Ethics Statement

This work proposed a novel method for certifying the individual fairness of models operating on high-dimensional data. Progress on this challenging problem could enable fairness auditing for high-risk computer vision applications, such as facial recognition. Recent work [74] argues that facial recognition algorithms can have undesirable, socially toxic, and divisive consequences. For instance, it was demonstrated that they may perpetuate and reinforce racial and gender bias [7,13]. Therefore, they must be applied carefully, considering the social dynamics and context in which they occur. Accordingly, following prior work [13], we refrained from using unstable social constructs, such as gender, or normatively judgemental attributes, such as "attractive" or "chubby", in our research.

One way to limit the potential harms of facial analysis technologies is to control and regulate their usage. Our work aims to help fill this gap by presenting a methodology for enforcing individual fairness via certification. As highlighted in our paper, we acknowledge that the quality of the generative models is a significant bottleneck of our certificates. For example, they may encode various biases present in the data. Another possible source of bias is the human perception and social constructs which can potentially impact the validity of our similarity specifications. Nevertheless, we believe that we can still leverage generative models and their latent space to construct more meaningful individual fairness specifications on high-dimensional data than those allowed by prior work. More broadly, developing rigorous, standardized processes for auditing and certifying facial recognition models (including human inspection, e.g., by considering the reconstructed images in App. G) should complement the contributions presented in our work. Finally, future quality advancements in generative modelling and normalizing flows can directly translate into stronger guarantees of our method, enabling certified fair application of models using rich, high-dimensional data.

A Proof of Thm. 1

This section provides a formal proof of the following:

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 $1 - \alpha_{cs} - \alpha_{rs}$.*

To prove Thm. 1, we will make use of the following randomized and center smoothing theorems proved in the literature:

Theorem 2 (Adapted from [10]). *Let $C \colon \mathbb{R}^k \to \mathcal{Y}$ be a classifier and let $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma_{rs}^2 I)$. Let $\widehat{C}$ be defined such that $\widehat{C}(\boldsymbol{r}) = \arg\max_{c \in \mathcal{Y}} \mathbb{P}_{\boldsymbol{\varepsilon}}(C(\boldsymbol{r} + \boldsymbol{\varepsilon}) = c)$.*

Suppose $c_A \in \mathcal{Y}$ *and* $\underline{p_A}, \overline{p_B} \in [0,1]$ *satisfy:*

$$\mathbb{P}_{\boldsymbol{\varepsilon}}(C(\boldsymbol{r}+\boldsymbol{\varepsilon}) = c_A) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c_B \neq c_A} \mathbb{P}_{\boldsymbol{\varepsilon}}(C(\boldsymbol{r}+\boldsymbol{\varepsilon}) = c_B). \tag{7}$$

Then $\widehat{C}(\boldsymbol{r}+\boldsymbol{\delta}) = c_A$ *for all* $\boldsymbol{\delta}$ *satisfying* $\|\boldsymbol{\delta}\|_2 < d_{rs}$*, where* $d_{rs} \coloneqq \frac{\sigma_{rs}}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B}))$.

Here, $\mathcal{Y}$ denotes the set of class labels, Φ is the cumulative distribution function (CDF) of the standard normal distribution $\mathcal{N}(0,1)$, and Φ^{-1} is its inverse.

Theorem 3 (Adapted from [44]). *Let* $g\colon \mathbb{R}^a \to \mathbb{R}^k$ *and* $\hat{g}\colon \mathbb{R}^a \to \mathbb{R}^k$ *is an approximation of the smoothed version of* g*, which maps* $\boldsymbol{t} \in \mathbb{R}^a$ *to the center point* $\hat{g}(\boldsymbol{t})$ *of a minimum enclosing ball containing at least half of the points* $\boldsymbol{r}_i \sim g(\boldsymbol{t} + \mathcal{N}(0, \sigma_{cs}^2 I))$*,* $i \in \{1, \ldots, m\}$*. Then, for* $\epsilon > 0$*, with probability at least* $1 - \alpha_{cs}$ *we have,*

$$\forall \boldsymbol{t}' \text{ s.t. } \|\boldsymbol{t} - \boldsymbol{t}'\|_2 \leq \epsilon, \|\hat{g}(\boldsymbol{t}) - \hat{g}(\boldsymbol{t}')\|_2 \leq d_{cs}. \tag{8}$$

We now proceed to proving Thm. 1:

Proof. Assume that Alg. 1 returns CERTIFIED for the input $\boldsymbol{x}$. We need to show that with probability at least $1 - \alpha_{cs} - \alpha_{rs}$

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

where $M = \widehat{C} \circ \widehat{R} \circ E$. By the definition of $S^{\text{in}}(\boldsymbol{x})$ and E being the inverse of D, we have for all $\boldsymbol{x}' \in S^{\text{in}}(\boldsymbol{x})$, $\boldsymbol{z}' = E(\boldsymbol{x}') \in S(\boldsymbol{x})$, hence it suffices to prove

$$\forall \boldsymbol{z}' \in S(\boldsymbol{x}) : \widehat{C} \circ \widehat{R}(\boldsymbol{z}_G) = \widehat{C} \circ \widehat{R}(\boldsymbol{z}'), \tag{9}$$

where $\boldsymbol{z}_G = E(\boldsymbol{x})$.

Next, recall the definition of $g_{\boldsymbol{z}}(t) \coloneqq R(\boldsymbol{z} + t \cdot \boldsymbol{a})$ and note that for $\boldsymbol{z}' = \boldsymbol{z} + t' \cdot \boldsymbol{a}$, the center smoothing of

$$\begin{aligned}
&\widehat{g_{\boldsymbol{z}'}}(t)\text{: samples from } g_{\boldsymbol{z}'}\left(t + \mathcal{N}(0, \sigma_{cs}^2)\right) = R\left(\boldsymbol{z}' + \left(t + \mathcal{N}(0, \sigma_{cs}^2)\right) \cdot \boldsymbol{a}\right); \\
&\widehat{g_{\boldsymbol{z}}}(t + t')\text{: samples from } g_{\boldsymbol{z}}\left(t + t' + \mathcal{N}(0, \sigma_{cs}^2)\right) = R\left(\boldsymbol{z} + \left(t + t' + \mathcal{N}(0, \sigma_{cs}^2)\right) \cdot \boldsymbol{a}\right).
\end{aligned}$$

Since $\boldsymbol{z}' = \boldsymbol{z} + t' \cdot \boldsymbol{a}$, the sampling distributions are the same, hence $\widehat{g_{\boldsymbol{z}'}}(t) = \widehat{g_{\boldsymbol{z}}}(t + t')$, and in particular $\widehat{R}(\boldsymbol{z}') = \widehat{g_{\boldsymbol{z}'}}(0) = \widehat{g_{\boldsymbol{z}}}(t')$.

Now, let us get back to Eq. (9). By definition of $S(\boldsymbol{x})$, for all $\boldsymbol{z}' \in S(\boldsymbol{x})$, $\boldsymbol{z}' = \boldsymbol{z}_G + t' \cdot \boldsymbol{a}$ for some $t' \in [-\epsilon, \epsilon]$. Moreover, $\boldsymbol{r}_{cs} = \widehat{R}(\boldsymbol{z}_G) = \widehat{g_{\boldsymbol{z}_G}}(0)$ and $\widehat{R}(\boldsymbol{z}') = \widehat{g_{\boldsymbol{z}_G}}(t')$. Thm. 3 tells us that with probability at least $1 - \alpha_{cs}$

$$\begin{aligned}
&\forall t' \in [-\epsilon, \epsilon]\,.\, \|\widehat{g_{\boldsymbol{z}_G}}(0) - \widehat{g_{\boldsymbol{z}_G}}(t')\|_2 \leq d_{cs} \\
\iff &\forall \boldsymbol{z}' \in S(\boldsymbol{x})\,.\, \|\boldsymbol{r}_{cs} - \widehat{R}(\boldsymbol{z}')\|_2 \leq d_{cs},
\end{aligned} \tag{10}$$

provided that the center smoothing computation of $\boldsymbol{r}_{cs}$ does not abstain.

Finally, we consider the last component of the pipeline – the smoothed classifier $\widehat{C}$. Provided that $\widehat{C}$ does not abstain at the input $\boldsymbol{r}_{cs}$, Thm. 2 provides us with a radius d_{rs} around $\boldsymbol{r}_{cs}$ such that with probability at least $1 - \alpha_{rs}$

$$\begin{aligned} &\forall \boldsymbol{\delta} \text{ s.t. } \|\boldsymbol{\delta}\|_2 < d_{rs},\ \widehat{C}\left(\boldsymbol{r}_{cs}\right) = \widehat{C}\left(\boldsymbol{r}_{cs} + \boldsymbol{\delta}\right) \\ \iff &\forall \boldsymbol{r}' \text{ s.t. } \|\boldsymbol{r}_{cs} - \boldsymbol{r}'\|_2 < d_{rs},\ \widehat{C}\left(\boldsymbol{r}_{cs}\right) = \widehat{C}\left(\boldsymbol{r}'\right). \end{aligned} \tag{11}$$

If Alg. 1 returns CERTIFIED, that is $d_{cs} < d_{rs}$, combining Eq. (10) and (11) and applying the union bound, we obtain that with probability at least $1 - \alpha_{cs} - \alpha_{rs}$ we have $\widehat{C}\left(\boldsymbol{r}_{cs}\right) = \widehat{C}\left(\widehat{R}\left(\boldsymbol{z}'\right)\right)$ for all $\boldsymbol{z}' \in S\left(\boldsymbol{x}\right)$. That is,

$$\forall \boldsymbol{z}' \in S\left(\boldsymbol{x}\right) : \widehat{C} \circ \widehat{R}\left(\boldsymbol{z}_G\right) = \widehat{C} \circ \widehat{R}\left(\boldsymbol{z}'\right), \tag{12}$$

as required by Eq. (9). The same proof technique can also be extended to the multiple attribute vectors case. □

B Datasets and Dataset Statistics

In this section we provide further information and statistics about the datasets used in this work. CelebA[1] [52] is restricted to non-commercial research and education purposes and its authors [52] do not own the copyrights. FairFace [34] is licensed under CC BY 4.0. Tab. 5 contains statistics about the sensitive attributes and their corresponding attribute vectors. The lengths of the CelebA attribute vectors are computed for 64×64 images.

