Spuriosity Rankings for Free:
A Simple Framework for Last Layer Retraining Based on Object Detection

Mohammad Azizmalayeri [* 1] **Reza Abbasi** [* 1] **Amir Hosein Haji Mohammad rezaie** [* 1] **Reihaneh Zohrabi** [* 1]
Mahdi Amiri [* 1] **Mohammad Taghi Manzuri** [1] **Mohammad Hossein Rohban** [1]

Abstract

Deep neural networks have exhibited remarkable performance in various domains. However, the reliance of these models on spurious features has raised concerns about their reliability. A promising solution to this problem is last-layer retraining, which involves retraining the linear classifier head on a small subset of data without spurious cues. Nevertheless, selecting this subset requires human supervision, which reduces its scalability. Moreover, spurious cues may still exist in the selected subset. As a solution to this problem, we propose a novel ranking framework that leverages an open vocabulary object detection technique to identify images without spurious cues. More specifically, we use the object detector as a measure to score the presence of the target object in the images. Next, the images are sorted based on this score, and the last-layer of the model is retrained on a subset of the data with the highest scores. Our experiments on the ImageNet-1k dataset demonstrate the effectiveness of this ranking framework in sorting images based on spuriousness and using them for last-layer retraining.

1. Introduction

A prominent issue that has recently gained substantial attention is the dependency of deep models on shortcuts (Geirhos et al., 2020). An illustrative example of such shortcuts is spurious correlation which is statistically informative heuristics for predicting the labels of most examples in the training data. However, they are not actually relevant to the true labeling function, and do not necessarily have a causal re-

lationship with the class label (Kirichenko et al., 2023). A model that captures these spurious correlations may have high average accuracy on a given test set, but encounters significant accuracy drop on data subsets without these correlations (Arjovsky et al., 2019; Sagawa et al., 2019; Alcorn et al., 2019). Theoretical arguments suggest that a bias towards model simplicity contributes to the reliance on spurious features (Shah et al., 2020).

A wide range of solutions have been proposed to address the model robustness against spurious correlations. One approach involves optimizing for the worst case group accuracy, which necessitates the annotations indicating the respective group affiliation of each training example (Sagawa et al., 2019; Zhang et al., 2021; Liu et al., 2021; Hu et al., 2018). Additionally, alternative strategies involve leveraging causality (Aubin et al., 2021; Seo et al., 2022), and learning invariant latent spaces (Arjovsky et al., 2019; Robey et al., 2021). Recent researches have also highlighted the effectiveness of fine-tuning the last layer of neural networks on samples without spurious correlations to the label, to mitigate the model sensitivity to spurious cues (Kirichenko et al., 2023; Rosenfeld et al., 2022; Lee et al., 2022). However, they often rely on human supervision to determine the presence of spurious features, which may limit their applicability, particularly for large datasets like ImageNet (Deng et al., 2009). For example, in (Kirichenko et al., 2023), human supervision is necessary to re-weight the spurious features through retraining the last linear layer on a small dataset without spurious correlations between the background and foreground.

To overcome this limitation, (Singla & Feizi, 2022; Singla et al., 2022) have introduced an approach that utilizes the neurons of robust models as detectors for visual attributes, enabling identification of the spurious features at a large scale while minimizing the need for human supervision. Moreover, (Neuhaus et al., 2022) have proposed an automated pipeline for detecting spurious features using class-wise neural PCA components, which still requires human supervision. Building upon the concept of robust neural features, (Moayeri et al., 2022) introduced a novel ranking framework that assesses the prominence of spurious cues

