Beyond Skin Tone: A Multidimensional Measure of Apparent Skin Color

William Thong[*]
Sony AI

Przemyslaw Joniak[†]
The University of Tokyo

Alice Xiang
Sony AI

Abstract

This paper strives to measure apparent skin color in computer vision, beyond a unidimensional scale on skin tone. In their seminal paper Gender Shades, Buolamwini and Gebru have shown how gender classification systems can be biased against women with darker skin tones. Subsequently, fairness researchers and practitioners have adopted the Fitzpatick skin type classification as a common measure to assess skin color bias in computer vision systems. While effective, the Fitzpatick scale only focuses on the skin tone ranging from light to dark. Towards a more comprehensive measure of skin color, we introduce the hue angle ranging from red to yellow. When applied to images, the hue dimension reveals additional biases related to skin color in both computer vision datasets and models. We then recommend multidimensional skin color scales, relying on both skin tone and hue, for fairness assessments.

1. Introduction

This paper focuses on measuring apparent skin color in images for fairness benchmarking, and moves toward a multidimensional score beyond the skin tone depth. Adverse decisions can arise in common computer vision models, as strikingly identified by Buolamwini and Gebru [8]. This has an impact on real-life applications, as models can produce wrong skin lesion diagnostics [15] or incorrect heart rate measurements [5, 27, 62] for individuals with darker skin tones. It is therefore critical to identify to what extent visual datasets and models are affected by changes in skin color. To achieve this, it becomes necessary to develop comprehensive skin color scores to characterize images of individuals, otherwise a fairness evaluation based on skin color would not be feasible.

Describing the apparent skin color remains an open challenge, as the final visual color perception results from a complex physical and biological phenomenon [34]. The skin is a multilayered structure, which varies among individuals: not every individual will have the same amount and
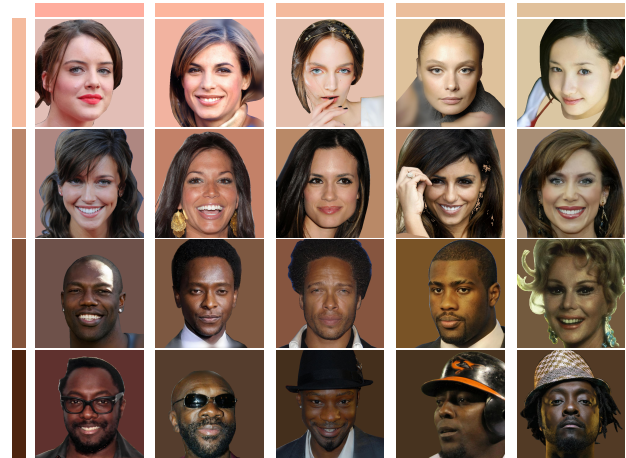

Figure 1: **Apparent skin color** in images results from a complex phenomenon, where it varies in tone from light to dark (vertical axis) and hue from red to yellow (horizontal axis). This figure depicts examples from CelebAMask-HQ. In this paper, we introduce a measure of the skin hue, besides the commonly used skin tone, to quantify apparent skin color variations.

distribution of carotene, hemoglobin, or melanin throughout the different layers [6, 10]. Hence, modeling the intrinsic skin color through the amount of reflected, scattered, or absorbed light for every individual is a difficult and complex task in images [67]. Additionally, color perception from the human visual system adds more complexity to the problem, as colors can be differently perceived depending on their context or people's cultures [26]. Instead, it is simpler and more appropriate to develop representative color scales to capture the variations in apparent skin color.

The commonly accepted standard for skin color scale is the Fitzpatrick skin type classification [24], which categorizes skin color into six different types based on *skin tone*, ranging from *light* to *dark*. It has unsurprisingly become a useful tool for fairness analysis [8,57] because skin tone annotations may serve as a proxy for race or ethnicity annotations. Indeed, such sensitive attributes are subject to significant data privacy protections [22], and are often unavailable or inferred in visual datasets [2, 36], making the measure of

---

[*]Corresponding author: william.thong@sony.com
[†]Work done while an intern at Sony AI

skin tone an alternative fairness tool. Yet, while practical and effective, reducing the skin color to its tone is limiting given the skin constitutive complexity. As illustrated in Figure 1, apparent skin color also varies along other axis, such as the *skin hue*. For example, when aging, Asian skin becomes darker and *more yellow* while Caucasian skin becomes darker and *redder* [17]. Focusing on skin tone would not capture such change in hue as it only assesses the skin lightness or darkness. In this paper, we therefore promote a multidimensional scale to better represent apparent skin color variations among individuals in images.

Akin to previous works on fairness analysis in computer vision (*e.g.*, [8, 57, 75, 78]), we are interested in characterizing *apparent* skin color rather than *true* skin color. The *apparent* skin color is the one depicted in images, and the one that a computer vision model would see, while the *true* skin color characterizes the constitutive skin color without the influence of external factors such as illumination or color cast. Assessing the *true* skin color is more important for dermatology [41, 60] or cosmetics [6, 46] applications as the constitutive color leads to more specific diagnostics or treatments and requires an active involvement with practitioners to avoid any misusage or mistrust [28]. In this paper, we rather focus on the assessment of computer vision models, which are fed images in the wild, and thus only consider the *apparent* skin color in images.

Our main contribution is to demonstrate the relevance and benefits of a multidimensional skin color scale for fairness assessments in computer vision. *First*, we introduce a step towards more comprehensive apparent skin color scores. Rather than classifying skin color in types, as done with the Fitzpatrick scale, we measure automatically and quantitatively skin color in a multidimensional manner in images. We propose to focus on the perceptual light L^*, as a measure of skin tone, and the hue angle h^*, as a measure of skin hue; which results in a multidimensional measure for every image. *Second*, we showcase the benefits of a multidimensional measure of skin color by (i) quantifying to what extent common image datasets are skewed towards light-red skin color and under-represent dark-yellow skin color, and how generative models trained on these datasets reproduce a similar bias; (ii) revealing multidimensional skin color biases in saliency-based image cropping and face verification models; and (iii) measuring the causal effect of skin color in attribute prediction in multiple commercial and non-commercial models. Overall, our contributions to assessing skin color in a multidimensional manner offer novel insights, previously invisible, to better understand biases in the fairness assessment of both datasets and models.

2. Background on Skin Color Scales

Scoring the skin color of individuals has wide implications in various fields. We focus in this section on the Fitz-

patrick skin type classification scale [24], which is widely used in practice but starts to raise ethical concerns due to its original design and its lack of representativeness [55]. We describe below its categorization methodology, as well as its current usage, limitations and impact for fairness assessment in computer vision.

