Do Visual Bias Mitigation Methods Generalize? A Preliminary Study Across Domains and Modalities

Anonymous Author(s)
Affiliation
Address
email

Abstract

Although spurious correlations constitute a modality-agnostic phenomenon that can emerge across diverse domains, research on this area gives a strong emphasis on vision datasets, such as ImageNet, CelebA, and Waterbirds, which comprise natural images. In our work, we look into the question of whether bias mitigation methods originally designed for and tested on natural images can generalize to other domains and modalities, where spurious correlations arise under similar principles. To this end, we conduct an extensive cross-domain and modality comparative analysis involving 11 methods and 7 datasets spanning text, medical imaging, audio, and video. Our findings indicate that most of the evaluated approaches yield models exhibiting reduced bias compared to vanilla baselines, while a statistical analysis provides evidence for statistically significant differences relative to the baselines for only two methods.

1 Introduction

Spurious correlations, i.e., high co-occurrence between certain data attributes and target labels that do not reflect true causality, pose a major challenge to the robust generalization and fairness in deep learning [1, 2]. While extensive research explores these phenomena in computer vision datasets such as ImageNet, CelebA, and Waterbirds [3, 4, 5, 6, 7], work on other modalities remains sparse or highly task-specific [8, 9]. Despite the modality-agnostic nature of such biases, there is scarce empirical evidence about the applicability of bias mitigation methods across domains and modalities.

To address this gap, we systematically investigate the *cross-domain and modality* application of bias mitigation methods originally designed for natural images. We evaluate 11 representative approaches (BB [10], Debian [11], DI [12], End [13], GroupDRO [3], JTT [4], LfF [5], SD [14], FLAC [6], BAdd [15], and MAVias [7]) on 7 datasets involving text (Bias in Bios [16] and Jigsaw Toxic Comments [17]), medical imaging (CheXpert [18] and NIH-Chest-X-ray [19]), audio (UrbanSound8K [20] and Speech Accent Archive [21]), and videos (UCF101 [22]). Our experiments show that most methods reduce bias relative to vanilla baselines. Furthermore, a statistical analysis—including these 7 datasets and 3 additional standard image benchmarks (CelebA, Waterbirds, and UrbanCars)—provides evidence of significant improvements w.r.t. the vanilla baselines for only two methods, namely BAdd [15] and MAVias [7], though this does not preclude the potential effectiveness of other methods under alternative conditions. Overall, our results highlight the promise of transferring bias mitigation principles across modalities and contribute to a more unified perspective on mitigating spurious correlations.

2 Comparative Analysis

Each dataset defines bias as a *spurious attribute* correlated with the *target attribute*. In Bias in Bios, the model predicts a *profession* while *gender* serves as the bias attribute. In Jigsaw Toxic Comments,

Submitted to Unifying Perspectives on Learning Biases EurIPS 2025 Workshop. Do not distribute.

Table 1: Cross-modal performance comparison of bias mitigation methods. Results are reported on Jigsaw Toxic Comments (JTC), Bias in Bios (BiB), UrbanSound8K (US8K), Speech Accent Archive (SAA), and CheXpert & NIH Chest X-ray (CX&NIH). *Stat. Sign.* indicates statistical significance compared to vanilla. Blue and red indicate higher and lower performance than vanilla, respectively.

Method	JTC (Acc.)	BiB (WGA)	US8K (WGA)	SAA (WGA)	CX&NIH (WGA)	UCF101 (WGA)	Avg Rank	Stat. Sign.
Vanilla	73.5±0.4	16.9±14.7	47.5±3.4	64.9±29.8	31.0±3.9	21.0±3.2	9.5	-
GroupDro	94.0±2.7	52.5±3.5	39.7±6.7	75.0±2.2	12.5±10.2	0.0±0.0	9.3	✗
JTT	80.8±1.5	18.9±5.9	49.3±2.8	72.7±2.5	20.1±9.5	21.9±7.2	8.3	✗
SD	72.8±0.3	19.1±2.9	44.7±5.7	79.4±3.1	25.2±9.6	26.9±1.6	7.7	✗
LfF	81.8±0.6	27.9±4.0	53.0±7.9	79.6±2.0	25.0±10.3	14.6±2.6	6.9	✗
End	81.6±0.7	18.7±12.6	54.0±9.2	78.9±2.2	31.5±2.5	21.0±3.2	6.8	✗
Debian	81.6±0.7	25.7±1.6	51.3±5.1	78.9±2.3	31.8±8.2	23.5±4.3	6.4	✗
BB	14.0±0.8	0.0±0.0	65.3±8.5	80.5±2.1	61.1±7.3	20.7±5.5	6.0	✗
DI	74.2±1.3	31.3±4.2	60.5±4.4	79.7±1.7	59.3±4.4	6.6±8.8	5.3	✗
FLAC	95.8±2.5	43.1±1.4	65.5±7.3	78.4±6.1	46.3±5.8	24.8±4.1	4.3	✗
MAVias[1]	94.0±1.1	46.8±0.6	62.7±6.3	84.7±0.2	55.0±3.8	13.7±6.2	3.3	✓
BAdd	95.5±0.9	29.9±3.5	60.2±5.6	81.6±1.1	64.4±5.3	27.3±5.7	2.9	✓

the target is to train an *identity hate* comment classifier. Considering that this dataset involves lexical biases [23], we identify words that frequently appear in hateful comments but are not inherently offensive (e.g., "muslim" or "black") and create an LLM-generated test set of hard negatives that include the identified words in neutral contexts. In UrbanSounds8K, we use two classes (i.e., *engine idle* and *siren*) exhibiting the strongest correlation with sound *salience*. For Speech Accent Archive, we opt for a training subset where 90% of *male* (or *female*) voices (i.e., bias attribute) correspond to *native* (or *non-native*) English accents (i.e., target attribute). For the medical domain, we consider the union of CheXpert and NIH-Chest-X-ray datasets with *pneumothorax* being the target attribute, with 90% of the *pneumothorax*-positive samples coming from CheXpert, while 90% of the negative samples come from NIH-Chest-X-ray, making *dataset source* the bias attribute. Finally, in UCF101, where the *action* is the target, we extract scene annotations using a Places365-pretrained model, map them to indoor/outdoor classes, which serve as the bias attribute.

