# Matched sample selection with GANs for mitigating attribute confounding

Chandan Singh [1,4] Guha Balakrishnan [2,4] Pietro Perona [3,4]

[1] University of California at Berkeley
[2] Massachusetts Institute of Technology
[3] California Institute of Technology
[4] Amazon Web Services

## Abstract

*Measuring biases of vision systems with respect to protected attributes like gender and age is critical as these systems gain widespread use in society. However, significant correlations between attributes in benchmark datasets make it difficult to separate algorithmic bias from dataset bias. To mitigate such attribute confounding during bias analysis, we propose a* ***matching*** *[1] approach that selects a subset of images from the full dataset with balanced attribute distributions across protected attributes. Our matching approach first projects real images onto a generative adversarial network (GAN)'s latent space in a manner that preserves semantic attributes. It then finds image matches in this latent space across a chosen protected attribute, yielding a dataset where semantic and perceptual attributes are balanced across the protected attribute. We validate projection and matching strategies with qualitative, quantitative, and human annotation experiments. We demonstrate our work in the context of gender bias in multiple open-source facial-recognition classifiers and find that bias persists after removing key confounders via matching.*[1]

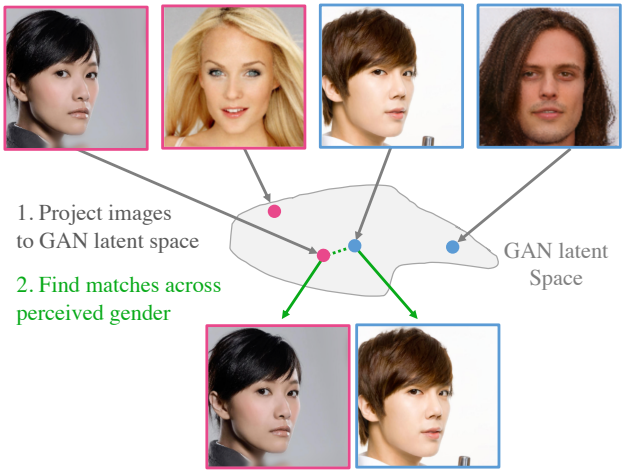


Figure 1. **Finding matched samples by GAN latent-space projection.** Given face images we look for a 'matched sample', i.e. a pair of face images that has similar attributes apart from one attribute (perceived gender here) which needs to be different. We propose a method for matching in GAN latent space (Sec 3.2) which requires projecting real face images to that space (Sec 3.1). The resulting matched samples can be used for benchmarking bias with respect to the protected attribute.

## 1. Introduction

Computer vision systems have applications in the entertainment, education, consumer, medical, security, and policing fields. In many applications, these systems can be an important factor used by humans to make impactful decisions. It is therefore important to minimize their potential biases with respect to protected attributes such as sex, gender, national origin, ethnicity, and age.

A key step in minimizing bias is measuring it. Benchmarking bias is not straightforward due to the presence of confounders in datasets. For example, male celebrities in the CelebA-HQ dataset [2, 3] are on average older and darker-skinned than their female counterparts; lighting, pose, expression and background may also be systematically different. These confounding factors can spuriously allocate inaccuracies in the model to the wrong attribute.

Random sampling of test data 'in the wild' is well-known to produce test sets that are not appropriate for causal inference due to confounding as explained above. A central concept in causal inference, used across domains is *causal matching* [1, 4], which selects unconfounded image pairs (i.e. the pairs are similar in aspects besides a specified protected attribute; see Fig 1). These pairs are then aggregated to yield a subset of the original dataset where semantic attributes are balanced across groups, thereby mitigating potential confounding. However, the matching process is difficult because, unlike with low-dimensional tabular data, image features are difficult and too numerous to specify and explicitly annotate.

[1] Code to reproduce the results here and apply the methods to new data is available at github.com/csinva/matching-with-gans.

Our approach addresses this challenge by exploiting the properties of the latent space of a pretrained GAN in a domain of interest, e.g., faces. We improve upon an existing technique for projecting real images onto StyleGAN2's latent "style" space [5, 6] by introducing a regularization penalty. We then show how to perform matching in this latent space and demonstrate with experiments that our approach can balance protected attributes such as race and age across subgroups, despite having no explicit access to these attribute labels. We then briefly show how the balanced data can be used to measure algorithmic bias – we benchmark bias of several open-source facial recognition systems with respect to a person's perceived gender[2]. We find that these systems are worse at identifying faces perceived as female, solidifying evidence measured on the original data (Fig 5).

This work's main contribution is a GAN-based method for producing a dataset of matched samples culled from a larger dataset of *real* face images. Each pair of samples differs by one selected attribute (e.g., gender) and is as similar as possible with respect to all other attributes (Sec 3.2). Additional contributions include (a) A method for projecting real images onto the latent space of StyleGAN2 (Sec 3.1), evaluated with detailed human experiments, (b) an experimental evaluation of our matching strategy showing that it successfully balances key covariates, and (c) an application to measuring algorithmic bias in face recognition (Sec 4.3) where we detect gender-based bias in popular models.

## 2. Related work

**Matching and causal inference** Matching is a popular and well-established technique in the field of causal inference [4]; it is used broadly across a variety of fields [7–12]. Our work builds on these ideas, but measures distance for matching using a pretrained neural network (StyleGAN2). This choice is motivated by observations that the StyleGAN2 latent space captures a wide variety of useful semantic attributes without any supervision.

Alternative causal analyses may require explicit values of the features over which we are trying to balance, such as age, background, and pose, which may be extremely difficult to collect and accurately measure. Moreover, most approaches, such as propensity score matching [13], do not work well when applied directly to low-level features like raw pixels. Classifying the probability of membership to a group, e.g., images perceived as male, directly from images is a difficult task for which models have shown considerable bias. Our approach overcomes this challenge by first embedding images in the StyleGAN2 latent space, in which a simple linear classifier can easily predict group membership. To our knowledge, this work is the first to propose image matching with GANs. Besides matching, recent works have attempted to use neural networks to aid in causal inference, e.g. by learning more invariant / balanced representations [14–22], using generative models [23–27], learning causal features [28, 29], or calculating adjustments [30–32].

**Analyzing bias in computer vision models** A long line of work has analyzed bias in computer vision [33–48], often focusing on discrepancies in performance (e.g. error rates) across protected attributes (e.g. gender). The datasets used in these works often contain many sources of possible confounding: combinations of attributes are disproportionately represented and/or correlated [49–51]. Recent approaches to mitigating dataset bias include collecting more comprehensive samples [43], synthesizing images to compensate for distribution gaps [42], weighting examples [52], and explicit annotation [53], photoshopping images while preserving various attributes [54], and synthetically manipulating images [55].

