LATENT SPACE SMOOTHING FOR INDIVIDUALLY FAIR REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Fair representation learning encodes user data to ensure fairness and utility, regardless of the downstream application. However, learning individually fair representations, i.e., guaranteeing that similar individuals are treated similarly, remains challenging in high-dimensional settings such as computer vision. In this work, we introduce LASSI, the first representation learning method for certifying individual fairness of high-dimensional data. Our key insight is to leverage recent advances in generative modeling to capture the set of similar individuals in the generative latent space. This allows learning an individually fair representation where similar individuals are mapped close together, by using adversarial training to minimize the distance between the representations of similar individuals. Finally, we employ randomized smoothing to provably map similar individuals close together, in turn ensuring that local robustness verification of the downstream application results in end-to-end fairness certification. Our experimental evaluation on challenging real-world image data demonstrates that our method increases certified individual fairness by more than 60%, without significantly affecting task utility.

1 INTRODUCTION

Deep learning models are increasingly being deployed in important domains such as credit scoring (Khandani et al., 2010), crime risk assessment (Brennan et al., 2009), and others. Unfortunately, both models and datasets employed in these settings were shown to be biased (Klare et al., 2012; Buolamwini & Gebru, 2018), causing regulators to increasingly hold organizations accountable for the discriminatory effects of their models (EU, 2019; 2021; FTC, 2020; 2021; UN, 2021).

Fair representation learning (Zemel et al., 2013) is a promising bias mitigation approach that transforms user data to prevent discrimination regardless of the downstream application, while maintaining high task utility. The approach is highly modular (McNamara et al., 2019), with a *data regulator* defining the fairness notion, a *data producer* learning the fair representation, and *data consumers* employing the transformed data in downstream tasks. However, although recent work successfully learned representations with fairness guarantees (Ruoss et al., 2020; Gitiaux & Rangwala, 2021), the application to high-dimensional data, such as images, remains challenging.

Key challenge: scaling to high-dimensional data and real-world models The two central challenges in *individually* fair representation learning are: (i) designing a suitable input similarity metric (Zemel et al., 2013; Yurochkin et al., 2020), and (ii) enforcing that similar individuals are provably treated similarly (according to the designed metric). For low-dimensional tabular data, prior work typically measures input similarity in terms of input features (age, income, etc.), using, e.g., logical constraints (Ruoss et al., 2020) or weighted ℓ_p-metrics (Yeom & Fredrikson, 2020). However, for high-dimensional data, such as images, characterizing the similarity on the input-level, e.g., by comparing pixels, is impractical. Moreover, proving that all points in the infinite set of similar individuals receive the same outcome entails propagating this set through the model, which is out of reach for prior approaches that rely on (mixed-integer) linear programming solvers (Ehlers, 2017; Tjeng et al., 2019), which only scale to small networks that are not useful for high-dimensional data.

This work In this work, we introduce Latent Space Smoothing for Individually Fair Representations (LASSI), a method which addresses both of these challenges. A high-level overview of our approach

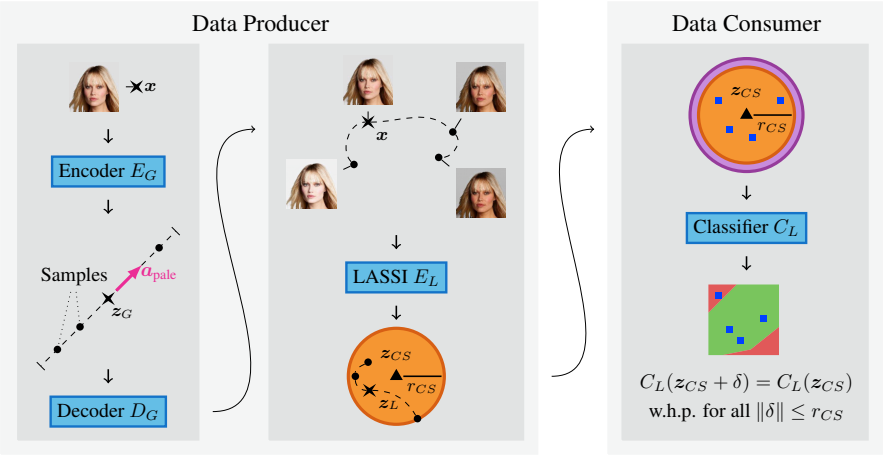


Figure 1: Overview of LASSI. The left part shows the data producer who captures the set of individuals similar to $\boldsymbol{x}$ by interpolating along the attribute vector $\boldsymbol{a}_{\text{pale}}$. The data producer then uses adversarial training (Madry et al., 2018) and center smoothing (Kumar & Goldstein, 2021) to compute a representation that provably maps all similar points into the ℓ_2-ball of radius r_{CS} around $\boldsymbol{z}_{CS}$. The right part shows the data consumer who can then certify individual fairness, i.e., prove that all similar individuals receive the same classification outcome, of the end-to-end model by checking whether the certified radius obtained via randomized smoothing (Cohen et al., 2019) exceeds r_{CS}.

is shown in Fig. 1. Concretely, we use recent advances in generative modeling (Kingma & Dhariwal, 2018) to enable data regulators to define input similarity by varying some continuous attribute of the image, such as pale skin in Fig. 1. To enforce that similar individuals are provably treated similarly, we base our approach on smoothing: (i) the data producer uses center smoothing (Kumar & Goldstein, 2021) in order to learn a representation that provably maps similar individuals close together, and (ii) the data consumer certifies local ℓ_2-robustness using randomized smoothing (Cohen et al., 2019), thereby proving individual fairness of the end-to-end model.

To measure input similarity, the data producer leverages the ability of a bijective generative model to interpolate in the latent space along the direction of the attribute vector. Consequently, the set of similar individuals is given by a line segment (lower left of Fig. 1), corresponding via the bijection to an elaborate curve in the input space (top middle of Fig. 1). However, this curve in the input space cannot be easily captured by an ℓ_p-ball. Therefore, the data producer learns a representation $E_L \circ D_G$ that maps all points of the latent line segment close together by using adversarial training to minimize the distance between similar individuals. As adversarial training does not provide guarantees on the maximum distance, the data producer uses center smoothing (Kumar & Goldstein, 2021) to adjust the representation such that the *smoothed* encoder $\widehat{E_L \circ D_G}$ provably maps all similar points into an ℓ_2-ball of radius r_{CS} around a center $\boldsymbol{z}_{CS}$ with high probability (lower middle of Fig. 1). Finally, the data consumer only needs to prove that the certified radius (violet, top right in Fig. 1) of its *smoothed* classifier $\overline{C_L}$ around $\boldsymbol{z}_{CS}$ is larger than r_{CS} to obtain an individual fairness certificate for the end-to-end model $\overline{C_L} \circ \widehat{E_L \circ D_G} \circ E_G$.

Our experimental evaluation on real-world image classification tasks shows that our method significantly increases the number of individuals for which we can certify individual fairness by up to 68% compared to the baseline. We also use a procedurally generated dataset to confirm that certificates obtained using the generative model are sound and transfer to the ground truth dataset.

Main contributions Our main contributions are:

- A novel input similarity metric for high-dimensional data via latent space interpolation.
- A scalable representation learning method with individual fairness certification for real-world models operating on high-dimensional data via randomized smoothing.
- An efficient implementation and an exhaustive evaluation of our method on a variety of real-world image classification tasks. We will release our code and all pretrained models.

