Supplementary Material: Fair Attribute Classification through Latent Space De-biasing

Vikram V. Ramaswamy, Sunnie S. Y. Kim, Olga Russakovsky
Princeton University
{vr23, suhk, olgarus}@cs.princeton.edu

In this supplementary document, we provide additional details on certain sections of the main paper.

Section 1: We derive a closed form solution for $\mathbf{z}'$ which allows us to easily manipulate latent vectors in the latent space (Section 3).

Section 2 We provide attribute-level results and further analysis of our main experiments (Section 4.1).

Section 3: We discuss some factors that influence (or not) our method's effectiveness.

Section 4: We provide more details on the ablation studies (Section 4.2).

Section 5: We investigate how many images with protected attribute labels our method requires to achieve the desired performance.

1. Derivation

In Section 3 of the main paper, we describe a method to compute perturbations within the latent vector space, such that the protected attribute score changes, while the target attribute score remains the same. More formally, if h_t is a function that approximates the target attribute score, and h_g is a function that approximates the protected attribute score, for every latent vector $\mathbf{z}$, we want to compute $\mathbf{z}'$ such that

$$h_t(\mathbf{z}') = h_t(\mathbf{z}), \quad h_g(\mathbf{z}') = -h_g(\mathbf{z}). \tag{1}$$

We assume that the latent space $\mathcal{Z}$ is approximately linearly separable in the semantic attributes. h_t and h_g thus can be represented as linear models $\mathbf{w_t}$ and $\mathbf{w_g}$, normalized as $||\mathbf{w_t}|| = 1, ||\mathbf{w_g}|| = 1$, for the target and protected attribute respectively, with intercepts b_t and b_g.

Equation 1 thus reduces to

$$\mathbf{w_t}^T\mathbf{z} + b_t = \mathbf{w_t}^T\mathbf{z}' + b_t, \quad \mathbf{w_g}^T\mathbf{z}' + b_g = -\mathbf{w_g}^T\mathbf{z} - b_g. \tag{2}$$

Simplifying, we get

$$\mathbf{w_t}^T(\mathbf{z}' - \mathbf{z}) = 0, \quad \mathbf{w_g}^T(\mathbf{z}' + \mathbf{z}) + 2b_g = 0. \tag{3}$$

These equations have infinitely many solutions, we choose the solution that minimizes the distance between $\mathbf{z}$ and $\mathbf{z}'$. This is true if $\mathbf{z}' - \mathbf{z}$ is in the span of $\{\mathbf{w_g}, \mathbf{w_t}\}$. Hence, we can represent $\mathbf{z}' - \mathbf{z} = \alpha\mathbf{w_t} + \beta\mathbf{w_g}$, and we get:

$$\mathbf{w_t}^T(\mathbf{z}' - \mathbf{z}) = 0 \tag{4}$$
$$\mathbf{w_t}^T(\alpha\mathbf{w_t} + \beta\mathbf{w_g}) = 0 \tag{5}$$
$$\Rightarrow \alpha = -\beta\mathbf{w_t}^T\mathbf{w_g} \tag{6}$$
$$\mathbf{w_g}^T((\mathbf{z}' - \mathbf{z}) + 2\mathbf{z}) + 2b_g = 0 \tag{7}$$
$$\mathbf{w_g}^T(\alpha\mathbf{w_t} + \beta\mathbf{w_g} + 2\mathbf{z}) + 2b_g = 0 \tag{8}$$
$$-\beta(\mathbf{w_t}^T\mathbf{w_g})^2 + \beta + 2\mathbf{w_g}^T\mathbf{z} + 2b_g = 0 \tag{9}$$
$$\Rightarrow (1 - (\mathbf{w_t}^T\mathbf{w_g})^2)\beta = -2(\mathbf{w_g}^T\mathbf{z} + b_g) \tag{10}$$
$$\Rightarrow \beta = -2\frac{(\mathbf{w_g}^T\mathbf{z} + b_g)}{(1 - (\mathbf{w_t}^T\mathbf{w_g})^2)} \tag{11}$$
$$\Rightarrow \alpha = 2\frac{(\mathbf{w_g}^T\mathbf{z} + b_g)(\mathbf{w_t}^T\mathbf{w_g})}{(1 - (\mathbf{w_t}^T\mathbf{w_g})^2)} \tag{12}$$