## 3. Methods

### 3.1. Projecting images onto StyleGAN2's style space

We build on the approach taken in a recent work [6] which optimizes the latent code to minimize the perceptual distance (measured using VGG16 embeddings [56, 57]) between the original image and the projected image. The optimization can be performed in StyleGAN2's original restricted style space $\mathbf{z} \in \mathbb{R}^{512}$ or in an expanded space formed by concatenating the style spaces of each decoder level, $Z_E \in \mathbb{R}^{18\times512}$. To explore this tradeoff, we cast the projection task as an optimization problem with the following objective: $Z_E^* = \text{argmin}_{Z_E} D(G(Z_E), x) + \lambda \sum_j ||Z_{E,j} - \bar{Z}_E||_2^2$, where $D(\cdot,\cdot)$ is the perceptual distance measure, $G$ is the GAN generator, $x$ is the image, $||Z_{E,j} - \bar{Z}_E||_2^2$ is a regularizer, and $\lambda$ is a positive scalar. The regularizer penalizes the deviation of each row of the expanded style matrix $Z_E$ towards the mean of its rows.

We quantify this tradeoff by sweeping $\lambda$ in Fig 2. Reconstructions using the restricted style space are poor in terms of both perceptual distance (blue curve, measured by VGG perceptual distance) and facial id distance (orange curve, measured by the public dlib facial recognition model [58]) between the original and reconstructed images. Unregularized reconstructions $Z_E$ yield projections far from the original restricted domain (green curve), making distance comparisons less reliable for matching. Setting $\lambda = 0.1$ achieves both good reconstruction and semantic structure in the latent space. Appendix A shows examples of how reconstruction quality and semantic structure are retained.

[2]We use the gender labels that were manually annotated in the curation of the CelebA dataset [3]. These annotations do not necessarily reflect the *gender identity* of the person, rather they refer to *binarized gender as perceived by a casual observer*. While this difference is immaterial to the techniques we propose, we prefer to make this transparent and therefore we refer to 'perceived gender' rather than 'gender'.

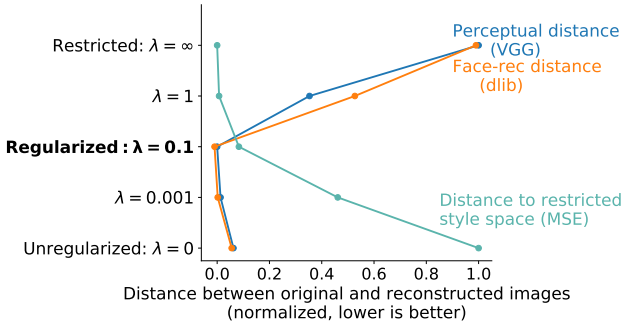


Figure 2. **Regularization improves projection quality**. Setting $\lambda = 0.1$ yields a modest improvement in both the perceptual distance between the original/reconstructed images (blue curve) and the facial identity distance (orange curve). Increasing regularization brings the latent projections closer to the restricted style space (green curve), as measured by the regularized term $\sum_j ||Z_{E,j} - \bar{Z}_E||_2^2$. This allows the latent space to better preserve semantic properties, which are useful for matching. Each curve is normalized to take minimum value 0 and maximum value 1. These results are averaged over the first 300 images in CelebA-HQ.

### 3.2. GAN-distance matching

We assume a dataset (e.g., face images) $x_1, ..., x_n$ and specify a single attribute $a$, sometimes referred to as the "matching attribute" or "treatment variable," we would like to analyze for bias. In our experiments, $a$ is gender and we find matches based on the following objective:

$$\text{Match}(x_i, A) = \underset{j}{\text{argmin}} \overbrace{||Z_{E,i} - Z_{E,j}||_{\text{F}}}^{\text{GAN-space dist}} \quad (1)$$

$$\text{subject to} \quad A_j \neq A_i \quad (2)$$

$$\text{(optional)} \quad \text{subject to} \quad \text{Face-Rec}(x_i) \approx \text{Face-Rec}(x_j) \quad (3)$$

$$(4)$$

where $A$ is a vector containing the binary matching attribute value for each example, $Z_{E,i}$ and $Z_{E,j}$ are latent representations in the expanded latent space for a pair of images $x_i$ and $x_j$. The objective (1) ensures that matches are near each other in this space.

The first constraint (2) requires that each match consists of exactly one observation from each group. While the objective (1) effectively captures similarity, subtle identity information can be hard to capture using the GAN latent space alone. To address this, we add a second optional constraint (3) that enforces that a pair of images are similar with regards to a pretrained facial recognition embedding model, Face-Rec$(\cdot)$ (i.e., their distance is below a prespecified threshold). In our experiments, we use a CNN from the dlib library [58] which achieves 99.38% accuracy on the Labeled Faces in the Wild benchmark [59].

We select matches sequentially, starting with the match obtaining the smallest GAN-distance, i.e. "optimal matching" [7]. Once a match is selected, all other images with the same identity of both images in the pair are removed from consideration for further matching. As an alternative to matching using the GAN distance described in this section, we also match using propensity scores computed via logistic regression on the GAN latent space (see Appendix C).

## 4. Results

We evaluate our methods on the CelebA-HQ [2] dataset, which contains 30,000 images of 6,216 unique celebrities. We use the public, pretrained StyleGAN2 model trained on the Flickr-Faces-HQ dataset [60].

### 4.1. Projection evaluation

We first evaluate how well our projection approach preserves identity using human annotation experiments. We find that human annotators are able to distinguish between a GAN reconstruction and real image at essentially the same rate as two real pairs: 66.6%, (64.1-69.1) versus 68.3%, (65.8-70.7), where the parentheses give the Wilson confidence interval. This suggests the reconstructions are very close to the real images. Moreover, we find that (i) the gap between these rates increases when annotators are familiar with the faces in given images, (ii) discrepancies between demographic groups are small, and more (see Appendix A).

### 4.2. GAN-distance matching

Our matching approach produces a subset of Celeba-HQ consisting of 1000 images across perceived gender (500 for each group, and 2 photos for each celebrity – we required 2 per identity for the face recognition benchmarking experiment in Sec 4.3). Fig 3 shows sample matches. The top row shows faces perceived as female, and the bottom row shows their corresponding matches. Matches accurately preserve attributes of the face (e.g., skin color, hair length), pose (e.g., yaw, pitch), and background (e.g., color, texture), both qualitatively and quantitatively (see Table B1).

Fig 4 shows the effect of matching on the distribution of (binary) key covariates. Fig 4A shows the mean value of different covariates for each group in the original dataset. The means are significantly different between groups, indicating the presence of confounding between perceived gender and attributes such as race, age, and smiling. Gaps between these means can skew a downstream analysis of gender bias. After matching, the gaps between these covariates shrink considerably (Fig 4B). The covariate means shift closer together except for *Makeup*, which is likely because there exist very few celebrity in CelebA-HQ who are both perceived as male and wear makeup. Fig 4 shows that GAN-based matching alone can match many important face attributes without requiring access to attribute labels except


Figure 3. **Example matches across perceived gender attribute using nearest neighbors matching.** Many attributes, such as skin color, hair length, pose, and background texture are preserved. Many more matches shown in Fig D1. Note that the perceived gender may not correspond to a celebrity's self-identified gender.

for the matching attribute (gender in this case). Importantly, since this procedure matches each observation one-to-one, it does more than simply match the means of individual covariates; it also matches joint distributions between covariates (see Fig B1).

