Appendix: Fair Contrastive Learning for Facial Attribute Classification

Sungho Park[1] Jewook Lee[1] Pilhyeon Lee[1] Sunhee Hwang[2] Dohyung Kim[3] Hyeran Byun[1*]

[1]Yonsei University [2]LG Uplus [3]SK Inc.

✉ qkrtjdgh18@yonsei.ac.kr

A. Definition of Ideally Biased Dataset

To confine a wide variety of data bias, we first define an ideally biased dataset that satisfies the following conditions.

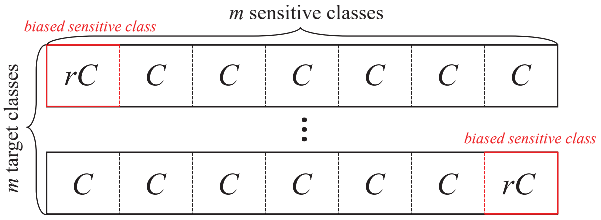
1. The dataset has m target and sensitive classes (*i.e.*, $N_t = N_s = m$). Each target and sensitive class contains the same number of data.

2. Target classes are biased to sensitive classes with a one-to-one mapping. That is, each target class has only one *biased sensitive class*, and no more than one target class has the same *biased sensitive class*.

3. In each target class, *biased sensitive class* has r times more data than other sensitive classes.

4. Target classes are highly biased to sensitive classes (*i.e.*, $r \geq m^2$).

We illustrated it in Figure 1, where the number of data in non-biased classes is set to C. All the proof below is based on this ideally biased dataset.



rC	C	C	C	C	C	C
			⋮			
C	C	C	C	C	C	rC

Figure 1. **The composition of the ideally biased dataset.** It has m target and sensitive classes, and each target class has only one *biased sensitive class*. C represents the number of data in non-biased classes and r is the bias for *biased sensitive classes*.

B. Mathematical Proof on Theorem 1

In the main paper, we demonstrated that *SupCon* will lead the encoding networks to learn sensitive attribute information based on Theorem 1. We provide the mathematical proof for the theorem below.

Assumption 1

Let input data come from the ideally biased dataset (refer to Sec. A), where $\tilde{X}$, $\tilde{Y}$, $\tilde{S}$ denote input images, target class labels, and sensitive attribute labels, respectively. We note that target classes are highly correlated with sensitive attributes in the dataset ($r \geq m^2$).

Definition 1

Learning of sensitive attribute information indicates an increase of $I(Z;\tilde{S})$, where $I(Z;\tilde{S}) = \mathbb{E}_{P(z,\tilde{s})} \log \frac{P(z,\tilde{s})}{P(z)P(\tilde{s})}$ and Z denotes the visual representation.

Assumption 2

Let t_l, t_m be random points in training time when $I(Z^{t_l};\tilde{S}) < I(Z^{t_m};\tilde{S})$.

Axiom 1

Given $\tilde{X}$, $\tilde{Y}$, and $\tilde{S}$, for all z_i, $|Z_p(i)| = Cr + (m-1)C - 1$, which is a constant.

Definition 2

$$L_a^{sup} = \sum_{z_i \in Z} \sum_{z_p \in Z_p(i)} \left[\log \left(\sum_{z_a \in Z_a(i)} \phi_a \right) \right].$$

Definition 3

$$L_p^{sup} = \sum_{z_i \in Z} \sum_{z_p \in Z_p(i)} \log \phi_p.$$

Proposition 1

$L^{Sup} = \hat{C}(-L_p^{sup} + L_a^{sup})$, where $\hat{C}$ is a constant.

Proof.

$$\begin{aligned} L^{Sup} &= -\sum_{z_i\in Z}\frac{1}{|Z_p(i)|}\sum_{z_p\in Z_p(i)}\log\frac{\phi_p}{\sum_{z_a\in Z_a(i)}\phi_a} \\ &= -\sum_{z_i\in Z}\frac{1}{|Z_p(i)|}\sum_{z_p\in Z_p(i)}\log\phi_p \\ &\quad + \sum_{z_i\in Z}\frac{1}{|Z_p(i)|}\sum_{z_p\in Z_p(i)}\log\sum_{z_a\in Z_a(i)}\phi_a \\ &= \frac{1}{|Z_p(i)|}\Big(-L_p^{sup}+L_a^{sup}\Big) \\ &= \hat{C}\Big(-L_p^{sup}+L_a^{sup}\Big)(\because \text{Axiom1}). \end{aligned} \tag{1}$$

Definition 4

Let $V_x^{t_l}$ and $V_x^{t_m}$ be the values of $L_x^{sup},\ x\in\{p,a\}$, at t_l and t_m, respectively.

For example, the value of L_a^{sup} at a certain point in training time, t_l, can be represented as:

$$V_a^{t_l}=\sum_{z_i\in Z}\sum_{z_p\in Z_p(i)}\left[\log\Big(\sum_{z_a\in Z_a(i)}\phi_a^{t_l}\Big)\right], \tag{2}$$

where $\phi_x^{t_k}=\exp(z_i^{t_k}\cdot z_x^{t_k}/\tau),\ x\in\{p,a\},\ k\in\{l,m\}$.

Definition 5

Let $Z_x(i)$=$Z_x^s(i)\cup Z_x^d(i)$, where $Z_x^s(i)=\{z_x\in Z_x(i)|\ \tilde{s}_x=\tilde{s}_i,\}$ and $Z_x^d(i)=\{z_x\in Z_x(i)|\tilde{s}_x\neq\tilde{s}_i\},\ x\in\{p,a\}$.

