MITIGATING SPURIOUS CORRELATIONS IN IMAGE RECOGNITION MODELS USING PERFORMANCE-BASED FEATURE SAMPLING

Aarav Monga*, Rajakrishnan Somou*, Sonia Zhang*, Antonio Ortega

Department of Computer Science
University of Southern California
Los Angeles, California 90089
{armonga, somou, skzhang, aortega}@usc.edu

ABSTRACT

Existing methods for detecting and correcting spurious correlations in image recognition models often fail to identify biasing features due to incoherent groupings of biased images. There is also little exploration of targeted removal of spurious correlations in a low-dimensional feature space. To address these gaps, we propose Performance-Based Feature Sampling (PBFS), a systematic method for producing image recognition models that are debiased w.r.t. a given feature space. We introduce a method for producing coherent bias group proposals (i.e., semantically related images potentially sharing biasing feature(s)) and decorrelating biasing features from the target label using adaptive resampling. We demonstrate that our framework is able to correct for known spurious correlations, and through both established and our proposed metrics, we show that our method is able to de-bias image recognition models both w.r.t a high-dimensional feature space capturing complex representations and w.r.t. low-dimensional feature spaces representing simple physical properties.

1 INTRODUCTION

Image recognition models can develop biases due to training data. Biases in image recognition models typically refer to the learning of spurious correlations (Zhao et al., 2021; Peng et al., 2022; Adeli et al., 2021; Stock & Cisse, 2018; Kim et al., 2024b) between features unrelated to the classification task (e.g., woman) and the target class (e.g., blond). These biases can be a consequence of representation bias (Wang et al., 2019; Shahbazi et al., 2023; Mitchell, 2017) in the training data.

Models trained on biased data exhibit many kinds of undesirable performance. They can reflect gender (Bhargava & Forsyth, 2019; Mandal et al., 2023; Wang et al., 2019) or racial biases (Zhao et al., 2021; Huang et al., 2022) that are present in the dataset. Further, models trained on biased data are particularly vulnerable to degrading performance from perturbations or distribution shifts (Hendrycks & Dietterich, 2019), due to spurious correlations learned between distribution-specific features (such as contrast, texture, and size of the object (Hendrycks & Dietterich, 2019; Geirhos et al., 2018)) and the target label.

Approaches to mitigating spurious correlations typically start with identifying bias groups (i.e., training examples that contain feature(s) spuriously correlated with the target label) in the training data. Formally, given an image classifier f and a target class C, the aim is to identify features $\boldsymbol{a}_1, \boldsymbol{a}_2, \dots \boldsymbol{a}_k$ unrelated to the target label C such that $P[f(\boldsymbol{x}) = C \mid \boldsymbol{a}_1, \boldsymbol{a}_2, \dots \boldsymbol{a}_k] \neq P[f(\boldsymbol{x}) = C]$. Then, steps can be taken to decorrelate these *extra-class* features from the target class, such that $P[f(\boldsymbol{x}) = C]$ is independent of $\boldsymbol{a}_1, \boldsymbol{a}_2, \dots \boldsymbol{a}_k$. In practice, with an n-dimensional input feature space F, the images sharing biasing extra-class features occupy a slice $S \leq F \leq \mathbb{R}^n$.

Identifying the slices of the input space that are spuriously correlated with target label(s) is a difficult task. Often, the biasing features are not labeled, and they can be highly arbitrary depending on class representation in the training data. Existing works generally attempt to identify underperforming

slices in a feature space (Jain et al., 2022; Eyuboglu et al., 2022; Kim et al., 2024b; Krishnakumar et al., 2021). However, these slices may be large and incoherent, i.e., their contents are not strongly semantically related. Prior work attempted to promote coherence in identified slices by considering classifier (task-specific) embeddings and class labels alongside classifier performance (Eyuboglu et al., 2022), or by using captioning and keyword extraction to form interpretable coherent slices (Kim et al., 2024b). However, the first approach did not outperform naive approaches to identifying spurious correlations in slices, while the second limited the complexity of biasing features that can be leveraged for decorrelation. Further, while several works propose methods for slice discovery that are agnostic to the feature space, to the best of our knowledge, none test their methods with low-dimensional feature spaces for targeted bias detection and correction.

