Abstract

Fairness in visual recognition is becoming a prominent and critical topic of discussion as recognition systems are deployed at scale in the real world. Models trained from data in which target labels are correlated with protected attributes (e.g., gender, race) are known to learn and exploit those correlations. In this work, we introduce a method for training accurate target classifiers while mitigating biases that stem from these correlations. We use GANs to generate realistic-looking images, and perturb these images in the underlying latent space to generate training data that is balanced for each protected attribute. We augment the original dataset with this generated data, and empirically demonstrate that target classifiers trained on the augmented dataset exhibit a number of both quantitative and qualitative benefits. We conduct a thorough evaluation across multiple target labels and protected attributes in the CelebA dataset, and provide an in-depth analysis and comparison to existing literature in the space. Code can be found at https://github.com/princetonvisualai/gan-debiasing.

1. Introduction

Large-scale supervised learning has been the driving force behind advances in visual recognition. Recently, however, there has been a growing number of concerns about the disparate impact of these visual recognition systems. Face recognition systems trained from datasets with an under-representation of certain racial groups have exhibited lower accuracy for those groups [9]. Activity recognition models trained on datasets with high correlations between the activity and the gender expression of the depicted person have over-amplified those correlations [46]. Computer vision systems are statistical models that are trained to maximize accuracy on the majority of examples, and they do so by exploiting the most discriminative cues in a dataset, potentially learning spurious correlations. In this work, we introduce a new framework for training computer vision models that aims to mitigate such concerns, illustrated in Figure 1.

One proposed path for building ‘fairer’ computer vision systems is through a ‘fairer’ data collection process. Works such as [9, 43] propose techniques for better sampling data to more accurately represent all people. Creating a perfectly balanced dataset, however, is infeasible in many cases. With the advances in Generative Adversarial Networks (GANs) [17], several works propose using generated data to augment real-world datasets [12, 35, 42]. These methods have been growing in computational and algorithmic complexity (e.g., [35, 42] adding multiple loss functions to GAN training), necessitating access to a sufficient number of inter-sectional real-world samples. In contrast, we demonstrate a simple and novel data augmentation technique that uses a single GAN trained on a biased real-world dataset.

Illustrative example: Consider our example from Figure 1. Our goal is to train a visual recognition model that recognizes the presence of an attribute, such as wearing a hat. Suppose in the real world wearing a hat is correlated with wearing glasses—for example, because people often wear both hats and sunglasses outside and take them off inside. This correlation may be reflected in the training data, and a classifier trained to recognize a hat may rely on the presence of glasses. Consequently, the classifier may fail to recognize a hat in the absence of glasses, and vice versa.

We propose using a GAN to generate more images with hats but not glasses and images with glasses but not hats, such that WearingHat is de-correlated from Glasses in the training data, by making perturbations in the latent space. Building on work by Denton et al. [14], which demonstrates...
a method for learning interpretable image manipulation directions, we propose an improved latent vector perturbation method that allows us to preserve the WearingHat attribute while changing the Glasses attribute (Figure 2).

**Protected attributes:** Our goal is to examine and mitigate biases of sensitive attributes such as gender expression, race, or age in visual classifiers. However, visual manipulations or explicit classifications along these dimensions have the potential to perpetuate harmful stereotypes (see [23]). Hence in our illustrations, we use Glasses as the protected attribute, as it has a clear visual signal. In the quantitative experimental results, we report our findings on the more sensitive protected attributes of gender expression and age.

**Contributions:** We propose a method for perturbing vectors in the GAN latent space that successfully de-correlates target and protected attributes and allows for generating a de-biased dataset, which we use to augment the real-world dataset. Attribute classifiers trained with the augmented dataset achieve quantitative improvements in several fairness metrics over both baselines and prior work [35, 36, 41], while maintaining comparable average precision. Furthermore, we analyze the CelebA [28] attributes with respect to label characteristics¹, discriminability, and skew, and discuss how these factors influence our method’s performance. We also evaluate our design choices with ablation studies and the results demonstrate the effectiveness of our augmentation method.²

### 2. Related Work

**De-biasing models:** The effect of gender and racial bias on AI models has been well documented [8, 9, 22, 40, 41]. Models trained on biased data sometimes even amplify the existing biases [46]. Tools such as AI Fairness 360 [6] and REVISe [38] surface such biases in large-scale datasets and enable preemptive analysis. In parallel, various work proposes methods for mitigating unwanted dataset biases from influencing the model. Oversampling techniques [7, 15] duplicate minority samples in imbalanced data to give them higher weight in training. Some work propose to mitigate bias through adversarial learning [40, 45] or through learning separate classifiers for each protected attribute [33, 41]. Other work improve fairness by introducing constraints [29] or regularization terms [3] during training. Contrary to these algorithmic approaches, our work aims to mitigate biases by training the model with a generated de-biased dataset.