Proposition 2

From Definition 2, 3, 4 and 5,

$$\begin{aligned} V_a^{t_k} &= \sum_{z_i\in Z}\sum_{z_p\in Z_p(i)}\left[\log\Big(\sum_{z_a\in Z_a(i)}\phi_a^{t_k}\Big)\right] \\ &= \sum_{z_i\in Z}\sum_{z_p\in Z_p(i)}\left[\log\Big(\sum_{z_a\in Z_a^s(i)}\phi_a^{t_k}+\sum_{z_a\in Z_a^d(i)}\phi_a^{t_k}\Big)\right]. \end{aligned} \tag{3}$$

$$\begin{aligned} V_p^{t_k} &= \sum_{z_i\in Z}\left[\sum_{z_p\in Z_p(i)}\log\phi_p^{t_k}\right] \\ &= \sum_{z_i\in Z}\left[\sum_{z_p\in Z_p^s(i)}\log\phi_p^{t_k}+\sum_{z_p\in Z_p^d(i)}\log\phi_p^{t_k}\right]. \end{aligned} \tag{4}$$

Conjecture 1

a) $\sum_{z_a\in Z_a^s(i)}\phi_a^{t_l}<\sum_{z_a\in Z_a^s(i)}\phi_a^{t_m}$

b) $\sum_{z_a\in Z_a^d(i)}\phi_a^{t_l}>\sum_{z_a\in Z_a^d(i)}\phi_a^{t_m}$

c) $\sum_{z_p\in Z_p^s(i)}\log\phi_p^{t_l}<\sum_{z_p\in Z_p^s(i)}\log\phi_p^{t_m}$

d) $\sum_{z_p\in Z_p^d(i)}\log\phi_p^{t_l}>\sum_{z_p\in Z_p^d(i)}\log\phi_p^{t_m}$

From Assumption 2, $I(Z^{t_l};\tilde{S})<I(Z^{t_m};\tilde{S})$, hence the similarity between z_i and $Z_k^s(i)$ is larger at t_m than t_l. Meanwhile, the similarity between z_i and $Z_k^d(i)$ is smaller at t_m than at t_l.

Proposition 3

Let $\alpha_{z_x},\beta_{z_x}\in R^+,\ x\in\{p,a\}$, then

$$\begin{aligned} V_a^{t_m} &= \sum_{z_i\in Z}\sum_{z_p\in Z_p(i)}\left[\log\Big(\sum_{z_a\in Z_a^s(i)}(1+\alpha_{z_a})\phi_a^{t_l}\right. \\ &\qquad\left.+\sum_{z_a\in Z_a^d(i)}(1-\beta_{z_a})\phi_a^{t_l}\Big)\right], \end{aligned} \tag{5}$$

$$\begin{aligned} V_p^{t_m} &= \sum_{z_i\in Z}\left[\sum_{z_p\in Z_p^s(i)}\log(1+\alpha_{z_p})\phi_p^{t_l}\right. \\ &\qquad\left.+\sum_{z_p\in Z_p^d(i)}\log(1-\beta_{z_p})\phi_p^{t_l}\right], \end{aligned} \tag{6}$$

where α_{z_x} is an increasing rate of similarity between an anchor and each sample from t_l to t_m. Conversely, β_{z_x} is the decreasing rate of similarity.

proof.

By Conjecture 1,

$$\begin{aligned} \sum_{z_a\in Z_a^s(i)}\phi_a^{t_m} &= \sum_{z_a\in Z_a^s(i)}(1+\alpha_{z_a})\phi_a^{t_l}, \\ \sum_{z_a\in Z_a^d(i)}\phi_a^{t_m} &= \sum_{z_a\in Z_a^d(i)}(1-\beta_{z_a})\phi_a^{t_l}, \\ \sum_{z_p\in Z_p^s(i)}\log\phi_p^{t_m} &= \sum_{z_p\in Z_p^s(i)}\log(1+\alpha_{z_p})\phi_p^{t_l}, \\ \sum_{z_p\in Z_p^d(i)}\log\phi_p^{t_m} &= \sum_{z_p\in Z_p^d(i)}\log(1-\beta_{z_p})\phi_p^{t_l}. \end{aligned} \tag{7}$$

Therefore,

$$\begin{aligned} V_a^{t_m} &= \sum_{z_i\in Z}\sum_{z_p\in Z_p(i)}\left[\log\Big(\sum_{z_a\in Z_a^s(i)}\phi_a^{t_m}+\sum_{z_a\in Z_a^d(i)}\phi_a^{t_m}\Big)\right] \\ &= \sum_{z_i\in Z}\sum_{z_p\in Z_p(i)}\left[\log\Big(\sum_{z_a\in Z_a^s(i)}(1+\alpha_{z_a})\phi_a^{t_l}\right. \\ &\qquad\left.+\sum_{z_a\in Z_a^d(i)}(1-\beta_{z_a})\phi_a^{t_l}\Big)\right], \end{aligned} \tag{8}$$

$$
\begin{aligned}
V_p^{t_m} &= \sum_{z_i \in Z} \left[\sum_{z_p \in Z_p^s(i)} \log \phi_p^{t_m} + \sum_{z_p \in Z_p^d(i)} \log \phi_p^{t_m} \right] \\
&= \sum_{z_i \in Z} \left[\sum_{z_p \in Z_p^s(i)} \log(1+\alpha_{z_p}) \phi_p^{t_l} \right. \\
&\qquad \left. + \sum_{z_p \in Z_p^d(i)} \log(1-\beta_{z_p}) \phi_p^{t_l} \right].
\end{aligned}
\tag{9}
$$

Assumption 3

Let $\overline{\alpha_{z_x}}$ be the mean increasing rate of similarity (*i.e.*, α_{z_x}) over $Z_x^s(i)$ and $\overline{\beta_{z_x}}$ be the mean decreasing rate of similarity (*i.e.*, β_{z_x}) over $Z_x^d(i)$, then $\overline{\alpha_{z_x}} \approx \overline{\beta_{z_x}}$.

Definition 6

In Eq. 5, let the mean $\phi_a^{t_l}$ over $Z_a^s(i)$ be $\overline{\phi_a^{t_l}}^s$ and that over $Z_a^d(i)$ be $\overline{\phi_a^{t_l}}^d$.

Assumption 4

Let the difference between $\overline{\phi_a^{t_l}}^s$ and $\overline{\phi_a^{t_l}}^d$ by sensitive attribute information be $\epsilon \in R^+$. Then, $\overline{\phi_a^{t_l}}^s \approx \overline{\phi_a^{t_l}}^d + \epsilon$, where $\epsilon \ll \overline{\phi_a^{t_x}}^d$, $\overline{\phi_a^{t_x}}^s$.

Lemma 1

Given $\tilde{X}$, $\tilde{Y}$, and $\tilde{S}$, for all t_l, t_m, $V_a^{t_l} \geq V_a^{t_m}$.

proof.

