Fair SA: Sensitivity Analysis for Fairness in Face Recognition

Aparna R. Joshi   APARNA_JOSHI@APPLE.COM
Xavier Suau   XSUAUCUADROS@APPLE.COM
Nivedha Sivakumar   NIVEDHA_S@APPLE.COM
Luca Zappella   LZAPPELLA@APPLE.COM
Nicholas Apostoloff   NAPOSTOLOFF@APPLE.COM
Apple

Abstract

As the use of deep learning in high impact domains becomes ubiquitous, it is increasingly important to assess the resilience of models. One such high impact domain is that of face recognition, with real world applications involving images affected by various degradations, such as motion blur or high exposure. Moreover, images captured across different attributes, such as gender and race, can also challenge the robustness of a face recognition algorithm. While traditional summary statistics suggest that the aggregate performance of face recognition models has continued to improve, these metrics do not directly measure the robustness or fairness of the models. Visual Psychophysics Sensitivity Analysis (VPSA) (19) provides a way to pinpoint the individual causes of failure by way of introducing incremental perturbations in the data. However, perturbations may affect subgroups differently. In this paper, we propose a new fairness evaluation based on robustness in the form of a generic framework that extends VPSA. With this framework, we can analyze the ability of a model to perform fairly for different subgroups of a population affected by perturbations, and pinpoint the exact failure modes for a subgroup by measuring targeted robustness. With the increasing focus on the fairness of models, we use face recognition as an example application of our framework and propose to compactly visualize the fairness analysis of a model via AUC matrices. We analyze the performance of common face recognition models and empirically show that certain subgroups are at a disadvantage when images are perturbed, thereby uncovering trends that were not visible using the model's performance on subgroups without perturbations.

Keywords: Sensitivity, Fairness, Robustness

1. Introduction

Face recognition has been routinely used in private and government organizations to automate complex decision making processes (22). The adoption of this technology into criminal justice systems (6; 25), consumer applications, such as smartphones, and healthcare, heightens the potential negative implications that unfair algorithms can have on society. As cases of racial, gender and other biases surface, the inequity in face recognition algorithms have prompted calls for stricter regulation (3; 2; 18). This inequity motivates both, the development of algorithms that treat sub-populations equally, and the need for companies to continually assess their technology for bias in order to reduce consumer harm.

The development of face recognition models relies on large scale datasets, and progress is often measured using summary statistics, such as accuracy, true positive rate, etc over the entire dataset (for brevity, we will refer to these statistics as *traditional* summary statistics).

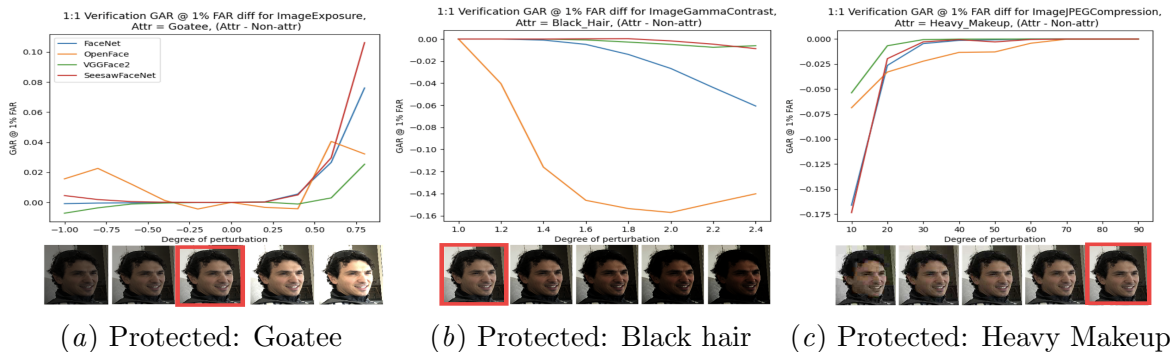(*a*) Protected: Goatee   (*b*) Protected: Black hair   (*c*) Protected: Heavy Makeup

Figure 1: Example Fair SA curves for face verification. In each plot, the original unperturbed image is highlighted. Models have a positive bias with increasing perturbation in Fig. 1(*a*) favoring the *goatee* subgroup under high exposure. Similarly, models disfavor *black hair* (Fig. 1(*b*)) and *heavy makeup* (Fig. 1(*c*)) subgroups (with negative bias values) under gamma contrast and JPEG compression respectively.

However, traditional summary statistics do not provide enough information to explain individual failure modes, and thus provide little insight into how to diagnose and address issues in the models (19). Moreover, face recognition models can have classification accuracies as high as 99% on benchmark datasets, but these models are often susceptible to failing under image degradations (24), and the outcomes can be dissimilar for different sub-populations (3). Performance of face recognition systems can be impacted by changes in the image processing pipeline (white balancing, enhancement, etc.), by artifacts related to the camera (noise, distortion, etc.) and other degradation artifacts due to motion such as motion blur or lighting conditions like exposure. Thus, along with the traditional evaluation of face recognition models, studying these algorithms with input images subject to perturbations can provide useful insights in evaluating robustness to visual distortions that can occur in real-world environments.

A number of works have shown divergent error rates of face recognition models across demographics (3). This discrepancy in performance across sub-populations elevates the need to assess face recognition models for fairness. Many metrics have been proposed to assess the fairness of an algorithm (7). One line of work studies *individual fairness*, which relies on the notion that similar inputs should yield similar predictions (8). On the other hand, *group fairness* focuses on grouping examples based on a protected attribute, and measures whether the protected and the unprotected groups are treated similarly. Proposals for group fairness include metrics such as demographic parity, which requires the predictions of the model to be independent of the protected attribute. However, evaluation based on conventional fairness metrics does not account for robustness to image degradations, which frequently appear in real-world face recognition systems.

Fields studying vision have employed a technique called visual psychophysics to evaluate algorithms. Psychophysics is a set of experimental procedures from psychology and

neuroscience that involve examination of relationships between perturbed test images and the responses they elicit in models (27). The methodology described in (19) (VPSA) derives ideas from visual psychophysics to measure the robustness of a face recognition model to incremental perturbations. Model performance is evaluated at various perturbation levels, providing explainability for failure cases beyond summary statistics. However, VPSA does not measure the robustness of the model with respect to fairness, since image degradations can affect different demographics differently. For instance, blurred images may cause face recognition models to favor certain subgroups (e.g. young) over others (e.g. old). Such a discrepancy can be observed when face recognition is performed on images captured by an out-of-focus lens, or when there is relative motion between the object and the camera. In this paper, we make the following contributions:

(1) We propose Fair Sensitivity Analysis (Fair SA) (Fig. 1) as a new way of measuring group fairness through robustness. Fair SA is an evaluation framework that allows machine learning practitioners to jointly evaluate the vulnerability of models to (i) image perturbations, and (ii) variations in demographics. This approach allows us to analyze how a model treats different subgroups under perturbations of varying levels.

