Detecting Spurious Correlations via Robust Visual Concepts in Real and AI-Generated Image Classification

Preetam Prabhu Srikar Dammu
University of Washington
Seattle, WA, USA
preetams@uw.edu

Chirag Shah
University of Washington
Seattle, WA, USA
chirags@uw.edu

Abstract

Often machine learning models tend to automatically learn associations present in the training data without questioning their validity or appropriateness. This undesirable property is the root cause of the manifestation of *spurious correlations*, which render models unreliable and prone to failure in the presence of distribution shifts. Research shows that most methods attempting to remedy spurious correlations are only effective for a model's known spurious associations. Current spurious correlation detection algorithms either rely on extensive human annotations or are too restrictive in their formulation. Moreover, they rely on strict definitions of visual artifacts that may not apply to data produced by generative models, as they are known to hallucinate contents that do not conform to standard specifications. In this work, we introduce a general-purpose method that efficiently detects potential spurious correlations, and requires significantly less human interference in comparison to the prior art. Additionally, the proposed method provides intuitive explanations while eliminating the need for pixel-level annotations. We demonstrate the proposed method's tolerance to the peculiarity of AI-generated images, which is a considerably challenging task, one where most of the existing methods fall short. Consequently, our method is also suitable for detecting spurious correlations that may propagate to downstream applications originating from generative models.

1 Introduction

Spurious correlations in machine learning (ML) models arise from uncontrolled confounding biases present in data collection [5]. They render ML models vulnerable to distribution shifts and lead to poor generalization [43]. Several fundamental flaws of ML models can be traced back to spurious correlations, such as bias and poor robustness. Therefore, to make these models reliable in deployment, a necessary first step is to identify spurious correlations accurately.

With the recent surge in the usage of generative models such as DALL-E 2 [31], CLIP [30], and others, AI-generated data is increasing its footprint on the internet. Eventually, we can expect training data for ML models to significantly comprise such content as they are already being used for performing data augmentation [11, 4, 45]. Hence, it is paramount to devise a method that supports detecting spurious correlations in downstream models that were trained on AI-generated data. Problematic content generated by a single foundation model could affect numerous applications, leading to a considerable propagation of undesirable traits.

Vision-language models were shown to generate images with bizarre visuals, especially when presented with prompts that entail compositional reasoning and complex spatial relationships

XAI in Action: Past, Present, and Future Applications @ NeurIPS 2023.

between objects [44, 32]. When dealing with such images, the existing spurious correlation detection algorithms may prove to be ineffective. This outcome is expected due to the reliance of these methods on segmentation [27], or strict definitions of visual artifacts [38].

Notably, a method that can automatically distinguish between causal and spurious patterns without any human intervention is the most desirable solution. However, this would require a causal machine learning model that is trustworthy and without flaws – and such models do not yet exist for image classification. All of the existing methods for detecting spurious patterns rely on some form of human knowledge, either directly or through algorithmic encoding, and this dependence may not be solved without major breakthroughs in causal machine learning. Until then, the primary goal is to reduce the manual effort involved, along with a few more desirable properties that we address through the proposed method.

This paper addresses many of the existing shortcomings. Specifically, the contributions of the proposed method are:

- Automatically detects potential spurious correlations.
- Provides intuitive visual explanations which help in confirming valid spurious correlations.
- Automatically flags similar instances of an identified spurious correlation.
- Compatibility with AI-generated images.
- Eliminates the need for object segmentation and pixel-wise annotations.

2 Related Work

In recent years, a few papers have been published in the literature that study spurious correlations in text [49, 47, 41] and a few on vision models [38, 27, 2, 23, 50]. However, none of these methods consider the complications that arise when working with generative models. Furthermore, the existing spurious correlation detection methods in vision either rely on extensive human interference or pixel-level annotations.

While explainability techniques could be used to spot spurious correlations in theory, they are not sufficiently suitable for this task as they typically generate local explanations for each sample [8, 36, 35, 34, 37, 13, 15, 46]. They do not aggregate information by detecting repeated occurrences of spurious patterns, and making manual observations by probing each explanation is infeasible. Moreover, studies have shown that these methods may not always be reliable [2, 1]. To address these challenges, a few dedicated techniques for identifying spurious correlations have been proposed which build on top of explainability methods [38, 16, 27].

In [27], the authors propose a method that identifies spurious patterns on the basis of sensitivity to changes from counterfactual versions of the original instance, and have a human label the identified patterns as spurious or valid. In [38], the authors show that obtaining human annotations for strategically procured representative images can generalize well for unseen images. Yet, this approach still required a significant amount of manual intervention that warranted crowd-sourcing. In [16], the authors proposed using explanation-based learning to identify spurious correlations, however, this study was conducted on toy datasets and it is unclear how well it would generalize.

