Evaluating Robustness of Pre-Trained Deep Neural Networks against Spurious Correlations

Alireza Hoseinpour
alirezzzhp1378@gmail.com

Majid Taherkhani
majidtaherkhani555@gmail.com

Fahimeh HosseiniNoohdani
fhosseini@ce.sharif.edu

Hesam Asadollahzadeh
hesam.asadzadeh26@researcher.sharif.edu

Mahdieh Soleymani Baghshah
soleymani@sharif.edu

Sharif University of Technology
Tehran, Iran

Abstract

This paper addresses the challenge of sub-optimal performance in deep neural networks trained for image classification under non-matching distribution scenarios. Spurious correlations—patterns in training data irrelevant to objects—can lead to accuracy loss during testing. We assess the robustness of pre-trained models to spurious correlations by subjecting them to datasets containing such correlations. We compare the effectiveness of two training methods—fine-tuning and backbone freezing. Additionally, we explore the impact of robust training on our model collection by applying the DFR method introduced in [14] to both frozen and fine-tuned model backbones.

1. Introduction

Despite the notable progress made in transfer learning for image classification using deep neural networks, these models continue to exhibit sub-optimal performance when faced with non-matching distributions in the training and test sets. A prominent instance of this reduced performance is the accuracy loss during test time attributed to the spurious correlations present in the training data. Spurious correlation, in the context of image classification, refers to the occurrence of patterns or features within the training datapoints that are unrelated to the true underlying semantics of the objects or scenes being represented. These models tend to learn incorrect associations and dependencies by relying on spurious features during training. Consequently, when faced with data in which such spurious correlations are no longer present, the models exhibit a decrease in performance and as a result, fail to generalize effectively. As an example, [5] shows that a model trained for classifying cows and camels is prone to use the background as a shortcut for prediction, instead of the object in the foreground.

Although numerous approaches have been put forth to address spurious correlations during model training, the exploration of this issue remains limited when it comes to fine-tuning pre-trained models for downstream tasks. This study aims to assess the resilience of pre-trained Convolutional Neural Networks (CNNs) and Vision Transformers [7] when confronted with datasets that exhibit imbalances, specifically focusing on scenarios where certain minority groups are underrepresented. To this end, we conducted comprehensive training and evaluation experiments on the aforementioned models. In this study, we conducted an investigation to assess how the last layer (classification layer) of our model zoo influences their resilience in the presence of spurious correlations. To achieve this, we methodically evaluated two training approaches, namely fine-tuning and freezing of the pre-trained backbone. We utilized established datasets that are acknowledged for their inherent spurious attributes.

Furthermore, to investigate the impact of robust training on our model collection, we conducted assessments using the DFR method. This evaluation involved two distinct approaches: first, fine-tuning the model's backbone with a designated spurious dataset before applying the DFR method, and second, applying the DFR method directly to the model's frozen backbone.

In pursuit of a thorough assessment of these pre-trained models, we curated a novel dataset. This dataset integrates spurious color patches into images across ten distinct classes derived from the COCO [18] dataset.

Based on our experimental findings, when working with real-world datasets featuring a higher ratio of minority examples, fine-tuning not only boosts initial results but also enhances outcomes after applying DFR to the model backbone. Moreover, our research indicates that pre-trained

models excel in extracting intricate features. In situations where computational limitations come into play, fine-tuning might be dispensable, thereby mitigating associated computational and complexity costs.

2. Problem Description

Spurious correlations involve features that display notable yet unstable connections with the label throughout the dataset. The presence of these features tempts deep models to rely on them rather than the essential predictive characteristics linked to the label. Consequently, models relying on these features often struggle when encountering test sets where these correlations don't hold. To illustrate, consider the Waterbirds dataset. The labels, "waterbird" and "landbird," exhibit spurious correlations with background features, "water" and "land," respectively. This division forms four distinct groups: waterbirds on water background (G1), waterbirds on land background (G2), landbirds on water background (G3), and landbirds on land background (G4). The group sizes are 3498, 184, 56, and 1057, respectively. Due to this imbalance, the background becomes a spurious factor. Consequently, deep neural networks may achieve high average test results, but falter on the minority groups, namely G2 and G3. To evaluate the resilience of CNNs and Visual Transformers on datasets containing spurious correlations, and to examine the impact of the models' classification layer on this resilience, we conducted experiments. We applied both training methods, namely fine-tuning and backbone freeze, to our model zoo. Additionally, we utilized datasets characterized by pronounced spurious correlations as a central element of our experimentation.