(2) The analysis of multiple models, across many attributes and perturbations generates a large amount of data. We propose the use of area under the curve (AUC) matrices in order to summarize the performance of each model. AUC matrices provide a at-a-glance overview and can be efficiently analyzed row-, column-, and matrix-wise to extract explicit information about the bias induced by individual perturbations and the impact on individual subgroups.

(3) We survey the performance of common face recognition models from the literature, such as VGGFace2 (4), OpenFace (1), FaceNet, (20) and SeesawFaceNet (29) on the widely used public dataset CelebA (14) to demonstrate the efficacy of Fair SA. We evaluate the robustness of these algorithms to protected subgroups under varying levels of 9 types of perturbations on the tasks of verification and self-matching. Using this analysis, we uncover consistent biases against certain subgroups under specific data perturbations (e.g. the effect of *exposure* on the *pale skin* subgroup).

2. Related Work

Perturbed Inputs and Psychophysics — Several studies have explored simulating real world conditions that impact the performance of face recognition algorithms. These simulations are often incremental perturbations that are applied to transform images. Karahan et al. (10) present an evaluation to assess the impact of incrementally perturbing face images on deep CNN face recognition models and found that while Gaussian blur, noise, and occlusion cause a significant decrease in performance, models were more robust to color distortions and changes in color balance. While these findings were for the closed set identification task, Grm et al. (9) propose a similar evaluation framework for face verification, and conclude that noise, blur, and brightness greatly impact the performance of models, whereas the effects of compression and contrast are limited. As in VPSA, perturbations constitute an important component in our methodology.

Psychophysics studies the relationships between controlled stimuli and resulting model responses. Several fields in vision have looked at psychophysics as an alternate way of evaluating algorithms (12). Psyphy (27) is a comprehensive evaluation framework for object recognition models based on psychophysics and item-response theory. VPSA (19) builds on Psyphy to support a similar item-response analysis for face recognition. VPSA comprises of a set of procedures developed to make face recognition algorithms more explainable. Fair SA, in turn, enhances VPSA to analyze the implications of variations in responses evoked for different data subgroups. These subgroups can be determined based on the sensitive attributes thereby yielding a fairness perspective to algorithmic robustness evaluation.

Robustness — Robustness is particularly important for face recognition due to the inevitable influence of camera and real world conditions on the inputs. Studies have focused on evaluating models for robustness against digital attacks, such as variations in pose, illumination (13), contrast and blur (10), adversarial attacks (28), and against physical attacks (24). Robustness against bias is a relatively recent area of research that has garnered much interest due to its widespread impact in the community. The idea of *robustness bias*, which indicates that certain subgroups may be less robust to adversarial attacks, has been discussed in (15). However, (15) proposes measuring the robustness bias by estimating the distance of subgroup members to the (closest) decision boundary relative to other members using bound approximations, and does not use perturbations to measure robustness. On the contrary, our method offers information on whether a model favors (disfavors) a subgroup under a perturbation, and can offer insights into potential mitigation strategies (e.g. informed training data augmentations). Thus, even though this experimental study shares an underlying motivation with this work, it is qualitatively and quantitatively different from the psychophysics-inspired approach we describe.

Fairness — Demographic differences in model performance have been studied over the past decades. In 2002, the seminal Face Recognition Vendor Test (FVRT) (17) already showed results signaling higher performance for males compared to females, and for older persons compared to younger persons. Apart from FRVT, there are several smaller scale evaluations of bias in trained algorithms on public datasets (23; 5). More recently, Buolamwini and Gebru (3) analyzed commercial face recognition software on per-class accuracy for gender and skin tone classes, and demonstrated poorer performance of three commercial systems for dark skinned females than for lighter skinned males. Khiyari et al. (11) measure the verification performance of VGGFace (16) on one, two, and three-class demographic groups based on race, gender and age with the best performance for middle-age white males and worst performance for young black females. The work in (26) provides insights into how genuine and imposter score distributions differ between African-American and Caucasian image cohorts. The 2020 ChaLearn "Looking at People" challenge (22) summarizes the top submissions for evaluating accuracy and bias on gender and skin tones on the task of face verification. However, these studies do not study how performance varies across demographics for degraded input data.

3. Method

VPSA (19) applies visual psychophysics to assess face recognition models by introducing incremental perturbations, such as blur and rotation to input images. An example of each

---

Algorithm 1: $C(D, t, n, b_l, b_u, a, v, L, M, T)$: Fair SA curve generation function for perturbation type T

Input: D: input image data, t: similarity threshold, n: number of stimulus levels, b_l and b_u: the lower and upper bounds of stimulus levels, a and v: protected attribute and value, L: identity labels, M: metadata (attribute annotations), T: perturbation type

Output: k, the Fair SA curve

1 $\Delta \leftarrow n$ stimulus levels from b_l to b_u

2 $k \leftarrow \bigcup\limits_{\delta \in \Delta} \phi(D, t, \delta, a, v, L, M, T)$

Algorithm 2: $\phi(D, t, \delta, a, v, L, M, T)$: Fair SA point generation function for stimuli $T(i, \delta)$

Input: D: input image data, t: similarity threshold, δ: stimulus level, a and v: protected attribute and value, L: identity labels, M: metadata (attribute annotations), T: perturbation type

Output: $\{\delta, \alpha\}$, an x,y coordinate pair (stimulus level, bias)

1 $D' \leftarrow i \in D : T(i, \delta)$

2 $pred \leftarrow \gamma(D', D, t)$

3 $\alpha \leftarrow \beta(a, v, D, M, L, pred)$

---

of the perturbations we consider is shown in Appendix A. Each of these perturbations is varied in magnitude thus controlling the difficulty of recognizing faces. VPSA assesses model robustness by comparing a perturbed image to itself (whilst ensuring it is sufficiently distinct from other images). Note that VPSA assumes one image per identity. It generates item-response curves (IRC) to quantify robustness as match rates (ratio of number of correct self-matches to the total number of images) on the y-axis against perturbation levels on the x-axis. Instead of using summary statistics to gauge performance, VPSA leverages visual psychophysics to compute these IRCs that map stimulus responses to performance, thereby allowing one to pinpoint the exact condition or stimulus type and level at which the algorithm can no longer reliably perform the task (27).

Although VPSA does an excellent job at evaluating model robustness with respect to external adversaries and incremental degradations, it does not assess if a model is also robustly fair. Incremental degradations may affect a specific subgroup much more than the others (e.g. *exposure* may treat certain skin tones unfairly). Therefore, we propose to enhance VPSA by enabling measurement of the effect that incremental perturbations can have on different demographics by way of computing the bias between protected and unprotected subgroups indicating if the algorithm favors (positive value) or disfavors (negative value) a specific population. This analysis can uncover discrepancies in model performance across subgroups under simulations of real world conditions, and gives us a way to carefully analyze these dissimilar behaviors to correct algorithmic deficiencies.