This gives us a closed form solution for $\mathbf{z}'$:

$$\mathbf{z}' = \mathbf{z} - 2\left(\frac{\mathbf{w_g}^T\mathbf{z} + b_g}{1 - (\mathbf{w_g}^T\mathbf{w_t})^2}\right)\left(\mathbf{w_g} - (\mathbf{w_g}^T\mathbf{w_t})\mathbf{w_t}\right). \tag{13}$$

As a quick verification, we confirm that this value of $\mathbf{z}'$ maintains changes the protected attribute score, and maintains the target attribute score:

$$\begin{aligned}
&h_g(\mathbf{z}') \\
&= \mathbf{w_g}^T\mathbf{z}' + b_g \\
&= \mathbf{w_g}^T\left[\mathbf{z} - 2\left(\frac{\mathbf{w_g}^T\mathbf{z} + b_g}{1 - (\mathbf{w_g}^T\mathbf{w_t})^2}\right)\left(\mathbf{w_g} - (\mathbf{w_g}^T\mathbf{w_t})\mathbf{w_t}\right)\right] + b_g \\
&= \mathbf{w_g}^T\mathbf{z} - 2\left(\frac{\mathbf{w_g}^T\mathbf{z} + b_g}{1 - (\mathbf{w_g}^T\mathbf{w_t})^2}\right)\left(1 - (\mathbf{w_g}^T\mathbf{w_t})\mathbf{w_g}^T\mathbf{w_t}\right) + b_g \\
&= \mathbf{w_g}^T\mathbf{z} - 2(\mathbf{w_g}^T\mathbf{z} + b_g) + b_g = -\mathbf{w_g}^T\mathbf{z} - b_g = -h_g(\mathbf{z})
\end{aligned}$$

$$
\begin{aligned}
& h_a(\mathbf{z}') \\
&= \mathbf{w_t}^T \mathbf{z}' + b_t \\
&= \mathbf{w_t}^T \left[\mathbf{z} - 2 \left(\frac{\mathbf{w_g}^T \mathbf{z} + b_g}{1 - (\mathbf{w_g}^T \mathbf{w_t})^2} \right) \left(\mathbf{w_g} - (\mathbf{w_g}^T \mathbf{w_t}) \mathbf{w_t} \right) \right] + b_t \\
&= \mathbf{w_t}^T \mathbf{z} - 2 \left(\frac{\mathbf{w_g}^T \mathbf{z}}{1 - (\mathbf{w_g}^T \mathbf{w_t})^2} \right) \left(\mathbf{w_t}^T \mathbf{w_g} - (\mathbf{w_g}^T \mathbf{w_t}) \right) + b_t \\
&= \mathbf{w_t}^T \mathbf{z} + b_t = h_t(\mathbf{z})
\end{aligned}
$$

2. Attribute-level results

We provide attribute-level results and further analysis of our main experiments (Section 4.1 of the main paper).

2.1. Linear separability of latent space

Our paired augmentation method assumes that the latent space is approximately linearly separable in the semantic attributes. Here we investigate to what extent this assumption holds for different attributes. As described in the main paper, the attribute hyperplanes were estimated with 10,000 samples using linear SVM.

In Table 1, we report hyperplane accuracy and AP, measured on 160,000 synthetic samples, as well as the percentage of positive samples and the skew of the CelebA training set. The skew is calculated as $\frac{\max(N_{g=-1,a=1}, N_{g=1,a=1})}{N_{g=-1,a=1} + N_{g=1,a=1}}$ where $N_{g=-1,a=1}$ is the number of samples with protected attribute label $g=-1$ (perceived as not male) and target label 1 (positive) and $N_{g=1,a=1}$ defined likewise. The protected attribute class with more positive samples is noted in the skew column. We observe that most attributes are well separated with the estimated hyperplanes, except for those with high skew that have too few examples from underrepresented subgroups.

For completeness, we also report our model's improvement over the baseline model on the four evaluation metrics. We did not find immediate correlations between the hyperplane quality with the downstream model performance.

2.2. Changes in baseline score

We next evaluate how well we are able to maintain the target attribute score when perturbing the latent vector. We use the change in the baseline classifier as a proxy to measure the target attribute score. We note that this measurement is flawed because the baseline classifier is known to perform worse on minority examples, however, we believe that this measurement still leads to some valuable insights. For each attribute, we measure the the absolute change in baseline score $|f_t(G(\mathbf{z}) - f_t(G(\mathbf{z}'))|$ over 5000 images, and compute averages based on what we expect the target and protected attribute values of $G(\mathbf{z}')$ to be. We plot this versus the fraction of images in the real world dataset that have these target and protected values (Figure 1). We find that there is a strong negative correlation. This could be because the target attribute is harder to maintain in this case, or because the baseline classifier has a tendency to misclassify minority samples.

