Towards causal benchmarking of bias in face analysis algorithms

Guha Balakrishnan[1,3], Yuanjun Xiong[3], Wei Xia[3], and Pietro Perona[2,3]

[1] Massachusetts Institute of Technology
[2] California Institute of Technology
[3] Amazon Web Services

Abstract. Measuring algorithmic bias is crucial both to assess algorithmic fairness, and to guide the improvement of algorithms. Current bias measurement methods in computer vision are based on *observational* datasets, and so conflate algorithmic bias with dataset bias. To address this problem we develop an *experimental* method for measuring algorithmic bias of face analysis algorithms, which directly manipulates the attributes of interest, e.g., gender and skin tone, in order to reveal causal links between attribute variation and performance change. Our method is based on generating synthetic image grids that differ along specific attributes while leaving other attributes constant. Crucially, we rely on the perception of human observers to control for synthesis inaccuracies when measuring algorithmic bias. We validate our method by comparing it to a traditional observational bias analysis study in gender classification algorithms. The two methods reach different conclusions. While the observational method reports gender and skin color biases, the experimental method reveals biases due to gender, hair length, age, and facial hair. We also show that our synthetic transects allow for more straightforward bias analysis on minority and intersectional groups.

Keywords: Faces, fairness, bias, causality, counterfactuals, image synthesis, generative adversarial networks (GANs).

1 Introduction

Automated machine learning methods are increasingly used to support decisions in industry, medicine and government. While their performance is often excellent, accuracy is not guaranteed, and needs to be assessed through careful measurements. Measuring *biases*, i.e., performance differences, across protected attributes such as age, sex, gender, and ethnicity, is particularly important for decisions that may affect peoples' lives.

The prevailing technique for measuring the performance of computer vision algorithms is to measure error frequencies on a test set sampled *in the wild* [9, 7, 26, 30] that hopefully mirrors the application domain. Each test image is annotated for attributes of interest, and split into groups that have homogeneous attribute values. Comparing error rates across such groups yields predictions

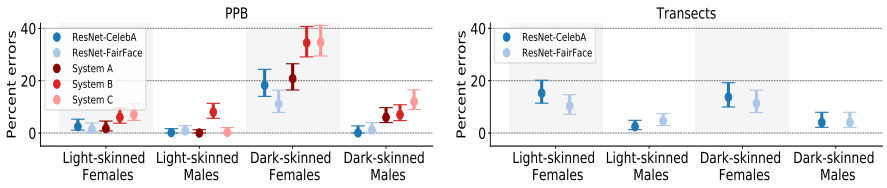

Fig. 1. Algorithmic bias measurements are test set dependent. (Left) Gender classification error rates of three commercial face analysis systems (System A–C) on the Pilot Parliaments Benchmark (PPB) [9] of portrait pictures. Error rates for dark-skinned females were found to be significantly higher than for other groups. We observed the same qualitative behavior when replicating the study with a standard classifier (ResNet-50) on two public face datasets (CelebA, FairFace). (Right) Our experimental investigation using our synthetic Transects dataset, where faces are matched across attributes, reveals a different picture of algorithmic bias (see Fig. 8 for a more complete analysis).

of bias. As an example, Fig. 1-left shows the results of a recent study of algorithmic bias in gender classification of face images from the Pilot Parliaments Benchmark (PPB) dataset. This type of study is called *observational*, because the independent variables (e.g., skin color and gender) are sampled from the environment, rather than controlled by the investigator.

One reason to measure bias is to determine the actions one should take to remove it. For example, based on the results of Fig. 2-left, engineers of systems A-C may decide that incorporating more training examples of dark-skinned women is needed. In this case, measuring bias has one main goal: revealing *causal* connections between attributes of interest and algorithmic performance. Unfortunately, observational studies are ill-suited for drawing such conclusions. When one samples data in the wild, hidden variables that correlate with the variable of interest may have an influence on the performance of the algorithm. As the saying goes, *"correlation does not imply causation."*

For example, in PPB very few males have long hair and almost no light-skinned females have short hair (Fig. 7, and [37]). The fact that hair length (a variable that may affect gender classification accuracy) is correlated in PPB with skin color (a variable of interest) complicates the analysis. Furthermore, the test dataset used to measure bias is often not representative of the population of interest. E.g., the middle-aged Scandinavians and Africans of PPB are not representative of, say, the broad U.S. Caucasian and African-American population [33]. While observational methods do yield useful information on disparate impact within a given test set population, generalizing observational performance predictions to different target populations is hit-or-miss [51] and can negatively impact underrepresented, or minority populations [35, 48]. One would want a method that systematically identifies algorithmic bias independent of the peculiarities of specific test sets.

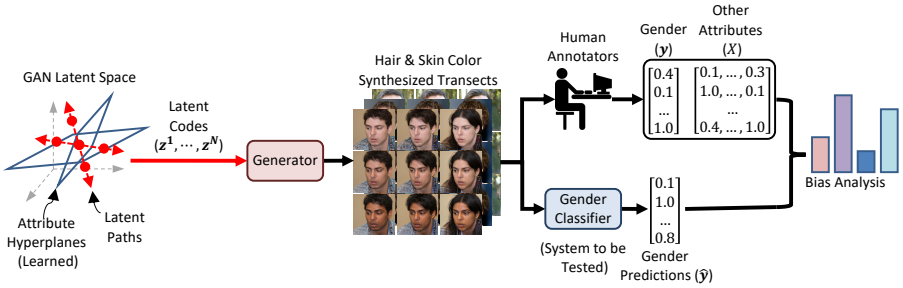

Fig. 2. Method overview. A GAN Generator is used to synthesize "transects," or grids of images along selected attributes by traversing the latent space in specific directions. Human annotations on the transects provide ground truth to be compared with algorithm output to measure errors. Attribute-specific bias measurements are obtained by comparing the algorithm's predictions with human annotations as the attributes are varied. The depicted example may study the question: *Does hair length, skin tone, or any combination of the two have a causal effect on classifier errors?*

