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

Anonymous authors
Paper under double-blind review

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., 2024) 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 spuriously correlated 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., 2024; 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., 2024). 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., 2024; 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., 2024), 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). 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 B for NNK-Means configurations.

We use the CLIP embedding space (Radford et al., 2021) as the feature space for bias group proposal identification for 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 instances in bias group b_i. See Appendix B for more details on the sampling ratio.

The sampling ratio of an image is the expected number of times each instance appears per epoch in fine-tuning a new classifier f' that has been pre-trained on Imagenet-1k. The sampling ratio acts to resample extra-class features that are present in low-performing bias groups, thus decorrelating the presence of these low-performing features from the class prediction. For bias group proposals that perform close to the mean performance (i.e., no biasing towards or against 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 between gender (woman) and the target hair color (blond) in CelebA. See Appendix C for clustering and training configurations and Appendix D for a rationale of the metrics used below. All experiment results are available in Appendix E.

Validation Accuracy: We observed improvements in validation accuracy across all datasets, with relative improvements of 1.9%, 1.5%, and 1.5% for Tiny-ImageNet, CIFAR-100, and CelebA respectively using CLIP feature space. On CelebA, accuracy on the worst group (blonde men) jumped from 56% to 76%.

Perturbation Performance: We observed consistent improvements in performance on perturbed data 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, performance 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.

Test Set Cluster Accuracy Variance: We propose this metric in order to measure de-biasing w.r.t the feature space F (see Appendix D for details). For the CLIP feature space, we observed significant relative change in cluster accuracy variance of 35%, 28%, and 33% for Tiny-ImageNet, CIFAR-100, and respectively.

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 use coherent bias group proposals and adaptive resampling of training images. Our framework demonstrates success in addressing spurious correlations, with steep reduction in model performance variance across extra-class features using both high-level (CLIP) and low-level feature spaces, which suggests 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 correction of biased representations 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.

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, 2024.

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.

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.

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.

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.

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 the feature space of a class, we expect training examples to be grouped by extra-class features that should not be relevant to classification, since all training examples share class features. 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 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 the model may be correlating with the class label. In Figure 2, most penguins have their head turned up, which may negatively impact classification performance. Orientation appears to be one of the extra-class features used to group these examples.

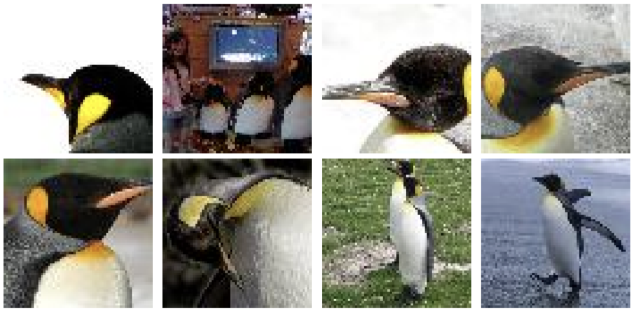

Figure 1: High-performing cluster in King Penguin class.

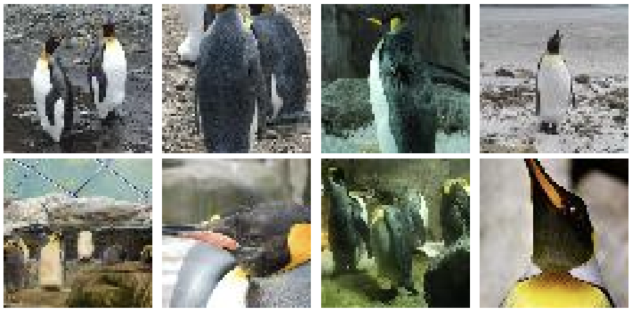

Figure 2: Low-performing cluster in King Penguin class.

To test our assumption that our clustering method does indeed group by *extra-class* features more robustly, we make the following observation: features are represented by directions in the 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 feature space (i.e., a dimension), and we plot a linear trendline. For the corresponding dimension values 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 that represent class features for C would be significantly more distant 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 with the 512-dimensional CLIP feature space, respectively. 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.

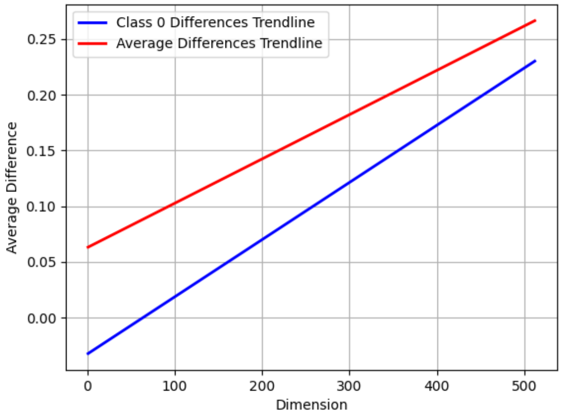

Figure 3: Average difference in CLIP dimensions for centroid 0 of class 0 w.r.t all other centroids of class 0 and w.r.t centroids of all other classes.