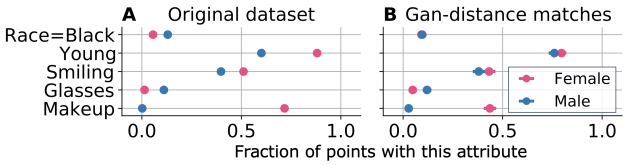


Figure 4. **After matching, distributions of key covariates are more similar across subgroups.** This is shown by the fact the gap between the means for different gender subgroups is smaller after matching (B) than on the original dataset (A). Error bars are 95% Wilson confidence intervals (often within the points).

### 4.3. Benchmarking facial recognition models on matched samples

Finally, we show how matching may be used when benchmarking bias in facial recognition systems. We use the GAN-distance matched samples from Sec 4.2 and benchmark the following popular, open-source systems: dlib [58], Inception ResNet v1 [61] trained on VGGFace2 [62] and CASIA-WebFace [63][3]. Each model returns a face embedding vector for a face, and the distance between any two embeddings is a measure of identity similarity. An accurate recognition model should report a small distance between images of the same identity.

For each perceived gender group, we average the distances reported by each recognition model between images of the same identity. Fig 5 presents the difference between mean recognition distances for females and males. Differences are positive both before and after matching, indicating that all models perform worse for females. Furthermore, the magnitude of the differences are greater using the matched samples, suggesting that confounding factors in the full Celeba-HQ data may mask gender bias. This result shows how the composition of the benchmarking dataset is critical to accurately measure bias.

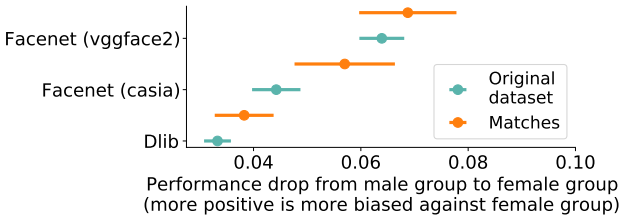


Figure 5. **All classifiers perform worse for female celebrities**, both before matching (Original dataset) and after GAN-distance matching (Matches). We quantify the performance of a classifier by its mean facial embedding distance for image pairs with the same identity, which should be small. Error bars are standard errors of the mean for the female subgroup.

## 5. Discussion and conclusions

We propose a GAN-based matching method which returns matched samples that are both visually accurate (Fig 3) and balanced across attributes (Fig 4, Fig B1). Our approach relies on no supervision except for labels of the matched attribute. We focused on faces in this work because of their prevalence in sensitive applications, and the impressive performance of StyleGAN for this domain. However, our ideas will likely be applicable to any causal domain where GANs can learn meaningful latent spaces, such as in bioimaging.

We applied our method to the measurement of algorithmic bias in facial recognition and found that these models have higher error rates on celebrities perceived as female. Unlike previous methods for measuring algorithmic bias, our method mitigates spurious attribute correlations that may bias the measurement, as is the case with many state-of-the-art observational studies. There are many future challenges to improve causal analysis of images, and the study here helps set a course for more rigorous benchmarking of bias on images "in the wild".

[3]Facenet models retrieved from https://github.com/davidsandberg/facenet.

## Acknowledgements

The authors would like to thank Luis Goncalves for very useful discussions and comments. Additionaly, we would like to thank De'Aira Bryant, Nashlie Sephus, Wei Xia, Yuanjun Xiong and the rest of the faces team and fairness team at Amazon for thoughtful feedback and discussions.

## References

[1] Donald B Rubin. Matching to remove bias in observational studies. *Biometrics*, pages 159–183, 1973. (Cited on page 1.)

[2] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. (Cited on pages 1 and 3.)

[3] 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)*, 2015. (Cited on pages 1 and 2.)

[4] Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010. (Cited on pages 1 and 2.)

[5] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE international conference on computer vision*, pages 4432–4441, 2019. (Cited on pages 2 and 8.)

[6] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *arXiv preprint arXiv:1912.04958*, 2019. (Cited on pages 2 and 8.)

[7] Paul R Rosenbaum. Overt bias in observational studies. In *Observational studies*, pages 71–104. Springer, 2002. (Cited on pages 2 and 3.)

[8] Donald B Rubin. *Matched sampling for causal effects*. Cambridge University Press, 2006.

[9] M Alan Brookhart, Sebastian Schneeweiss, Kenneth J Rothman, Robert J Glynn, Jerry Avorn, and Til Stürmer. Variable selection for propensity score models. *American journal of epidemiology*, 163(12):1149–1156, 2006.

[10] Stephen L Morgan and David J Harding. Matching estimators of causal effects: Prospects and pitfalls in theory and practice. *Sociological methods & research*, 35(1):3–60, 2006.

[11] Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

[12] Daniel E Ho, Kosuke Imai, Gary King, and Elizabeth A Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, 15(3):199–236, 2007. (Cited on page 2.)

[13] Rajeev H Dehejia and Sadek Wahba. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, 84(1):151–161, 2002. (Cited on pages 2 and 14.)

[14] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029, 2016. (Cited on page 2.)

[15] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

[16] Serge Assaad, Shuxi Zeng, Chenyang Tao, Shounak Datta, Nikhil Mehta, Ricardo Henao, Fan Li, and Lawrence Carin. Counterfactual representation learning with balancing weights. *arXiv preprint arXiv:2010.12618*, 2020.

[17] Qingyu Zhao, Ehsan Adeli, and Kilian M Pohl. Training confounder-free deep learning models for medical applications. *Nature communications*, 11(1):1–9, 2020.

[18] Claudia Shi, Victor Veitch, and David Blei. Invariant representation learning for treatment effect estimation. *arXiv preprint arXiv:2011.12379*, 2020.

[19] Liu Qidong, Tian Feng, Ji Weihua, and Zheng Qinghua. A new representation learning method for individual treatment effect estimation: Split covariate representation network. In *Asian Conference on Machine Learning*, pages 811–822. PMLR, 2020.

[20] Giambattista Parascandolo, Alexander Neitz, Antonio Orvieto, Luigi Gresele, and Bernhard Schölkopf. Learning explanations that are hard to vary. *arXiv preprint arXiv:2009.00329*, 2020.

[21] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. *arXiv preprint arXiv:2006.07500*, 2020.

[22] Nathan Kallus. Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. In *International Conference on Machine Learning*, pages 5067–5077. PMLR, 2020. (Cited on page 2.)

