Through a Fair Looking-Glass: Mitigating Bias in Image Datasets

Amirarsalan Rajabi[†], Mehdi Yazdani-Jahromi[†], Ozlem Ozmen Garibay[†], Gita Sukthankar[†]

[†]Department of Computer Science, University of Central Florida

An Important Problem

Bias in computer vision is a major problem, often stemming from the training datasets used for computer vision models [6]. This work addresses the issue of a decision-making process being dependent on *protected attributes*, where this dependence should ideally be avoided. We propose a model to reconstruct an image dataset to reduce statistical dependency between a protected attribute and target attribute. We modify a U-net [5] to reconstruct the image dataset and apply the Hilbert-Schmidt norm of the cross-covariance operator [2] between reproducing kernel Hilbert spaces of the target attribute and the protected attribute, as a measure of statistical dependence.

Methodology

Consider a dataset $D = (\mathcal{X}, \mathcal{S}, \mathcal{Y})$, where $\mathcal{X}$ is the set of images, $\mathcal{Y} = \{+1, -1\}$ is the target attribute such as attractiveness, and $\mathcal{S} = \{A, B, C, ...\}$ is the protected attribute such as gender. Assume there exists a classifier $f : (\mathcal{X}) \rightarrow \mathcal{Y}$, such that the classifier's prediction for target attribute is not independent from the protected attribute, i.e. $f(\mathcal{X}) \not\perp \mathcal{S}$. Our objective is to design a transformation $g : \mathcal{X} \rightarrow \widetilde{\mathcal{X}}$, such that 1) $f(\widetilde{\mathcal{X}}) \perp \mathcal{S}$, i.e. the classifier's predictions for target attribute is independent of the protected attribute , and 2) $f(\widetilde{\mathcal{X}}) \approx f(\mathcal{X})$, i.e. the classifier still achieves high accuracy.

We seek to modify a set of images, such that 1) the produced images are close to the original images, and 2) the predicted target attribute is independent from the predicted protected attribute. In the optimization problem, image quality (1) is measured by pixel-wise MSE loss. For independence (2), consider our U-net network as a mapping from original image to the transformed image, i.e. $U_w(\mathbf{x}) = \widetilde{\mathbf{x}}$. Consider also a function $h : \mathcal{X} \rightarrow [0,1] \times [0,1]$, where $h(\mathbf{x}_i) = (h_1(\mathbf{x}_i), h_2(\mathbf{x}_i)) = (\mathrm{P}(y_i = 1|\mathbf{x}_i), \mathrm{P}(s_i = 1|\mathbf{x}_i))$. Our objective is to train the parameters of U_w such that $h_1(U_w(\mathbf{x}))h_2(U_w(\mathbf{x}))$, i.e. $h_1(U_w(\mathbf{x}))$ is independent of $h_2(U_w(\mathbf{x}))$.

Given X representing a batch of N training images and $\widetilde{X}$ representing the transformed batch, our formal optimization problem is as follows:

$$\underset{U_w}{\text{minimize}} \underbrace{\frac{1}{NCWH} \sum_{n=1}^{N} \sum_{i,j,k} (\mathbf{x}_{ijk}^{n} - \widetilde{\mathbf{x}}_{ijk}^{n})^2}_{\text{image accuracy}} + \lambda \times \underbrace{HSIC(h_1(\widetilde{X}), h_2(\widetilde{X}))}_{\text{independence}} \tag{1}$$

where N is the number of samples, C is the number of channels of an image, W is the width of an image, H is the height of an image, and λ is the parameter that controls the trade-off between accuracy of the transformed images and independence (fairness). In practice, the mapping function U_w that we use is a U-net, the function $h(\cdot)$ is a pre-trained classifier with two outputs h_1 and h_2, each being the output of a Sigmoid function within the range of $[0, 1]$, where $h_1 = \mathrm{P}(Y = 1|X)$ (a vector of size N), and $h_2 = \mathrm{P}(S = 1|X)$ (also a vector of size N), and $HSIC(\cdot, \cdot)$ denotes Hilbert-Schmidt Independence Criteria.