In Tab. 6 we report the base accuracies of two standard classifiers trained on the `Smiling` and `Earrings` CelebA tasks. The first classifier is a ResNet-18 network trained directly on the original images. The other one is a fully connected network operating on their Glow latent representations, $\boldsymbol{z}_G = E\left(\boldsymbol{x}\right)$. We remark that none of these classifiers involves representation learning. We report the means and standard deviations, averaged over 5 runs with different random seeds, on the validation and test sets, where the test set is the same subset on which we report the results in the main paper. The base accuracies on the downstream tasks used for the transfer learning experiments are reported in App. D.

In order to estimate the relative "unfairness" associated with each sensitive attribute, in Tab. 7 we compute the empirical individual fairness of the two classifiers. For each data point $\boldsymbol{x}$, we sample 9 points from $S^{\text{in}}\left(\boldsymbol{x}\right)$ evenly (15 points for `Pale+Young+Blond`). If all samples are classified the same, we add the original data point $\boldsymbol{x}$ to the empirical fairness counter. Note that this procedure cannot certify that all points from $S^{\text{in}}\left(\boldsymbol{x}\right)$ are classified the same. Therefore, these results come with no provable guarantees and serve as upper bounds of the certified individual fairness of the classifiers.

[1] https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

Table 5: Sensitive attribute statistics. The positive and negative sample ratio is reported for the training set, as the attribute vectors are computed on it.

Dataset	Sensitive attribute	Pos (%)	Neg (%)	$\|\boldsymbol{z}_{G,pos} - \boldsymbol{z}_{G,neg}\|_2$
CelebA	`Pale_Skin`	4.3	95.7	11.5
	`Young`	77.9	22.1	7.8
	`Blond_Hair`	14.9	85.1	15.8
	`Heavy_Makeup`	38.4	61.6	11.9
FairFace	`Race=Black`	14.1	85.9	10.9

Table 6: Baseline accuracies for the `Smiling` and `Earrings` CelebA tasks. The ResNet-18 classifier takes the original images as an input, while the $\boldsymbol{z}_G$ classifier is a fully connected network classifying their Glow latent representations. Neither of these classifiers involves representation learning.

	Majority class		Acc (ResNet-18)		Acc ($\boldsymbol{z}_G$)	
Task	Valid	Test	Valid	Test	Valid	Test
`Smiling`	51.7	52.6	**92.1 ± 0.2**	**90.9 ± 0.7**	89.4 ± 0.1	87.2 ± 1.1
`Earrings`	80.9	78.2	**86.2 ± 0.8**	**88.2 ± 1.1**	84.7 ± 0.1	85.2 ± 0.9

Table 7: Empirical individual fairness of the base classifiers evaluated via sampling. These results come with no provable guarantees and serve as upper bounds of the certified individual fairness.

		Emp. Fair (ResNet-18)		Emp. Fair ($\boldsymbol{z}_G$)	
Task	Sensitive attribute(s)	Valid	Test	Valid	Test
`Smiling`	`Pale_Skin`	74.1 ± 1.0	75.2 ± 1.1	75.8 ± 0.6	79.9 ± 1.2
	`Young`	87.7 ± 0.5	90.1 ± 0.7	95.2 ± 0.6	96.8 ± 0.9
	`Blond_Hair`	89.1 ± 1.1	89.4 ± 1.5	81.9 ± 1.6	84.6 ± 2.9
	`Heavy_Makeup`	82.3 ± 1.0	82.5 ± 1.2	74.9 ± 1.6	78.2 ± 2.3
	`Pale+Young`	71.5 ± 0.9	72.6 ± 1.1	75.8 ± 0.6	79.9 ± 1.2
	`Pale+Young+Blond`	70.3 ± 0.5	70.6 ± 0.9	72.5 ± 0.6	76.9 ± 1.1
`Earrings`	`Pale_Skin`	92.8 ± 0.9	90.4 ± 1.2	91.5 ± 1.0	91.5 ± 1.6
	`Young`	90.6 ± 1.8	87.7 ± 1.5	93.0 ± 1.2	94.7 ± 1.0
	`Blond_Hair`	89.7 ± 2.2	86.9 ± 2.5	88.3 ± 1.9	89.7 ± 2.2
	`Heavy_Makeup`	85.8 ± 3.4	82.2 ± 3.3	74.4 ± 4.4	73.3 ± 3.8

C Hyperparameter Tuning

In this section, we perform an extensive hyperparameter search in order to select suitable values for the hyperparameters. We evaluate on 311 samples from the *validation* set of CelebA (again, every 64-th), on the `Smiling` task with sensitive attributes `Pale_Skin` and `Young`. Afterwards, we reuse the same hyperparameter values for all tasks with very minor changes (which we verify by running the experiments on the validation set first). The tunable hyperparameters, as well as the range of values that we consider about them, are as follows:

- Adversarial loss weight:
  $\lambda_2 \in \{0, 0.001, 0.0025, 0.005, 0.01, 0.025, 0.05, 0.1, 0.25\}$
- Gaussian noise added during center smoothing of R:
  $\sigma_{cs} \in \{0.5, 0.55, 0.6, 0.65, 0.7, 0.75\}$
- Gaussian noise added during randomized smoothing of C:
  $\sigma_{rs} \in \{0.1, 0.25, 0.5, 1, 2.5, 5, 10, 25\}$

Tuning σ_{cs} and the baselines We begin with selecting the value for σ_{cs}. It is not used during the training of R and C, but is an integral part of the center smoothing computation which is performed during inference and is the most time-consuming component of the model pipeline. More concretely, both $\boldsymbol{r}_{cs} = \widehat{R}(\boldsymbol{z}_G)$ and d_{cs} depend on σ_{cs}, in turn affecting both the accuracy and the certified individual fairness. We evaluate the Naive model with all candidate values for σ_{cs} and show the results in Tab. 8. We observe very little variation in accuracy, while the best certified individual fairness and the smallest average center smoothing radii are obtained at $\sigma_{cs} = 0.6$ and 0.65. While there is no significant difference in performace between these two configurations, we expect that the slightly larger value for σ_{cs} would generally produce smaller center smoothing radii, leading to increased certified fairness. Therefore, we set $\sigma_{cs} = 0.65$ for all experiments (except for FairFace, where we use $\epsilon = 0.5$ and scale σ_{cs} correspondingly, i.e., $\sigma_{cs} = 0.325$). Using the same σ_{cs} values for both the baselines and LASSI allows us to attribute the improvements of the results to the additional training mechanisms that we apply and not merely to different hyperparameter values.

We perform a similar evaluation on the validation set of the other baseline, DataAug, and from the results in Tabs. 8 and 9 we set $\sigma_{rs} = 10$ for both Naive and DataAug. Although $\sigma_{rs} = 5$ seems to work slightly better for `Young`, we remark that `Young` is also the fairest of all considered sensitive attributes, so we choose a more conservative value that would be suitable for all of them.

Tuning λ_2 Next, we incorporate the adversarial loss weight λ_2 to the training and explore its impact on the model in Tab. 10. The certified individual fairness increases with increasing λ_2, until $\lambda_2 = 0.05$, and stays at the same level afterwards. Interestingly, the accuracy is mostly unaffected. We set $\lambda_2 = 0.05$ and $\sigma_{rs} = 2.5$ for LASSI, as they give most of the fairness boost obtained from adversarial training, while keeping the accuracy high. Notably, the hyperparameter tuning demonstrates that LASSI successfully enforces and certifies individual fairness for a wide range of hyperparameter values and is not highly sensitive to them.

Table 8: Results of Naive on the validation subset of CelebA for different values of σ_{cs} and σ_{rs}. The third column contains the mean center smoothing radii corresponding to the different σ_{cs} values. Smaller is generally better for certified individual fairness (see the condition in Alg. 1).

				σ_{rs}							
Sens. attribute	σ_{cs}	Mean d_{cs}	Metric	0.1	0.25	0.5	1	2.5	5	10	25
Pale_Skin	0.5	42.25	Acc	87.8	87.8	87.5	88.4	89.1	88.7	88.7	84.6
			Fair	0	0	0	0	0	0	0	0
	0.55	34.19	Acc	87.8	87.8	87.8	88.4	88.7	88.7	88.4	84.6
			Fair	0	0	0	0	0	0	0	0
	0.6	**33.34**	Acc	87.8	87.5	87.8	88.4	88.7	88.7	88.7	84.6
			Fair	0	0	0	0	0	0	**1.0**	0
	0.65	33.37	Acc	87.5	87.5	87.8	88.4	88.8	88.7	88.4	84.6
			Fair	0	0	0	0	0	0	**1.0**	0
	0.7	33.72	Acc	87.5	87.5	87.5	88.4	88.4	88.7	88.1	84.6
			Fair	0	0	0	0	0	0	**1.0**	0
	0.75	34.18	Acc	87.8	88.1	88.1	88.4	88.7	89.1	88.1	84.6
			Fair	0	0	0	0	0	0	**1.0**	0
Young	0.6	**8.16**	Acc	88.1	88.1	87.8	87.8	88.7	88.7	88.1	85.2
			Fair	0	0	0	5.1	36.3	**58.8**	58.5	39.9
	0.65	**8.16**	Acc	88.1	88.1	87.8	87.8	88.7	88.7	88.1	84.9
			Fair	0	0	0	4.8	36.3	**58.8**	58.2	39.5

Table 9: Results of the DataAug baseline on the validation set of CelebA for $\sigma_{cs} = 0.65$ and different values of σ_{rs}.