VPSA begins by pruning the identities that lead to false matches or false non-matches for a face recognition algorithm. By retaining only the identities that match well to themselves

---

Algorithm 3: $\gamma(D_p, D_g, t)$: function that generates task-dependent predictions for face recognition model f

---

Input: f: a function to extract model responses (representations from selected layers) from a face recognition model, D_p: probe (perturbed) data, D_g: gallery (original) data, t: similarity threshold

Output: $pred$, the task-dependent predictions.

1 $R' \leftarrow i \in D_p : f(i)$
2 $R \leftarrow i \in D_g : f(i)$
3 $S \leftarrow r' \in R', r \in R : dist(r, r')$
4 **if** *TASK is CLASSIFICATION* **then**
5 $\quad pred \leftarrow S$
6 **else if** *TASK is FACE RECOGNITION* **then**
7 $\quad M \leftarrow S \geq t$
8 $\quad pred \leftarrow diag(M)$

Algorithm 4: $\beta(a, v, D, M, L, pred)$: Bias generation function for protected attribute a and value v

---

Input: b: function to compute bias w.r.t. protected attribute, a and v: protected attribute and value, D: input image data, M: metadata (attribute annotations), L: identity labels, $pred$: set of task-dependent predictions for the data

Output: α, the bias.

1 $P_{a,v} \leftarrow i \in D$ s.t. $M_i[a] = v$
2 $P_{a,\bar{v}} \leftarrow D \setminus P_{a,v}$
3 $\alpha \leftarrow b(pred_i | i \in P_{a,v}, pred_i | i \in P_{a,\bar{v}}, L)$ (Note: for face verification we have $pred_{i,j} | i, j \in P_a$)

---

and poorly to others, the effect of applied perturbations on the remaining input images is disentangled from model failures at no perturbation. On the other hand, Fair SA treats this initial pruning step as optional due to the possibility that the pruning may result in the removal of more identities belonging to some subgroups than others, which can adversely affect our fairness assessment. However, for self-matching, experiments with the original VPSA pruning step have also been performed and presented in Appendix C. In our experiments, we observe similar trends for VPSA with and without the pruning step. For Fair SA with pruning, we largely observe the same signedness albeit different magnitudes for the attributes as without pruning.

Let D be original input data, a and v be the protected attributes and values (e.g. a can be the attribute *gender*, and v can be the value *female*). Let $P_{a,v} \in D$ be the protected subgroup of the population (as defined by a and v), and $P_{a,\bar{v}} = D \setminus P_{a,v}$ be the unprotected subgroup. The Fair SA algorithm starts with function C (Alg. 1) that generates the final Fair SA curves in the form of item-response values. C is called once for each perturbation type T and it repeatedly calls the function ϕ (Alg. 2) for stimulus levels ranging from b_l to b_u for each input perturbation type. Function ϕ generates a point on the Fair SA curve, where

the point represents the bias α of the face recognition algorithm f on protected subgroup $P_{a,v}$. A bias measure can be task-dependent e.g. for face verification, we use Genuine Acceptance Rate (GAR) at 1% False Acceptance Rate (FAR) (21), and for self-matching, we use statistical imparity.

The perturbation function T in Alg. 2 produces perturbed images given the original images and stimulus level δ. In this paper, we refer to the perturbed images as the probes ($D_p = D'$), and the original images as the gallery ($D_g = D$). Similar to VPSA, the γ function (Alg. 3) is a wrapper to a given face recognition model f that generates task-dependent model predictions. γ computes cosine similarities between model responses (of a layer) for the probe set D_p, and the gallery set D_g. For verification, we use the subgroup similarities and ground truth subgroup identities to compute Genuine Acceptance Rate (GAR) at 1% False Acceptance Rate (FAR). For self-matching, the predictions are the similarities of perturbed images with corresponding original images (e.g. diagonal of the thresholded similarity matrix). β (Alg. 4) then computes the statistical imparity, the difference between the probability that a random sample drawn from $P_{a,v}$ is a match with its perturbed version and the probability that a random sample from the complement $P_{a,\bar{v}}$ is a match.

AUC Matrices. Drawing conclusions about multiple models across different perturbations using curves generated by Alg. 1 can be challenging. Therefore, we propose the use of AUC matrices for each model to visualize Fair SA (or VPSA) results. Furthermore, we use matrix norms to understand the average model fairness, and row-wise or column-wise analysis to extract further information about fairness with respect to certain perturbations or subgroups. For Fair SA (Fig. 3), an element in the AUC matrix depicts (for a model), the signed AUC of the Fair SA curves[1]. We propose to use the L1 norm to marginalize the AUC matrix, row, column or matrix-wise.

4. Experiments

We perform experiments on two tasks: face verification (21) (comparing a perturbed image with original images to verify whether a perturbed-original image pair belongs to the same identity), and self-matching (matching a perturbed image to its original image). We use the following pre-trained models without additional fine-tuning: VGGFace2 (4), FaceNet (20), OpenFace, (1) and SeesawFaceNet (29). For all the models, we obtain responses from one of the final feature layers.

Data and perturbations. We investigate model performances on the following 9 perturbations: *Gaussian blur*, *gamma contrast*, *rotation*, *speckle noise*, *exposure*, *saturation*, *motion blur*, *JPEG compression*, and *vignette* (Examples in Appendix A). We evaluate our models on the CelebA test set which consists of ∼20k samples. Each sample has an identity label, and binary annotations for 40 attributes e.g. *male*, *wavy hair*, etc.

Verification. This task involves comparing the model responses for every perturbed image with every original image of all identities. A pair is a match if both images belong to the same identity. The overall performances of the models are evaluated in terms of Genuine Acceptance Rate (GAR) at False Acceptance Rate (FAR) (21) (Appendix B, Fig.