3. Experiments

3.1. Datasets

Waterbirds Initially introduced in [25], Bird images sourced from the Caltech-UCSD Birds-200-2011 (CUB) [29] were selected, with labels categorized as $y \in \{waterbird, landbird\}$. These images were then paired with backgrounds from the places [34] dataset, denoted by $e \in E = \{water, land\}$, creating four distinct groups in total. The dataset's spurious correlation emerges from the notable frequency of waterbirds appearing against water backgrounds and landbirds against land backgrounds, in contrast to other groups. This dataset consists of n = 4976 training examples with the smallest group containing 56 images featuring waterbirds on land backgrounds.

CelebA This dataset consists of n = 162770 celebrity images. The label space is Y = blond hair, non-blond hair and the gender serves as a spurious feature, creating 4 distinct groups that are non-blond females, blond females, non-blond males and blond males with proportions 44%, 14%, 41% and 1% of the data respectively.

COCO-on-Colors In line with previous work [1], we incorporated 10 segmented COCO [18] objects onto backgrounds with distinct colors. Our methodology for colorizing the background in each image was as follows: For each class, 80 percent of the images were adorned with a specific background color, while the remaining 20 percent were assigned an alternative, unique background color. During the evaluation phase of the dataset, we assigned each class a distinct background color with a 90 percent probability. This color was clearly distinguishable from the color utilized in the training phase for the predominant 80 percent of the same class. In the remaining 10 percent of cases, we employed the same color that was used for the majority 80 percent of that class during training. In scenarios where an image contained multiple instances of the same object, we followed a strategy akin to that of [1]. Notably, all images were ultimately resized to dimensions of 64 × 64. It is important to note that the training set encompassed 800 images per category, while both the validation and test sets comprised 100 images each.

COCO-with-Patches We selected the identical set of 10 segmented COCO objects as those in the Coco-on-colors dataset. These objects were superimposed onto black backgrounds. Subsequently, color patches were positioned at the upper left corner of each image. Our coloration approach for these patches mirrored the scheme employed for coloring the backgrounds in the Coco-on-colors dataset. The very same set of colors used for the background of each image in that dataset was also used to color these patches.

3.2. Model zoo

In our experiments, we trained and evaluated pre-trained versions of various models. We did this by both freezing and fine-tuning their backbones and adding a nonlinear classifier to each. The models included VGG16 [13], ResNet50 [10], ViT Base, and two versions of DINO [6] using ViT Base. Additionally, we explore two variants of the vision encoder segment within the CLIP [21] model: one that adopts a ViT Base and another that employs ResNet50. All these vision transformer models, including the one utilized in the CLIP model, share a common input patch resolution of 16 × 16, except for the one utilized in DINO version two, which utilizes a 14 × 14 input patch.

3.3. Implementation Details

In both freezing and fine-tuning the backbones of our model zoo, we opted for the Adam optimizer coupled with a cosine annealing scheduler. For fine-tuning the backbone, we set the initial learning rate at 1e-5, with a minimum of 1e-6. Meanwhile, for freezing the backbone, we initialized the learning rate to 1e-3, with a minimum of 1e-5. In both training scenarios, whether it involved freezing or fine-tuning the backbone, the scheduler's initialization allowed