From Proposition 3,

$$
\begin{aligned}
V_a^{t_m} &= \sum_{z_i \in Z} \sum_{z_p \in Z_p(i)} \left[\log \Big(\sum_{z_a \in Z_a^s(i)} (1+\alpha_{z_a}) \phi_a^{t_l} \right. \\
&\qquad \left. + \sum_{z_a \in Z_a^d(i)} (1-\beta_{z_a}) \phi_a^{t_l} \Big) \right] \\
&\approx \sum_{z_i \in Z} \sum_{z_p \in Z_p(i)} \left[\log \Big(\sum_{z_a \in Z_a^s(i)} (1+\overline{\alpha_{z_a}}) \phi_a^{t_l} \right. \\
&\qquad \left. + \sum_{z_a \in Z_a^d(i)} (1-\overline{\beta_{z_a}}) \phi_a^{t_l} \Big) \right].
\end{aligned}
\tag{10}
$$

Note that $\overline{\alpha_{z_a}}$ and $\overline{\beta_{z_a}}$ are defined in Assumption 3. Then we compare $V_a^{t_m}$ and $V_a^{t_l}$ as follows.

$$
\begin{aligned}
\Delta V_a &= V_a^{t_m} - V_a^{t_l} \\
&\approx \sum_{z_i \in Z} \sum_{z_p \in Z_p(i)} \left[\log \left(\frac{\sum_{z_a \in Z_a^s(i)} (1+\overline{\alpha_{z_a}}) \phi_a^{t_l}}{\sum_{z_a \in Z_a(i)} \phi_a^{t_l}} \right. \right. \\
&\qquad \left. \left. + \frac{\sum_{z_a \in Z_a^d(i)} (1-\overline{\beta_{z_a}}) \phi_a^{t_l}}{\sum_{z_a \in Z_a(i)} \phi_a^{t_l}} \right) \right] \\
&= \sum_{z_i \in Z} \sum_{z_p \in Z_p(i)} \left[\log \left(1 + \frac{\sum_{z_a \in Z_a^s(i)} \overline{\alpha_{z_a}} \phi_a^{t_l}}{\sum_{z_a \in Z_a(i)} \phi_a^{t_l}} \right. \right. \\
&\qquad \left. \left. - \frac{\sum_{z_a \in Z_a^d(i)} \overline{\beta_{z_a}} \phi_a^{t_l}}{\sum_{z_a \in Z_a(i)} \phi_a^{t_l}} \right) \right].
\end{aligned}
\tag{11}
$$

By Definition 6, it is rephrased as follows.

$$
\begin{aligned}
\Delta V_a = \sum_{z_i \in Z} \sum_{z_p \in Z_p(i)} & \left[\log \left(1 + \frac{\sum_{z_a \in Z_a^s(i)} \overline{\alpha_{z_a}} \, \overline{\phi_a^{t_l}}^s}{\sum_{z_a \in Z_a(i)} \phi_a^{t_l}} \right. \right. \\
&\qquad \left. \left. - \frac{\sum_{z_a \in Z_a^d(i)} \overline{\beta_{z_a}} \, \overline{\phi_a^{t_l}}^d}{\sum_{z_a \in Z_a(i)} \phi_a^{t_l}} \right) \right].
\end{aligned}
\tag{12}
$$

From Assumption 4, $\overline{\phi_a^{t_x}}^s \approx \overline{\phi_a^{t_x}}^d + \epsilon$. Based on this, ΔV_a is approximated as follows.

$$
\begin{aligned}
\Delta V_a \approx \sum_{z_i \in Z} \sum_{z_p \in Z_p(i)} & \left[\log \left(1 + \frac{\Big(\sum_{z_a \in Z_a^s(i)} \overline{\alpha_{z_a}}}{\sum_{z_a \in Z_a(i)} \phi_a^{t_l}} \right. \right. \\
&\qquad \left. \left. \frac{- \sum_{z_a \in Z_a^d(i)} \overline{\beta_{z_a}} \Big) \overline{\phi_a^{t_l}}^d}{} \right) \right],
\end{aligned}
\tag{13}
$$

where we omit ϵ for readability since $\epsilon \ll \overline{\phi_a^{t_x}}^d$, $\overline{\phi_a^{t_x}}^s$. In the ideally biased dataset, regardless of z_i and z_p, $|Z_a^s(i)|$=$rC + (m-1)C$ and $|Z_a^d(i)| = (m-1)rC + (m-1)^2 C$. Thus, we can reformulate Eq. 13 as follows.

$$\Delta V_a = \sum_{z_i \in Z} \sum_{z_p \in Z_p(i)} \Bigg[\log \Bigg(1 + \frac{\Big((rC + (m-1)C)\overline{\alpha_{z_a}} - ((m-1)rC + (m-1)^2 C)\overline{\beta_{z_a}} \Big) \overline{\phi_a^{t_l}}^d}{\sum_{z_a \in Z_a(i)} \phi_a^{t_l}} \Bigg) \Bigg]$$
$$= \sum_{z_i \in Z} \sum_{z_p \in Z_p(i)} \Bigg[\log \Bigg(1 + \frac{(m+r-1)C\Big((\overline{\alpha_{z_a}} - (m-1)\overline{\beta_{z_a}})\overline{\phi_a^{t_l}}^d \Big)}{\sum_{z_a \in Z_a(i)} \phi_a^{t_l}} \Bigg) \Bigg]. \tag{14}$$

By Assumption 3, it is approximated as follows.

$$\Delta V_a \approx \sum_{z_i \in Z} \sum_{z_p \in Z_p(i)} \Bigg[\log \Bigg(1 + \frac{(m+r-1)C \times \Big(((2-m)\overline{\alpha_{z_a}})\overline{\phi_a^{t_l}}^d \Big)}{\sum_{z_a \in Z_a(i)} \phi_a^{t_l}} \Bigg) \Bigg] \leq 0. \tag{15}$$

Here, $m \geq 2$, $r > m^2$, $C > 0$ ($\because$ Assumption 1), $\overline{\alpha_{z_a}} > 0$ ($\because$ Proposition 3), and $\phi_a^{t_l} > 0$ ($\because$ Definition 4). Therefore, $\Delta V_a \leq 0$.

