE can be applied to tasks other than object detection and we illustrated two potential ways to apply SPIRE when there are no dataset annotations to use to create counterfactuals (creating counterfactual representations and using external models).

[4] Note that, because the dataset does not contain enough malignant images with stickers to produce a reliable accuracy estimate, we had to use counterfactual images for this evaluation. This may influence the results for balanced AP and the recall gap metric. However, SPIRE-EM is still an improvement over the baseline because of the original AP and hallucination gap metric results.

7 Ethics Statement

While developing tools to identify and mitigate SPs should be useful for helping address some of the ethical concerns around applying machine learning, these tools should not be viewed as a panacea for addressing such concerns. Moreover, it is not impossible that similar ideas could be used to identify and then exacerbate harmful patterns.

8 Reproducibility Statement

We provide a description of the data processing and model training process in Appendix D. Additionally, we provide a link to the source code that was used to run all of the experiments in the supplemental material.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31:9505–9515, 2018.

Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. Debugging tests for model explanations. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 700–712. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/075b051ec3d22dac7b33f788da631fd4-Paper.pdf.

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9690–9698, 2020.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91, 2018.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10800–10809, 2020a.

Yining Chen, Colin Wei, Ananya Kumar, and Tengyu Ma. Self-training avoids using spurious features under domain shift. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21061–21071. Curran Associates, Inc., 2020b. URL https://proceedings.neurips.cc/paper/2020/file/f1298750ed09618717f9c10ea8d1d3b0-Paper.pdf.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, 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.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Exchanging lessons between algorithmic fairness and domain generalization. *arXiv preprint arXiv:2010.07249*, 2020.

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *Advances in neural information processing systems*, 31:592–603, 2018.

Karan Goel, Albert Gu, Yixuan Li, and Christopher Re. Model patching: Closing the subgroup performance gap with data augmentation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=9YlaeLfuhJF.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *International Conference on Machine Learning*, pp. 2376–2384, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 771–787, 2018.

Jungseock Joo and Kimmo Kärkkäinen. Gender slopes: Counterfactual fairness for computer vision models by attribute manipulation. In *Proceedings of the 2nd International Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia*, pp. 1–5, 2020.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

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

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pp. 5338–5348. PMLR, 2020.

Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. Learning to contrast the counterfactual samples for robust visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3285–3292, 2020.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. In *International Conference on Learning Representations*, 2020.

Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*, 2019.

Elias Chaibub Neto. Counterfactual confounding adjustment for feature representations learned by deep models: with an application to image classification tasks. *arXiv preprint arXiv:2004.09466*, 2020.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Weakly-supervised learning of visual relations. In *ICCV*, 2017.

Gregory Plumb, Jonathan Terhorst, Sriram Sankararaman, and Ameet Talwalkar. Explaining groups of points in low-dimensional representations. In *International Conference on Machine Learning*, pp. 7762–7771. PMLR, 2020.

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*, pp. 1135–1144, 2016.

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*, pp. 8116–8126. PMLR, 2020.

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*, pp. 2662–2670, 2017.

Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=ryxGuJrFvS`.

Axel Sauer and Andreas Geiger. Counterfactual generative networks. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=BXewfAYMmJw`.

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*, pp. 618–626, 2017.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *NeurIPS*, 2020.

Rakshith Shetty, Bernt Schiele, and Mario Fritz. Not using the car to see the sidewalk–quantifying and controlling the effects of context in classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8218–8226, 2019.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Chandan Singh, W James Murdoch, and Bin Yu. Hierarchical interpretations for neural network predictions. In *International Conference on Learning Representations*, 2018.

Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don't judge an object by its context: Learning to overcome contextual bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11070–11078, 2020.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *arXiv preprint arXiv:2007.00644*, 2020.

Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. Learning what makes a difference from counterfactual examples and gradient supervision. *arXiv preprint arXiv:2004.09034*, 2020.

Angelina Wang, Arvind Narayanan, and Olga Russakovsky. Revise: A tool for measuring and mitigating bias in visual datasets. In *European Conference on Computer Vision*, pp. 733–751. Springer, 2020.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5310–5319, 2019.

Jun Wen, Changjian Shui, Kun Kuang, Junsong Yuan, Zenan Huang, Zhefeng Gong, and Nenggan Zheng. Interventional domain adaptation. *arXiv preprint arXiv:2011.03737*, 2020.

Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=gl3D-xY7wLq`.

Kaiyu Yang, Olga Russakovsky, and Jia Deng. Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2051–2060, 2019.

Yi Zhang and Jitao Sang. Towards accuracy-fairness paradox: Adversarial example-based data augmentation for visual debiasing. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 4346–4354, 2020.