2 Related Work

In this work, we consider individual fairness, which requires that similar individuals be treated similarly (Dwork et al., 2012). In contrast, group fairness enforces that specific classification statistics are equal across different groups of the population (Dwork et al., 2012; Hardt et al., 2016). While both fairness notions are desirable, they also both suffer from certain shortcomings. For instance, models satisfying group fairness may still discriminate against individuals (Dwork et al., 2012) or subgroups (Kearns et al., 2018). In contrast, the central challenge limiting practical adoption of individual fairness is the lack of a widely accepted similarity metric (Yurochkin et al., 2020). While recent work has made progress in developing similarity metrics for tabular data (Wang et al., 2019a; Mukherjee et al., 2020; Maity et al., 2021; Yurochkin & Sun, 2021; Ilvento, 2020), defining a concise similarity for high-dimensional data remains challenging and is a key goal of our work.

Fair representation learning A wide range of methods have been proposed to learn fair representations of user data. Most of these works consider group fairness and employ techniques such as adversarial learning (Edwards & Storkey, 2016; Madras et al., 2018; Liao et al., 2019; Kehrenberg et al., 2020), disentanglement (Creager et al., 2019; Locatello et al., 2019; Sarhan et al., 2020), duality (Song et al., 2019), low-rank matrix factorization (Oneto et al., 2020), and distribution alignment (Louizos et al., 2016; Zhao et al., 2020; Balunovic et al., 2021). Individually fair representation learning has recently gained attention, with similarity metrics based on logical formulas (Ruoss et al., 2020), Wasserstein distance (Lahoti et al., 2019a; Feng et al., 2019), fairness graphs (Lahoti et al., 2019b), and weighted ℓ_p-norms (Zemel et al., 2013). Unfortunately, these approaches cannot capture similarity between individuals for the high-dimensional data we consider in our work.

Bias in high-dimensional data A long line of work has investigated the biases of models operating on high-dimensional data, such as images (Wang et al., 2020; Wilson et al., 2019) and text (Bolukbasi et al., 2016; Tatman, 2017; Park et al., 2018; Liang et al., 2021), showing, e.g., that black women obtain lower accuracy in commercial face classification (Klare et al., 2012; Buolamwini & Gebru, 2018; Raji & Buolamwini, 2019). Importantly, these models not only learn but also amplify the biases of the training data (Zhao et al., 2017; Hendricks et al., 2018), even for balanced datasets (Wang et al., 2019b). A key challenge for investigating and mitigating bias in high-dimensional data is that, unlike tabular data, sensitive attributes such as gender or skin color are not directly encoded as features. Thus, prior work often relies on generative models (Kim et al., 2018; Denton et al., 2019; Sattigeri et al., 2019; Dash & Sharma, 2020; Balakrishnan et al., 2020; Joo & Kärkkäinen, 2020; Ramaswamy et al., 2021; Kim et al., 2021) or computer simulations (McDuff et al., 2018) to manipulate data attributes and check whether the perturbed instances are classified the same. However, unlike our work, these methods only empirically test for bias and do not provide certification guarantees.

Fairness certification Regulatory agencies are increasingly holding organizations accountable for the discriminatory effects of their machine learning models (EU, 2019; 2021; FTC, 2020; 2021; UN, 2021). Accordingly, designing algorithms with fairness guarantees has become an active area of research (Balunovic et al., 2021; Gitiaux & Rangwala, 2021; Albarghouthi et al., 2017; Bastani et al., 2019; Segal et al., 2021; Choi et al., 2021). However, unlike our work, most approaches for individual fairness certification consider pretrained models and thus cannot be employed in fair representation learning (Yeom & Fredrikson, 2020; John et al., 2020; Urban et al., 2020). In contrast, Ruoss et al. (2020) learn individually fair representations with provable guarantees for low-dimensional tabular data, providing a basis for our approach. However, neither the similarity notions nor the certification methods employed by Ruoss et al. (2020) scale to the high-dimensional data we consider.

3 Background

In this section we provide the necessary background on fair representation learning, generative modelling, and randomized smoothing.

LCIFR The LCIFR framework (Ruoss et al., 2020) learns representations with individual fairness guarantees for low-dimensional tabular data, ensuring that for a point $\boldsymbol{x} \in \mathbb{R}^n$ all similar individuals are treated similarly. To that end, Ruoss et al. (2020) define a family of similarity notions and leverage

(mixed-integer) linear programming methods to propagate all similar individuals through the model. Then, if all the points satisfying the given similarity notion obtain the same classification, LCIFR has provably shown that individual fairness is satisfied at $\boldsymbol{x}$. However, high-dimensional applications remain out of reach for LCIFR since both the similarity notions and linear programming methods are tailored to low-dimensional tabular data. Concretely, similarity is defined via logical formulas operating on the features of $\boldsymbol{x}$, which is infeasible for e.g., images, which cannot be compared solely on the pixel level. Moreover, while the linear programming employed by Ruoss et al. (2020) are known to work well for small neural networks, they do not scale to real-world computer vision models. In this work, we thus demonstrate how to resolve these two key concerns to generalize the high-level idea of LCIFR to real-world, high-dimensional applications.

Glow Normalizing flows, such as Glow (Kingma & Dhariwal, 2018), recently emerged as a promising approach for generative modeling due to their exact log-likelihood evaluation, efficient inference and synthesis, and useful latent space for downstream tasks. Unlike GANs (Goodfellow et al., 2014) or VAEs (Kingma & Welling, 2014), normalizing flows are bijective models consisting of an encoder $E_G : \mathbb{R}^n \to \mathbb{R}^q$ and a decoder $D_G : \mathbb{R}^q \to \mathbb{R}^n$ for which $\boldsymbol{x} = D_G\left(E_G\left(\boldsymbol{x}\right)\right)$. The latent space of Glow captures important attributes of the data, which enables latent space interpolation such as, e.g., manipulating the skin color of a person in an image. While attribute manipulation via latent space interpolation has also been investigated in the fairness context for GANs and VAEs (Kim et al., 2018; Denton et al., 2019; Balakrishnan et al., 2020; Joo & Kärkkäinen, 2020; Ramaswamy et al., 2021), the key advantage of Glow is the existence of an encoder (unlike GANs, which cannot represent an input point in the latent space efficiently) and the bijectivity of the end-to-end model (VAEs cannot reconstruct the input point exactly). Our key idea is to leverage Glow to define the similarity between two images by interpolating along certain sensitive attributes in the latent space.

Randomized Smoothing Given a standard classifier, recent work (Cohen et al., 2019) constructs smooth classifiers together with a probabilistic certification method. Cohen et al. (2019) defines the smoothed classifier $\bar{f}$ of a standard classifier $f\colon \mathbb{R}^m \to \mathcal{Y}$ by

$$\bar{f}(\boldsymbol{x}) := \arg\max_c \mathbb{P}_{\epsilon\sim\mathcal{N}(0,\sigma^2 I)}(f(\boldsymbol{x}+\epsilon) = c).$$

Further, Cohen et al. (2019) provide the following theorem yielding robustness certificates:

Theorem 3.1 (from (Cohen et al., 2019)). *Suppose $c_A \in \mathcal{Y}$, $\underline{p_A}, \overline{p_B} \in [0,1]$. If*

$$\mathbb{P}_\epsilon(f(\boldsymbol{x}+\epsilon) = c_A) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c\neq c_A} \mathbb{P}_\epsilon(f(\boldsymbol{x}+\epsilon) = c_B),$$

then $\bar{f}(\boldsymbol{x}+\delta) = c_A$ for all δ satisfying $\|\delta\|_2 \leq R$ with $R := \frac{\sigma}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B}))$.

Since calculating the above quantities algebraically is not feasible in cases where f is a large neural network, the quantities c_A and $\mathbb{P}_\epsilon(f(\boldsymbol{x}+\epsilon) = c_A)$ have to be estimated by sampling, resulting in a $1-\alpha_s$ confidence to lower bound $\underline{p_A}$. Thus, the overall certificate holds with confidence $1-\alpha_s$.