---

[*]Equal contribution [1]Department of Computer Engineering, Sharif University of Technology, Tehran, Iran. Correspondence to: Mohammad Azizmalayeri <m.azizmalayeri@sharif.edu>, Reza Abbasi <reza.abbasi@sharif.edu>.
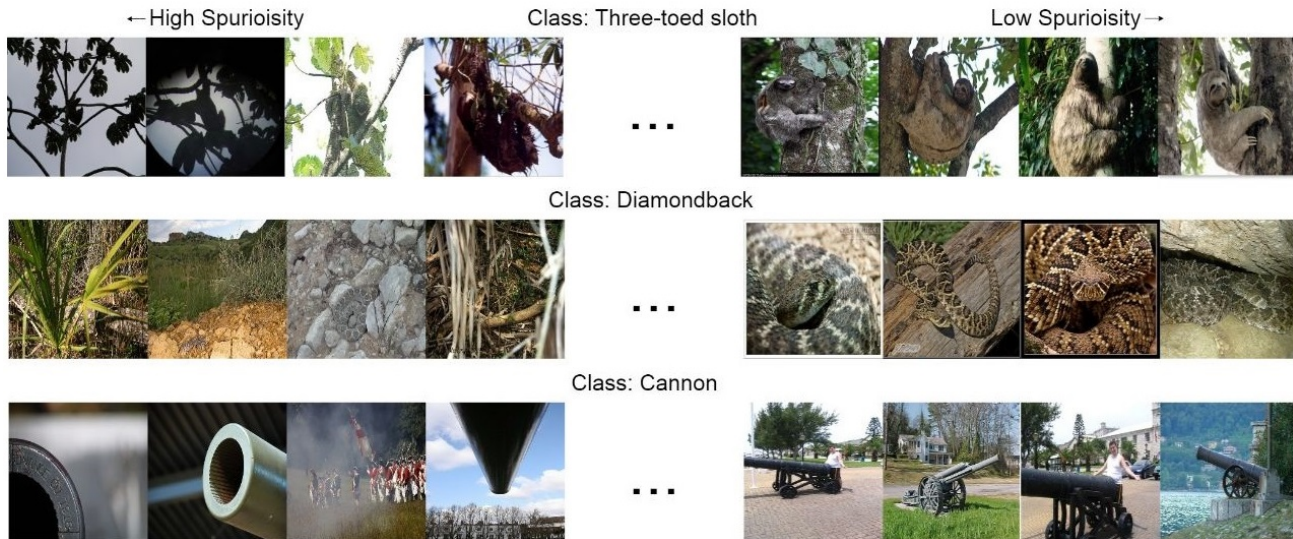
Figure 1. Spuriosity ranking reveals core object detectability and the presence of spurious cues, distinguishing between clear representations such as a recognizable three-toed sloth, Diamondback, and Cannon (low spurious cues - right) and more ambiguous instances featuring spurious cues like trees with a less prominent three-toed sloth, Diamondbacks in hard-to-distinguish locations, and Cannon barrel (high spurious cues - left).

within image classes. In order to reduce the impact of spurious features, their approach involves training linear heads on carefully selected subsets of data determined by the ranking.

Despite these notable attempts to reduce the reliance on human supervision, the complete elimination of such supervision remains challenging. The mentioned approaches still require human involvement in annotating feature-class dependencies, indicating that previous works may face limitations when employing a human-in-the-loop strategy. Therefore, it is important to acknowledge that the need for human involvement in these methods introduces scalability concerns. To overcome these challenges and further enhance the autonomy of the system, we propose a novel method for eliminating the need for human supervision in the detection and mitigation of the effect of spurious features.

In this work, we introduce a novel approach to identify the core and spurious features within classes, utilizing an open vocabulary object detection technique (Minderer et al., 2022). By examining the images with the highest and lowest spuriosity ranks, as illustrated in Figure 1, we are able to detect images with the high spurious cues and minority subgroups. In our framework, images that clearly have the object, such as a clear identifiable three-toed sloth, receive high scores and are ranked as low spurious. Conversely, images where the spurious cues are present, and the object could not be easily detected, like trees with a less evident three-toed sloth, are assigned low scores and ranked as highly spurious. This approach allows for precise differentiation between images based on their level of spuriousness and object detectability, enabling us to rank images based on

their spuriosity without any reliance on human supervision.

2. Proposed Sorting Approach

One of the most effective methods for reducing the reliance of models on spurious features is the last layer re-training approach (Kirichenko et al., 2023). To leverage this approach effectively, images within a dataset should be ranked based on their spuriosity in order to select a subset that contains low spurious cues for retraining. For large datasets, the existing ranking method (Moayeri et al., 2022) requires human supervision, which may not be feasible. Hence, we propose a method for determining feature-class dependencies without requiring human supervision. Our ranking method is based on large-scale pre-trained models that can generalize to out-of-distribution (OOD) samples.

