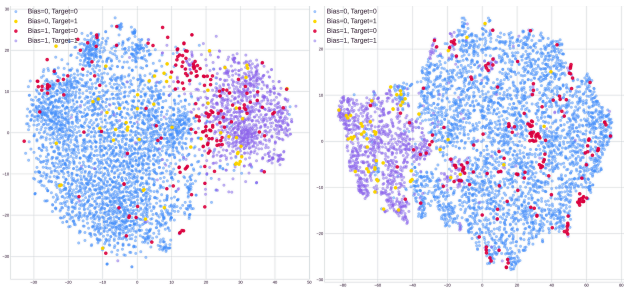
Supplementary for - Rank-Guided Pseudo-Bias Learning for Robust Black-Box Adaptation

1. Feature Representation

A core claim of our work is that standard foundation models often learn to rely on spurious correlations, while our method can successfully mitigate this reliance by reorganizing the feature space. To provide a clear, qualitative validation of this claim, we visualize the feature embeddings from the Waterbirds dataset in Figure 1 using t-SNE. This technique allows us to project the high-dimensional feature vectors into a two-dimensional space, revealing the dominant clustering structures within the data. In an ideal, unbiased representation, we would expect the data to cluster based on the primary task label (i.e., 'waterbird' vs. 'landbird'). However, in a biased representation, the features may instead cluster based on the spurious attribute (i.e., 'water' vs. 'land' background). Figure 1 presents a direct comparison of the feature space before and after our debiasing pipeline is applied, illustrating a fundamental shift from bias-driven to task-driven organization.



(a) Raw features from encoder. (b) Features after our debiasing method.

Figure 1. Comparison of feature spaces on the Waterbirds dataset. ((**??**)) In the raw feature space, data points are primarily clustered by the spurious background attribute (bias), not the true class label. For example, minority groups like waterbirds on land (yellow) are pulled towards landbirds on land (light blue). ((**??**)) After applying our method, the features are successfully reorganized. The data is now clearly separated by the target label (waterbird vs. landbird), demonstrating that the influence of the spurious correlation has been effectively mitigated.

2. Rank Regularization

2.1. Background

Rank Regularization was introduced by [2] in which they defined it as: Let $\bar{Z}$ denote the mean-centered representations Z along the batch dimension. The normalized auto-correlation matrix $C \in \mathbb{R}^{d\times d}$ of $\bar{Z}$ is defined as follow:

$$C_{i,j} = \frac{\sum_{b=1}^{n} \bar{Z}_{b,i}\bar{Z}_{b,j}}{\sqrt{\sum_{b=1}^{n} \bar{Z}_{b,i}^2}\sqrt{\sum_{b=1}^{n} \bar{Z}_{b,j}^2}} \quad \text{for } 1 \leq \forall i,j \leq d, \tag{1}$$

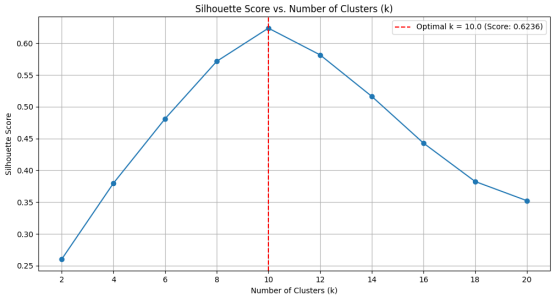
where b is an index of sample and i, j are index of each vector dimension. Then the regularization term is defined as the negative of a sum of squared off-diagonal terms in C:

