Finding and Fixing Spurious Patterns with Explanations

Gregory Plumb[1], Marco Tulio Ribeiro[2], and Ameet Talwalkar[1]

CMU[1], Microsoft Research[2]

Summary

- Image classifiers often use spurious patterns (SPs) such as "relying on the presence of a person to detect a tennis racket."
- We present a pipeline for identifying and mitigating such SPs.
- We identify patterns such as "the model's prediction for tennis racket changes 63% of the time if we hide the people."
- If a pattern is spurious, we mitigate it via a novel form of data augmentation.
- This approach identifies a diverse set of SPs and mitigates them by producing a model that is 1) more accurate on a distribution where the SP is not helpful and 2) more robust to distribution shift.

Background

We focus on spurious patterns (SPs) where an image classifier is relying on one object, called *Spurious*, to detect another object, called *Main*.

We view a dataset as a probability distribution over a set of *image splits* (Fig. 1).

Balanced Distribution

Relying on a SP is neither helpful nor harmful on the *balanced distribution* (Fig. 2).

We measure a model's performance on this distribution to see how effectively this SP was mitigated.

Gap Metrics

We use two *gap metrics* to measure how robust a model is to distribution shifts related to a SP.

For example, the model has a large *Recall Gap* because it is 45.4% more likely to detect a tennis racket when a person is present (Fig. 1).

Our Method: SPIRE

Counterfactuals. By adding or removing either Main or Spurious, we can move an image from one split to another. For example, we can move an image from Both to Just Main by removing Spurious.

Identifying SPs. SPIRE identifies potential SPs by measuring how often removing Spurious from an image in Both changes the model's prediction.

Mitigating SPs. The key to SPIRE's mitigation strategy is that it shifts the original distribution towards the balanced distribution (Fig. 2).

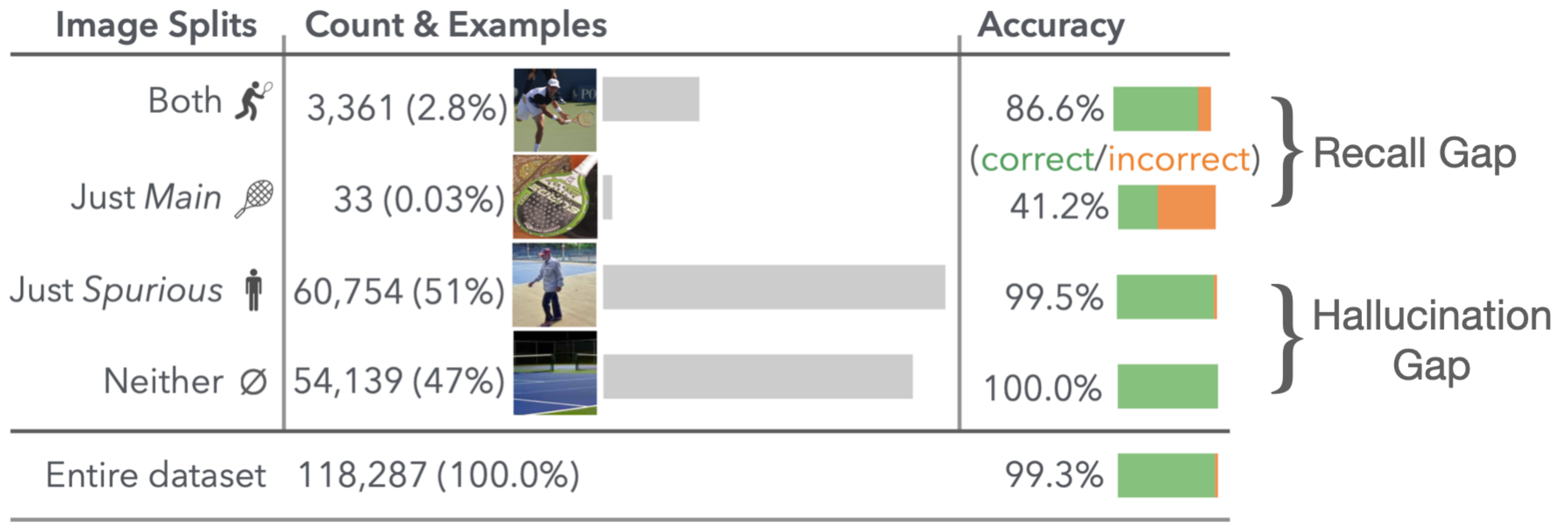

Image Splits	Count & Examples	Accuracy
Both	3,361 (2.8%)	86.6%
Just Main	33 (0.03%)	41.2%
Just Spurious	60,754 (51%)	99.5%
Neither ∅	54,139 (47%)	100.0%
Entire dataset	118,287 (100.0%)	99.3%

