A Binary Classifier Architecture

For all experiments, we train and use binary classifiers with the following architecture: We implement feedforward neural networks with 10 main layers, where each main layer consists of a convolution layer followed by batch normalization and ReLU activation. Additionally, we add pooling and dropout layers between main layers in an alternating fashion such that there are a total of 5 pooling layers and 3 dropout layers, where the dropout percentage is set to 0.5.

B Fairness Analysis of the StyleGAN2 FFHQ Model

To evaluate the fairness of the StyleGAN2 model trained on the FFHQ dataset, we started by generating 1000 random images. Then we used binary classifiers to label each image for the attributes *gender, smiling, eyeglasses*, and *young* for marginal and joint distributions (Table 1, Table 2). As can be seen, the StyleGAN2 model generates images that are slightly biased towards *Male=False*, moderately biased towards *Smiling=True* and strongly biased towards *Young=True* and *Eyeglasses=False* attributes. We also examine the joint distribution of attribute pairs such as *gender + eyeglasses*, *gender + smiling* and *eyeglasses + smiling*. As can be seen, the joint probability distribution of the attributes can be extremely imbalanced even if the marginal probability distributions of the individual attributes are not, such as the ratio of *women + eyeglasses* to *men + eyeglasses*. In Figure 1 and Figure 2, respectively, we show the percentage of assigned binary labels for single and multiple attributes.

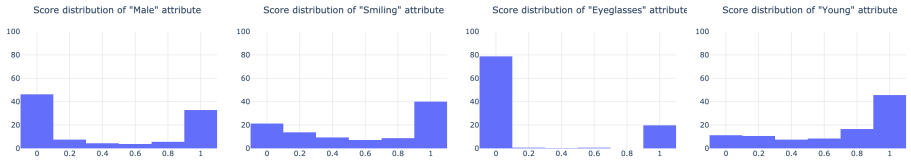

Fig. 1: Marginal probability distributions of ‘male‘, ‘smiling‘, ‘eyeglasses‘, ‘young‘ attributes sampled from images generated by StyleGAN2 pre-trained on the FFHQ dataset.

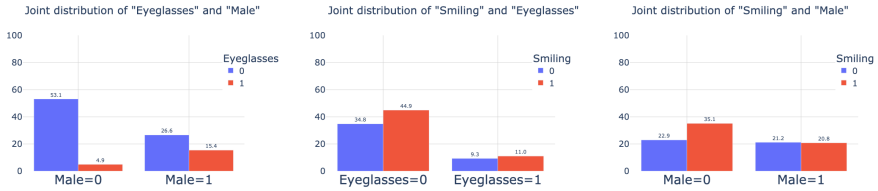

Fig. 2: Joint probability distributions of (‘male‘, ‘eyeglasses‘), (‘eyeglasses‘, ‘smiling‘), (‘male‘, ‘smiling‘) attribute pairs sampled from images generated by StyleGAN2 trained on the FFHQ dataset.

Table 1: Marginal distributions of attributes measured on the FFHQ dataset and images generated by StyleGAN2 pretrained on the FFHQ dataset.

Attribute	FFHQ	StyleGAN2
Eyeglasses	F=0.78, T=0.22	F=0.80, T=0.20
Young	F=0.28, T=0.72	F=0.30, T=0.70
Smiling	F=0.43, T=0.57	F=0.44, T=0.56
Male	F=0.58, T=0.42	F=0.58, T=0.42

Table 2: Joint distributions of attribute pairs measured on the FFHQ dataset and images generated by StyleGAN2 pretrained on the FFHQ dataset.

Attributes	FFHQ	StyleGAN2
Eyegl.-Smile	FF=0.34, FT=0.44 TF=0.09, TT=0.13	FF=0.35, FT=0.45 TF=0.09, TT=0.11
Smile-Male	FF=0.22, FT=0.36 TF=0.21, TT=0.21	FF=0.23, FT=0.35 TF=0.21, TT=0.21
Male-Eyegl.	FF=0.50, FT=0.08 TF=0.28, TT=0.14	FF=0.53, FT=0.05 TF=0.27, TT=0.15

C Additional debiasing results

We also performed debiasing for *eyeglasses* (Figure 3) and *afro hair* attribute (Figure 4) on the same latent codes showing the before/after of our debiasing method.