In our work, we present *Performance-Based Feature Sampling*, a novel method for identifying and correcting spurious correlations. We propose the use of bias group proposals, where training examples of a class are first grouped only by extra-class features using sparse clustering in a task-independent feature space F, in order to identify candidate coherent slices that may represent biasing features. "Task-independent" means we do not use the embedding space of a model trained on the classification task. To decorrelate extra-class features from the target label, we then propose a resampling strategy for training a new classifier f' based on the performance of a prior classifier f on these group proposals . We further introduce a novel metric for testing how well a bias correcting framework decorrelates extra-class features in a given feature space F from the target label. We test the generalizability of our framework on both high and low-dimensional feature spaces.

2 Related Work

Bias Group Identification: Bias groups contain features that are spuriously correlated with the class label. Some previous work manually identified bias groups in the training data (Calmon et al., 2017; Kamiran & Calders, 2012; Adeli et al., 2021; Alvi et al., 2018; Jain et al., 2022), which typically limits the number and nature of bias groups that can be identified. Other works identify potential bias groups by slicing and interpreting the training data in a latent feature space (Jain et al., 2022; Eyuboglu et al., 2022; Kim et al., 2024b; Krishnakumar et al., 2021). Within this category, various methods are used to identify slices with biasing features, including training linear classifiers on the feature space of a class to separate correct and incorrect classifications via a hyperplane (Jain et al., 2022), producing keyword descriptions of underperforming slices (Kim et al., 2024b), and using an error mixture model to produce coherent slices (Eyuboglu et al., 2022).

Bias Mitigation Following the identification of bias groups, bias mitigation attempts to decorrelate the predicted label from the biasing feature(s). Some existing methods mitigate spurious correlations by altering gradients or the loss function (Calmon et al., 2017; Adeli et al., 2021; Alvi et al., 2018; Bahng et al., 2020) to penalize statistical dependence of the predicted label on the biasing feature. Other works have altered the data exposed to the image classifier during training, either by adaptively resampling training data (Qraitem et al., 2023; Li & Vasconcelos, 2019; Curi et al., 2020) or by generating new training samples (Chen et al., 2024; Chawla et al., 2002).

3 Performance-based Feature Sampling

3.1 Bias Group Proposal

We form bias group proposals using sparse clustering in the task-independent feature space F with training examples of a particular class C. By clustering only on task-independent input embeddings, we prioritize semantic coherence in the resulting proposals, unlike previous work that used the classifier's embedding space and formed clusters based on model accuracy. The regions occupied by the proposals $b_1, b_2, \ldots, b_k$ form k non-overlapping slices of the feature space $F \geq \bigsqcup_{i \in k} b_i$. Since all instances of a class share class features (e.g., all instances of 'dog' have 'ears'), we expect instances of the class to be grouped by extra-class features (e.g., orientation) that should not be relevant to classification. Each bias group proposal thus may contain a shared extra-class feature that could form spurious correlations to the target class. See Appendix A for more information supporting this claim. For forming bias group proposals, we use sparse soft clustering based on non-negative kernel

regression (NNK), NNK-Means (Shekkizhar & Ortega, 2021); soft assignments achieve better input space representation than classic methods such as K-Means, promoting semantic coherence in the identified bias group proposals. See Appendix C.4 for NNK-Means configurations.

We use the CLIP embedding space (Radford et al., 2021) as the feature space for bias group proposal identification in general spurious correlation identification tasks. We also test targeted identification and correction of spurious correlations using low-dimensional feature spaces that contain a few summarizing features: the contrast value and the image $\%$ composition of white, black, gray, red, orange, yellow, green, blue, purple, pink, and brown. This is the Low-Level (LL) feature space.

3.2 Adaptive Resampling

For k bias group proposals $b_1, b_2, \ldots, b_k$ formed from class C, the average accuracy a_{avg} of the prior classifier f on class C is used to compute the sampling ratio $S_i = clip(e^{\frac{a_{avg}-a_i}{0.7}}, 1, 2.5)$ of bias group b_i. The prior classifier is a ResNet-50 base trained from scratch on the train set.

The cluster sampling ratio S_i is the expected number of times each instance of the cluster appears per epoch in fine-tuning a new classifier f' that has been pre-trained on Imagenet-1k. See Appendix B for more details on deriving the sampling ratio. The sampling ratio resamples extra-class features in low-performing bias groups, thus decorrelating these low-performing features from the class prediction. For bias group proposals that perform near the mean (i.e., no biasing w.r.t label prediction), the sampling ratio is unchanged. We thus expect debiasing of f' w.r.t the feature space F.