Fitzpatrick skin type classification scale categorizes the skin response to ultraviolet A light [24]. This was originally used to assess the suntanning pathways of skin, *i.e.*, how sensitive skin is to sun exposure, and how it evolves over time. While the classification was first proposed to categorize Caucasian skin into four different types, it has later been extended to include brown and darker skins, resulting in six different skin types [24]. The scale starts from type I lighter skin, which "always burns and never tans", and continues to type VI darker skin, which "never burns".

Since its original introduction, the Fitzpatrick skin type has been, and still is, a central role in the literature. It now goes beyond the characterization of suntanning pathways, and serves as a color classification scale for skin in fields such as cosmetics to measure the efficacy of aesthetic cosmetology (*e.g.*, [6, 46]) or clinical dermatology (*e.g.*, [41, 60]) for skin analysis and treatment. In this paper, akin to the seminal work of Buolamwini and Gebru [8], we focus on leveraging apparent skin color scores to study potential discrimination coming from data and model biases.

However, such widespread use of the Fitzpatrick skin type raises ethical questions as it was originally developed for Caucasian skin tones, implying a lack of representativeness and utility for other groups. As highlighted by Pichon *et al*. [55], the Fitzpatrick skin type has limited utility in dermatology for several ethnic groups such as Asians [12], Arabs [68] or African-Americans [74], because characterizing how skin "burns" or "tans" might be restrictive. Adding race labels does not provide better predictors for skin characteristics either, as it is still not enough to represent its variation [32]. This representativeness limitation accentuates the need for more comprehensive classification types to characterize skin color. In this paper, we go beyond the assessment of skin tone and propose the skin hue to achieve a multidimensional score for apparent skin color.

Another limitation of the Fitzpatrick skin type comes from the annotation process. While the usage of a spectrophotometer to measure the skin reflectance remains the standard to provide the ground truth score of the true or constitutive skin color [46], it might not be available in practice. As such skin type annotation is usually self-reported by the subject [24], labeled by domain experts [72], or labeled by non-domain experts [8]. Unfortunately, in all the above cases, it appears that misclassification can occur in the Fitzpatrick scale, as current skin type definitions are not objective and descriptive enough, and do not capture the skin color variability among different ethnicities [20]. An-

other reason for such misclassification comes from the annotator bias, where the demographics and background of an annotator influence the skin color score assignment [9]. These annotation limitations indicate the need for quantitative and more objective measures for skin color. In this paper, we extract skin color scores in facial images automatically, without the need of external human annotators, to obtain a representative scalar scores.

Social impact of skin color scales. In his seminal paper, Fitzpatrick [24] considers race and ethnicity as cultural and political constructs. Current scales have been developed to describe skin color rather than providing a race or ethnic label. Still, the Fitzpatrick skin type has deviated from its original usage as up to fifty percent of dermatologists use the Fitzpatrick types to describe race and ethnicity [72], and doing so can be harmful as such simplification ignores their complexity [30]. Furthermore, the lack of representativeness in the Fitzpatrick skin type can lead to mis-representing some ethnical groups [29, 54], as skin color can have a wide spectrum within an ethnic group or between two groups. For example, two groups can have a similar skin tone but differ on other axes. In light of this, this paper focuses on understanding performance discrepancy related to skin color, with more comprehensive scores to better represent human skin color variation.

Applications in fairness assessment. In recent years, there has been growing awareness on the limitations of computer vision models to be biased against under-represented groups [8]. It is thus critical to develop fairness tools that can help assess potential biases and document them in datasheets [25] and model cards [50]. In the context of this paper, we are interested in developing a fairness tool to better assess and quantify biases related to skin color. Towards this goal, collecting skin tone annotations has enabled bias identification in facial recognition [8, 35, 57], image captioning [78], person detection [75], skin image analysis in dermatology [15, 41], face reconstruction [23] and detecting deepfakes [31] among other tasks. In this paper, we build on this line of work and propose complementary and novel tools for measuring and extracting multidimensional skin color scores, and showcase their relevance and effectiveness for revealing dataset and model biases.

While the paper focuses on bias identification and measurement, our multidimensional skin color scale could also enable how to address bias mitigation. Such related works commonly rely on data augmentation [59], adversarial debiasing [69], contrastive learning [52], independent classifiers [71], or debiasing both feature and label embeddings [65]. Our skin color scores could inform which samples models are struggling with, and provide a solution by augmenting images or mitigating model representations in a multidimensional manner.

3. Multidimensional Skin Color Scores

Given the limitations of the Fitzpatrick scale in its definition or annotation process, deriving quantitative metrics enables more reliable skin color scores. Instead of asking a subject for a self-identification of the skin color type or collecting the skin color type from an annotator, it is preferable to compute a skin color score from a point measurement. Indeed, this mitigates the subjectivity of the (self-)annotator as well as the inter-rater reliability [10, 46, 73]. In this section, we introduce the importance of the color space for representing images and define metrics for skin color scores.

Colorimetry aims to represent faithfully the human perception of colors. Towards this goal, the Commission Internationale de l'Eclairage (CIE) establishes standards regarding illuminants, tristimulus values, or color spaces [16]. While images are usually represented in the standard RGB space, it might be more relevant to represent them in a color space that better reflects the human perception when assessing skin color variation.

In this paper, we are particularly interested in the CIE $L^*a^*b^*$ (CIELAB) color space, originally introduced in 1976 [16], which correlates with the response of the human eye by covering its entire range of color perception: the L^* component corresponds to the perceptual lightness and ranges from black at value 0 to white at value 100; the a^* component describes the green-red opponent colors, with negative values corresponding to green and positive values to red; and the b^* component refers to the blue-yellow opponent colors, with negative values corresponding to blue and positive values to yellow.

Individual typology angle (ITA) provides a quantitative alternative to the Fitzpatrick scale [10]. The individual typology angle is commonly used to describe the skin color on spectrophotometer measurements for aesthetic cosmetology [46] or clinical dermatology [74]. In fairness analyses, it has notably been applied to natural images of faces images [36,48], and skin analysis [41]. Concretely, the individual typology angle is defined in the CIELAB color space as follows:

$$\text{ITA} = \arctan\left(\frac{L^* - 50}{b^*}\right) \times \frac{180°}{\pi} \qquad (1)$$

where a perceptual lightness at value 50 corresponds to a maximum chroma. Only the L^* and b^* components are selected as they are the ones that explain best the variation of the suntanning pathways of the skin.

Despite its advantages of providing a quantitative measure, the individual typology angle is not a comprehensive skin color score. It was originally developed for the suntanning pathways of Caucasian skin [10], similar to the Fitzpatrick scale [24]. In their original paper, Chardon *et al*. [10] even raise this limitation as they acknowledge the

absence of the a^* component, which should be taken into account to better represent other types of population. Furthermore, while there exists a mapping between angles and Fitzpatrick skin color types (*e.g.*, values above 28° correspond to light skin tones, type I to III; and to dark skin tones, type IV to VI, otherwise), the low correlation with expert annotations makes it less reliable [28]. Given these limitations, there is a need to provide a more comprehensive assessment of skin color.