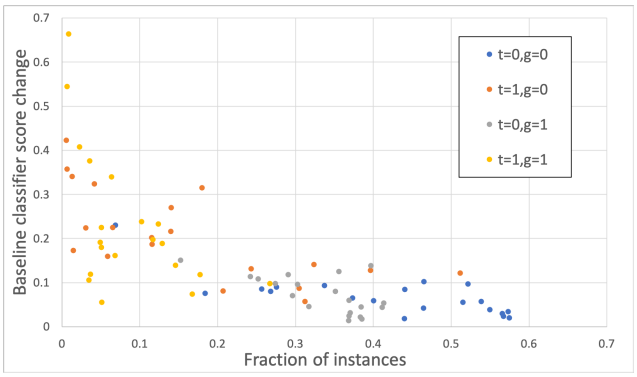

Figure 1: We plot average absolute change in the baseline classifier score versus the fraction of images in the dataset that have the corresponding ground truth labels. We separate them based on what the new ground truth values should be, for each attribute. We find that the score change is larger when creating an image with minority labels. This could be because we are unable to maintain the target attribute in this case or because the baseline classifier performs worse on minority images.

Another question that we were interested in was interactions between different attributes as we create balanced synthetic datasets for different attributes. We measured the change in baseline classifier score for different targets t' when trying to maintain target attribute t and found that some attributes changed drastically when creating a balanced dataset for any attribute (Table 2). For example, the attribute `Attractive` changed by a large amount irrespective of which target attribute we were trying to preserve. This suggests that some of these attributes are more sensitive to latent space manipulations.

3. Factors of influence

In this section, we discuss in more detail how some factors influence (or not) our method's effectiveness (Section 4.1 of the main paper).

3.1. Skew of attributes

For some attributes, the majority of the positive samples come from one gender expression. For example, `ArchedBrows` has a skew of 0.92 towards $g=-1$, that is, 92% of positive `ArchedBrows` samples have gender expression label $g=-1$. To understand the effect of data skew on our method's performance, we ran experiments with differently skewed data. From the 162,770 CelebA training set images, we created slightly smaller training sets where the attribute of interest (e.g. `HighCheeks`) has different values of skew. Specifically, we created three versions of training data each with skew 0.5, 0.7, 0.9, while keeping the total number of images fixed. We trained a GAN on each training set, created a synthetic de-biased dataset with our method, and trained an attribute classifier with the training set and 160,000 pairs of synthetic images. For comparison, we also