				σ_{rs}							
Sens. attribute	σ_{cs}	Mean d_{cs}	Metric	0.1	0.25	0.5	1	2.5	5	10	25
Pale_Skin	0.65	14.52	Acc	87.5	87.5	87.8	87.8	88.7	89.4	88.7	84.9
			Fair	0	0	0	0	0	28.3	**31.5**	10.0
Young	0.65	7.09	Acc	87.5	87.8	87.8	89.1	88.7	88.7	88.7	84.9
			Fair	0	0	0	1.6	46.6	**65.6**	65.0	48.9

Table 10: Results of LASSI on the validation subset of CelebA for different values of λ_2 and σ_{rs}, while keeping $\sigma_{cs} = 0.65$. The certified individual fairness increases with increasing λ_2, until the $\lambda_2 = 0.05$ level.

			σ_{rs}							
Sens. attribute	λ_2	Metric	0.1	0.25	0.5	1	2.5	5	10	25
Pale_Skin	0.001	Acc	86.5	86.8	87.1	87.5	89.1	89.4	88.1	84.9
		Fair	0	0	0	0	0	**13.2**	12.9	1.9
	0.0025	Acc	87.8	88.1	88.4	88.7	90.4	89.1	86.5	83.3
		Fair	0	0	0	0	15.4	**27.3**	24.8	5.5
	0.005	Acc	87.8	87.8	88.1	87.8	89.7	89.1	87.5	82.0
		Fair	0	0	0	0	35.0	**40.5**	37.0	15.4
	0.01	Acc	88.1	88.1	87.8	88.1	89.4	90.0	87.5	82.0
		Fair	0	0	0	9.3	46.0	**49.5**	47.6	27.7
	0.025	Acc	88.4	88.1	88.1	88.4	89.1	89.7	87.5	82.3
		Fair	0	1.9	9.6	49.2	64.3	**66.2**	64.0	47.9
	0.05	Acc	87.8	87.8	88.1	88.1	89.7	89.4	86.8	83.0
		Fair	45.0	97.1	97.7	**98.1**	96.1	96.1	95.5	93.6
	0.1	Acc	86.5	86.5	86.5	86.8	86.5	85.9	83.3	76.8
		Fair	57.9	93.6	93.9	94.5	**96.8**	96.1	94.9	88.1
	0.25	Acc	87.1	87.1	87.1	87.5	87.5	85.9	79.4	67.8
		Fair	96.8	96.1	96.1	96.1	**98.1**	97.4	93.2	79.7
Young	0.05	Acc	89.1	89.1	88.1	89.4	89.4	89.1	88.7	84.6
		Fair	97.1	97.7	98.4	**99.0**	**99.0**	98.7	96.5	96.1
	0.1	Acc	88.1	88.7	89.4	89.4	88.7	88.7	87.8	82.6
		Fair	58.5	94.9	94.9	96.8	**97.1**	95.8	96.1	92.3
	0.25	Acc	88.4	88.4	88.4	88.7	88.4	88.1	86.8	77.8
		Fair	98.4	98.1	98.4	**99.4**	**99.4**	98.7	95.2	89.4

Selected experiment hyperparameters Here, we summarize the hyperparameter values selected for the final experiments. We use $\epsilon = 1$ for all similarity set definitions except the experiments with: (i) the alternative attribute vectors from [13,48], where $\epsilon = 10$, and (ii) FairFace, where $\epsilon = 0.5$. We maintain the ϵ/σ_{cs} ratio, which impacts center smoothing, setting $\sigma_{cs} = 0.65$ by default (as stated in the sections above) and using $\sigma_{cs} = 6.5$ and 0.325 when $\epsilon = 10$ and 0.5 respectively. Our smoothing arguments are consistent with prior work [10,44]:

- Randomized smoothing [10]: $\alpha_{rs} = 0.001$, $N_{rs} = 100{,}000$, $N_{0,rs} = 2000$.
- Center smoothing [44]: $\alpha_{cs} = 0.01$, $N_{cs} = 10{,}000$, $N_{0,cs} = 10{,}000$.

The rest of the model hyperparameters are listed in Tab. 11. In the CelebA 64×64 and 128×128 setups, we run LASSI with $\lambda_2 = 0.25$ for the (target=`Earrings`, sensitive=`Makeup`) pair of attributes because of the high correlation between them. We train the representation R for 20 epochs in the transfer experiments (CelebA, FairFace) and 5 epochs otherwise. The linear classifier C is trained for 1 epoch. We generally set a lower value to σ_{rs} when the task is more difficult and the downstream classifier is therefore less confident. Overall, we remark that the hyperparameter values are similar and within the same range for all models and experiments, meaning that our approach does not require substantial fine-tuning.

Table 11: Hyperparameters used for the different model and experiment setups.

Dataset	Model / Experiment(s)	Hyperparameters
CelebA	64×64 and 128×128	$\lambda_1 = 1$; $\lambda_2 = 0$ (Naive, DataAug) and 0.05 (LASSI); $\lambda_3 = 0$; $\sigma_{rs} = 10$ (Naive, DataAug) and 2.5 (LASSI); $s = 10$ (DataAug, LASSI).
	Transfer	$\lambda_1 = 0$; $\lambda_2 = 0.05$; $\lambda_3 = 0.1$; $\sigma_{rs} = 0.5$; $s = 10$.
FairFace	Naive	$\lambda_1 = 1$; $\lambda_2 = \lambda_3 = 0$; $\sigma_{rs} = 5$ (`Age-2`) and 0.1 (`Age-3`, `Age` (all)).
	LASSI	$\lambda_1 = 1$; $\lambda_2 = 0.1$; $\lambda_3 = 0$; $\sigma_{rs} = 0.25$; $s = 10$.
	Transfer	$\lambda_1 \in \{0, 0.001, 0.01\}$; $\lambda_2 = \lambda_3 = 0.1$; $\sigma_{rs} = 0.1$; $s = 10$.
3D Shapes (App. F)	Naive	$\lambda_1 = 1$; $\lambda_2 = \lambda_3 = 0$; $\sigma_{rs} = 5$.
	LASSI	$\lambda_1 = 1$; $\lambda_2 = 0.1$; $\lambda_3 = 0$; $\sigma_{rs} = 1$; $s = 100$.

D More Experimental Results on CelebA

This section provides further details about the experiments on the CelebA dataset with the standard attribute vector from [41], $\boldsymbol{a} = \boldsymbol{z}_{G,pos} - \boldsymbol{z}_{G,neg}$ (Sec. 4.1).

64×64 images Tab. 12 contains the means and the standard deviations of the accuracies and the certified individual fairness of the CelebA 64×64 experiments summarized in Tab. 1, averaged over 5 runs. The standard deviation of Naive and

DataAug's fairness is high, while LASSI consistently enforces certified individual fairness with low variance.

Table 12: Means and standard deviations of the accuracies and the certified individual fairness reported in Tab. 1, averaged over 5 runs with different random seeds on the `Smiling` (rows 1-6) and `Earrings` (rows 7-10) tasks.

	Naive		DataAug		LASSI (ours)	
Sens. attribs.:	Acc	Fair	Acc	Fair	Acc	Fair
`Pale_Skin`	**86.3 ± 1.5**	0.6 ± 0.5	85.7 ± 1.2	12.2 ± 14.7	85.9 ± 1.3	**98.0 ± 0.5**
`Young`	**86.3 ± 1.8**	38.2 ± 23.4	85.9 ± 1.6	43.0 ± 30.7	**86.3 ± 1.3**	**98.8 ± 0.6**
`Blond_Hair`	86.3 ± 1.6	3.4 ± 3.1	**86.6 ± 1.0**	9.4 ± 10.0	86.4 ± 1.0	**94.7 ± 1.5**
`Heavy_Makeup`	**86.3 ± 1.1**	0.4 ± 0.4	85.3 ± 1.7	13.7 ± 8.8	85.6 ± 1.6	**91.3 ± 8.1**
`P+Y`	**86.0 ± 1.5**	0.4 ± 0.4	85.8 ± 1.4	9.9 ± 12.7	85.8 ± 0.9	**97.3 ± 0.9**
`P+Y+B`	86.2 ± 1.7	0.0 ± 0.0	**86.4 ± 1.0**	3.6 ± 3.8	85.5 ± 0.4	**86.5 ± 2.7**
`Pale_Skin`	81.3 ± 2.2	24.3 ± 35.6	81.0 ± 2.3	40.4 ± 32.6	**85.0 ± 0.5**	**98.5 ± 0.9**
`Young`	81.4 ± 2.2	59.2 ± 18.0	79.9 ± 1.4	72.0 ± 24.1	**84.5 ± 1.0**	**98.0 ± 1.1**
`Blond_Hair`	81.4 ± 2.2	9.2 ± 17.5	82.2 ± 2.8	30.5 ± 40.9	**84.8 ± 0.5**	**96.2 ± 2.6**
`Heavy_Makeup`	81.6 ± 1.9	20.5 ± 13.0	80.3 ± 1.9	49.2 ± 37.0	**82.3 ± 0.6**	**98.7 ± 0.7**

Table 13: Empirical evaluation of the individual fairness of the models computed by comparing their predictions on the original test samples to the model predictions on the endpoints of the corresponding similarity sets.