---

1. Similarly, for VPSA (Fig. 9), an element in the AUC matrix indicates the AUC of the IRC generated by VPSA for a particular perturbation.
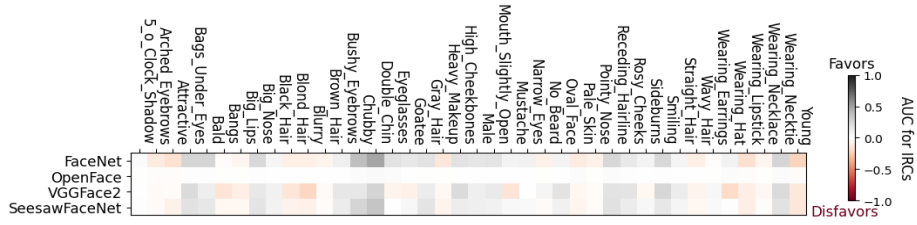
Figure 2: Depiction of GAR difference at 1% FAR between protected and unprotected subgroups at no perturbation.
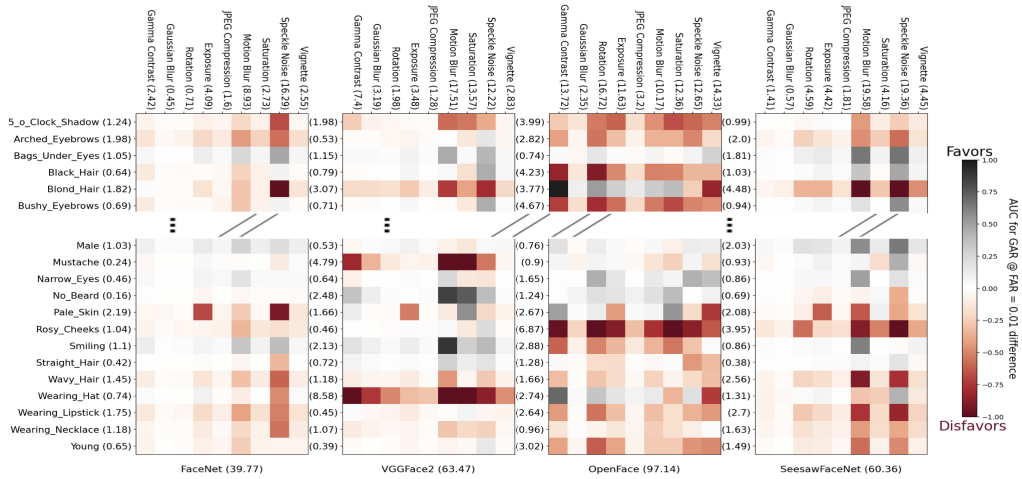

Figure 3: Depiction of AUC for GAR difference at 1% FAR for selected subgroups. Note that the numbers in attribute labels denote the attribute norms. Matrix L1 norms show that, on average, FaceNet is the most fair (lowest norm), and OpenFace is the least fair model.

7 shows that VGGFace2 performs the best and OpenFace the worst at no perturbation). We consider the GAR at 1% FAR metric which is commonly used for verification (21), and plot Fair SA curves representing the differences between GAR at 1% FAR across protected and unprotected subgroups at varying levels of perturbation (Fig. 1). For Fair SA verification experiments, we implement a pruning procedure that eliminates the pairs of images that are false matches at no perturbation, so that the GAR at 1 % FAR metric aligns for the protected and unprotected subgroups and we can use AUC to compare models. The smaller the absolute value of AUC for the Fair SA curves, the more fair is the algorithm. Note that we allow the AUC in Fair SA curves to be negative. In such cases, the unprotected subgroup is favored over the protected one. Fig. 2 indicates that OpenFace is the fairest model at no perturbation, with the lowest GAR difference at 1% FAR. However, matrix L1 norms on the Fair SA AUC model matrices[2] (Fig. 3) indicate that OpenFace is the least robustly fair

2. see complete verification results and L1 norms in Appendix B, Fig. 8

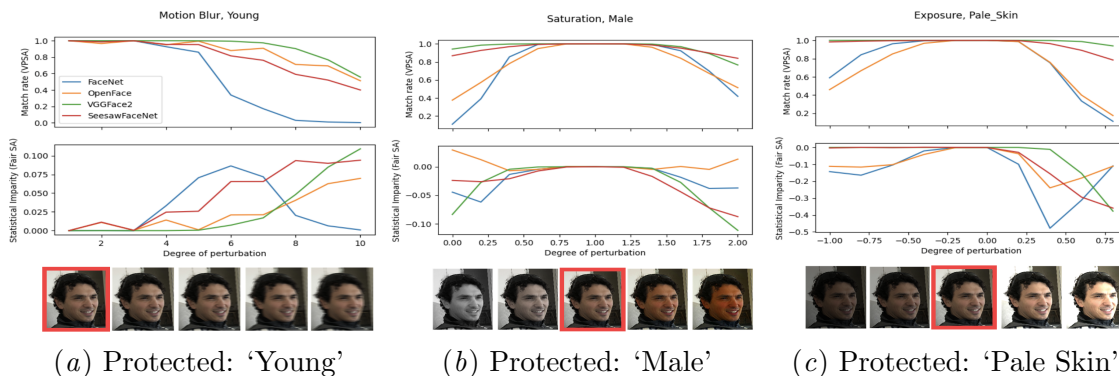(a) Protected: 'Young' (b) Protected: 'Male' (c) Protected: 'Pale Skin'

Figure 4: VPSA (top) and Fair SA (bottom) curves for self-matching. VPSA curves show that VGGFace2 is the most robust across *motion blur*, *saturation*, and *exposure* perturbations. However, Fair SA curves show that the models are biased upon subjecting the data to perturbations e.g., all the models are above 0 in Fig. 4(a), indicating that they favor the *young* subgroup when subjected to *motion blur*. Similarly, in the case of low and high *saturation* (Fig. 4(b)) and *exposure* (Fig. 4(c)), Fair SA indicates that models largely disfavor the *male*, and *pale skin* subgroups, respectively. Contrary to VPSA, Fair SA shows that at the highest level of perturbation, VGGFace2 is the least fair model.

model (based on the Fair SA score). Interestingly, while we see that at no perturbation, models (e.g. VGGFace2 and SeesawFaceNet) are fair with respect to different subgroups (e.g. *pale skin*) (Fig. 2), we see in Fig. 3 that the models can disfavor those subgroups significantly in the presence of certain perturbations (e.g. *exposure*).

Self-matching. This task is more generic in that it ignores the identities for the inputs, and can be used when a model is intended to be used for multiple tasks. It compares the model response for a perturbed image with that of the corresponding original image. VPSA shows that VGGFace2 largely outperforms the other models in robustness by withstanding perturbations to a greater degree (Fig. 4). Matrix L1 norms on the Fair SA AUC model matrices[3] (Fig. 5) confirm that VGGFace2 is the most robustly fair model. However, from a more detailed analysis, we can see that at the highest levels of perturbations like *motion blur*, *saturation*, and *exposure*, VGGFace2 is the least fair with respect to certain subgroups (*young*, *male*, *pale skin*). Note from Fig. 4(a) that Fair SA analysis on FaceNet indicates that with increasing *motion blur*, first the bias increases and then decreases. However, the fairness of FaceNet at higher perturbation is likely due to the fact that the performance of FaceNet is drastically degraded for strong motion blur. This indicates that fairness should be analyzed in conjunction with robustness for a complete outlook on the models.