Attribute type	Attribute statistics			Hyperplane acc.		Hyperplane AP		Improvement over baseline			
Inconsistently labeled	Positive	Skew		$g{=}-1$	$g{=}-1$	$g{=}-1$	$g{=}1$	AP	DEO	BA	KL
BigLips	24.1%	0.73	$g{=}-1$	80.3	92.0	49.7	28.9	-0.35	-0.79	1.23	-0.03
BigNose	23.6%	0.75	$g{=}1$	91.7	74.5	51.1	82.4	-0.66	11.03	2.52	1.04
OvalFace	28.3%	0.68	$g{=}-1$	75.4	74.2	85.3	63.1	-1.82	7.53	3.33	0.77
PaleSkin	4.3%	0.76	$g{=}-1$	94.4	96.9	48.4	30.9	-1.90	4.26	0.31	0.26
StraightHair	20.9%	0.52	$g{=}-1$	87.7	69.8	25.0	58.8	-1.76	0.94	0.53	-0.08
WavyHair	31.9%	0.81	$g{=}-1$	73.0	92.1	79.4	23.5	-0.65	7.59	1.33	0.26
Gender-dependent	Positive	Skew		$g{=}-1$	$g{=}1$	$g{=}-1$	$g{=}1$	AP	DEO	BA	KL
ArchedBrows	26.6%	0.92	$g{=}-1$	72.3	92.1	82.6	25.5	-0.69	-3.31	-0.09	0.02
Attractive	51.4%	0.77	$g{=}-1$	88.4	81.0	97.9	81.9	-0.33	3.25	0.98	0.41
BushyBrows	14.4%	0.71	$g{=}1$	94.5	79.6	37.6	62.0	-1.20	8.49	1.14	0.25
PointyNose	27.6%	0.75	$g{=}-1$	73.6	82.9	84.4	59.9	-1.32	3.25	0.99	-0.40
RecedingHair	8.0%	0.62	$g{=}1$	94.5	88.3	41.8	57.7	-1.44	2.32	0.40	0.17
Young	77.9%	0.66	$g{=}-1$	96.2	84.1	99.7	95.3	-0.24	0.78	0.49	0.31
Gender-independent	Positive	Skew		$g{=}-1$	$g{=}1$	$g{=}-1$	$g{=}1$	AP	DEO	BA	KL
Bangs	15.2%	0.77	$g{=}-1$	90.3	94.9	81.5	58.9	-0.50	0.62	0.38	0.09
BlackHair	23.9%	0.52	$g{=}1$	89.3	83.2	78.9	79.2	-1.00	2.25	0.44	0.00
BlondHair	14.9%	0.94	$g{=}-1$	88.9	97.1	82.7	19.8	-0.77	1.04	0.23	-0.12
BrownHair	20.3%	0.69	$g{=}-1$	66.4	80.4	45.5	38.8	-0.51	-0.57	-0.01	0.01
Chubby	5.8%	0.88	$g{=}1$	99.1	89.9	7.6	33.8	-1.95	4.08	0.01	0.13
EyeBags	20.4%	0.71	$g{=}1$	90.7	74.4	64.1	74.4	-1.74	8.30	1.91	0.58
Glasses	6.5%	0.80	$g{=}1$	97.8	92.5	60.3	77.8	-0.24	-0.07	0.05	-0.27
GrayHair	4.2%	0.86	$g{=}1$	98.4	92.6	10.4	32.9	-2.60	7.02	0.32	0.54
HighCheeks	45.2%	0.72	$g{=}-1$	86.3	86.3	95.2	83.5	-0.33	-1.06	0.24	0.04
MouthOpen	48.2%	0.63	$g{=}-1$	88.6	87.0	96.4	93.1	-0.08	0.69	0.34	-0.03
NarrowEyes	11.6%	0.56	$g{=}-1$	93.8	92.1	29.6	26.4	-0.97	3.10	-0.53	0.12
Smiling	48.0%	0.65	$g{=}-1$	91.5	90.7	98.0	96.5	-0.09	1.01	0.67	0.03
Earrings	18.7%	0.97	$g{=}-1$	71.8	96.3	56.9	3.0	-0.63	8.18	0.64	1.40
WearingHat	4.9%	0.70	$g{=}1$	97.4	94.0	45.0	60.6	-0.95	2.67	0.14	-0.06
Average	24.1%	0.73		87.4	86.9	62.9	55.7	-0.95	3.18	0.69	0.21

Table 1: Attribute-level information. The columns are (from left to right) target attribute name, percentage of positive samples, skew, hyperplane accuracy, hyperplane AP, and our model's improvement over the baseline model on the four evaluation metrics.

Attribute	Change	Attribute	Change
ArchedBrows	**0.314**	Glasses	0.109
Attractive	**0.336**	GrayHair	0.056
Bangs	0.120	HighCheeks	**0.233**
BlackHair	0.153	MouthOpen	0.187
BlondHair	0.180	NarrowEyes	0.066
BrownHair	0.158	PointyNose	0.152
BushyBrows	0.136	RecedingHair	0.069
Chubby	0.067	Smiling	0.176
Earrings	0.176	WearingHat	0.065
Eyebags	**0.212**	Young	**0.268**

Table 2: We report the average classifier score change in an attribute when trying to create balanced datasets for other attributes. Classifier scores are between 0 and 1, and changes above 0.2 are bolded. We find that some attributes (e.g. Attractive, Young) change by a lot, whereas others (e.g. GrayHair, WearingHat) do not change much.

trained baseline models on just the differently skewed training sets. The classifiers were evaluated on the CelebA validation set. Table 3 summarizes the results. Compared to the baseline, our model has lower AP as expected, better DEO for skew 0.5 and 0.7, worse DNAP, and better or on par BA. Overall, classifiers trained on more imbalanced data with higher skew perform worse on all metrics.