Recent advances have applied language capabilities to object detection (Gu et al., 2022; Kamath et al., 2021; Minderer et al., 2022), integrating image and text in a shared representation space(Frome et al., 2013). OWL-ViT (Minderer et al., 2022), a cutting-edge method, combines image and text encoders using contrastive learning on a large dataset. Inspired by the CLIP model, OWL-ViT exhibits remarkable generalization and detects objects in open-vocabulary environments. Its generalization capability can help us detect high-spurious samples.

OWL-ViT architecture consists of a Vision Transformer (Dosovitskiy et al., 2021) for image encoding and a Transformer architecture for text encoding. The model image and text encoders are trained on a large-scale image-text data us-
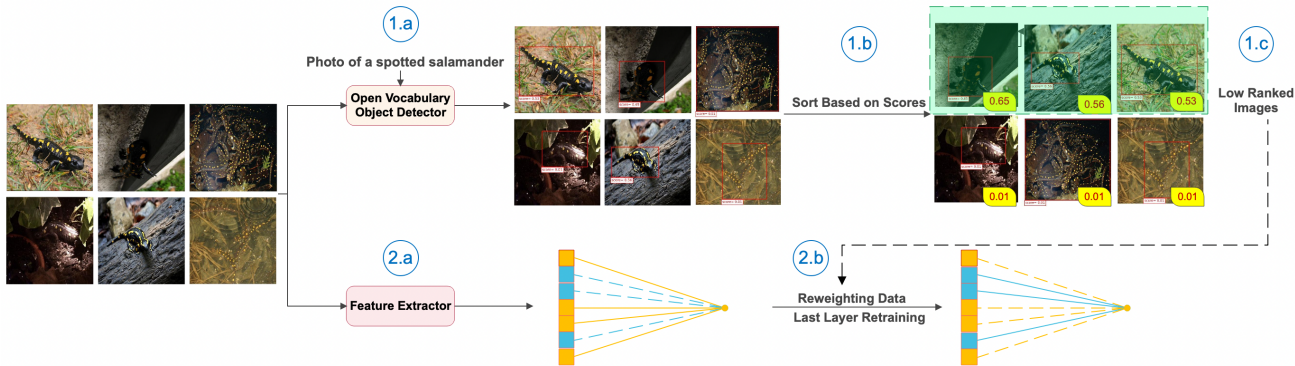
Figure 2. Schematic representation of our proposed 2-step approach. 1: a) utilizing an open-vocabulary object detector, b) sorting based on detection scores, and c) selecting low-ranked images. 2: a) training a model on the entire dataset and b) retraining the last layer.

ing contrastive learning. Then, detection heads are added to the model, and it is trained on a medium-sized detection data. OWL-ViT modifies the image encoder for object detection by projecting token representations and using a small MLP to determine the box coordinates. Predicted bounding boxes are ranked based on the confidence scores. These scores are calculated by transforming logits from a fully connected layer with a sigmoid function to obtain class probability scores. Aggregating the scores across classes produces a single score for each predicted bounding box. Notably, this score serves as a valuable metric for assessing the level of spurious features within an image and can effectively rank images within their respective classes. Specifically, higher scores correspond to lower rankings of spuriosity, indicating a more clear and more identifiable object in the image.

On this basis, as shown in Figure 2, our proposed method utilizes an open-vocabulary object detector to detect and classify objects in images. First, the model assigns detection scores to the objects based on its confidence. The images are then sorted based on these scores to refine the results and prioritize them according to their spuriosity.

To mitigate the impact of spurious correlations learned by the model during the initial detection process, the last layer of the model is retrained. This technique allows for fine-tuning the parameters in the last layer, enhancing the model ability to recognize meaningful features while disregarding irrelevant or misleading information. This two-step approach provides an efficient and effective solution for addressing spurious correlations in image classification.

The combination of accurate detection scores, automated sorting based on the spuriosity, and targeted retraining of the last layer ensures improved reliability and robustness in object classification tasks. This technique also enables the determination of the degree of spuriosity in each image, providing valuable insights into the reliability and accuracy of the detected objects. Additionally, the novel sorting technique eliminates human supervision, which is particularly
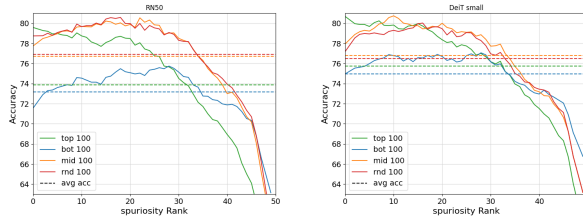

Figure 3. Impact of Training Subsets on Classification Accuracy for Validation Samples with Varying spuriosity Ranks. The dashed horizontal lines depict the average accuracy levels achieved by each model.

advantageous when working with large datasets.