A powerful approach to discovering causal relationships is the *experimental method*, used in other disciplines like medicine and social sciences, which involves manipulating the variable of interest while fixing all the other inputs [5, 39]. In this work, we offer a practical way forward to systematically measure bias in computer vision algorithms using the experimental method. Our approach (Fig. 2) generates the test images synthetically, rather than sampling them from the wild, so that they are varied selectively along attributes of interest. This is enabled by recent progress in controlled and realistic image synthesis [22, 23], along with methods for collecting large amounts of accurate human evaluations [8] to quantify the perceptual effect of image manipulations. Our synthesis approach can alter multiple attributes at a time to produce grid-like matched samples of images we call *transects*. We quantify the image manipulations with detailed human evaluations which we then compare with algorithm output to estimate algorithmic bias.

We evaluate our methodology with experiments on two gender classification algorithms. We first find that our transect generation strategy creates significantly more balanced data across key attributes compared to "in the wild" datasets. Next, inspired by [9], we use this synthetic data to explore the effects of various attributes on gender classifier errors. Our findings reveal that the experimental method can change the picture of algorithmic bias (Fig. 8), which will affect the strategy of algorithm improvement, particularly concerning groups that are often underrepresented in training and test sets. This work is a first step to developing experimental methods for algorithmic bias testing in computer vision, and so much remains to be done both in design and experimentation to achieve broadly-applicable and reliable techniques. In Sec. 5 we discuss limitations of the current method, and next steps in this research area.

2 Related Work

Benchmarking in computer vision has a long history [4, 6, 13] including face recognition [30, 40–42, 15, 16] and face analysis [9]. Some of these studies examine biases in performance, i.e., error rates across variation of important parameters (e.g. racial background in faces). Since these studies are purely observational, they raise the question of whether the biases they measure depend on algorithmic bias, or on correlations in the test data. Our work addresses this question.

A dataset is said to be biased when combinations of features are disproportionately represented. Computer vision datasets are often found to be biased [44, 51]. Human face datasets are particularly scrutinized [2, 12, 26, 28, 29, 36] because methods and models trained on these data can end up being biased along attributes that are protected by the law [27]. Approaches to mitigating dataset bias include collecting more thorough examples [36], using image synthesis to compensate for distribution gaps [29], and example resampling [31].

Studies of face analysis systems [9, 26, 34] and face recognition systems [17, 30] attempt to measure bias across gender and skin-color (or ethnicity). However, the evaluations are based on observational rather than interventional techniques – and therefore any conclusions from these studies should be treated with caution. A notable exception is a recent study [37] using the experimental method to investigate the effect of skin color in gender classification. In that study, skin color is modified artificially in photographs of real faces to measure the effects of differences in skin color, all else being equal. However, the authors observe that generalizing the experimental method to other attributes, such as hair length, is too onerous if one is to modify existing photographs. Our goal is to develop a generally applicable experimental method, where *any* attribute may be studied independently.

Recent work uses generative models to explore face classification system biases. One study explores how variations in pose and lighting affect classifier performance [2, 28, 29]. A second study uses a generative model to synthesize faces along particular attribute directions [11]. These studies rely on the strong assumption that their generative models can modify one attribute at a time. However, this assumption relies on having unbiased training data, which is almost always not practical. In contrast, our framework uses human annotations to account for residual correlations produced by our generative model.

3 Method

Our framework consists of two components: a technique to synthesize *transects*, or grid-like constructs of synthesized images with control over semantic attributes (Sec. 3.1), and a procedure using these synthesized images, along with human annotators, to analyze biases of a classifier (Sec. 3.2).

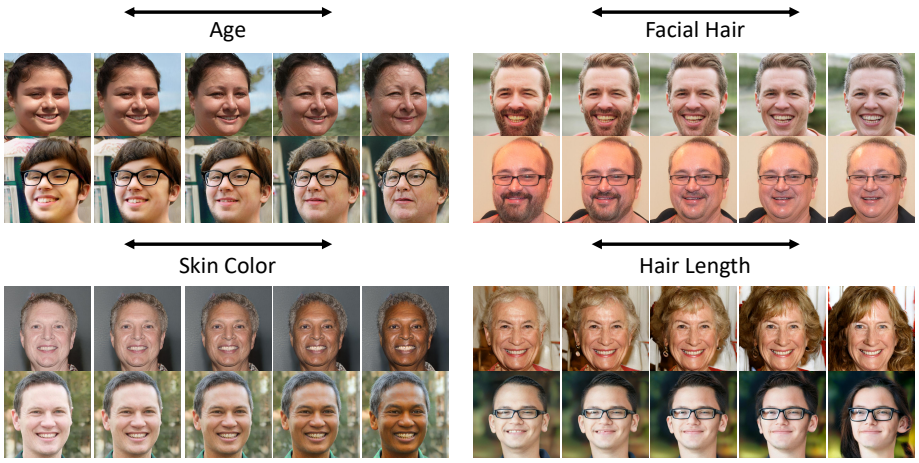

Fig. 3. 1D transects. 1×5 sample transects synthesized by our method for various attributes. Orthogonalization was used (see Fig. **??**).