Lemma 2

Given $\tilde{X}$, $\tilde{Y}$, and $\tilde{S}$, for all t_l, t_m, $V_p^{t_l} < V_p^{t_m}$.

proof.

By Proposition 3,

$$V_p^{t_m} = \sum_{z_i \in Z} \Bigg[\sum_{z_p \in Z_p^s(i)} \log(1 + \alpha_{z_p})\phi_p^{t_l} + \sum_{z_p \in Z_p^d(i)} \log(1 - \beta_{z_p})\phi_p^{t_l} \Bigg]$$
$$\approx \sum_{z_i \in Z} \Bigg[\sum_{z_p \in Z_p^s(i)} \log(1 + \overline{\alpha_{z_p}})\phi_p^{t_l} + \sum_{z_p \in Z_p^d(i)} \log(1 - \overline{\beta_{z_p}})\phi_p^{t_l} \Bigg]. \tag{16}$$

Similar to Eq. 11, we compare $V_p^{t_m}$ and $V_p^{t_l}$ as follows.

$$\Delta V_p = V_p^{t_m} - V_p^{t_l} = \sum_{z_i \in Z} \Bigg[\sum_{z_p \in Z_p^s(i)} \log \frac{(1 + \overline{\alpha_{z_p}})\phi_p^{t_l}}{\phi_p^{t_l}} + \sum_{z_p \in Z_p^d(i)} \log \frac{(1 - \overline{\beta_{z_p}})\phi_p^{t_l}}{\phi_p^{t_l}} \Bigg]. \tag{17}$$

Here, $\log(1 - \overline{\alpha_{z_p}}) \approx \log(1 - \overline{\beta_{z_p}})$ ($\because$ Assumption 3), and $\log(1 - \overline{\alpha_{z_p}}) \approx -\log(1 + \overline{\alpha_{z_p}})$ since $\log(1) = 0$ and $\frac{\mathrm{d}\log(1)}{\mathrm{dx}} = 1$. Therefore, $\log(1 + \overline{\alpha_{z_p}}) \approx -\log(1 - \overline{\beta_{z_p}})$. Based on this, we can approximate ΔV_p as follows.

$$\Delta V_p \approx \log(1 + \overline{\alpha_{z_p}}) \Big(\sum_{z_i \in Z} \sum_{z_p \in Z_p^s(i)} \mathbb{1} - \sum_{z_i \in Z} \sum_{z_p \in Z_p^d(i)} \mathbb{1} \Big), \tag{18}$$

where $\mathbb{1}$ is an indicator function. In the ideally biased dataset, $\sum_{z_i \in Z} \sum_{z_p \in Z_p^s(i)} \mathbb{1} = (rC)^2 + (m-1)C^2$ and $\sum_{z_i \in Z} \sum_{z_p \in Z_p^d(i)} \mathbb{1} = 2(m-1)rC^2 + (m-1)(m-2)C^2$. Therefore, we rephrase it as follows.

$$\Delta V_p = \Big(\Big((rC)^2 + (m-1)C^2 \Big) - \Big(2(m-1)rC^2 + (m-1)(m-2)C^2 \Big) \Big) \log(1 + \overline{\alpha_{z_p}})$$
$$= C^2 \Big(r^2 + (-2m+1)r - m^2 + 4m - 3 \Big) \log(1 + \overline{\alpha_{z_p}})$$
$$= C^2 \Big((r + \lambda m)(r - ((2+\lambda)m - 1)) + (4-\lambda)m - 3 \Big) \times \log(1 + \overline{\alpha_{z_p}}) > 0 \quad \text{s.t.} \quad r > (2+\lambda)m - 1 \tag{19}$$

where $\lambda = -1 + \sqrt{2}$. Finally, $\Delta V_p > 0$ since $m > 2$, $r > m^2$, and $C > 0$ ($\because$ Assumption 1).

Theorem 1

Given $\tilde{X}$, $\tilde{Y}$, and $\tilde{S}$, for all t_l, t_m, $V^{t_l} > V^{t_m}$.

proof.

From Lemma 1 and 2, $V_a^{t_l} \geq V_a^{t_m}$ and $V_p^{t_l} < V_p^{t_m}$ for all t_l, t_m. Since $V^{t_k} = \hat{C}(-V_p^{t_k} + V_a^{t_k})$ by Proposition 1, $V^{t_l} > V^{t_m}$ for all t_l, t_m.

Corollary 1

Learning sensitive attribute information decreases L^{Sup}, given $\tilde{X}$, $\tilde{Y}$, and $\tilde{S}$.

proof.

From Definition 1, learning of sensitive attribute information equals to the increase of $I(Z; \tilde{S})$. In addition, the increase of $I(Z; \tilde{S})$ corresponds to a transition from t_l to

t_m since $I(Z;\tilde{S})$ is always higher at t_m than at t_l ($\because$ Assumption 2). Finally, V^{t_m} is always smaller than V^{t_l} ($\because$ Theorem 1), therefore, learning sensitive attribute information decreases L^{Sup}.

Method	Adversarial Training	EO (↓)	Acc. (↑)
SupCon	✗	30.5±1.3	80.5±0.7
	✓	20.0±0.3	77.2±0.1
FSCL+	✗	**6.5±0.4**	79.1±0.1
	✓	20.5±0.4	77.8±0.2

Table 1. **Effectiveness of adversarial training in classifier training stage on CelebA.** We set *attractiveness* and *male* to the target class and sensitive attribute, respectively.

C. Fairness Strategy in Classifier Training Stage

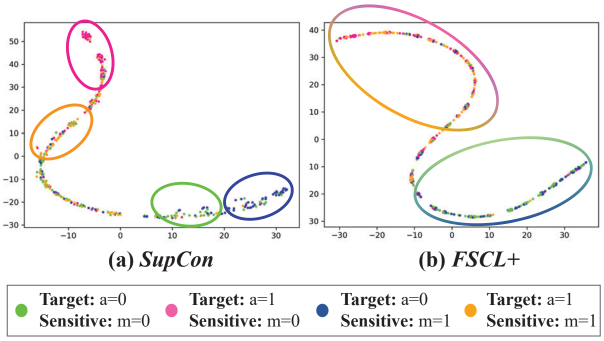
In Table 1, we explore the effectiveness of applying *GRL* [17] in the classifier training stage, after finishing the representation learning with *SupCon* and *FSCL+*. To this end, we deploy an additional classifier for the sensitive attribute and do not freeze the encoder and projection networks in the second stage. As might be expected, *GRL* improves the fairness of *SupCon* by sacrificing the classification accuracy. Meanwhile, it degrades EO as well as the classification accuracy in ours. We speculate that it is because the fair representation learned by *FSCL+* becomes biased by re-training the encoding networks with the cross entropy loss and *GRL*. The similar results of EO and top-1 accuracy between *SupCon* with *GRL* and *FSCL+* with *GRL* support that the learned representation is almost renewed in the classifier training stage. In conclusion, the results show that applying the additional strategy for fairness in the classifier training stage is not effective to our method.

