Latent Space Smoothing for Individually Fair Representations

Momchil Peychev[1], Anian Ruoss[2]*, Mislav Balunović[1], Maximilian Baader[1], and Martin Vechev[1]

[1]Department of Computer Science, ETH Zurich [2]DeepMind, London
{momchil.peychev,mislav.balunovic,mbaader,martin.vechev}@inf.ethz.ch
anianr@deepmind.com

Abstract. Fair representation learning transforms user data into a representation that ensures 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 enables us to learn individually fair representations that map similar individuals close together by using adversarial training to minimize the distance between their representations. 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 up to 90% without significantly affecting task utility.

Keywords: fair representation learning, individual fairness, smoothing

1 Introduction

Deep learning models are increasingly deployed in critical domains, such as face detection [74], credit scoring [38], or crime risk assessment [6], where decisions of the model can have wide-ranging impacts on society. Unfortunately, the models and datasets employed in these settings are biased [7,43], which raises concerns against their usage for such tasks and causes regulators to hold organizations accountable for the discriminatory effects of their models [18,19,22,23,77].

In this regard, fair representation learning [88] is a promising bias mitigation approach that transforms data to prevent discrimination regardless of the concrete downstream application while simultaneously maintaining high task utility. The approach is highly modular [60]: the *data regulator* defines the fairness notion, the *data producer* learns a fair representation that encodes the data, and the

[*] Work partially done while the author was at ETH Zurich.

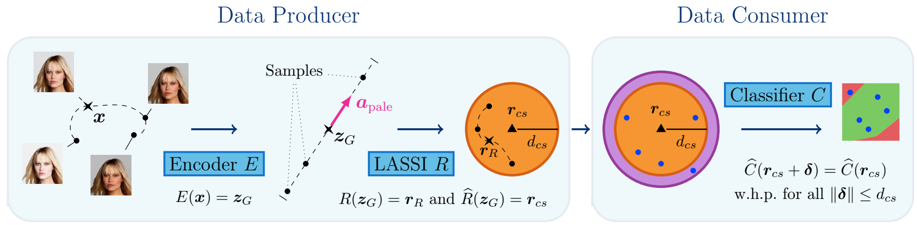

Data Producer
Data Consumer
$\boldsymbol{x}$
Encoder E
$E(\boldsymbol{x}) = \boldsymbol{z}_G$
Samples
$\boldsymbol{a}_{\text{pale}}$
$\boldsymbol{z}_G$
LASSI R
$R(\boldsymbol{z}_G) = \boldsymbol{r}_R$ and $\widehat{R}(\boldsymbol{z}_G) = \boldsymbol{r}_{cs}$
$\boldsymbol{r}_{cs}$
d_{cs}
$\boldsymbol{r}_R$
$\boldsymbol{r}_{cs}$
d_{cs}
Classifier C
$\widehat{C}(\boldsymbol{r}_{cs} + \boldsymbol{\delta}) = \widehat{C}(\boldsymbol{r}_{cs})$
w.h.p. for all $\|\boldsymbol{\delta}\| \leq d_{cs}$

Fig. 1: Overview of our framework 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 and center smoothing to compute a representation that provably maps all similar points into the ℓ_2-ball of radius d_{cs} around $\boldsymbol{r}_{cs}$. The right part shows the data consumer who can 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 exceeds d_{cs}.

data consumers employ the transformed data in downstream tasks. Recent work successfully augmented fair representation learning with guarantees [24,68], but its 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 of *individually* fair representation learning, which requires similar individuals to be treated similarly, are: (i) designing a suitable input similarity metric [86,88] and (ii) enforcing that similar individuals are *provably* treated similarly according to that metric. For low-dimensional tabular data, prior work has typically measured input similarity in terms of the input features (age, income, etc.), using, e.g., logical constraints [68] or weighted ℓ_p-metrics [85]. However, characterizing the similarity of high-dimensional data, such as images, at the input-level, e.g., by comparing pixels, is infeasible. Moreover, proving that all points in the infinite set of similar individuals obtain the same classification requires propagating this set through the model. Unfortunately, for high-dimensional applications this is unattainable for prior work using (mixed-integer) linear programming solvers [16,76], which only scale to small networks.

This work In this work, we introduce latent space smoothing for individually fair representations (LASSI), a method that addresses both of the above challenges. Our approach leverages two recent advances: the emergence of powerful generative models [41], which enable the definition of image similarity for individual fairness, and the scalable certification of deep models [10], which allows proving individual fairness. A high-level overview of our approach is shown in Fig. 1. Concretely, we use generative modeling [41] to enable data regulators to define input similarity by varying a continuous attribute of the image, such as pale skin in Fig. 1. To enforce that similar individuals are provably treated similarly, we further base our approach on smoothing: (i) the data producer uses center smoothing [44] to learn a representation that provably maps similar individuals close together, and (ii) the data consumer certifies local ℓ_2-robustness using randomized smoothing [10], thereby proving individual fairness of the end-to-end model. Therefore, our approach enables data regulators to impose fairness notions of the form: *"For a given person, all people differing only in skin tone should receive the same classification"* and allows data producers and consumers to independently learn a representation and classification models that provably enforce this notion.

To measure input similarity, the data producer leverages the ability of a bijective generative model to interpolate along the direction of an attribute vector in the latent space, which is impractical in the pixel space. As a result, the set of similar individuals can be defined by a line segment in the latent space (center part of the data producer in Fig. 1), corresponding to an elaborate curve in the input space (left part of the data producer in Fig. 1), which cannot be concisely captured by, e.g., an ℓ_p-ball. Thus, the data producer learns a representation R that maps all points of the latent line segment close together in the representation space by using adversarial training to minimize the distance between similar individuals. However, as adversarial training cannot provide guarantees on this maximum distance, the data producer uses center smoothing [44] to adjust the representation such that its *smoothed* version $\widehat{R}$ provably maps all similar points into an ℓ_2-ball of radius d_{cs} around a center $\boldsymbol{r}_{cs}$ with high probability (right part of the data producer in Fig. 1). Finally, the data consumer only needs to prove that the certified radius (violet in the data consumer part of Fig. 1) of its *smoothed* classifier $\widehat{C}$ around $\boldsymbol{r}_{cs}$ is larger than d_{cs} to obtain an individual fairness certificate for the end-to-end model $M := \widehat{C} \circ \widehat{R} \circ E$.

Our experimental evaluation on several image classification tasks shows that training with LASSI significantly increases the number of individuals for which we can certify individual fairness, with respect to multiple different sensitive attributes, as well as their combinations. Overall, we certify up to 90% more than the baselines. Furthermore, we demonstrate that the representations obtained by LASSI can be used to solve classification tasks that were unseen during training.

Main contributions We make the following contributions:

- A novel input similarity metric for high-dimensional data defined via interpolation in the latent space of generative models.
- A scalable representation learning method with individual fairness certification for models using high-dimensional data via randomized smoothing.
- A large-scale evaluation of our method on various image classification tasks.