3.1 Transects: A Walk in Face Space

We assume a black-box face generator G that can transform a latent vector $\mathbf{z} \in \mathcal{R}^D$ into an image $I = G(\mathbf{z})$, where $p(\mathbf{z})$ is a distribution we can sample from. In our study, G is the generator of a pre-trained, publicly available GAN, "StyleGAN2" [22, 23]. We base our approach on a recent study [56] for single attribute traversals in GAN latent spaces. That method trains a linear model to predict a particular image attribute from $\mathbf{z}$, and uses the model to traverse the $\mathbf{z}$-space in a discriminative direction. We generalize this idea to synthesize image grids, i.e., *transects*, spanning arbitrarily many attributes, unlike related work that operate on only one or two attributes at a time [11, 47, 49, 55, 56].

Estimating Latents-to-Attributes Linear Models. Let there be a list of N_a image attributes of interest (age, gender, skin color, etc.). We generate an annotated training dataset $\mathcal{D}_\mathbf{z} = \{\mathbf{z}^i, \mathbf{a}^i\}_{i=1}^{N_z}$, where $\mathbf{a}^i$ is a vector of scores, one for each attribute, for generated image $G(\mathbf{z}^i)$. The score for attribute j, $\mathbf{a}^i_j$, may be continuous or binary. We sample a generous number of values of $\mathbf{z}^i$ from $p(\mathbf{z})$ and obtain $\mathbf{a}^i$ from human annotators.

For each attribute j, we use $\mathcal{D}_\mathbf{z}$ to compute a $(D-1)$-dimensional linear hyperplane $h_j = (\mathbf{n}_j, b_j)$, where $\mathbf{n}_j$ is the normal vector and b_j is the offset. For continuous attributes like age or skin color, we train a ridge regression model [20]. For binary attributes we train a support vector machine (SVM) classifier [10].

Multi-attribute Transect Generation. Each hyperplane h_j specifies the subspace of $\mathcal{R}^D$ with boundary values of attribute j, and the normal vector $\mathbf{n}_j$ specifies

Fig. 4. 2D transects. 5×5 transects varying simultaneously hair length and skin tone. Multidimensional transects allow for intersectional analysis, i.e. analysis across the joint distribution of multiple attributes. Orthogonalization was used (see Fig. **??**).

a direction along which that attribute primarily varies. To construct a one-dimensional, length-L transect for attribute j, we first start with a random point $\mathbf{z}^i$ and project it onto h_j. We then query $L-1$ evenly-spaced points along $\mathbf{n}_j$, within fixed distance limits on both sides of the h_j. Fig. 3 presents some single transect examples (with orthogonalization, a concept introduced in the next section). Further details on querying points are found in Sec. 3.1.

The 1D transect does not allow us to explore the joint space of several attributes, or to fix other attributes in precise ways. We generalize to K-dimensional transects to address this (see supplementary material for algorithm). The main extensions are: (1) we project $\mathbf{z}^i$ onto the intersection of K attribute hyperplanes, and (2) we move in a K-dimensional grid in $\mathbf{z}$-space (see Fig. 4).

Orthogonalization of Traversal Directions. The hyperplane normals $\{\mathbf{n}_j\}_{j=1}^{N_a}$ are not orthogonal to one another. If we set the traversal direction vectors equal to these normal vectors, i.e., $\mathbf{v}_j = \mathbf{n}_j$, we will likely observe unwanted correlations between attributes. We reduce this effect by producing a set of modified direction vectors such that $\mathbf{v}_j \perp \mathbf{n}_k, \forall k \neq j$. See supplementary material for our orthogonalization algorithm and image examples.

Setting Step Sizes and Transect Dimensions. We set L to small values to reduce annotation cost. For example, $L = 5$ for the 1D transects in Fig. 3 and 2D transects in Fig. 4, and $L = 2$ for the 3D transects in Fig. 5. For each attribute j, we manually set min/max signed decision values with respect to h_j, and linearly interpolate L_j points between these extremes. We set per-attribute min/max values so that transects depict a full dynamic range for most random samples.

3.2 Analyses Using Transects

An ideal transect will modify only selected attributes at a time, but in practice, unintended attributes will be accidentally modified. In addition, the degree to

which an attribute is altered varies across transects. To measure and control for these factors we annotate each image of each transect, resulting in a second dataset $\mathcal{D}_{transect} = \{I^i, \mathbf{a}^i\}_{i=1}^{N_{images}}$ of images and human annotations.

We assume a target attribute of interest, e.g., gender, and a target attribute classifier C. We denote the ground truth gender of image I^i (as reported by humans) by y^i, and C's prediction by $\hat{y}^i$. For ease of analysis, we discretize the remaining attributes into bins, and assign an independent binary variable to each bin [14]. We denote the vector of concatenated binary covariates for image i by $\mathbf{x}^i$, and the classification error by $e^i = \ell(\hat{y}^i, y^i)$, where $\ell(\cdot, \cdot)$ is an error function.

Our first analysis strategy is to simply compare error rates across different subpopulations [9, 7, 26, 30]. Let E_j^s be the average error of C over test samples for which covariate j is equal to $s \in \{0, 1\}$. If the data is generated from a perfectly randomized or controlled study, the quantity $E_j^1 - E_j^0$ is a good estimate of the "average treatment effect"(ATE) [3, 19, 38, 46] of covariate j on e, a causal measure of the expected change in e when covariate j is flipped from 0 to 1, with other covariates fixed. Note that exactly computing the ATE from an observational dataset is virtually never possible, because we do not observe the counterfactual case(s) for each data point, e.g., the same person with both light and dark skin tones. Though our transects come closer to achieving an ideal study than real datasets do (see Sec. 4.3), there may still be confounding between covariates (see Fig. 10 for an example).