Recently, Kumar & Goldstein (2021) presented a method to smooth vector valued functions. This is particularly useful in our setting to certify a smoothed version of the encoder $f = E_L \circ D_G$. The smoothed function $\bar{f}$ evaluated at x returns the center of the minimum enclosing ball covering at least $1/2$ the probability mass of $f(x + \mathcal{N}(0, \sigma^2 I))$. As evaluating $\bar{f}$ directly is infeasible, we evaluate a proxy $\hat{f}$ relying on sampling and approximation: Specifically, $\hat{f}(x)$ evaluates to the center $\tilde{z}$ of a relaxed version of a minimum enclosing ball containing at least half the points in $Z = \{z_i\}_{i=1}^n$ where $z_i \sim f(x + \mathcal{N}(0, \sigma^2 I))$. The robustness certificate results from the following theorem:

Theorem 3.2 (adapted from (Kumar & Goldstein, 2021)). *With probability at least $1-\alpha_c$ we have,*

$$\forall x' \text{ s.t. } \|x - x'\|_2 \leq \epsilon, \|\hat{f}(x) - \hat{f}(x')\|_2 \leq r_{CS}.$$

Here, r_{CS} depends on the given ϵ, α_c and on a quantile of the distances $\|\hat{f}(x) - z_i\|_2$ for $z_i \in Z$.

4 Individually Fair Representations of High-Dimensional Data

In this section we first define the set of individuals similar to $\boldsymbol{x}$ (Section 4.1). We then describe our approach for learning individually fair representations of these individuals (Section 4.2) and finally

demonstrate how we can certify individual fairness for them (Section 4.3). We emphasize that our approach is general, but for presentational purposes we focus on the case where $\boldsymbol{x}$ is an image.

4.1 SIMILARITY VIA GENERATIVE MODEL

We consider two individuals $\boldsymbol{x}$ and $\boldsymbol{x}'$ to be similar if they differ only in their sensitive attributes, e.g., skin color, and all the other attributes are the same. However, such semantic attributes cannot be conveniently captured directly in the input (pixel) space of the data, so we leverage the latent space of a generative model G. Our first step is to compute a vector $\boldsymbol{a}$ associated with the attribute, such that interpolating along the direction of $\boldsymbol{a}$ in the latent space, and reconstructing back to the input space results in a meaningful semantic transformation of that attribute. This in itself is an active research area with various approaches for computing $\boldsymbol{a}$ (Higgins et al., 2017; Denton et al., 2019).

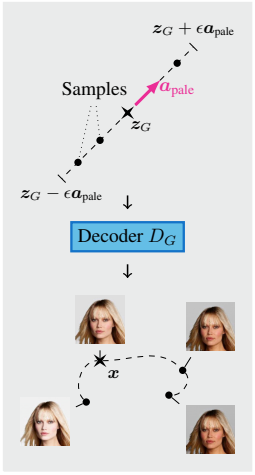


Figure 2: Similarity

Individual similarity via Glow In this work, we propose to use Glow (Kingma & Dhariwal, 2018) to define individual fairness using its latent space. Let $\boldsymbol{z}_G = E_G(\boldsymbol{x})$ be the latent code of $\boldsymbol{x}$ in the generative model latent space. We calculate the average latent vectors $\boldsymbol{z}_{G,pos}$ for samples with the attribute and $\boldsymbol{z}_{G,neg}$ for samples without the attribute, and set the attribute vector to the difference between them: $\boldsymbol{a} = \boldsymbol{z}_{G,pos} - \boldsymbol{z}_{G,neg}$. As observed by Kingma & Dhariwal (2018), moving in the direction of $\boldsymbol{a}$ in the latent space increases the presence of the attribute and interpolating in the opposite direction decreases its strength. Once we have G and $\boldsymbol{a}$, we define the set of individuals similar to $\boldsymbol{x}$ *in the latent space of* G to be $S_l(\boldsymbol{x}) = \{\boldsymbol{z}_G + t \cdot \boldsymbol{a} \mid t \in [-\epsilon, \epsilon]\}$ (see top of Fig. 2). Here, ϵ defines the maximum perturbation level, which we can apply to the attribute. Together with G and $\boldsymbol{a}$, it is considered to be a part of the similarity specification. Crucially, the similarity set S_l contains an infinite number of points but is compactly represented in the latent space of G in the form of the line segment ranging from $\boldsymbol{z}_G - \epsilon \cdot \boldsymbol{a}$ to $\boldsymbol{z}_G + \epsilon \cdot \boldsymbol{a}$. In contrast, the same set represented directly in the input space of the data, $S_i(\boldsymbol{x}) \coloneqq \{D_G(\boldsymbol{z}) \mid \boldsymbol{z} \in S_l(\boldsymbol{x})\}$, i.e., obtained by propagating the latent representations in $S_l(\boldsymbol{x})$ through the decoder of the generative model, cannot be abstracted conveniently without leveraging any kind of latent information (see bottom of Fig. 2). Moreover, our approach for constructing $S_l(\boldsymbol{x})$ can also be extended to multiple sensitive attributes by simply interpolating along their attribute vectors simultaneously.

4.2 LEARNING INDIVIDUALLY FAIR REPRESENTATIONS

Assuming that the generative model $G = D_G \circ E_G$ is pretrained and given, in this section we describe the training of the encoder $E_L : \mathbb{R}^n \to \mathbb{R}^k$, which together with G, is part of the data producer of our end-to-end model. E_L will be trained separately from the data consumer, the classifier C_L, the training of which is explained in the next section.

Adversarial loss We encourage similar treatment for all points in $S_i(\boldsymbol{x})$ by training E_L such that the data producer maps all points in $S_i(\boldsymbol{x})$ close to each other in $\mathbb{R}^k$. This can be achieved by training E_L to minimize the loss function

$$\mathcal{L}_{adv}(\boldsymbol{x}) = \max_{\boldsymbol{z}' \in S_l(\boldsymbol{x})} \| (E_L \circ D_G)(\boldsymbol{z}_G) - (E_L \circ D_G)(\boldsymbol{z}') \|_2.$$

Minimizing $\mathcal{L}_{adv}(\boldsymbol{x})$ is a min-max optimization problem and adversarial training (Madry et al., 2018) has been demonstrated to work well in such setups. Since the underlying domain of the inner maximization problem is simply the line segment $S_l(\boldsymbol{x})$, we perform a random adversarial attack in which we sample k points $\boldsymbol{z}_i \sim \mathcal{U}(S_l(\boldsymbol{x}))$ uniformly at random from $S_l(\boldsymbol{x})$ and approximate $\mathcal{L}_{adv}(\boldsymbol{x}) \approx \max_{i=1}^{k} \| (E_L \circ D_G)(\boldsymbol{z}_G) - (E_L \circ D_G)(\boldsymbol{z}_i) \|_2$.