Training Method	Models	COCO-on-Colors	COCO-on-Colors Patches	Waterbirds		CelebA	
				Average Acc.	Worst-Group Acc.	Average Acc.	Worst-Group Acc.
Backbone Fine-Tune	VGG16	67.4 ± 1.56	58.23 ± 2.1	83.69 ± 0.39	45.37 ± 3.34	96.03 ± 0.03	38.7 ± 2.76
	ResNet50	81.43 ± 0.28	71.36 ± 0.52	87.03 ± 0.21	65.89 ± 5.36	95.20 ± 0.56	40.92 ± 8.95
	ViT-B/16	80.9 ± 0.88	67.46 ± 5.27	89.98 ± 0.91	62.56 ± 4.22	95.49 ± 0.08	41.48 ± 0.52
	CLIP	85.36 ± 5.01	72.43 ± 6.17	88.7 ± 2.57	50.93 ± 13.38	96.26 ± 0.04	46.66 ± 0.45
	CLIP RN50	79.33 ± 1.51	66.9 ± 0.78	68.17 ± 1.39	13.45 ± 3.2	94.97 ± 0.08	42.22 ± 8.65
	DINO	84.03 ± 0.85	75.9 ± 2.11	86.83 ± 1.13	69.1 ± 1.61	95.52 ± 0.13	48.14 ± 2.14
	DINO v2	90.76 ± 0.44	80.1 ± 1.72	91.63 ± 0.87	56.53 ± 15.56	96.16 ± 0.035	42.58 ± 4.99
Backbone Freeze	VGG16	69.8 ± 1.86	58.9 ± 1.81	81.92 ± 0.45	55.86 ± 2.84	92.95 ± 0.07	32.40 ± 4.06
	ResNet50	68.4 ± 0.57	49.93 ± 1.39	81.01 ± 1.02	48.54 ± 3.26	94.66 ± 0.69	18.53 ± 2.09
	ViT-B/16	78.7 ± 0.28	60.02 ± 1.01	88.08 ± 0.44	65.10 ± 3.34	95 ± 0.03	30.73 ± 0.94
	CLIP	89.2 ± 0.32	67.06 ± 0.46	85.57 ± 0.35	62.3 ± 1.27	95.36 ± 0.04	32.58 ± 0.69
	CLIP RN50	39.9 ± 6.76	27.76 ± 0.28	73.56 ± 0.56	33.45 ± 4.22	94.95 ± 0.15	30.13 ± 0.45
	DINO	88.36 ± 0.49	82.53 ± 0.38	86.64 ± 0.35	64.79 ± 1.52	95.16 ± 0.15	29.81 ± 1.88
	DINO v2	93.6 ± 0.21	88.36 ± 0.26	93.84 ± 0.26	81.72 ± 1.15	93.9 ± 0.009	23.51 ± 0.69

Table 1. Average accuracies for the model zoo over the selected datasets. Pre-trained checkpoints of the models equipped with a linear classifier were employed. The reported results (mean and standard deviation) are averaged across three separate runs for each configuration.

Training Method	Models	Waterbirds		CelebA	
		Average Acc.	Worst-Group Acc.	Average Acc.	Worst-Group Acc.
DFR Backbone Fine-Tune	VGG16	87.48 ± 0.27	85.99 ± 0.61	89.15 ± 2.97	86.81 ± 2.92
	ResNet50	90.4 ± 0.66	75.08 ± 1.2	94.57 ± 2.2	42.22 ± 1.33
	CLIP	93.12 ± 0.66	91.17 ± 0.89	93.16 ± 0.07	91.65 ± 1.17
	CLIP RN50	62.88 ± 3.97	59.47 ± 6.49	88.88 ± 0.12	79.26 ± 2.89
	DINO	92.63 ± 0.2	90.91 ± 0.36	91.72 ± 0.3	89.5 ± 0.35
	DINO v2	96.18 ± 0.56	95.00 ± 0.26	89.9 ± 3.92	87.93 ± 4.62
DFR Pre-Trained	VGG16	84.2 ± 0.16	83.3 ± 0.34	85.08 ± 0.28	82.18 ± 0.05
	ResNet50	90.26 ± 0.09	88.71 ± 0.3	86.32 ± 0.4	80.18 ± 0.25
	CLIP	89.88 ± 0.18	89.53 ± 0.22	90.59 ± 0.36	89.44 ± 0.59
	CLIP RN50	87.26 ± 0.36	88.28 ± 0.32	91.66 ± 0.24	88.14 ± 1.55
	DINO	90.04 ± 0.3	88.91 ± 0.38	89.37 ± 0.18	84.07 ± 0.69
	DINO v2	95.06 ± 0.12	93.87 ± 0.25	85.55 ± 0.45	82.77 ± 0.9