2 Related Work

In this work, we consider individual fairness, which requires that similar individuals be treated similarly [14]. In contrast, group fairness enforces specific classification statistics to be equal across different groups of the population [14,28]. While both fairness notions are desirable, they also both suffer from certain shortcomings. For instance, models satisfying group fairness may still discriminate against individuals [14] or subgroups [36]. In contrast, the central challenge limiting practical adoption of individual fairness is the lack of a widely accepted similarity metric [86]. While recent work has made progress in developing similarity metrics for tabular data [31,57,62,79,87], defining similarity concisely for high-dimensional data remains challenging and is a key contribution of our work.

Fair representation learning A wide range of methods has been proposed to learn fair representations of user data. Most of these works consider group fairness and employ techniques such as adversarial learning [15,37,50,55], disentanglement [11,53,69], duality [73], low-rank matrix factorization [63], and distribution alignment [3,54,89]. Fair representation learning for individual fairness has recently gained attention, with similarity metrics based on logical formulas [68], Wasserstein distance [20,45], fairness graphs [46], and weighted ℓ_p-norms [88]. Unfortunately, none of these approaches can capture the 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 [81,83] and text [5,49,64,75], showing, e.g., that black women obtain lower accuracy in commercial face classification [7,43,66]. Importantly, these models not only learn but also amplify the biases of the training data [29,90], even for balanced datasets [80]. A key challenge for bias mitigation in high-dimensional settings is that, unlike tabular data, sensitive attributes such as age or skin tone are not directly encoded as features. Thus, prior work has often relied on generative models [2,12,13,33,39,40,47,48,67,70] or computer simulations [59] to manipulate these sensitive attributes and check whether the perturbed instances are classified the same. However, unlike our work, these methods only tested for bias empirically and do not provide fairness guarantees. Recent work also explored using generative models to define [27,84] or certify [61] robustness, but without focusing on fairness.

Fairness certification Regulatory agencies are increasingly holding organizations accountable for the discriminatory effects of their machine learning models [18,19,22,23,77]. Accordingly, designing algorithms with fairness guarantees has become an active area of research [1,3,4,9,24,71]. However, unlike our work, most approaches for individual fairness certification consider pretrained models and thus cannot be employed in fair representation learning [32,78,85]. In contrast, [68] 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 [68] scale to high-dimensional data, which is the primary focus of our work.

3 Background

This section provides the necessary background on individual fairness, fair representation learning, generative modeling, and randomized smoothing.

Individual fairness The seminal work of [14] defined individual fairness as "treating similar individuals similarly". In this work, we consider the concrete instantiation of this notion from [68]: an individual $\boldsymbol{x}'$ is similar to $\boldsymbol{x}$ with respect to a binary input similarity metric $\phi\colon \mathbb{R}^n \times \mathbb{R}^n \to \{0,1\}$ if and only if $\phi(\boldsymbol{x},\boldsymbol{x}') = 1$. A model $M\colon \mathbb{R}^n \to \mathcal{Y}$ is individually fair at $\boldsymbol{x} \in \mathbb{R}^n$ if it classifies all individuals similar to $\boldsymbol{x}$ (as measured by ϕ) the same, i.e.,

$$\forall \boldsymbol{x}' \in \mathbb{R}^n \colon \phi\left(\boldsymbol{x},\boldsymbol{x}'\right) \implies M\left(\boldsymbol{x}\right) = M\left(\boldsymbol{x}'\right). \tag{1}$$

For example, a credit rating algorithm is individually fair for a given person if all similar applicants (e.g., similar income and repayment history) receive the same credit rating. Our goal is to learn a model M that maximizes the number of points $\boldsymbol{x}$ from the distribution for which we can *guarantee* that Eq. (1) is satisfied. Defining a suitable input similarity metric ϕ is one of the key challenges limiting practical applications of individual fairness, and in Sec. 4.1 we will show how to employ generative modeling to overcome this obstacle for high-dimensional data.

Fair representation learning Fair representation learning [88] partitions the model $M\colon \mathbb{R}^n \to \mathcal{Y}$ into a data producer $P\colon \mathbb{R}^n \to \mathbb{R}^k$, which maps input points $\boldsymbol{x} \in \mathbb{R}^n$ into a representation space $\mathbb{R}^k$ that satisfies a given fairness notion while maintaining downstream utility, and a data consumer $C\colon \mathbb{R}^k \to \mathcal{Y}$ that solves a downstream task taking only the transformed data points $\boldsymbol{r} \coloneqq P\left(\boldsymbol{x}\right) \in \mathbb{R}^k$ as inputs. Importantly, the consumers (potentially indifferent to fairness) can employ standard training methods to obtain fair classifiers that are useful across a variety of different tasks. We base our approach on the LCIFR framework [68], which learns representations with individual fairness guarantees for low-dimensional tabular data. LCIFR defines a family of similarity notions and leverages (mixed-integer) linear programming methods for fairness certification. However, high-dimensional applications are out of reach for LCIFR because both the similarity notions and linear programming methods are tailored to low-dimensional tabular data. In particular, 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 at the pixel level. Moreover, while linear programming methods work well for small networks, they do not scale to real-world computer vision models. In this work, we show how to resolve these two key concerns to generalize the high-level idea of LCIFR to real-world, high-dimensional applications.

Generative modeling Normalizing flows, such as Glow [41], recently emerged as a promising generative modeling approach due to their exact log-likelihood evaluation, efficient inference and synthesis, and useful latent space for downstream tasks. Unlike GANs [25] or VAEs [42], normalizing flows are bijective models consisting of an encoder $E\colon \mathbb{R}^n \to \mathbb{R}^q$ and a decoder $D\colon \mathbb{R}^q \to \mathbb{R}^n$ for which $\boldsymbol{x} = D\left(E\left(\boldsymbol{x}\right)\right)$. Glow's input space $\mathbb{R}^n$ and latent space $\mathbb{R}^q$ have the same dimensionalities $n = q$. Its latent space captures important data attributes, thus enabling latent space interpolation such as changing the age 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 [2,13,33,39,48,67], Glow's key advantages are 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 image similarity by interpolating along the directions defined by certain sensitive attributes in the latent space.

