Distributionally Generative Augmentation for Fair Facial Attribute Classification

Supplementary Material

The supplementary materials are organized as follows:

- In Appendix A, we give the proof for Theorem 1. Theorem 1 guarantees the existence of optimal combination coefficients, so that we can use grid search to find them;
- In Appendix B, as an empirical supplement to Theorem 1, we show our observations on synthetic dataset to reveal the relationship between β_{clf} (the bias of learned classifier in latent space) and λ (the regularization strength);
- In Appendix C, we present the additional results of bias detection on real facial dataset to more intuitively show why and how our approach works.
- In Appendix D, we present the implementation details.

Appendix A. Proof for theoretical justification

Proof: We first define the sample ratio of majority group and minority group as $p_{maj} = n_{maj}/(n_{maj} + n_{min})$ and $p_{min} = n_{min}/(n_{maj}+n_{min})$ respectively. The optimization objective $R(\boldsymbol{w})$ can be written as

$$\begin{aligned}
R(\boldsymbol{w}) =& \mathbb{E}_{(\boldsymbol{z},y)}[\log(1+\mathrm{e}^{-\mathrm{y}\boldsymbol{w}\boldsymbol{z}})] + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2 \\
=& \frac{p_{maj}}{2}\mathbb{E}_{\boldsymbol{z}_y\sim N(\mathbf{1},\sigma_y^2 I_d)}\mathbb{E}_{\boldsymbol{z}_s\sim N(\mathbf{1},\sigma_s^2 I_d)}[\log(1+\mathrm{e}^{-\boldsymbol{w}\boldsymbol{z}})] \\
&+\frac{p_{maj}}{2}\mathbb{E}_{\boldsymbol{z}_y\sim N(\text{-}\mathbf{1},\sigma_y^2 I_d)}\mathbb{E}_{\boldsymbol{z}_s\sim N(\text{-}\mathbf{1},\sigma_s^2 I_d)}[\log(1+\mathrm{e}^{\boldsymbol{w}\boldsymbol{z}})] \\
&+\frac{p_{min}}{2}\mathbb{E}_{\boldsymbol{z}_y\sim N(\mathbf{1},\sigma_y^2 I_d)}\mathbb{E}_{\boldsymbol{z}_s\sim N(\text{-}\mathbf{1},\sigma_s^2 I_d)}[\log(1+\mathrm{e}^{-\boldsymbol{w}\boldsymbol{z}})] \\
&+\frac{p_{min}}{2}\mathbb{E}_{\boldsymbol{z}_y\sim N(\text{-}\mathbf{1},\sigma_y^2 I_d)}\mathbb{E}_{\boldsymbol{z}_s\sim N(\mathbf{1},\sigma_s^2 I_d)}[\log(1+\mathrm{e}^{\boldsymbol{w}\boldsymbol{z}})] \\
&+\frac{\lambda}{2}\|\boldsymbol{w}\|_2^2.
\end{aligned} \tag{6}$$

Without loss of generality, we let $d = 1$. Then we have

$$\begin{aligned}
R(\boldsymbol{w}) =& \frac{p_{maj}}{2}\mathbb{E}_{z_y\sim N(1,\sigma_y^2),z_s\sim N(1,\sigma_s^2)}[\log(1+\mathrm{e}^{-\mathrm{w_y z_y}-\mathrm{w_s z_s}})] \\
&+\frac{p_{maj}}{2}\mathbb{E}_{z_y\sim N(-1,\sigma_y^2),z_s\sim N(-1,\sigma_s^2)}[\log(1+\mathrm{e}^{\mathrm{w_y z_y}+\mathrm{w_s z_s}})] \\
&+\frac{p_{min}}{2}\mathbb{E}_{z_y\sim N(1,\sigma_y^2),z_s\sim N(-1,\sigma_s^2)}[\log(1+\mathrm{e}^{-\mathrm{w_y z_y}-\mathrm{w_s z_s}})] \\
&+\frac{p_{min}}{2}\mathbb{E}_{z_y\sim N(-1,\sigma_y^2),z_s\sim N(1,\sigma_s^2)}[\log(1+\mathrm{e}^{\mathrm{w_y z_y}+\mathrm{w_s z_s}})] \\
&+\frac{\lambda}{2}\|\boldsymbol{w}\|_2^2 \\
=& p_{maj}\mathbb{E}_{z_y\sim N(1,\sigma_y^2),z_s\sim N(1,\sigma_s^2)}[\log(1+\mathrm{e}^{-\mathrm{w_y z_y}-\mathrm{w_s z_s}})] \\
&+p_{min}\mathbb{E}_{z_y\sim N(1,\sigma_y^2),z_s\sim N(1,\sigma_s^2)}[\log(1+\mathrm{e}^{-\mathrm{w_y z_y}+\mathrm{w_s z_s}})] \\
&+\frac{\lambda}{2}\|\boldsymbol{w}\|_2^2.
\end{aligned} \tag{7}$$