Table 2. Average and worst group accuracy of DFR on pre-trained models and models with fine-tuned backbone. The reported results (mean and standard deviation) are averaged across three separate runs for each configuration

it to transition from the initial learning rate to the minimum learning rate over the course of a single epoch. At the onset of the subsequent epoch, the scheduler would reset to the initial learning rate. Throughout our various training processes, we employed early stopping mechanisms. Specifically, training was halted if the absolute difference between the accuracies of two consecutive epochs was less than 0.2 percent, or if the accuracy of the latest epoch exceeded 99 percent. Across all our training configurations, we consistently opted for a linear classifier for all our chosen models. It is noteworthy that a consistent resolution of 224×224 was employed across all models within our model zoo.

4. Results

We will divide our discussion on results into two parts. For the first part, we will discuss real datasets such as CelebA and Waterbirds. It's worth noting that even though Waterbirds is generated synthetically, it contains a real or natural spurious feature in the images. In the second part, we will discuss COCO-on-Colors and COCO-with-Patches, where the spurious feature is a synthetic color background or color patch. These are not real spurious features like the ones mentioned earlier.

Real datasets: As shown in Table 3, for the CelebA dataset, the fine-tuning strategy consistently improves the worst-group accuracy in all of the benchmark models. This is contrary to our intuition, where we might expect the model to learn shortcut features while being fine-tuned, leading to a dip in performance for the worst group. However, it's important to consider that fine-tuning can lead to the extraction of dataset-specific features. In the case of CelebA, these dataset-specific features may capture characteristics unique to the dataset, enhancing the model's overall performance.

Regarding the Waterbirds dataset, the fine-tuning strategy consistently reduces the worst-group accuracy in most cases. This aligns with our earlier explanation, especially considering that the Waterbirds dataset comprises fewer

than 100 data points from the minority group. However, when we extend our observations to the DFR, it becomes evident that the fine-tuned model exhibits significantly better worst-group accuracy. This suggests that while the model initially experiences a drop in worst-group performance after fine-tuning, it has also managed to extract valuable features. These features contribute to the DFR's improved performance, as it can reweight these dataset-specific features to its advantage.

Synthetic datasets: Overall, results on the various COCO dataset variations were lower than those on Waterbirds and CelebA. Notably, the models exhibited higher failure rates on the COCO-with-Patches dataset in comparison to COCO-on-Colors. It's worth highlighting that for the majority of models, the fine-tuning strategy proved more effective for the COCO-with-Patches dataset, possibly due to harder-to-learn shortcuts, which causes less overfitting during fine-tuning.

Additionally, following the findings in [**?**], we train the final layer of the model on a balanced dataset with respect to group annotation and label. It's observed that the worst-group performance is dramatically enhanced in all architectures.

5. Related Work

Spurious correlations The problem of spurious correlations in supervised deep learning has come to attention in the recent years and has been studied in various vision [5, 24, 27], video [33], text [8, 30, 28], and graph [17, 31] domains. Works in the field of invariant learning propose a novel training regime to enhance models robustness to spurious correlations by enforcing them to work equally well on different data distributions[4, 2, 23]. Also, a recent, yet popular line of works has emerged which suggests sample selection and reweighting as an effective means to mitigate this problem[19, 20]. [14] observations reveal that pre-trained CNN-based vision models which are fine-tuned on biased datasets are capable of extracting the core features of the images, even if they perform poorly on out-of-distribution data.

pre-trained models robustness The study of effectiveness of pre-training vision models on their generalization ability has always been a center of interest [22, 12, 3]. [11] states that pre-training improves models robustness to adversarial examples, label corruption, class imbalance, and out-of-distribution detection. [16] makes the observation that linear probing has a better performance compared to fine-tuning pre-trained models on large distribution shifts. Some works have evaluated the performance of pre-trained models on datasets with spurious correlations [32, 28]. [9] observes that factors such as the number of parameters and size of the pre-training datasets positively affect these model's robustness. [26] reports that features extracted by self-supervised learned models and autoencoders are more robust to distribution shifts compared to features learned by supervised pre-training. As a complement to the previous studies, in this work we strive to examine the effect of architecture and multi-modal pre-training on their ability to generalize to datasets with correlation shifts.