Row- and column-wise analysis. Such an analysis can help with a broad and quick evaluation of the impact on subgroups and effect of perturbations. Across models for verification, the row-wise L1 norms indicate that models are most biased against *wearing*

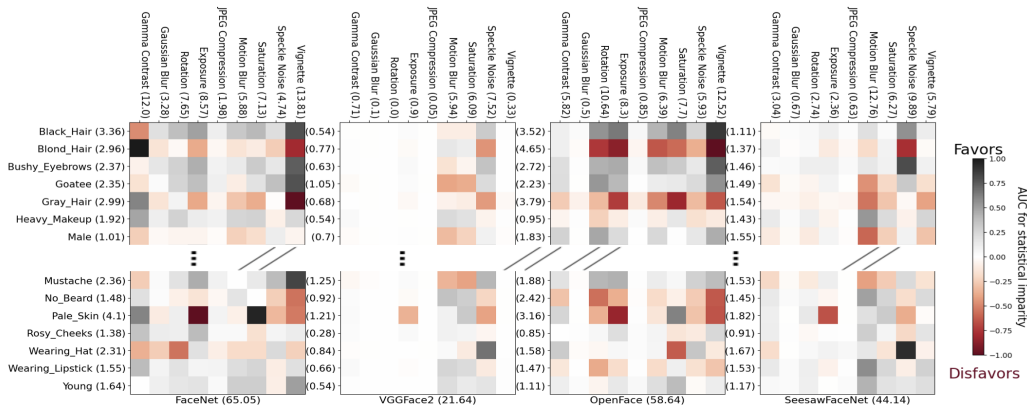3. see complete self-matching results and L1 norms in Appendix C, Fig. 11

Figure 5: Depiction of AUC for Fair SA curves (self-matching) for selected attributes. Note that the numbers in attribute labels denote the attribute L1 norms. Matrix L1 norms indicate that, on average, VGGFace2 is the most fair (with the lowest L1 norm), and FaceNet the least fair model.

hat, *blond hair*, and *rosy cheeks* (Fig. 3). Similarly, column-wise L1 norms across models indicate that the perturbations that most affect the model fairness are *speckle noise*, *motion blur*, and *saturation*. Note that this analysis can also be performed per model to help understand the impact on subgroups and effect of perturbations for individual models.

Effect of perturbations on salient features. Unsurprisingly, perturbations that degrade salient details adversely affect the performance on those sub-groups that have more salient features than those that do not. In the case of self-matching, increasing *Gaussian blur*, *motion blur*, or *JPEG compression* results in models favoring the *young* attribute. We hypothesize that blurring images of faces of older people may result in loss of details (that are not as prominent in the faces of young people) that are important for face recognition. Some observations are common to both self-matching and verification tasks, such as, increasing image *exposure* results in models favoring dark features, such as *black hair*, *bushy eyebrows*, and disfavoring features such as *pale skin*.

5. Conclusion

We propose an evaluation framework called Fair SA to measure group fairness in the presence of data perturbations and demonstrate its use on two tasks: verification and self-matching. An absolutely fair algorithm will have performance that is even on (all pairs of) protected and unprotected subgroups for incremental perturbations, and unfairness is any deviation from this fair algorithmic behavior. We provide experimental evidence that this unfairness can exist in some commonly used models trained on real world datasets, and we propose to visualize this unfairness via AUC matrices. Such matrices can be marginalized row and column wise to provide insights with respect to attributes or input perturbations. We show that even if models exhibit performance equity between protected and unprotected

subgroups when there is no data degradation, in the presence of perturbations the models may start to favor (or disfavor) certain data subgroups. In future works, we will explore how Fair SA can be utilized to mitigate biases via targeted training data augmentation driven by the type and the magnitude of perturbations that cause biased behavior.

References

[1] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, 6(2), 2016.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: there's software used across the country to predict future criminals. and it's biased against blacks. propublica 2016, 2016.

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

[4] 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.

[5] 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? *IEEE transactions on biometrics, behavior, and identity science*, 3(1): 101–111, 2020.

[6] Bethan Davies, Martin Innes, and Andrew Dawson. *An evaluation of South Wales police's use of automated facial recognition*. Universities' Police Science Institute, 2018.

[7] Mengnan Du, Fan Yang, Na Zou, and Xia Hu. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, 2020.

[8] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[9] Klemen Grm, Vitomir Štruc, Anais Artiges, Matthieu Caron, and Hazım K Ekenel. Strengths and weaknesses of deep learning models for face recognition against image degradations. *Iet Biometrics*, 7(1):81–89, 2018.

[10] Samil Karahan, Merve Kilinc Yildirum, Kadir Kirtac, Ferhat Sukru Rende, Gultekin Butun, and Hazim Kemal Ekenel. How image degradations affect deep cnn-based face recognition? In *2016 international conference of the biometrics special interest group (BIOSIG)*, pages 1–5. IEEE, 2016.

[11] Hachim El Khiyari and H. Wechsler. Face verification subject to varying (age, ethnicity, and gender)demographics using deep learning. *Journal of biometrics & biostatistics*, 7:1–5, 2016.

[12] Joel Z Leibo, Cyprien de Masson d'Autume, Daniel Zoran, David Amos, Charles Beattie, Keith Anderson, Antonio García Castañeda, Manuel Sanchez, Simon Green, Audrunas Gruslys, et al. Psychlab: a psychology laboratory for deep reinforcement learning agents. *arXiv preprint arXiv:1801.08116*, 2018.

[13] Greg Little, Sreekar Krishna, John Black, and Sethuraman Panchanathan. A methodology for evaluating robustness of face recognition algorithms with respect to variations in pose angle and illumination angle. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 2, pages ii–89. IEEE, 2005.

[14] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.

[15] Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P Dickerson. Fairness through robustness: Investigating robustness disparity in deep learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 466–477, 2021.

[16] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015.

[17] P J Phillips, Patrick J Grother, Ross J Micheals, D M Blackburn, Elham Tabassi, and Mike Bone. Face recognition vendor test 2002: Evaluation report. 2003.

[18] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 145–151, 2020.

[19] Brandon RichardWebster, So Yon Kwon, Christopher Clarizio, Samuel E Anthony, and Walter J Scheirer. Visual psychophysics for making face recognition algorithms more explainable. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 252–270, 2018.

[20] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[21] Maneet Singh, Richa Singh, Mayank Vatsa, Nalini K Ratha, and Rama Chellappa. Recognizing disguised faces in the wild. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(2):97–108, 2019.

[22] Tomáš Sixta, Julio CS Jacques Junior, Pau Buch-Cardona, Eduard Vazquez, and Sergio Escalera. Fairface challenge at eccv 2020: analyzing bias in face recognition. In *European Conference on Computer Vision*, pages 463–481. Springer, 2020.

[23] Nisha Srinivas, Karl Ricanek, Dana Michalski, David S Bolme, and Michael King. Face recognition algorithm bias: Performance differences on images of children and adults. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[24] Liang Tong, Zhengzhang Chen, Jingchao Ni, Wei Cheng, Dongjin Song, Haifeng Chen, and Yevgeniy Vorobeychik. Facesec: A fine-grained robustness evaluation framework for face recognition systems. *arXiv preprint arXiv:2104.04107*, 2021.