Since any observable confounder may be annotated in $\mathcal{D}_{transect}$, our second strategy is to use a covariate-adjusted ATE estimator [43, 45, 53]. One simple adjustment approach is to train a linear regression model predicting e^i from $\mathbf{x}^i$: $e^i = \epsilon^i + \beta_0 + \sum_j \beta_j \mathbf{x}_j^i$, where β's are parameters, and ϵ^i is a per-example noise term. β_j captures the ATE, the average change in e given one unit change in $\mathbf{x}_j$ holding all other variables constant, provided: (1) a linear model is a reasonable fit for the relationship between the dependent and independent variables, (2) all relevant attributes are included in the model (i.e., no hidden confounders), and (3) no attributes that are influenced by $\mathbf{x}_j$ are included in the model, otherwise these other factors can "explain away" the impact of $\mathbf{x}_j$. An experimenter can never be completely sure that (s)he has satisfied these conditions but (s)he can strive to do so through careful consideration.

Finally, when the outcome lies in a fixed range, as is the case in our experiments with $e^i \in [0, 1]$, we use logistic instead of linear regression. β_j then represents the expected change in the log odds of e for a unit change in $\mathbf{x}_j$.

3.3 Human Annotation

We collect human annotations on the synthetic faces to construct $\mathcal{D}_{\mathbf{z}}$ and $\mathcal{D}_{transect}$, using Amazon Mechanical Turk [8] through the AWS SageMaker Ground Truth service [1]. Annotators evaluated each image for seven attributes: gender, facial hair, skin color, age, makeup, smiling, hair length and realism. Each attribute was evaluated on a discrete scale. For complete details about our annotation

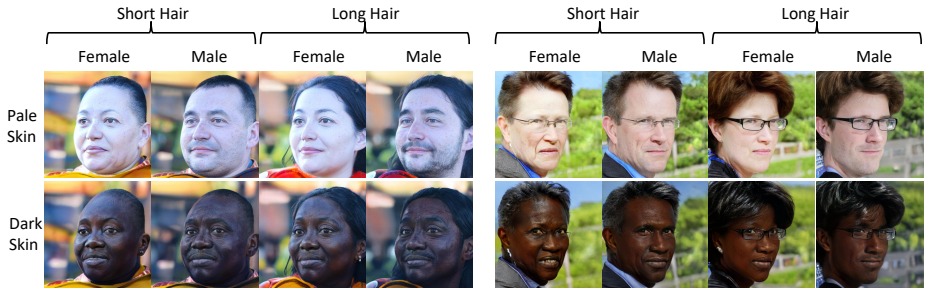

Fig. 5. Examples of transects used in our experiments. We created 1,000 transects spanning pale-to-dark skin tones, short-to-medium hair lengths, and male-to-female genders – two transects are shown here. Other face attributes are approximately held constant. For each image we collected human annotations to measure the perceived attributes. Intended attributes do not always agree with human perception (see Fig. 6, below).

Fig. 6. Human perception of transect attributes. Humans label the first face as a male, though we intended to produce a female. In all our experiments we used human perception, rather than intended attributes, as the ground truth.

process, samples of our survey layouts and analysis, please see supplementary material.

4 Experiments

We evaluate our method on benchmarking bias of gender classifiers. The Pilot Parliaments Benchmark (PPB) [9] was the first wild-collected test dataset to balance gender and skin color with the goal of fostering the study of gender classfication bias across both attributes. The authors of that study found a much larger error rate on dark-skinned females, as compared to other groups and conjectured that this is due to bias in the algorithms, i.e., that the performance of the algorithm changes when gender and skin color are changed, all else being equal. Our method allows us to test this hypothesis.

4.1 Gender Classifiers

We trained two research-grade gender classifier models, each using the ResNet-50 architecture [18]. The first was trained on the CelebA dataset [32], and the

second on the FairFace dataset [21]. CelebA is the most popular public face dataset due to its size and rich attribute annotations, but is known to have severe imbalances [11]. The FairFace dataset was introduced to mitigate some of these biases. See supplementary material for training details. We decided not to test commercial systems for reproducibility reasons — models we test may be re-implemented and retrained by other researchers, while commercial systems are black boxes which may change unpredictably over time.

4.2 Transect Data

We used the generator from StyleGAN2 trained on Flickr-Faces-HQ (FFHQ) [22, 23]. We use the generator's "style space" as the latent space in our method, because we found it better suited for disentangling semantic attributes than the input noise space. We trained linear regression models to predict age, gender, skin color and hair length attributes from style vectors. For the remaining attributes — facial hair, makeup and smiling — we found that binarizing the ranges and training a linear SVM classifier works best.

We generated 3D transects across subgroups of skin color, hair length, and gender as described in Sec. 3.1. We use a transect size of $2 \times 2 \times 2 \times 1$, with grid decision values (input $\mathbf{c}$ of transect generation algorithm in supplementary material) spaced to generate pale-to-dark skin colors, short-to-medium hair lengths, male-to-female genders, and adult ages. We set the decision values by trial-and-error, and made them equal for all transects: $(-1.5, 1.7)$ for skin color, $(-0.5, 0)$ for hair length, $(-1.75, 1.75)$ for gender, and 0.5 for age. We generated 1000 such transects, resulting in 8000 total images. Fig. 5 presents two example transects.

Not all synthesized images are ideal for our analysis. Some elicit ambiguous human responses or are unrealistic, and others may not belong clearly to the intended category of each attribute. See supplementary material for details on how we prune non-ideal images.