3 Generic Framework for Spurious Correlation Detection

3.1 Core and Spurious Attributes

In vision tasks, the definition of spurious attributes is a bit vague, but previous methods have established a usable interpretation [38, 27]. Similar to the definition presented in [38], we consider

- visual features always expected to be part of the object definition as *core attributes*
- features that are likely to co-occur, but not a part of the object, as *spurious attributes*.

However, we extend the definition of *core attributes* to include features that are perceivably a part of the object but may not always be expected to be present. This modification enables us to account for the hallucinations or imperfections of generative models. For instance, if a generative

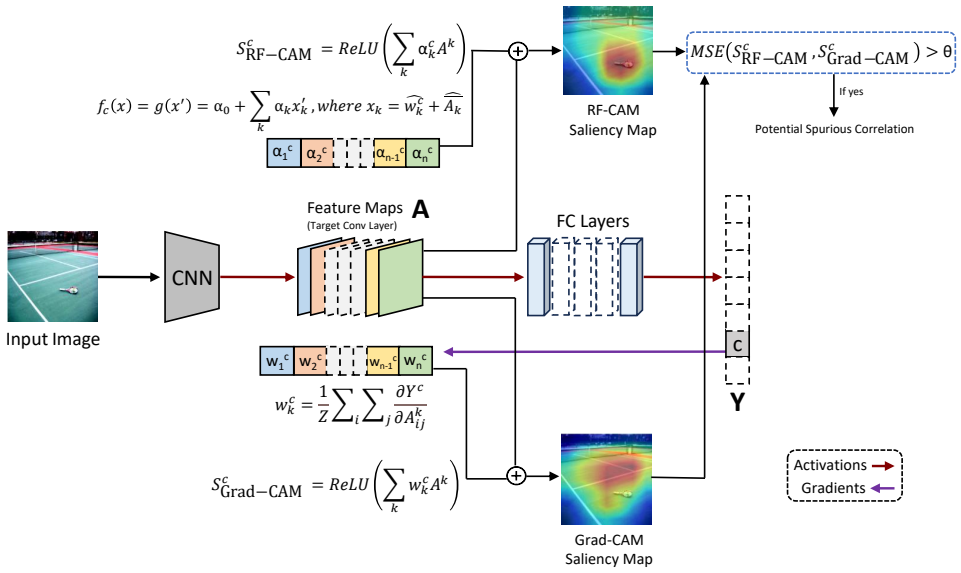

Figure 1: Overview of the proposed method for detecting *spurious correlations* via robust features.

model depicts a gymnast as a person with three legs, all of the limbs should be considered *core attributes* even though this deviates from physical reality – as it reflects the generative model's perception of reality. As a result, the flexibility of this formulation allows us to understand the generative model's worldview and examine its spurious associations.

3.2 Approach Overview

We propose building local surrogate models (LSMs), one for each image class, trained to learn robust features specific to its respective class. As these robust features are representative of core attributes, disagreement between the image classifier and the LSM of the predicted class can reveal spurious associations. In order to accommodate deviations from conventional definitions, we utilize prototypical explanations that convey spurious behavior. This makes the framework sufficiently flexible to support any data distribution, even when the contents of the images are not photorealistic, as we only need to point out the spurious concepts through heatmaps without requiring to define them. We operationalize this approach by leveraging feature activations in the final convolutional layer, gradients, and the explanations from LSMs designed to learn robust features and the behavior of the image classifier for each class.

3.3 Methodology

Here, we present the details of our method in logical parts, and then tie it all together in 3.3.6.

3.3.1 Spurious Correlation Identification

As per the definitions in 3.1, each input image $x \epsilon X$ has core attributes $c \epsilon C$ and spurious attributes $c' \epsilon\ U \setminus C$, where U is the universal set of all features. Consider an image classifier $f(x) : X \rightarrow Y$, where each $x \epsilon X$ has to be assigned a label $y \epsilon Y$. Our goal is to identify if any c' attributes were activated when x is correctly classified, and flag such instances for potential spurious correlations.

3.3.2 Local Surrogate Models (LSMs)

Without groundtruths, either through pixel-level annotations or human review, detecting if a visual attribute is spurious is non-trivial. Therefore, we build an LSM for each class that bases its decisions on robust features, as they tend to represent core attributes. Global models that leverage robust features were shown to significantly minimize reliance on spurious attributes

[50, 39], however, they demonstrate lower clean accuracy. In contrast, our approach leaves the original model unaltered and builds LSMs that help us detect and visualize spurious correlations.