Hue angle provides a colorimetric measure to describe the perceived gradation of color [16]. In the CIELAB color space, the hue angle is defined as follows:

$$h^* = \arctan(b^*/a^*) \qquad (2)$$

where h^* goes from 0 to 360° as a^* and b^* components are unbounded. That said, we are mainly interested in angles between 0 and 90° (*i.e.*, positive values of a^* and b^*), as the skin color is expressed through red and yellow colors [10]. Initially proposed by Weatherall and Coombs [73] for skin color measurements, the hue angle has proven to be suitable as an additional dimension to design proprietary skin color scales for aesthetics cosmetology [17].

While these other fields have been exploring a multidimensional measure for skin color, fairness benchmarks in computer vision have mainly focused on a unidimensional measure of skin tone. In this paper, we propose to consider both the skin tone and the skin hue as a multidimensional measure for skin color analysis in images. To the best of our knowledge, the hue angle has not been used for fairness analysis in computer vision. We take inspiration from de Rigal *et al.* [17], and focus on the measurement of L^* for skin tone and h^* for skin hue. Measuring L^* instead of ITA avoids having correlated measures as both ITA and h^* contain the b^* component.

To extract the apparent skin color scores from images, we build on the algorithm initially proposed by Merler *et al.* [48] for the Diversity in Faces dataset (see Section 4.6 in their paper), and generalize their method to handle any scalar scoring value, any face pose and facial variation. We provide the details of the methodology and a robustness analysis to different illuminations in Figure S3. We focus in the next section on understanding the effect of skin color in image datasets on computer vision models.

4. Fairness Benchmarking with Multidimensional Skin Color Scores

First, we quantify the skin color bias in face datasets, and in generative models trained on such datasets. This reveals a skewness towards light-red skins and an underrepresentation of dark-yellow skins. Second, we break down results by skin color of saliency-based image cropping and face verification algorithms. This reveals that
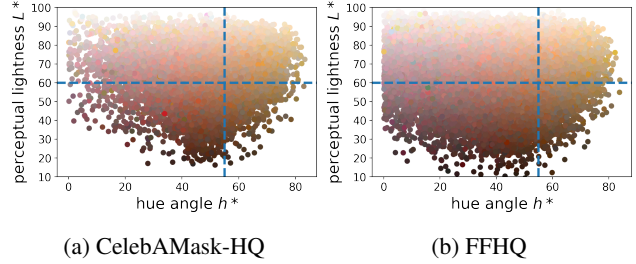

(a) CelebAMask-HQ (b) FFHQ

Figure 2: **Skin color distribution** on common face datasets. Every dot in the scatter plot corresponds to an image sample in the dataset. The skin tone threshold is at value 60 (light *vs*. dark), and the hue threshold at value 55° (red *vs*. yellow).

model bias not only exists for skin tone, but also for skin hue. Third, we investigate the causal effect of skin color in attribute prediction. This reveals performance differences when skin color changes, as classifiers tend to predict people with lighter skin tones as more feminine, and people with redder skin hue as more smiley. Code can be found at https://github.com/SonyResearch/apparent_skincolor.

4.1. Skin color bias in datasets

Skin color scores enable the assessment of potential biases in a given image dataset. By extracting them for every sample in the dataset, it is possible to estimate the distribution of every subgroup and characterize how unbalanced the distribution might be. We focus on the CelebAMask-HQ [43], composed of 30,000 images available with a non-commercial research agreement, and FFHQ-Ageing [51], composed of 70,000 images available with a Creative Commons BY-NC-SA 4.0 license. Both datasets provide ground truth segmentation masks of the skin, which we use to extract skin color scores. Note that both datasets are derivative datasets from CelebA [44] and FFHQ [38], respectively, with images crawled from social media platforms.

Figure 2 depicts the distribution in terms of perceptual lightness L^* and hue angle h^*. When L^* is over 60, it corresponds to a light skin tone (and conversely for a dark skin tone). When h^* is over 55°, it corresponds to a skin turning towards yellow (and conversely for a skin turning towards red). Threshold values are taken from Ly *et al.* [46].

Table 1 provides a quantitative assessment of the skin tone and the skin hue. As commonly hypothesized in the literature [37, 38], both datasets are skewed towards light skins. This paper enables to quantify such skin color bias. Measuring the hue angle further shows that both datasets are also skewed towards red skins. We also explore in Section A.3 the skin color distribution per ethnicity along the tone and hue dimensions. We show the relevance of skin hue to better represent different ethnical groups, as some

		Skin tone		
		Light	Dark	Total
Hue	Red	**46.95**	19.28	**66.23**
	Yellow	27.48	06.29	33.77
	Total	**74.43**	25.57	100

(a) CelebAMask-HQ

		Skin tone		
		Light	Dark	Total
Hue	Red	**52.82**	29.06	**81.87**
	Yellow	13.44	04.68	18.12
	Total	**66.26**	33.74	100

(b) FFHQ

Table 1: **Skin color bias** in common face datasets (in %). Images are skewed towards a light skin tone and a red skin hue while dark and yellow skin colors are underrepresented.

		Skin tone		
		Light	Dark	Total
Hue	Red	**53.05**	28.71	**81.76**
	Yellow	13.46	04.78	18.24
	Total	**66.51**	33.49	100

(a) GAN – StyleGAN3

		Skin tone		
		Light	Dark	Total
Hue	Red	**53.29**	30.08	**83.37**
	Yellow	12.01	04.62	16.63
	Total	**65.30**	34.70	100

(b) Diffusion – P2

Table 2: **Skin color bias** in the output of generative models when trained on FFHQ (in %, for 10,000 images). Both StyleGAN3 and P2 reproduce the skin color bias of FFHQ.

ethnicities can have a similar skin tone but different skin hue profile. We also provide additional histogram plots in Section A.4, where we go beyond a simple binary thresholding for both L^* and h^*. The different thresholds confirm the over-representation of light skin tones and red skin hues.

Additionally, we measure the skin color bias on the output of generative models trained on FFHQ. To achieve this, we generate 10,000 images with a generative adversarial network (StyleGAN3 [37]) and a diffusion model (P2 [13]). Table 2 shows that both StyleGAN3 and P2 are reproducing the skin color bias present in the FFHQ dataset, with P2 slightly amplifying the bias over the original dataset distribution. In future dataset collections, we recommend tracking both skin tone and skin hue for skin color scores, such that all subgroups are well balanced, which in turn makes the dataset more diverse.