$$\ell_{\text{reg}}(X;\theta) = -\sum_{i}\sum_{j\neq i} C_{i,j}^2 \tag{2}$$

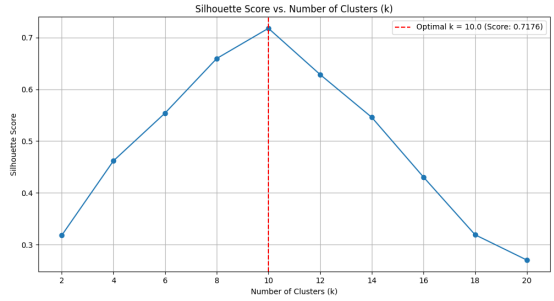
When ℓ_{reg} combined with CE loss is used to train a neural network, promotes the network to learn a representation which which is semantically restricted by the rank loss while performing sufficiently in the classification task. This results in the network to rely on biases in the dataset which are relatively easier to learn and correlate highly with the target label. [2] used this network to identify the bias-conflicting points by hypothesising that the misclassified points by this network will mostly be the bias-conflicting points. We on the other hand leverage the biased feature space by performing clustering to get pseudo bias labels and later use the pseudo labels to define the contrastive loss.

2.2. Clustering for Pseudo-Bias Labels

The feature space of a Rank-regularized network encodes bias on the basis of which we make use of clustering to extract pseudo bias labels. We have primarily made use of DBSCAN clustering for Waterbirds and CelebA datasets and K-means clustering for CMNIST to extract bias labels. The optimal number of clusters, K, was determined by evaluating the silhouette score for values in the set $K \in \{2, 4, 6, ..., 18, 20\}$. This analysis identified $K = 10$ as the optimal choice for all encoders and bias splits, yielding the maximum score, as illustrated in Figure 2



(a) k vs silhouette for CMNIST 0.9



(b) k vs silhouette for CMNIST 0.995

Figure 2. Determining the optimal number of clusters (k) by analyzing the silhouette score. The figure illustrates the score for different values of k on features extracted by a ResNet-18 model from the CMNIST dataset. The plots correspond to bias ratios of (a) 0.9 and (b) 0.995. For both levels of spurious correlation, the silhouette score is maximized at k=10, suggesting ten distinct underlying groups in the feature space.

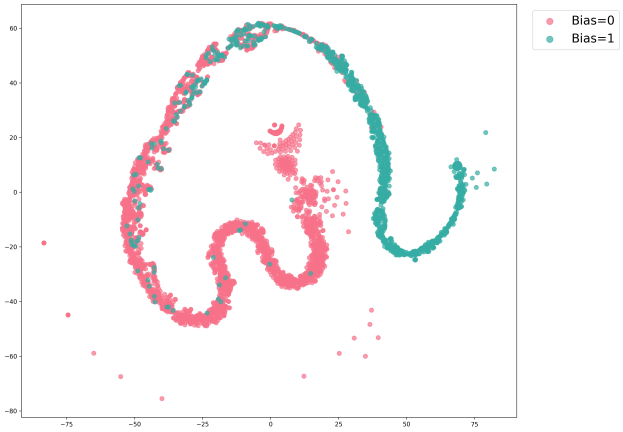


Figure 3. This Figure shows the feature representation of the rank regularized network. The rank loss effectively encourages the model to increase its feature correlation and hence the datapoints are clustered together based on their biased attribute which is much easier and simpler to learn.

3. Effective Rank

To evaluate the semantic diversity of a given representation matrix, we introduce *effective rank* [3], which is a widely used metric to quantify the effective dimensionality of a matrix. It is especially useful for analyzing the spectral properties of learned features in neural networks.

Definition 1. *Given a matrix $X \in \mathbb{R}^{m\times n}$ and its singular values $\{\sigma_i\}_{i=1}^{\min(m,n)}$, the effective rank $\rho(X)$ is defined as the Shannon entropy of normalized singular values:*

$$\rho(X) = -\sum_{i=1}^{\min(m,n)} \bar{\sigma}_i \log \bar{\sigma}_i, \tag{3}$$

where $\bar{\sigma}_i = \sigma_i / \sum_k \sigma_k$ is the i-th normalized singular value.

Effective rank, also referred to as *spectral entropy*, reaches its maximum value when all singular values are equal—indicating a uniform distribution of information—and is minimized when only a few singular values dominate. This property makes it a powerful tool for understanding how information is distributed across the feature dimensions of a learned representation.