Classification loss To learn representations useful for downstream tasks, we train E_L jointly with an auxiliary classifier C_L^{aux} to predict a ground-truth target label y by having an additional term in the loss function:

$$\mathcal{L}_{cls}(\boldsymbol{x}, y) = \text{cross\_entropy}\left((C_L^{aux} \circ E_L \circ D_G)(\boldsymbol{z}_G), y \right).$$

Contrastive loss We observe that, in theory, we can make $\mathcal{L}_{adv}$ arbitrarily small while preserving $\mathcal{L}_{cls}$ by multiplying the output of E_L and dividing the input of C_L by a large constant. This effectively pushes all points $(E_L \circ D_G)(\boldsymbol{z})$ close to each other regardless of their ground-truth class y. To prevent this from happening and to promote better separability of the classes, we consider a modification of the contrastive loss (Chopra et al., 2005). Denoting $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \| (E_L \circ G)(\boldsymbol{x}_i) - (E_L \circ G)(\boldsymbol{x}_j) \|_2$, we define the contrastive loss over a batch

$$\mathcal{L}_{contr}(\boldsymbol{x}, y) = \sum_{i<j} \mathbb{I}[y_i = y_j] \max(0, d(\boldsymbol{x}_i, \boldsymbol{x}_j) - \delta) + \mathbb{I}[y_i \neq y_j] \max(0, 2\delta - d(\boldsymbol{x}_i, \boldsymbol{x}_j))$$

where $\mathbb{I}[\cdot]$ is indicator function and δ is a hyperparameter. This loss term effectively penalizes pairs of inputs from the same class which are encoded further than δ from each other and pairs from different classes which are encoded closer than 2δ from each other. In the end, E_L is trained by solving

$$\underset{E_L, C_L^{aux}}{\arg\min}\, \mathbb{E}_{(\boldsymbol{x},y)\sim D}\left[\mathcal{L}_{cls}(\boldsymbol{x}, y) + \lambda_1 \mathcal{L}_{adv}(\boldsymbol{x}) + \lambda_2 \mathcal{L}_{contr}(\boldsymbol{x}, y)\right]$$

with λ_1 and λ_2 being hyperparameters balancing the adversarial and contrastive loss terms.

4.3 Certifying Individual Fairness via Latent Space Smoothing

At this point we have a pretrained generative model $G = D_G \circ E_G$, which allows us to compute the attribute vector $\boldsymbol{a}$. $S_l(\boldsymbol{x})$ and in turn $S_i(\boldsymbol{x})$ are defined based on $\boldsymbol{a}$ and the maximum perturbation level ϵ. Assume that we have trained E_L as discussed above, as well as C_L, the training of which is described at the end of this section. The goal now is to construct an end-to-end model P for which we can certify individual fairness of the form

$$\forall \boldsymbol{x}' \in S_i(\boldsymbol{x}) \,.\, P(\boldsymbol{x}) = P(\boldsymbol{x}')$$

for a given input $\boldsymbol{x}$. We denote $\boldsymbol{z}_G = E_G(\boldsymbol{x})$ and define the function $f(t) = E_L(D_G(\boldsymbol{z}_G + t \cdot \boldsymbol{a}))$ with $t \in \mathbb{R}$. We apply the center smoothing procedure presented by Kumar & Goldstein (2021) to obtain the smoothed version of f, namely $\hat{f}$. Instantiating Theorem 3.2 to $\hat{f}$ with ϵ tells us that for $t = 0$, with probability at least $1 - \alpha_c$ we have that $\forall t'$ s.t. $\|t - t'\| \leq \epsilon$, $\|\hat{f}(t) - \hat{f}(t')\|_2 \leq r_{CS}$ (see Fig. 3). We obtain the center smoothing radius r_{CS} computed for $t = 0$ and then by expanding back the definition of f (with $t = 0$) the guarantee we have is that with probability at least $1 - \alpha_c$

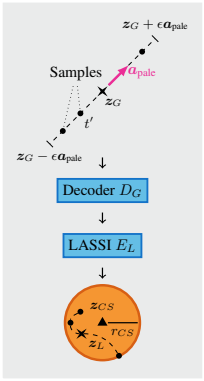


Figure 3: Center-smoothing

$$\forall t' \in [-\epsilon, \epsilon] \,.\, \left\| (\widehat{E_L \circ D_G})(\boldsymbol{z}_G) - (\widehat{E_L \circ D_G})(\boldsymbol{z}_G + t' \cdot \boldsymbol{a}) \right\|_2 \leq r_{cs} \tag{1}$$

$$\iff \forall \boldsymbol{z} \in S_l(\boldsymbol{x}) \,.\, \left\| (\widehat{E_L \circ D_G})(\boldsymbol{z}_G) - (\widehat{E_L \circ D_G})(\boldsymbol{z}) \right\|_2 \leq r_{cs} \tag{2}$$

Moreover, center smoothing provides us an estimate of the center $\boldsymbol{z}_{CS} = (\widehat{E_L \circ D_G})(\boldsymbol{z}_G)$. On the other hand, by smoothing the classifier C_L, Theorem 3.1 gives us a radius R and certifies that with probability $1 - \alpha_s$ the prediction $\overline{C_L}(\boldsymbol{z}_{CS} + \delta)$ is the same for all δ s.t. $\|\delta\|_2 \leq R$. Combining this with Eq. (2), we derive that if $r_{CS} \leq R$, then

$$\forall \boldsymbol{z} \in S_l(\boldsymbol{x}) \,.\, \overline{C_L}\big((\widehat{E_L \circ D_G})(\boldsymbol{z}_G)\big) = \overline{C_L}\big((\widehat{E_L \circ D_G})(\boldsymbol{z})\big)$$

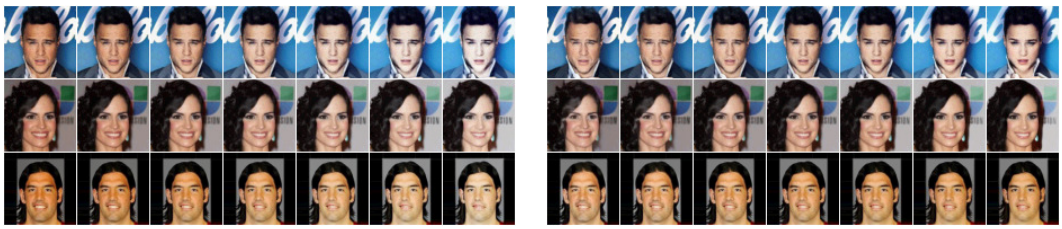
Taking into account the bijectivity of Glow, $\forall \boldsymbol{x}' \in S_i(\boldsymbol{x}) \,.\, E_G(\boldsymbol{x}') \in S_l(\boldsymbol{x})$, we obtain Theorem 4.1 stating that our end-to-end model is provably individually fair.

Theorem 4.1 (Informally)**.** *If neither center smoothing nor randomized smoothing abstain during the computation of the smoothed model $P = \overline{C_L} \circ (\widehat{E_L \circ D_G}) \circ E_G$ with input $\boldsymbol{x}$, and $r_{CS} \leq R$, then P is certifiably individually fair for $\boldsymbol{x}$ w.r.t. the similarity set $S_i(\boldsymbol{x})$ with probability $1 - \alpha_c - \alpha_s$.*

The procedure for certifying individual fairness is summarized in Algorithm 1. We note that the certificate obtained by the function Certify is probabilistic – it holds with probability at least $1 - \alpha_c - \alpha_s$ (being conservative here, as the abstentions of the center and the classification smoothing might not be independent) to account for the fact that the center smoothing and the classification smoothing give probabilistic certificates as well.

Algorithm 1 Certifying the individual fairness of $\overline{C_L} \circ (\widehat{E_L \circ D_G}) \circ E_G$ for the input $\boldsymbol{x}$.

function CERTIFY(E_G, D_G, E_L, C_L, $\boldsymbol{x}$)
  $\boldsymbol{z}_{CS} = (\widehat{E_L \circ D_G})(E_G(\boldsymbol{x}))$ and r_{CS} from center smoothing (Kumar & Goldstein, 2021)
  if center smoothing abstained **then return** ABSTAIN
  Do classification smoothing (Cohen et al., 2019) to obtain the certified radius R around $\boldsymbol{z}_{CS}$ for which the classification stays the same
  if $r_{CS} \leq R$ **then return** CERTIFIED
  else return NOT CERTIFIED



(a) Varying `Pale_Skin` attribute.

(b) Varying `Young` attribute.