Skew	AP ↑		DEO ↓	
	Base	Ours	Base	Ours
0.5	**95.1 ± 0.3**	93.6 ± 0.4	7.0 ± 1.7	**6.6 ± 1.8**
0.7	**94.8 ± 0.3**	94.1 ± 0.3	19.6 ± 1.9	**19.4 ± 1.9**
0.9	**94.1 ± 1.7**	93.1 ± 0.4	**31.3 ± 2.0**	32.9 ± 1.9
Skew	BA ↓		KL ↓	
	Base	Ours	Base	Ours
0.5	-1.9 ± 0.5	**-3.0 ± 0.5**	0.4 ± 0.1	**0.3 ± 0.1**
0.7	**3.4 ± 0.5**	**3.4 ± 0.5**	**0.9 ± 0.1**	**0.9 ± 0.1**
0.9	7.1 ± 0.5	**7.0 ± 0.5**	**1.7 ± 0.1**	1.9 ± 0.1

Table 3: Comparison of HighCheeks attribute classifiers trained on differently skewed data.

3.2. Discriminability of attributes

Nam et al. [2] recently observed that correlations among attributes affect a classifier only if the protected attribute is 'easier' to learn than the target attribute. Inspired by their

	ArchedBrows	Attractive	Bangs	BlackHair	BlondHair	BrownHair	BushyBrows	Chubby	Earrings	EyeBags	Glasses	GrayHair	HighCheeks	MouthOpen	NarrowEyes	PointyNose	RecedingHair	Smiling	WearingHat	Young
Gender	y	y	y	y	y	y	y	y	y	y	n	y	y	y	y	y	y	y	n	y
Glasses	y	y	y	y	y	y	y	y	y	y	–	y	y	y	y	y	y	y	y	y
Young	n	n	n	n	n	y	n	n	y	n	n	n	n	n	n	y	n	n	y	–

Table 4: Discriminability of attributes. We compare attributes on the row to those in the columns. y indicates that the attribute in the row is easier to learn than that in the column and n indicates the opposite. We find that gender expression is one of the easiest attributes to learn, while Young is relatively hard.

observation, we design an experiment where we put a pair of CelebA attributes in competition to assess their relative discriminability. We create a fully skewed dataset in which half of the images have both attributes and the other half have neither. With this dataset, we train a classifier to predict if an image has both attributes or neither. At test time, we evaluate the classifier on a perfectly balanced subset of the CelebA validation set (where each of the four possible hat-glasses combinations occupies a quarter of the dataset), and compute AP for each attribute. If one attribute has a higher AP than the other, it suggests that this attribute is 'easier' to learn than the other. We repeat this experiment with a second dataset skewed in a different way (i.e. half of the images have one attribute but not the other).

The results for gender-dependent and gender-independent attributes are in Table 4. We report that an attribute is 'easier' to learn than the other if it has a higher AP for both created datasets. We find that gender expression is one of the easiest attributes to learn, which may be why gender bias is prevalent in many models. On the other hand, Young is relatively hard for a model to learn, so its correlation with other attributes may not be as influential. We find that gender expression is one of the easiest attributes to learn (with gender expression having a higher AP than every attribute we tested except WearingHat and Glasses), which may be why gender bias is prevalent in many models. On the other hand, Young is relatively hard for a model to learn (Young is harder to learn than all but 4 other attributes), so its correlation with other attributes may not be as influential.

4. Ablation studies

In this section, we describe in more detail the ablation studies we have conducted to investigate how improved hyperplanes and use of different labels for synthetic images impact (or not) our method's performance (Section 4.2 of the main paper).

We first investigate if hyperplanes estimated with better balanced samples improve the performance of downstream attribute classifiers. We test this hypothesis by training models using hyperplanes that are estimated with different fractions of positive or negative samples.

For the attribute HighCheeks, we estimate hyperplanes with different fractions of positive and negative samples, while keeping the total number of samples constant at 12,000 and the number of positive samples same for each gender expression. We then train attribute classifiers with the CelebA training set and synthetic pair images augmented with these different hyperplanes. In Table 5, we report results evaluated on the CelebA validation set. We find that although the fairness metrics deteriorate as the target attribute hyperplanes were estimated with less balanced samples, this rate is relatively slow, and the downstream classifier still performs reasonably well.

Fraction	AP ↑	DEO ↓	BA ↓	KL ↓
50.0%	95.1 ± 0.3	13.2 ± 1.7	0.5 ± 0.5	0.7 ± 0.1
12.5%	95.1 ± 0.3	14.0 ± 1.7	0.8 ± 0.5	$\mathbf{0.6 \pm 0.1}$
6.3%	95.1 ± 0.3	15.1 ± 1.8	1.3 ± 0.5	0.8 ± 0.2
3.1%	95.1 ± 0.3	14.2 ± 1.7	1.0 ± 0.5	0.7 ± 0.1
1.6%	95.1 ± 0.3	$\mathbf{12.9 \pm 1.8}$	$\mathbf{0.3 \pm 0.5}$	0.7 ± 0.1

Table 5: The amount of underrepresentation in samples used for hyperplane estimation doesn't appear to affect the performancee of the downstream classsification model much.