Smoothing Unlike (mixed-integer) linear programming [16,76], smoothing approaches [10] can compute local robustness guarantees for any type of classifier $C\colon \mathbb{R}^k \to \mathcal{Y}$, regardless of its complexity and scale. To that end, [10] construct a smoothed classifier $\widehat{C}\colon \mathbb{R}^k \to \mathcal{Y}$, which returns the most probable classification of C for an input $\boldsymbol{r} \in \mathbb{R}^k$ when perturbed by random noise from $\mathcal{N}(0, \sigma_{rs}^2 I)$. Using a sampling-based approach, [10] establish a local robustness guarantee of the form: $\forall \boldsymbol{\delta} \in \mathbb{R}^k$ such that $\|\boldsymbol{\delta}\|_2 < d_{rs}$ we have $\widehat{C}\,(\boldsymbol{r}+\boldsymbol{\delta}) = \widehat{C}\,(\boldsymbol{r})$ with probability $1-\alpha_{rs}$, where α_{rs} can be made arbitrarily small. Thus, $\widehat{C}$ will classify all points in the ℓ_2-ball of radius d_{rs} around $\boldsymbol{r}$ the same with high probability. Recently, [44] introduced center smoothing, which extends this approach from classification to multidimensional regression. Concretely, for a function $R\colon \mathbb{R}^q \to \mathbb{R}^k$, center smoothing uses sampling and approximation to compute a smooth version $\widehat{R}\colon \mathbb{R}^q \to \mathbb{R}^k$, which maps $\boldsymbol{z} \in \mathbb{R}^q$ to the center point $\boldsymbol{r}_{cs} \coloneqq \widehat{R}\,(\boldsymbol{z})$ of a minimum enclosing ball containing at least half of the points $\boldsymbol{r}_i \sim R(\boldsymbol{z} + \mathcal{N}(0, \sigma_{cs}^2 I))$ for $i \in \{1, \dots, m\}$. Then, for $\epsilon > 0$ and $\forall \boldsymbol{z}' \in \mathbb{R}^q$ such that $\|\boldsymbol{z}-\boldsymbol{z}'\|_2 \leq \epsilon$, we have $\|\widehat{R}\,(\boldsymbol{z}) - \widehat{R}\,(\boldsymbol{z}')\,\|_2 \leq d_{cs}$ with probability at least $1-\alpha_{cs}$. That is, center smoothing computes a sound upper bound d_{cs} on the ℓ_2-ball of the function outputs of $\widehat{R}$ for all points in the ℓ_2-ball of radius ϵ around $\boldsymbol{z}$.

4 High-Dimensional Individually Fair Representations

In this section, we describe how our method defines a set of similar individuals (Sec. 4.1), learns individually fair representations for these points (Sec. 4.2), and finally, certifies individual fairness for them (Sec. 4.4). Our approach is general, but we focus on images for presentational purposes.

4.1 Similarity via a Generative Model

We consider two individuals $\boldsymbol{x}$ and $\boldsymbol{x}'$ to be similar if they differ only in their continuous sensitive attributes. However, semantic attributes, such as skin color, cannot be captured conveniently via the input features of $\boldsymbol{x}$. Thus, our key idea is to define similarity in the latent space of a generative model G. We compute a vector $\boldsymbol{a} \in \mathbb{R}^q$ associated with the sensitive 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. There is active research investigating different ways of computing $\boldsymbol{a}$ [13,30,41,48,67], and we will empirically show that our method is compatible with any such method.

Computing $\boldsymbol{a}$ We define individual similarity in the latent space of Glow [41]. Our method is independent of the actual computation of $\boldsymbol{a}$, which we demonstrate by instantiating four different attribute vector types. Let $\boldsymbol{z}_G = E(\boldsymbol{x})$ be the latent code of $\boldsymbol{x}$ in the generative latent space. First, following [41], we compute $\boldsymbol{a}$ by calculating the average latent vectors $\boldsymbol{z}_{G,pos}$ for samples with the attribute and $\boldsymbol{z}_{G,neg}$ for samples without it and set $\boldsymbol{a}$ to their difference, $\boldsymbol{a} = \boldsymbol{z}_{G,pos} - \boldsymbol{z}_{G,neg}$. Second, following [13], we train a linear classifier $\operatorname{sign}(\boldsymbol{a}^\top \boldsymbol{z}_G + b)$ to predict the presence of the attribute from $\boldsymbol{z}_G$ and take $\boldsymbol{a}$ to be the vector orthonormal to the decision boundary of the linear classifier. Finally, we employ [48] and [67] who build on these methods, accounting for the possible correlations between the sensitive and target attributes. In all cases, moving in one direction of $\boldsymbol{a}$ in the latent space increases the presence of the attribute and interpolating in the opposite direction decreases it. LASSI is independent of the sensitive attribute vector computation and will immediately benefit from all advancements in this area. We evaluate with vectors computed by [41] and [13] in the main paper (Sec. 5) and present further results with vectors from [48,67] in App. E.

Individual similarity in latent space Using the generative model G and the attribute vector $\boldsymbol{a}$, we define the set of individuals similar to $\boldsymbol{x}$ in the latent space of G as $S(\boldsymbol{x}) \coloneqq \{\boldsymbol{z}_G + t \cdot \boldsymbol{a} \mid |t| \leq \epsilon\} \subseteq \mathbb{R}^q$ (bottom of Fig. 2). Here, ϵ denotes the maximum perturbation level applied to the attribute. We consider G, $\boldsymbol{a}$, and ϵ to be a part of the similarity specification set by the data regulator. Crucially, $S(\boldsymbol{x})$ contains an infinite number of points but is compactly represented in the latent space of G as a line segment. In contrast, the same set represented directly in the input space, $S^{\text{in}}(\boldsymbol{x}) \coloneqq D(S(\boldsymbol{x})) \subseteq \mathbb{R}^n$, obtained by decoding the latent representations in $S(\boldsymbol{x})$ with D, cannot be abstracted conveniently (top of Fig. 2). Moreover, this approach for constructing $S(\boldsymbol{x})$ can be extended to multiple sensitive attributes by interpolating along their attribute vectors simultaneously. Referring back to the notation in Sec. 3, we formally define the input similarity metric ϕ to satisfy $\phi(\boldsymbol{x}, \boldsymbol{x}') \iff \boldsymbol{x}' \in S^{\text{in}}(\boldsymbol{x})$.

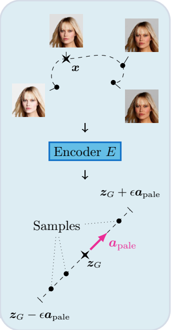

$\boldsymbol{x}$
Encoder E
$\boldsymbol{z}_G + \epsilon \boldsymbol{a}_{\text{pale}}$
Samples
$\boldsymbol{a}_{\text{pale}}$
$\boldsymbol{z}_G$
$\boldsymbol{z}_G - \epsilon \boldsymbol{a}_{\text{pale}}$

Fig. 2: Similarity in latent space.

4.2 Learning Individually Fair Representations

Assuming that the generative model $G = (E, D)$ is pretrained and given (e.g., by the data regulator), in this section we describe the learning of the representation $R\colon \mathbb{R}^q \to \mathbb{R}^k$, which maps from the generative latent space $\mathbb{R}^q$ directly to the representation space $\mathbb{R}^k$. The representation R is trained separately from the data consumer, the classifier C, whose training is explained in the next section.

Adversarial loss We encourage similar treatment for all points in $S^{\text{in}}(\boldsymbol{x})$ by training R to map them close to each other in $\mathbb{R}^k$, minimizing the loss

$$\mathcal{L}_{adv}(\boldsymbol{x}) = \max_{\boldsymbol{z}' \in S(\boldsymbol{x})} \|R(\boldsymbol{z}_G) - R(\boldsymbol{z}')\|_2. \tag{2}$$

Minimizing $\mathcal{L}_{adv}(\boldsymbol{x})$ is a min-max optimization problem, and adversarial training [56] is known to work well in such settings. Because the underlying domain of the inner maximization problem is simply the line segment $S(\boldsymbol{x})$, we perform a random adversarial attack in which we sample s points $\boldsymbol{z}_i \sim \mathcal{U}\left(S\left(\boldsymbol{x}\right)\right)$ uniformly at random from $S\left(\boldsymbol{x}\right)$ and approximate $\mathcal{L}_{adv}\left(\boldsymbol{x}\right) \approx \max_{i=1}^{s} \|R\left(\boldsymbol{z}_G\right) - R\left(\boldsymbol{z}_i\right)\|_2$. This efficient attack is typically more effective [17] than the first-order methods such as FGSM [26] and PGD [56] when the search space is low-dimensional.