[23] Murat Kocaoglu, Christopher Snyder, Alexandros G Dimakis, and Sriram Vishwanath. Causalgan: Learning causal implicit generative models with adversarial training. *arXiv preprint arXiv:1709.02023*, 2017. (Cited on page 2.)

[24] Yash Goyal, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*, 2019.

[25] Matthew James Vowels, Necati Cihan Camgoz, and Richard Bowden. Targeted vae: Structured inference and targeted learning for causal parameter estimation. *arXiv preprint arXiv:2009.13472*, 2020.

[26] Ioana Bica, James Jordon, and Mihaela van der Schaar. Estimating the effects of continuous-valued interventions using generative adversarial networks. *arXiv preprint arXiv:2002.12326*, 2020.

[27] Amelia J Averitt, Natnicha Vanitchanant, Rajesh Ranganath, and Adler J Perotte. The counterfactual *chi*-gan. *arXiv preprint arXiv:2001.03115*, 2020. (Cited on page 2.)

[28] Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Causal feature learning: an overview. *Behaviormetrika*, 44(1):137–164, 2017. (Cited on page 2.)

[29] David Kinney and David Watson. Causal feature learning for utility-maximizing agents. *arXiv preprint arXiv:2005.08792*, 2020. (Cited on page 2.)

[30] Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, pages 2507–2517, 2019. (Cited on page 2.)

[31] Mehrdad Farajtabar, Andrew Lee, Yuanjian Feng, Vishal Gupta, Peter Dolan, Harish Chandran, and Martin Szummer. Balance regularized neural network models for causal effect estimation. *arXiv preprint arXiv:2011.11199*, 2020.

[32] Elias Chaibub Neto. Counterfactual confounding adjustment for feature representations learned by deep models: with an application to image classification tasks. *arXiv preprint arXiv:2004.09466*, 2020. (Cited on page 2.)

[33] John L Barron, David J Fleet, and Steven S Beauchemin. Performance of optical flow techniques. *International journal of computer vision*, 12(1):43–77, 1994. (Cited on page 2.)

[34] Kevin Bowyer and P Jonathon Phillips. *Empirical evaluation techniques in computer vision*. IEEE Computer Society Press, 1998.

[35] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.

[36] Martim Brandao. Age and gender bias in pedestrian detection algorithms. *arXiv preprint arXiv:1906.10490*, 2019.

[37] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018.

[38] Brendan F Klare, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge, and Anil K Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6):1789–1801, 2012.

[39] Boyu Lu, Jun-Cheng Chen, Carlos D Castillo, and Rama Chellappa. An experimental evaluation of covariates effects on unconstrained face verification. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(1):42–55, 2019.

[40] Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *AAAI/ACM Conf. on AI Ethics and Society*, volume 1, 2019.

[41] Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, 2020.

[42] Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. (Cited on page 2.)

[43] Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. Diversity in faces. *arXiv preprint arXiv:1901.10436*, 2019. (Cited on page 2.)

[44] P Jonathon Phillips, Amy N Yates, Ying Hu, Carina A Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, et al. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176, 2018.

[45] Jacqueline G Cavazos, P Jonathon Phillips, Carlos D Castillo, and Alice J O'Toole. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *arXiv preprint arXiv:1912.07398*, 2019.

[46] K. S Krishnapriya, K. Vangara, M.C. King, V. Albiero, and K. Bowyer. Characterizing the variability in face recognition accuracy relative to race, 4 2019.

[47] Hachim El Khiyari and Harry Wechsler. Face verification subject to varying (age, ethnicity, and gender) demographics using deep learning. *Journal of Biometrics and Biostatistics*, 7(323):11, 2016.

[48] Patrick Grother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test (frvt)part 3: Demographic effects. IR 8280, NIST, https://doi.org/10.6028/NIST.IR.8280, 2019. (Cited on page 2.)

[49] 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. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5310–5319, 2019. (Cited on page 2.)

[50] Jean Ponce, Tamara L Berg, Mark Everingham, David A Forsyth, Martial Hebert, Svetlana Lazebnik, Marcin Marszalek, Cordelia Schmid, Bryan C Russell, Antonio Torralba, et al. Dataset issues in object recognition. In *Toward category-level object recognition*, pages 29–48. Springer, 2006.

[51] Antonio Torralba, Alexei A Efros, et al. Unbiased look at dataset bias. In *CVPR*, volume 1, page 7, 2011. (Cited on page 2.)

[52] Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9572–9581, 2019. (Cited on page 2.)

[53] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019. (Cited on pages 2 and 12.)

[54] Vidya Muthukumar, Tejaswini Pedapati, Nalini Ratha, Prasanna Sattigeri, Chai-Wah Wu, Brian Kingsbury, Abhishek Kumar, Samuel Thomas, Aleksandra Mojsilovic, and Kush R Varshney. Understanding unequal gender classification accuracy from face images. *arXiv preprint arXiv:1812.00099*, 2018. (Cited on page 2.)

[55] Guha Balakrishnan, Yuanjun Xiong, Wei Xia, and Pietro Perona. Towards causal benchmarking of bias in face analysis algorithms, 2020. (Cited on pages 2 and 8.)

[56] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. (Cited on page 2.)

[57] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. (Cited on page 2.)

[58] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009. (Cited on pages 2, 3, and 4.)

[59] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments, 2008. (Cited on page 3.)

[60] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. (Cited on page 3.)

[61] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017. (Cited on page 4.)

[62] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018. (Cited on page 4.)

[63] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. (Cited on page 4.)

[64] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing, 2020. (Cited on page 8.)

[65] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European conference on computer vision*, pages 597–613. Springer, 2016.

[66] Guim Perarnau, Joost Van De Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016.

[67] David Bau, Hendrik Strobelt, William Peebles, Bolei Zhou, Jun-Yan Zhu, Antonio Torralba, et al. Semantic photo manipulation with a generative image prior. *arXiv preprint arXiv:2005.07727*, 2020. (Cited on page 8.)

[68] Zachary C Lipton and Subarna Tripathi. Precise recovery of latent vectors from generative adversarial networks. *arXiv preprint arXiv:1702.04782*, 2017. (Cited on page 8.)

[69] Fangchang Ma, Ulas Ayaz, and Sertac Karaman. Invertibility of convolutional generative networks from partial measurements. In *Advances in Neural Information Processing Systems*, pages 9628–9637, 2018. (Cited on page 8.)

[70] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun 2018. (Cited on page 12.)

[71] Abhishek Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2017. (Cited on page 12.)

[72] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. (Cited on page 14.)

[73] Alberto Abadie and Guido W Imbens. Matching on the estimated propensity score. *Econometrica*, 84(2):781–807, 2016. (Cited on page 14.)

# Appendix



## A. Projection results continued

Image projection, sometimes referred to as GAN inversion, involves finding a vector in the GAN's latent space that can generate a desired image. There are broadly two approaches to do so. The first trains an encoder on synthetic training samples to map from the image domain to the latent space [64–67]. A second, optimization-based approach [5, 68, 69] uses gradient descent to find a code in the latent space which best generates the image. We build on the second approach, which has been shown to be more stable and generalizable [5, 6].