**Generating and perturbing images using GANs:** Generative Adversarial Network (GAN) [17] is a popular class of generative models composed of a generator and a discriminator trained in an adversarial setting. Over the past few years, a number of works [18, 24, 25, 27, 34] improved GANs to generate more realistic images with better stability. Shen et al. [37] show that the latent space of GANs have semantic meaning and demonstrate facial attributes editing through latent space manipulation. Denton et al. [14] propose a method to evaluate how sensitive a trained classifier is to such image manipulations, and find several attributes that affect a smiling classifier trained on CelebA. Balakrishnan et al. [4] use GANs to generate synthetic images that differ along specific attributes while preserving other attributes, and use them to measure algorithmic bias of face analysis algorithms. Unlike [4, 14] who use the GAN-generated images to evaluate models, our work uses these generated images to train better attribute classification models.

**Using GANs to augment datasets:** Several works use GANs to augment datasets for low-shot [20] and long-tail [48] recognition tasks, whereas our work focuses specifically on de-biasing classifiers affected by dataset bias. More related to our work are [12, 35, 36] which leverage GANs to generate less biased data. Choi et al. [12], given access to a small, unlabeled, and unbiased dataset, detect bias in a large and potentially biased dataset, and learn a generator that generates unbiased data at test time. Sattigeri et al. [35] train a GAN with a modified loss function to achieve demographic parity or equality of odds in the generated dataset. Sharmanska et al. [36] use an image-to-image translation GAN to generate more minority samples and create a balanced dataset. While [12, 35, 36] require training a new GAN for each bias they want to correct, our method uses a single GAN trained on a biased dataset to augment all attributes.

### 3. Method

We study a class of problems where a protected attribute is correlated with a target label in the data $X$, influencing target label prediction. Let $t$ be the target label (e.g., WearingHat in the running example from Figure 1) and $g$ be the protected attribute (e.g., gender expression or Glasses from our running example) with $t, g \in \{-1, 1\}$.

To mitigate the effect of unwanted dataset bias, we aim to generate a balanced set of synthetic images $X_{syn}$ where the protected attribute and target label are de-correlated.

---

¹We observe several discrepancies in the CelebA [28] attribute labels and categorize the attributes into three categories: inconsistently labeled, gender-dependent, and gender-independent.

²Code for all our experiments can be found at https://github.com/princetonvisualai/gan-debiasing.
we now want to generate synthetic data \( X' \) with the property that for \( x \in X' \):

\[ P[f_t(x) = 1|f_g(x) = 1] = P[f_t(x) = 1] , \]

such that attributes \( t \) and \( g \) are de-correlated.

**De-biased dataset creation:** To create \( X'_\text{syn} \), we use a GAN trained on real images \( X' \) whose generator \( G \) generates a synthetic image \( x \) from a random latent vector \( z \in \mathcal{Z} \). We can assign semantic attribute labels to these images using the learned functions \( f_t(x) \) and \( f_g(x) \). However, as the GAN inherits correlations from its training data, a random sampling of \( z \) will produce an \( X'_\text{syn} \) with similar correlations and biases as \( X' \). Hence, we propose a latent vector perturbation method that allows us to generate a de-biased \( X'_\text{syn} \).

We sample a random set of latent vectors \( Z' \subset \mathcal{Z} \) (inheriting the biases) and train classifiers \( h_t, h_g : \mathcal{Z} \rightarrow [-1,1] \) in the latent space that approximate \( f_t \circ G \) and \( f_g \circ G \), respectively. That is, we train classifiers \( h_t \) with input \( z \) and output \( f_t(G(z)) \), and \( h_g \) with input \( z \) and output \( f_g(G(z)) \).

Given a vector \( z \), we generate a complementary vector \( z' \) with the same (predicted) target label but the opposite (predicted) protected attribute label, or

\[ h_t(z') = h_t(z), \quad h_g(z') = -h_g(z). \]

We note that this data generation method is agnostic to the type of classifier used to compute \( h \).