Classification loss To ensure that the learned representations remain useful for downstream tasks, we introduce an auxiliary classifier C_{aux} to predict a ground truth target label y by adding an additional classification loss term:

$$\mathcal{L}_{cls}\left(\boldsymbol{x}, y\right) = \text{cross\_entropy}\left(C_{aux} \circ R\left(\boldsymbol{z}_G\right), y\right). \tag{3}$$

Reconstruction loss The downstream task may not always be known to the data producer a priori, and thus our representations should ideally transfer to a variety of such tasks. To that end, we optionally utilize a reconstruction loss, which is designed to preserve the signal from the original data [55,68]:

$$\mathcal{L}_{recon}\left(\boldsymbol{x}\right) = \|\boldsymbol{z}_G - Q\left(R\left(\boldsymbol{z}_G\right)\right)\|_2, \tag{4}$$

where $Q \colon \mathbb{R}^k \to \mathbb{R}^q$ denotes a reconstruction network.

The representation R, the auxiliary classifier C_{aux}, and the reconstruction network Q are trained jointly using stochastic gradient descent to minimize the combined objective

$$\lambda_1 \mathcal{L}_{cls}\left(\boldsymbol{x}, y\right) + \lambda_2 \mathcal{L}_{adv}\left(\boldsymbol{x}\right) + \lambda_3 \mathcal{L}_{recon}\left(\boldsymbol{x}\right). \tag{5}$$

Trading off fairness, accuracy, and transferability is a multi-objective optimization problem, an active area of research. Here, we follow [55,68] and use a linear scalarization scheme, with the hyperparameters λ_1, λ_2 and λ_3 balancing the three losses, but our method is also compatible with other schemes [51,58,82].

4.3 Training Classifier C

Once we have learned the representation R, we can use it to train any classifier C (often different from the auxiliary one C_{aux}). As we will apply smoothing to C, we train it by adding isotropic Gaussian noise to its inputs during the training process, as in [10]. We use the outputs of $R \circ E$ (and not the smoothed version $\widehat{R} \circ E$) as inputs to train C, since repeatedly smoothing the pipeline at this step is computationally expensive and because the distance between the smoothed and the unsmoothed outputs is generally small [44].

4.4 Certifying Individual Fairness via Latent Space Smoothing

With R and C trained as described above, we now construct the end-to-end model $M \colon \mathbb{R}^n \to \mathcal{Y}$ for which, given an input $\boldsymbol{x}$, we can certify individual fairness of the form

$$\forall \boldsymbol{x}' \in S^{\text{in}}\left(\boldsymbol{x}\right) : M\left(\boldsymbol{x}\right) = M\left(\boldsymbol{x}'\right), \tag{6}$$

with arbitrarily high probability.

Algorithm 1 Certifying the individual fairness of $\widehat{C} \circ \widehat{R} \circ E$ for the input $\boldsymbol{x}$.

```
function CERTIFY($E$, $R$, $C$, $\boldsymbol{x}$)
    Let $\boldsymbol{z}_G = E(\boldsymbol{x})$. Then, $\boldsymbol{r}_{cs} = \widehat{R}(\boldsymbol{z}_G)$ and $d_{cs}$ from center smoothing [44].
    if center smoothing abstained then return ABSTAIN
    Smooth $C$ [10]: obtain the certified radius $d_{rs}$ around $\boldsymbol{r}_{cs}$ (i.e., same classification)
    if $d_{cs} < d_{rs}$ then return CERTIFIED
    else return NOT CERTIFIED
```

Given a point $\boldsymbol{z}$ in the latent space of G, we define the function $g_{\boldsymbol{z}}(t) := R(\boldsymbol{z} + t \cdot \boldsymbol{a})$ for $t \in \mathbb{R}$. We apply the center smoothing procedure presented by [44] to obtain $\widehat{g_{\boldsymbol{z}}}$, the smoothed version of $g_{\boldsymbol{z}}$, and define $\widehat{R}(\boldsymbol{z}) := \widehat{g_{\boldsymbol{z}}}(0)$ such that for all $\boldsymbol{z}' \in S(\boldsymbol{x})$, $\|\widehat{R}(\boldsymbol{z}) - \widehat{R}(\boldsymbol{z}')\|_2 \leq d_{cs}$ (see Fig. 3). Next, we smooth the classifier C to obtain its ℓ_2-robustness radius d_{rs}. If $d_{cs} < d_{rs}$, then the end-to-end model $M = \widehat{C} \circ \widehat{R} \circ E$ certifiably satisfies individual fairness at $\boldsymbol{x}$ (as defined in Eq. (6)) with high probability. Concretely, if we instantiate center smoothing with confidence α_{cs} and randomized smoothing with confidence α_{rs}, then the individual fairness certificate holds with probability at least $1 - \alpha_{cs} - \alpha_{rs}$ (union bound). The compositional certification procedure is summarized in Alg. 1. Its correctness is formalized in Thm. 1 with a detailed proof in App. A.

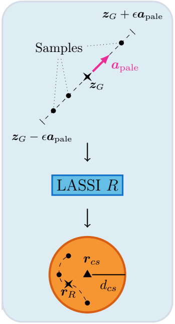

$\boldsymbol{z}_G + \epsilon \boldsymbol{a}_{\text{pale}}$
Samples
$\boldsymbol{a}_{\text{pale}}$
$\boldsymbol{z}_G$
$\boldsymbol{z}_G - \epsilon \boldsymbol{a}_{\text{pale}}$
LASSI R
$\boldsymbol{r}_{cs}$
d_{cs}
$\boldsymbol{r}_R$

Fig. 3: Center smoothing the similarity set.

Theorem 1. *Assume that we have a bijective generative model $G = (E, D)$ used to define the similarity set $S^{\text{in}}(\boldsymbol{x})$ for a given input $\boldsymbol{x}$. Let Alg. 1 perform center smoothing [44] with confidence $1 - \alpha_{cs}$ and randomized smoothing [10] with confidence $1 - \alpha_{rs}$. If Alg. 1 returns* CERTIFIED *for the input $\boldsymbol{x}$, then the end-to-end model $M = \widehat{C} \circ \widehat{R} \circ E$ is individually fair for $\boldsymbol{x}$ with respect to $S^{\text{in}}(\boldsymbol{x})$ with probability at least $1 - \alpha_{cs} - \alpha_{rs}$.*

5 Experiments

We now evaluate LASSI and present the key findings: (i) LASSI enforces individual fairness and keeps accuracy high, (ii) LASSI handles various sensitive attributes and attribute vectors, and (iii) LASSI representations transfer to unseen tasks.