Fig A1 shows how reconstruction quality differs in each space for one example face. The middle column shows the original image and its reconstruction in the restricted (top row), and unregularized, expanded (bottom row) spaces. Each row shows the same face manipulated by age using a linear model in the latent space. We trained the linear model using the approach proposed in a previous study [55] (see Fig A2 for another example manipulating skin color). Effective image manipulation with a simple linear model implies that the latent space has good semantic structure, and that distance measurements in the space will be meaningful. Reconstructions using the restricted style space (top row) are poor in terms of perceptual and identity similarity to the original image. Reconstructions using the unregularized, expanded space (bottom row) agree closely with the input image, but semantic manipulation is not possible. A balance between the two extremes is needed.

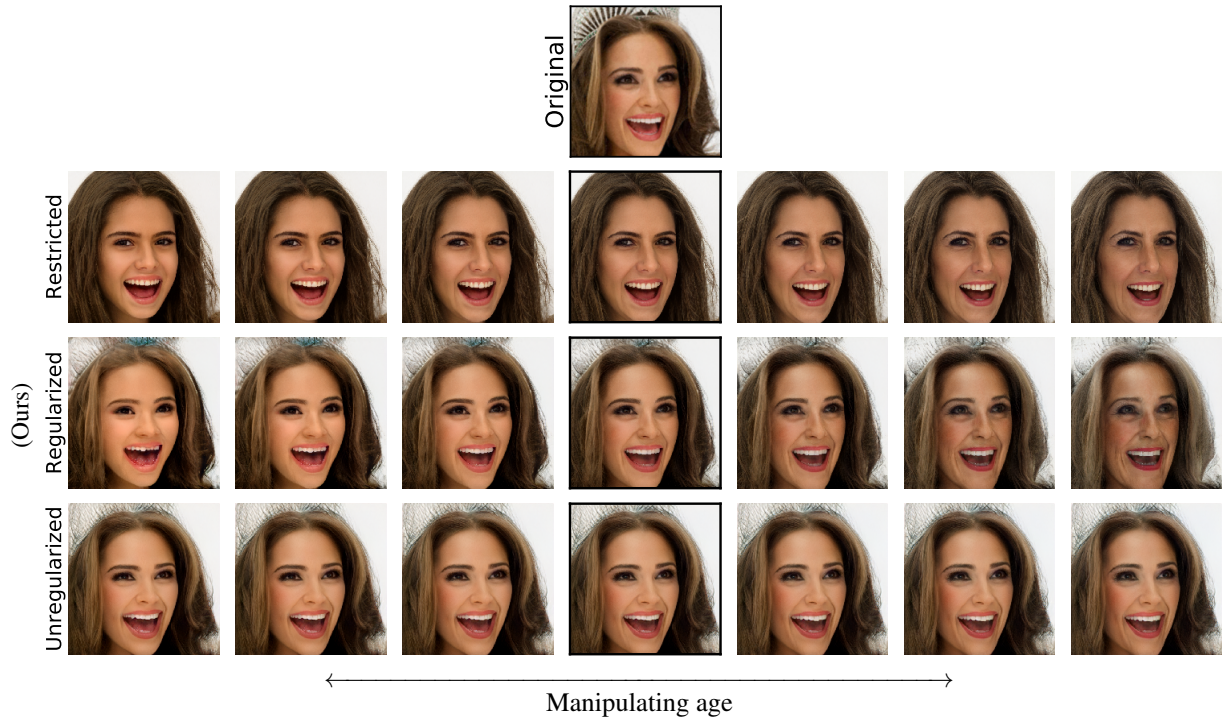


Figure A1. **Regularized projections achieve good visual reconstruction and sensible semantic manipulation**, when traversing the latent direction associated with age (middle row). The restricted projection (top row) achieves poor visual reconstruction but good semantic manipulation. The unregularized projection in the expanded latent space (bottom row) achieves good visual reconstruction but poor semantic manipulation. Proper regularization ($\lambda = 0.1$, middle row) achieves the best of both worlds. Restricted, regularized, and unregularized latent spaces correspond to those in Fig 2.

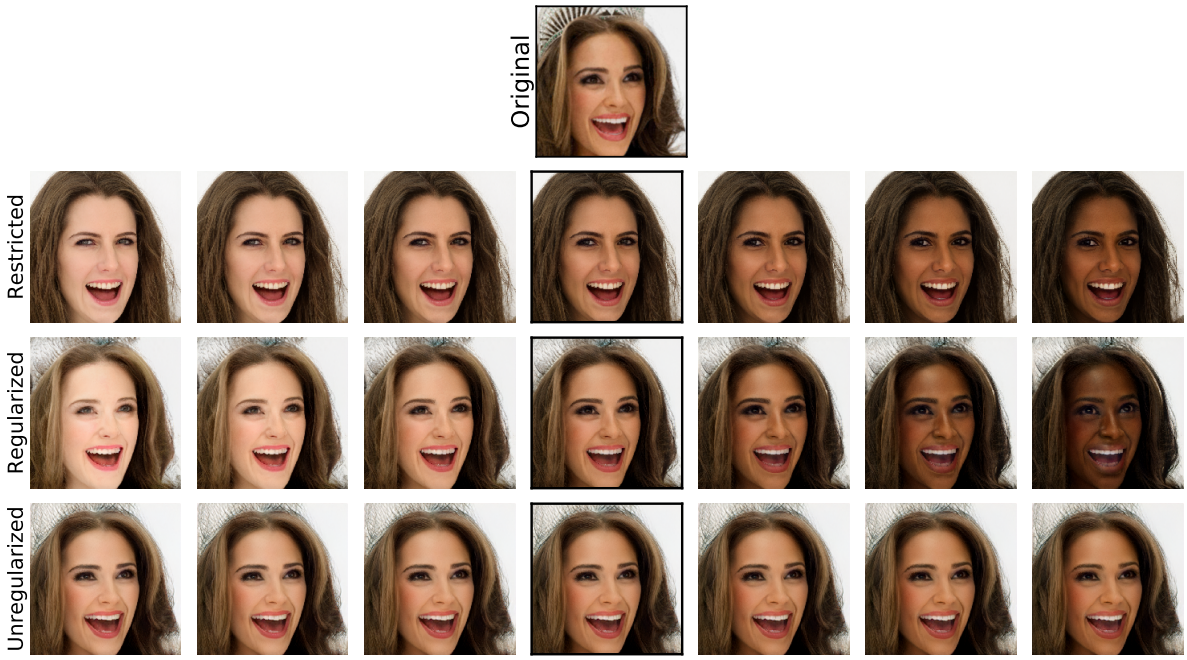


Figure A2. Regularization achieves good visual reconstruction and allows for semantic image manipulation. Middle column shows that the restricted latent space does not qualitatively look like the original image, but the expanded latent space (with and without regularization) both do. In the regularized latent space, traversing the latent direction associated with skin color does change the face skin color, whereas the unregularized version does not.