For convenience, we write $\mathbb{E}_{z_y\sim N(1,\sigma_y^2)}$ and $\mathbb{E}_{z_s\sim N(1,\sigma_s^2)}$ as $\mathbb{E}_{z_y}$ and $\mathbb{E}_{z_s}$ respectively without causing any ambiguity. Our goal is to minimize $R(w_y, w_s)$. So we focus on the gradients of classifier parameters w_y and w_s:

$$\begin{aligned}
&\nabla_{w_y} R(w_y, w_s) \\
=& p_{maj}\mathbb{E}_{z_y}\mathbb{E}_{z_s}[\frac{1}{(1+e^{w_y z_y + w_s z_s})}(-z_y)] \\
&+p_{min}\mathbb{E}_{z_y}\mathbb{E}_{z_s}[\frac{1}{(1+e^{w_y z_y - w_s z_s})}(-z_y)] \\
&+\lambda w_y
\end{aligned} \tag{8}$$

and

$$\begin{aligned}
&\nabla_{w_s} R(w_y, w_s) \\
=& p_{maj}\mathbb{E}_{z_y}\mathbb{E}_{z_s}[\frac{1}{(1+e^{w_y z_y + w_s z_s})}(-z_s)] \\
&+p_{min}\mathbb{E}_{z_y}\mathbb{E}_{z_s}[\frac{1}{(1+e^{w_y z_y - w_s z_s})}z_s] \\
&+\lambda w_s.
\end{aligned} \tag{9}$$

We use proof by contradiction. Let w_s^* be zero. Then we have

$$\begin{aligned}
&\nabla_{w_s} R(w_y^*, 0) \\
=& p_{maj}\mathbb{E}_{z_y}\mathbb{E}_{z_s}[\frac{1}{(1+e^{w_y^* z_y})}(-z_s)] \\
&+p_{min}\mathbb{E}_{z_y}\mathbb{E}_{z_s}[\frac{1}{(1+e^{w_y^* z_y})}z_s] \\
=& (-p_{maj})\mathbb{E}_{z_y}\mathbb{E}_{z_s}[\frac{1}{(1+e^{w_y^* z_y})}z_s] \\
&+(1-p_{maj})\mathbb{E}_{z_y}\mathbb{E}_{z_s}[\frac{1}{(1+e^{w_y^* z_y})}z_s] \\
=& (1-2p_{maj})\mathbb{E}_{z_y}\mathbb{E}_{z_s}[\frac{1}{(1+e^{w_y^* z_y})}z_s] \\
=& (1-2p_{maj})\mathbb{E}_{z_y}[\frac{1}{(1+e^{w_y^* z_y})}]\mathbb{E}_{z_s}[z_s] \\
=& (1-2p_{maj})\mathbb{E}_{z_y}[\frac{1}{(1+e^{w_y^* z_y})}] \\
<& 0.
\end{aligned} \tag{10}$$

Note that $\mathbb{E}_{z_y}[\frac{1}{(1+e^{w_y^* z_y})}] > 0$, so that the $\nabla_{w_s} R(w_y^*, 0) = 0$ if and only if the majority group sample ratio $p_{maj} = 1/2$ (*i.e.*, the data is unbiased). The above equation shows that the solution w_s^* cannot be zero. Similarly, we also have