In our work, we assume that the latent spaces is approximately linearly separable in the semantic attributes, as observed and empirically validated by Denton et al. [14]. In this case, \( h_t \) and \( h_g \) can be represented as linear models (hyperplanes) \( w_t \) and \( w_g \) with intercepts \( b_t \) and \( b_g \) for the target and protected attributes respectively. We can derive a closed-form solution for \( z' \) as

\[ z' = z - 2 \left( \frac{w_g^T z + b_g}{1 - (w_g^T w_t)^2} \right) \left( w_g - (w_g^T w_t)w_t \right). \]

This latent vector perturbation method is illustrated in Figure 3 (Top left). A similar idea of hyperplane projection was presented in Zhang et al. [45], although for a different goal of adversarial training. The sampling process results in a complementary image pair:

- \( x = G(z) \) with target label \( f_t(G(z)) \) and protected attribute label \( f_g(G(z)) \)

- \( x' = G(z') \) with target label \( f_t(G(z)) \) and protected attribute label \( -f_g(G(z)) \),

creating de-biased data \( X'_\text{syn} \). We train our target attribute classifier with \( x \) and \( X'_\text{syn} \), as shown in Figure 3.

We label the generated images \( x \) and \( x' \) both with \( f_t(x) \) because it allows us to capture the target attribute labels better than using \( f_t(x) \) and \( f_t(x') \). It is likely that the accuracy of \( f_t \) is higher for the overrepresented group, and \( x \) will more often belong to the overrepresented group and \( x' \) to the underrepresented group. However, other design choices are possible in our approach—for example, we could use \( h_t(z) \) and \( h_t(z') \) instead (after thresholding appropriately) or only use \( z \) for which \( f_t(x) = f_t(x') \). We compare these different design choices experimentally in Section 4.2.

**Advantages:** Our data augmentation method has several attractive properties:

1. We use a single GAN trained on the biased real-world dataset to augment multiple target labels and protected attributes. This is in contrast to prior works like [35, 12] that require training a GAN for every pair of target and protected attributes.

2. By augmenting samples \( z \) generated from (approximately) the original data distribution the GAN was trained on and maintaining their target attribute scores, our method preserves the intra-class variation of the images.

3. The samples \( z \) and \( z' \) are generated to simulate the independence goal of Equation 1. By construction, \( z' \) maintains \( z \)'s target label \( f_t(G(z)) \) and takes on the opposite protected attribute label \( -f_g(G(z)) \).

4. Our method generalizes to multiple protected attributes. We demonstrate how our method can simultaneously augment two protected attributes in Section 4.3 when we compare our work to Sharmanska et al. [36].

---

\( ^3 \) \( f_t \) is equivalent to the baseline classifier in Section 4.1.

\( ^4 \) Derivations are in the supplementary material. \( \|w_t\| = \|w_g\| = 1. \)
4. Experiments

In this section, we study the effectiveness of our data augmentation method on training fairer attribute classifiers. We first describe our experiment setup and compare our results to those of a baseline classifier. We then discuss how different factors influence our method’s performance, and finally compare our work to several prior works.

Dataset and attributes categorization: Given the task of training attribute classifiers that are not dependent on gender expression, we require a dataset that has target labels, as well as gender expression labels. CelebA [28] is a dataset with 2,022,599 images of celebrity faces, each with 40 binary attributes labels. We assume the Male attribute corresponds to gender expression. Among the other 39 attributes, we use 26 of them that have between 1% and 99% fraction of positive images for each gender expression. However, we noticed several discrepancies among the attribute labels, and decided to categorize the attributes into three categories: inconsistently labeled, gender-dependent, and gender-independent.

We categorized attributes as inconsistently labeled when we visually examined sets of examples and found that we often disagreed with the labeling and could not distinguish between positive and negative examples. This category includes StraightHair shown in Figure 4, as well as BigLips, BigNose, OvalFace, PaleSkin, and WavyHair. While we report results on these attributes for completeness in Section 4.1, classifiers trained on these attributes may behave erratically.

Of the remaining attributes with more consistent labeling, we found that some attribute labels are gender-dependent. That is, images are labeled to have (or not have) these attributes based on the perceived gender. For example in...
measure the accuracy of the classifiers. AP is a threshold-invariant accuracy metric that summarizes the precision and recall curve. We use this metric to ensure that our models learn a reasonable classification rule. AP, however, does not capture a classifier’s behavior on different protected classes, and in fact, we expect to see a slight dip in overall AP when our model improves on some of the fairness metrics.