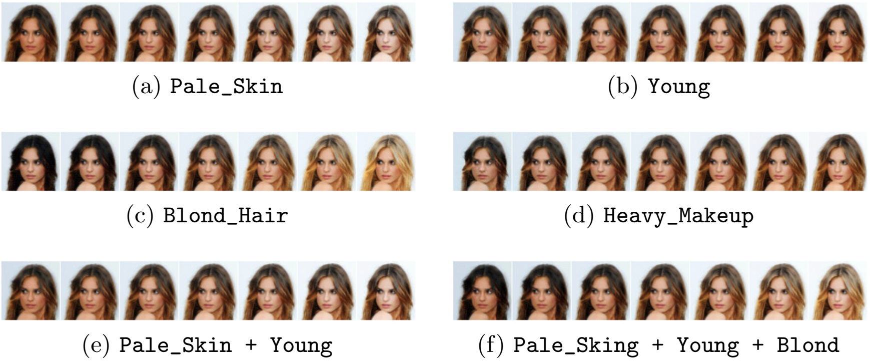
Datasets We evaluate LASSI on two datasets. CelebA [52] contains 202,599 aligned and cropped face images of real-world celebrities. The images are annotated with the presence or absence of 40 face attributes with various correlations between them [13]. As CelebA is highly imbalanced, we also experiment with FairFace [34]. It is balanced on race and contains 97,698 released images (padding 0.25) of individuals from 7 race and 9 age groups. We split the training set randomly (80:20 ratio) and evaluate on the validation set because the test set is not publicly shared. Further information about the datasets (including experimental "unfairness" of different attributes computed on CelebA) is in App. B.

Experimental setup The following setup is used for all experiments, unless stated otherwise. We use images of size 64×64, and for each dataset pretrain a Glow model G with 4 blocks of 32 flows, using an open-source PyTorch [65] implementation [72]. We use $\boldsymbol{a} = \boldsymbol{z}_{G,pos} - \boldsymbol{z}_{G,neg}$ and set $\epsilon = 1$ such that $S^{\text{in}}(\boldsymbol{x})$ contains realistic high-quality reconstructions (confirmed by manual inspection). Thus, the similarity specification (Sec. 4.1) for enforcing individual fairness is determined by G and the radius ϵ. We implement the representation R as a fully-connected network that propagates Glow's latent code of an input $\boldsymbol{x}$ through two hidden layers of sizes 2048 and 1024, mapping to a 512-dimensional space. The final layer applies zero mean and unit variance normalization ensuring that all components of R's output are in the same range when Gaussian noise is added during smoothing. A linear classifier C is used for predicting the target label.

Our fairness-unaware baseline (denoted as Naive) is standard representation learning of R without adversarial and reconstruction losses ($\lambda_2 = \lambda_3 = 0$). When training LASSI, we set the classification loss weight $\lambda_1 = 1$, except for the transfer learning experiments. A recent work [67] proposed generating synthetic images with a ProGAN [35] to balance the dataset. Their method is not concerned with individual fairness and their transformation of latent representations may change other, non-sensitive attributes. Nevertheless, we employ [67]'s high-level idea of augmenting the training set with synthetic samples from a generative model (Glow in our case). For each training sample $\boldsymbol{x}$, we synthesize and randomly sample s additional images from $S^{\text{in}}(\boldsymbol{x})$ in every epoch. Then, we proceed with representation learning of R on the augmented dataset. We denote this baseline, addapted to the individual fairness setting, as DataAug. We do not compare with LCIFR [68] as our individual similarity specifications cannot be directly encoded as logical formulas over the input features of $\boldsymbol{x}$ and because its certification is based on expensive solvers that do not scale to Glow and large models.

We list all selected hyperparameters for all experiments, based on an an extensive hyperparameter search on the validation sets, in App. C (details provided for the CelebA dataset). The hyperparameter study shows that LASSI works for a wide range of hyperparameter values and demonstrates that λ_2 controls the trade-off between accuracy and fairness. We report the accuracy and the certified individual fairness of the models measured on 312 samples from CelebA's test set (every 64-th) and 343 samples from FairFace's test set (every 32-nd). The certified fairness refers to the percentage of test samples for which Alg. 1 returns CERTIFIED, i.e., for which we can prove that Eq. (6) holds, guaranteeing that all similar individuals (according to our similarity definition) are classified the same. This metric is denoted as "Fair" in the tables. The evaluation of a single data point takes up to 6 seconds due to the sampling required by the smoothing procedures, which is why we do not report results on the whole test sets. We ran the experiments on GeForce RTX 2080 Ti GPUs and release all the code and models to reproduce our results at `https://github.com/eth-sri/lassi`.

Single sensitive attribute We experiment with 4 different continuous sensitive attributes from CelebA: `Pale_Skin`, `Young`, `Blond_Hair` and `Heavy_Makeup` on two tasks: predicting `Smiling` and `Earrings`. We chose attributes with different



(a) `Pale_Skin`

(b) `Young`

(c) `Blond_Hair`

(d) `Heavy_Makeup`

(e) `Pale_Skin + Young`

(f) `Pale_Sking + Young + Blond`

Fig. 4: Similar points from $S^{\text{in}}(\boldsymbol{x})$, as reconstructed by Glow, for multiple sensitive attribute combinations. Central images correspond to the original input. We vary t uniformly (left to right) in the $[-\frac{\epsilon}{\sqrt{n}}, \frac{\epsilon}{\sqrt{n}}]$ range, $n =$ number of sensitive attributes, $\epsilon = 1$. For $n > 1$, all attribute vectors are multiplied by the same t.

balance ratios that have been used in prior work [13], while avoiding attributes that perpetuate harmful stereotypes [13] (e.g., avoiding `Male`). Glow can also be used to generate discrete attributes, but then fairness certification can be done via enumeration because partial eyeglasses or hats, for example, are not plausible. Fig. 4 provides example images from $S^{\text{in}}(\boldsymbol{x})$ for a single $\boldsymbol{x}$. The `Earrings` task is considerably more imbalanced than `Smiling`, with 78.21% majority class accuracy on our test subset. Because of the high correlation between `Earrings` and `Makeup`, we run LASSI with increased λ_2 for this pair of attributes.

We show the results in Tab. 1 averaged over 5 runs with different random seeds. The results indicate that data augmentation helps, but is not enough. LASSI significantly improves the certified fairness, compared to the baselines, with a minor loss of accuracy on `Smiling` and even acts as a helpful regularizer on the imbalanced `Earrings` task. In App. D we report the standard deviations demonstrating that LASSI consistently enforces individual fairness with low variance and further evaluate empirical (i.e., non-certifiable) fairness metrics.

Multiple sensitive attributes In the next experiment, we combine the sensitive attributes `Pale_Skin`, `Young` and `Blond_Hair` and predict `Smiling`. The similarity sets w.r.t. which we certify individual fairness are defined as $S(\boldsymbol{x}) = \{E(\boldsymbol{x}) + \sum_i t_i \cdot \boldsymbol{a}_i \mid \|\boldsymbol{t}\|_2 \leq \epsilon\}$. The results in Tab. 1 (rows 5 – 6) show that the certified fairness drops as the similarity sets become more complex, as expected, but LASSI still successfully enforces individual fairness in these cases.