Experiments

we test the methodology described in Section Methodology on CelebA dataset [3]. CelebA is a popular dataset that is widely used for training and testing models for face detection, particularly recognising facial attributes. It consists of 202,599 face images of celebrities, with 10,177 identities. Each image is annotated with 40 different binary attributes describing the image. The CelebA dataset is reported to be biased. In this experiment, we consider `Male` attribute as the protected attribute (with $\texttt{Male} = 0$ showing the image does not belong to a man and $\texttt{Male} = 1$ showing the image belongs to a man), and `Attractive` to be the target attribute.

Results

After removing the attributes with less than 5% positive images, the remaining 26 attributes are categorized into three groups. *inconsistently-labeled*, *gender-dependent*, and *gender-independent* attributes. For attribute classifiers, we use ResNet-18 pre-trained on ImageNet, in which the last layer is replaced with a layer of size one, along with a Sigmoid activation for binary classification. We compare our results with Ramaswamy et al.'s method, described in their paper 'Fair Attribute Classification through Latent Space De-biasing' [4], and 'explicit removal of biases from neural network embeddings', presented in [1].

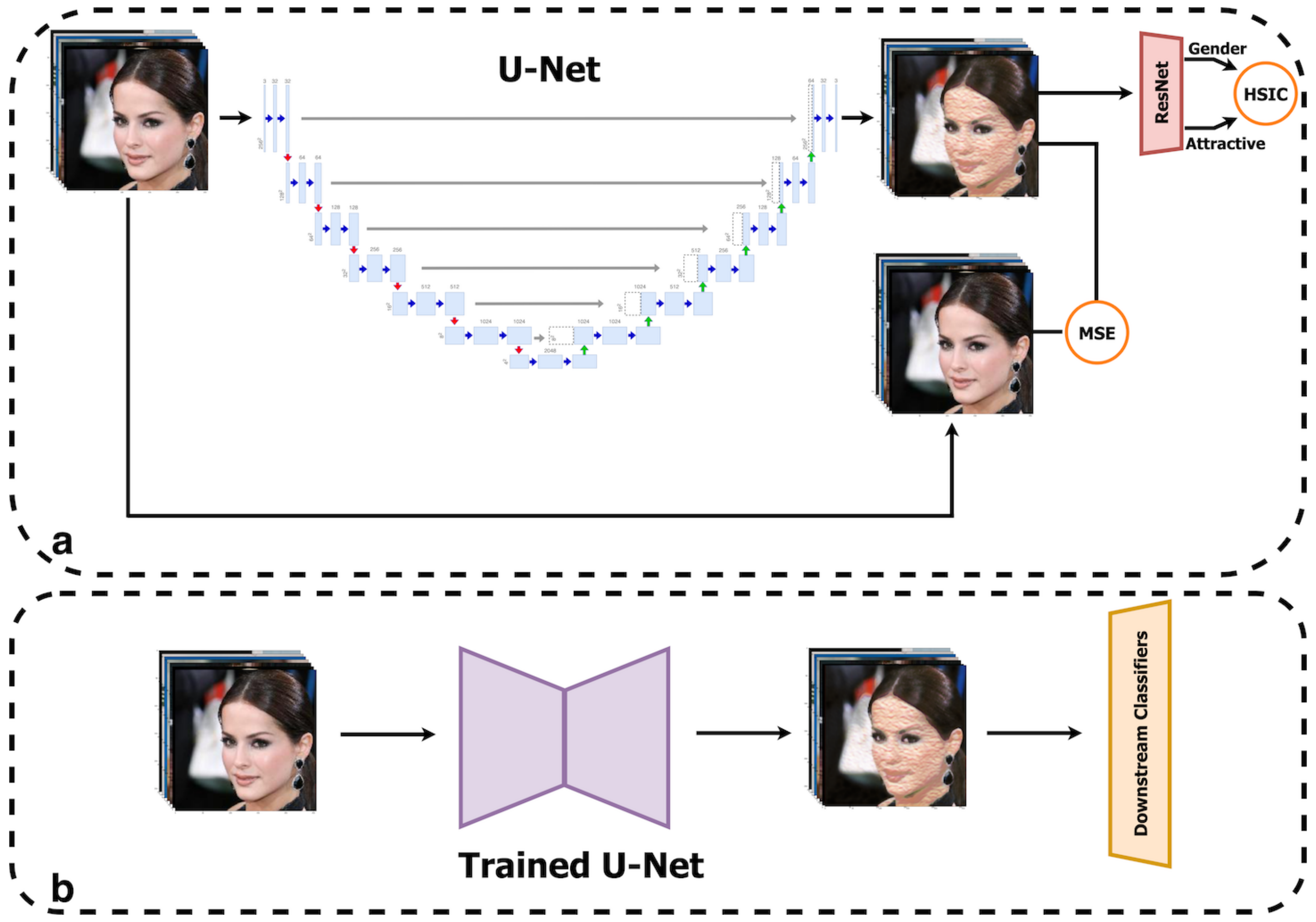