Recent works [2] have connected effective rank to the discriminative capacity of representations, showing that high-rank representations typically encode richer and more diverse semantic information. When a model learns from biased datasets, strong spurious correlations can cause the network to collapse its representations into a low-dimensional subspace, relying heavily on a few dominant features aligned with the bias. This results in a lower effective rank and diminished generalization capability.

Through spectral analysis, it has been observed that the singular values of biased representations decay faster compared to those of unbiased ones. The top few singular values may remain large, but the rest drop sharply—indicating that many feature directions contribute little to the representation. This not only limits the semantic diversity but also undermines the network's ability to discriminate between target classes. Thus, effective rank serves as a useful proxy for measuring the extent of bias in neural representations. In our work, we use this metric to show that our proposed method improves feature diversity and reduces the encoding of spurious correlations.

4. Hyperparameters

In this section, we present the implementation details required to reproduce our work. We utilized validation accuracy for model selection and hyperparameter tuning. The primary hyperparameters in our study are the Batch Size, for which we experimented with values in the set $\{256, 512, 1024, 2048\}$, and the contrastive loss weight,

λ_{con}, with values selected from $\{0.1, 0.01, 0.001, 0.0001\}$. The rationale for choosing smaller values for λ_{con} is that the contrastive loss does not provide direct classification guidance to the model. Instead, it serves as an auxiliary signal to enhance the fairness of the learned features. A higher value of λ_{con} could potentially overshadow the primary classification signal, thereby diminishing the model's discriminative ability.

Other key hyperparameters include Weight Decay (WD) and Learning Rate (LR). We explored a range of values for these, with $WD \in \{0, 1 \times 10^{-6}, 1 \times 10^{-5}, \ldots, 0.01\}$ and $LR \in \{0.0001, 0.001, 0.01\}$. For the CMNIST dataset, lower values of LR were generally preferred, whereas for the Waterbirds and CelebA datasets, relatively higher values were employed. All models were trained for 100 epochs with a fixed Temperature (τ) of 0.5.

Tables 1 and 3 present the optimal hyperparameter values identified in our experiments. Any additional hyperparameters related to the Adaptive Margin loss are set to the values specified in the original paper by Basu et al. [1].

Dataset	Encoder	BS	λ_{con}	WD	LR	Epoch
Waterbirds	ResNet-18	1024	0.01	0	0.01	100
	CLIP	1024	0.001	0.01	0.01	100
CelebA	ResNet-18	2048	0.001	0.01	0.01	100

Table 1. Debiasing training configurations for Waterbirds and CelebA dataset

Table 2. Stage 1 hyperparameters for each dataset.

Dataset	LR	Weight Decay	Batch Size
CelebA	3e-4	0	256
Waterbirds	3e-4	0	256
CMNIST	1e-4	0	256

Dataset	Encoder	BS	λ_{con}	WD	LR	Epoch
CMNIST 0.9	ResNet-18	1024	0.01	0	0.01	100
	CLIP	1024	0.01	1×10^{-6}	0.0001	100
	ViT-B	512	0.001	0	0.0001	100
CMNIST 0.995	ResNet-18	512	0.001	1×10^{-5}	0.001	100
	CLIP	1024	0.001	1×10^{-5}	0.0001	100
	ViT-B	1024	0.001	1×10^{-4}	0.0001	100

Table 3. Training configurations for CMNIST datasets with varying spurious correlation levels.

References

[1] Abhipsa Basu, Saswat Subhajyoti Mallick, and Venkatesh Babu Radhakrishnan. Mitigating biases in blackbox feature extractors for image classification tasks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3

[2] Geon Yeong Park, Chanyong Jung, Sangmin Lee, Jong Chul Ye, and Sang Wan Lee. Self-supervised debiasing using low rank regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12395–12405, 2024. 1, 2

[3] Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European Signal Processing Conference*, pages 606–610, 2007. 2