$$\nabla_{w_y} R(0, w_s^*) < 0. \tag{11}$$

So the bias degree of the classifier $\beta_{clf} = \|w_s^*\|/\|w_y^*\| > 0$ if the data is biased (*i.e.*, $\beta = p_{maj} => 1/2$). Different values of λ will scale the impact of the regularization term, affecting the solution $\boldsymbol{w}^\star = (w_y^*, w_s^*)$ of logistic regression. Denote the solutions under regularization strength λ_1 and λ_2 are $\boldsymbol{w_1^\star} = (w_{y1}^*, w_{s1}^*)$ and $\boldsymbol{w_2^\star} = (w_{y2}^*, w_{s2}^*)$ respectively. As we have proven before, w_{y1}^*, w_{s1}^*, w_{y2}^*, and w_{s2}^* are not zero. Then we construct $c_1^* = w_{y2}^*/(w_{y2}^* w_{s1}^* - w_{y1}^* w_{s2}^*)$ and $c_2^* = w_{y1}^*/(w_{y2}^* w_{s1}^* - w_{y1}^* w_{s2}^*)$ such that $\boldsymbol{w}_{cmb} := c_1^* \boldsymbol{w}_1^* - c_2^* \boldsymbol{w}_2^* = [0, 1]$. Here we have completed the proof of the existence of the optimal combination coefficients. □

Appendix B. Observations on synthetic dataset

In this section, as an empirical supplement to Theorem 1, we explore the relationship between β_{clf} (bias of learned linear classifier in the latent space) and λ (regularization strength used in logistic regression) on synthetic dataset.

Experimental Setup. Following the previous studies [62], we use the same settings as in the theoretical justification. Specifically, target attribute $y \in \{1, -1\}$ and spurious attribute $s \in \{1, -1\}$ are binary. The training dataset contains $n = 20000$ samples, which can be divided into four groups: two majority groups with $s = y$, each containing $n_{maj}/2$ samples, and two minority groups with $s = -y$, each containing $n_{min}/2$ samples. In the latent space of generative models, each group has its own distribution over latent codes $\boldsymbol{z} = [\boldsymbol{z}_y, \boldsymbol{z}_s] \in \mathbb{R}^{200}$ consisting of stable features $\boldsymbol{z}_y \in \mathbb{R}^{100}$ generated from the target attribute y, and spurious features $\boldsymbol{z}_s \in \mathbb{R}^{100}$ generated from the spurious attribute s: $\boldsymbol{z}_y \mid y \sim N(y\mathbf{1}, \sigma_y^2 I_{100})$ and $\boldsymbol{z}_s \mid s \sim N(s\mathbf{1}, \sigma_s^2 I_{100})$. To get the classification boundary, we use logistic regression with regularization strength λ. Recall that the bias degree of the classifier as $\beta_{clf} = \|\boldsymbol{w}_s^*\|/\|\boldsymbol{w}_y^*\| \in [0, +\infty)$. We set different data bias by using different ratios $n_{maj} : n_{min}$. We also set different standard deviations for $\boldsymbol{z}_y$ and $\boldsymbol{z}_s$. All results were averaged over 100 random repetitions.

Observations. As shown in Table 6, in most cases, if we increase the regularization strength λ in logistic regression, the classifier bias β_{clf} will be larger. This observation motivates us to design a ***simple*** but ***effective*** method to obtain two different biased semantic directions in the latent space, that is to set different regularization strength λ.

Appendix C. Additional results on real dataset

In response to the above findings, we show the images edited by different semantic directions, obtained with different regularization strengths λ. The training dataset (sampled from CelebA) is biased where the target attribute $Smiling$ is spuriously correlated with the spurious attributes $Female$ and $Young$. We first use a trained generative model to encode the images into latent codes. Then we train linear classifiers in latent space using logistic regression with different λ. The semantic directions are normal