For an image classifier $f(x) : X \rightarrow Y$, where each $x \epsilon X$ is assigned a label $y \epsilon Y$, we build an LSM for each class c, $f_c(\phi) : \Phi \rightarrow Y'$, where each $\phi \epsilon \Phi$ is representative of an image x in a transformed latent space Φ, and $y\prime \epsilon Y'; y\prime \in \{0, 1\}$ represents if the original model has misclassified. Each class c will have an LSM f_c, which is trained and evaluated only on instances belonging to c.

Each image x is appropriately represented by its corresponding ϕ, as it is the element-wise sum of the global-average-pooled gradients of original model's prediction w.r.t. activation maps A^k

$$w_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k} \tag{1}$$

(Z = number of pixels in A^k with dimensions $i * j$)

and the mean of activation map $\bar{A}^k$. Both w_k^c and $\bar{A}^k$ are normalized to ensure equal weightage.

$$\phi_k^c = \widehat{w_k^c} + \widehat{\bar{A}^k} \tag{2}$$

Intuitively, $\widehat{\bar{A}^k}$ represents any visual attributes present in an image, and $\widehat{w_k^c}$ represents the visual attributes contributing to the decision. We want our LSM $f_c(\phi)$ to learn patterns from both of these entities in relation to misclassification, and therefore we perform an element-wise sum. While we use XGBoost [9] as the model choice for LSMs, any ML method could be used.

3.3.3 Computing SHAP values

When an input image is classified as class c, we employ the LSM for that class, $f_c(\phi)$, for generating SHAP [22] values corresponding to visual attributes in the image. SHAP assigns each feature an importance value for a particular prediction

$$f_c(\phi) = g(\phi') = \alpha_0 + \sum_k \alpha_k \phi_k' \tag{3}$$

The explanation model $g(\phi')$ approximates the surrogate model $f_c(\phi)$ and computes contributions α_k by every feature, and these values are used to generate the robust features saliency maps.

3.3.4 Robust Features Saliency Map (RF-CAM)

To produce robust features class activation maps (RF-CAM), a weighted combination is performed on SHAP values computed in equation 3 and class activation maps, followed by a ReLU [3]:

$$S^c_{\text{RF-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \tag{4}$$

RF-CAM saliency maps highlight robust features characteristic to the predicted class, which tend to be core attributes. In contrast, explainability methods such as Grad-CAM [35], highlight visual features that triggered the prediction by leveraging w_k^c computed in equation 1, and observing the difference between these saliency maps reveals information regarding potential spurious attributes.

$$S^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k w_k^c A^k\right) \tag{5}$$

ReLU [3] is used on the weighted combination to preserve positively associated robust features toward a particular class. These saliency maps will be the size of the featuremaps in the final convolution layer, and they are upscaled to the input image size, and overlayed as the final step.

3.3.5 Spurious Correlations from Saliency Maps

Consider that the RF-CAM accurately highlights core attributes of the predicted class in an image, and an explainability map, such as Grad-CAM, accurately highlights the attributes that triggered the prediction. In the presence of a spurious correlation, a correctly classified image will have these two saliency maps that differ beyond a threshold θ. In an ideal scenario when there is no spurious correlation, these saliency maps should look very similar. We can measure the dissimilarity using Mean Squared Error: $MSE(S^c_{\text{RF-CAM}}, S^c_{\text{Grad-CAM}}) > \theta$ The appropriate threshold θ limit will vary depending on several factors such as model, dataset, and surrogates.

3.3.6 Flow of Operations

The overview of the proposed method is presented in Figure. 1. As a first step, using the original image classifier on each sample, we compute the predicted label, feature maps A^k, and the gradients w^c_k for the predicted class y_c. All of these entities are required for training LSMs.

Using the labels predicted by the image classifier and comparing them with true labels, a new misclassification label is generated for all samples in both training and test sets. Next, for each class c, a LSM f_c is trained using the misclassification labels on the train split and evaluated using the test split. The input for these surrogate models is the combination of unit gradients and mean activations, given by θ (discussed in detail in 3.3.2). SHAP values generated on these local surrogates represent importance values for robust features (refer 3.3.3), and these values are used to produce robust feature saliency maps (RF-CAM) (refer 3.3.4).

Subsequently, comparing RF-CAM and Grad-CAM maps using mean square error, potential spurious correlations are identified. These findings on potential spurious correlations are then presented to the model developer, as human judgment and domain knowledge are vital for the confirmation of spurious behavior. For a confirmed spurious correlation, the most activating neural feature can be utilized to identify other instances of this spurious correlation. In [38], the authors show that neural feature annotations generalize very well to many images.