D. Modification for Incomplete Supervised Setting

To apply our method to the environment where target class labels are not provided, we introduce $FSCL^{\dagger}$, which a modified version of *FSCL*. We set a positive sample to another patch from the same image with an anchor and negative samples to Z_{ig} and Z_{tg}. It is formulated as follows.

$$FSCL^{\dagger} = -\sum_{z_i \in Z} \log \frac{exp(z_i \cdot z_p/\tau)}{\sum_{z_f^* \in Z_f^*(i)} exp(z_i \cdot z_f^*/\tau)}, \quad (20)$$

where $Z_f^*(i) = \{z_f^* \in Z | \hat{s}_f^* = \hat{s}_i\}$. Except for the loss function, the overall structure is the same as the original.



(a) *SupCon* (b) *FSCL+*

Figure 2. **t-SNE visualizations with random intialization.**

E. Details of t-SNE Visualization

For the t-SNE [23] visualization, we exploit the models pre-trained on CelebA dataset [14] for 100 epochs. Then we obtain 50 random samples (*i.e.*, representation) per data group with the pre-trained models. Before applying the t-SNE algorithm, we reduce the dimensionality of the samples using PCA reduction. We tune the hyperparameters in the scikit-learn implementation as follows.

- Perplexity: from 10 to 40 by 1
- Learning rate: 10 or 100
- Iteration= 100, 1000, or 10000

We set the perplexity, learning rate, and iterations 10, 10, and 10000 respectively, but in all the cases, we note that representation learned by *FSCL+* is more agnostic to the sensitive attribute than that learned by *SupCon*. Furthermore, we provide t-SNE plots without PCA reduction in Figure 2 since it considerably affects the structure of representations.

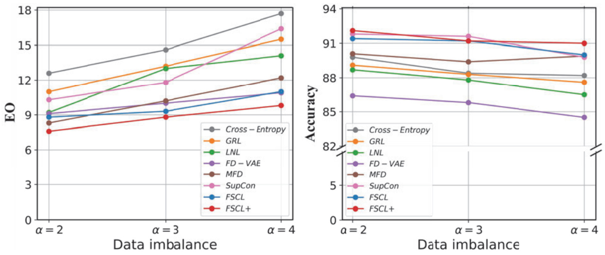


Figure 3. **Classification results on UTK Face dataset.** We set *gender* and *age* to the target class and sensitive attribute, respectively. It shows trends of classification accuracy and equalized odds (EO) at different α.

F. Further Experiments on UTK Face

In Figure 3, we provide experimental results on UTK Face with the other sensitive attribute, *age*. It shows that *FSCL+* maintain the fairest EO and the best top-1 accuracy at all α.

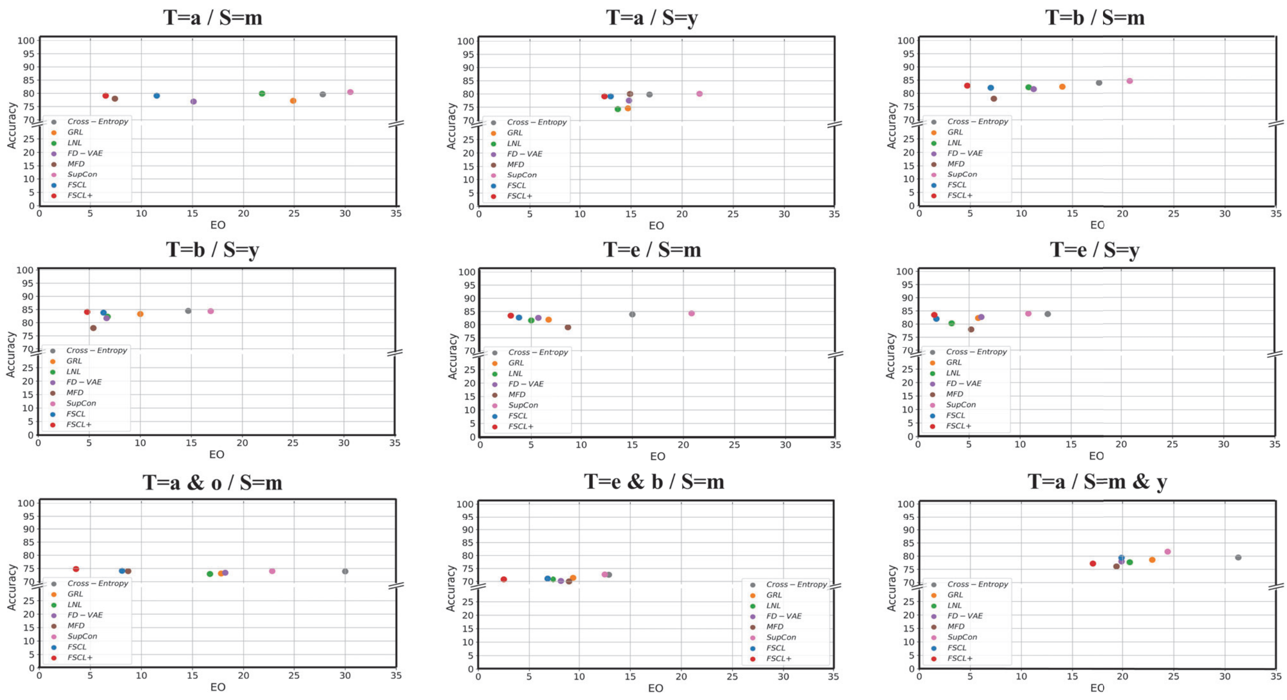


Figure 4. **Experimental results in figure form on CelebA dataset.** It shows the trade-off performances between ACC. and EO more clearly. The upper left corner of the plots corresponds to the optimal trade-off performance.