Figure 1: The tennis racket example, where Main = "Tennis Racket" and Spurious = "Person" on COCO.

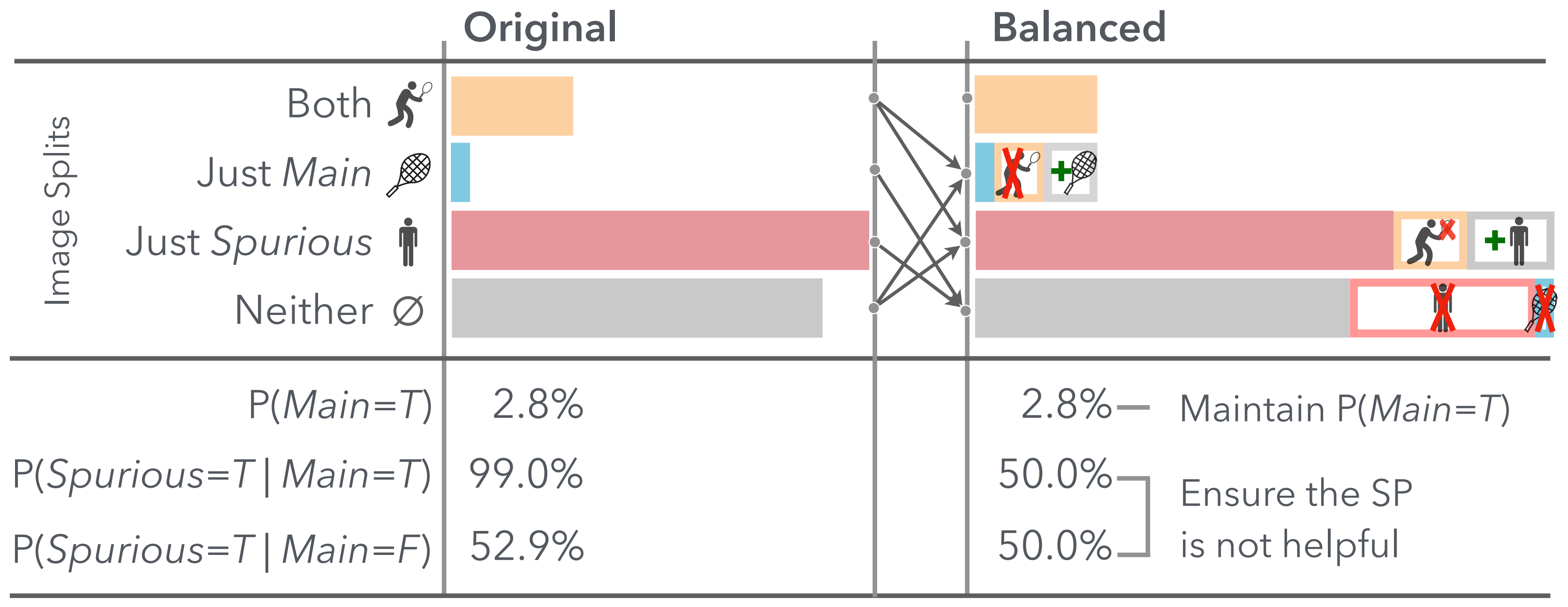

Image Splits	Original	Balanced
Both		
Just Main		
Just Spurious		
Neither ∅		
P(*Main=T*)	2.8%	2.8% — Maintain P(*Main=T*)
P(*Spurious=T* | *Main=T*)	99.0%	50.0%
P(*Spurious=T* | *Main=F*)	52.9%	50.0%

Figure 2: A comparison of the original and balanced distributions for the tennis racket example.