Multiple metrics have been proposed to measure fairness of a model [19, 44, 46, 10, 11] and each of these measures a different notion of fairness. In our work, we use three metrics for comprehensive understanding. First, we measure the difference in equality of opportunity (DEO), i.e. the absolute difference between the false negative rates for both gender expression, as in Lokhande et al. [29].

As our second fairness metric, we use the bias amplification (BA) metric proposed by Wang and Russakovsky [39]. Intuitively, BA measures how much more often a target attribute is predicted with a protected attribute than the ground truth value. Let $P_{t|g}$ be the fraction of images with protected attribute $g$ that have target attribute $t$, $P_{\bar{t}|g}$ be the fraction of images with protected attribute $g$ that are predicted to have target attribute $t$, $P_t$, $g$ be the fraction of images with target $t$ and protected attribute $g$, and $P_t$ and $P_g$ be the fraction of images with attribute $t$ and $g$ respectively. For each pair of target and protected attribute values, we add $(P_{t|g} - P_{\bar{t}|g})$ if $P_t, g > P_t, g$ and $-P_{t|g} - P_{\bar{t}|g}$ otherwise. A negative value implies that bias now exists in a different direction than in the training data.

Both DEO and BA fluctuate based on the chosen classification threshold. Hence, as our final fairness metric, we use a threshold-invariant metric that measures the divergence between score distributions (KL) [11] defined as follows: Suppose $s_{g,t}$ represents a smoothed histogram of classifier scores of a certain protected attribute label and a target label, appropriately normalized as a probability distribution of the scores. For each target attribute label $t$, we measure $KL[s_{g=1,t} \| s_{g=1,t}] + KL[s_{g=0,t} \| s_{g=0,t}]$. That is, we measure the divergence of $g=1$ and $g=0$ score distributions, separately for positive and negative attribute samples. This is a stricter notion of equalized odds [19].

4.1. Comparison with the baseline

To start, we compare our model (i.e. target classifiers trained using both the balanced synthetic datasets $\mathcal{X}_{\text{syn}}$ and the real dataset $\mathcal{X}$) with a baseline model trained using just $\mathcal{X}$. In Table 1, we show results on the four metrics, averaged for each of the three attribute categories. As expected, our model performs better on all three fairness metrics, DEO, BA and KL, while maintaining comparable AP. For gender-independent attributes, AP drops from 83.9 to 83.0, while DEO improves from 16.7 to 13.9. BA improves from 0.3 to 0.0 and KL improves from 1.1 to 0.9. For gender-dependent attributes, the fairness metrics improve over the baseline, but the improvements are smaller compared to those of gender-independent attributes. Later in Section 5, we demonstrate an extension of our augmentation method with an improved performance on the gender-dependent attributes.

Additionally, we conduct score change evaluations suggested by Denton et al. [14] and measure the change in target attribute score as we perturb the protected attribute in images. Specifically, we measure the classifier score difference between $G(z)$ and $G(z')$. This evaluation helps understand how the protected attribute influences a trained classifier’s output. We find that the model trained with our augmentation method consistently has a smaller change in score than the baseline: 0.09 vs. 0.12 for inconsistently labeled, 0.07 vs. 0.11 for gender-dependent, and 0.06 vs. 0.09 for gender-independent attributes. We also observe that the baseline score changes are higher when we try to construct underrepresented samples. Consider the attribute ArchedBrows where only 2.3% of the training set images are labeled to have ArchedBrows, and appear masculine. When we construct a $z'$ with this target and protected value, the baseline classifier’s score changes by 0.41. On the other hand, when we try to construct an image that is without ArchedBrows and appears feminine, which comprises 33.7% of the training set, the baseline classifier score only changes by 0.094. This could be due to the errors that the baseline classifier makes on underrepresented images during synthetic image labeling, or could imply that underrepresented attributes are harder to maintain during image manipulations.

We next examine several factors that could influence our method, including how easy the protected attribute is to learn compared to the target attribute and how data skew affects our method. We discuss the former here and provide more information about the latter in the supplementary material.