Larger images and different attribute vectors Next, we explore if LASSI can also work with larger images. We increase the dimensionality of the CelebA images to 128×128, pretrain Glow with 5 blocks and keep the rest of the hyperparameters the same. The results are consistent with those already presented in Tab. 1: LASSI increases the certified individual fairness by up to 77% on the

Table 1: Evaluation of LASSI on the CelebA dataset, showing that LASSI significantly increases certified individual fairness compared to the baselines without affecting the classification accuracy, even increasing it for imbalanced tasks. Reported means averaged over 5 runs, see App. D for standard deviations.

		Naive		DataAug		LASSI (ours)	
Task	Sensitive attribute(s)	Acc	Fair	Acc	Fair	Acc	Fair
`Smiling`	`Pale_Skin`	**86.3**	0.6	85.7	12.2	85.9	**98.0**
	`Young`	**86.3**	38.2	85.9	43.0	**86.3**	**98.8**
	`Blond_Hair`	86.3	3.4	**86.6**	9.4	86.4	**94.7**
	`Heavy_Makeup`	**86.3**	0.4	85.3	13.7	85.6	**91.3**
	`Pale+Young`	**86.0**	0.4	85.8	9.9	85.8	**97.3**
	`Pale+Young+Blond`	86.2	0.0	**86.4**	3.6	85.5	**86.5**
`Earrings`	`Pale_Skin`	81.3	24.3	81.0	40.4	**85.0**	**98.5**
	`Young`	81.4	59.2	79.9	72.0	**84.5**	**98.0**
	`Blond_Hair`	81.4	9.2	82.2	30.5	**84.8**	**96.2**
	`Heavy_Makeup`	81.6	20.5	80.3	49.2	**82.3**	**98.7**

`Smiling` task (see App. D for detailed results). We also instantiate LASSI with the alternative attribute vector type [13] introduced in Sec. 4.1 (with $\epsilon = 10$). Although interpolating along the vector which is perpendicular to the linear decision boundary of the sensitive attribute possibly reduces the correlations leaked into the similarity sets, Tab. 2 shows that LASSI still improves the certified fairness by up to 16% compared to the baselines. This improvement is 9.7% and 6.1% for the attribute vectors proposed by [67] and [48] respectively, further demonstrating that LASSI can be useful for various attribute vector types. More details about these experiments are provided in App. E.

Transfer learning To demonstrate the modularity of our approach, we show that LASSI can learn fair and transferable representations which are useful for unseen downstream tasks. To that end, we turn off the classification loss, consistent with prior work [55] ($\lambda_1 = 0$, i.e., the representation R is trained unsupervised), and enable the reconstruction loss ($\lambda_3 = 0.1$). The reconstruction network Q has an architecture symmetric to that of R. In Tab. 3 we report the accuracies and the certified fairness on 7 different, relatively well-balanced, downstream tasks. The models perform slightly worse compared to the case where the downstream task is known in advance, but the obtained certified individual fairness is still consistently high – more than 80% for the most complex similarity specification (`P+Y+B`) and above 90% for the simpler ones. Standard deviations and baseline accuracies on these tasks are reported in App. D.

Training on FairFace dataset To verify that LASSI works well in different settings, we also evaluate on the FairFace [34] dataset. We select `Race=Black` as a sensitive attribute and predict `Age`. This is a very challenging multi-class task

Table 2: Evaluation with $\boldsymbol{a}$ perpendicular to the linear decision boundary of the sensitive attribute [13] (Sec. 4.1) on the Smiling task, showing that LASSI is not limited to a specific attribute vector type.

	Naive		DataAug		LASSI (ours)	
Sensitive attribute(s)	Acc	Fair	Acc	Fair	Acc	Fair
Pale_Skin	86.4	34.0	85.9	90.3	**86.5**	**98.8**
Young	86.3	73.1	86.2	90.3	**86.8**	**97.9**
Blond_Hair	86.2	71.4	86.1	88.8	**86.7**	**98.8**
Heavy_Makeup	86.2	11.5	86.3	87.4	**86.8**	**98.8**
Pale+Young	86.2	28.6	85.8	84.7	**86.5**	**98.6**
Pale+Young+Blond	86.2	23.7	85.9	82.2	**86.4**	**98.7**

Table 3: Transfer learning results, demonstrating that LASSI can still achieve high certified individual fairness even when the downstream tasks are not known.

Sens. attrib.:	Pale (P)		Young (Y)		Blond (B)		P + Y		P + Y + B	
Transfer task	Acc	Fair	Acc	Fair	Acc	Fair	Acc	Fair	Acc	Fair
Smiling	86.2	93.1	86.0	95.4	85.1	93.8	85.9	92.2	85.1	87.0
High_Cheeks	81.7	92.6	82.3	96.0	81.3	92.2	80.8	93.0	80.6	84.5
Mouth_Open	81.5	91.2	82.4	94.3	82.4	87.5	81.6	90.1	82.5	80.8
Lipstick	88.3	94.0	85.8	95.8	86.8	91.2	85.1	90.6	86.2	81.0
Heavy_Makeup	86.5	93.0	83.5	95.3	85.6	89.3	83.7	90.0	83.3	80.4
Wavy_Hair	79.2	93.3	77.5	95.8	78.0	91.3	77.6	91.5	78.8	85.3
Eyebrows	78.3	92.1	78.3	94.7	78.9	89.6	77.8	92.2	78.7	85.6

with around 60% state of the art accuracy. Therefore, we create two easier tasks: Age-2, predicting if an individual is younger or older than 30, and Age-3 with three target ranges: $[0-19]$, $[20-39]$, and 40+. Tab. 4 reports the results for $\epsilon = 0.5$. We verify that transfer learning also works in this setup by training on Age-2 and then transferring the representations to all three tasks. As the tasks are related, increasing the classification loss weight λ_1 on the base task from 0 to 0.01, increases both the transfer downstream accuracy and the certified fairness. The highest certified fairness is generally obtained when the downstream task is known and the model is trained on it (LASSI, $\lambda_1 = 1$).

6 Limitations and Future Work

We now discuss some of the limitations of LASSI. First, our method trains individually fair models, but it does not guarantee that models satisfy other

Table 4: Results on FairFace, showing that LASSI can significantly improve the certified individual fairness even on balanced datasets. The adversarial loss weight is $\lambda_2 = 0.1$ for all models except Naive, the transfer models are trained on `Age-2` with reconstruction loss weight $\lambda_3 = 0.1$. LASSI is trained on the corresponding tasks with adversarial but without reconstruction loss ($\lambda_1 = 1$, $\lambda_3 = 0$).