Figure 4: Points from the similarity set $S_i(\boldsymbol{x})$ for various $\boldsymbol{x}$, as reconstructed by our Glow model. Central images are reconstructions of the original inputs. The variations go uniformly (from left to right) in the range of $t \in [-\epsilon, \epsilon]$.

Training C_L Since we will be applying smoothing over the classifier C_L, following Cohen et al. (2019), once we have trained E_L, we train C_L separately by adding isotropic Gaussian noise to its inputs during the training process. The inputs of C_L here are the outputs of $E_L \circ D_G \circ E_G$. We do not smooth the pipeline at this step as it is computationally expensive and because the distance between the smoothed and the unsmoothed outputs is generally small (Kumar & Goldstein, 2021).

5 EXPERIMENTAL EVALUATION

In this section we perform experimental evaluation of LASSI on several image classification tasks.

Experimental Setup We demonstrate the effectiveness of our approach on a real world dataset consisting of faces of celebrities, CelebA (Liu et al., 2015), annotated with the presence or absence of 40 face attributes. Following the setup of Kingma & Dhariwal (2018), we pretrain a Glow model with a number of flows $K = 32$, number of blocks $L = 4$ and additive coupling on 64×64 rescaled versions of the images. We select `Pale_Skin` and `Young` as our sensitive attributes, and we want to train a model which is certifiably individually fair for perturbations of these two sensitive attributes. We compute corresponding attribute vectors, $\boldsymbol{a}_{\text{pale}}$ and $\boldsymbol{a}_{\text{young}}$, as discussed in Section 4.1. We set the maximum perturbation level in the latent space, used to define individual fairness, to $\epsilon = 1$. Fig. 4 provides examples of images from the $S_i(\boldsymbol{x})$ similarity sets for various original inputs $\boldsymbol{x}$. We consider the binary classification task of predicting the `Smiling` attribute.

We train the LASSI encoder E_L as described in Section 4.2. To show the scalability of our method, we use the ResNet-18 (He et al., 2016) architecture for E_L. We provide the following ablations: a standard representation learning baseline ($\lambda_1 = \lambda_2 = 0$), training with adversarial loss ($\lambda_1 = 0.1$, $\lambda_2 = 0$), and training with both adversarial and contrastive losses ($\lambda_1 = 0.1$, $\lambda_2 = 0.1$, $\delta = 10$). For each encoder E_L, we train C_L with random Gaussian noise augmentation $\mathcal{N}(0, \sigma_{cls}^2 I)$ where $\sigma_{cls} \in \{1, 2.5, 5, 10, 25, 50\}$. At inference time, computing P includes performing center smoothing for $\widehat{E_L \circ D_G}$. We tried 3 different values for the standard deviation $\sigma_{enc} \in \{0.5, 0.75, 1\}$ of the Gaussian noise used in center smoothing and picked the one which maximizes the certified fairness of the baselines, $\sigma_{enc} = 0.75$, and which we use in the rest of the CelebA experiments. We execute center smoothing with $\alpha_c = 0.01$ and randomized smoothing with $\alpha_s = 0.001$. We will release our code, pretrained models and scripts to reproduce the results presented in this paper.

Table 1: Evaluation of LASSI on CelebA dataset, showing that LASSI significantly increases certified individual fairness compared to the baseline without affecting the classification accuracy.

Attribute	Method	σ_{cls}	Accuracy (%)	Certified Fairness (%)
Pale_Skin	Baseline	10	**93.59**	10.90
	LASSI + Adversarial	1	91.03	55.77
	LASSI + Adv + Contrastive	10	90.38	**78.85**
Young	Baseline	10	**91.67**	17.31
	LASSI + Adversarial	1	**91.67**	66.03
	LASSI + Adv + Contrastive	10	**91.67**	**80.77**
Pale_Skin + Young	Baseline	10	91.67	7.05
	LASSI + Adversarial	1	**92.95**	48.72
	LASSI + Adv + Contrastive	10	88.46	**65.38**

Results We show the results in Table 1, measured on a subset of 156 samples from CelebA's test set. For an input point $\boldsymbol{x}$, we are interested if our end-to-end model P classifies $\boldsymbol{x}$ correctly and if it is individually fair for $\boldsymbol{x}$ w.r.t. the similarity set $S_i(\boldsymbol{x})$, i.e., whether we can provably certify that *all points* in $S_i(\boldsymbol{x})$ are classified the same by P. For each input sample $\boldsymbol{x}$, we consider the prediction to be accurate if $P(\boldsymbol{x})$ matches the ground-truth label. P is considered to be individually fair for $\boldsymbol{x}$ if the CERTIFY method from Algorithm 1 returns CERTIFIED. We report the value for the σ_{cls} which maximizes the individual fairness without major sacrifice in accuracy (note that the constant function trivially has 100% individual fairness).

We observe that adversarial training significantly improves the certified fairness, compared to the baseline, at the cost of a minor drop in accuracy. The results are further improved by combining adversarial training with our proposed contrastive loss, confirming our intuition from Section 4.2. The baseline obtains its highest certified fairness for a relatively large values of σ_{cls}. This is the case because by default nothing enforces the points from $S_i(\boldsymbol{x})$ to be mapped closely together in the output space of E_L (i.e., the input of C_L), and we need a big σ_{cls} to obtain a certified radius R of $\overline{C_L}$ which exceeds r_{CS}. However, as we increase σ_{cls} too much, the smoothed classifier $\overline{C_L}$ becomes more and more uncertain and begins to abstain more.

The combination of adversarial training and contrastive loss also obtains its best certified fairness for a value of σ_{cls} bigger than that for adversarial training only. This is likely because the adversarial loss in itself pushes not only the points from $S_i(\boldsymbol{x})$ close together but also possibly decreases the distance between points from different classes (in the E_L output space). This issue is partially alleviated by the contrastive loss which enables us to keep increasing σ_{cls} and still certify more.

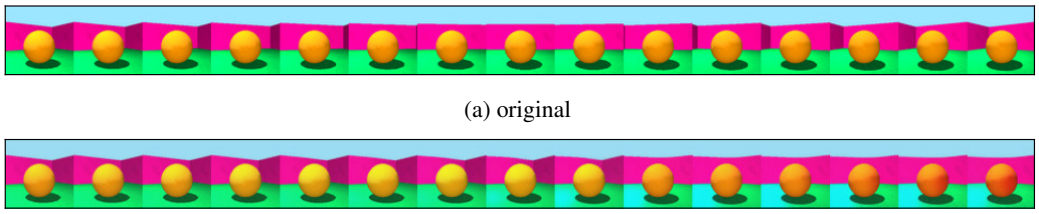
We also performed an experiment which combines the two sensitive attributes Pale_Skin and Young together by defining the similarity set as $S_l(\boldsymbol{x}) = \left\{E_G(\boldsymbol{x}) + t_1 \cdot \boldsymbol{a}_{\text{pale}} + t_2 \cdot \boldsymbol{a}_{\text{young}} \mid \sqrt{t_1^2 + t_2^2} \leq \epsilon\right\}$, i.e., (t_1, t_2) is at most ϵ away from the origin. Results in Table 1 show that LASSI also successfully enforces individual fairness in this case.

Performance As performing center smoothing for $\widehat{E_L \circ D_G}$ is costly, we perform evaluation on a subset of the test set. Center smoothing involves not only propagating through the large generative model D_G, but doing so for each random sample needed by the center smoothing to produce the final output. Certifying a single sample takes around 64 seconds after parallelizing the center smoothing procedure over 4 GeForce RTX 2080 Ti GPUs. On the other hand, we are capable of handling large networks, such as ResNet, which are out of reach for the prior work such as Ruoss et al. (2020).