3. Experiments

To demonstrate the efficacy of our ranking framework, we conduct several experiments which will be discussed in the following sections.

3.1. Effect of spuriosity Rankings

In this section, we utilize two pre-trained models on the ImageNet-1k (ResNet-50 or DeiT-small) and try to re-train only their last layer as a linear classification head on a subset from the same data from scratch. This process is similar to the method used in (Kirichenko et al., 2023). The training data subsets are selected based on our proposed approach for spuriosity rankings. More specifically, we consider four subsets each containing k samples of each class: *bot k* (lowest spuriosity rank), *mid k* (closest to median rank), *top k* (highest spuriosity rank), and *rnd k* (randomly selected samples). By incorporating these various subsets of data, we aim to capture the impact of our proposed method for ranking images in the last layer re-training method. For assessing the models' performance, each model is evaluated on varying spuriosity level test data. For this purpose, we chose the image with the i-th highest spuriosity rank from
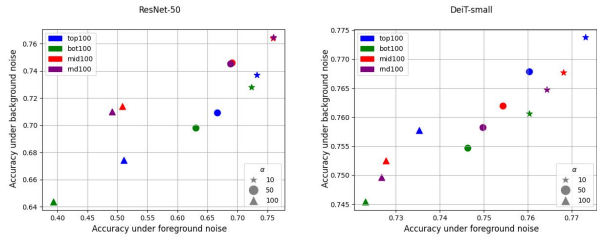
Figure 4. Experimental results show *top 100* models are more robust than *bot 100* under foreground and background noise for both ResNet and VIT models, with *top 100* VIT outperforming *rnd 100* and *mid 100* models. Results suggest *top 100* approach enhances ResNet's robustness on high-level noisy foreground samples.

each class in the test data, resulting in a set of images with varying spurious rankings.

The evaluation results for k=100 are presented in Figure 3, indicating that the model trained on *top k* data performs better on low-ranked ones, whereas the model trained on *bot k* data is more resilient to the accuracy change when assessed on data with varying spurious rankings. It is worth noting that our method achieves similar results to (Moayeri et al., 2022), but ranks images without human supervision. The results for k=50 and k=200 could be found in the appendix.

3.2. Foreground and Background Noise Sensitivity Analysis: Comparing Model Robustness

In this section, we present a sensitivity analysis on the models trained as described in section 3.1 by adding noise to the foreground or background of the images. First, we define the foreground region of ImageNet images as the area inside the detection rectangle determined by the OWL-ViT model. We create a binary mask for each sample by assigning a value of 1 to pixels within the foreground region and 0 otherwise. To create the noisy samples, a noise tensor n is generated in which the pixel values are independently drawn from $\mathcal{N}(0, 1)$. Afterward, the noisy foreground sample $\bar{x}_{fg}$ and noisy background sample $\bar{x}_{bg}$ are:

$$\bar{x}_{fg} = x + \alpha \cdot f(n \odot m_x)$$
$$\bar{x}_{bg} = x + \alpha \cdot f(n \odot (1 - m_x)),$$

where $\odot$ is the hadamard product, m_x is the binary mask for sample x, $f(.)$ is the noise ℓ_2 normalization function, and α is a scalar parameter controlling noise level. ℓ_2 normalization is applied to ensure that the magnitude of noise in the foreground and background remains consistent, thereby enabling fair comparisons between them. We consider three different levels of noise $\alpha \in \{10, 100, 250\}$ for evaluating the models' sensitivity. Figure 4 shows the results of our experiment, where the *top 100* models consistently exhibit greater robustness than the *bot 100* models under foreground and background noise for both ResNet and VIT models. The *top 100* VIT model outperforms the *rnd 100* and *mid 100*
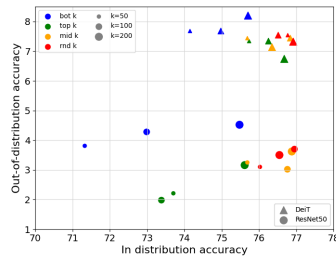