Task	Sensitive attribute(s)	Naive	DataAug	LASSI (ours)
`Smiling`	`Pale_Skin`	78.4 ± 2.1	90.1 ± 1.9	**99.6 ± 0.2**
	`Young`	95.3 ± 0.4	96.7 ± 0.5	**99.6 ± 0.2**
	`Blond_Hair`	83.3 ± 0.7	93.9 ± 1.5	**99.2 ± 0.4**
	`Heavy_Makeup`	75.8 ± 2.4	88.3 ± 0.8	**97.9 ± 1.6**
	`Pale+Young`	78.0 ± 2.0	89.0 ± 2.2	**99.4 ± 0.5**
	`Pale+Young+Blond`	77.9 ± 2.1	87.4 ± 0.9	**96.9 ± 0.7**
`Earrings`	`Pale_Skin`	97.1 ± 1.6	99.1 ± 0.7	**99.5 ± 0.4**
	`Young`	98.5 ± 1.4	**99.5 ± 0.5**	99.2 ± 0.4
	`Blond_Hair`	96.7 ± 3.4	98.5 ± 0.4	**99.1 ± 0.7**
	`Heavy_Makeup`	92.2 ± 6.6	98.1 ± 1.1	**99.7 ± 0.3**

Moreover, in Tab. 13 we check for what fraction of the test subset the models classify the similarity set endpoints the same as the original data point. Note

that this is again another empirical estimate, serving as an upper bound of the certified individual fairness of the models. Nevertheless, LASSI outperforms the baselines on that metric as well. More importantly, out of all 150 combinations of models, tasks and sensitive attributes (3 model types, 10 task-attribute pairs, 5 random seeds), in 8 combinations there is only 1 test sample which we certify as individually fair but the endpoints classifications mismatch. In all other combinations, no such situation occurs, serving as another test for the correctness of our certificates. One test sample out of 312 is 0.32%, which is within our confidence of $1 - \alpha_{cs} - \alpha_{rs} = 98.9\%$.

128×128 images Keeping all hyperparameters the same, we evaluate LASSI on images of size 128×128. The results in Tab. 14 indicate that LASSI increases the certified individual fairness in this setting as well, while also slightly improving the classification accuracy. We attribute this to the richer and larger latent space of Glow, which is potentially more easily separable in this case.

Table 14: Evaluation of LASSI on 128×128-dimensional images, demonstrating that it significantly increases the certified individual fairness for larger images as well. Evaluated tasks: `Smiling` (rows 1-6) and `Earrings` (rows 7-10).

	Naive		DataAug		LASSI (ours)	
Sens. attribs.:	Acc	Fair	Acc	Fair	Acc	Fair
`Pale_Skin`	88.8 ± 1.0	0.0 ± 0.0	89.6 ± 0.5	0.0 ± 0.0	**90.0 ± 1.1**	**70.6 ± 14.2**
`Young`	88.7 ± 0.7	46.0 ± 16.2	88.8 ± 1.0	47.6 ± 20.2	**89.7 ± 0.7**	**97.2 ± 1.6**
`Blond_Hair`	88.8 ± 0.9	0.1 ± 0.1	89.4 ± 1.1	0.0 ± 0.0	**90.1 ± 0.8**	**77.8 ± 10.2**
`Heavy_Makeup`	89.0 ± 0.9	2.5 ± 3.5	89.6 ± 1.1	30.4 ± 20.7	**90.2 ± 0.3**	**87.6 ± 3.9**
`P+Y`	88.8 ± 1.0	0.0 ± 0.0	89.4 ± 1.3	8.7 ± 16.5	**90.2 ± 0.5**	**69.4 ± 9.7**
`P+Y+B`	88.7 ± 0.8	0.0 ± 0.0	89.9 ± 1.5	4.4 ± 9.6	**90.2 ± 0.7**	**72.7 ± 5.0**
`Pale_Skin`	80.1 ± 1.4	0.0 ± 0.0	80.1 ± 2.5	0.1 ± 0.1	**84.4 ± 0.9**	**90.4 ± 2.5**
`Young`	80.2 ± 1.4	73.5 ± 20.4	80.3 ± 1.5	78.2 ± 18.1	**85.5 ± 1.4**	**96.4 ± 1.7**
`Blond_Hair`	80.2 ± 1.4	0.0 ± 0.0	80.6 ± 2.0	0.0 ± 0.0	**83.9 ± 0.9**	**89.7 ± 4.0**
`Heavy_Makeup`	80.3 ± 1.4	42.1 ± 15.9	80.1 ± 1.9	65.1 ± 31.1	**81.7 ± 1.3**	**98.3 ± 1.3**

Transfer learning Tab. 15 contains the base standard accuracies on the transfer tasks. Tab. 16 reports the means and the standard deviations of LASSI on the `Smiling` task when solved in a transfer learning setting.

Table 15: Baseline accuracies on the transfer CelebA tasks. As before, the ResNet-18 classifier takes as an input the original images, while the $\boldsymbol{z}_G$ classifier is a fully connected network classifying their Glow latent representations. Neither of these classifiers involves representation learning.

	Majority class		Acc (ResNet-18)		Acc ($\boldsymbol{z}_G$)	
Task	Valid	Test	Valid	Test	Valid	Test
`Smiling`	51.7	52.6	**92.1 ± 0.2**	**90.9 ± 0.7**	89.4 ± 0.1	87.2 ± 1.1
`High_Cheeks`	55.1	51.9	**87.2 ± 0.2**	**86.8 ± 0.4**	84.3 ± 0.1	83.8 ± 0.7
`Mouth_Open`	51.8	53.8	**92.7 ± 0.3**	**92.9 ± 0.7**	88.1 ± 0.2	89.6 ± 1.1
`Lipstick`	55.4	54.8	**91.5 ± 0.2**	90.5 ± 0.8	89.2 ± 0.1	**90.6 ± 1.1**
`Heavy_Makeup`	61.0	58.7	**90.2 ± 0.4**	**89.9 ± 0.4**	87.8 ± 0.1	88.6 ± 1.1
`Wavy_Hair`	72.3	65.1	**82.7 ± 1.8**	76.3 ± 3.3	80.9 ± 0.5	**81.7 ± 0.4**
`Eyebrows`	74.2	71.8	**83.5 ± 0.5**	**81.1 ± 0.6**	80.1 ± 0.1	79.4 ± 1.6

Table 16: Mean and standard deviation of the accuracies and the certified individual fairness of LASSI on `Smiling` in a transfer learning setting (Tab. 3).

Task	Sensitive attribute(s)	Acc	Fair
	`Pale_Skin`	86.2 ± 1.1	93.1 ± 2.4
	`Young`	86.0 ± 1.2	95.4 ± 1.0
`Smiling`	`Blond_Hair`	85.1 ± 1.6	93.8 ± 1.8
	`Pale+Young`	85.9 ± 0.3	92.2 ± 0.7
	`Pale+Young+Blond`	85.1 ± 0.7	87.0 ± 2.3

E Different Attribute Vector Types

In this section, we demonstrate that LASSI is independent of the actual computation of the attribute vector $\boldsymbol{a}$ and that it can improve the individual fairness for various attribute vector types.

Denton et al. [13] First, in Tab. 17 we report the means and the standard deviations of the accuracies and the certified individual fairness from Tab. 2. The attribute vector $\boldsymbol{a}$ used here is orthogonal to the decision boundary of the linear classifier $\text{sign}(\boldsymbol{a}^\top \boldsymbol{z}_G + b)$ [13] (Sec. 4.1), with its length set to $\epsilon = 10$.

Ramaswamy et al. [67] Next, we adapt the attribute vector computation proposed by [67] by computing sample-specific vectors $\boldsymbol{a}_i = \boldsymbol{z}_{G,i} - \boldsymbol{z}'_{G,i}$ for every $\boldsymbol{x}_i$ from the training set, where $\boldsymbol{z}_{G,i} = E(\boldsymbol{x}_i)$ and $\boldsymbol{z}'_{G,i}$ is as defined in [67, Eq. (3)]. All sample-specific $\boldsymbol{a}_i$'s share the same direction, so we can average them to obtain the global attribute vector $\boldsymbol{a} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{a}_i$ and set $\epsilon = 1$.

Table 17: Means and standard deviations of the accuracies and the certified individual fairness reported in Tab. 2, averaged over 5 runs with different random seeds on the `Smiling` task.

	Naive		DataAug		LASSI (ours)	
Sens. attribs.:	Acc	Fair	Acc	Fair	Acc	Fair
`Pale_Skin`	86.4 ± 1.7	34.0 ± 5.4	85.9 ± 1.5	90.3 ± 3.9	**86.5 ± 1.3**	**98.8 ± 1.2**
`Young`	86.3 ± 1.8	73.1 ± 3.5	86.2 ± 1.5	90.3 ± 3.3	**86.8 ± 1.0**	**97.9 ± 1.2**
`Blond_Hair`	86.2 ± 1.8	71.4 ± 4.0	86.1 ± 1.8	88.8 ± 2.7	**86.7 ± 1.4**	**98.8 ± 0.7**
`Heavy_Makeup`	86.2 ± 1.6	11.5 ± 2.5	86.3 ± 1.1	87.4 ± 1.6	**86.8 ± 1.0**	**98.8 ± 0.9**
`P+Y`	86.2 ± 1.8	28.6 ± 3.4	85.8 ± 1.5	84.7 ± 4.1	**86.5 ± 1.2**	**98.6 ± 1.8**
`P+Y+B`	86.2 ± 1.7	23.7 ± 2.1	85.9 ± 1.8	82.2 ± 5.2	**86.4 ± 1.1**	**98.7 ± 0.5**

Li and Xu [48] Finally, [48] discover biased attributes of pre-trained classifiers. To that end, we train a ResNet-18 on the `Smiling` task. Then, we run [48]'s optimization procedure to iteratively find 3 biased attribute vectors (each orthogonal to the target and to the other attribute vectors) for that model using Glow as the generative model. We use $\epsilon = 10$ for these vectors.

Tab. 18 shows that LASSI significantly improves the certified individual fairness while maintaining the same high accuracy level for [67] and [48], as with the attribute vectors from [13,41], when evaluated on the `Smiling` task.

Table 18: Evaluation of LASSI on CelebA using sensitive attribute vectors from [48,67]. We denote [48]'s vectors as $\boldsymbol{a}_0$, $\boldsymbol{a}_1$, and $\boldsymbol{a}_2$ since they are not necessarily associated with a sensitive attribute (unlike [13,41,67]). As for the vectors from [13,41] (Tabs. 1 and 2), LASSI significantly increases certified fairness without affecting the accuracy.