Fig. 1: Model consists of an encoder-decoder (U-net) and a double-output pre-trained ResNet classifier. First, the output batch of the U-net (reconstructed images) is compared with the original batch of images by calculating MSE loss. Then, the output batch of the U-net passes through the ResNet and statistical dependency of the two vectors is calculated by HSIC. Detailed architecture of the U-net is described in the supplementary material.

In evaluating the results of our model with the baseline models, three metrics are used. To capture the accuracy of the classifiers, we measure the *average precision* (**AP**). To measure fairness, we use *demographic parity* (**DP**). This metric captures the disparity of receiving a positive decision among different protected groups ($|P(\hat{Y} = 1|S = 0) - P(\hat{Y} = 1|S = 1)|$), and *difference in equality of opportunity* (**DEO**), i.e. the absolute difference between the true positive rates for both gender expressions ($|TPR(S = 0) - TPR(S = 1)|$).

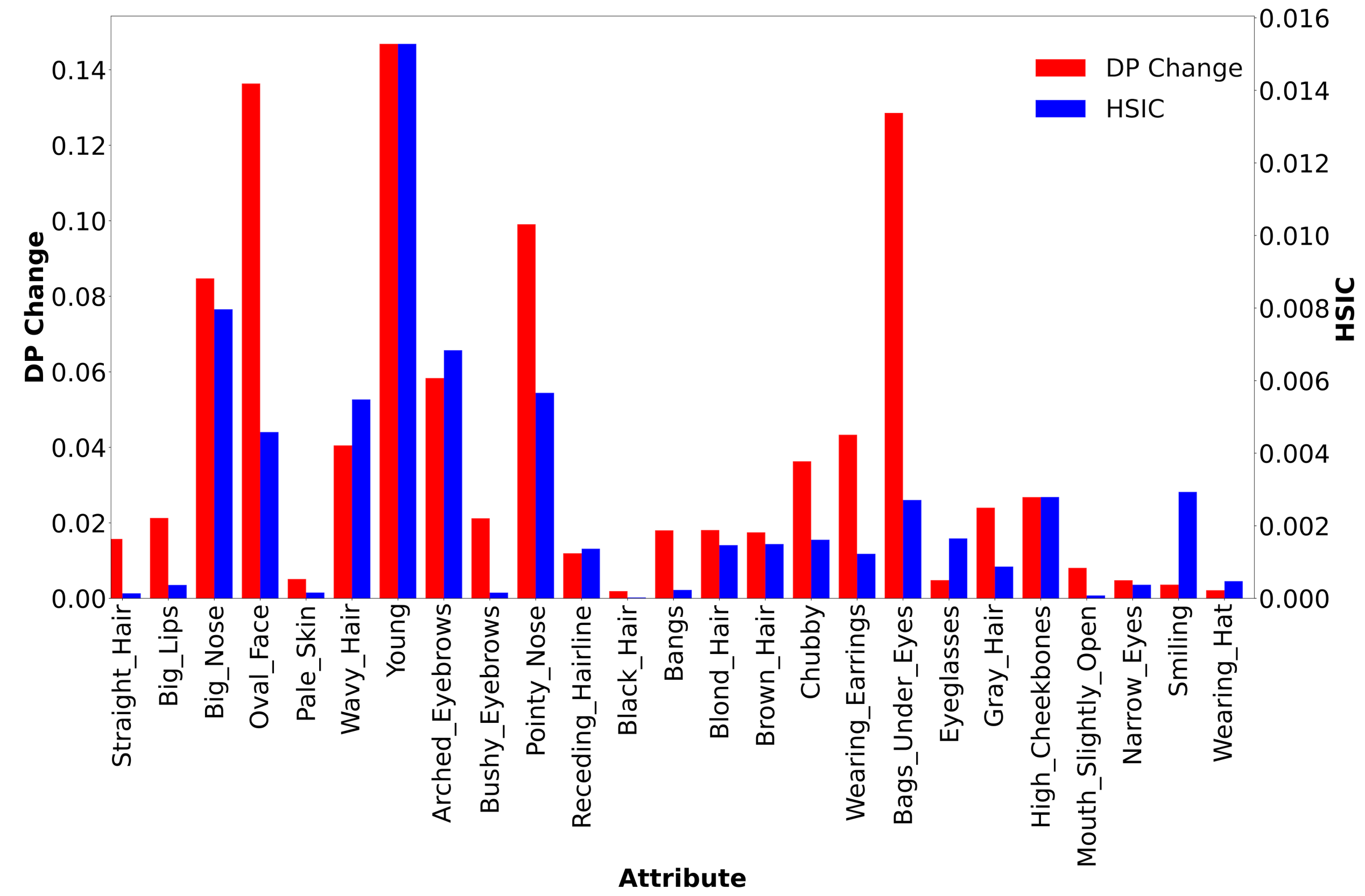
	AP ↑			**DP ↓**			**DEO ↓**		
	Incons.	G-dep	G-indep	Incons.	G-dep	G-indep	Incons.	G-dep	G-indep
Baseline	0.667	0.79	0.843	0.147	0.255	0.137	0.186	0.243	0.163
GanDeb	0.641	0.763	0.831	0.106	0.233	0.119	0.158	0.24	0.142
AdvDb	0.243	0.333	0.218	0.091	0.169	0.121	0.136	0.149	0.098
Ours	0.618	0.732	0.839	0.097	0.146	0.118	0.124	0.172	0.114