Figure 5. The performance comparison of different models on OOD dataset (ImageNet-A) and in-distribution setting. VIT models outperform others on OOD data, while the *bot k* subset models achieve superior results on ImageNet-A; *mid k* and *rnd k* subset models demonstrate strong performance across both in-distribution and OOD scenarios.

models, while the *mid 100* and *rnd 100* ResNet models demonstrate greater robustness due to their ability to learn more robust features of the foreground region in the presence of low-level noise. The results suggest that the *top 100* approach enables ResNet to achieve greater robustness on high-level noisy foreground samples, which confirms the effectiveness of our approach.

3.3. Model Performance on OOD Data

In this experiment, we aim to evaluate the performance of various models on OOD datasets, specifically ImageNet-A. The models used are the same as those trained in section 3.1, and their accuracy is assessed on both in-distribution and OOD data. The results are presented in Figure 5, indicating that VIT models outperformed convolutional models on OOD dataset. Moreover, models trained on *bot k* achieved higher accuracy on ImageNet-A, possibly due to the similarity in image style between ImageNet-A and the training data for *bot k* models. This suggests that the *bot k* models were able to effectively capture the underlying patterns and characteristics of the ImageNet-A images, resulting in improved performance on this specific OOD dataset. Additionally, models trained on *mid k* and *rnd k*, which include different image styles, demonstrate superior overall performance compared to other models in both in-distribution and OOD settings.

4. Conclusion

In conclusion, our study introduced a novel method for ranking images based on their spuriosity within classes without human supervision. Through comprehensive experiments, we examined the impact of our approach on model training, assessed model sensitivity to foreground and background noise, and evaluated model performance on OOD data. The results demonstrate the efficacy of our method in enhancing model's robustness and performance in the presence

of spurious cues. By eliminating the need for human supervision, our approach presents a promising solution for tackling spurious features and strengthening the robustness of deep neural networks in practical applications.

References

Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W., and Nguyen, A. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2019. doi: 10.1109/CVPR.2019.00498. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Alcorn_Strike_With_a_Pose_Neural_Networks_Are_Easily_Fooled_by_CVPR_2019_paper.html.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *CoRR*, abs/1907.02893, 2019. URL http://arxiv.org/abs/1907.02893.

Aubin, B., Slowik, A., Arjovsky, M., Bottou, L., and Lopez-Paz, D. Linear unit-tests for invariance discovery. *CoRR*, abs/2102.10867, 2021. URL https://arxiv.org/abs/2102.10867.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR*, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://doi.org/10.1109/CVPR.2009.5206848.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., and Mikolov, T. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems, NeurIPS*, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/7cce53cf90577442771720a370c3c723-Abstract.html.

Geirhos, R., Jacobsen, J., Michaelis, C., Zemel, R. S., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *CoRR*, abs/2004.07780, 2020. URL https://arxiv.org/abs/2004.07780.

Gu, X., Lin, T., Kuo, W., and Cui, Y. Open-vocabulary object detection via vision and language knowledge dis-tillation. In *The Tenth International Conference on Learning Representations, ICLR*, 2022. URL https://openreview.net/forum?id=lL3lnMbR4WU.

Hu, W., Niu, G., Sato, I., and Sugiyama, M. Does distributionally robust supervised learning give robust classifiers? In *Proceedings of the 35th International Conference on Machine Learning, ICML*, 2018. URL http://proceedings.mlr.press/v80/hu18a.html.

Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., and Carion, N. MDETR - modulated detection for end-to-end multi-modal understanding. In *IEEE/CVF International Conference on Computer Vision, ICCV*, 2021. doi: 10.1109/ICCV48922.2021.00180. URL https://doi.org/10.1109/ICCV48922.2021.00180.

Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations, ICLR*, 2023. URL https://openreview.net/forum?id=Zb6c8A-Fghk.

Lee, Y., Chen, A. S., Tajwar, F., Kumar, A., Yao, H., Liang, P., and Finn, C. Surgical fine-tuning improves adaptation to distribution shifts. *CoRR*, abs/2210.11466, 2022. doi: 10.48550/arXiv.2210.11466. URL https://doi.org/10.48550/arXiv.2210.11466.

Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn, C. Just train twice: Improving group robustness without training group information. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, 2021. URL http://proceedings.mlr.press/v139/liu21f.html.

Minderer, M., Gritsenko, A. A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., Wang, X., Zhai, X., Kipf, T., and Houlsby, N. Simple open-vocabulary object detection with vision transformers. In *European Conference on Computer Vision, ECCV*, 2022. URL https://doi.org/10.48550/arXiv.2205.06230.

Moayeri, M., Wang, W., Singla, S., and Feizi, S. Spuriosity rankings: Sorting data for spurious correlation robustness. *CoRR*, abs/2212.02648, 2022. doi: 10.48550/arXiv.2212.02648. URL https://doi.org/10.48550/arXiv.2212.02648.

Neuhaus, Y., Augustin, M., Boreiko, V., and Hein, M. Spurious features everywhere - large-scale detection of harmful spurious features in imagenet. *CoRR*, abs/2212.04871, 2022. doi: 10.48550/arXiv.2212.04871. URL https://doi.org/10.48550/arXiv.2212.04871.

Robey, A., Pappas, G. J., and Hassani, H. Model-based domain generalization. In *Advances in Neural Information Processing Systems, NeurIPS*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/a8f12d9486cbcc2fe0cfc5352011ad35-Abstract.html.

Rosenfeld, E., Ravikumar, P., and Risteski, A. Domain-adjusted regression or: ERM may already learn features sufficient for out-of-distribution generalization. *CoRR*, abs/2202.06856, 2022. URL https://arxiv.org/abs/2202.06856.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *CoRR*, abs/1911.08731, 2019. URL http://arxiv.org/abs/1911.08731.

Seo, S., Lee, J., and Han, B. Information-theoretic bias reduction via causal view of spurious correlation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence*, 2022. ISBN 978-1-57735-876-3. URL https://ojs.aaai.org/index.php/AAAI/article/view/20115.

Shah, H., Tamuly, K., Raghunathan, A., Jain, P., and Netrapalli, P. The pitfalls of simplicity bias in neural networks. In *Advances in Neural Information Processing Systems, NeurIPS*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6cfe0e6127fa25df2a0ef2ae1067d915-Abstract.html.

Singla, S. and Feizi, S. Salient imagenet: How to discover spurious features in deep learning? In *The Tenth International Conference on Learning Representations, ICLR*, 2022. URL https://openreview.net/forum?id=XVPqLyNxSyh.

Singla, S., Moayeri, M., and Feizi, S. Core risk minimization using salient imagenet. *CoRR*, abs/2203.15566, 2022. doi: 10.48550/arXiv.2203.15566. URL https://doi.org/10.48550/arXiv.2203.15566.

Zhang, J., Menon, A. K., Veit, A., Bhojanapalli, S., Kumar, S., and Sra, S. Coping with label shift via distributionally robust optimisation. In *9th International Conference on Learning Representations, ICLR*, 2021. URL https://openreview.net/forum?id=BtZhsSGNRNi.

A. Appendix

A.1. Experimental Analysis of Training Subsets and spuriosity Ranking on Model Performance

We conducted comprehensive experiments using two specific values of k, k=50 and k=200, to further investigate the impact of training subsets on model's performance. The models were trained using the same method described in section 3.1, on the proportion of the data sorted based on spuriosity ranking. For k=50, the evaluation results revealed that the model trained on the *top 50* subset exhibited superior performance on low-ranking data, showing its ability to effectively capture and generalize patterns from such samples. On the other hand, the model trained on the *bot 50* subset demonstrated increased resilience to accuracy changes when assessed on data with varying spurious rankings. This indicates that the model trained on the *bot 50* subset is more robust and less influenced by spurious cues present in the test data.

Similarly, for k=200, we observed a similar trend. The model trained on the *top 200* subset displayed better performance on low-ranking data, while the model trained on the *bot 200* subset exhibited higher resilience to accuracy changes when evaluated on data with varying spurious rankings.