## A.1. Human annotation experiment details

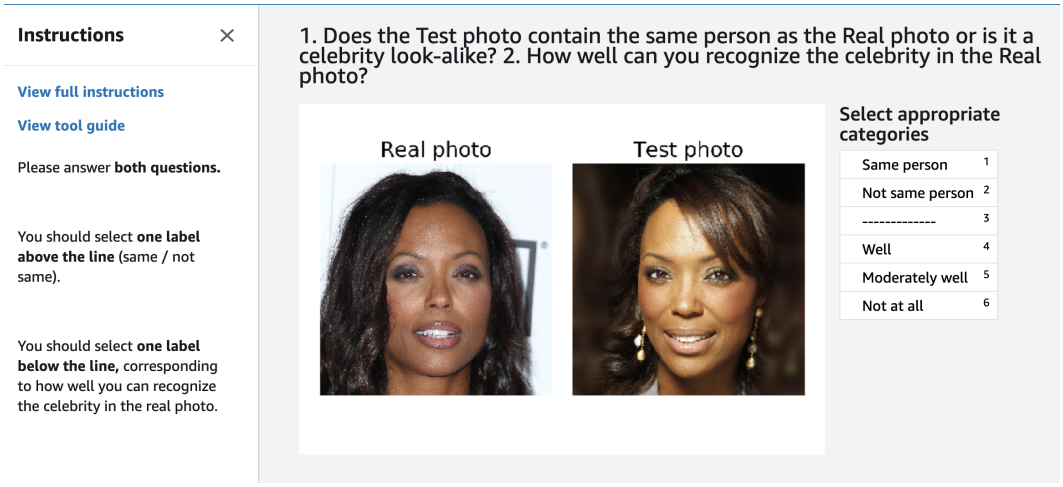


Figure A3. Online interface for benchmarking annotation projections.

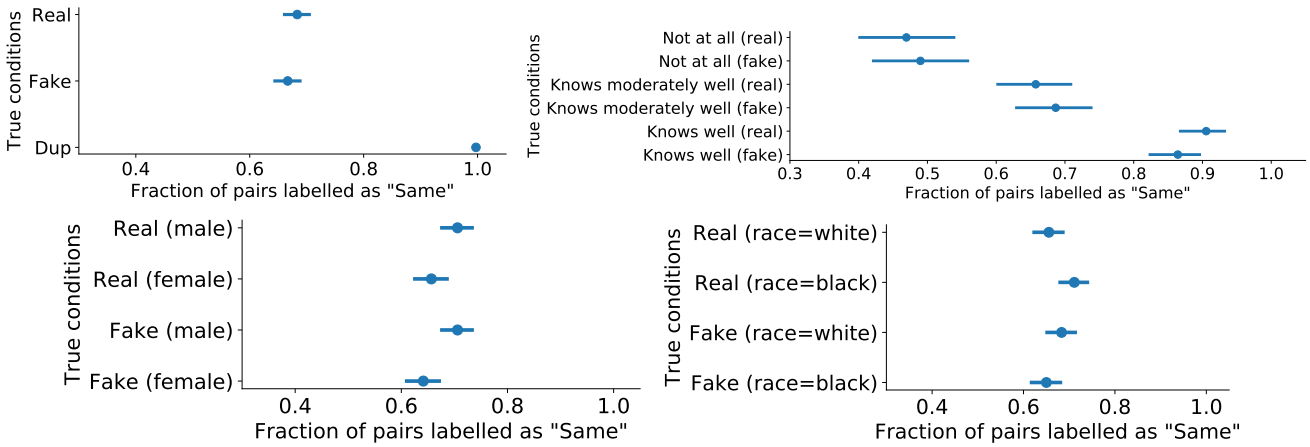


Figure A4. Annotation breakdown. **Top-left.** Duplicate image ('Dup') shows that people accurately detect when the photos are duplicates with small crops. There is almost no difference between real and fake. People who know celebs better do better. **Top-right** People who knew the celebrity well were slightly more accurate at identifying the real celebrity photo as opposed to the GAN reconstructed photo. Interestingly, they are also more likely to label the pair of images as the same (regardless of whether it was GAN-reconstructed or not). **Bottom-left** Error rates are roughly equal across male/female celebrities, although female celebrities are slightly less likely to be labelled as the "Same". **Bottom-right** Error rates for black celebrities are slightly lower than for white celebrities. All error bars are Wilson confidence intervals.

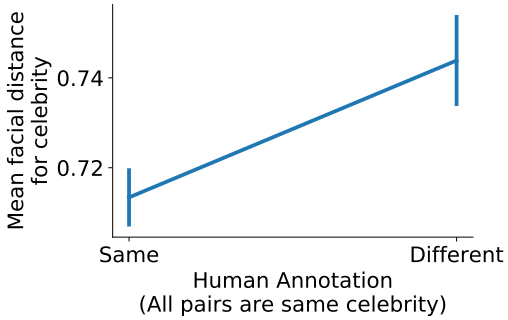


Figure A5. Facial recognition and humans struggle with the same celebrities. Pairs that human annotators annotated as being the 'Same' person have a lower mean facial distance, as measured by a facial recognition classifier. Mean facial distance measured by FaceNet trained on VGGFace2.

Our regularization-based technique for projecting images onto a GAN latent space improves visual reconstruction over past approaches, as measured using perceptual metrics (Fig 2) and through human annotation experiments (Table A1), an important step missing in existing work. Additionally, our projections preserve semantic properties useful for matching (Fig 2, Fig A1) and can be used for downstream tasks such as GAN-based semantic image editing and style transfer, independent of their use for matching.

We first test how well our projection approach preserves identity using human annotations. In each annotation trial, we show an annotator a pair of images (A and B), both of the same celebrity. Image A is always a real image. Image B is one of two possibilities: a different real image of the same celebrity depicted in image A, or StyleGAN2's reconstruction of another real image of the celebrity. We pose the question "Does the test photo contain the same person as the real photo or is it a celebrity look-alike?" (see user interface in Fig A3). We purposefully choose this wording so that annotators have a high bar for judging the GAN reconstructions – a reconstruction must not only be similar to the original, it must be sufficiently similar such that it could not be a different person who looks very much like the original.