4.2. Skin color bias in models

4.2.1 Saliency-based image cropping

Task. The saliency-based image cropping task produces a thumbnail based on a saliency map. Such cropping algorithms are notably useful to select which region of a large image to display on a smaller screen [19,70,76]. In the context of this paper, we are inspired by previous works [7,77] and focus on cropping images with two faces. This setting can then be related to a pairwise comparison where the output decides which face to keep to produce the thumbnail.

Discrimination can happen in case the cropping algorithm favors a specific member of a protected attribute. For example, a method could consistently prefer a facial image of a light-skinned person over a dark-skinned one. In fact,

Yee *et al*. [77] have shown that the Twitter cropping algorithm is prone to gender or skin tone biases, as well as male-gaze-like artifacts. In this paper, we focus on assessing the skin color bias on the saliency-based image cropping task.

Method. To build a benchmark for saliency-based image cropping, we start from the one initially proposed by Birhane *et al*. [7]. We extend it to multiple race labels and propose a quantitative score to measure the probability of a face to be selected by the cropping algorithm. We are given a dataset $\{x_i, a_i\}_{i=1}^N$ of N facial images x with their associated race label a. We also extract skin color scales corresponding to perceptual lightness y^{L^*} and the hue angle y^{h^*} for every facial image. To evaluate the cropping algorithm, we build images to include a unique pair of two facial images x_i and x_j with $i \neq j$ (see Figure 3a).

We propose to compute the Elo rating [21] for every facial image. Every pairwise comparison is considered as a game between two facial images x_i and x_j. The objective is to get a rating for every facial image, *i.e*., R_i and R_j, which indicates the probability of the face to be selected by the cropping algorithm. The probability for image x_i to be chosen is defined as $p_i = 1/(1 + 10^{(R_j - R_i)/M})$, where M acts as a temperature for the sigmoid function. Intuitively, if there is a difference of M points between i and j, this means that i has 10 times more chance to be chosen. Conversely, for image x_i we have $p_j = 1 - p_i$. The outcome of the cropping algorithm S_i for x_i is equal to 1 if i wins and 0 if j looses, and can be used to update the player score with $R'_i = R_i + K(S_i - p_i)$, and conversely for x_j. Following common practice in chess playing [21], we set $M = 400$ and $K = 16$ and we initialize scores at 1400.

For benchmarking, we rely on Chicago Face Dataset (CFD) [47] and CFD-India [42], for a total of 739 unique facial images acquired in a controlled setting available for non-commercial research purposes. The dataset includes a self-reported gender label: 359 females and 380 males; as well as self-reported ethnic labels: 109 Asians, 197 Blacks, 142 Indians, 108 Latinos, 183 Whites. Individuals have given their informed consent for data collection. We extract the skin masks with DeepLabV3 [11], trained on CelebAMask-HQ [43] as done in FFHQ-Ageing [51]. From CFD, we sample pairs of facial images that are equally distributed with respect to gender and ethnicity. The product of both gender and ethnicity label sets results in 10 intersectional groups, forming a total of 45 pairwise combinations. We sample 500 pairwise comparisons for each combination, for a total of 22,500. To form the final image, we concatenate both facial images vertically with a white space in between and preserve their aspect ratio. This setting is highly inspired by Birhane *et al*. [7]; it differs by the inclusion of many ethnic groups instead of only two and it preserves the aspect ratio of images, as we noticed effects on the cropping algorithm when resizing images.

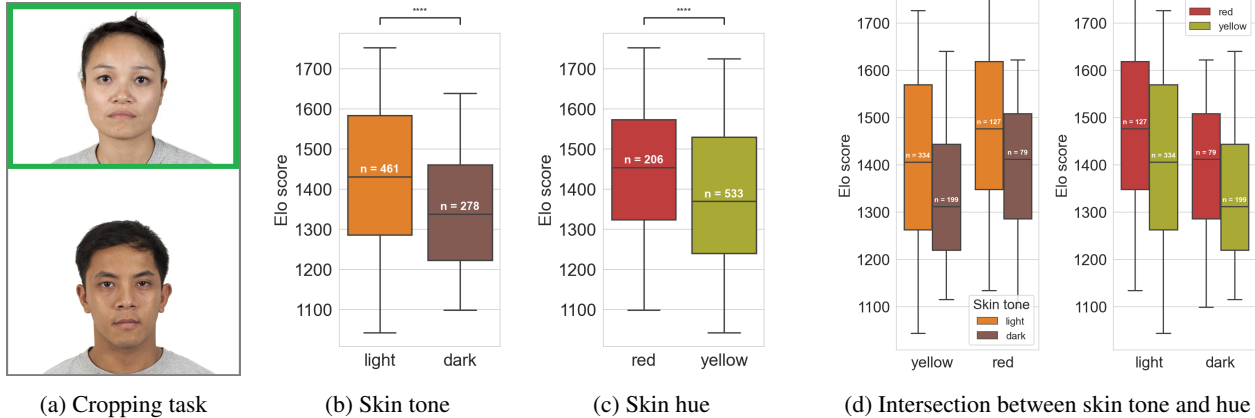(a) Cropping task   (b) Skin tone   (c) Skin hue   (d) Intersection between skin tone and hue

Figure 3: **Saliency-based image cropping** on CFD. (a) Illustration of cropping task where the model is fed an image with two faces and predicts which one to keep (green box). (b) Performance differences between light and dark skin tones are statistically significant ($p < 0.0001$). (c) Differences between red and yellow skin hues are also statistically significant ($p < 0.0001$). (d) The intersectional analysis also reveals statistically significant differences ($p < 0.01$), except between light-yellow and dark-red skin colors. Complementary to the skin tone, the skin hue reveals additional differences in performance.

Results. Figure 3 presents the results of the saliency-based cropping algorithm of Twitter on CFD. We report the Elo score of the 739 individuals after 22,500 pairwise comparisons for image cropping. A high Elo score indicates a preference of the algorithm to select the individual to be kept for cropping. We find that light skin tones are preferred over dark skin tones (Figure 3b) with an average Elo score indicating a 60.73% preference, which is confirmed with a statistically significant independent t-test ($p < 0.0001$). The algorithm also prefers red skin hues over yellow ones with an average Elo score indicating a 58.25% preference, which again is statistically significant (Figure 3c). We further consider intersectional groups of skin tone and hue (Figure 3d). When performing independent t-tests with a Bonferroni correction, we find that all pairwise groups are statistically different ($p < 0.01$), except for the light-yellow and dark-red skin colors. Overall, this benchmark reveals a multidimensional hierarchy on the skin color preference of the Twitter cropping algorithm, with light-red skin colors being favored and dark-yellow skin colors being disfavored.