4 Datasets

We evaluate our method on three different datasets. The first one is the widely-used ImageNet [12], which was also used in other studies exploring spurious correlations [38, 50, 39]. As our method supports content produced by generative models – an improvement over extant work, we would require synthetic datasets generated by an AI model to demonstrate this property. Therefore, we create two datasets that are AI-generated, using Stable Diffusion 2.1 [33].

GenAI Images Dataset 1: 9,750 images are generated for 75 classes of *MS-COCO* [21] using dataset captions. All classes are balanced, with 100 training images and 30 test images each.

GenAI Images Dataset 2: 52,000 images are generated for 40 classes of *ImageNet People SubTree* [52] using Vit-GPT2 [14] captions. Each class has 1,000 training and 300 testing images.

Further details of GenAI images datasets are discussed in Appendix (refer A.1).

5 Experiments and Results

Table 1: Accuracies of Image Classifiers (IC) and averaged acc. of Local Surrogate Models (LSM).

Classifier	Dataset	Classes	IC Acc	LSM Avg. Acc
ResNet50	ImageNet	1,000	74.54	99.93
ResNet50	GenAI Images 1	75	88.56	96.58
ResNet50	GenAI Images 2	40	83.19	95.28

We present the findings of our method on three datasets with significantly different data distributions. While ImageNet is representative of real-world images, the GenAI datasets are suitable for testing the compatibility of our method with AI-generated images. *GenAI Images 1*

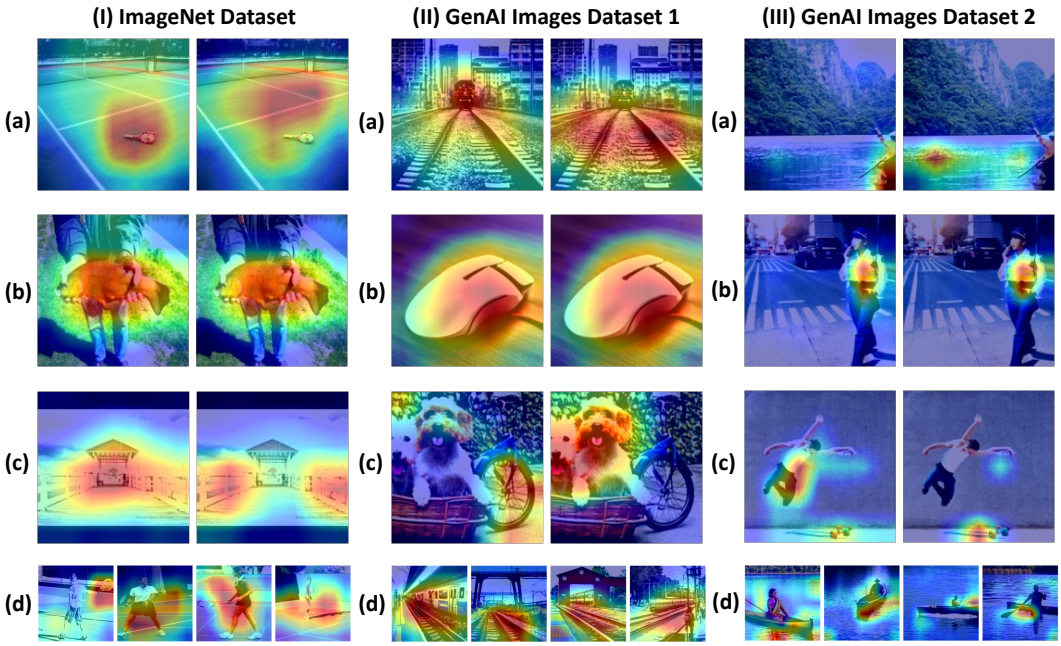

Figure 2: RF-CAM and Grad-CAM (left and right respectively, in each pair for [a-b]). Correctly classified instances based on (a) Spurious Attributes (b) Core Attributes; (c) Misclassified instances (d) Grad-CAM of instances activating spurious feature (I) 1890 for class *racquet* (II) 936 for class *train* (III) 1017 for class *boatperson*.

Figure 3: RF-CAM highlights deformities as representative and robust attributes for class *Gymnast*

consists of relatively more photorealistic images, while images in *GenAI Images 2* are characterized by peculiar imperfections since Stable Diffusion does not perform very well for human images.

We choose pretrained ResNet50 [17] as the image classifier, but any CNN-based vision model is supported. Default weights were used on ImageNet, and only the last FC layer was retrained to match the number of classes in the other two datasets. We tabulate the performance of the image classifier and the average accuracy of the LSMs in Table. 1. The high LSMs average accuracies indicate that the robust features learned by them are reliable.