	Naive		DataAug		Transfer$_{\lambda_1=0}$		Transfer$_{\lambda_1=0.01}$		LASSI	
Task	Acc	Fair	Acc	Fair	Acc	Fair	Acc	Fair	Acc	Fair
`Age-2`	69.0	5.7	68.9	4.8	66.4	91.7	**74.9**	91.7	72.0	**95.0**
`Age-3`	67.0	0.0	67.1	0.6	63.0	85.6	**67.7**	88.0	65.1	**90.8**
`Age` (all)	**42.2**	0.0	39.9	0.0	34.3	72.0	37.1	**77.5**	41.5	65.9

fairness notions, e.g., group fairness. While individual fairness is a well-studied research area, recent work argues that it does not qualify as a valid fairness notion as it can be insufficient to guarantee fairness in certain instances and risks encoding implicit human biases [21]. Moreover, the validity of our fairness certificates depends heavily on the generative model used by LASSI. In particular, the similarity sets $S(\boldsymbol{x})$ considered in our work may not be exhaustive enough as there can be latent points outside $S(\boldsymbol{x})$ that correspond to input points that would be perceived as similar to $\boldsymbol{x}$ by a human observer. This can also happen if the generative model is not powerful enough to generate all possible instances and combinations of similar individuals. For the above reasons, it is hard to obtain formal guarantees about G and the computed certificates may not always transfer from G to the real world. We explore this issue further in App. F where we experiment with 3D Shapes [8], a procedurally generated dataset with known ground truth similarity sets. Future work can consider addressing these challanges by performing extensive manual human inspection of reconstructions produced by G (similar to App. G). Moreover, all future advancements in the active research area of normalizing flows will immediately improve the quality of our certificates.

7 Conclusion

We proposed LASSI, which defines image similarity with respect to a generative model via attribute manipulation, allowing us to capture complex image transformations 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 on several datasets and illustrates the practicality of our approach.

Acknowledgments We thank Seyedmorteza Sadat for his help with preliminary investigations and the anonymous reviewers for their insightful feedback.

References

1. Albarghouthi, A., D'Antoni, L., Drews, S., Nori, A.V.: Fairsquare: probabilistic verification of program fairness. Proc. ACM Program. Lang. (2017)
2. Balakrishnan, G., Xiong, Y., Xia, W., Perona, P.: Towards causal benchmarking of bias in face analysis algorithms. In: Computer Vision - ECCV 2020 - 16th European Conference (2020)
3. Balunovic, M., Ruoss, A., Vechev, M.T.: Fair normalizing flows. CoRR (2021)
4. Bastani, O., Zhang, X., Solar-Lezama, A.: Probabilistic verification of fairness properties via concentration. Proc. ACM Program. Lang. (2019)
5. Bolukbasi, T., Chang, K., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: Advances in Neural Information Processing Systems 29 (2016)
6. Brennan, T., Dieterich, W., Ehret, B.: Evaluating the predictive validity of the compas risk and needs assessment system. Criminal Justice and Behavior (2009)
7. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on Fairness, Accountability and Transparency (2018)
8. Burgess, C., Kim, H.: 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/ (2018)
9. Choi, Y., Dang, M., den Broeck, G.V.: Group fairness by probabilistic modeling with latent fair decisions. In: Thirty-Fifth AAAI Conference on Artificial Intelligence (2021)
10. Cohen, J.M., Rosenfeld, E., Kolter, J.Z.: Certified adversarial robustness via randomized smoothing. In: Proceedings of the 36th International Conference on Machine Learning (2019)
11. Creager, E., Madras, D., Jacobsen, J., Weis, M.A., Swersky, K., Pitassi, T., Zemel, R.S.: Flexibly fair representation learning by disentanglement. In: Proceedings of the 36th International Conference on Machine Learning (2019)
12. Dash, S., Sharma, A.: Counterfactual generation and fairness evaluation using adversarially learned inference. CoRR (2020)
13. Denton, E., Hutchinson, B., Mitchell, M., Gebru, T.: Detecting bias with generative counterfactual face attribute augmentation. CoRR (2019)
14. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.S.: Fairness through awareness. In: Innovations in Theoretical Computer Science (2012)
15. Edwards, H., Storkey, A.J.: Censoring representations with an adversary. In: 4th International Conference on Learning Representations (2016)
16. Ehlers, R.: Formal verification of piece-wise linear feed-forward neural networks. In: Automated Technology for Verification and Analysis - 15th International Symposium (2017)
17. Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., Madry, A.: Exploring the landscape of spatial robustness. In: Proceedings of the 36th International Conference on Machine Learning (2019)
18. EU: Ethics guidelines for trustworthy ai (2019)
19. 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)
20. Feng, R., Yang, Y., Lyu, Y., Tan, C., Sun, Y., Wang, C.: Learning fair representations via an adversarial framework. CoRR (2019)

21. Fleisher, W.: What's fair about individual fairness? In: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event (2021)
22. FTC: Using artificial intelligence and algorithms (2020)
23. FTC: Aiming for truth, fairness, and equity in your company's use of ai (2021)
24. Gitiaux, X., Rangwala, H.: Learning smooth and fair representations. In: The 24th International Conference on Artificial Intelligence and Statistics (2021)
25. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems 27 (2014)
26. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: 3rd International Conference on Learning Representations (2015)
27. Gowal, S., Qin, C., Huang, P., Cemgil, A.T., Dvijotham, K., Mann, T.A., Kohli, P.: Achieving robustness in the wild via adversarial mixing with disentangled representations. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
28. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems 29 (2016)
29. Hendricks, L.A., Burns, K., Saenko, K., Darrell, T., Rohrbach, A.: Women also snowboard: Overcoming bias in captioning models. In: Computer Vision - ECCV 2018 - 15th European Conference (2018)
30. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. In: 5th International Conference on Learning Representations (2017)
31. Ilvento, C.: Metric learning for individual fairness. In: 1st Symposium on Foundations of Responsible Computing (2020)
32. John, P.G., Vijaykeerthy, D., Saha, D.: Verifying individual fairness in machine learning models. In: Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence (2020)
33. Joo, J., Kärkkäinen, K.: Gender slopes: Counterfactual fairness for computer vision models by attribute manipulation. CoRR (2020)
34. Kärkkäinen, K., Joo, J.: Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: IEEE Winter Conference on Applications of Computer Vision (2021)
35. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018), https://openreview.net/forum?id=Hk99zCeAb
36. Kearns, M.J., Neel, S., Roth, A., Wu, Z.S.: Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In: Proceedings of the 35th International Conference on Machine Learning (2018)
37. Kehrenberg, T., Bartlett, M., Thomas, O., Quadrianto, N.: Null-sampling for interpretable and fair representations. In: Computer Vision - ECCV 2020 - 16th European Conference (2020)
38. Khandani, A.E., Kim, A.J., Lo, A.W.: Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance (2010)
39. Kim, B., Wattenberg, M., Gilmer, J., Cai, C.J., Wexler, J., Viégas, F.B., Sayres, R.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: Proceedings of the 35th International Conference on Machine Learning (2018)