4.2.2 Face verification

Task. The face verification task compares a pair of facial images to verify whether they belong to the same individual or not. In the context of this paper, we evaluate non-commercial and publicly available facial recognition models. Note that, similar to the previous experiment, we focus on exposing potential skin color biases rather than improving the selected face verification models.

Method. We adopt a standard benchmark in face verification, where models are evaluated on their accuracy to

Model	Overall	Independent groups				Intersectional groups			
		Skin tone		Skin hue		L+R	L+Y	D+R	D+Y
		Light	Dark	Red	Yellow				
ArcFace	95.20	**95.52**	94.39	**95.36**	95.07	**96.12**	94.82	92.55	**96.77**
FaceNet	94.40	**94.55**	94.04	**94.89**	93.81	**95.61**	93.30	93.17	**95.16**
Dlib	94.20	**94.41**	93.68	**94.71**	93.58	**94.83**	93.90	**94.41**	92.74

Table 3: **Face verification** on LFW. Individuals with a light skin tone or a red skin hue are better identified by all face verification models. Intersectional results confirm this pattern for lighter tones, but differ for darker tones.

predict whether a pair of facial images corresponds to the same individual. Following previous works (*e.g.*, [18, 61]), we evaluate on the Labeled Faces in the Wild (LFW) dataset [33], which contains 1,000 test pairs (500 positives and 500 negatives). We rely on the pre-processed version of LFW available in scikit-learn [53], and on the DeepFace repository[1] to run ArcFace [18], FaceNet [61] and Dlib [40] face recognition models.

Results. Table 3 presents the face verification results of several methods on LFW, broken down by skin tone (light *vs*. dark) and by skin hue (red *vs*. yellow). All models tend to prefer light skin tones and red skin hues. Specifically, both ArcFace and Dlib models are affected by skin tone differences, as they better verify the identify of light-skinned individuals. FaceNet has a different behaviour, it is more robust to skin tone differences but as prone to skin hue differences as Dlib. When looking at intersectional groups, ArcFace and FaceNet perform lower for light and yellow skin colors, as well as dark and red skin colors; Dlib has

---

[1]https://github.com/serengil/deepface

| $L^*=67.29$ | $L^*=61.92$ | $L^*=71.74$ | $L^*=48.20$ | $L^*=64.09$ | $L^*=61.50$ | $L^*=69.56$ | $L^*=54.91$ |
| $h^*=42.72$ | $h^*=46.76$ | $h^*=45.76$ | $h^*=46.69$ | $h^*=36.76$ | $h^*=55.27$ | $h^*=47.87$ | $h^*=43.54$ |

| $L^*=45.51$ | $L^*=44.43$ | $L^*=58.73$ | $L^*=23.78$ | $L^*=52.05$ | $L^*=46.70$ | $L^*=60.48$ | $L^*=27.55$ |
| $h^*=68.98$ | $h^*=70.19$ | $h^*=68.82$ | $h^*=73.71$ | $h^*=57.32$ | $h^*=77.13$ | $h^*=77.55$ | $h^*=61.59$ |

| $L^*=74.19$ | $L^*=72.15$ | $L^*=77.43$ | $L^*=62.40$ | $L^*=71.34$ | $L^*=75.92$ | $L^*=80.60$ | $L^*=61.40$ |
| $h^*=59.07$ | $h^*=55.96$ | $h^*=55.88$ | $h^*=56.84$ | $h^*=42.16$ | $h^*=66.32$ | $h^*=59.47$ | $h^*=51.70$ |

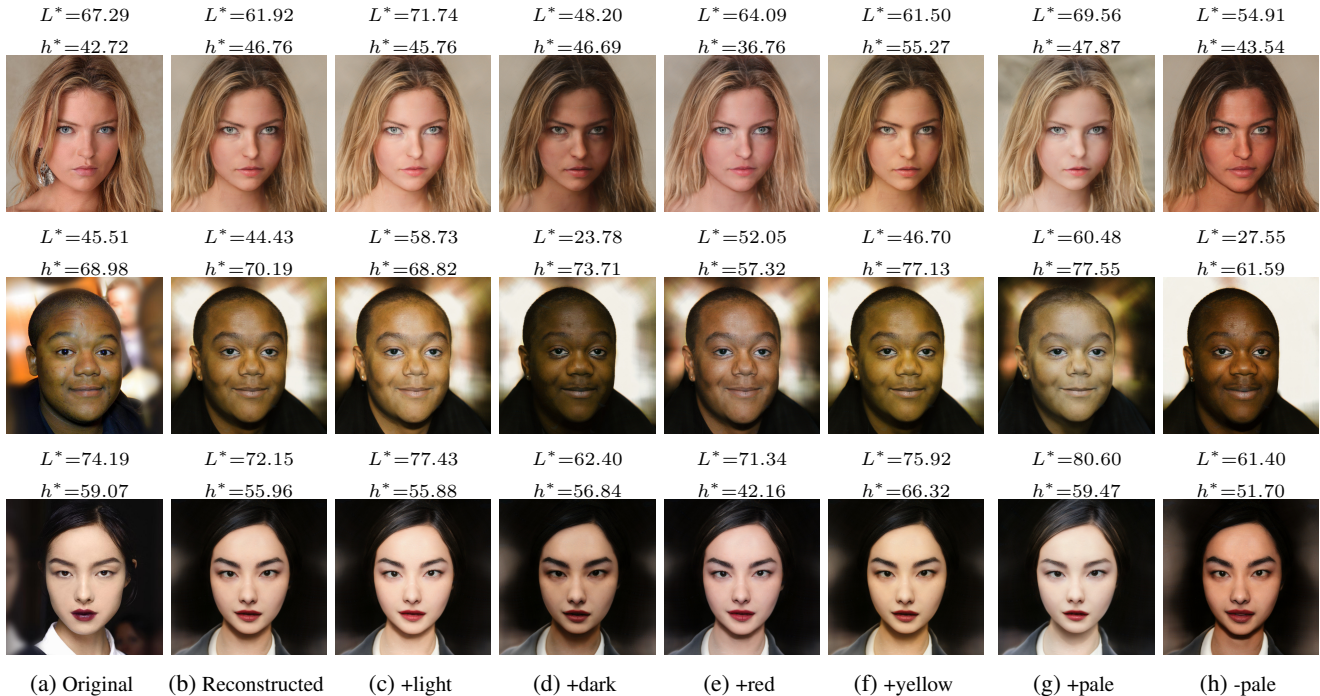| (a) Original | (b) Reconstructed | (c) +light | (d) +dark | (e) +red | (f) +yellow | (g) +pale | (h) -pale |

Figure 4: **Skin color manipulation** on CelebAMask-HQ. By moving along specific directions in the latent space of Style-GAN3 (b), we manipulate the skin tone (c-d) or skin hue (e-f). When modifying the skin tone, this affects the perceptual lightness L^* and preserves the hue angle h^* (and conversely when modifying the skin hue). This differs from the attribute "pale" in CelebA, where a manipulation leads to changes in both L^* and h^* (g-h). As such, our proposal (c-f) is better for measuring the multidimensional causal effect of skin color as we have an independent control over L^* and h^*.

a decreased performance of dark and yellow skin colors. Overall, this benchmark confirms that in well-established methods for face verification, there exist performance differences in both skin tones and skin hues, which reiterates the importance of a multidimensional measure of skin color.