In summary, in all three datasets, the proposed algorithm successfully identified potential spurious correlations, generated intuitive explanations to help aid decisions, and retrieved similar instances for identified spurious correlations. Roughly, 35% to 40% potential spurious correlations were detected in each test set. Notably, many of these instances are repetitions belonging to a few spurious associations. When the model developer confirms a detected potential spurious correlation, several similar spurious instances can be remedied automatically. Therefore, the manual labor involved is significantly reduced, and notably lower than other existing methods. Similar to [38], Human Intelligence Tasks (HITs) would be required to compute the conversion ratio of potential spurious correlations to confirmed ones, and is a viable task for future work.

In all experiments, the dissimilarity between RF-CAM and Grad-CAM heatmaps (section 3.3.5) was calculated on pixels with intensity higher than 0.78, which roughly correspond to the red regions in the heatmaps. This prevents noise from significantly affecting the dissimilarity score.

Additionally, all pixels with more than 0.78 intensity are rounded to 1 while computing the MSE, in order to increase tolerance to minor variations. The MSE threshold (θ) is set to 15.

In the first row of Figure. 2, we observe instances where the classifier has based its predictions on spurious attributes. The first heatmap in each pair is RF-CAM, which highlights core features and the second heatmap highlights the features that were actually used by the classifier, and differences between these uncover spurious attributes. For instance, in (Ia), we notice that the classifier relies on the tennis court background rather than the racquet itself. By using the most activated neural feature, we obtain other instances of this spurious association between the class *racquet* and tennis court, and the top four retrieved images are shown in (Id). Similarly, in (IIa), we observe that the classifier relies on railway tracks to identify trains, and other instances of this spurious association are shown in (IId), some of which have no train visible at all. In (IIIa), we notice that only water background is used for identifying boatperson and other instances of this can be seen in (IIId). In the second row, we observe that when there are no spurious associations present, RF-CAM and Grad-CAM heatmaps look almost identical, as expected.

In images with multiple objects, there can be several valid labels. Most classification tasks, however, only designate one of these as the 'groundtruth' label. When an image classifier identifies a valid label that is different from the groundtruth, RF-CAM can be utilized to emphasize the key features of the groundtruth class. This is particularly helpful for debugging models, as it shows which areas the model should have focused on for a correct prediction. For example, in (Ic), RF-CAM highlights the area the classifier should have considered to classify it as *'pier'*, while Grad-CAM illustrates the area that led to its incorrect classification as *'worm fence'*.

In Figure. 3, we notice that multiple RF-CAM heatmaps show malformed limbs as robust features representative of a gymnast. This demonstrates that the proposed method can be useful for exploring and uncovering spurious associations projected by the generative model as well.

Additional details on results for each dataset are discussed in Appendix A.3.

6 Limitations

There is no one solution that fits all requirements for spurious correlation detection. The proposed method, for instance, will perform better when there are sufficient misclassifications in the dataset that were caused by spurious associations. Noticeably, this is commonly observed in most datasets with unconstrained real-world settings, as they would have long tails of data distribution and memorization becomes unlikely. Conversely, if a spurious correlation is so prevalent in the dataset and without counter-examples, the LSMs might pick it up as a robust feature. In such cases, where the dataset is not sufficiently diverse, evaluating on a different dataset built for the same purpose would be helpful in overcoming the shortcomings.

In scenarios where an image contains multiple objects, the proposed method may flag for potential spurious correlations when the predicted label deviates from the groundtruth label, as illustrated in Figure 2 (I-IIIc). However, this issue arises from the restrictive formulation of classification tasks where only a single groundtruth label is considered even when multiple objects are present in the image.

7 Conclusion

In this paper, we proposed a method that effectively detects spurious correlations while providing several advantages over the existing methods, such as: (1) not requiring pixel-level annotations; (2) unconstrained by restrictive object definitions; (3) intuitive visual and comparative explanations; (4) automatic flagging of similar spurious correlation instances; and (5) compatibility with AI-generated images. Furthermore, we discussed the importance of this line of work in the currently evolving landscape of visual data, driven by a surge in the adoption of generative models.

Acknowledgements

This work was supported in part by the cloud computing credits made available by the University of Washington eScience Institute in partnership with Microsoft Azure.

References

[1] Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. Debugging tests for model explanations. *arXiv preprint arXiv:2011.05429*, 2020.

[2] Julius Adebayo, Michael Muelly, Harold Abelson, and Been Kim. Post hoc explanations may be ineffective for detecting unknown spurious correlation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xNOVfCCvDpM.

[3] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.

[4] Mohamed Akrout, Bálint Gyepesi, Péter Holló, Adrienn Poór, Blága Kincső, Stephen Solis, Katrina Cirone, Jeremy Kawahara, Dekker Slade, Latif Abid, et al. Diffusion-based data augmentation for skin disease classification: Impact across original medical datasets to fully synthetic images. *arXiv preprint arXiv:2301.04802*, 2023.