settings			regularization strength λ				
$n_{maj} : n_{min}$	σ_y	σ_s	1	10	100	1000	10000
2:1	0.1	0.1	0.027	0.032	0.039	0.051	0.072
	0.1	1.0	0.027	0.032	0.040	0.051	0.072
	1.0	0.1	0.026	0.031	0.039	0.051	0.073
	1.0	1.0	0.030	0.033	0.040	0.051	0.073
3:1	0.1	0.1	0.043	0.051	0.063	0.082	0.116
	0.1	1.0	0.043	0.051	0.063	0.082	0.116
	1.0	0.1	0.041	0.050	0.062	0.082	0.117
	1.0	1.0	0.044	0.051	0.063	0.082	0.117
4:1	0.1	0.1	0.054	0.065	0.080	0.104	0.148
	0.1	1.0	0.052	0.063	0.079	0.104	0.148
	1.0	0.1	0.052	0.063	0.079	0.104	0.150
	1.0	1.0	0.055	0.064	0.079	0.104	0.149
5:1	0.1	0.1	0.063	0.076	0.094	0.122	0.175
	0.1	1.0	0.063	0.075	0.093	0.122	0.174
	1.0	0.1	0.061	0.074	0.093	0.122	0.176
	1.0	1.0	0.064	0.075	0.093	0.122	0.175
6:1	0.1	0.1	0.071	0.085	0.105	0.137	0.197
	0.1	1.0	0.070	0.084	0.104	0.137	0.195
	1.0	0.1	0.069	0.083	0.104	0.137	0.199
	1.0	1.0	0.071	0.084	0.104	0.136	0.197
7:1	0.1	0.1	0.077	0.092	0.115	0.150	0.216
	0.1	1.0	0.077	0.092	0.114	0.149	0.214
	1.0	0.1	0.075	0.090	0.113	0.150	0.218
	1.0	1.0	0.077	0.091	0.113	0.149	0.216
8:1	0.1	0.1	0.082	0.099	0.123	0.162	0.233
	0.1	1.0	0.082	0.099	0.122	0.160	0.231
	1.0	0.1	0.080	0.097	0.122	0.161	0.235
	1.0	1.0	0.082	0.097	0.121	0.160	0.233
9:1	0.1	0.1	0.087	0.105	0.131	0.172	0.248
	0.1	1.0	0.087	0.105	0.130	0.171	0.246
	1.0	0.1	0.085	0.103	0.129	0.172	0.250
	1.0	1.0	0.087	0.103	0.129	0.171	0.248
10:1	0.1	0.1	0.092	0.110	0.138	0.181	0.262
	0.1	1.0	0.092	0.110	0.137	0.180	0.260
	1.0	0.1	0.089	0.108	0.136	0.181	0.264
	1.0	1.0	0.092	0.109	0.136	0.180	0.262
11:1	0.1	0.1	0.096	0.115	0.144	0.189	0.275
	0.1	1.0	0.096	0.115	0.143	0.188	0.272
	1.0	0.1	0.093	0.113	0.142	0.189	0.277
	1.0	1.0	0.096	0.114	0.142	0.188	0.275
12:1	0.1	0.1	0.100	0.120	0.150	0.197	0.286
	0.1	1.0	0.099	0.119	0.149	0.196	0.284
	1.0	0.1	0.097	0.118	0.148	0.197	0.289
	1.0	1.0	0.099	0.118	0.148	0.196	0.287
13:1	0.1	0.1	0.103	0.124	0.155	0.205	0.298
	0.1	1.0	0.103	0.124	0.154	0.203	0.294
	1.0	0.1	0.101	0.122	0.154	0.205	0.301
	1.0	1.0	0.103	0.123	0.153	0.203	0.298
14:1	0.1	0.1	0.107	0.128	0.160	0.211	0.308
	0.1	1.0	0.106	0.128	159	0.210	0.306
	1.0	0.1	0.104	0.126	0.159	0.212	0.311
	1.0	1.0	0.106	0.127	0.158	0.210	0.309

Table 6. Results of classifier bias β_{clf} on synthetic dataset. Empirically, in most cases, the classifier bias β_{clf} will be larger, if we increase the regularization strength λ in logistic regression.

vectors of the learned classification boundaries. As shown in Figure 10, a larger λ produces a larger bias in direction, resulting in a more obvious change in spurious attributes.

Appendix D. Implementation Details

For generative modeling, we utilize StyleGAN2 [32] for generator and e4e [67] for encoder. We use HFGI [69] algorithm to train generative models on training dataset with image size of 256 for 30 epochs. The size of features encoded by e4e is (18, 512), and we average over the channels to get latent codes with size of 512. We use regularized logistic regression to obtain directions, and the values of regularization strength λ are 1e+4 and 1e-4 respectively. To get the optimal combination coefficients, we perform grid search and use CLIP [57] as a reference model. More details about combination coefficients are shown in the next subsection. For representation model, we use ResNet-18 [23] for encoder and the representation dimensions are 512. We train the encoder for 135 epochs. We use Adam [36] as optimizer with learning rate 3e-4. We set the editing range [α_l,α_u] as [3,5]. For efficiency, we approximate the sampled degree as an integer. To complete the classification, we fix the encoder and train a linear classifier with Adam until convergence. The learning rate is 1e-2 with 1e-6 weight decay.