4.3. Skin color causal effect in models

Task. Given a facial image x, the objective is to predict the presence or absence of an attribute a. In the context of this paper, and following previous works [25, 57, 58], we focus on commercial systems, as well as publicly available models, to predict the gender and the presence of a smile.

Empowered with the ability to measure skin color quantitatively, we propose to manipulate an existing dataset by changing its skin tone and skin hue. We are inspired by the idea of transects [4] to reveal causal effects by manipulating one particular property in the image at a time. In the context of this paper, we are interested in manipulating skin color and observing its causal effect on attribute prediction performance. For example, we modify all images in a dataset to have a lighter skin tone and compare its performance with respect to the original dataset version. Any discrepancy would then corresponds to the effect of a bias towards light skin tone in the model.

Method. To build a benchmark for attribute prediction, we propose to manipulate images with an image generation method. Specifically, we consider an encoder-decoder scheme, which encodes an image x into a latent vector w, and then decodes w to provide a reconstructed image $\hat{x}$. Modifying x then consists of moving in the latent space towards specific and meaningful directions D, *i.e.*, $w + D$.

Following Alahuf *et al.* [1], we rely on e4e [66] to encode images and on StyleGAN3 [37] to generate images. Finding directions in the latent space relies on the InterfaceGAN [63] method. Concretely, we first train a classifier on CelebAMask-HQ [43] to predict the skin tone (light *vs*. dark) and skin hue (red *vs*. yellow). We then sample 500,000 images from StyleGAN3 randomly and use the trained classifier to infer skin color labels. Two linear SVMs, one for skin tone and one for skin hue, are finally trained on the most confident positive and negative predictions to produce decision boundaries to be used to compute the latent directions D.

For benchmarking, we consider face recognition models of commercial systems such as AWS Rekognition [3], Microsoft Azure [49] and Clarifai [14]. Additionally, we consider the publicly available model trained on the FairFace dataset [36]. Note that all models have been released af-

Model	Reconstructed			Skin tone						Skin hue					
				+light			+dark			+red			+yellow		
	F	M	*All*	F	M	*All*	F	M	*All*	F	M	*All*	F	M	*All*
AWS	99.67	94.82	97.88	99.84	90.66	96.45	98.84	97.73	98.27	99.69	94.88	97.92	99.63	94.56	97.76
Azure	99.78	95.80	98.31	99.91	92.59	97.22	99.11	97.72	98.60	99.75	96.01	98.37	99.78	95.45	98.18
Clarifai	95.31	85.94	91.85	96.04	80.14	90.18	93.04	92.20	92.73	94.72	86.52	91.70	96.02	84.53	91.79
FairFace	99.49	92.91	97.06	99.77	87.56	95.27	98.21	96.57	97.61	99.52	93.10	96.97	99.42	92.77	97.16

(a) **Gender classification accuracy**, where F denotes female samples, M denotes male samples and *All* denotes all samples.

Model	Reconstructed			Skin tone						Skin hue					
				+light			+dark			+red			+yellow		
	S	NS	*All*	S	NS	*All*	S	NS	*All*	S	NS	*All*	S	NS	*All*
AWS	92.80	93.66	93.26	94.22	91.58	92.82	88.80	96.02	92.63	96.88	86.38	91.32	87.97	96.70	92.60
Azure	97.29	80.84	88.57	98.13	76.36	86.60	95.59	86.11	90.57	99.00	69.76	83.51	95.18	87.22	90.97

(b) **Smile classification accuracy**, where S denotes smiling samples, NS denotes non-smiling samples and *All* denotes all samples.

Table 4: **Attribute prediction** on CelebAMask-HQ, with disaggregated results for positive and negative samples. (a) In gender classification, models classify individuals as more feminine when their skin tone becomes lighter. (b) In smile classification, models classify individuals as smiling when their skin tone becomes lighter or when their skin hue becomes redder.

ter the seminal Gender Shades paper [8], which highlighted bias issues in commercial systems, and we use their latest available version for attribute prediction. As a result, biases we observe in this paper might be reduced compared to the Gender Shades paper.

We focus on gender and smile classifications, as both attributes can be predicted with the above systems and come with ground truth labels annotated by a professional company in the CelebA dataset [44]. We follow previous literature initiated by the Gender Shades paper [8] and follow-up works [35, 57, 58], which perform gender and smile classification to highlight ethical concerns. Such fairness benchmarking promotes model transparency, which in turn creates accountability that could lead to the discontinuation of harmful models [58]. As such, we stress that we do not condone the gender classification task, as it causes harm to non-binary and transgendered individuals by reducing gender to a binary value [39], but rather report it to examine additional ethical issues around this commonly used task in the fairness literature.

We consider CelebAMask-HQ [43] for benchmarking. Modifications of the images involve making the skin tone lighter or darker, as well as making the skin hue more red or yellow. Figure 4 depicts qualitative samples of the reconstructed image with e4e [66] and StyleGAN3 [37], as well as manipulations in the latent space with Interface-GAN [63]. We observe that when skin tone changes, the hue angle stays the same (4c-4d); and conversely when manipulating the hue angle (4e-4f). Modifying the image in one direction does not affect the skin color score in the other direction, making them orthogonal. The same does not hold when selecting the "pale" attribute available in CelebA [44] as a comparison. While the *pale* direction is effective at manipulating the skin tone, it actually also alters the skin hue

(4g-4h). In other words, both L^* and h^* shift when modifying the *pale* attribute in the images, which makes such metric impractical for measuring the causal effect of skin color as we cannot control its effect. Thus for benchmarking, we modify all images in the CelebAMask-HQ datasets by increasing and decreasing directions of L^* and h^*, for a total of $4 \times 30,000$ images, and report the binary accuracy of evaluated models on gender and smile classifications.

A current limitation of this setup lies in the image reconstruction and manipulation methods, which struggle with spurious correlations. This results in attributes not being preserved, which can be depicted by an altered background or eye gaze (see original *vs.* reconstructed in Figure 4a–4b). As such, causal experiments in this section are dependent on the quality of synthetic data, akin to related previous works (*e.g.*, [4, 59]). To avoid capturing performance discrepancies due to imperfect reconstruction, we compare the performance of manipulated images with the *reconstructed* version rather than the *original* one in Table 4.