We select images from CelebA-HQ by first sorting the celebrities by their number of unique photos in the dataset. We then select the 30 top celebrities in each of the following demographic sets: Black female, Black male, White female, White male. For each celebrity, we have 1 real pair and 1 "fake" pair (i.e., one of the photos is a GAN reconstruction). Each pair is annotated 3 times, resulting in a total of 720 annotations. Annotators were paid $0.024 USD per annotation. Across all

| | Fake pairs | Real pairs |
|---|---|---|
| All | 66.7 (64.14, 69.10) | 68.3 (65.83, 70.73) |
| Well | 86.4 (82.17, 89.78) | 90.5 (86.58, 93.42) |
| Moderately well | 68.7 (62.75, 74.01) | 65.8 (60.02, 71.06) |
| Not at all | 49.0 (41.93, 56.07) | 46.9 (39.94, 54.06) |

Table A1. **Percentage of pairs judged by human annotators to be the same person.** Real pairs contain two real images of a celebrity whereas fake pairs contain one real image and one GAN reconstruction. Parentheticals give 95% Wilson confidence intervals. **Top row**: The overall difference between percentages for fake pairs and real pairs is extremely small, well within 95% Wilson confidence intervals. **Bottom 3 rows**: Annotators who report that they recognize a celebrity 'Well' can better discern real pairs from fake pairs than annotators who say they recognize the celebrity 'Moderately well' or 'Not at all'.

annotations, inter-rater agreement (i.e., the probability that two annotators agree on the label) is 0.737, suggesting that the annotations for individual image pairs are fairly reliable.

Table A1 shows the resulting percentage of pairs judged by human annotators to be the same person. The overall difference between fake pairs and real pairs is very small, well within 95% Wilson confidence intervals (first row). This indicates that the reconstructions preserve identity very well. Furthermore, rows 2-4 show that people who say they recognize the celebrity "Well" can better discern the GAN reconstructions, suggesting that identity may not be preserved well enough to fool a very familiar observer. Interestingly, these familiar observers are much more likely to report that the pair is a real pair, regardless of whether the pair contains a GAN reconstruction.

## B. Quantitative matching results continued

### B.1. GAN-distance measure comparisons

We first present quantitative analyses in Table B1 of the two key components of our distance measure: GAN latent space distance (Eq. 1) and face-recognition embedding distance (Eq. 3). Please see Appendix D for qualitative results. For a given image, we retrieve the 10 closest matches in CelebA-HQ based on a particular distance, and measure how well certain attributes of an image are preserved. We approximate most of the attributes in Table B1 with an algorithm rather than with human annotators. We calculate pose attributes (yaw, pitch, and roll) using a CNN proposed in a recent work [70]. We calculate background statistics by segmenting the background with a CNN trained for semantic segmentation [71] and averaging the pixels corresponding to the background segment. We obtain race from a CNN trained to classify four race categories from a diverse dataset of face images [53].

Matches obtained using GAN latent space distance alone (Table B1, top row) tend to preserve coarse attributes of an image, such as the background and the pose. Matches obtained using GAN-distance between embeddings from a pre-trained facial-recognition classifier tend to preserve identity and attributes related to identity, such as gender and race (middle row). Finally, we combine both distances by finding the closest matches in GAN space subject to the facial-recognition distance being below a threshold of 0.6 (the recommended threshold for the model when classifying whether two faces have the same identity). The combined distance best preserves both global attributes of the image as well as identity-related attributes (bottom row).

| Distance measure | Background Mean | Yaw | Pitch | Roll | ID (top1) | Gender | Race |
|---|---|---|---|---|---|---|---|
| GAN | $33.2 \pm 6.0$ | $6.6 \pm 2.0$ | $5.6 \pm 2.0$ | $1.7 \pm 1.0$ | $80.4 \pm 20.0$ | $16.0 \pm 11.0$ | $36.5 \pm 15.0$ |
| Facial recognition | $40.3 \pm 9.0$ | $13.8 \pm 4.0$ | $6.9 \pm 2.0$ | $2.3 \pm 1.0$ | $7.8 \pm 13.0$ | $1.0 \pm 3.0$ | $8.2 \pm 9.0$ |
| Combined | $38.9 \pm 8.0$ | $10.2 \pm 3.0$ | $6.5 \pm 2.0$ | $2.0 \pm 1.0$ | $36.0 \pm 24.0$ | $0.9 \pm 3.0$ | $14.8 \pm 11.0$ |

Table B1. **Errors of attributes between matched images using different distance measures**. Values are errors between an image and its 10 nearest matches in CelebA-HQ based on each distance measure (see Appendix B.1). The values left of the vertical line are mean absolute errors, and the values to the right are percentages. **Top row**: GAN latent space distance does a good job preserving attributes related to an image's style (i.e., background, face pose). **Middle row**: In contrast, distance based on a face recognition embedding better preserves identity and related attributes like gender and race. **Bottom row**: Combining both distances via Eq. (1) and Eq. (3) strikes a balance. Intervals are standard errors of the mean.

## B.2. Propensity-score matching

We extract a total of 1,210 unique images (605 per group) using propensity-score matching. Fig 4C shows how means of key covariates also shift closer together after matching on propensity scores. While both GAN-distance matching and propensity-score matching substantially reduce covariate gaps between the groups, propensity matching seems to slightly improve balance for attributes such as *Eyeglasses* and *Makeup*.

Again, we are interested in not only matching individual distributions of covariates but also their joint distributions. Fig B1B shows the effect of matching on the joint distributions for key covariates. Across the board, propensity score matching reduces the gap between the groups for different covariate combinations, sometimes quite substantially. For example, propensity matching removes the large imbalance for the *Young & Race* $\neq$ *Black* category between gender groups.

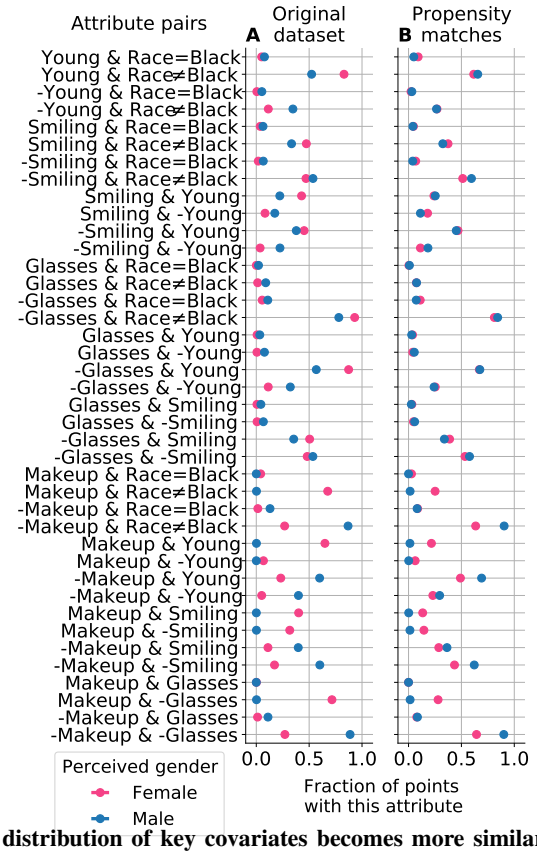


Figure B1. **After matching, the joint distribution of key covariates becomes more similar between gender groups**. Samples are matched using propensity scores. Error bars are 95% Wilson confidence intervals (within the points). "-Young" means that the binary attribute "Young" is false. Fig B2 shows similar results for the nearest-neighbor matches.