4.3 Comparison of Transects to Real Face Datasets

Fig. 7-top analyzes attribute distributions for the CelebA-HQ, FFHQ and PPB datasets, along with our transects, stratified by gender. The wild-collected datasets contain significant imbalances across gender, particularly with hair length and age. In contrast, our transects exhibit more balance across gender. They depict more males with medium-to-long hair, and fewer females with very long hair. Our transects also have a bimodal skin color distribution, and an older population by design, since we are interested in mimicking those population characteristics of PPB. All datasets are imbalanced along the "Beard" and "Makeup" attributes — this is reasonable since we expect these to have strong correlations with gender.

In an ideal matched study, sets of images stratified by a sensitive attribute will exhibit the same distribution over remaining attributes. Fig. 7-bottom stratifies by skin color. We see correlations of hair length distributions and skin colors in all the wild-collected data, while the synthetic transects exhibit much better balance.

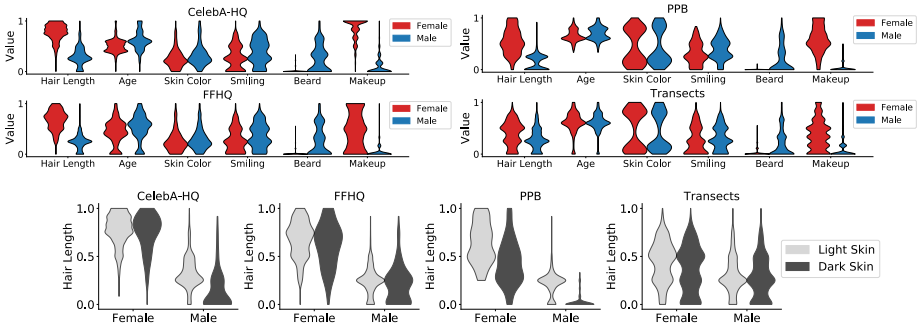

Fig. 7. (Top) Attribute distributions ("violin plots") by gender groups for different datasets. Wild-collected datasets have greater attribute imbalances across gender than synthetic transects, e.g., longer hair and younger ages for women. We designed our transects to mirror PPB's skin color and age distributions, while mitigating hair length imbalance. (Bottom) Hair length distributions by gender and skin color groups. In the wild-collected datasets hair length is correlated with skin color, when gender is held constant. Transects may be designed to minimize this correlation.

4.4 Analysis of Bias

We now analyze the performance of the classifiers on PPB and our transects. Our classifiers exhibit similar error patterns to the commercial classifiers already evaluated on PPB [9]. Because PPB only consists of adults, we remove children and teenagers (age < 0.4 in the normalized $[0, 1]$ scale) from our transects to make a more direct comparison, leaving us with 5335 total images.

Fig. 1 presents classification errors split by gender (M/F) and skin color (L/D). We replicated the reported errors of the commercial classifiers in [9], and report the errors of our classifiers on our in-house version of PPB. All classifiers perform significantly worse on dark-skinned females. Fig. 8 presents classification errors, stratified by gender/hair length/skin color combinations.

We can make a number of broad-stroke, qualitative observations. First, the pattern of errors is similar across PPB and transects, with more errors on the left (females) than on the right (males). Second, transect errors are either comparable or higher than in PPB, indicating that synthetic faces can be at least as challenging as real faces. Most significantly, errors are nonzero on males, which allows the study of relative difficulties when attributes are varied. Third, in PPB, there are few males with long hair and few females with short hair and light skin, making measurements unreliable for these categories. This is not a problem with transects. Fourth, transect errors are higher when hair is shorter for women. However, hair length has a negligible effect for males (see Fig. 10 for a possible explanation). Fifth, there is no consistent transect error pattern in skin tone: within homogeneous groups changing skin tone does not seem to affect the performance of either algorithm. Looking at PPB alone, we could not make this observation, since skin tone is so strongly correlated with hair length.

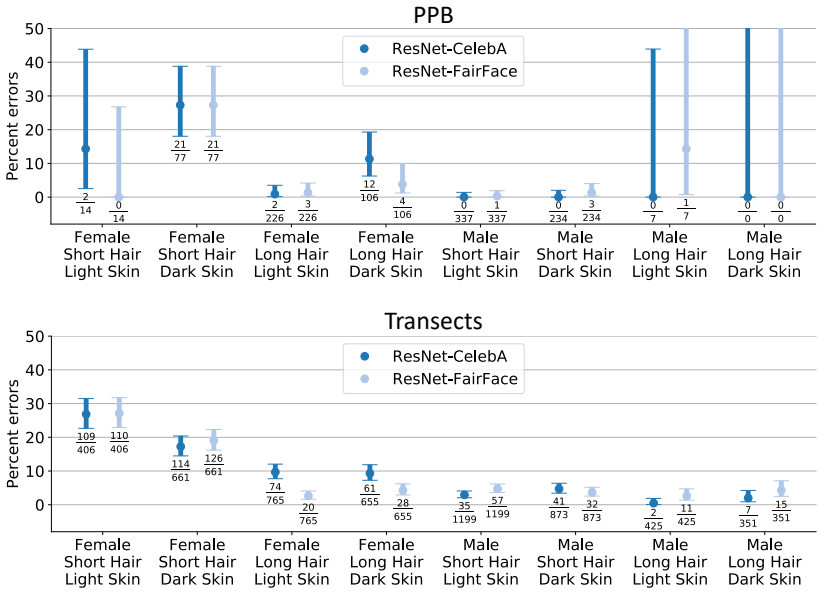

Fig. 8. Algorithmic errors, disaggregated by intersectional groups for wild-collected (PPB, top) and synthetic (transects, bottom). Wilson score 95% confidence intervals [54] are indicated by vertical bars, and the misclassification count and total number of samples are written below each bar. PPB has few samples for several groups, such as short-haired, light-skinned females and long-haired males (see Fig. 7). Synthetic transects provide numerous test samples for all groups. The role of the different attributes in causing the errors is studied in Sec. 4.4 and Fig. 9.