Attributes	CE [7]		*GRL* [17]		*LNL* [13]		*FD-VAE* [16]		*MFD* [11]		*SupCon* [12]		*FSCL*		*FSCL+*	
	EO	Acc.	EO	Acc.	EO	Acc.	EO	Acc.	EO	Acc.	EO	Acc.	EO	Acc.	EO	Acc.
T=*a* / S=*m*	27.8$_{\pm0.2}$	79.6$_{\pm0.5}$	24.9$_{\pm0.3}$	77.2$_{\pm0.5}$	21.8$_{\pm0.4}$	79.9$_{\pm0.5}$	15.1$_{\pm0.1}$	76.9$_{\pm0.0}$	7.4$_{\pm0.3}$	78.0$_{\pm0.3}$	30.5$_{\pm1.3}$	80.5$_{\pm0.7}$	11.5$_{\pm0.3}$	79.1$_{\pm0.4}$	**6.5**$_{\pm0.4}$	79.1$_{\pm0.4}$
T=*a* / S=*y*	16.8$_{\pm0.3}$	79.8$_{\pm0.4}$	14.7$_{\pm0.4}$	74.6$_{\pm0.4}$	13.7$_{\pm0.3}$	74.3$_{\pm0.4}$	14.8$_{\pm0.2}$	77.5$_{\pm0.1}$	14.9$_{\pm0.4}$	80.0$_{\pm0.3}$	21.7$_{\pm1.0}$	80.1$_{\pm0.8}$	13.0$_{\pm0.6}$	79.1$_{\pm0.5}$	**12.4**$_{\pm0.5}$	79.1$_{\pm0.5}$
T=*b* / S=*m*	17.6$_{\pm0.3}$	84.0$_{\pm0.3}$	14.0$_{\pm0.3}$	82.5$_{\pm0.5}$	10.7$_{\pm0.2}$	82.3$_{\pm0.4}$	11.2$_{\pm0.1}$	81.6$_{\pm0.3}$	7.3$_{\pm0.2}$	78.0$_{\pm0.3}$	20.7$_{\pm0.5}$	84.6$_{\pm0.6}$	7.0$_{\pm0.4}$	82.1$_{\pm0.3}$	**4.7**$_{\pm0.5}$	82.9$_{\pm0.4}$
T=*b* / S=*y*	14.7$_{\pm0.1}$	84.5$_{\pm0.3}$	10.0$_{\pm0.2}$	83.3$_{\pm0.5}$	6.8$_{\pm0.3}$	82.3$_{\pm0.5}$	6.7$_{\pm0.2}$	81.7$_{\pm0.0}$	5.4$_{\pm0.1}$	78.0$_{\pm0.2}$	16.9$_{\pm0.9}$	84.4$_{\pm0.8}$	6.4$_{\pm0.4}$	83.8$_{\pm0.4}$	**4.8**$_{\pm0.3}$	84.1$_{\pm0.5}$
T=*e* / S=*m*	15.0$_{\pm0.3}$	83.9$_{\pm0.2}$	6.7$_{\pm0.4}$	81.9$_{\pm0.6}$	5.0$_{\pm0.3}$	81.6$_{\pm0.3}$	5.7$_{\pm0.0}$	82.6$_{\pm0.1}$	8.7$_{\pm0.3}$	79.0$_{\pm0.4}$	20.8$_{\pm1.1}$	84.3$_{\pm0.5}$	3.8$_{\pm0.3}$	82.7$_{\pm0.3}$	**3.0**$_{\pm0.4}$	83.4$_{\pm0.6}$
T=*e* / S=*y*	12.7$_{\pm0.2}$	83.8$_{\pm0.3}$	5.9$_{\pm0.4}$	82.3$_{\pm0.4}$	3.3$_{\pm0.4}$	80.3$_{\pm0.6}$	6.2$_{\pm0.1}$	84.0$_{\pm0.2}$	5.2$_{\pm0.2}$	78.0$_{\pm0.2}$	10.8$_{\pm1.0}$	84.0$_{\pm0.7}$	1.8$_{\pm0.3}$	82.0$_{\pm0.4}$	**1.6**$_{\pm0.3}$	83.5$_{\pm0.3}$
T=*a & o* / S=*m*	30.0$_{\pm0.2}$	73.9$_{\pm0.5}$	17.8$_{\pm0.2}$	73.1$_{\pm0.5}$	16.7$_{\pm0.4}$	72.9$_{\pm0.5}$	18.2$_{\pm0.1}$	73.4$_{\pm0.1}$	8.7$_{\pm0.4}$	74.0$_{\pm0.3}$	22.8$_{\pm0.7}$	74.0$_{\pm0.5}$	8.1$_{\pm0.3}$	74.1$_{\pm0.3}$	**3.6**$_{\pm0.3}$	74.8$_{\pm0.4}$
T=*b & e* / S=*m*	12.9$_{\pm0.2}$	72.6$_{\pm0.4}$	9.4$_{\pm0.3}$	71.4$_{\pm0.4}$	7.4$_{\pm0.2}$	70.8$_{\pm0.5}$	8.2$_{\pm0.1}$	70.2$_{\pm0.2}$	9.0$_{\pm0.1}$	70.0$_{\pm0.1}$	12.5$_{\pm0.8}$	72.7$_{\pm0.9}$	6.8$_{\pm0.4}$	71.1 $_{\pm0.2}$	**2.5**$_{\pm0.6}$	70.8$_{\pm0.5}$
T=*a* / S=*m & y*	31.3$_{\pm0.3}$	79.5$_{\pm0.4}$	22.9$_{\pm0.4}$	78.6$_{\pm0.5}$	20.7$_{\pm0.3}$	77.7$_{\pm0.5}$	19.9$_{\pm0.0}$	78.0$_{\pm0.1}$	19.4$_{\pm0.2}$	76.1$_{\pm0.3}$	24.4$_{\pm1.3}$	81.7$_{\pm0.7}$	19.9$_{\pm0.5}$	79.4$_{\pm0.3}$	**17.0**$_{\pm0.5}$	77.2$_{\pm0.5}$

Table 2. **Classification results on CelebA.** We further specify the standard deviation in this table.