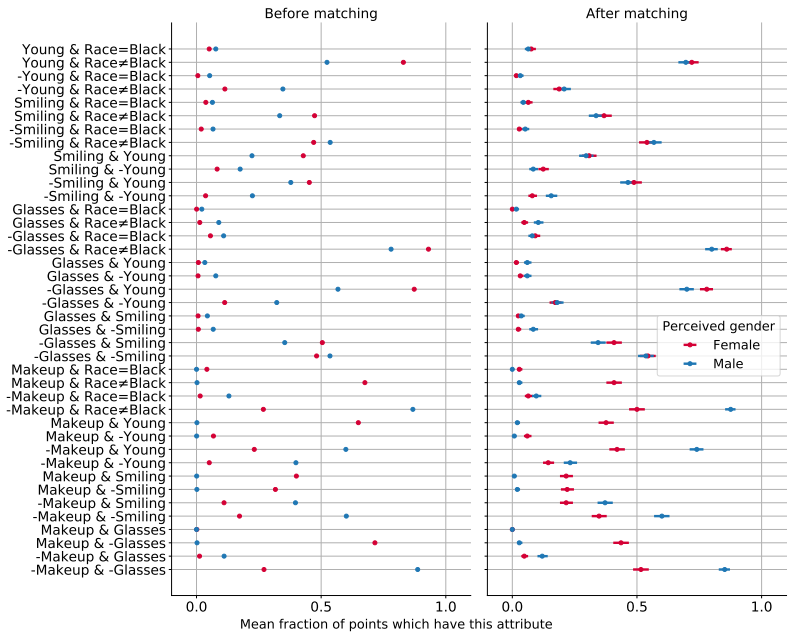


Figure B2. After matching, key covariates (such as the proportion of Black celebrities) are more similar across subgroups. Error bars are 95% Wilson confidence intervals (often within the points).

## C. Propensity score matching details

A well-studied alternative matching approach is propensity-score matching [13, 72, 73]: matching samples based on their predicted probability for a binary protected attribute (e.g., "perceived maleness" in our experiments). Matching examples based on this probability, known as the propensity score, can reduce distribution imbalances caused by variables correlated with the protected attribute. Existing work using propensity scores often assume tabular raw data [73], which is not immediately applicable to image datasets. We overcome this issue by again leveraging the GAN latent space.

For each image, we compute its (regularized) expanded latent matrix $Z_E \in \mathbb{R}^{18\times512}$ and then project this matrix to a vector $\bar{Z}_E \in \mathbb{R}^{512}$ in the restricted style space, by averaging over the expanded dimension. We then train a logistic regression model to predict the probability of the protected attribute value being positive (i.e., the propensity score) from these restricted style space vectors. We verify that these propensity scores are accurate (a prerequisite for matching), finding that they effectively separate perceived gender groups, achieving 98.1% accuracy when fitted to the whole dataset and 97.15% 5-fold cross-validated accuracy (full propensity score distribution given in Fig C1).

Next, we sequentially label matches using the propensity scores. We loop through the smaller of the two protected attribute groups in random order, find the example in the second group with minimal propensity score distance, and accept the match if this distance is within a fixed threshold (0.1 in our experiments). If the match is accepted, we add both images to a list of matches and remove them from further consideration. Otherwise, we discard the original image. In our propensity-score matching experiments, we did not use the facial recognition distance constraint (3) from our nearest neighbors matching approach, although they may be added if desired.

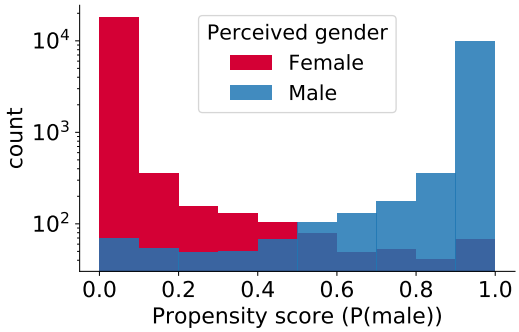


Figure C1. Propensity scores accurately divide the classes.

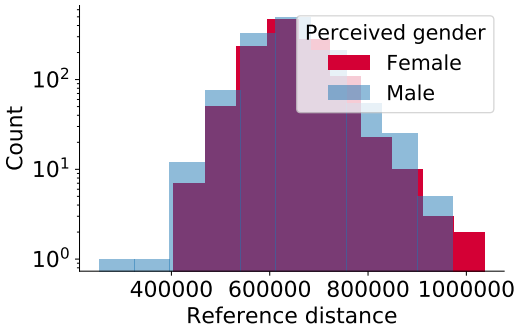


Figure C2. Distribution of reference distances for both groups are similar.

## D. Qualitative matching results


Figure D1. Top matches across perceived gender. Note that the perceived gender may not correspond to the celebrities self-identified gender.

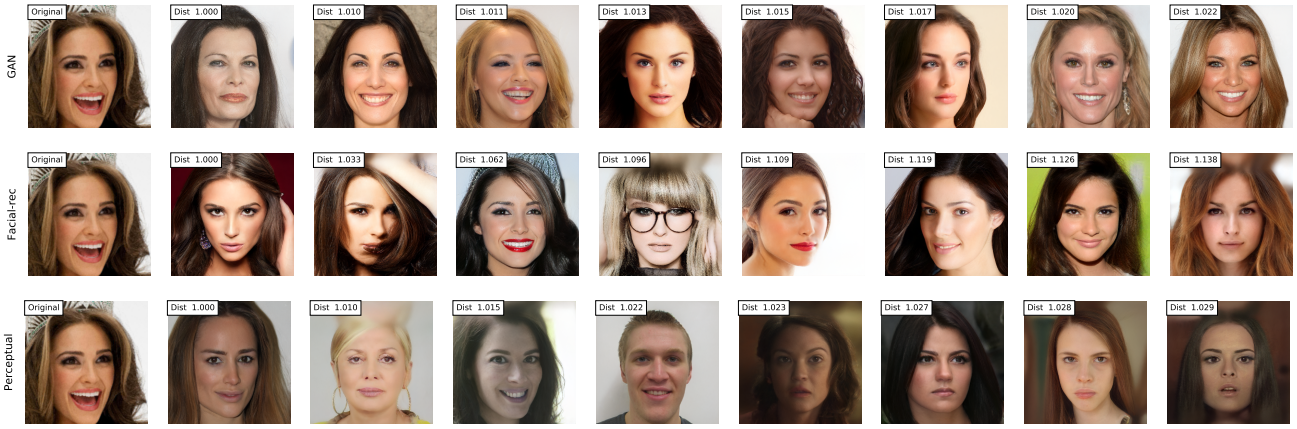


Figure D2. Comparing different distance metrics for matching. Leftmost column shows the original image. Next columns show the top matches (box shows distance from original image, where the distances are normalized by dividing the distance to the closest matches to make distance values comparable across rows.