[25] Jennifer Valentino-DeVries. How the police use facial recognition, and where it falls short. *New York Times*, 12, 2020.

[26] Kushal Vangara, Michael C King, Vitor Albiero, Kevin Bowyer, et al. Characterizing the variability in face recognition accuracy relative to race. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[27] Brandon Richard Webster, Samuel E Anthony, and Walter J Scheirer. Psyphy: A psychophysics driven evaluation framework for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2280–2286, 2018.

[28] Xiao Yang, Dingcheng Yang, Yinpeng Dong, Wenjian Yu, Hang Su, and Jun Zhu. Delving into the adversarial robustness on face recognition. *arXiv preprint arXiv:2007.04118*, 2020.

[29] Jintao Zhang. Seesawfacenets: sparse and robust face verification model for mobile platform. *arXiv preprint arXiv:1908.09124*, 2019.

Appendices

Appendix A. Data and perturbations

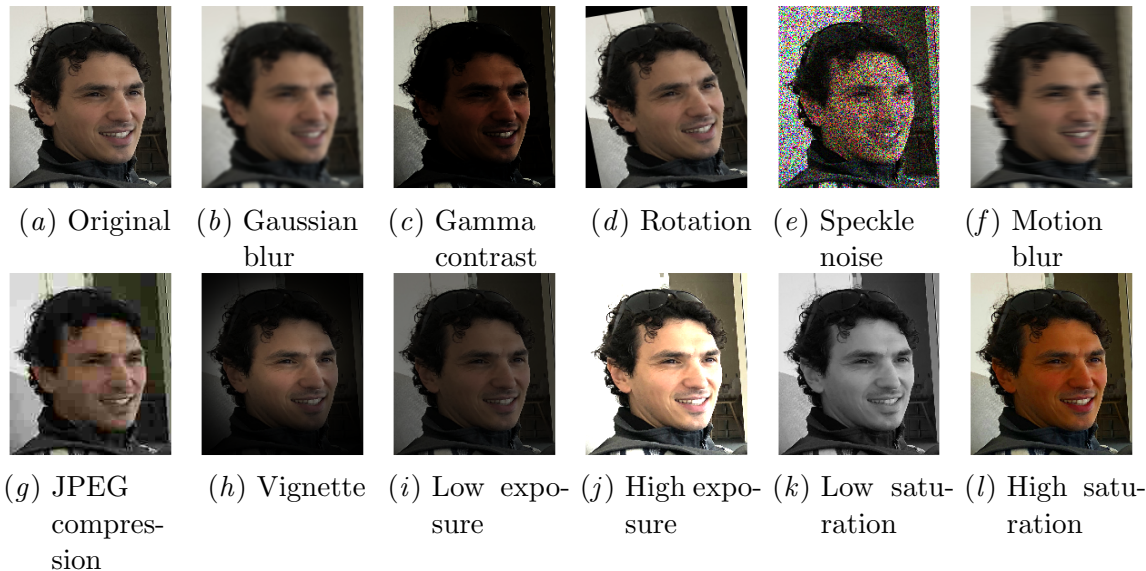In this section, the perturbations applied on a sample image are presented.

(*a*) Original (*b*) Gaussian blur (*c*) Gamma contrast (*d*) Rotation (*e*) Speckle noise (*f*) Motion blur

(*g*) JPEG compression (*h*) Vignette (*i*) Low exposure (*j*) High exposure (*k*) Low saturation (*l*) High saturation

Figure 6: Perturbations applied to a sample image 6(*a*). For exposure and saturation, the range is varied from low to high.

Appendix B. Verification

In this section, additional results for verification are presented. Fig. 7 shows the GAR at FAR curves for different models at no perturbation. Fig. 8 depicts the AUC matrices for all attributes.

Appendix C. Self-matching

In this section, additional results for self-matching are presented including the complete set of results on all perturbations and attributes for Fair SA, along with the results for experiments using VPSA's pruning step. Fig. 9 shows the AUC matrix for IRCs obtained by VPSA without and with pruning, and we see that the trends obtained with pruning (Fig. 9(*b*)) are largely similar to those without pruning (Fig. 9(*a*)). Fig. 10 compares the plots obtained with and without the pruning step for a sample perturbation and protected subgroup. Figures 11 and Fig. 13 depict the AUC matrices for Fair SA for all attributes without and with pruning respectively.
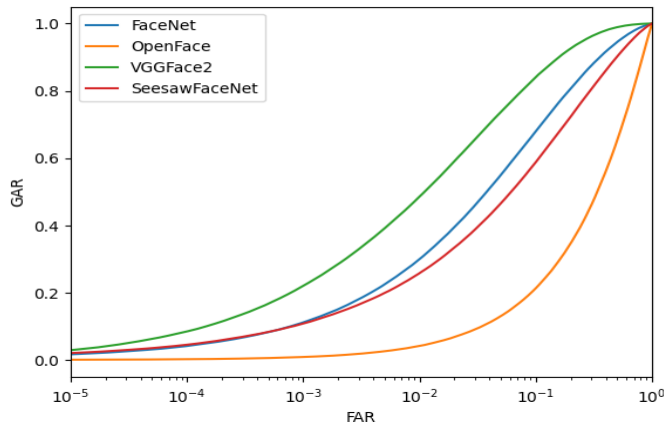
Figure 7: Genuine Acceptance Rate (GAR) at False Acceptance Rate (FAR) for verification on CelebA at no perturbation. Higher the AUC, the better the performance of the model. Thus, VGGFace2 performs the best, and OpenFace the worst amongst the models compared.
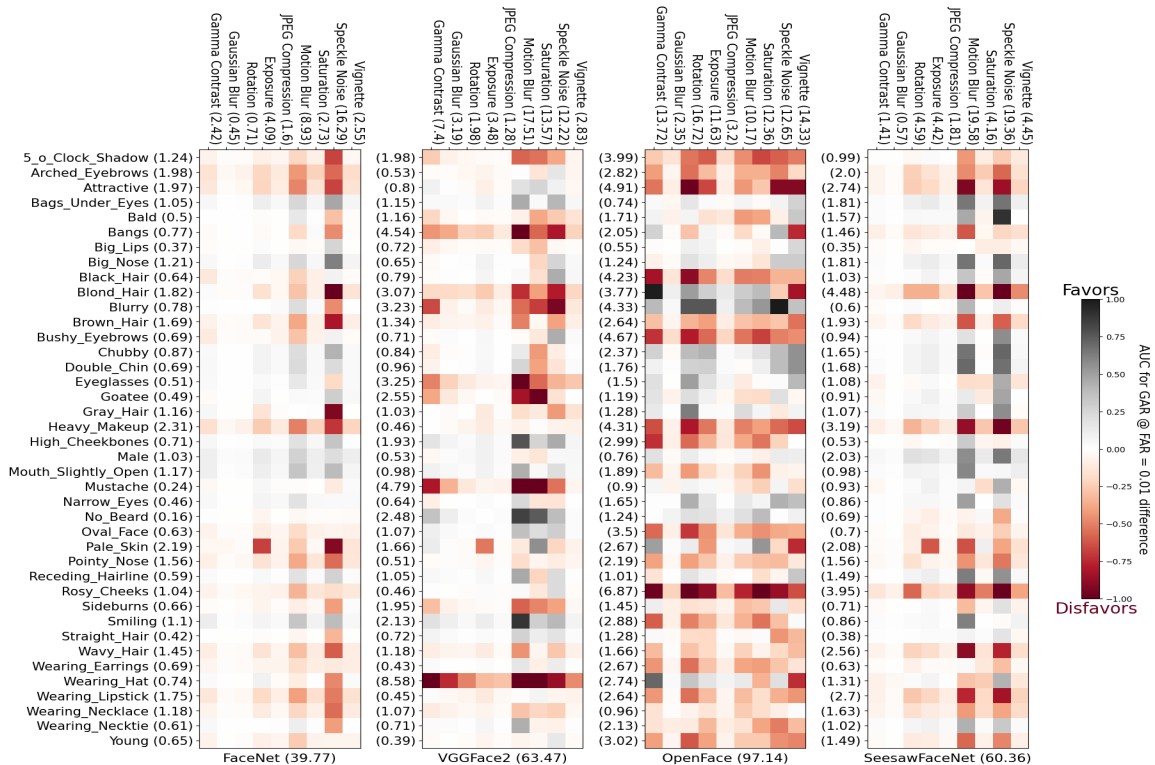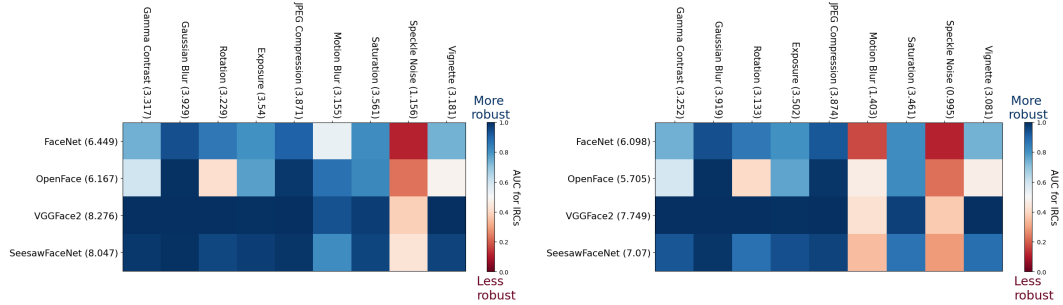