Next, we tried training models with synthetic images with the same hallucinated target labels, i.e. using only $G(\mathbf{z})$ and $G(\mathbf{z}')$ such that $f_t(G(\mathbf{z}))=f_t(G(\mathbf{z}'))$, and labeling synthetic images with $h_t(\mathbf{z})$ in place of $f_t(G(\mathbf{z}))$. Table 6 contains all results. We report average results over all gender-dependent and gender-independent attributes. We find that both these ablations are comparable to ours, with in a slight loss in AP (79.8 and 82.1 versus 82.6), and worse fairness metrics in general (average DEO is 18.1 and 17.4 vs 16.1, BA is 0.9 and 0.7 vs 0.5).

5. Number of required labeled images

Choi et al. [1] use a method that is unsupervised. Assuming access to a small unbiased dataset, as well as a large (possibly biased) dataset, they estimate the bias in the larger dataset, and learn a generative model that generates unbiased data at test time. Using these generated images, as well as real images, they train a downstream classifier for the attribute Attractive, and achieve an accuracy of 75%. Since most of the protected attributes that we care about are sensitive (for example gender or race), not requiring protected attribute labels prevents perpetuation of harmful stereotypes. In order to understand how much our model depends on the protected attribute labels, we investigate where our model depends on the protected attributes labels. We use protected attribute labels only to compute the linear separator in the latent space ($\mathbf{w_g}$ and b_g from section 1 in this document). We now train classifiers for gender expression, using different numbers of labeled images, and use these classifiers to train target attribute classifiers for 4 different attributes (EyeBags, BrownHair, GrayHair and HighCheeks). Most of the fairness metrics improve slightly when using more labeled examples (DEO improves from 11.1 when using just 10 samples to 9.6 when using all

	AP ↑	DEO ↓	BA ↓	KL ↓
$f_t(G(\mathbf{z})) = f_t(G(\mathbf{z}'))$	79.8 ± 1.6	17.4 ± 4.5	0.9 ± 0.4	$\mathbf{1.0} \pm \mathbf{0.3}$
Labels computed using h_t	82.1 ± 1.5	18.1 ± 4.2	0.7 ± 0.4	1.4 ± 0.8
Ours	$\mathbf{82.6} \pm \mathbf{1.5}$	$\mathbf{16.1} \pm \mathbf{4.2}$	$\mathbf{0.5} \pm \mathbf{0.4}$	1.3 ± 0.7

Table 6: Mean performances over all gender-dependent and gender-independent attributes on the validation set when using different methods to pick and label synthetic images. We find that most performances are comparable, with our method having a slightly higher AP, and slightly better DEO and KL.

Metric	Num. of samples used to compute f_g				
	10	100	1000	10000	162,770
AP ↑	78.8 ± 1.5	78.8 ± 1.5	78.8 ± 1.5	$\mathbf{78.9} \pm \mathbf{1.6}$	78.7 ± 1.6
DEO ↓	11.1 ± 3.4	11.3 ± 3.0	10.5 ± 3.7	10.8 ± 3.7	$\mathbf{9.6} \pm \mathbf{3.1}$
BA ↓	0.6 ± 0.5	1.0 ± 0.5	0.5 ± 0.5	0.7 ± 0.5	$\mathbf{0.4} \pm \mathbf{0.5}$
KL ↓	0.6 ± 0.2	0.8 ± 0.3	0.7 ± 0.3	0.7 ± 0.3	$\mathbf{0.5} \pm \mathbf{0.6}$

Table 7: Average over 4 attributes when using different numbers of labeled examples to compute gender expression. Results are reported on the validation set. We find that while the fairness metrics improve slightly by using more labelled examples, this is gradual, and within the error bars, in all cases.

162k samples in the CelebA training set, BA improves from 0.6 to 0.4, and KL improves from 0.6 to 0.5), however, these are all gradual, and within the error bars. Full results are in Table 7.

References

[1] Kristy Choi, Aditya Grover, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020. 4

[2] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. In *Advances in Neural Information Processing Systems*, 2020. 3