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

[6] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[7] Chris Berg. The case for generative ai in scholarly practice. *Available at SSRN 4407587*, 2023.

[8] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[9] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[10] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. Talebrush: sketching stories with generative pretrained language models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.

[11] Preetam Prabhu Srikar Dammu, Yunhe Feng, and Chirag Shah. Addressing weak decision boundaries in image classification by leveraging web search and generative models. In Edith Elkind, editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 5941–5949. International Joint Conferences on Artificial Intelligence Organization, 8 2023. doi: 10.24963/ijcai.2023/659. URL https://doi.org/10.24963/ijcai.2023/659. AI for Good.

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

[13] 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, 2018.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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

[16] Misgina Tsighe Hagos, Kathleen M Curran, and Brian Mac Namee. Identifying spurious correlations and correcting them with an explanation-based learning. *arXiv preprint arXiv:2211.08285*, 2022.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] 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)*, pages 771–787, 2018.

[19] Mengxiao Hu and Jinlong Li. Exploring bias in gan-based data augmentation for small samples. *arXiv preprint arXiv:1905.08495*, 2019.

[20] Nikita Jaipuria, Xianling Zhang, Rohan Bhasin, Mayar Arafa, Punarjay Chakravarty, Shubham Shrivastava, Sagar Manglani, and Vidya N Murali. Deflating dataset bias using synthetic data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 772–773, 2020.

[21] 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*, pages 740–755. Springer, 2014.

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

[23] Yifei Ming, Hang Yin, and Yixuan Li. On the impact of spurious correlation for out-of-distribution detection. In *The AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[24] Michael Muller, Lydia B Chilton, Anna Kantosalo, Charles Patrick Martin, and Greg Walsh. Genaichi: Generative ai and hci. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–7, 2022.

[25] Jonas Oppenlaender. A taxonomy of prompt modifiers for text-to-image generation. *arXiv preprint arXiv:2204.13988*, 2022.

[26] P Phillips, Amanda Hahn, Peter Fontana, David Broniatowski, and Mark Przybocki. Four principles of explainable artificial intelligence (draft), 2020-08-18 2020.

[27] Gregory Plumb, Marco Tulio Ribeiro, and Ameet Talwalkar. Finding and fixing spurious patterns with explanations. *Transactions on Machine Learning Research*, 2022. URL `https://openreview.net/forum?id=whJPugmP5I`.

[28] Cale Plut and Philippe Pasquier. Generative music in video games: State of the art, challenges, and prospects. *Entertainment Computing*, 33:100337, 2020.

[29] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[31] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[32] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[34] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*, 2017.

[35] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

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

[37] Chandan Singh, W James Murdoch, and Bin Yu. Hierarchical interpretations for neural network predictions. *arXiv preprint arXiv:1806.05337*, 2018.

[38] Sahil Singla and Soheil Feizi. Salient imagenet: How to discover spurious features in deep learning? In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=XVPqLyNxSyh`.

[39] Sahil Singla, Besmira Nushi, Shital Shah, Ece Kamar, and Eric Horvitz. Understanding failures of deep networks via robust feature extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12853–12862, 2021.

[40] Ramya Srinivasan and Kanji Uchino. Biases in generative art: A causal look from the lens of art history. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 41–51, 2021.

[41] Megha Srivastava, Tatsunori Hashimoto, and Percy Liang. Robustness to spurious correlations via human annotations. In *International Conference on Machine Learning*, pages 9109–9119. PMLR, 2020.

[42] Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. The biased artist: Exploiting cultural biases via homoglyphs in text-guided image generation models. *arXiv preprint arXiv:2209.08891*, 2022.

[43] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

[44] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022.

[45] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*, 2023.

[46] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020.

[47] Tianlu Wang, Diyi Yang, and Xuezhi Wang. Identifying and mitigating spurious correlations for improving robustness in nlp models. *arXiv preprint arXiv:2110.07736*, 2021.

[48] Yunlong Wang, Shuyuan Shen, and Brian Y Lim. Reprompt: Automatic prompt editing to refine ai-generative art towards precise expressions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–29, 2023.

[49] Zhao Wang and Aron Culotta. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14024–14031, 2021.

[50] Eric Wong, Shibani Santurkar, and Aleksander Madry. Leveraging sparse linear layers for debuggable deep networks. In *International Conference on Machine Learning*, pages 11205–11216. PMLR, 2021.