Figure 8: Depiction of AUC for difference between GAR @ 1% FAR values for all protected and unprotected subgroups.

(*a*) AUC matrix for VPSA without pruning

(*b*) AUC matrix for VPSA with pruning

Figure 9: Fig. 9(*a*) Depiction of AUCs for IRCs as computed by VPSA without pruning. Note that the numbers in the labels denote the L1 norms for the models and perturbations. Computing L1 norms for the rows indicates that, on average, VGGFace2 is the most robust (with the highest L1 norm), and OpenFace the least robust amongst the models compared. Fig. 9(*b*) Depiction of AUCs for IRCs as computed by VPSA with pruning. Trends observed are largely similar to those obtained without pruning. Computing L1 norms for the rows also indicates the same order of robustness of the models i.e., on average, VGGFace2 is the most robust (with the highest L1 norm), and OpenFace the least robust amongst the models compared, on CelebA.
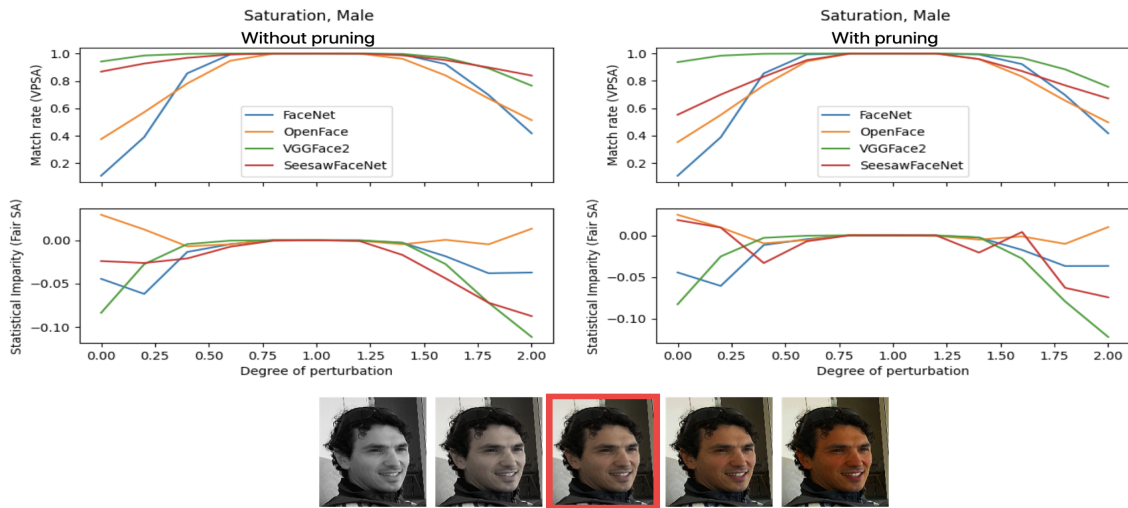