We investigated further by calculating covariate-adjusted causal effects using an $L2$-regularized logistic regression model to predict that classifier's error conditioned on all attributes. See supplementary material for details on how we trained the regression models. Fig. 9 presents coefficients for both models, which represent the change in log odds of the classifier's error for a change of one unit of each covariate (see Sec. 3.2). Facial hair, gender, makeup, hair length and age all have significant effects on classification error, while skin color has a negligible effect. We made a simplifying assumption that each covariate has an independent, linear effect on classification error, which we know is not true. Please see supplementary material for further discussion on this topic.

5 Discussion

Our synthesis-based experimental method offers a number of attractive properties over traditional observational methods. First, it generates approximately matched samples along selected attributes, allowing for counterfactual synthesis. Observational image data are almost never matched. Second, image synthesis al-

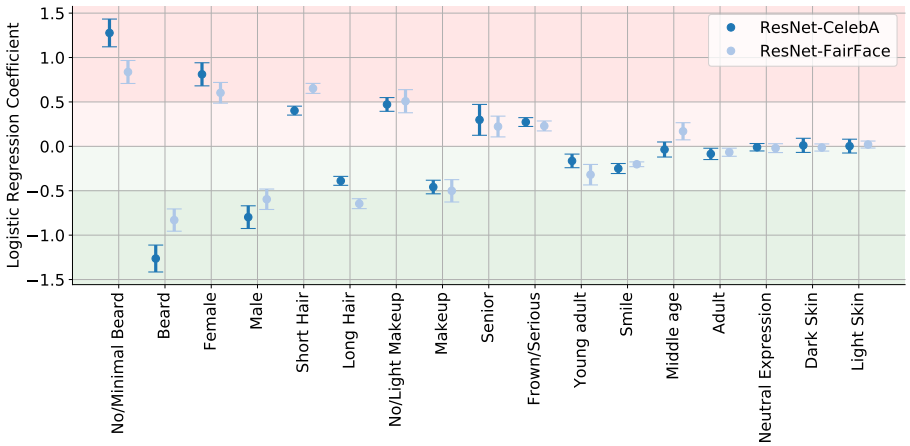

Fig. 9. Logistic regression coefficient values. The regression model is trained to predict absolute errors of the gender classifiers on our transect images given binary attributes as input. Coefficients represent the change in log odds of the error for a change of one unit of each attribute. Larger magnitudes indicate more important attributes, and positive(red)/negative(green) values correspond to attributes that increase/decrease classifier error. We order attributes here by large-to-small coefficient magnitudes.

lows, to a greater degree, uniform sampling of the space of attributes of interest. This is very difficult to do when one relies on images that are sampled from in-the-wild distributions, where some groups are underrepresented. Third, bias may be measured for intersectional groups defined by specific attribute combinations. Single-attribute analysis may conceal biases affecting groups defined by the combination of multiple attributes [25]. Some such combinations are often vastly undersampled in natural data. Fourth, image synthesis is fast and inexpensive, and crowdsourced image annotation is also relatively fast and affordable. By contrast, assembling large datasets of natural images is laborious and expensive. Thus, synthetic data has the potential to democratize testing for bias. And finally, ethical and legal concerns are greatly reduced. Collecting real face images in the wild requires great care to respect the privacy and dignity of individuals, the rights of minors and other vulnerable groups, as well as copyright laws. By contrast, synthetic datasets are freer from such risks because they do not depict real people.

The experimental analysis (transects) and traditional observational analysis (using PPB) diverged most significantly on the effect of skin color, which the observational study flagged as significant and the experimental method found to be not significant in determining algorithmic bias. The experimental method reveals a number of additional sources of bias: age, hair length and facial hair (Fig. 9). The two methods agree on gender.

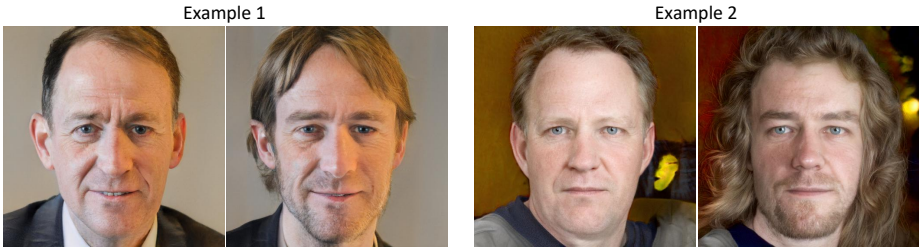

Fig. 10. Correlated attribute modifications. We found that our method sometimes adds a beard to a male face when attempting to only modify hair length. This is an example of an imprecise intervention which can complicate downstream bias analyses. This bias may be due to the training data itself (men with long hair tend to have facial hair), or injected by the algorithm.

Fig. 11. Hidden confounders. There may always be a hidden confounder lurking in a dataset. As an example, we found that our method tends to add earrings when transitioning from dark-skinned men to dark-skinned women, a cue that a gender classifier might use to perform disproportionately well on the latter group. Interestingly, one male in this image also has an earring; that earring becomes larger for his female counterpart.

Bias measurements guide scientists and engineers towards effective corrective measures for improving the performance of their algorithms. It is instructive to view the different predictions of the two methods through this lens. The correlational study based on PPB (Fig. 1) may suggest that, in order to reduce biases in our classifiers, more images of dark-skinned women should be added to their training sets. The experimental method leads engineers in a different direction. First, more training images of long-haired men and short-haired women of all races are needed. Second, correcting age bias requires more training images in the child-teen and, possibly, senior age groups.