**Discriminability of attributes:** Nam et al. [30] recently ob-
Table 2: Improvement over baseline for different fairness metrics when using different protected attributes. Next to the protected attribute are numbers of attributes that are ‘easier’ and ‘harder’ to learn, compared to the protected attribute. Columns ‘Easy’ (‘Hard’) show the averages of all non-inconsistent target attributes that are easier (harder) for a classifier to learn. We note that our method works better when the target attribute is ‘harder’ to learn.

served that correlations among attributes affect a classifier only if the protected attribute is ‘easier’ to learn than the target attribute. Inspired by their observation, we conduct a two-step experiment to understand how the relative discriminability of attributes affects our method’s effectiveness.

First, we put a pair of CelebA attributes in competition to assess their relative discriminability. Experiment details are in the supplementary material. We find that gender expression is one of the easiest attributes to learn (Gender is easier than all but Glasses and WearingHat), 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.

Next, to understand how the relative discriminability of attributes affects our method’s performance, we train target attribute classifiers for gender-dependent and gender-independent attributes, using Young and Glasses as protected attributes. In Table 2, we report our method’s improvement over baseline in the three fairness metrics. For each protected attribute, we report the average improvement separately for ‘easier’ and ‘harder’ target attributes. While training with our augmentation method generally outperforms the baseline on the three fairness metrics, as expected, the improvement is greater for target attributes that are harder to learn than the protected attribute, for example, for Young, the improvement in DEO over baseline is -0.2 for easy target attributes, and 2.1 for hard target attributes.

Skew of the dataset: The skew of a target attribute $t$ is measured following the literature [41] as $\frac{\max(P_{1:t}, P_t)}{P_{1:t} + P_t}$, where $P_{1:t}$ is the number of images with $t=1$ and protected attribute label $g=-1$, and $P_t$ is the number of images with $t=1$ and protected attribute label $g=1$. We find that our augmentation method is most effective on attributes with low to moderate skew. Full details are in the supplementary material.

4.2. Ablation studies

We now examine the design choices made in our method. **Removal of z’ samples:** First, we evaluate the effect of $G(z')$ on the classifier. We train a classifier with just $G(z)$ and the real dataset $X$, and compare its performance against the performance of our model, trained with $G(z)$, $G(z')$, and $X$ on the gender-dependent and gender-independent attributes. While the new classifier’s AP is higher than that of our model (82.9 vs. 82.6), all fairness metrics are worse: DEO is higher (19.7 vs. 16.1), BA is higher (1.1 vs. 0.5) and KL is higher (1.6 vs 1.3). All numbers were calculated on the validation set. In fact, it performs worse on the fairness metrics than the baseline model trained on $X$. This result suggests that simply synthesizing more images with a GAN and adding them to the training data does not improve the model but rather hurts performance. Possible reasons include the image and label noise of $G(z)$ and the skew of $G(z)$ being worse than the original data the GAN was trained on. The fairness metrics improve only when we add $G(z')$, and make the training data more balanced.

**Choice of z’:** Next, we evaluate our choice of $z'$ through examining a number of alternative perturbation choices visualized in Figure 6. We train classifiers on just the generated data for gender-dependent and gender-independent attributes and compare the overall AP on the validation set. As expected, training with $z'$ (our choice) has the highest AP.

**Filtering z’s and using different labels for synthetic images:** Since we hallucinate labels for the synthetic images, some of these labels may be incorrect and harm our classifier. We try three different ways of addressing this issue: First, we try learning hyperplanes with different fractions of positive and negative samples. We find that while this improves the hyperplane accuracy, the downstream classifiers trained with samples generated using different hyperplanes have similar performances. For the second and third methods, we use the original hyperplanes learned in our method, but vary the vectors/labelling used. We remove points that are incorrectly classified by the baseline classifier after perturbing the latent vector from $z$ to $z'$, i.e., we remove all points wherein $f_t(G(z)) \neq f_t(G(z'))$, and use the remaining synthetic images and the real dataset to train the classifiers. Third, we label the synthetic images $G(z)$ and $G(z')$ with $h_t(z)$, and use these labels to train the classifiers. We compare their performance to our method on the validation set. We find that these two methods result in a slight drop in AP (79.8 when using $h_t$ scores, 82.1 when removing incorrectly classified points, and 82.6 for our method), as well as a small drop in the fairness metrics (the average DEO is 18.1 when using $h_t$ scores, 17.4 when removing incorrectly classified points).