Controlled Experiments

In this experiment, we sub-sample the COCO dataset in order to artificially induce SPs of various strengths.

Compared to a range of existing methods (RRR [2], CDEP [1], GS [5], QCEC [3], FS [4]), SPIRE is better able to improve Balanced Accuracy and to shrink the Gap Metrics (Fig. 3).

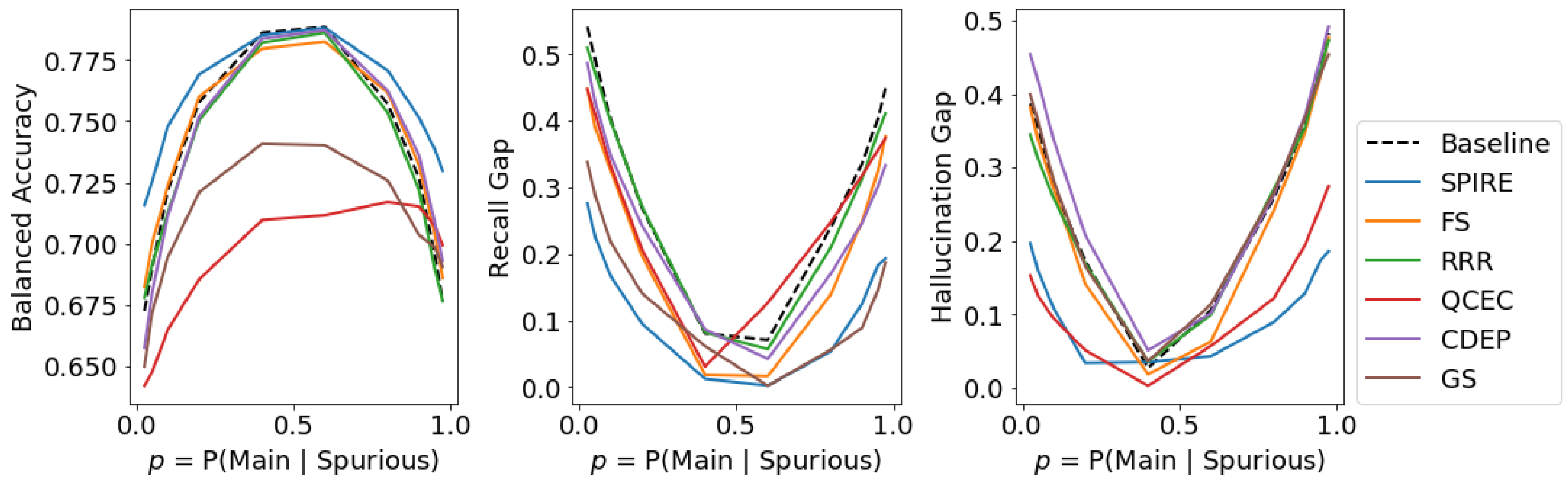

Figure 3: A summary of the results for the Controlled Experiments.

Full Experiment

When used on a model trained normally on the entire COCO dataset (*i.e.*, the "baseline" model), SPIRE identifies a wide range of SPs (Table 1) and is more effective than the best method from the Controlled experiments for at mitigating those SPs (Table 2).

Generalization Experiments

We also ran three experiments exploring how SPIRE generalizes along two different dimension:

- Tasks other than object detection.
- Settings without annotations to use to construct counterfactual images.

Table 1: A few examples of the SPs identified by SPIRE.

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 2: A summary of the mitigation results (averaged across the SPs identified by SPIRE and eight trials).

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

References

[1] L. Rieger, C. Singh, W. Murdoch, and B. Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International Conference on Machine Learning*, pages 8116–8126. PMLR, 2020.

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

[3] R. Shetty, B. Schiele, and M. 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*, pages 8218–8226, 2019.

[4] K. K. Singh, D. Mahajan, K. Grauman, Y. J. Lee, M. Feiszli, and D. 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*, pages 11070–11078, 2020.

[5] D. Teney, E. Abbasnedjad, and A. v. d. Hengel. Learning what makes a difference from counterfactual examples and gradient supervision. *arXiv preprint arXiv:2004.09034*, 2020.