Certification with disentangled ground truth data We showed that LASSI successfully learns representations with certified individual fairness for high-dimensional data (see Table 1). However, as we leverage generative models to capture the similarity between inputs, we now demonstrate that that the fairness certificates obtained using the generative model transfer to ground truth data. Since the CelebA does not contain images of the same individual with different attributes, e.g., the same

Table 2: Evaluation on the 3D shapes. Certification using the generative model transfers to ground truth data as the certification rate and percentage of unfair ground truth data sum up below 100%.

Attribute	Target	Method	Accuracy (%)	Certified Fairness (%)	Unfair (%)
`Orientation`	`Object_Hue`	Baseline	92.67	0	44.67
		LASSI + Avd + Contrastive	**100**	**88**	**0**



(a) original

(b) interpolated

Figure 5: A sample shape at 15 different ground truth orientations (original) and the corresponding reconstructions obtained from interpolating along the generative model's attribute vector (interpolated).

individual with different skin colors, we use the 3D Shapes dataset (Burgess & Kim, 2018), which consists of images of 3D shapes that are procedurally generated from 6 independent latent factors: `floor hue`, `wall hue`, `object hue`, `scale`, and `orientation`. The 3D shapes dataset is typically used to investigate disentanglement properties of unsupervised learning methods (e.g., in the context of fairness (Locatello et al., 2019)). Our goal is to show that the attribute vector learned by Glow captures a given latent factor, and thus certification with respect to the similarity set defined via interpolation along that attribute vector will result in certification of the ground truth data. To that end, we consider `orientation` as the continuous sensitive attribute, for which we have 15 similar ground truth data points, i.e., the same shape at 15 different orientations (fixing all other factors). Thus, our certification transfers to ground truth data if for every certified data point from the test set, all 15 similar ground truth data points obtain the same classification. Indeed, Table 2 shows that the percentage of unfair ground truth data points is always below certification rate. Moreover, Fig. 5 illustrates that interpolation along Glow's attribute vector closely mimics the ground truth data.

6 Conclusion

We defined image similarity with respect to a generative model via attribute manipulation, allowing us to capture complex image manipulations such as changing the age or skin color, which are otherwise difficult to characterize. Further, we were able to scale certified representation learning for individual fairness to real world high dimensional datasets by using randomized smoothing based techniques. Our extensive evaluation yields promising results and illustrates the practicality of our approach.

Ethics Statement

Certified individual fairness is important for the application of machine learning systems in real world applications and required by regulators. As this work certifies individual fairness with respect to a generative model approximating the ground truth, potential biases encoded in the generative model can prevail and skew certification. However, quality advancements in generative models can directly translate into stronger guarantees of our method. This work enables a certified fair application of models making use of rich, high dimensional data.

REFERENCES

Aws Albarghouthi, Loris D'Antoni, Samuel Drews, and Aditya V. Nori. Fairsquare: probabilistic verification of program fairness. *Proc. ACM Program. Lang.*, 2017.

Guha Balakrishnan, Yuanjun Xiong, Wei Xia, and Pietro Perona. Towards causal benchmarking of bias in face analysis algorithms. In *Computer Vision - ECCV 2020 - 16th European Conference*, 2020.

Mislav Balunovic, Anian Ruoss, and Martin T. Vechev. Fair normalizing flows. *CoRR*, 2021.

Osbert Bastani, Xin Zhang, and Armando Solar-Lezama. Probabilistic verification of fairness properties via concentration. *Proc. ACM Program. Lang.*, 2019.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29*, 2016.

Tim Brennan, William Dieterich, and Beate Ehret. Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and Behavior*, 2009.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, 2018.

Chris Burgess and Hyunjik Kim. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

YooJung Choi, Meihua Dang, and Guy Van den Broeck. Group fairness by probabilistic modeling with latent fair decisions. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, 2021.

S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pp. 539–546 vol. 1, 2005. doi: 10.1109/CVPR.2005.202.

Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A. Weis, Kevin Swersky, Toniann Pitassi, and Richard S. Zemel. Flexibly fair representation learning by disentanglement. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Saloni Dash and Amit Sharma. Counterfactual generation and fairness evaluation using adversarially learned inference. *CoRR*, 2020.

Emily Denton, Ben Hutchinson, Margaret Mitchell, and Timnit Gebru. Detecting bias with generative counterfactual face attribute augmentation. *CoRR*, 2019.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science*, 2012.

Harrison Edwards and Amos J. Storkey. Censoring representations with an adversary. In *4th International Conference on Learning Representations*, 2016.

Rüdiger Ehlers. Formal verification of piece-wise linear feed-forward neural networks. In *Automated Technology for Verification and Analysis - 15th International Symposium*, 2017.

EU. Ethics guidelines for trustworthy ai, 2019.

EU. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, 2021.

Rui Feng, Yang Yang, Yuehan Lyu, Chenhao Tan, Yizhou Sun, and Chunping Wang. Learning fair representations via an adversarial framework. *CoRR*, 2019.

FTC. Using artificial intelligence and algorithms, 2020.

FTC. Aiming for truth, fairness, and equity in your company's use of ai, 2021.

Xavier Gitiaux and Huzefa Rangwala. Learning smooth and fair representations. In *The 24th International Conference on Artificial Intelligence and Statistics*, 2021.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, 2014.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29*, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Computer Vision - ECCV 2018 - 15th European Conference*, 2018.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations*, 2017.

Christina Ilvento. Metric learning for individual fairness. In *1st Symposium on Foundations of Responsible Computing*, 2020.

Philips George John, Deepak Vijaykeerthy, and Diptikalyan Saha. Verifying individual fairness in machine learning models. In *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence*, 2020.

Jungseock Joo and Kimmo Kärkkäinen. Gender slopes: Counterfactual fairness for computer vision models by attribute manipulation. *CoRR*, 2020.

Michael J. Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Thomas Kehrenberg, Myles Bartlett, Oliver Thomas, and Novi Quadrianto. Null-sampling for interpretable and fair representations. In *Computer Vision - ECCV 2020 - 16th European Conference*, 2020.

Amir E. Khandani, Adlar J. Kim, and Andrew W. Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 2010.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Hyemi Kim, Seungjae Shin, JoonHo Jang, Kyungwoo Song, Weonyoung Joo, Wanmo Kang, and Il-Chul Moon. Counterfactual fairness with disentangled causal effect variational autoencoder. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, 2021.

Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems 31*, 2018.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations*, 2014.

Brendan Klare, Mark James Burge, Joshua C. Klontz, Richard W. Vorder Bruegge, and Anil K. Jain. Face recognition performance: Role of demographic information. *IEEE Trans. Inf. Forensics Secur.*, 2012.

Aounon Kumar and Tom Goldstein. Center smoothing for certifiably robust vector-valued functions. *CoRR*, 2021.

Preethi Lahoti, Krishna P. Gummadi, and Gerhard Weikum. ifair: Learning individually fair data representations for algorithmic decision making. In *35th IEEE International Conference on Data Engineering*, 2019a.

Preethi Lahoti, Krishna P. Gummadi, and Gerhard Weikum. Operationalizing individual fairness with pairwise fair representations. *Proc. VLDB Endow.*, 2019b.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Jiachun Liao, Chong Huang, Peter Kairouz, and Lalitha Sankar. Learning generative adversarial representations (GAP) under fairness and censoring constraints. *CoRR*, 2019.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision*, 2015.

Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. In *Advances in Neural Information Processing Systems 32*, 2019.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. The variational fair autoencoder. In *4th International Conference on Learning Representations*, 2016.

David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations*, 2018.

Subha Maity, Songkai Xue, Mikhail Yurochkin, and Yuekai Sun. Statistical inference for individual fairness. In *9th International Conference on Learning Representations*, 2021.