		Naive		DataAug		LASSI (ours)	
$\boldsymbol{a}$	Sens. attribs.:	Acc	Fair	Acc	Fair	Acc	Fair
[67]	`Pale_Skin`	86.3 ± 1.8	89.0 ± 3.9	86.0 ± 1.5	92.4 ± 2.6	**86.8 ± 1.2**	**98.6 ± 1.0**
	`Young`	86.3 ± 1.8	95.1 ± 1.5	86.2 ± 1.6	95.6 ± 1.8	**86.9 ± 1.2**	**99.5 ± 0.5**
	`Blond_Hair`	86.2 ± 1.8	90.8 ± 3.5	86.2 ± 1.6	89.7 ± 3.0	**86.8 ± 1.1**	**98.8 ± 0.3**
	`Heavy_Makeup`	86.3 ± 1.8	92.8 ± 1.4	86.0 ± 1.6	94.4 ± 1.4	**86.7 ± 1.2**	**99.4 ± 0.3**
	`P+Y`	86.3 ± 1.8	88.0 ± 3.9	86.2 ± 1.9	91.5 ± 4.1	**86.7 ± 1.1**	**98.8 ± 0.9**
	`P+Y+B`	86.3 ± 1.8	85.6 ± 4.3	86.5 ± 1.5	88.7 ± 5.4	**86.7 ± 1.3**	**98.4 ± 0.9**
[48]	$\boldsymbol{a}_0$	86.2 ± 1.8	92.3 ± 2.1	86.3 ± 1.6	94.8 ± 3.7	**86.9 ± 1.4**	**99.3 ± 0.9**
	$\boldsymbol{a}_0+\boldsymbol{a}_1$	86.3 ± 1.8	90.7 ± 2.7	86.4 ± 1.5	93.4 ± 1.2	**86.9 ± 1.1**	**98.3 ± 1.3**
	$\boldsymbol{a}_0+\boldsymbol{a}_1+\boldsymbol{a}_2$	86.3 ± 1.8	90.1 ± 2.8	86.3 ± 1.7	92.4 ± 1.6	**86.8 ± 1.0**	**98.5 ± 0.6**

F Certification with Ground Truth Data

An essential part of the evaluation is demonstrating that the fairness certificates obtained using the generative model can transfer to ground truth data. However, CelebA does not contain images of the same individual with different attributes, e.g., the same individual with different skin colors. Thus, we experiment with the 3D Shapes dataset (Apache-2.0 license) [8], which provides images of 3D shapes that are procedurally generated from 6 independent latent factors: `floor hue`, `wall hue`, `object hue`, `scale`, `shape`, and `orientation`. Therefore, we can obtain ground truth images of the same object with varying latent factors. The 3D Shapes dataset is typically used to investigate disentanglement properties of unsupervised learning methods, e.g., in the context of fairness [53].

The goal is to show that the similarity set computed by Glow captures a given latent factor (as in Fig. 12) and that certification with respect to this set will result in certification of the ground truth. To that end, we experiment with `orientation` as the continuous sensitive attribute. It has $v = 15$ possible values, the most among all latent factors, providing for the most rigorous evaluation. The target attribute is set to `object hue`, which has 10 different classes.

We filter the original training set to create a biased one, correlating `orientation` and `object hue`. We only keep those samples in the training set for which: (i) `hue` ≤ 5 and `orient` ≤ 7, or (ii) `hue` ≥ 6 and `orient` ≥ 9. We extend the attribute vector computation from Sec. 4.1 [41] (performed on the original, unfiltered training set) to non-binary attributes, defining $\boldsymbol{a}_{ij} = \boldsymbol{z}_{G,i} - \boldsymbol{z}_{G,j}$, where $1 \leq i, j \leq v$ are sensitive attribute values. Based on the construction of the biased training set, we let the similarity set $S(\boldsymbol{x})$ to be defined by all attribute vectors $\{\boldsymbol{a}_{ij}\}$ for which $i < 8 < j$ ($7 \cdot 7 = 49$ vectors) and set $\epsilon = 1$. We train Naive ($\lambda_1 = 1$; $\lambda_2 = \lambda_3 = 0$; $\sigma_{rs} = 5$) and LASSI ($\lambda_1 = 1$; $\lambda_2 = 0.1$; $\lambda_3 = 0$; $\sigma_{rs} = 1$; $s = 100$) models and report results on 300 samples from the test set. When running LASSI on 3D Shapes, we sample more points ($s = 100$) compared to the other datasets in order to accommodate for the more complex similarity set, defined by many more attribute vectors.

In the evaluation, apart from reporting the accuracy and the certified fairness (CertFair) on the (unbiased) test subset, for each sample we also obtain the v similar ground truth data points, i.e., the same shape at v different orientations, while fixing all other factors. The empirical unfairness (EmpUnfair) in this case is the percentage of test samples for which the downstream classifier does not classify all v ground truth individually similar images the same. Moreover, if any of the v similar data points is *certified*, we check whether *all* v similar ground truth data points obtain the same classification, indicating ground truth fairness.

Tab. 19 shows that LASSI substantially increases the accuracy and the certified individual fairness (w.r.t. the similarity set computed using Glow), while being nearly 100% empirically fair on the ground truth images. That is, in 0.3% of the test samples there were different classification outcomes among their v similar (ground-truth) samples. Crucially, in all of these cases, our method did not certify individual fairness for *any* of the v similar data points, showing that the certificates transfer to the ground truth.

Table 19: Evaluation on 3D Shapes for the task `object hue`. The certification rate (CertFair) and the percentage of ground truth empirically unfair data points (EmpFair) sum up below 100%.

Method:	Naive			LASSI (ours)		
Sens. attrib.	Acc	CertFair	EmpUnfair (↓)	Acc	CertFair	EmpUnfair (↓)
`orientation`	32.0	0	69.3	**100**	**81.3**	**0.3**

G More Examples of Similar Individuals

Here, we provide further samples from the similarity sets $S^{\text{in}}(\boldsymbol{x})$ (defined with $\boldsymbol{a} = \boldsymbol{z}_{G,pos} - \boldsymbol{z}_{G,neg}$), as reconstructed by Glow, for various inputs $\boldsymbol{x}$ randomly drawn from our evaluation subsets. A summary of all configurations is listed in Tab. 20. The images in the middle of the CelebA and FairFace reconstructions correspond to the original inputs. The perturbations range uniformly between $[-\frac{\epsilon}{\sqrt{n}}, \frac{\epsilon}{\sqrt{n}}]$, where n is the number of sensitive attributes. For $n > 1$, all attribute vectors are multiplied by the same t before adding them to the latent representation of the original inputs. $\epsilon = 1$ for CelebA and 3D Shapes and $\epsilon = 0.5$ for FairFace.

Table 20: Example image reconstructions from the similarity sets in this work.

Dataset	Sensitive attribute(s)	Figure
CelebA	`Pale_Skin`	Fig. 5
	`Young`	Fig. 6
	`Blond_Hair`	Fig. 7
	`Heavy_Makeup`	Fig. 8
	`Pale + Young`	Fig. 9
	`Pale + Young + Blond`	Fig. 10
FairFace	`Race=Black`	Fig. 11
3D Shapes	`orientation`	Fig. 12