---

<table>
<thead>
<tr>
<th>Protected Attribute</th>
<th>Improvement over baseline ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DEO</td>
</tr>
<tr>
<td></td>
<td>Easy</td>
</tr>
<tr>
<td>Glasses (0,19)</td>
<td>-0.2</td>
</tr>
<tr>
<td>Gender (2,17)</td>
<td>0.8</td>
</tr>
<tr>
<td>Young (15,4)</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Figure 6: Comparison of different perturbation choices. We train attribute classifiers using only synthetic images generated from the perturbations, and measure the mean AP over all target attributes on the validation set. The classifier trained with $z'$ (our choice) has the highest AP.
points, and 16.1 for our method), suggesting that our current labeling of the synthetic images works well. Full results are in the supplementary material.

4.3. Comparison with prior work

In this section, we compare our method to few recent works [35, 36, 41]. One of the current challenges in the space of AI fairness is the lack of standardized benchmarks and metrics. While some of this stems from the complexity of the problem at hand (where it is difficult and even counter-productive to use a single fairness definition), in the computer vision community, we believe that more effort should be made to provide thorough comparison between methods. Each work we consider here uses slightly different evaluation protocols and benchmarks. We made comparisons to the best of our ability, and hope that our work helps enable more standardization and empirical comparisons.

**Fairness GAN:** Sattigeri et al. [35] use GANs to create datasets that achieve either demographic parity (Dem. Par.) or equality of opportunity (Eq. Opp.). They train classifiers for the Attractive attribute on just the generated data, using gender expression as the protected attribute. We train classifiers with our pair-augmented synthetic data to mimic the conditions of Fairness GAN, and evaluate both on the CelebA test data. Comparison results are in Table 3. Our model performs better on most metrics, even though we use a single GAN to augment all attributes.

**Contrastive examples generated by image-to-image translation GANs:** Sharmanska et al. [36] propose a different method for balancing a biased dataset using StarGAN [13], a class of image-to-image translation GANs. They use two protected attributes, age and gender expression, and create a balanced dataset by creating contrastive examples, i.e. images of different ages and gender, for each image in the training set. They train a Smiling classifier with the augmented dataset, and propose making a prediction at test time only when the classifier makes the same prediction on the image and their contrastive examples. We extend our method to incorporate multiple protected attributes, and use gradient descent to find three points \( \{ z_i \}_{i \in \{1,2,3\}} \) in the latent space that preserve the target attribute score and flip either the gender expression score, the age score, or both. This process gives us three synthetic images per training image, with which we train a Smiling classifier.

### Table 3: Comparison of the Attractive classifier trained using synthetic data from Fairness GAN [35] and the classifier trained using our pair-augmented synthetic data. The latter (ours) outperforms on most metrics.

<table>
<thead>
<tr>
<th></th>
<th>Fairness GAN [35]</th>
<th>Ours (Synthetic only)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender exp. g</td>
<td>( g = 1 )</td>
<td>( g = 1 )</td>
</tr>
<tr>
<td>FPR ↓</td>
<td>0.52</td>
<td>0.22</td>
</tr>
<tr>
<td>FNR ↓</td>
<td>0.18</td>
<td>0.06</td>
</tr>
<tr>
<td>Error ↓</td>
<td>0.30</td>
<td>0.21</td>
</tr>
<tr>
<td>Error Rate ↓</td>
<td>0.22</td>
<td>0.20</td>
</tr>
</tbody>
</table>

### Table 4: Comparison of our method with domain independent training [31]. Numbers reported are the mean over all gender-dependent and gender-independent attributes on the test set. We note that we perform better than domain-independent training for attributes with low to moderate skew.

<table>
<thead>
<tr>
<th>Method</th>
<th>AP ↑</th>
<th>DEO ↓</th>
<th>BA ↓</th>
<th>KL ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low/Mod.</td>
<td>83.0 ± 1.0</td>
<td>7.0 ± 3.1</td>
<td>0.1 ± 0.3</td>
<td>0.8 ± 0.7</td>
</tr>
<tr>
<td>High/Mod.</td>
<td>80.7 ± 1.6</td>
<td>14.9 ± 5.6</td>
<td>0.4 ± 0.5</td>
<td>0.8 ± 1.0</td>
</tr>
<tr>
<td>Weighted</td>
<td>79.6 ± 1.6</td>
<td>5.7 ± 4.2</td>
<td>2.8 ± 0.5</td>
<td>0.5 ± 0.4</td>
</tr>
<tr>
<td>Adversarial</td>
<td>81.3 ± 1.6</td>
<td>23.9 ± 4.4</td>
<td>1.5 ± 0.5</td>
<td>0.6 ± 0.5</td>
</tr>
<tr>
<td>Ours</td>
<td>81.5 ± 1.5</td>
<td>16.7 ± 4.7</td>
<td>0.5 ± 0.5</td>
<td>1.0 ± 0.5</td>
</tr>
</tbody>
</table>