4 Experiments

We test on the Tiny-ImageNet (Le & Yang, 2015), CIFAR-100 (Krizhevsky et al., 2009), and CelebA (Liu et al., 2015) datasets for the CLIP feature space, and on CIFAR-100 for the LL feature space. For our classifiers, we use a ResNet-50 (He et al., 2015) base, and we benchmark against a baseline that has been fine-tuned without resampling. To verify our method against known biases, we test against a known spurious correlation in CelebA. See Appendices C.1 & C.2 for details.

4.1 Validation Accuracy

We use this metric to test general de-biasing using the CLIP feature space. We expect our de-biasing method to promote robust representation learning with fewer spurious correlations, thus improving out-of-sample (test) accuracy. For CelebA, we use hair color as the target (CelebA-blond), and we report the validation accuracy of the worst-performing group. We expect our method to improve on the accuracy of the worst-group due to decorrelation between the spurious feature (gender) and target. Table 1 highlights improvements in validation accuracy across all datasets. On CelebA-blond, worst-group accuracy jumped from 56% to 76%. See Appendix D for CelebA-blond benchmarks.

Table 1: Comparison of accuracy on validation set with and without adaptive resampling

Dataset	**Feature Space**	**Baseline**	**With Adaptive Resampling**
Tiny-ImageNet	CLIP	70.67%	71.99%
CIFAR-100	CLIP	80.25%	81.46%
CelebA	CLIP	91.96%	93.38%

4.2 Performance Under Perturbations

We use this metric to test both general de-biasing using the CLIP feature space and targeted de-biasing using the Low-Level (LL) feature space. We expect our de-biasing method to promote robust representation learning, thus preventing the correlation of distribution-specific features with the target label and improving performance on perturbed samples. In particular, we test perturbations relevant to the Low-Level (LL) feature space, since this represents targeted debiasing w.r.t a selected set of features. See Appendix C.3 for details on perturbation types and degree. Table 2

shows consistent improvements in performance for all perturbation settings on CIFAR-100 using the CLIP feature space, with the most notable relative improvements being 1.6% for the "Brightness" perturbation and 0.9% for the "Contrast" perturbation. For Tiny-ImageNet using the CLIP feature space, results were not consistent across different types of perturbations. For the LL feature space, we observed a more significant consistent improvement in performance across perturbations of CIFAR-100, with an average relative improvement of 3.0% across perturbations.

Table 2: Comparison of test set accuracy on perturbed data with and without adaptive resampling

Dataset	Feature Space	Perturbation	Baseline	With Adaptive Resampling
Tiny-ImageNet	CLIP	Brightness	55.90%	56.48%
Tiny-ImageNet	CLIP	Fog	53.34%	52.95%
Tiny-ImageNet	CLIP	Contrast	41.84%	39.97%
Tiny-ImageNet	CLIP	Snow	46.87%	48.02%
CIFAR-100	CLIP	Brightness	76.61%	77.87%
CIFAR-100	CLIP	Fog	69.23%	69.67%
CIFAR-100	CLIP	Contrast	62.76%	63.34%
CIFAR-100	CLIP	Gaussian Blur	57.66%	57.80%
CIFAR-100	Low-Level	Brightness	76.61%	77.63%
CIFAR-100	Low-Level	Fog	69.23%	71.18%
CIFAR-100	Low-Level	Contrast	62.76%	64.93%
CIFAR-100	Low-Level	Gaussian Blur	57.66%	60.28%

4.3 Test Set Cluster Accuracy Variance

We use this metric to test general de-biasing using the CLIP feature space. As with the training set, we form clusters $c_1, c_2, \ldots, c_k$ in the same feature space F of the test set using NNK-Means. Since these clusters represent groupings based on extra-class features, we expect that our de-biased classifier f' would have a lower variance in performance across these clusters than the baseline classifier. This signifies reduced dependence of classifier performance on features irrelevant to classification. We propose this metric in order to measure de-biasing w.r.t the feature space F. For the CLIP feature space, Table 3 shows significant relative decrease in test set cluster accuracy variance of 35%, 28%, and 33% for Tiny-ImageNet, CIFAR-100, and CelebA respectively.