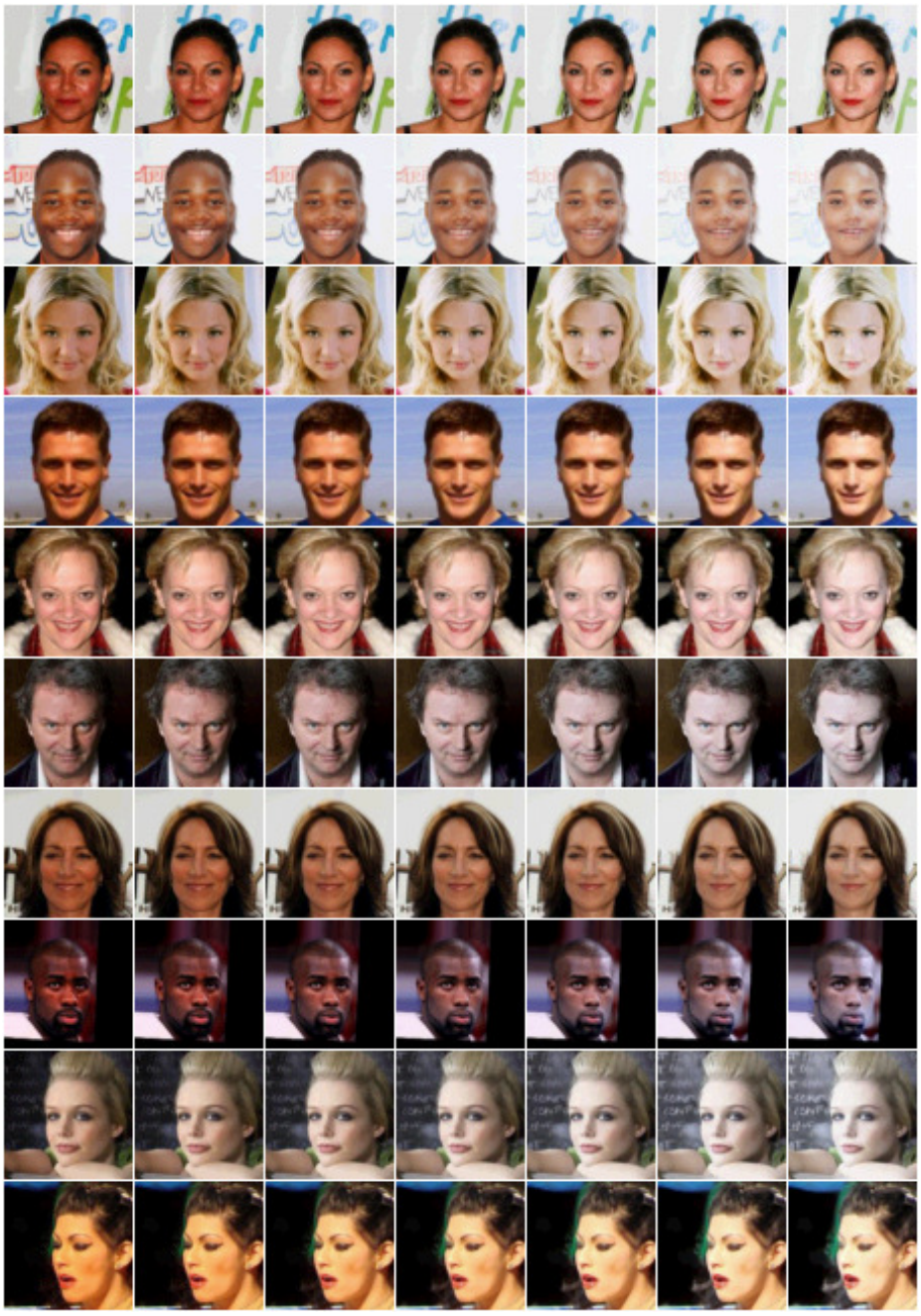

Fig. 5: Similar individuals from $S^{\text{in}}(\boldsymbol{x})$, for $\boldsymbol{x}$ in the CelebA dataset, obtained by varying the sensitive attribute `Pale_Skin`.

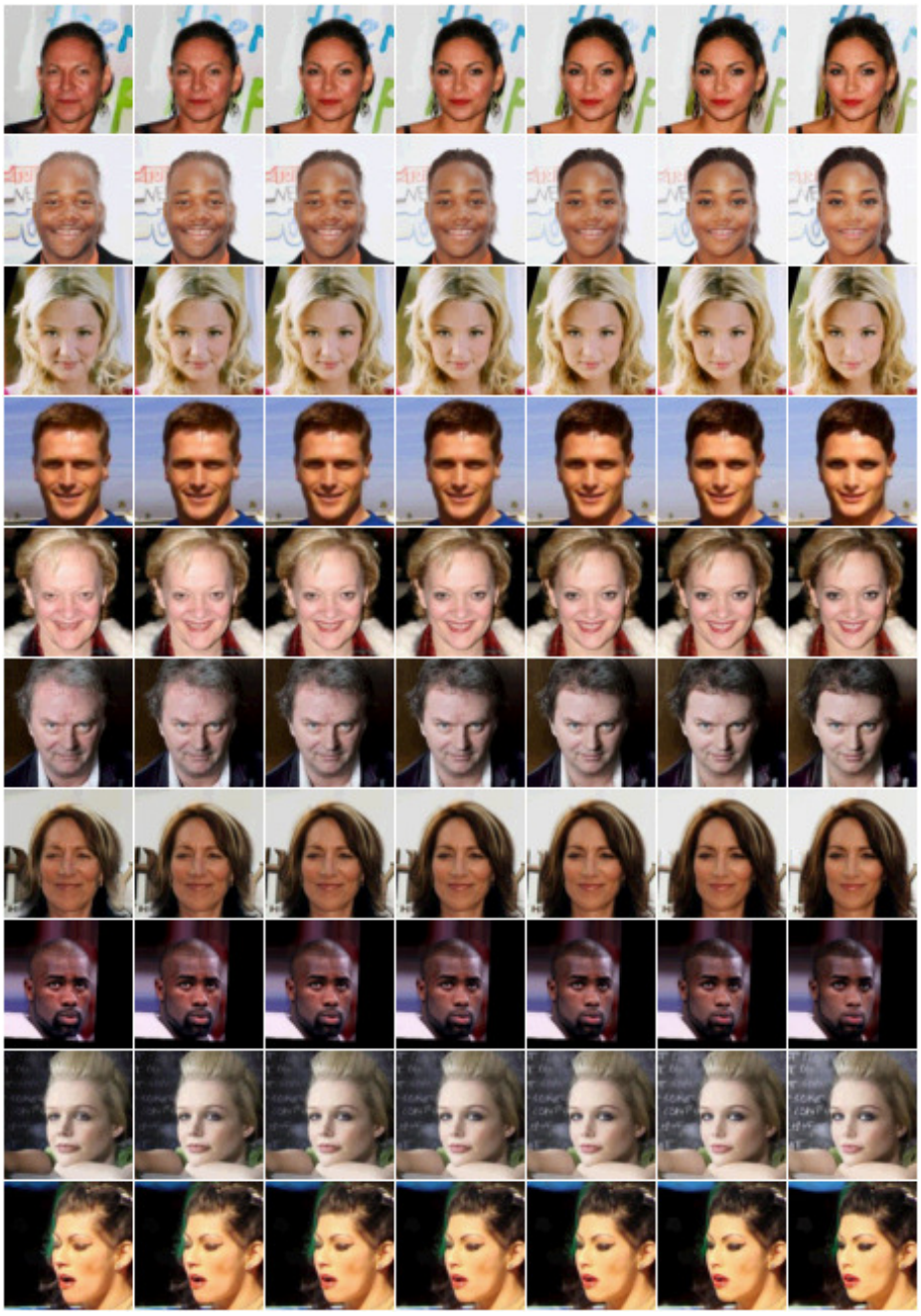

Fig. 6: Similar individuals from $S^{\text{in}}(\boldsymbol{x})$, for $\boldsymbol{x}$ in the CelebA dataset, obtained by varying the sensitive attribute `Young`.

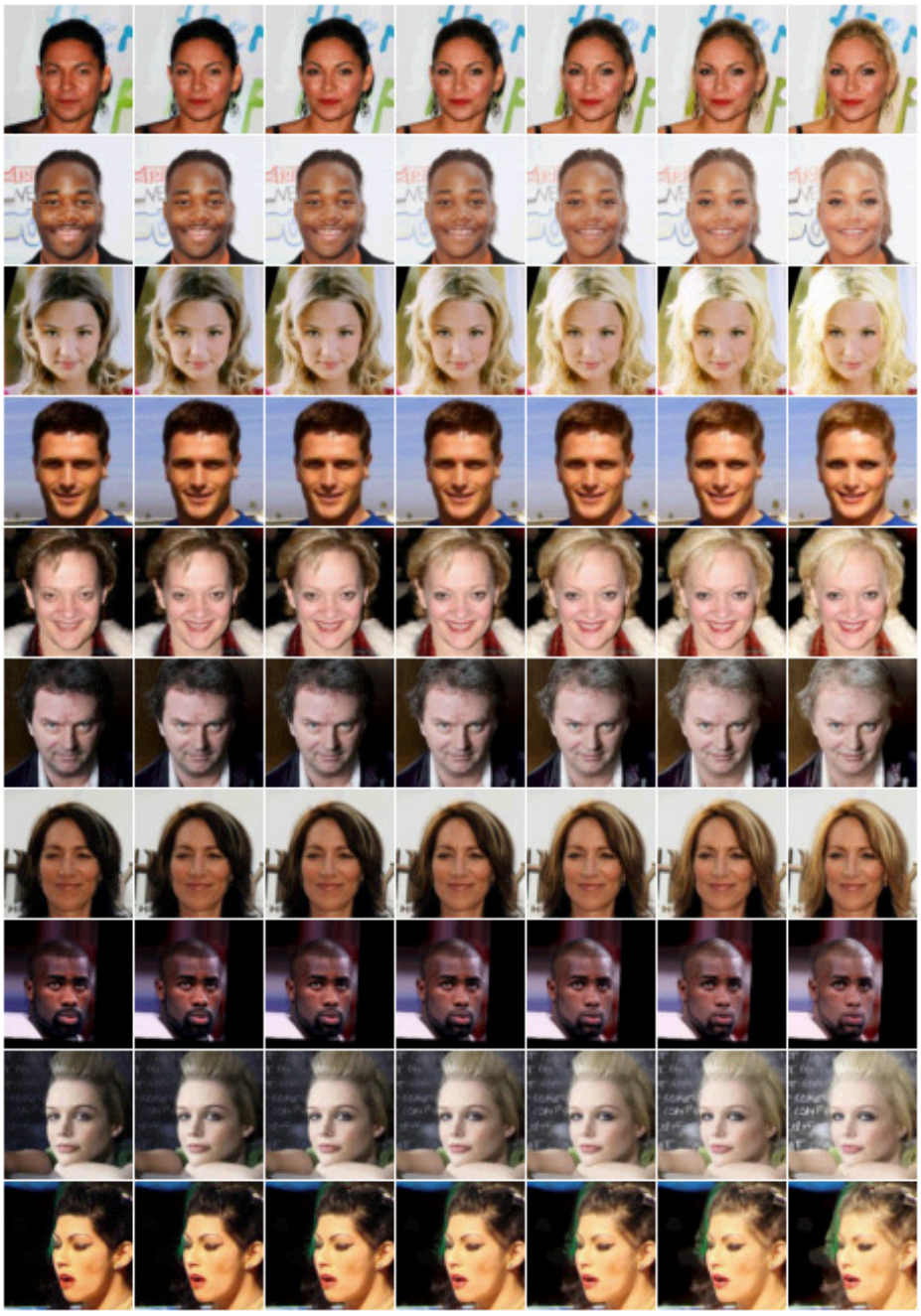

Fig. 7: Similar individuals from $S^{\text{in}}(\boldsymbol{x})$, for $\boldsymbol{x}$ in the CelebA dataset, obtained by varying the sensitive attribute `Blond_Hair`.