### Table 5: Comparison of our method with weighted and adversarial training from [41]. Numbers reported are the mean over all gender-dependent and gender-independent attributes on the test set. We find that our model performs better in terms of the mean difference in FNR (0.34 versus their 0.54) and FPR (0.23 compared to their 0.46).

<table>
<thead>
<tr>
<th>Method</th>
<th>AP ↑</th>
<th>DEO ↓</th>
<th>BA ↓</th>
<th>KL ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weighted</td>
<td>79.6 ± 1.6</td>
<td>5.7 ± 4.2</td>
<td>2.8 ± 0.5</td>
<td>0.5 ± 0.4</td>
</tr>
<tr>
<td>Adversarial</td>
<td>81.3 ± 1.6</td>
<td>23.9 ± 4.4</td>
<td>1.5 ± 0.5</td>
<td>0.6 ± 0.5</td>
</tr>
<tr>
<td>Ours</td>
<td>81.5 ± 1.5</td>
<td>16.7 ± 4.7</td>
<td>0.5 ± 0.5</td>
<td>1.0 ± 0.5</td>
</tr>
</tbody>
</table>

**Effective training strategies for bias mitigation:** Wang et al. [41] quantitatively compare different techniques for bias mitigation, including weighted training [7, 15], adversarial training with losses inspired by [2, 45], and their proposed domain discriminative and domain independent training. We compare our method to their best performing domain independent training method where they learn separate classifiers for each protected attribute class and combine them to leverage any shared information. We report results for all gender-dependent and gender-independent attributes in Table 4. We find that our method performs better for attributes with low to moderate skew (<0.7)—DEO is 6.0 compared to 7.0, KL is 0.3 compared to 0.8—whereas domain independent training performs better for attributes with high skew—DEO is 23.9 compared to 14.9, KL is 1.5 compared to 0.8. This result is consistent with our earlier observation that our method works well for low to moderately skewed datasets. Wang et al also use a simpler weighted training method that reweights samples such that the protected attribute classes have equal weight and an adversarial training method that uses a minimax objective to maximize the classifier’s accuracy on the objective while minimizing an adversary’s ability to predict the protected attribute from the learned features. For weighted and adversarial training methods, we report results in Table 5. We find that while the weighted model overall performs well on the fairness metrics, it has a strongly negative BA (-2.7 versus our 0.5) indicating that bias is now in
the opposite direction, and a low AP (79.6 versus our 81.5) suggesting that it makes incorrect predictions to reduce bias. For adversarial training, our method does better overall, with lower DEO (16.7 versus 23.9) and lower BA (0.5 versus 1.5).

5. Extensions of our method

In this final section, we study two natural extensions of our method: using domain-dependent hyperplanes in place of the current domain-independent hyperplanes, and directly augmenting a real image dataset with GAN-inversion.

**Domain-dependent hyperplanes:** Our method implicitly assumes the learned hyperplane \( w \) behaves equally well for all \( z \), irrespective of the value of \( f_g(G(z)) \). However, for gender-dependent attributes, the hyperplane learned using samples with \( f_g(G(z)) = 1 \) may be very different from that learned using samples with \( f_g(G(z)) = -1 \).

For these attributes, we extend our method to learn per-domain target attribute hyperplanes: \( w_{t_1}, b_{t_1} \) for points with \( f_g(G(z)) = 1 \) and \( w_{t-1}, b_{t-1} \) for points with \( f_g(G(z)) = -1 \). For \( z \) with \( f_g(G(z)) = 1 \), we find \( z' \) such that

\[
\begin{align*}
    w_{t-1}T(z') + b_{t-1} &= w_{t_1}T(z) + b_{t_1}, \\
    w_gTz' + b_g &= -w_gT(z) - b_g
\end{align*}
\]

as shown in Figure 7. In order to compute \( z' \) that satisfies the above constraints, while minimizing \( ||z - z'||_2 \), we note that all constraints are linear; hence the feasible region is the intersection of several hyperplanes. Starting from a point in this region, in each iteration, we find a new location of the point using gradient descent, then project it back onto the feasible region to maintain the constraints.