Although FD-VAE [16] achieves similar EO with *FSCL*, its accuracy is significantly inferior to ours. It indicates that ours highly outperform it in terms of the trade-off performance between fairness and accuracy.

G. Additional Experimental Results on CelebA

To clearly show the trade-off performances between classification accuracy and fairness, we plot the experimental results on CelebA in Figure 4. *FSCL+* achieves the best trade-off performances in all the results. Furthermore, we supplement the experimental results by reporting standard deviation in Table 2.

H. Dataset Composition

H.1. CelebA and UTK Face

In CelebA [14], we conduct experiments in terms of a variety of target and sensitive attribute pairs. Table 3 shows the specific composition of the training set in all the settings. In UTK Face [27], we involve 10,000, 2,400, and 2,400 data in the training, validation, and test sets, respectively. We provide the various compositions of the training set according to α in Table 4.

H.2. Dogs and Cats

Similar to UTK Face, we leverage 3,425 black cat and white dog images, and 685 white cat and black dog images for training. The test set includes 2,400 images which are

CelebA								
	a=0	a=1		b=0	b=1		e=0	e=1
m=0	29,920	64,589	m=0	84,954	9,555	m=0	84,963	9,546
m=1	49,247	19,014	m=1	39,475	28,786	m=1	44,527	23,734
	a=0	a=1		b=0	b=1		e=0	e=1
y=0	30,618	5,364	m=0	19,164	16,818	m=0	22,146	13,836
y=1	48,549	78,239	m=1	105,265	21,523	m=1	107,344	19,444
	a=0	a=1		m=0	m=1		m=0	m=1
m=0, y=0	7,522	3,645	a=0, o=0	13,995	27,966	b=0, e=0	78,613	30,481
m=1, y=0	23,096	1,719	a=1, o=0	30,943	11,380	b=1, e=0	6,350	14,046
m=0, y=1	22,398	60,944	a=0, o=1	15,925	21,281	b=0, e=1	6,341	8,994
m=1, y=1	26,151	17,295	a=1, o=1	33,646	7,634	b=1, e=1	3,205	14,740

Table 3. **Composition of the training set of CelebA.** a, b, e, o, m, and y denote *attractiveness*, *bignose*, *bags-under-eyes*, *mouth-slightly-open*, *male*, and *young*, respectively.

UTK Face				
	$\alpha = 2$ / $\alpha = 3$ / $\alpha = 4$			
	Ethinicity		Age	
	Caucasian	Others	More than 35	Others
Female	1,666 / 1,250 / 1,000	3,334 / 3,750 / 4,000	1,666 / 1,250 / 1,000	3,334 / 3,750 / 4,000
Male	3,334 / 3,750 / 4,000	1,666 / 1,250 / 1,000	3,334 / 3,750 / 4,000	1,666 / 1,250 / 1,000

Table 4. **Composition of the training set of UTK Face.** α denotes the intensities of data imbalance.

completely balanced. We note that it is different from the original setting in [13]. In the study, the target attribute and bias are completely correlated in the training set. For instance, cats are always black and dogs are always white. Although they solved the task by utilizing the pixel-level of bias labels (*i.e.*, RGB values of each pixel), it is an almost unsolvable problem with only the image-level of labels since the target attribute and bias labels are always the same at the image-level. Therefore, we designed the task more reasonable to validate fairness methods which mostly exploit the image-level of labels.

H.3. Discussion on License and Data Collection

Both CelebA [14] and UTK Face [27] have a non-standard license (i.e, Custom (non-commericial)), but the creators clarify the datasets are available for non-commercial research purposes only.

CelebA consists of the images collected from Celeb-Faces dataset [21] and attribute labels. According to [21], the images are collected by searching names of celebrities on the web. Also in UTK Face, the creators combine the images from CACD [3] and Morph [10] datasets with the images crawled in Bing and Google search engines. In both CACD and Morph, the images are gathered by searching on the web.

I. Implementation Details

I.1. Structure of Comparable Models

Cross-Entropy [7] , ***GRL*** [17], ***LNL*** [13]: The models utilize ResNet-18 [7] for backbone networks and a MLP with one hidden layer for classifiers. The dimensions of representation are the same as ours. *GRL* and *LNL* are reproduced based on [13, 17], and the hyperparameter to determine a weight for the reversed gradient is searched in the range from 0.01 to 0.1 in each experiment. For *LNL*, hyperparameter λ for regularization loss is searched in the range from 0.01 to 0.1 in each experiment. For all the models, we train them in an end-to-end manner for 100 epochs.

FD-VAE [16]: We build the model with the same structure as the original paper [16] without the encoder network. For a fair comparison, we substitute the encoder network to ResNet-18 and obtain better reproduction performances. Following the paper, we separate each latent space to have the same dimensions to each other and set hyperparameter β to 1. The other hyperparameters are found by grid searching and set to $\alpha = 1$, $\gamma = 5$, and $\lambda = 1$ for all the experiments. For representation learning, we train the encoder networks for 100 epochs. After that, we train the classifiers for downstream tasks for 10 epochs.

MFD [11]: We implement the model with source code released by the authors. The teacher and student models both leverage ResNet-18 for backbone networks and a MLP with one hidden layer for a classifier. Following the original paper, we train the models for 50 epochs and set hyperparameter λ to 7 and 5 for CelebA and UTK Face, respectively. For Dogs and Cats, λ is determined as 7 through grid searching.

SupCon [12], ***SimCLR*** [4], ***FSCL*** **(ours)**: We implement *SupCon* and *SimCLR* with source code released by the authors of [12], and *FSCL* is also based on the code (which is licensed under the terms of the MIT license). The models use ResNet-18 [7] for the encoder network and a MLP with two hidden layers for the projection network, which have 256 hidden nodes.

I.2. Augmentation Strategy and Experimental Setup

For the models based on contrastive loss, we augment two patches per image. Except for this, we use the same augmentation strategy [4] for all the models. Specifically, we sequentially and randomly apply cropping and resizing, horizontal flipping, color jittering, and gray scaling.

For all the models, we set the identical environments of SGD optimizer with momentum [18], batch sizes of 128, and learning rate of 0.1. All the experiments are based on the PyTorch library and are conducted in a Linux environment with 4 NVIDIA Titan Xp GPUs with 12GB of memory.