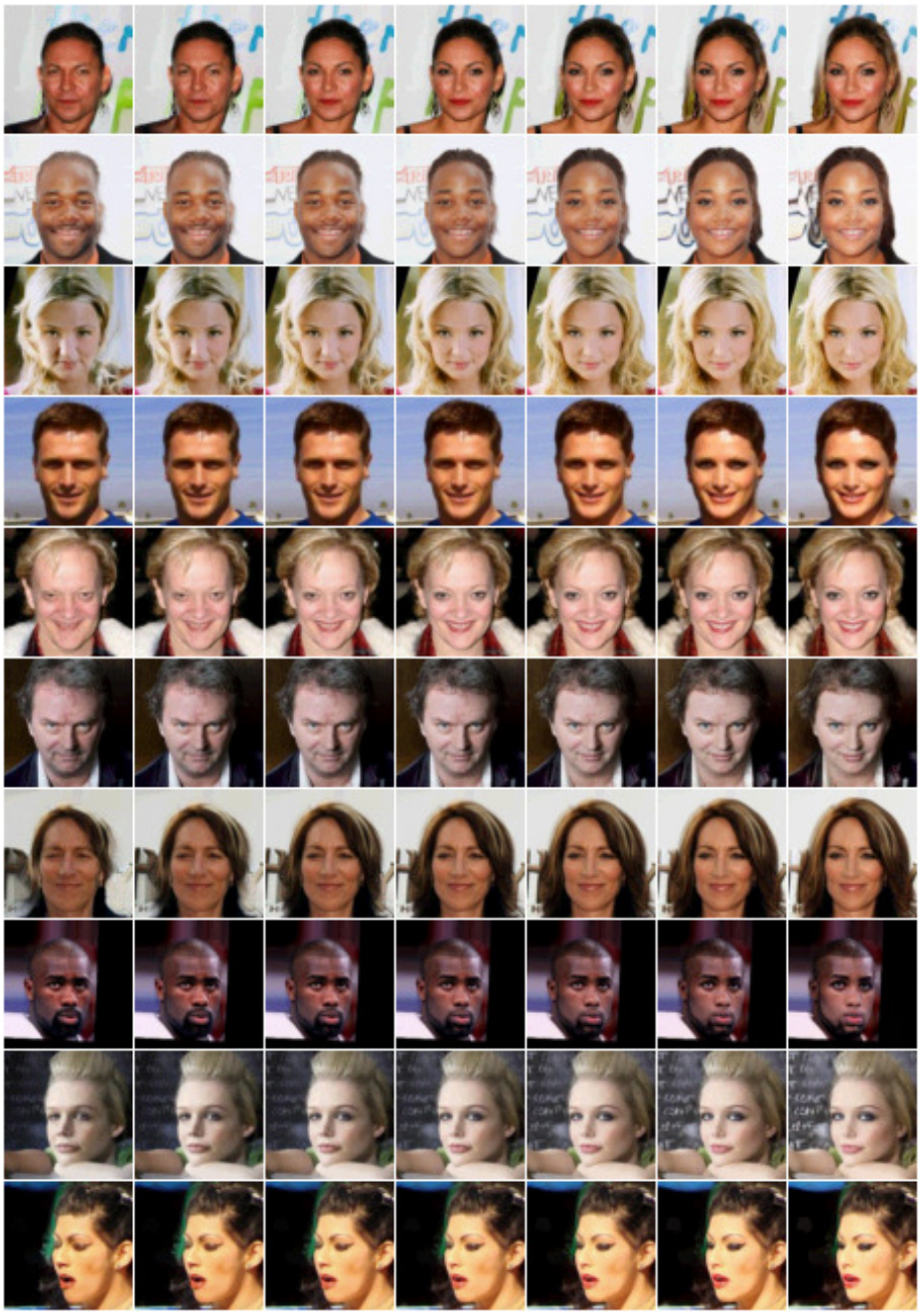

Fig. 8: Similar individuals from $S^{\text{in}}(\boldsymbol{x})$, for $\boldsymbol{x}$ in the CelebA dataset, obtained by varying the sensitive attribute `Heavy_Makeup`.

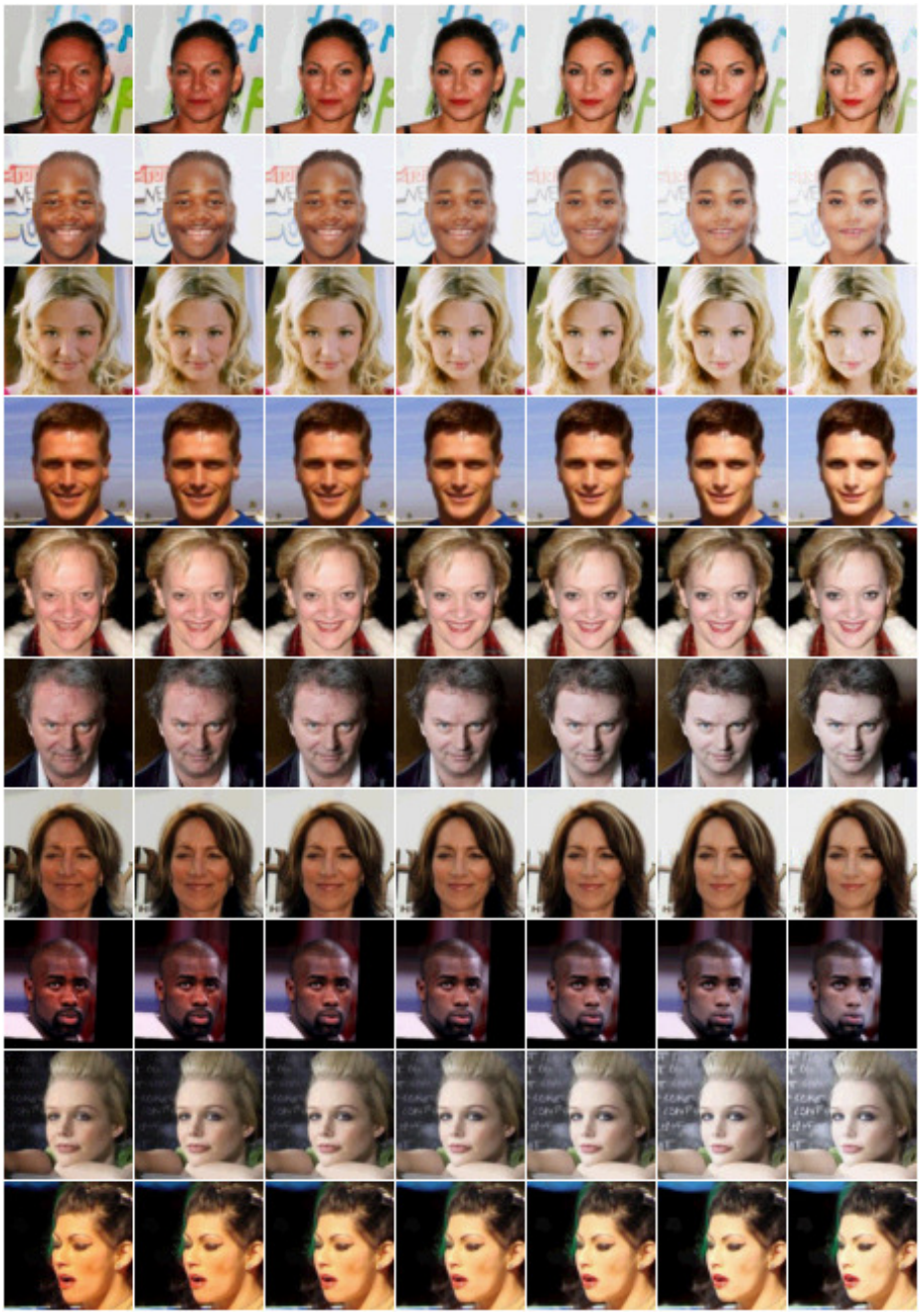

Fig. 9: Similar individuals from $S^{\text{in}}(\boldsymbol{x})$ obtained by simultaneously varying the sensitive attributes `Pale_Skin` + `Young`.