Results. Table 4 presents the attribute prediction performance of several methods on CelebAMask-HQ. When measuring the performance on gender classification in Table 4a, we observe that manipulating the skin color to have a lighter skin tone decreases the gender classification accuracy while the skin hue does not seem to have a large effect. The skin tone bias occurs because models are prone to classify people as feminine when the tone is lighter. For example, in the gender predictions of AWS for male samples, the accuracy drops from 94.82% to 90.66% while it stays relatively the same for female samples. Interestingly, manipulating the skin to be redder or darker in male sample results in a increase in accuracy for all models. When measuring the performance on smile classification in Table 4b, we observe that manipulating the skin color to have a lighter skin tone

or a redder skin hue decreases the accuracy in non-smiling individuals as they tend to be predicted as smiling. For example, the accuracy for non-smiling individuals with Azure drops from 80.84% to 69.76% when the skin hue becomes redder while it stays the same for smiling individuals. Conversely, a darker skin tone or a yellower skin hue decreases the accuracy in smiling individuals. Overall, this benchmark reveals a bias towards a light skin tone when predicting if the individual belongs to the female gender, and a bias towards light or red skin hue when predicting the presence of a smile, which illustrates the importance of a multidimensional measure of skin color.

5. Conclusion

Measuring apparent skin color requires a multidimensional score to capture its variation and provide a comprehensive representation of its constitutive complexity. In this paper, we first focus on the perceptual lightness L^*, as a measure of skin tone ranging from light to dark, and the hue angle h^*, as a measure of skin hue ranging from red to yellow. Our proposal serves as a simple, yet effective, first step towards a multidimensional skin color score. Second, we reveal biases related to skin color in image datasets and computer vision models, previously invisible. Our multidimensional skin color scale offers a more representative assessment to surface socially relevant biases due to skin color effects in computer vision. While the paper considers face-related tasks, assessing skin color could also be applied to other human-centric tasks (*e.g.*, pose estimation, segmentation, etc). This would help to (i) enhance the diversity in the data collection process, by encouraging specifications to better represent skin color variability; and (ii) improve the identification of dataset and model biases in fairness benchmarking, by highlighting their limitations and leading to fairness-aware training methods. Therefore, we recommend the usage of a multidimensional skin color measure as a fairness tool to assess the computer vision pipeline, from data collection to model deployment.

References

[1] Yuval Alaluf, Or Patashnik, Zongze Wu, Asif Zamir, Eli Shechtman, Dani Lischinski, and Daniel Cohen-Or. Third time's the charm? image and video editing with stylegan3. *arXiv:2201.13433*, 2022.

[2] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *FAccT*, 2021.

[3] AWS Rekognition. Detecting and analyzing faces. https://docs.aws.amazon.com/rekognition/latest/dg/faces.html.

[4] Guha Balakrishnan, Yuanjun Xiong, Wei Xia, and Pietro Perona. Towards causal benchmarking of bias in face analysis algorithms. In *ECCV*. 2020.

[5] Brinnae Bent, Benjamin A Goldstein, Warren A Kibbe, and Jessilyn P Dunn. Investigating sources of inaccuracy in wearable optical heart rate sensors. *NPJ digital medicine*, 3(1):1–9, 2020.

[6] Sandra Del Bino and Françoise Bernerd. Variations in skin colour and the biological consequences of ultraviolet radiation exposure. *British Journal of Dermatology*, 169, 2013.

[7] Abeba Birhane, Vinay Uday Prabhu, and John Whaley. Auditing saliency cropping algorithms. In *WACV*, 2022.

[8] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAccT*, 2018.

[9] Mary E. Campbell, Verna M. Keith, Vanessa Gonlin, and Adrienne R. Carter-Sowell. Is a picture worth a thousand words? an experiment comparing observer-based skin tone measures. *Race and Social Problems*, 12:266–278, 2020.

[10] Alain Chardon, Isabelle Cretois, and Colette Hourseau. Skin colour typology and suntanning pathways. *International Journal of Cosmetic Science*, 13, 1991.

[11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[12] YB Choe, SJ Jang, SJ Jo, KJ Ahn, and Jai Il Youn. The difference between the constitutive and facultative skin color does not reflect skin phototype in asian skin. *Skin Research and Technology*, 12(1):68–72, 2006.

[13] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *CVPR*, 2022.

[14] Clarifai. Ai face detection model. https://www.clarifai.com/models/ai-face-detection.

[15] Roxana Daneshjou, Kailas Vodrahalli, Roberto A Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, et al. Disparities in dermatology ai performance on a diverse, curated clinical image set. *arXiv:2203.08807*, 2022.

[16] Commission Internationale de l'Eclairage. *Colorimetry*. CIE technical report. 2004.

[17] Jean de Rigal, Marie-Laurence Abella, Franck Giron, Laurence Caisey, and Marc André Lefebvre. Development and validation of a new skin color chart®. *Skin Research and Technology*, 13(1):101–109, 2007.

[18] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

[19] Yubin Deng, Chen Change Loy, and Xiaoou Tang. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine*, 34(4):80–106, 2017.

[20] Steven Eilers, Daniel Q Bach, Rikki Gaber, Hanz Blatt, Yanina Guevara, Katie Nitsche, Roopal V Kundu, and June K Robinson. Accuracy of self-report in assessing fitzpatrick skin phototypes i through vi. *JAMA dermatology*, 149(11):1289–1294, 2013.

[21] Arpad Elo. *The Rating of Chessplayers, Past and Present*. Arco Publishing, 1978.

[22] European Commission. Guidance note on the collection and use of equality data based on racial or ethnic origin, 2021.

[23] Haiwen Feng, Timo Bolkart, Joachim Tesch, Michael J Black, and Victoria Abrevaya. Towards racially unbiased skin tone estimation via scene disambiguation. In *ECCV*, 2022.

[24] Thomas B. Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology*, 124 6:869–71, 1988.

[25] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[26] Theo Gevers, Arjan Gijsenij, Joost Van de Weijer, and Jan-Mark Geusebroek. *Color in computer vision: fundamentals and applications*. John Wiley & Sons, 2012.

[27] Eric Raphael Gottlieb, Jennifer Ziegler, Katharine Morley, Barret Rush, and Leo Anthony Celi. Assessment of racial and ethnic differences in oxygen supplementation among patients in the intensive care unit. *JAMA Internal Medicine*, 182(8):849–858, 2022.

[28] Matthew Groh, Caleb Harris, Roxana Daneshjou, Omar Badri, and Arash Koochek. Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm. *arXiv:2207.07942*, 2022.