Finally, it is important to consider a rich number of attributes and attribute combinations, besides the one(s) of immediate interest. This is for two reasons. First, unobserved confounders can have strong effects and need to be included in the analysis. Second, the combined effect of attributes can be strongly nonlinear (see the interaction of age and gender in supplementary material), and therefore an intersectional analysis [24, 9] is necessary. Selecting attributes combinations

is as much of an art as a science, and therefore one has to rely on good judgment and on a healthy multidisciplinary debate to progressively reveal missing ones.

6 Limitations and Future Work

Our method can not perfectly eliminate unwanted correlations with annotated variables, nor can it account for hidden confounders [52], and one will need to keep a sharp eye out for both. As an example of the first, we found that our method often adds facial hair to male faces when increasing hair length (see Fig. 10). This is likely a reason for why our classifiers did not have higher error rates for males with longer hair (see Fig. 8). As an example of the second, we found that our method tends to synthesize earrings when modifying a dark-skinned face to look female (see Fig. 11). Depending on culture, earrings may or may not be relevant to the definition of gender. If this is an unwanted correlation, one ought to add earrings to the annotation pipeline so that it may be "orthogonalized away" by the synthesis method. A significant advantage of an approach that is based on synthetic images and human annotation is thus the following: *as soon as one residual correlation is discovered it may be systematically annotated, compensated for in the analysis, and mitigated in the synthesis.*

A number of refinements in face synthesis will make our experimental method more practical and powerful. First, many of the faces we generated contained visible artifacts (see supplementary material), which we eliminated by human annotation – even subtle artifacts can affect classifier outputs, as revealed by the literature on adversarial examples [50]. Second, we do not yet have tools to estimate the sets of physiognomies and attribute combinations that can and cannot be produced by a given generator. Current GANs are known to have difficulties in generating data outside of their training distributions. Third, we observed a bias of StyleGAN2 towards generating Caucasian faces when sampling from its latent distribution. While our method can compensate for biases through carefully oriented traversals calibrated by human annotations, it would be clearly better to start from unbiased synthesis methods.

Our first-order technique for controlling synthesis can also be improved. A better understanding of the geometry of face space will hopefully yield more accurate global coordinate systems. These, in turn, will help reduce residual biases in synthetic transects, which we currently mitigate by having transects annotated by hand. Finally, extending our method beyond gender classification to more complex tasks, such as face recognition, is not straightforward in practice and will require further study.

Acknowledgments. We are grateful to Frederick Eberhardt, Bill Freeman, Lei Jin, Michael Kearns, R. Manmatha, Tristan McKinney, Sendhil Mullainathan, and Chandan Singh for insights and suggestions.

References

1. https://aws.amazon.com/sagemaker/groundtruth/
2. Albiero, V., KS, K., Vangara, K., Zhang, K., King, M.C., Bowyer, K.W.: Analysis of gender inequality in face recognition accuracy. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision Workshops. pp. 81–89 (2020)
3. Angrist, J.D., Imbens, G.W.: Identification and estimation of local average treatment effects. Tech. rep., National Bureau of Economic Research (1995)
4. Barron, J.L., Fleet, D.J., Beauchemin, S.S.: Performance of optical flow techniques. International journal of computer vision **12**(1), 43–77 (1994)
5. Bertrand, M., Mullainathan, S.: Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. American economic review **94**(4), 991–1013 (2004)
6. Bowyer, K., Phillips, P.J.: Empirical evaluation techniques in computer vision. IEEE Computer Society Press (1998)
7. Brandao, M.: Age and gender bias in pedestrian detection algorithms. arXiv preprint arXiv:1906.10490 (2019)
8. Buhrmester, M., Kwang, T., Gosling, S.D.: Amazon's mechanical turk: A new source of inexpensive, yet high-quality data? (2016)
9. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. pp. 77–91 (2018)
10. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3), 273–297 (1995)
11. Denton, E., Hutchinson, B., Mitchell, M., Gebru, T.: Detecting bias with generative counterfactual face attribute augmentation. arXiv preprint arXiv:1906.06439 (2019)
12. Drozdowski, P., Rathgeb, C., Dantcheva, A., Damer, N., Busch, C.: Demographic bias in biometrics: A survey on an emerging challenge. IEEE Transactions on Technology and Society (2020)
13. Fei-Fei, L., Fergus, R., Perona, P.: 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. pp. 178–178. IEEE (2004)
14. Gelman, A., Hill, J.: Data analysis using regression and multilevel/hierarchical models. Cambridge university press (2006)
15. Grother, P., Ngan, M., Hanaoka, K.: Ongoing face recognition vendor test (frvt) part 1: Verification. National Institute of Standards and Technology, Tech. Rep (2018)
16. Grother, P.J., Ngan, M.L., Hanaoka, K.K.: Ongoing face recognition vendor test (frvt) part 2: identification. Tech. rep. (2018)
17. Hanaoka, P.G.N.K.: Face recognition vendor test (frvt)part 3: Demographic effects. IR 8280, NIST, https://doi.org/10.6028/NIST.IR.8280 (2019)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. Heckman, J.J., Vytlacil, E.J.: Instrumental variables, selection models, and tight bounds on the average treatment effect. In: Econometric Evaluation of Labour Market Policies, pp. 1–15. Springer (2001)

20. Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics **12**(1), 55–67 (1970)
21. Kärkkäinen, K., Joo, J.: Fairface: Face attribute dataset for balanced race, gender, and age. arXiv preprint arXiv:1908.04913 (2019)
22. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
23. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. arXiv preprint arXiv:1912.04958 (2019)
24. Kearns, M., Neel, S., Roth, A., Wu, Z.S.: Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. arXiv preprint arXiv:1711.05144 (2017)
25. Kearns, M., Roth, A.: The Ethical Algorithm: The Science of Socially Aware Algorithm Design. Oxford University Press (2019)
26. Klare, B.F., Burge, M.J., Klontz, J.C., Bruegge, R.W.V., Jain, A.K.: Face recognition performance: Role of demographic information. IEEE Transactions on Information Forensics and Security **7**(6), 1789–1801 (2012)
27. Kleinberg, J., Ludwig, J., Mullainathany, S., Sunstein, C.R.: Discrimination in the age of algorithms. Published by Oxford University Press on behalf of The John M. Olin Center for Law, Economics and Business at Harvard Law School (2019), https://academic.oup.com/jla/article-abstract/doi/10.1093/jla/laz001/5476086
28. Kortylewski, A., Egger, B., Schneider, A., Gerig, T., Morel-Forster, A., Vetter, T.: Empirically analyzing the effect of dataset biases on deep face recognition systems. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 2093–2102 (2018)
29. Kortylewski, A., Egger, B., Schneider, A., Gerig, T., Morel-Forster, A., Vetter, T.: 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. pp. 0–0 (2019)
30. Krishnapriya, K.S., Vangara, K., King, M., Albiero, V., Bowyer, K.: Characterizing the variability in face recognition accuracy relative to race. ArXiv 1904.07325 (4 2019)
31. Li, Y., Vasconcelos, N.: Repair: Removing representation bias by dataset resampling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9572–9581 (2019)
32. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (2015)
33. Lohr, S.: Facial recognition is accurate, if you're a white guy. New York Times (February 9 2018), https://nyti.ms/2BNurVq
34. Lu, B., Chen, J.C., Castillo, C.D., Chellappa, R.: An experimental evaluation of covariates effects on unconstrained face verification. IEEE Transactions on Biometrics, Behavior, and Identity Science **1**(1), 42–55 (2019)
35. Merkatz, R.B., Temple, R., Sobel, S., Feiden, K., Kessler, D.A., on Women in Clinical Trials, W.G.: Women in clinical trials of new drugs–a change in food and drug administration policy. New England Journal of Medicine **329**(4), 292–296 (1993)
36. Merler, M., Ratha, N., Feris, R.S., Smith, J.R.: Diversity in faces. arXiv preprint arXiv:1901.10436 (2019)
37. Muthukumar, V., Pedapati, T., Ratha, N., Sattigeri, P., Wu, C.W., Kingsbury, B., Kumar, A., Thomas, S., Mojsilovic, A., Varshney, K.R.: Understanding unequal gender classification accuracy from face images. arXiv preprint arXiv:1812.00099 (2018)

38. Oreopoulos, P.: Estimating average and local average treatment effects of education when compulsory schooling laws really matter. American Economic Review **96**(1), 152–175 (2006)
39. Pearl, J.: Causality. Cambridge university press (2009)
40. Phillips, P.J., Grother, P., Micheals, R., Blackburn, D.M., Tabassi, E., Bone, M.: Face recognition vendor test 2002. In: 2003 IEEE International SOI Conference. Proceedings (Cat. No. 03CH37443). p. 44. IEEE (2003)
41. Phillips, P.J., Wechsler, H., Huang, J., Rauss, P.J.: The feret database and evaluation procedure for face-recognition algorithms. Image and vision computing **16**(5), 295–306 (1998)
42. Phillips, P.J., Yates, A.N., Hu, Y., Hahn, C.A., Noyes, E., Jackson, K., Cavazos, J.G., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., 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)
43. Pocock, S.J., Assmann, S.E., Enos, L.E., Kasten, L.E.: Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practiceand problems. Statistics in medicine **21**(19), 2917–2930 (2002)
44. Ponce, J., Berg, T.L., Everingham, M., Forsyth, D.A., Hebert, M., Lazebnik, S., Marszalek, M., Schmid, C., Russell, B.C., Torralba, A., et al.: Dataset issues in object recognition. In: Toward category-level object recognition, pp. 29–48. Springer (2006)
45. Robinson, L.D., Jewell, N.P.: Some surprising results about covariate adjustment in logistic regression models. International Statistical Review/Revue Internationale de Statistique pp. 227–240 (1991)
46. Rubin, D.B.: Matched sampling for causal effects. Cambridge University Press (2006)
47. Shen, Y., Gu, J., Tang, X., Zhou, B.: Interpreting the latent space of gans for semantic face editing. arXiv preprint arXiv:1907.10786 (2019)
48. Simon, V.: Wanted: women in clinical trials (2005)
49. Singla, S., Pollack, B., Chen, J., Batmanghelich, K.: Explanation by progressive exaggeration. arXiv preprint arXiv:1911.00483 (2019)
50. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
51. Torralba, A., Efros, A.A., et al.: Unbiased look at dataset bias. In: CVPR. vol. 1, p. 7 (2011)
52. VanderWeele, T.J., Shpitser, I.: On the definition of a confounder. Annals of statistics **41**(1), 196 (2013)
53. Willan, A.R., Briggs, A.H., Hoch, J.S.: Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. Health economics **13**(5), 461–475 (2004)
54. Wilson, E.B.: Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association **22**(158), 209–212 (1927)
55. Xiao, T., Hong, J., Ma, J.: Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In: Proceedings of the European conference on computer vision (ECCV). pp. 168–184 (2018)
56. Zhou, B., Bau, D., Oliva, A., Torralba, A.: Interpreting deep visual representations via network dissection. IEEE transactions on pattern analysis and machine intelligence (2018)