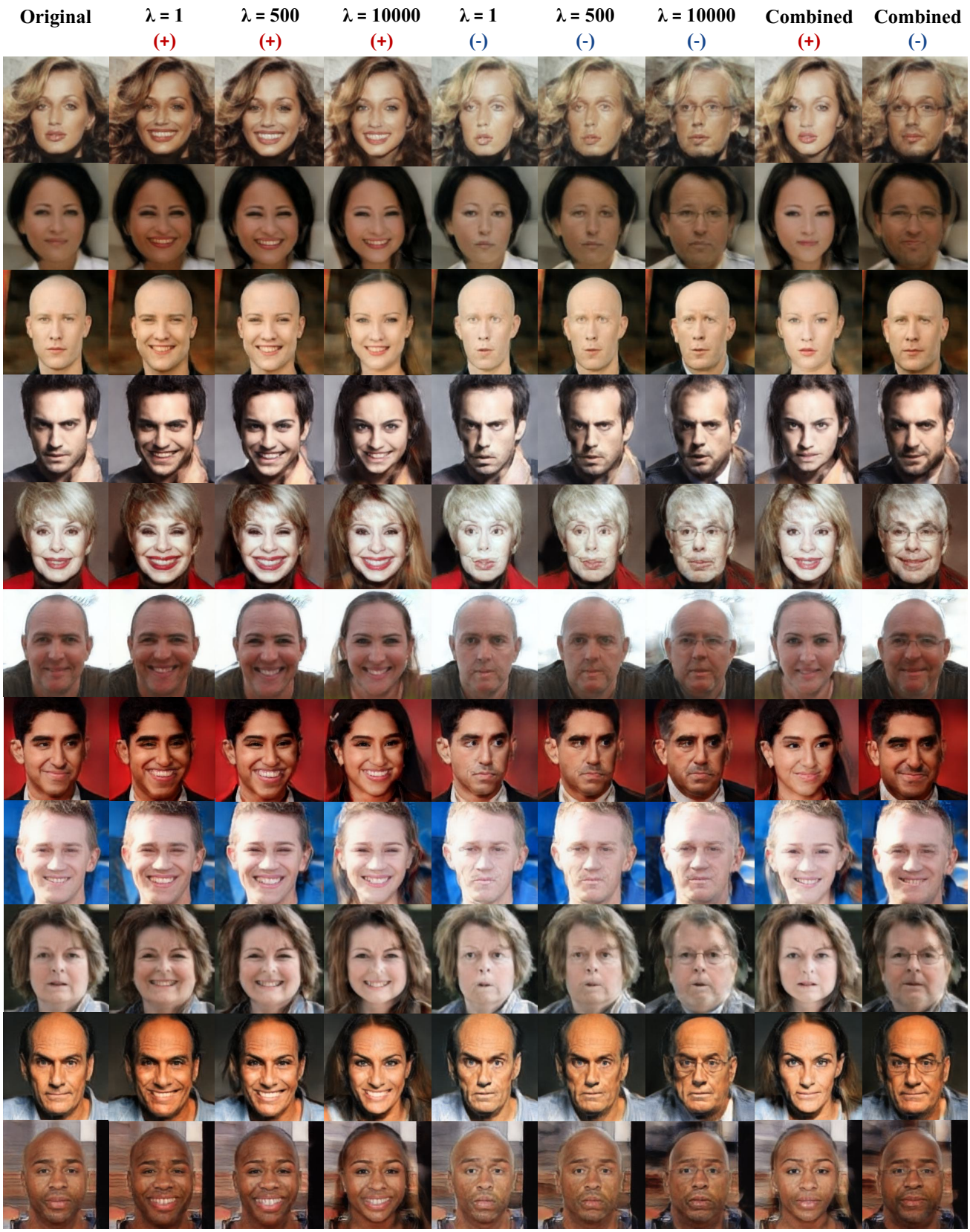

Figure 10. Illustration of images edited by different semantic directions, which are trained with different regularization strength λ.