If \( w_{t_1} \) and \( w_{t-1} \) are similar, these constraints are the same as Equation 2 and this method of computing \( z' \) collapses to the first. We compare results of training a classifier that is augmented with images computed with domain-independent hyperplanes and with that using images computed with domain-dependent hyperplanes for all gender-dependent and gender-independent attributes over the validation set. We find that for gender-dependent attributes, using domain-dependent hyperplanes improves the fairness metrics considerably (DEO reduces from 21.4 to 17.2, BA reduces from 1.5 to 0.4, KL reduces from 1.2 to 1.0), without losing accuracy. However, for gender-independent attributes, we do not see significant improvement, suggesting that \( w_t \) is similar to both \( w_{t_1} \) and \( w_{t-1} \). Full results are in Table 6.

**Augmenting real images with GAN-inversion:** Our method operates in the GAN latent space and can only augment images that are generated from latent vectors, and so, only the GAN-generated images. Recently, several GAN-inversion methods have been proposed [1, 5, 47]. These methods invert a real image \( x_{real} \in X \) to a vector \( z_{inv} \) in the latent space of a trained GAN. Using Zhu et al. [47], we tried directly augmenting the original dataset by perturbing \( z_{inv} \) to \( z'_{inv} \) with our method, creating \( x'_{real} = G(z'_{inv}) \) with the same target label and the opposite protected label of \( x_{real} \). When we trained classifiers with datasets augmented in this way, however, we did not see an appreciable improvement, despite the more complex procedure (Table 7).

### Table 6: Comparison of classifiers that use domain-dependent hyperplanes vs. domain-independent hyperplanes to compute \( z' \). We see a significant improvement among Gender-dependent attributes when we use Domain-dependent hyperplanes. Numbers are reported on the validation set.

<table>
<thead>
<tr>
<th>Attr. type</th>
<th>AP↑</th>
<th>DEO↓</th>
<th>BA↓</th>
<th>KL↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dom-dep</td>
<td>78.1 ±1.5</td>
<td>81.4 ±1.4</td>
<td>13.9 ±4.3</td>
<td>1.3 ±1.6</td>
</tr>
<tr>
<td>Dom-indep</td>
<td>84.5 ±1.5</td>
<td>84.6 ±1.7</td>
<td>13.1 ±2.7</td>
<td>1.2 ±2.7</td>
</tr>
</tbody>
</table>

### Table 7: Comparison of our classifiers (without) to classifiers trained using data augmented with a GAN-inversion module (with inv.). Numbers reported are the mean over all gender-dependent and gender-independent attributes on the validation set. We do not see an appreciable improvement.

<table>
<thead>
<tr>
<th></th>
<th>AP↑</th>
<th>DEO↓</th>
<th>BA↓</th>
<th>KL↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without</td>
<td>82.6 ±1.5</td>
<td>1.3 ±2.3</td>
<td>1.3 ±0.4</td>
<td>1.0 ±0.5</td>
</tr>
<tr>
<td>With inv.</td>
<td>82.4 ±1.5</td>
<td>1.4 ±2.3</td>
<td>1.3 ±0.4</td>
<td>1.0 ±0.5</td>
</tr>
</tbody>
</table>

### 6. Conclusions

We introduced a GAN-based data augmentation method for training fairer attribute classifiers when correlations between the target label and the protected attribute (such as gender expression) might skew the results. We report results across a large number of attributes and metrics, including comparisons with existing techniques. We also analyze in detail when our method is the most effective. Our findings show the promise of augmenting data in the GAN latent space in a variety of settings. We hope our detailed analyses and publicly available code serve as a stepping stone for future explorations in this very important space.

**Acknowledgements:** This work is supported by the National Science Foundation under Grant No. 1763642 and the Princeton First Year Fellowship to SK. We also thank Arvind Narayanan, Deniz Oktay, Angelina Wang, Zeyu Wang, Felix Yu, Sharon Zhang, as well as the Bias in AI reading group for helpful comments and suggestions.
References


[16] FAIR HDGAN. Pytorch GAN Zoo. 4


[23] Khari Johnson. Google Cloud AI removes gender labels from Cloud Vision API to avoid bias, 02 2020. 2