40. Kim, H., Shin, S., Jang, J., Song, K., Joo, W., Kang, W., Moon, I.: Counterfactual fairness with disentangled causal effect variational autoencoder. In: Thirty-Fifth AAAI Conference on Artificial Intelligence (2021)
41. Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. In: Advances in Neural Information Processing Systems 31 (2018)
42. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: 2nd International Conference on Learning Representations (2014)
43. Klare, B., Burge, M.J., Klontz, J.C., Bruegge, R.W.V., Jain, A.K.: Face recognition performance: Role of demographic information. IEEE Trans. Inf. Forensics Secur. (2012)
44. Kumar, A., Goldstein, T.: Center smoothing: Certified robustness for networks with structured outputs. Advances in Neural Information Processing Systems 34 (2021)
45. Lahoti, P., Gummadi, K.P., Weikum, G.: ifair: Learning individually fair data representations for algorithmic decision making. In: 35th IEEE International Conference on Data Engineering (2019)
46. Lahoti, P., Gummadi, K.P., Weikum, G.: Operationalizing individual fairness with pairwise fair representations. Proc. VLDB Endow. (2019)
47. Lang, O., Gandelsman, Y., Yarom, M., Wald, Y., Elidan, G., Hassidim, A., Freeman, W.T., Isola, P., Globerson, A., Irani, M., Mosseri, I.: Explaining in style: Training a gan to explain a classifier in stylespace. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 693–702 (October 2021)
48. Li, Z., Xu, C.: Discover the unknown biased attribute of an image classifier. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14970–14979 (October 2021)
49. Liang, P.P., Wu, C., Morency, L., Salakhutdinov, R.: Towards understanding and mitigating social biases in language models. In: Proceedings of the 38th International Conference on Machine Learning (2021)
50. Liao, J., Huang, C., Kairouz, P., Sankar, L.: Learning generative adversarial representations (GAP) under fairness and censoring constraints. CoRR (2019)
51. Lin, X., Zhen, H., Li, Z., Zhang, Q., Kwong, S.: Pareto multi-task learning. In: Advances in Neural Information Processing Systems 32 (2019)
52. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: IEEE International Conference on Computer Vision (2015)
53. Locatello, F., Abbati, G., Rainforth, T., Bauer, S., Schölkopf, B., Bachem, O.: On the fairness of disentangled representations. In: Advances in Neural Information Processing Systems 32 (2019)
54. Louizos, C., Swersky, K., Li, Y., Welling, M., Zemel, R.S.: The variational fair autoencoder. In: 4th International Conference on Learning Representations (2016)
55. Madras, D., Creager, E., Pitassi, T., Zemel, R.S.: Learning adversarially fair and transferable representations. In: Proceedings of the 35th International Conference on Machine Learning (2018)
56. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: 6th International Conference on Learning Representations (2018)
57. Maity, S., Xue, S., Yurochkin, M., Sun, Y.: Statistical inference for individual fairness. In: 9th International Conference on Learning Representations (2021)
58. Martínez, N., Bertrán, M., Sapiro, G.: Minimax pareto fairness: A multi objective perspective. In: Proceedings of the 37th International Conference on Machine Learning (2020)
59. McDuff, D.J., Cheng, R., Kapoor, A.: Identifying bias in AI using simulation. CoRR (2018)

60. McNamara, D., Ong, C.S., Williamson, R.C.: Costs and benefits of fair representation learning. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (2019)
61. Mirman, M., Hägele, A., Bielik, P., Gehr, T., Vechev, M.T.: Robustness certification with generative models. In: 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation (2021)
62. Mukherjee, D., Yurochkin, M., Banerjee, M., Sun, Y.: Two simple ways to learn individual fairness metrics from data. In: Proceedings of the 37th International Conference on Machine Learning (2020)
63. Oneto, L., Donini, M., Pontil, M., Maurer, A.: Learning fair and transferable representations with theoretical guarantees. In: 7th IEEE International Conference on Data Science and Advanced Analytics (2020)
64. Park, J.H., Shin, J., Fung, P.: Reducing gender bias in abusive language detection. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (2018)
65. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E.Z., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32 (2019)
66. Raji, I.D., Buolamwini, J.: 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)
67. Ramaswamy, V.V., Kim, S.S.Y., Russakovsky, O.: Fair attribute classification through latent space de-biasing. In: IEEE Conference on Computer Vision and Pattern Recognition (2021)
68. Ruoss, A., Balunovic, M., Fischer, M., Vechev, M.T.: Learning certified individually fair representations. In: Advances in Neural Information Processing Systems 33 (2020)
69. Sarhan, M.H., Navab, N., Eslami, A., Albarqouni, S.: Fairness by learning orthogonal disentangled representations. In: Computer Vision - ECCV 2020 - 16th European Conference (2020)
70. Sattigeri, P., Hoffman, S.C., Chenthamarakshan, V., Varshney, K.R.: Fairness GAN: generating datasets with fairness properties using a generative adversarial network. IBM J. Res. Dev. (2019)
71. Segal, S., Adi, Y., Pinkas, B., Baum, C., Ganesh, C., Keshet, J.: Fairness in the eyes of the data: Certifying machine-learning models. In: AAAI/ACM Conference on AI, Ethics, and Society (2021)
72. Seonghyeon, K.: Glow pytorch (commit: 97081ff1). https://github.com/rosinality/glow-pytorch (2020)
73. Song, J., Kalluri, P., Grover, A., Zhao, S., Ermon, S.: Learning controllable fair representations. In: The 22nd International Conference on Artificial Intelligence and Statistics (2019)
74. Sun, X., Wu, P., Hoi, S.C.H.: Face detection using deep learning: An improved faster RCNN approach. Neurocomputing (2018)
75. Tatman, R.: Gender and dialect bias in youtube's automatic captions. In: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing (2017)

76. Tjeng, V., Xiao, K.Y., Tedrake, R.: Evaluating robustness of neural networks with mixed integer programming. In: 7th International Conference on Learning Representations (2019)
77. UN: The right to privacy in the digital age (2021)
78. Urban, C., Christakis, M., Wüstholz, V., Zhang, F.: Perfectly parallel fairness certification of neural networks. Proc. ACM Program. Lang. (2020)
79. Wang, H., Grgic-Hlaca, N., Lahoti, P., Gummadi, K.P., Weller, A.: An empirical study on learning fairness metrics for COMPAS data with human supervision. CoRR (2019)
80. Wang, T., Zhao, J., Yatskar, M., Chang, K., Ordonez, V.: Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In: IEEE/CVF International Conference on Computer Vision (2019)
81. Wang, Z., Qinami, K., Karakozis, I.C., Genova, K., Nair, P., Hata, K., Russakovsky, O.: Towards fairness in visual recognition: Effective strategies for bias mitigation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
82. Wei, S., Niethammer, M.: The fairness-accuracy pareto front. CoRR (2020)
83. Wilson, B., Hoffman, J., Morgenstern, J.: Predictive inequity in object detection. CoRR (2019)
84. Wong, E., Kolter, J.Z.: Learning perturbation sets for robust machine learning. In: 9th International Conference on Learning Representations (2021)
85. Yeom, S., Fredrikson, M.: Individual fairness revisited: Transferring techniques from adversarial robustness. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (2020)
86. Yurochkin, M., Bower, A., Sun, Y.: Training individually fair ML models with sensitive subspace robustness. In: 8th International Conference on Learning Representations (2020)
87. Yurochkin, M., Sun, Y.: Sensei: Sensitive set invariance for enforcing individual fairness. In: 9th International Conference on Learning Representations (2021)
88. Zemel, R.S., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: Proceedings of the 30th International Conference on Machine Learning (2013)
89. Zhao, H., Coston, A., Adel, T., Gordon, G.J.: Conditional learning of fair representations. In: 8th International Conference on Learning Representations (2020)
90. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.: 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)