[51] Zhanghao Wu, Shuai Wang, Yanmin Qian, and Kai Yu. Data augmentation using variational autoencoder for embedding based speaker verification. In *INTERSPEECH*, pages 1163–1167, 2019.

[52] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 547–558, 2020.

[53] Seyma Yucer, Samet Akçay, Noura Al-Moubayed, and Toby P Breckon. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–19, 2020.

[54] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

A Appendix

A.1 Further Dataset Details

A.1.1 GenAI Images Dataset 1

To demonstrate that the proposed method works well with AI-generated images, we create a new dataset using Stable Diffusion 2.1. As this is the first work of its kind, a standard dataset of text prompts for generating images does not yet exist. To circumvent this issue, we leverage an image captioning dataset – MS-COCO [21], and generate the dataset by using the captions as the input prompts to Stable Diffusion. We noticed that, in many cases, the generated image and the original image which share the same caption/prompt are not very representative of each other. Also, each image in MS-COCO has 5 captions, all of which could yield very different results.

There are 80 object categories in the MS-COCO dataset, and each image has pixel-wise annotations for these objects, if present. As our goal is to generate a classification dataset, we use object categories as class labels for images that contain these objects. Note that an image can contain multiple objects, and in those cases, we randomly pick one of the object labels as the class label. This results in a labeling scheme very similar to ImageNet, as it also contains images with multiple objects and a single class label. This property is very representative of the real world in unconstrained settings.

It is imperative to ensure that the object corresponding to the class label is present in the generated image. To this effect, we filter and select the captions with the class label explicitly mentioned in them, forcing the generative model to produce visual features corresponding to the class label. Applying this constraint resulted in the removal of 4 categories, which are *hair drier*, *sport ball*, *handbag*, and *toaster*, as they have very few captions explicitly consisting of them as words. Additionally, we also removed the *person* category, as most images contain people in the background, and this adds a significant amount of perplexity to the task. For the remaining 75 categories, we generated 130 images each – 100 to be included in the training set and 30 in the test set. This results in a total of 9750 images, with 7500 in the train split and 2250 in the test split. Most of the generated images in this dataset were photorealistic, indicating that the image captions are suitable inputs to the generative model.

A.1.2 GenAI Images Dataset 2

Stable Diffusion 2.1 performs impressively for non-human images and achieves remarkable photorealism, but it seems to fall short when humans are involved as the main subject of the image. For a thorough evaluation, we require a synthetic dataset that is characterized by peculiar imperfections of AI-generated images. Therefore, we create a synthetic version of ImageNet people subtree.

ImageNet People SubTree [52] is a subset of ImageNet containing 2,832 people categories. Although, most of these categories are unusable as only 139 of them are considered safe and imageable [52]. From these 139 classes, we use ones that are either a profession or an occupation. Additionally, categories that share a lowest common hypernym and have a broader yet specific definition of an occupation were merged together. For instance, *captain*, *chief_of_staff*, *general*, *major*, *Navy_SEAL*, *military_personnel* were all mapped to *military_officer*, as some of these classes are an abstraction of others. Finally, we get 40 classes that are used in this study.

The synthetic version of ImageNet people subtree is created by first generating captions for each image using Vit-GPT2 [14, 29]. However, most of these captions are not granular enough to describe details about the person and simply use pronouns. In order to overcome this issue, we replace all pronouns with the person class label corresponding to the image, and this step results in rich text prompts that yield images representative of the person category. Finally, these captions are fed as text prompts to generate images using Stable Diffusion. For each of the 40 classes, we generate 1,000 training images and 300 test images, leading to a training split of 40,000 images and a test split of 12,000 images.

A.2 Difficulties in handling AI-generated images

Generative models are known to produce images that are incomplete, confusing, and sometimes incomprehensible. This is especially noticeable when the contents of the image are complex, require compositional reasoning, or relational constraints.

Figure 4: Flaws in AI-generated images. The first image in each pair is synthetic, and the second is a real image sharing the same caption.

In Figure. 4, the first pair of images share the caption: *'A tabby cat laying on a cat scratcher in front of a bicycle wheel'*, and Stable Diffusion generated a cat coming out of a bizarre-looking bicycle wheel. Second pair share the caption *'A cat resting inside a bowl next to different household items'*, and the generated image has a cat's head placed in a bowl, and none of the other objects are recognizable. These synthetic images are from GenAI dataset 1 [A.1.1].

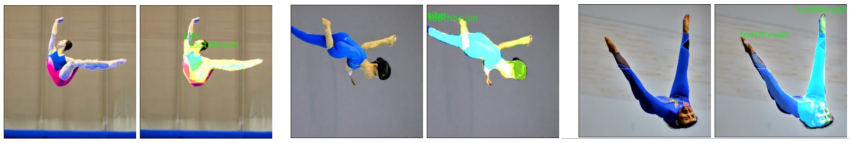

Figure 5: Segmentation Failure. Prompt: *'a gymnast in a blue shirt is on a blue and white board'*. Instead of *person*, detected as *kite* & *toothbrush*.