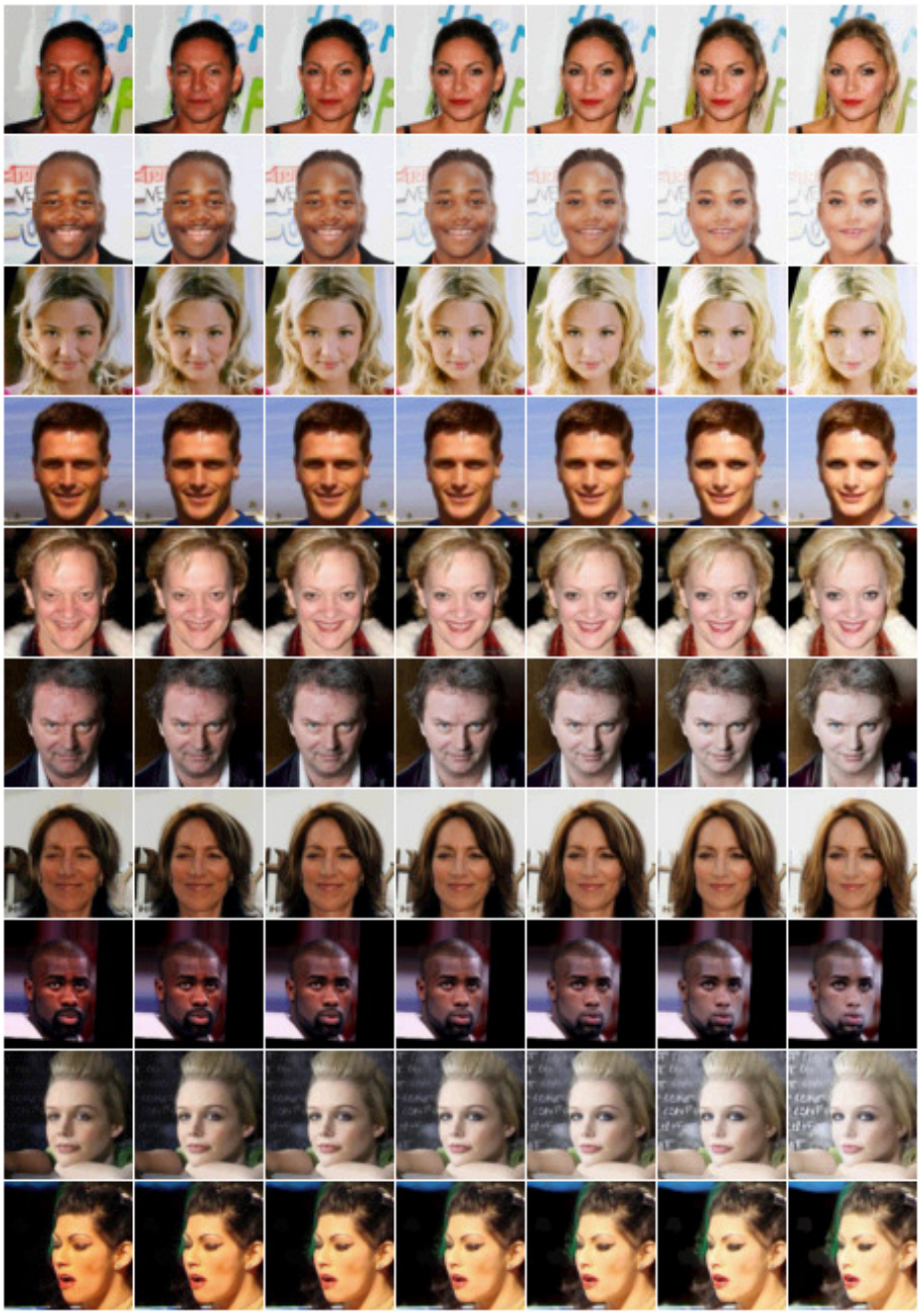

Fig. 10: Similar individuals from $S^{\text{in}}(\boldsymbol{x})$ obtained by simultaneously varying the sensitive attributes `Pale_Skin` + `Young` + `Blond`.

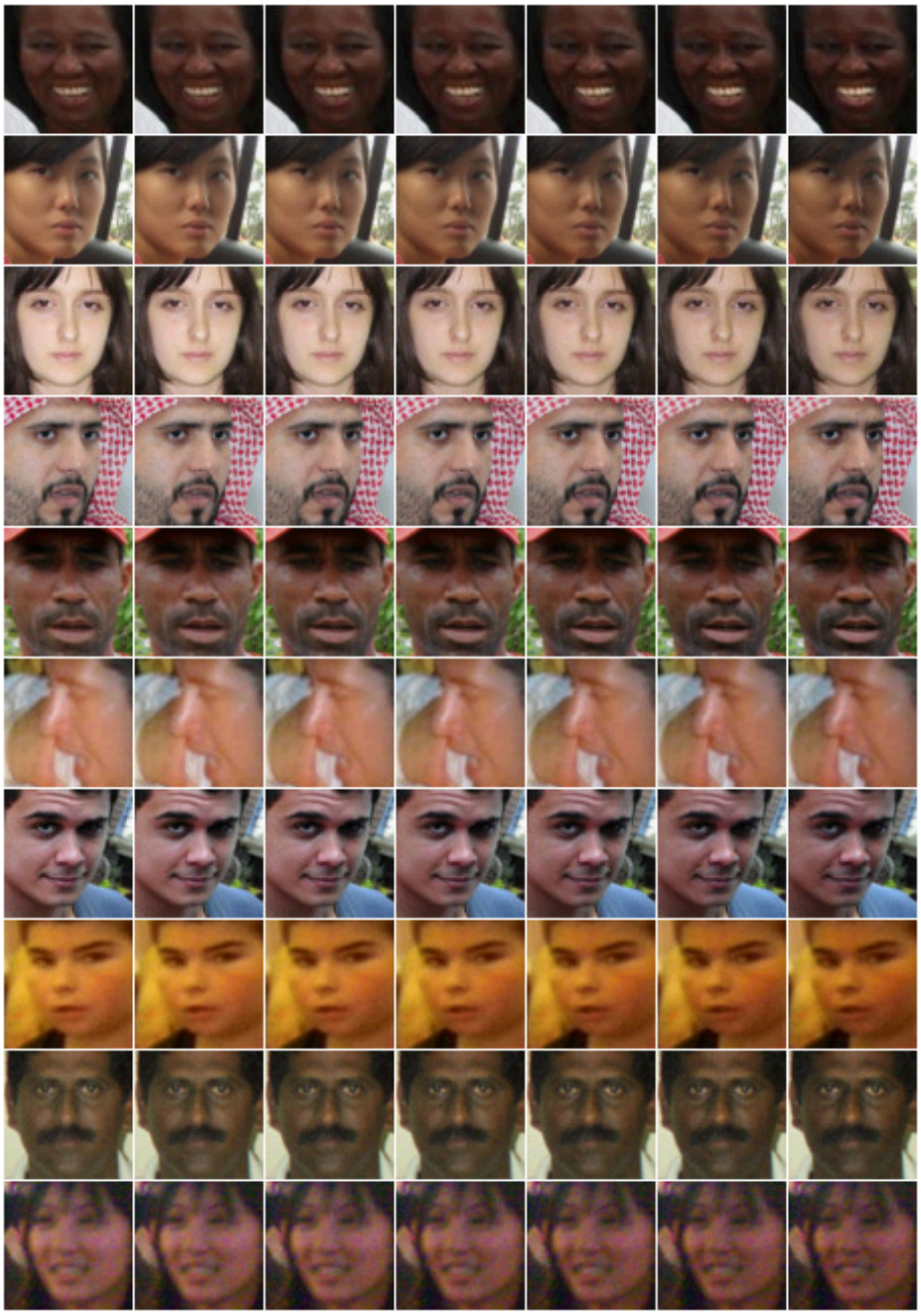

Fig. 11: Similar individuals from $S^{\text{in}}(\boldsymbol{x})$, for $\boldsymbol{x}$ in FairFace and $\epsilon = 0.5$, obtained by varying the sensitive attribute `Race=Black`.

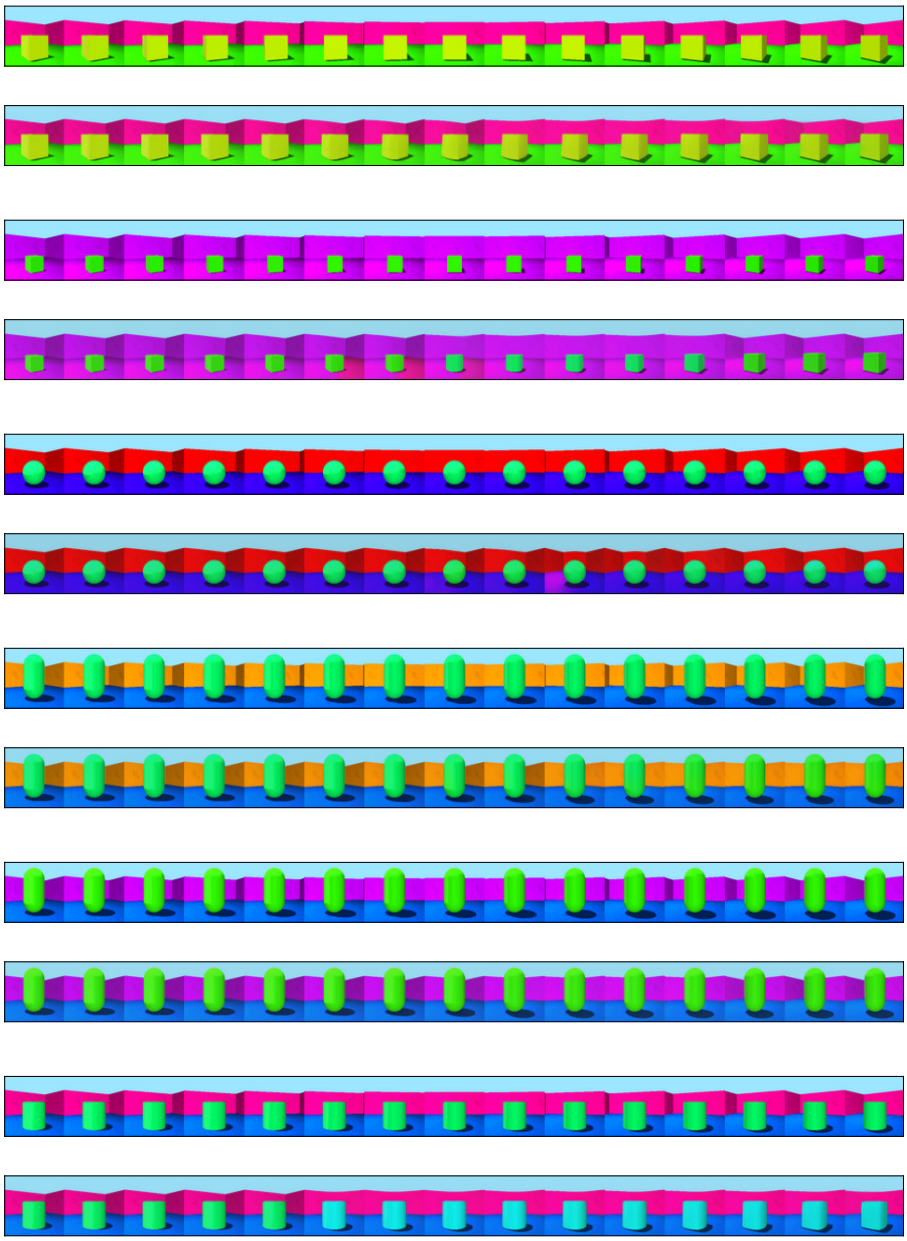

Fig. 12: Sampled shapes at 15 different ground truth orientations. The original (above) and the corresponding reconstructions (below) obtained from interpolating along one of the attribute vectors, $\boldsymbol{a}_{1,15}$ (see App. F for details), grouped together.