6. Conclusion

In summary, our study assessed the robustness of pre-trained Convolutional Neural Networks (CNNs) and Vision Transformers when confronted with datasets containing spurious correlations. We conducted a systematic comparison of two training methods: fine-tuning and freezing the backbone. Furthermore, we utilized DFR [15] to assess the ability of different models to extract non-shortcut features in both pre-trained and fine-tuned configurations. Surprisingly, our findings suggest that, for real-world datasets with a higher prevalence of minority examples, fine-tuning leads to improvements in both early-stage results and outcomes after applying DFR to the backbone. This outcome contrasts with our initial intuition. In conclusion, our study suggests that pre-trained models are proficient at extracting sufficiently rich features, and given computational constraints, fine-tuning may not be necessary, thereby the associated costs can be avoided.

References

[1] Faruk Ahmed, Yoshua Bengio1, Harm van Seijen, and Aaron Courville. Systematic generalisation with group invariant predictions. *conference paper at ICLR*, 2021.

[2] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 145–155. PMLR, 13–18 Jul 2020.

[3] Isabela Albuquerque, Nikhil Naik, Junnan Li, Nitish Keskar, and Richard Socher. Improving out-of-distribution generalization via multi-task self-supervised pretraining, 2020.

[4] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020.

[5] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI*, page 472–489, Berlin, Heidelberg, 2018. Springer-Verlag.

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is

worth 16x16 words: Transformers for image recognition at scale. *In International Conference on Learning Representations*, 2021.

[8] Jacob Eisenstein. Informativeness and invariance: Two perspectives on spurious correlations in natural language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4326–4331, Seattle, United States, July 2022. Association for Computational Linguistics.

[9] Soumya Suvra Ghosal, Yifei Ming, and Yixuan Li. Are vision transformers robust to spurious correlations ? In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv:1512.03385v1*, 2015.

[11] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2712–2721. PMLR, 09–15 Jun 2019.

[12] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, Online, July 2020. Association for Computational Linguistics.

[13] Andrew Zisserman Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *conference paper at ICLR*, 2015.

[14] P. Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *ArXiv*, abs/2204.02937, 2022.

[15] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

[16] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

[17] Haoyang Li, Xin Wang, Ziwei Zhang, and Wenwu Zhu. Ood-gnn: Out-of-distribution generalized graph neural network. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7328–7340, 2023.

[18] Tsung-Yi Lin, M. Maire, Serge J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. *ArXiv, abs/1405.0312*, 2014.

[19] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6781–6792. PMLR, 18–24 Jul 2021.

[20] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. In *Advances in Neural Information Processing Systems*, 2020.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *ArXiv:2103.00020*, 2021.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[23] Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18347–18377. PMLR, 17–23 Jul 2022.

[24] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

[25] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *In International Conference on Learning Representations (ICLR)*, 2020.

[26] Yuge Shi, Imant Daunhawer, Julia E Vogt, Philip Torr, and Amartya Sanyal. How robust are pre-trained models to distribution shift? In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022.

[27] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528, 2011.

[28] Lifu Tu, Garima Lalwani, Spandana Gella, and He He. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633, 2020.

[29] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. *Technical Report CNS-TR-2011-001, California Institute of Technology*, 2011.

[30] Zhao Wang and Aron Culotta. Identifying spurious correlations for robust text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3431–3440, Online, Nov. 2020. Association for Computational Linguistics.

[31] Yingxin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. Discovering invariant rationales for graph neural

networks. In *International Conference on Learning Representations*, 2022.

[32] Yihao Xue, Ali Payani, Yu Yang, and Baharan Mirzasoleiman. Eliminating spurious correlations from pre-trained models via data mixing, 2023.

[33] Shengyu Zhang, Xusheng Feng, Wenyan Fan, Wenjing Fang, Fuli Feng, Wei Ji, Shuo Li, Li Wang, Shanshan Zhao, Zhou Zhao, Tat-Seng Chua, and Fei Wu. Video-audio domain generalization via confounder disentanglement. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):15322–15330, Jun. 2023.

[34] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, pages 1452—-1464, 2017.