[29] Patrick Grother, Mei Ngan, and Kayee Hanaoka. *Face recognition vendor test (fvrt): Part 3, demographic effects*. National Institute of Standards and Technology Gaithersburg, MD, 2019.

[30] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *FAccT*, 2020.

[31] Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton Ferrer. Towards measuring fairness in ai: the casual conversations dataset. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2021.

[32] Steven Y He, Charles E McCulloch, W John Boscardin, Mary-Margaret Chren, Eleni Linos, and Sarah T Arron. Self-reported pigmentary phenotypes and race are significant but incomplete predictors of fitzpatrick skin phototype in an ethnically diverse population. *Journal of the American Academy of Dermatology*, 71(4):731–737, 2014.

[33] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[34] Nina G Jablonski. The evolution of human skin and skin color. *Annual Reviews Anthropology*, 33:585–623, 2004.

[35] Siddharth Jaiswal, Karthikeya Duggirala, Abhisek Dash, and Animesh Mukherjee. Two-face: Adversarial audit of commercial face recognition systems. In *AAAI Conference on Web and Social Media*, 2022.

[36] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *WACV*, 2021.

[37] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021.

[38] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

[39] Os Keyes. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–22, 2018.

[40] Davis E. King. Dlib-ml: A machine learning toolkit. *JMLR*, 10:1755–1758, 2009.

[41] Newton M Kinyanjui, Timothy Odonga, Celia Cintas, Noel CF Codella, Rameswar Panda, Prasanna Sattigeri, and Kush R Varshney. Fairness of classifiers across skin tones in dermatology. In *MICCAI*, 2020.

[42] Anjana Lakshmi, Bernd Wittenbrink, Joshua Correll, and Debbie S Ma. The india face set: International and cultural boundaries impact face impressions and perceptions of category membership. *Frontiers in psychology*, 12:161, 2021.

[43] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020.

[44] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[45] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[46] Bao Chau K Ly, Ethan B Dyer, Jessica L Feig, Anna L Chien, and Sandra Del Bino. Research techniques made simple: cutaneous colorimetry: a reliable technique for objective skin color measurement. *Journal of Investigative Dermatology*, 140(1):3–12, 2020.

[47] Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. The chicago face database: A free stimulus set of faces and norming data. *Behavior research methods*, 47(4):1122–1135, 2015.

[48] Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. Diversity in faces. *arXiv preprint arXiv:1901.10436*, 2019.

[49] Microsoft Azure. Face api. https://azure.microsoft.com/en-us/services/cognitive-services/face/.

[50] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *FAccT*, 2019.

[51] Roy Or-El, Soumyadip Sengupta, Ohad Fried, Eli Shechtman, and Ira Kemelmacher-Shlizerman. Lifespan age transformation synthesis. In *ECCV*, 2020.

[52] Sungho Park, Jewook Lee, Pilhyeon Lee, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. Fair contrastive learning for facial attribute classification. In *CVPR*, 2022.

[53] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu

Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *JMLR*, 12:2825–2830, 2011.

[54] P Jonathon Phillips, Fang Jiang, Abhijit Narvekar, Julianne Ayyad, and Alice J O'Toole. An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception (TAP)*, 8(2):1–11, 2011.

[55] Latrice C Pichon, Hope Landrine, Irma Corral, Yongping Hao, Joni A Mayer, and Katherine D Hoerster. Measuring skin cancer risk in african americans: is the fitzpatrick skin type classification scale culturally sensitive? *Ethnicity & disease*, 20(2):174–179, 2010.

[56] 3dtotal Publishing. *Artists' Master Series: Color and Light*. Artists' Master Series. 3DTotal Publishing, 2022.

[57] Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *AIES*, 2019.

[58] 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 *AIES*, 2020.

[59] Vikram V Ramaswamy, Sunnie SY Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. In *CVPR*, 2021.

[60] Silonie Sachdeva. Fitzpatrick skin typing: Applications in dermatology. *Indian journal of dermatology, venereology and leprology*, 75(1):93, 2009.

[61] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[62] Anna Shcherbina, C Mikael Mattsson, Daryl Waggott, Heidi Salisbury, Jeffrey W Christle, Trevor Hastie, Matthew T Wheeler, and Euan A Ashley. Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. *Journal of personalized medicine*, 7(2):3, 2017.

[63] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020.

[64] Herbert A Sturges. The choice of a class interval. *Journal of the american statistical association*, 21(153):65–66, 1926.

[65] William Thong and Cees GM Snoek. Feature and label embedding spaces matter in addressing image classifier bias. In *BMVC*, 2021.

[66] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021.

[67] Norimichi Tsumura, Nobutoshi Ojima, Kayoko Sato, Mitsuhiro Shiraishi, Hideto Shimizu, Hirohide Nabeshima, Syuuichi Akazaki, Kimihiko Hori, and Yoichi Miyake. Image-based skin color and texture analysis/synthesis by extracting hemoglobin and melanin information in the skin. In *ACM SIGGRAPH*. 2003.

[68] MN Venkataram and Al Amiri Haitham. Correlating skin phototype and minimum erythema dose in arab skin. *International journal of dermatology*, 42(3):191–192, 2003.

[69] 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 *ICCV*, 2019.

[70] Wenguan Wang, Jianbing Shen, and Haibin Ling. A deep network solution for attention and aesthetics aware photo cropping. *TPAMI*, 41(7):1531–1544, 2018.

[71] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *CVPR*, 2020.

[72] Olivia R Ware, Jessica E Dawson, Michi M Shinohara, and Susan C Taylor. Racial limitations of fitzpatrick skin type. *Cutis*, 105(2):77–80, 2020.

[73] Ian L Weatherall and Bernard D Coombs. Skin color measurements in terms of cielab color space values. *Journal of investigative dermatology*, 99(4):468–473, 1992.

[74] Marcus Wilkes, Caradee Y Wright, Johan L du Plessis, and Anthony Reeder. Fitzpatrick skin type, individual typology angle, and melanin index in an african population: steps toward universally applicable skin photosensitivity assessments. *JAMA dermatology*, 151(8):902–903, 2015.

[75] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. Predictive inequity in object detection. *arXiv:1902.11097*, 2019.

[76] Jianzhou Yan, Stephen Lin, Sing Bing Kang, and Xiaoou Tang. Learning the change for automatic image cropping. In *CVPR*, 2013.

[77] Kyra Yee, Uthaipon Tantipongpipat, and Shubhanshu Mishra. Image cropping on twitter: Fairness metrics, their limitations, and the importance of representation, design, and agency. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–24, 2021.

[78] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *ICCV*, 2021.

[79] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W. Jacobs. Deep single portrait image relighting. In *ICCV*, 2019.