Method	Regularization	EO (↓)	Acc. (↑)
GDRO	Standard	$21.3_{\pm 1.0}$	$76.3_{\pm 0.2}$
	Early Stopping	**4.0** $_{\pm 0.1}$	$74.7_{\pm 0.1}$
	Strong L_2 (lr=0.1)	$8.7_{\pm 2.6}$	$76.3_{\pm 0.1}$
	Strong L_2 & Group adjustments (C=5)	$8.0_{\pm 2.0}$	$77.1_{\pm 0.2}$
FSCL+	Standard	$6.5_{\pm 0.4}$	$79.1_{\pm 0.1}$

Table 5. **Comparison with GDRO on CelebA.** We set *attractiveness* and *male* to the target class and sensitive attribute, respectively.

J. Comparison with GDRO

GDRO [19] is one of the state-of-the-art methods to minimize the performance gaps between data groups and has a goal similar to our group-wise normalization. Thus, we report comparison results with GDRO in Table 5. Following the original paper, we search for the best C in the range of [0, 5]. The results show that ours achieves a better trade-off performance than GDRO.

K. Two kinds of Supervised Contrastive Losses

In this section, we summarize two kinds of supervised contrastive losses (*i.e.*, L^{sup}_{out} and L^{sup}_{in}) proposed in [12] and why we leverage L^{sup}_{out} as our baseline. Unlike L^{sup}_{out} (*i.e.*, L^{Sup} in the main paper), L^{sup}_{in} places the summation over positive samples and the normalization factor inside the log as follows.

$$L^{Sup}_{in} = -\sum_{z_i \in Z} \log \left(\frac{1}{|Z_p(i)|} \sum_{z_p \in Z_p(i)} \frac{\phi_p}{\sum_{z_a \in Z_a(i)} \phi_a} \right). \tag{21}$$

In the loss, the normalization factor works as a constant (*i.e.*, $-\sum_{z_i \in Z} \log \frac{1}{|Z_p(i)|}$), so it cannot normalize the imbalance in the positive samples. As the result, L^{sup}_{in} is more vulnerable to the data bias and shows inferior classification performances to L^{sup}_{out}. For these reasons, we utilize the latter as our baseline.

L. Discussion on Limitations

In this section, we discuss two limitations of our study. The first one is that our work is confined to the image classification task. We discuss it by explaining why we cover the task in this paper. One reason is that the superior performance of our baselines (*i.e.*, SupCon and SimCLR) has been experimentally validated in the image classification task [4, 12]. Therefore, through the task, we can make a fair comparison with the models and convincingly demonstrate our improvement over them. The other reason is that image classification is a fundamental and common task not only in contrastive representation learning [4, 12, 22, 26] but in fairness studies in the field of computer vision [5, 16, 20, 25]. Although fair visual representation can be exploited in other tasks, such as object recognition [24], image-to-image translation [8, 9], face recognition [2, 6], and object detection [1], each of them requires a suitable notion of fairness [1, 24] and specialized architectures [6, 8, 9]. Therefore, to achieve the best performance on the tasks, we also need to modify the proposed loss more appropriately for them. We leave the extension of *FSCL* to broader tasks for future work.

Second, our method essentially requires sensitive attribute labels to improve fairness. Even though supervision of the sensitive attribute labels is common in the literature on fair classification [5, 15, 16, 20], sometimes we cannot access the labels and it is laborious and expensive to annotate them. Although we show that our method can reduce such costs by effectively improving fairness using only a few labels, it cannot be utilized in the complete absence of the labels. Therefore, future works that develop a fair contrastive loss free of the sensitive attribute labels would make a significant contribution to the research community. We expect our study to be a bridgehead for them.

References

[1] Martim Brandao. Age and gender bias in pedestrian detection algorithms. *arXiv e-prints*, page arXiv:1906.10490, June 2019. 8

[2] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification.

In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR. 8

[3] Bor-Chun Chen, Chu-Song Chen, and Winston H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 768–783, Cham, 2014. Springer International Publishing. 7

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020. 8

[5] Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1436–1445, Long Beach, California, USA, 09–15 Jun 2019. PMLR. 8

[6] Sixue Gong, Xiaoming Liu, and Anil K. Jain. Jointly debiasing face recognition and demographic attribute estimation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIX*, volume 12374 of *Lecture Notes in Computer Science*, pages 330–347. Springer, 2020. 8

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7, 8

[8] Sunhee Hwang and Hyeran Byun. Unsupervised image-to-image translation via fair representation of gender bias. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1953–1957, 2020. 8

[9] Sunhee Hwang, Sungho Park, Dohyung Kim, Mirae Do, and Hyeran Byun. Fairfacegan: Fairness-aware facial image-to-image translation. In *BMVC*, volume 2020, 2020. 8

[10] S. M. Seitz I. Kemelmacher-Shlizerman, S. Suwajanakorn. Illumination-aware age progression. In *CVPR*, 2014. 7

[11] Sangwon Jung, Donggyu Lee, Taeeon Park, and Taesup Moon. Fair feature distillation for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12115–12124, June 2021. 6, 7

[12] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc., 2020. 6, 8

[13] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 6, 7

[14] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 5, 6, 7

[15] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3384–3393. PMLR, 10–15 Jul 2018. 8

[16] Sungho Park, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. Learning disentangled representation for fair facial attribute classification via fairness-aware information alignment. *Proceedings of AAAI-2021*, 2021. 6, 7, 8

[17] Edward Raff and Jared Sylvester. Gradient reversal against discrimination: A fair neural network learning approach. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 189–198, 2018. 5, 6, 7

[18] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016. 8

[19] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. 8

[20] P. Sattigeri, S. C. Hoffman, V. Chenthamarakshan, and K. R. Varshney. Fairness gan: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63(4/5):3:1–3:9, 2019. 8

[21] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Hybrid deep learning for face verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10):1997–2009, 2016. 7

[22] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 776–794, Cham, 2020. Springer International Publishing. 8

[23] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, June 2015. 5

[24] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5309–5318, 2019. 8

[25] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8916–8925, 2020. 8

[26] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance

discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 8

[27] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 6, 7