Such peculiar flaws in AI-generated images pose a challenge to existing spurious correlation detection algorithms, as they cause segmentation methods to fail. For instance, consider the images in Figure. 5, in all three images a gymnast was not detected as a person and instead as a *kite* or *toothbrush*. These synthetic images are from GenAI dataset 2 [A.1.2].

A.3 Further Result Details

ImageNet: The ImageNet [12] dataset is highly representative of images taken in the wild, and consists of strong signals that co-occur together in the real world, making it ideal for testing spurious correlation detection algorithms. Prior works addressing spurious correlations in image classification have conducted experiments on ImageNet [38, 50, 39]. On a test set of size 50,000, our method detected 18,062 (36.12%) instances of potential spurious correlations.

GenAI Images Dataset 1: In order to demonstrate the compatibility of our method with AI-generated images, we evaluate it on this dataset generated from MS-COCO image captions using Stable Diffusion. Details of this dataset are discussed in Section A.1.1. Stable Diffusion was able to generate photorealistic images based on COCO captions with few exceptions. Our method handled this dataset well, with a performance similar to ImageNet. Out of 2,250 images in the test set, 923 (41.02%) potential spurious correlations were detected.

As most of the images in this dataset are photorealistic, it is unsurprising that the method is able to perform equally well. However, it should be noted that traditional segmentation algorithms might still fail at higher rates for AI-generated images, even if they appear to be mostly photorealistic, as they are still out-of-distribution samples to the segmentation model.

GenAI Images Dataset 2: The tolerance toward the peculiarity of AI-generated images of the proposed algorithm is better demonstrated on this dataset, as Stable Diffusion does not perform very well for human images. Unlike GenAI images dataset 1 [A.1.1], this dataset is not photorealistic and contains evident imperfections. Despite these flaws, the proposed algorithm has performed similarly to other datasets, demonstrating its tolerance and compatibility with AI-generated images. Out of 12,000 test images, 4,103 (34.19%) were detected to have potential spurious correlations.

We note that defining core attributes for the classes of ImageNet person subtree is a complex task and without a clear solution, as many of these classes have an imageable score of less than 1 [52]. However, the purpose of this experiment is to test the ability of the proposed method in handling the imperfections of AI-generated images, and these flaws appear to be more prevalent in human-related content, making this dataset a suitable option.

A.4 Observations and Opportunities

Arguably, one of the most crucial problems of this decade that has to be addressed by the machine learning community, in light of the recent advances in generative models, is to upgrade current information systems to effectively process AI-generated content. Generative AI is now used in the creation of art [40, 48, 25], music [28, 7], writing [24, 10], and many more applications.

Notably, almost every modern information access system, either directly or indirectly, relies on machine learning. These include infotainment services like music recommendations, and applications with more serious implications such as job search and medical advice. It is paramount to ensure that these ML services are robust and reliable. A major step toward this goal is verifying that they are free from spurious correlations. Apart from compromising performance in deployment, researchers have shown that spurious correlations have societal consequences by causing the amplification and propagation of bias [42, 18, 54, 6].

Information retrieval systems, especially the ones that leverage neural ranking and neural recommendation, will be significantly affected by this new incursion of AI-generated content that is becoming ubiquitous on the web. The National Institute of Standards and Technology (NIST) has stressed the importance of identifying limitations and scenarios where machine learning models could fail [26]. This goal cannot be realized unless advances are made in detecting spurious correlations.

Once spurious correlations are accurately identified, there are several ways to remedy them. Model retraining is the most common approach, which is achieved by retraining the model on an augmented dataset – either by removing harmful samples or by introducing neutralizing samples. Procuring new samples that act as counterbalances can be achieved through different methods, and techniques such as counterfactual data generation, GANs, VAEs, simulation engines, and more recently generative models have been explored [11, 27, 20, 53, 19, 51].

Another future direction, one with potentially tremendous impact, is detecting and quantifying spurious correlations in foundation models. These models are adapted for various tasks, and any bias originating from them would have far-reaching effects in downstream applications. Compared to discriminative models, uncovering spurious associations of generative models appears to be more challenging. One approach to solving this problem could be through a proxy model, where a downstream model trained on the content generated by a foundation model is examined to understand the spurious behaviors of the main model. Addressing the problem of spurious correlations at the source, i.e., generative models, could translate into a large impact downstream and eventually lead to cleaner data in information systems and the web.