Figure 10: Example VPSA and Fair SA curves without and with pruning for saturation with *male* as the protected subgroup. We observe similar trends for individual models for without and with pruning.
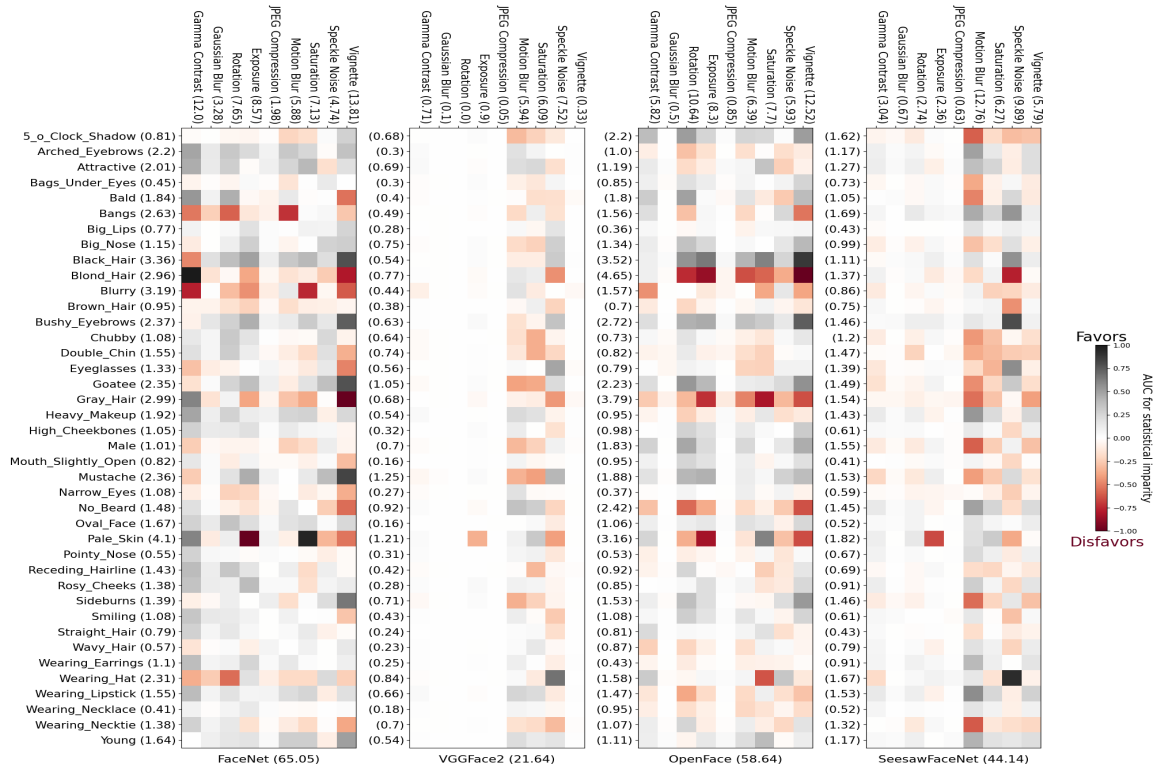
17

Figure 11: Depiction of AUC for statistical imparity curves all attributes without pruning. Note that the numbers in the labels denote the L1 norms.
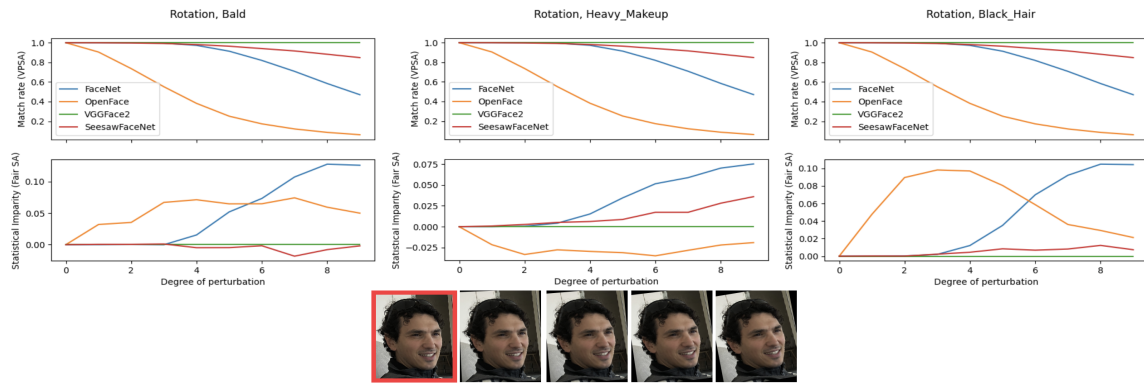

Figure 12: Depiction of VPSA and Fair SA curves for the rotation perturbation across sample protected subgroups. We can see that VGGFace2 is absolutely robust and fair since the AUC for the VPSA and Fair SA curves is 1 and 0 respectively.
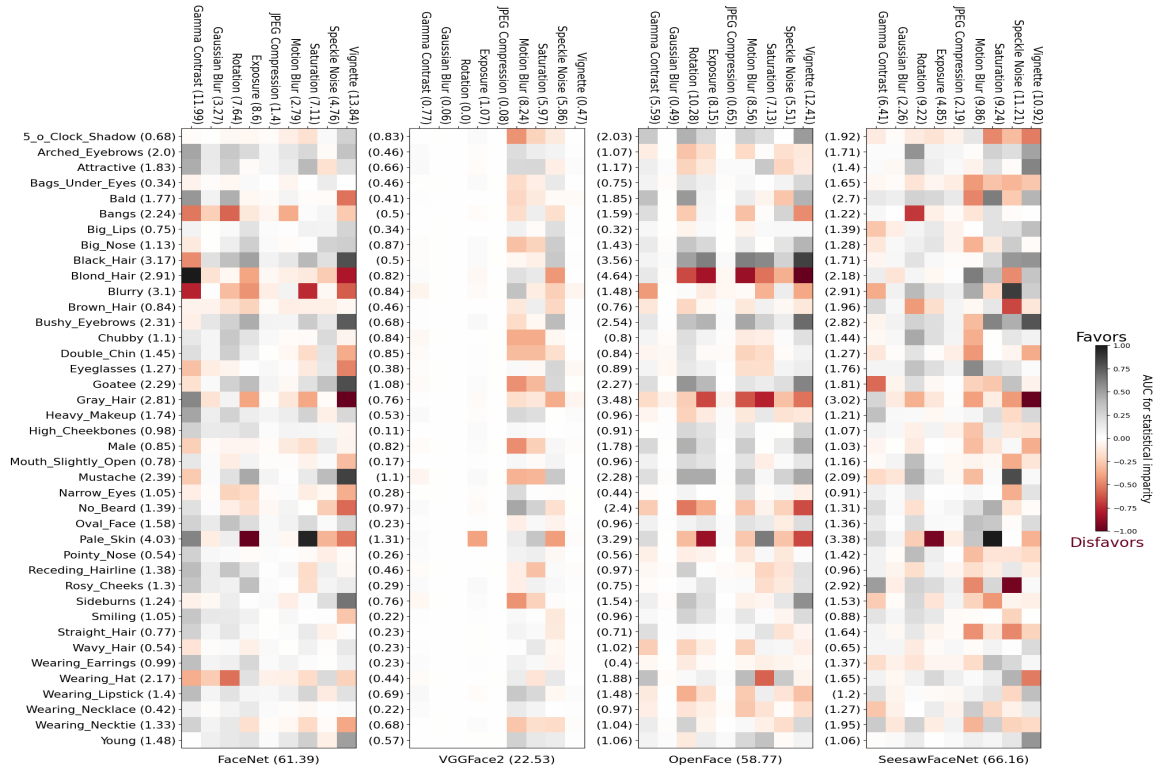
Figure 13: Depiction of AUC for statistical imparity curves for the CelebA dataset with pruning. Note that the numbers in the labels denote the L1 norms. Computing matrix L1 norms for the matrices indicates that, on average, VGGFace2 is the most fair (with the lowest L1 norm), and SeesawFaceNet the least fair amongst the models compared.