Fig. 2: Comparing the results of our model with Baseline, GAN debiasing (GanDeb), and Adversarial debiasing (AdvDb). Showing AP (Average Precision, higher the better), DP (Demographic Parity, lower the better), and DEO (Difference in Equality of Opportunity, lower the better) values for each attribute category. Each number is the average over all attributes within that specific attribute category.

The results show that Ours (our model) is close to GanDeb in terms of maintaining high average precision scores, which means higher accuracy of prediction, while beating GanDeb in terms of fairness metrics. Also, while AdvDb performance in terms of fairness enforcement is better than ours in 3 out of 6 cases, it falls behind significantly in terms of average precision.

Interpretation and Conclusion

For each attribute, we record two values, namely HSIC value between that attribute and the `Attractive` attribute, and the change in demographic parity. To calculate the change in demographic parity, we first calculate the demographic parity of the classifier for that specific attribute, when the classifier classifies the original testing set images. We then calculate the demographic parity of the classifier for that specific attribute, when the classifier receives the modified training images **Ours(5,0.07)**. We then subtract the two values, to get the change in demographic parity for that specific attribute. The results show that the absolute change in demographic parity is positively correlated with that attribute's statistical dependence with the attribute `Attractive`, with a Pearson correlation coefficient of 0.757. For instance, we observe large changes in demographic parity for attributes such as `Young`, `Big_Nose`, `Pointy_Nose`, `Oval_Face`, and `Arched_Eyebrows`, as they are typically associated with being attractive, and therefore reflected in the CelebA dataset labels.



The proposed model showed promising results in mitigating bias while maintaining high precision for classifiers. An interesting aspect of the results is that although we only explicitly train the U-net to remove dependence between the target attribute (`Attractive`) and the protected attribute (`Male`), classifiers related to many other attributes, most of which have a statistical dependency with the target attribute, become 'fairer'. An advantage of the proposed model is that it does not rely on modifying downstream classifiers, and rather includes only modifying the input data, hence making it suitable to be deployed in an automated machine learning pipeline more easily and with lower cost.

References

[1] 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*. 2018, pp. 0–0.

[2] Arthur Gretton et al. "Measuring statistical dependence with Hilbert-Schmidt norms". In: *International conference on algorithmic learning theory*. Springer. 2005, pp. 63–77.

[3] Ziwei Liu et al. "Deep Learning Face Attributes in the Wild". In: *Proceedings of International Conference on Computer Vision (ICCV)*. Dec. 2015.

[4] Vikram V Ramaswamy, Sunnie SY Kim, and Olga Russakovsky. "Fair attribute classification through latent space de-biasing". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 9301–9310.

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[6] Tatiana Tommasi et al. "A deeper look at dataset bias". In: *Domain adaptation in computer vision applications*. Springer. 2017, pp. 37–55.