Daniel J. McDuff, Roger Cheng, and Ashish Kapoor. Identifying bias in AI using simulation. *CoRR*, 2018.

Daniel McNamara, Cheng Soon Ong, and Robert C. Williamson. Costs and benefits of fair representation learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

Debarghya Mukherjee, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun. Two simple ways to learn individual fairness metrics from data. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Luca Oneto, Michele Donini, Massimiliano Pontil, and Andreas Maurer. Learning fair and transferable representations with theoretical guarantees. In *7th IEEE International Conference on Data Science and Advanced Analytics*, 2020.

Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

Vikram V. Ramaswamy, Sunnie S. Y. Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

Anian Ruoss, Mislav Balunovic, Marc Fischer, and Martin T. Vechev. Learning certified individually fair representations. In *Advances in Neural Information Processing Systems 33*, 2020.

Mhd Hasan Sarhan, Nassir Navab, Abouzar Eslami, and Shadi Albarqouni. Fairness by learning orthogonal disentangled representations. In *Computer Vision - ECCV 2020 - 16th European Conference*, 2020.

Prasanna Sattigeri, Samuel C. Hoffman, Vijil Chenthamarakshan, and Kush R. Varshney. Fairness GAN: generating datasets with fairness properties using a generative adversarial network. *IBM J. Res. Dev.*, 2019.

Shahar Segal, Yossi Adi, Benny Pinkas, Carsten Baum, Chaya Ganesh, and Joseph Keshet. Fairness in the eyes of the data: Certifying machine-learning models. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2021.

Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 2019.

Rachael Tatman. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 2017.

Vincent Tjeng, Kai Yuanqing Xiao, and Russ Tedrake. Evaluating robustness of neural networks with mixed integer programming. In *7th International Conference on Learning Representations*, 2019.

UN. The right to privacy in the digital age, 2021.

Caterina Urban, Maria Christakis, Valentin Wüstholz, and Fuyuan Zhang. Perfectly parallel fairness certification of neural networks. *Proc. ACM Program. Lang.*, 2020.

Hanchen Wang, Nina Grgic-Hlaca, Preethi Lahoti, Krishna P. Gummadi, and Adrian Weller. An empirical study on learning fairness metrics for COMPAS data with human supervision. *CoRR*, 2019a.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *IEEE/CVF International Conference on Computer Vision*, 2019b.

Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. Predictive inequity in object detection. *CoRR*, 2019.

Samuel Yeom and Matt Fredrikson. Individual fairness revisited: Transferring techniques from adversarial robustness. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 2020.

Mikhail Yurochkin and Yuekai Sun. Sensei: Sensitive set invariance for enforcing individual fairness. In *9th International Conference on Learning Representations*, 2021.

Mikhail Yurochkin, Amanda Bower, and Yuekai Sun. Training individually fair ML models with sensitive subspace robustness. In *8th International Conference on Learning Representations*, 2020.

Richard S. Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J. Gordon. Conditional learning of fair representations. In *8th International Conference on Learning Representations*, 2020.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.

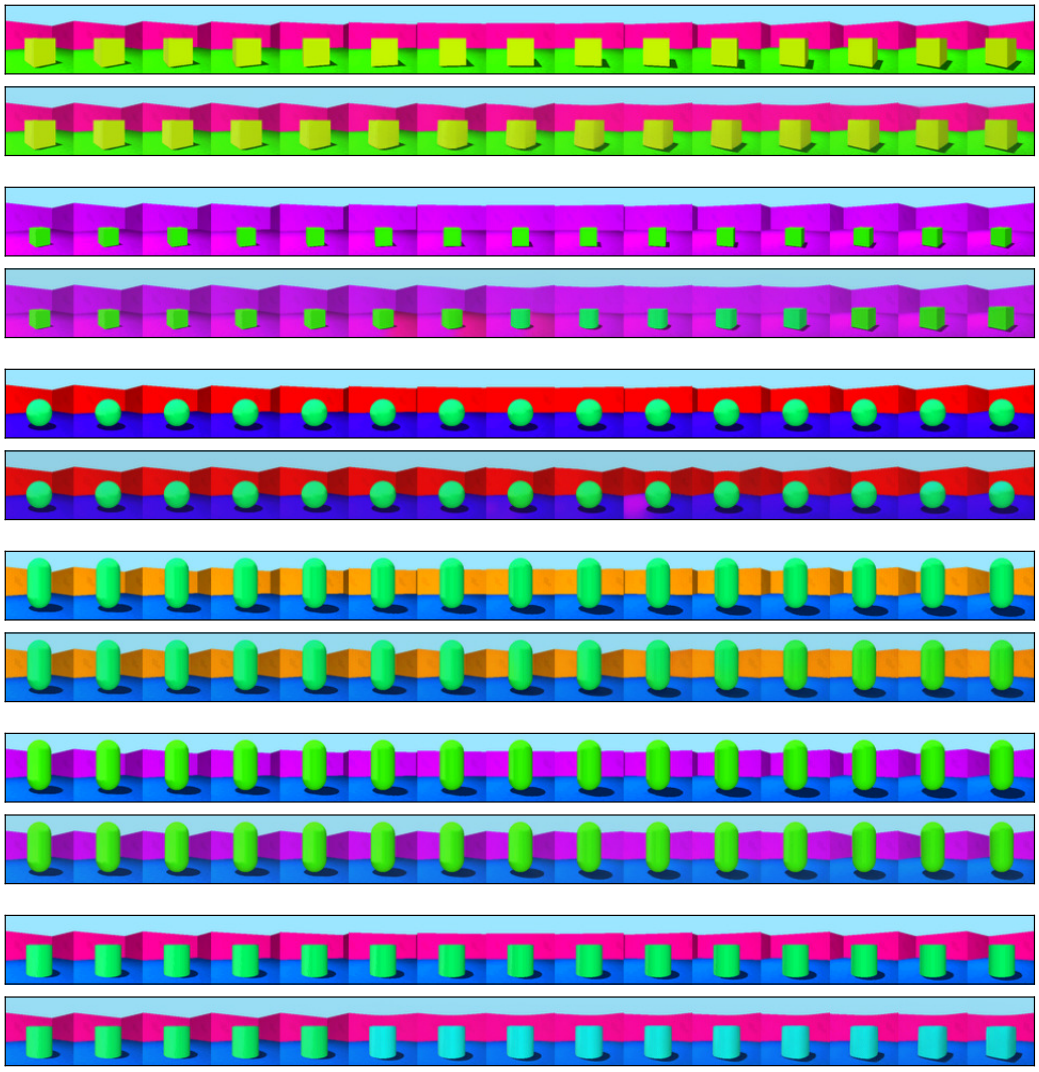


Figure 6: Sampled shapes at 15 different ground truth orientations. The original (above) and the corresponding reconstructions (below) obtained from interpolating along the generative model's attribute vector are grouped together.

A SIMILAR INDIVIDUALS

We provide more examples of points from the similarity set $S_i(\boldsymbol{x})$ for various inputs $\boldsymbol{x}$ randomly drawn from our evaluation test set. Fig. 6 illustrates the quality of the interpolation along the attribute vector with respect to the ground trouth. Fig. 7 and Fig. 8 visualize variations in the `Pale_Skin` attribute on CelebA, whereas Fig. 9 and Fig. 10 visualize variations of the `Young` attribute. The middle images in the figures correspond to reconstructions of the original inputs. The perturbations range equally from the $[-1, 1]$ range (going from left to right).