Table 3: Comparison of accuracy variance of test set clusters with and without adaptive resampling

Dataset	Feature Space	Baseline	With Adaptive Resampling
Tiny-ImageNet	CLIP	0.0040	0.0026
CIFAR-100	CLIP	0.0025	0.0018
CelebA	CLIP	0.0172	0.0129

5 Conclusion

In this work, we propose *Performance-Based Feature Sampling*, a novel method for mitigating spurious correlations in image classifiers w.r.t a task-independent feature space F. We produce coherent bias group proposals and leverage adaptive resampling of training images. Our framework demonstrates success in addressing spurious correlations, with steep reduction in model performance variance across extra-class features in the high-level (CLIP) feature space, suggesting reduced dependence of model performance on features not relevant to classification. We also observed consistent improvement in validation accuracy and performance on perturbed samples, with particularly notable improvement using a low-level feature space, demonstrating the utility of our method in both unsupervised and targeted robustification to spurious correlations in image recognition models.

REFERENCES

Ehsan Adeli, Qingyu Zhao, Adolf Pfefferbaum, Edith V Sullivan, Li Fei-Fei, Juan Carlos Niebles, and Kilian M Pohl. Representation learning with statistical independence to mitigate bias. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2513–2523, 2021.

Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pp. 0–0, 2018.

Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, pp. 528–539. PMLR, 2020.

Shruti Bhargava and David Forsyth. Exposing and correcting the gender bias in image captioning datasets and models. *arXiv preprint arXiv:1912.00578*, 2019.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30, 2017.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Ruizhe Chen, Jianfei Yang, Huimin Xiong, Jianhong Bai, Tianxiang Hu, Jin Hao, Yang Feng, Joey Tianyi Zhou, Jian Wu, and Zuozhu Liu. Fast model debias with machine unlearning. *Advances in Neural Information Processing Systems*, 36, 2024.

Sebastian Curi, Kfir Y Levy, Stefanie Jegelka, and Andreas Krause. Adaptive sampling for stochastic risk-averse learning. *Advances in Neural Information Processing Systems*, 33:1036–1047, 2020.

Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. Domino: Discovering systematic errors with cross-modal embeddings. *arXiv preprint arXiv:2203.14960*, 2022.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Yujin Han and Difan Zou. Improving group robustness on spurious correlation requires preciser group inference. *arXiv preprint arXiv:2404.13815*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL `https://arxiv.org/abs/1512.03385`.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Jonathan Huang, Galal Galal, Mozziyar Etemadi, and Mahesh Vaidyanathan. Evaluation and mitigation of racial bias in clinical machine learning models: scoping review. *JMIR Medical Informatics*, 10(5):e36388, 2022.

Saachi Jain, Hannah Lawrence, Ankur Moitra, and Aleksander Madry. Distilling model failures as directions in latent space. *arXiv preprint arXiv:2206.14754*, 2022.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee, Jaeho Lee, and Jinwoo Shin. Discovering and mitigating visual biases through keyword explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11082–11092, 2024a.

Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee, Jaeho Lee, and Jinwoo Shin. Discovering and mitigating visual biases through keyword explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11082–11092, 2024b.

Arvindkumar Krishnakumar, Viraj Prabhu, Sruthi Sudhakar, and Judy Hoffman. Udis: Unsupervised discovery of bias in deep visual recognition models. In *BMVC*, pp. 143, 2021.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9572–9581, 2019.

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*, pp. 6781–6792. PMLR, 2021.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Abhishek Mandal, Suzanne Little, and Susan Leavy. Multimodal bias: Assessing gender bias in computer vision models with nlp techniques. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pp. 416–424, 2023.

Benjamin R Mitchell. *The spatial inductive bias of deep learning*. PhD thesis, Johns Hopkins University, 2017.

Andi Peng, Besmira Nushi, Emre Kiciman, Kori Inkpen, and Ece Kamar. Investigations of performance and bias in human-ai teamwork in hiring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 12089–12097, 2022.

Maan Qraitem, Kate Saenko, and Bryan A Plummer. Bias mimicking: A simple sampling approach for bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20311–20320, 2023.

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*, pp. 8748–8763. PMLR, 2021.

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.

Nima Shahbazi, Yin Lin, Abolfazl Asudeh, and HV Jagadish. Representation bias in data: A survey on identification and resolution techniques. *ACM Computing Surveys*, 55(13s):1–39, 2023.

Sarath Shekkizhar and Antonio Ortega. Nnk-means: Dictionary learning using non-negative kernel regression. *arXiv preprint arXiv:2110.08212*, 2021.

Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352, 2020.

Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 498–512, 2018.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5310–5319, 2019.

Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*, 2022.

Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14830–14840, 2021.

A EXTRA-CLASS FEATURE GROUPING

By clustering in a feature space of a class, we expect training examples to be grouped by extra-class features that should not be relevant to classification, since class features are shared by all training examples. Figures 1 and 2 show examples of a high-performing and low-performing cluster of the King Penguin class in Tiny-ImageNet, respectively. In both cases, it seems that orientation of the head may be a feature that is spuriously correlated with the "King Penguin" label. In Figure 1, most penguins have their head turned to the side, which may be a feature spuriously correlated with the "King Penguin" label. In Figure 2, most penguins have their head turned up, which may negatively impact classification performance. Orientation of the head appears to be one of the extra-class features used to group these examples.

Figure 1: High-performing cluster in King Penguin class.

Figure 2: Low-performing cluster in King Penguin class.

To test our assumption that our clustering method does indeed group by *extra-class* features, we make the following observation: features are represented by directions in the latent feature space. Within a class, we expect cluster centroids to share some features (class features), but differ in extra-class features. Between a cluster centroid of one class and another, we expect more features to be different.

Consider cluster i in class C with centroid c_i. We compute the distance vector d_c, which is the mean distance between c_i and all other centroids of class C:

$$d_c = \frac{1}{N_c} \sum_{i \neq j} (c_i - c_j)^2$$

Let A contain all cluster centroids in the train dataset X. We then compute d_a, which is the mean distance between c_i and cluster centroids of all classes excluding class C.

$$d_{all} = \frac{1}{N_{all \setminus C}} \sum_{a \in A, a \notin C} (c_i - a)^2$$

We sort the values of d_c, each representing a direction in the latent feature space, and we plot a linear trendline. For the corresponding directions of d_{all}, we plot a linear trendline on the same graph.

If we do cluster on extra-class features, we would expect directions in the feature space representing class features to be common for all centroids in class C, thus giving low distances in d_c in these dimensions. In d_{all}, however, these dimensions of the feature space would be significantly more different between centroid c_i of class C and centroids of other classes. Thus, in the trendline plot for dimensions of d_c and d_{all} sorted by value in d_c, smaller values of d_c resulting from shared class features should result in a steeper trendline shifted down w.r.t that of d_{all}.

Figures 3, 4, and 5 represent this observation with c_i as cluster 0 of class 0, cluster 0 of class 49, and cluster 0 of class 99 in CIFAR-100, respectively, with the 512-dimensional CLIP feature space. While the degree to which the expected observation holds varies, all 3 figures reflect the expected characteristics that would suggest that class features are shared between clusters of the same class and, thus, that clustering is based on extra-class features.

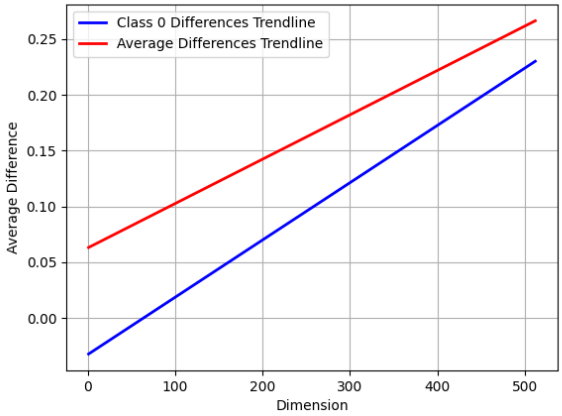

Figure 3: Average difference in CLIP dimensions for centroid 0 of class 0 w.r.t all other centroids of class 0 and w.r.t centroids of all other classes.

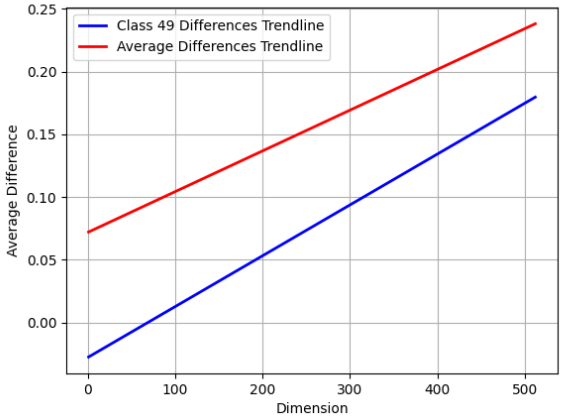

Figure 4: Average difference in CLIP dimensions for centroid 0 of class 49 w.r.t all other centroids of class 49 and w.r.t centroids of all other classes.

