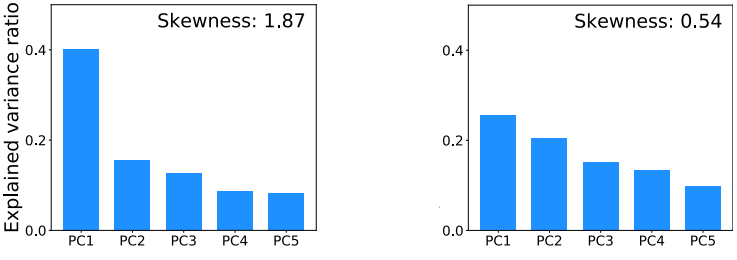
Feature and Label Embedding Spaces Matter in Addressing Image Classifier Bias

Supplementary material

Bias removal. Once the proposed model has been trained, we compute $\mathbf{\Delta}$ in Eq. 2 from the training set of CIFAR-10S. When performing a principal component analysis on $\mathbf{\Delta}$, we observe that there remains a main direction explaining the variance (Figure 4(a)). Though, compared with the baseline model (Figure 2(a)), our model with protected embeddings reduces the skewness from 2.63 to 1.87. This effect is even more noticeable after the bias removal (Figure 4(b)). Indeed, the skewness drops to 0.54 and there is no longer a main direction of variance. The bias removal operation reduces the presence of the bias in the feature space.



(a) Before. (b) After.

Figure 4: **Bias removal in the feature space effect** on the explained variance of the principal components of $\mathbf{\Delta}$ on CIFAR-10S. After the removal of the bias direction, there is no longer a main direction of variance as illustrated by a reduced skewness.

Samples from datasets. Figure 5 shows examples of the color bias in CIFAR-10S [50] while Figure 6 shows examples of the gender bias in CelebA [32].

Figure 5: **CIFAR-10S samples**, where five classes are skewed towards color images and five other classes are skewed towards gray images in the training set.

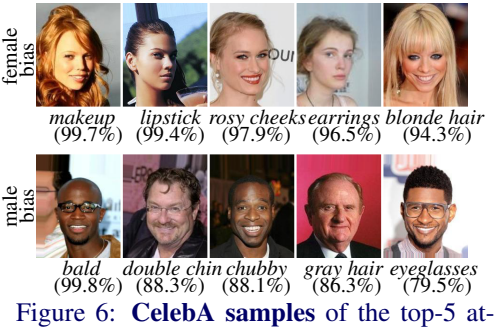


Figure 6: **CelebA samples** of the top-5 attributes skewed towards "female" and "male" genders in the training set.