Fig. 3: A set of images generated with the same latent codes before and after debiasing the StyleGAN2 model with respect to the 'Eyeglasses' attribute on a single channel with our method.

We performed experiments with all attributes in the CelebA dataset, but we only included a subset in the paper for comparison with other methods. We

Fig. 4: A set of images generated with the same latent codes before and after debiasing the StyleGAN2 model with respect to the 'a person with afro hairstyle' text-based attribute with our method.

note that the attributes we performed comparisons on are the attributes that are common across the competitor methods and for which, reasonably high performing binary classifiers or high-quality labeled images are available. We use our first method to debias the additional attributes *Lipstick*, *Bangs* and *Beard* attributes and present the results in Table 3.

Table 3: KL Divergence between a uniform distribution and the distribution of images generated with our method. StyleGAN2 is included to demonstrate un-debiased results.

Method	Lipstick	Bangs	Beard
StyleGAN2	0.04	0.54	0.44
FairStyle	1.28×10^{-6}	1.62×10^{-6}	0.001

D Runtime Analysis

Our method directly debias the StyleGAN2 model within a short period of time. More specifically, the average time to debias a single attribute is 2.25 minutes, while debiasing joint attributes takes 4.2 minutes. Moreover similarly, classifying 1000 images with the CLIP classifier takes 1.28 minutes while the Celeb-A classifiers take 1.49 minutes on average.

E Comparison of CLIP vs Binary Classifiers

In terms of computation time, CLIP-based classifiers and binary classifiers perform similarly. The CLIP classifier takes 1.28 minutes while the Celeb-A classifiers take 1.49 minutes on average to classify 1000 images. However, CLIP provides a richer attribute set due to its large-scale and text-based nature. We perform an additional experiment to debias the 'smiling' attribute using both our first method and CLIP-based method. The CLIP-based method results in a KLD of 5.96×10^{-8}, which is comparable to our first method, which yields a KLD of 8×10^{-8}.

References

1. Abdal, R., Zhu, P., Mitra, N.J., Wonka, P.: Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. ArXiv **abs/2008.02401** (2021)
2. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., Wallach, H.M.: A reductions approach to fair classification. ArXiv **abs/1803.02453** (2018)
3. Azadi, S., Olsson, C., Darrell, T., Goodfellow, I.J., Odena, A.: Discriminator rejection sampling. ArXiv **abs/1810.06758** (2019)
4. Bau, D., Liu, S., Wang, T., Zhu, J.Y., Torralba, A.: Rewriting a deep generative model. ArXiv **abs/2007.15646** (2020)
5. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. CoRR **abs/1809.11096** (2018), http://arxiv.org/abs/1809.11096
6. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: FAT (2018)
7. Feldman, M.: Computational Fairness: Preventing Machine-Learned Discrimination. Ph.D. thesis, Haverford College (2015)
8. Goetschalckx, L., Andonian, A., Oliva, A., Isola, P.: Ganalyze: Toward visual definitions of cognitive image properties. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5744–5753 (2019)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27, pp. 2672–2680. Curran Associates, Inc. (2014), http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf
10. Grover, A., Choi, K., Shu, R., Ermon, S.: Fair generative modeling via weak supervision. In: ICML (2020)
11. Grover, A., Song, J., Agarwal, A., Tran, K., Kapoor, A., Horvitz, E., Ermon, S.: Bias correction of learned generative models using likelihood-free importance weighting. In: DGS@ICLR (2019)
12. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: NIPS (2016)
13. Härkönen, E., Hertzmann, A., Lehtinen, J., Paris, S.: Ganspace: Discovering interpretable gan controls. arXiv preprint arXiv:2004.02546 (2020)
14. Jahanian, A., Chai, L., Isola, P.: On the" steerability" of generative adversarial networks. arXiv preprint arXiv:1907.07171 (2019)

15. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. CoRR **abs/1812.04948** (2018), http://arxiv.org/abs/1812.04948
16. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 8107–8116 (2020)
17. Kocasari, U., Dirik, A., Tiftikci, M., Yanardag, P.: Stylemc: Multi-channel based fast text-guided image generation and manipulation. 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) pp. 3441–3450 (2022)
18. 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. ArXiv **abs/2104.13369** (2021)
19. Li, S., Araujo, I.B., Ren, W., Wang, Z., Tokuda, E.K., Junior, R.H., Cesar-Junior, R., Zhang, J., Guo, X., Cao, X.: Single image deraining: A comprehensive benchmark analysis (2019)
20. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. 2015 IEEE International Conference on Computer Vision (ICCV) pp. 3730–3738 (2015)
21. Louizos, C., Swersky, K., Li, Y., Welling, M., Zemel, R.S.: The variational fair autoencoder. CoRR **abs/1511.00830** (2016)
22. McDuff, D., Ma, S., Song, Y., Kapoor, A.: Characterizing bias in classifiers using generative models. arXiv preprint arXiv:1906.11891 (2019)
23. Oneto, L., Chiappa, S.: Fairness in machine learning. ArXiv **abs/2012.15816** (2020)
24. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. arXiv preprint arXiv:2103.17249 (2021)
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krüger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. ArXiv **abs/2103.00020** (2021)
26. Ramaswamy, V.V., Kim, S.S.Y., Russakovsky, O.: Fair attribute classification through latent space de-biasing. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 9297–9306 (2021)
27. Shen, Y., Yang, C., Tang, X., Zhou, B.: Interfacegan: Interpreting the disentangled face representation learned by gans. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
28. Shen, Y., Zhou, B.: Closed-form factorization of latent semantics in gans. arXiv preprint arXiv:2007.06600 (2020)
29. Sun, W., Chen, Z.: Learned image downscaling for upscaling using content adaptive resampler. IEEE Transactions on Image Processing **29**, 4027–4040 (2020). https://doi.org/10.1109/tip.2020.2970248, http://dx.doi.org/10.1109/TIP.2020.2970248
30. Tan, S., Shen, Y., Zhou, B.: Improving the fairness of deep generative models without retraining. ArXiv **abs/2012.04842** (2020)
31. Tanaka, A.: Discriminator optimal transport. In: NeurIPS (2019)
32. Tanielian, U., Issenhuth, T., Dohmatob, E., Mary, J.: Learning disconnected manifolds: a no gans land. ArXiv **abs/2006.04596** (2020)
33. Voynov, A., Babenko, A.: Unsupervised discovery of interpretable directions in the gan latent space. In: International Conference on Machine Learning. pp. 9786–9796. PMLR (2020)
34. Wang, T., Yang, X., Xu, K., Chen, S., Zhang, Q., Lau, R.: Spatial attentive single-image deraining with a high quality real rain dataset (2019)

35. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans (2017)
36. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemometrics and intelligent laboratory systems **2**(1-3), 37–52 (1987)
37. Woodworth, B.E., Gunasekar, S., Ohannessian, M.I., Srebro, N.: Learning non-discriminatory predictors. ArXiv **abs/1702.06081** (2017)
38. Wu, Z., Lischinski, D., Shechtman, E.: Stylespace analysis: Disentangled controls for stylegan image generation. arXiv preprint arXiv:2011.12799 (2020)
39. Yüksel, O.K., Simsar, E., Er, E.G., Yanardag, P.: Latentclr: A contrastive learning approach for unsupervised discovery of interpretable directions. arXiv preprint arXiv:2104.00820 (2021)
40. Zafar, M.B., Valera, I., Gomez-Rodriguez, M., Gummadi, K.P.: Fairness constraints: Mechanisms for fair classification. In: AISTATS (2017)
41. Zemel, R.S., Wu, L.Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: ICML (2013)
42. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan++: Realistic image synthesis with stacked generative adversarial networks. CoRR **abs/1710.10916** (2017), http://arxiv.org/abs/1710.10916
43. Zhu, J., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. CoRR **abs/1703.10593** (2017), http://arxiv.org/abs/1703.10593