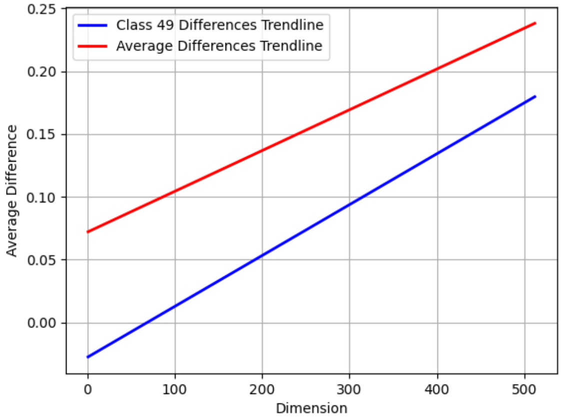

Figure 4: Average difference in CLIP dimensions for centroid 0 of class 49 w.r.t all other centroids of class 49 and w.r.t centroids of all other classes.

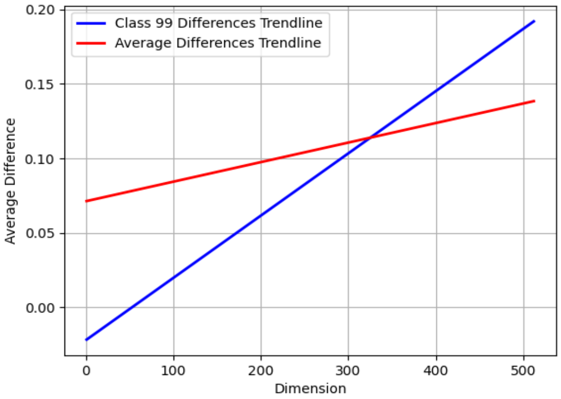

Figure 5: Average difference in CLIP dimensions for centroid 0 of class 99 w.r.t all other centroids of class 99 and w.r.t centroids of all other classes.

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 for small clusters (<20 samples). 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. In our early hyperparameter tuning, we found $\alpha = 0.7$ to promote reduction in cluster accuracy variance.

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. See Appendix D.1 for more details on this choice.

C.2 Perturbations

For perturbed datasets, 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). The degree of perturbation varies from 1-5, and the data we test on is an even mix of these degrees.

C.3 NNK-Means Configurations

For all train set clustering using NNK-Means, we set the initial number of atoms to 5% of the class size as a heuristic. We set the sparsity to half of the initial number of atoms (rounded down), and we set the entropy parameter to 0.001. The NNK-Means clustering is run for 15 epochs.

C.4 Initial Classifier Training

The initial classifier is a ResNet-50 base that has been trained from scratch on the training data of the dataset. The initial classifier is a ResNet-50 base that has been trained from scratch on the training data of the dataset.

C.5 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.

D METRICS

D.1 VALIDATION ACCURACY

We use this method 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 accuracy. As such, we expect that our method will result in better validation accuracy compared to the baseline. We also report the validation accuracy on CelebA's worst performing group (blonde men), because of a known spurious correlation in CelebA between the gender ”woman” and the hair color ”blond”. We expect our method to improve on the accuracy of classifying men with the ”blond hair” label as a consequence of decorrelation between gender (an irrelevant feature for classification) and hair color.

D.2 PERTURBATION PERFORMANCE

We use this method 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 use this metric to test the correction of spurious correlations in perturbations relevant to the Low-Level (LL) feature space, since this represents targeted debiasing. See Appendix C.2 for more details.

D.3 TEST SET CLUSTER ACCURACY VARIANCE

We use this method 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 $\mathrm{Var}(accuracy(c_i))$ than the baseline classifier that has been fine-tuned without adaptive sampling. This signifies reduced dependence of classifier performance on feaatures irrelevant to classification. We propose this metric in order to measure de-biasing w.r.t the feature space F.

E RESULTS

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%

Table 2: Comparison of 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%

Table 3: Comparison of accuracy variance of validation 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

Table 4: Comparison of test set accuracy on perturbed data with and without adaptive resampling using CLIP feature space

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%

Table 5: Comparison of test set accuracy on perturbed data with and without adaptive resampling using Low-Level feature space

Dataset	Feature Space	Perturbation	Baseline	With Adaptive Resampling
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%