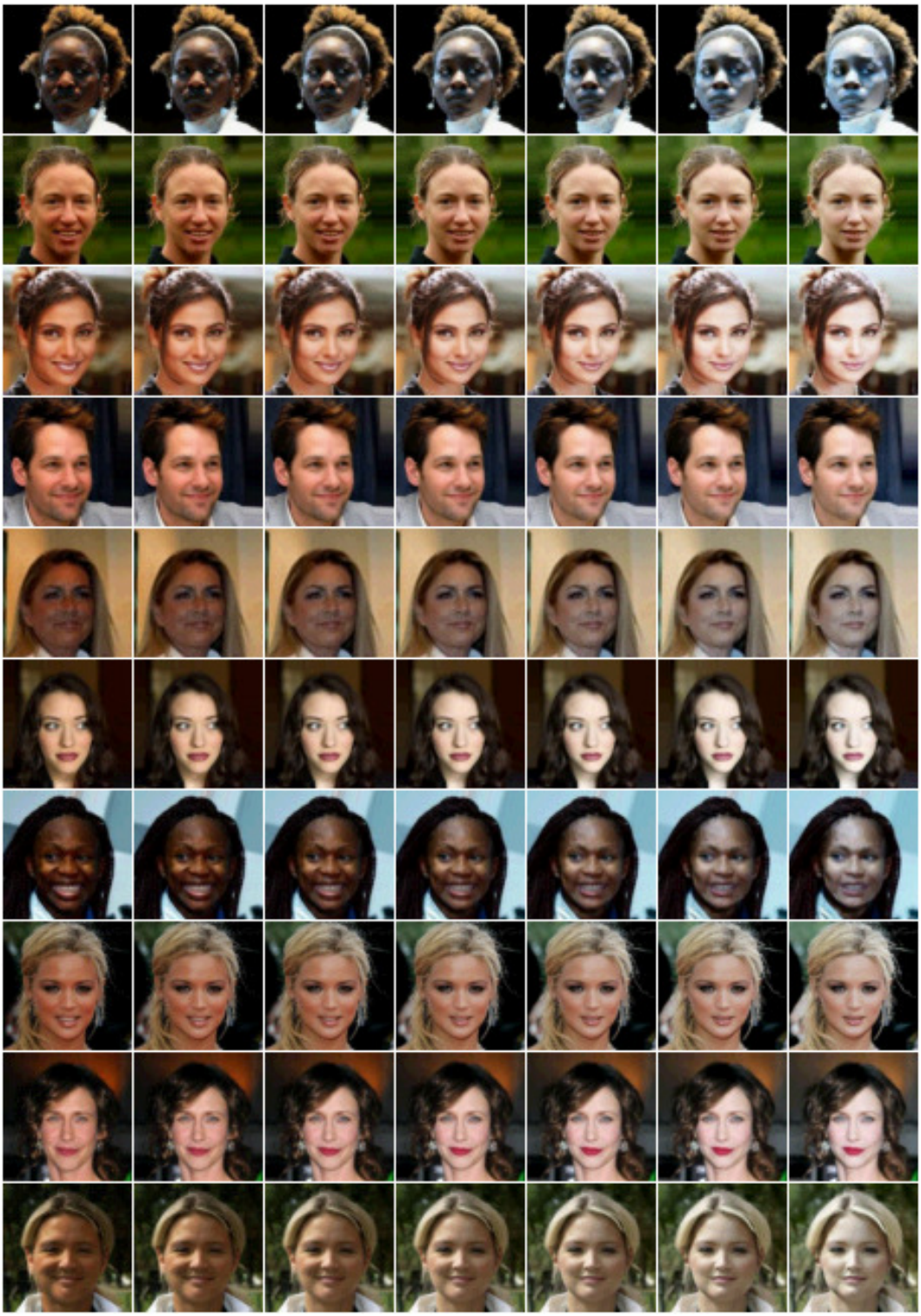


Figure 7: Varying the sensitive attribute `Pale_Skin`.

Figure 8: Varying the sensitive attribute `Pale_Skin`.

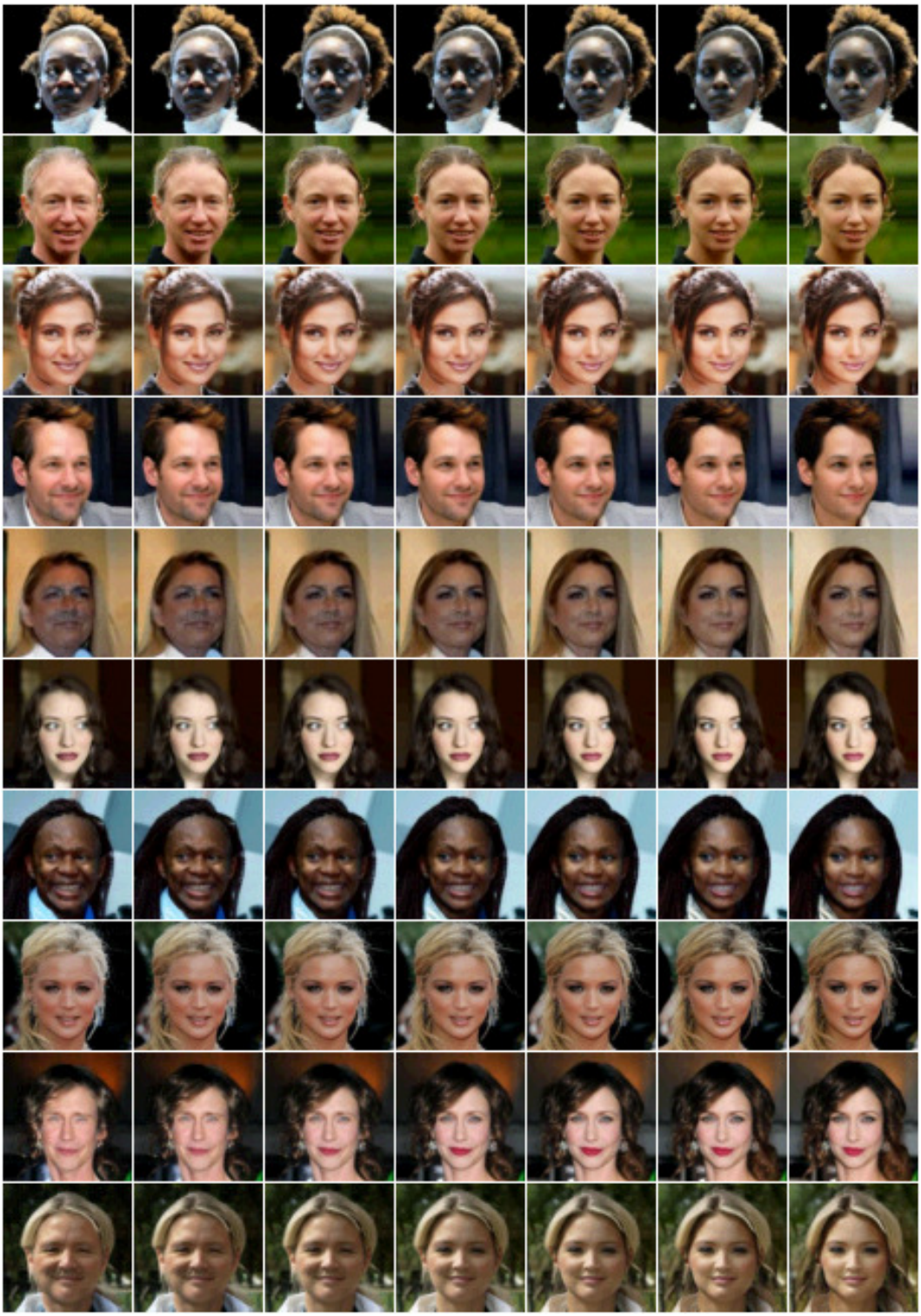


Figure 9: Varying the sensitive attribute `Young`.

Figure 10: Varying the sensitive attribute `Young`.