These findings highlight the impact of different training data subsets and spuriosity rankings on the model's performance. The models trained on subsets containing low-ranking samples prioritize accuracy on those data points, while the models trained on subsets with high-ranking samples are more robust to changes in spurious correlations. These insights contribute to a deeper understanding of the relationship between training data subsets, spuriosity rankings, and model performance. The detailed visualizations and analysis of the results for k=50 and k=200 are presented in Figure 6.

A.2. Further Illustrations of spuriosity Rankings

In this part, we present additional examples that highlight the concept of spuriosity rankings, their implications for object detectability, and the presence of spurious cues for some random classes. A wide range of images is shown in 7 ranging from clear images with low spuriosity to hard-to-distinguish cases with high spuriosity.
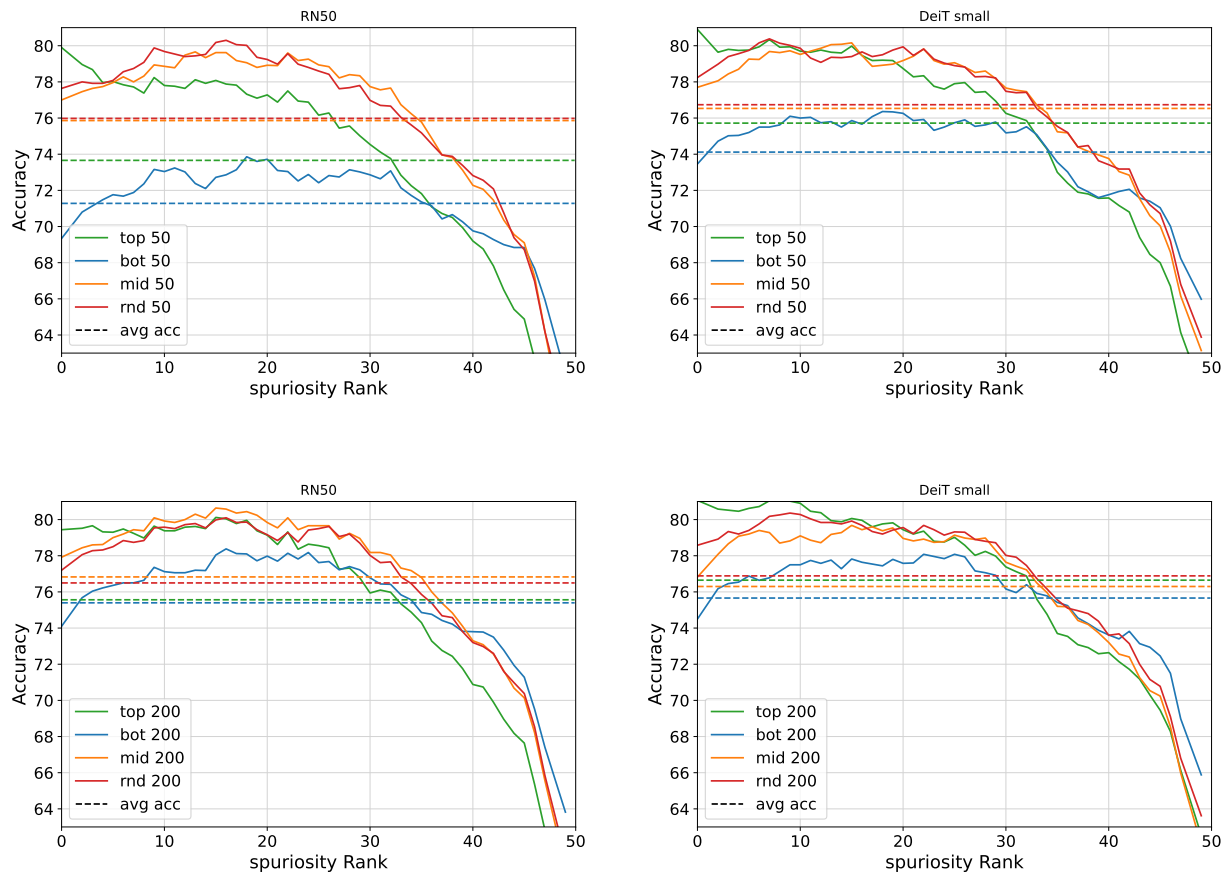
Figure 6. Impact of training subsets (k=50, 200) on classification accuracy for validation samples with Varying spuriosity ranks. The plot illustrates how different subsets of training data, each containing 50 and 200 samples per class respectively, influence the accuracy of classification. The dashed horizontal lines represent the average accuracy levels achieved by each model on different data subsets.
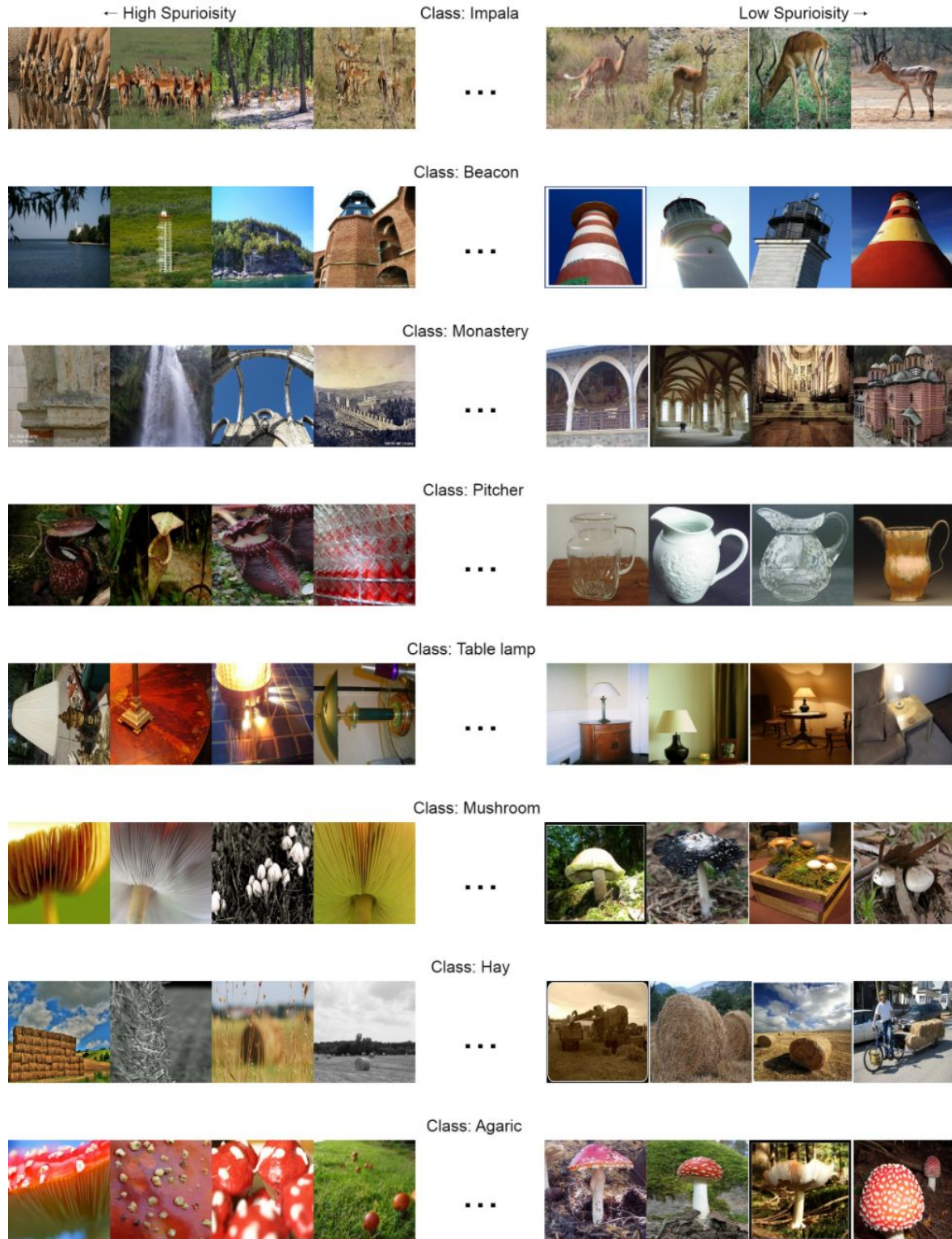
Figure 7. Examples of Spuriosity Rankings: Illustrating object detectability and spurious cues. The images on the left side demonstrate high spuriosity, showing cases where spurious cues are present, making the object difficult to distinguish. Conversely, the images on the right side exhibit low spurious cues, displaying clear images of the object.