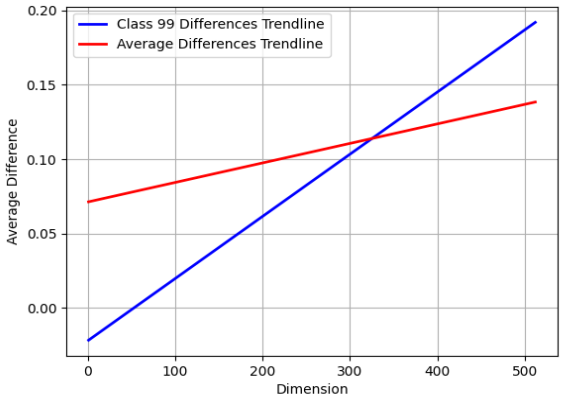

Figure 5: Average difference in CLIP dimensions for centroid 0 of class 99 w.r.t all other centroids of class 99 and w.r.t centroids of all other classes.

B SAMPLING RATIO

For k bias group proposals $b_1, b_2, \ldots, b_k$ formed from class C, the average accuracy a_{avg} of the prior classifier f on class C is used to compute the sampling ratio $S_i = clip(e^{\frac{a_{avg}-a_i}{0.7}}, 1, 2.5)$ of instances in bias group b_i.

The accuracy of the classifier f on class C, $E[accuracy(f(x))|x \in C] = \frac{1}{k}\sum_{i=1}^{k} n_i a_i$, where n_i is the number of instances in cluster i. We further have $\mathrm{Var}(accuracy(f(x))|x \in C) = \dfrac{1}{k}\displaystyle\sum_{i=1}^{k} n_i(a_i - a_{avg})^2$. To minimize the variance of the accuracy across clusters (i.e., to minimize the dependence of the performance of the classifier f' on extra-class features used for clustering), we must minimize $(a_i - a_{avg})$ across clusters. This is the basis for the sampling ratio we use.

To derive our sampling ratio, we assume log-scaling of accuracy with increasing cluster size, since our clusters are relatively small (20-40 instances on average). For a cluster with accuracy $a_i < a_{avg}$, we want to push the accuracy towards a_{avg} by increasing the size of the cluster through resampling of instances of that cluster. We thus express $a_k = a_i + \alpha * \log S_i$, where S_i is the sampling ratio of cluster i (i.e., the expected number of times an instance of the cluster will be seen per epoch in fine-tuning) and α is a hyperparameter. Rearranging gives $S_i = e^{\frac{a_k - a_i}{\alpha}}$, which is the formula we use to compute the sampling ratio. For stability and to prevent loss in accuracy via undersampling, we clip the sampling ratio between 1 and 2.5. We use $\alpha = 0.7$ to scale down the sampling ratio and prevent overcorrection of spurious correlations.

C CONFIGURATIONS

C.1 DATA DIVISIONS

For all datasets, we use the train and held-out validation set. Clustering is performed on the train set, while our metrics for debiasing are computed on the validation set. For CelebA specifically, we choose the "hair color" label as the classification target (CelebA-blond).

C.2 CLASSIFIER FINE-TUNING

The fine-tuned classifier is a ResNet-50 base that has been pre-trained on ImageNet-1k. Fine-tuning is conducted for exactly 5 epochs, with no adaptive sampling for the baseline and with adaptive sampling for the debiased classifier.

C.3 PERTURBATIONS

For perturbation testing, we use the perturbed CIFAR-100 test sets provided in (Hendrycks & Dietterich, 2019). We test on perturbations relevant to the Low-Level (LL) feature space we use in order to validate targeted correction of spurious correlations. Specifically, we test the "Brightness", "Fog", "Contrast", and "Gaussian Blur" perturbations for CIFAR-100, which affect the low level features we consider (color composition and contrast). Similarly, for Tiny-ImageNet, we test the "Brightness", "Fog", "Contrast", and "Snow" perturbations. The degree of perturbation varies from 1-5, and the data we test on is an even mix of these degrees.