For the Jigsaw Toxic Comments dataset, we report accuracy on the generated test set as the evaluation metric. For the remaining datasets, we use worst-group accuracy (WGA), where groups are defined by combinations of target classes and biased attributes. We omit detailed training configurations at this stage, as this is a short paper presenting ongoing work.

Overall, according to the results in Table 1, the evaluated bias mitigation methods consistently outperform the vanilla baselines across multiple modalities, suggesting that principles originally developed for vision datasets can transfer to other domains. However, domain-specific challenges persist. In medical imaging, dataset-specific artifacts are prevalent and difficult to mitigate, often exerting a dominant influence on model predictions. Likewise, the UCF101 dataset presents high complexity due to its large number of action classes, making it difficult even for the top-performing methods to achieve significant improvement.

We additionally consider the results reported in [24] for the same 12 methods on the CelebA, WaterBirds, and UrbanCars datasets to perform a comprehensive statistical analysis across all the domains. The Friedman test confirms significant differences among methods across datasets ($p < 0.001$), while the Nemenyi post-hoc test indicates that BAdd and MAVias achieve statistically significant improvements over the vanilla baselines, demonstrating their robustness across domains.

3 Conclusion and Future Work

Our preliminary analysis indicates that bias mitigation strategies originally developed for general-purpose image datasets can partially transfer to text, audio, medical imaging, and video domains. To complete the analysis, we plan to include domain-specific baseline approaches for each set of experiments, explore larger models for complex datasets like UCF101 (reported results correspond to 3D-ResNet18), and leverage domain-tailored benchmarks such as SCUBA [25] for video action recognition. Overall, by evaluating a large set of methods across diverse domains, this work aims to establish the first large-scale benchmark for understanding how bias mitigation principles generalize—or fail—across domains and modalities.

[1]The reported results pertain to the mitigation component of MAVias, i.e., the logit alignment term.

References

[1] Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, and Aidong Zhang. Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715*, 2024.

[2] Yu Yang, Eric Gan, Gintare Karolina Dziugaite, and Baharan Mirzasoleiman. Identifying spurious biases early in training through the lens of simplicity bias. In *International Conference on Artificial Intelligence and Statistics*, pages 2953–2961. PMLR, 2024.

[3] 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. *arXiv preprint arXiv:1911.08731*, 2019.

[4] 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 *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021.

[5] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020.

[6] Ioannis Sarridis, Christos Koutlis, Symeon Papadopoulos, and Christos Diou. Flac: Fairness-aware representation learning by suppressing attribute-class associations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[7] Ioannis Sarridis, Christos Koutlis, Symeon Papadopoulos, and Christos Diou. Mavias: Mitigate any visual bias. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

[8] Oscar Chew, Hsuan-Tien Lin, Kai-Wei Chang, and Kuan-Hao Huang. Understanding and mitigating spurious correlations in text classification with neighborhood analysis. In Yvette Graham and Matthew Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1013–1025, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[9] Wiebke Toussaint Hutiri and Aaron Yi Ding. Bias in automated speaker recognition. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 230–247, 2022.

[10] Youngkyu Hong and Eunho Yang. Unbiased classification through bias-contrastive and bias-balanced learning. *Advances in Neural Information Processing Systems*, 34:26449–26461, 2021.

[11] Zhiheng Li, Anthony Hoogs, and Chenliang Xu. Discover and mitigate unknown biases with debiasing alternate networks. In *European Conference on Computer Vision*, pages 270–288. Springer, 2022.

[12] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8919–8928, 2020.

[13] Enzo Tartaglione, Carlo Alberto Barbano, and Marco Grangetto. End: Entangling and disentangling deep representations for bias correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13508–13517, 2021.

[14] Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34:1256–1272, 2021.

[15] Ioannis Sarridis, Christos Koutlis, Symeon Papadopoulos, and Christos Diou. Badd: Bias mitigation through bias addition. *In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2025.

[16] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pages 120–133, 2019.

[17] Jigsaw. Toxic comment classification challenge, 2018.

[18] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):590–597, 2019.

[19] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, 2017.

[20] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *22nd ACM International Conference on Multimedia*, pages 1041–1044, Orlando, USA, 2014.

[21] Steven H. Weinberger and Stephen A. Kunath. The speech accent archive: towards a typology of english accents. 2011.

[22] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[23] Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. Handling bias in toxic speech detection: A survey. *ACM Computing Surveys*, 55(13s):1–32, 2023.

[24] Ioannis Sarridis, Christos Koutlis, Symeon Papadopoulos, and Christos Diou. VB-Mitigator: An open-source framework for evaluating and advancing visual bias mitigation. *In 3nd AEQUITAS Workshop on Fairness and Bias in AI | co-located with ECAI 2025*, 2025.

[25] Haoxin Li, Yuan Liu, Hanwang Zhang, and Boyang Li. Mitigating and evaluating static bias of action representations in the background and the foreground. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19911–19923, 2023.