C.4 NNK-MEANS CONFIGURATIONS

To select NNK-Means configurations, we run a grid-search on CIFAR-100 with the CLIP feature space over 2 hyperparameters that affect the clustering: the sparsity (i.e., number of atoms an instance can be assigned to) and initial number of atoms. We fix the entropy parameter at 0.001, and we run clustering for 15 epochs. After the NNK-Means run is complete, we assign each instance to its closest atom to generate non-overlapping clusters, and we compute the mean size and distance from centroid over all clusters. We use the mean distance from the centroid as a measure of coherence, with smaller mean distances suggesting tighter clusters that are more similar in the feature space. The results of the grid search are in Table 4.

In general, we prefer configurations that produce larger clusters and thus capture more prevalent biasing features, as well as configurations that produce more coherent clusters. To select the best configuration, we compute the ratio between the average size of clusters and the average distance to cluster centroids. We select the configuration with the highest ratio (emboldened row in Table 4), providing a balance between the prevalence and coherence of potential biasing features. From Table 4, it seems clear that reducing the initial number of atoms increases the mean size to mean distance from centroid ratio, producing a more favorable balance; however, we found that for initial atom settings smaller than 25, the variance in cluster sizes increased rapidly. This suggests more meaningless clusters with too many/few instances to have coherent or prevalent biasing features, respectively. As such, we do not consider these configurations.

To translate the selected configuration across dataset sizes, we maintain the ratio between the number of instances, the number of initial atoms, and the sparsity. As such, for both train and test clustering, the initial number of atoms and sparsity are set to 5% and 2% of the dataset size respectively.

Table 4: NNK-Means configuration grid search on CIFAR-100 with CLIP feature space

Initial Atoms	Sparsity	Avg Distance to Centroid	Avg Cluster Size	Avg Size/Avg Distance
25	**10**	**3.96**	**20.2**	**5.10**
25	15	3.97	20.2	5.08
25	20	3.98	20.2	5.08
40	10	3.80	12.6	3.32
40	15	3.81	12.6	3.31
40	20	3.82	12.6	3.30
60	10	3.61	8.4	2.32
60	15	3.63	8.4	2.31
60	20	3.63	8.4	2.31

D WORST-GROUP PERFORMANCE ON CELEBA

Table 5 compares our results on CelebA compared to other debiasing methods that also do not use pre-defined group labels. While PBFS worst-group performance shows significant improvement compared to our baseline, compared to methods like DRO-B2T, JTT, and CNC, our worst-group performance is not as strong. However, it is important to note that the aforementioned methods leverage training schemes specifically designed to improve worst-group accuracy, whereas our method focuses on general robustification of representation learning to spurious correlations, for which worst-group accuracy is one metric. This is reflected in Table 6, which demonstrates that we improve performance on all groups in CelebA-blond and maintain a high average accuracy. Compared to prior methods that produce significant gains in worst-group accuracy, we achieve the best average test accuracy despite only 5 epochs of training, which reinforces that our method is strong in promoting generalization through better representation learning.

Table 5: Comparison of accuracy on validation set with and without adaptive resampling

Method	**Average Accuracy**	**Worst-Group Accuracy**
ERM	94.9%	47.7%
GEORGE (Sohoni et al., 2020)	94.6%	54.9%
JTT (Liu et al., 2021)	88.1%	81.5%
CNC (Zhang et al., 2022)	89.9%	88.8%
GIC (Han & Zou, 2024)	92.1%	89.5%
DRO-B2T (Kim et al., 2024a)	93.2%	90.4%
Resnet-50 Baseline	92.0%	56.0%
PBFS on Resnet-50	93.8%	76.0%

Table 6: Comparison of PBFS accuracy of hair-color classification of subgroups in CelebA.

Subgroup	**Baseline**	**With Adaptive Resampling**
Dark-haired Female	89.98%	92.27%
Dark-haired Male	93.86%	95.12%
Blonde Female	97.62%	97.84%
Blonde Male	56.00%	76.00%

Our method can be adapted specifically for improving the worst-group accuracy using Distributionally Robust Optimization (Sagawa et al., 2019), a training scheme designed to minimize the worst-case loss over pre-defined groups, which would be the soft clusters produced by NNK-Means for PBFS. Using a more flexible training scheme rather than a fixed 5 epochs, considering alternate model architectures for classification, and using more benchmarks for worst-group accuracy (e.g. the WaterBirds dataset (Sagawa et al., 2019)) would also provide a more definite understanding of how PBFS can improve robustness to spurious correlations in state-of-the art models. We are unable to address these limitations in this paper due to resource and time constraints. This remains a promising direction for future research in addressing shortcut learning.