Beyond Skin Tone: A Multidimensional Measure of Apparent Skin Color

William Thong
Sony AI

Przemyslaw Joniak
The University of Tokyo

Alice Xiang
Sony AI

A. Supplementary Material

We provide additional materials to supplement our main paper. Section A.1 describes challenges in estimating skin color in in-the-wild images. Section A.2 details our method to extract skin color scores in images and performs a robustness analysis with different illuminations. Section A.3 explores the skin color distribution per ethnicity while Section A.4 characterizes skin color bias in datasets beyond binary thresholding.

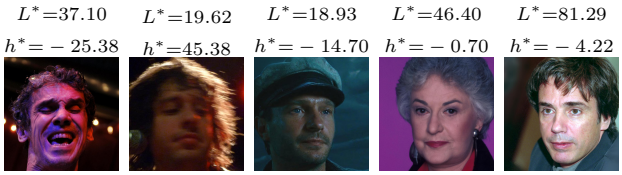
A.1. Factors influencing skin color in the wild

Compared with images acquired in a controlled setting, images in the wild have a wide range of variation. External factors can influence the visual output, which would affect in turn the measure of the "true" skin color of the individual. Characterizing the effect of external factors remains an open challenge. As such, similar to previous works, we focus in this paper on the "apparent" skin color observed in the image for our fairness evaluation.

Figure S1 highlights some of the external factors that affect the skin color in the CelebAMask-HQ dataset. We observe that the "apparent" skin color can be affected by external factors in the scene environment or the camera setting (*e.g.*, color cast, intensity and orientation of the illumination, low-light environment, etc), but also by structural factors of the subjects (*e.g.*, having makeup or face flushing). As a result, the skin color in these images can, for example, appear to be much darker or much redder than their "true" color. The effect of these external and structural factors make the measure of skin color in in-the-wild images an open challenging problem. Still, we consider the "apparent" skin color as this what computer vision models are seeing.

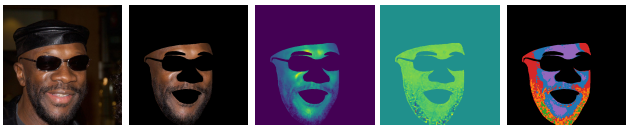
A.2. Extracting skin color from skin pixels

Skin color scores provide a quantitative measure to characterize the appearance of the skin in an image. Extracting such measures helps to identify potential biases towards skin color subgroups in model performance. Our objective differs from the cosmetics or dermatology fields, which requires an accurate assessment of constitutive skin color from cutaneous measurements [46]. We focus, instead, on



(a) Color cast (b) Illumination (c) Low-light (d) Makeup (e) Face flushing

Figure S1: **Factors influencing skin color measurement** in CelebAMask-HQ. External factors coming from scene affect the skin color (a-c), as well as structural ones from the individuals (d-e).



(a) Original image (b) Skin pixels (c) L^* and h^* (d) Clusters

Figure S2: **Extracting skin color scores** method overview. Given an input image, we isolate the skin pixels and compute a per-pixel skin color score measurement. Pixels are then clustered together and we perform a weighted average of skin color scores of each cluster to get the f

the "apparent" skin color in images acquired from any camera, with varying acquisition parameters or lighting conditions. The main challenge resides in extracting skin color scores from skin pixels in an image. We propose a framework that starts from a facial image x and outputs a final scalar scoring value y or a set of scalar scoring values $y = \{y^1, \cdots, y^Y\}$.

Method. To extract skin color scores from a facial image, we are inspired by 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 variations. Figure S2 presents the steps to extract a skin color score in human-centric images:

(a) We are given an input image of a subject.

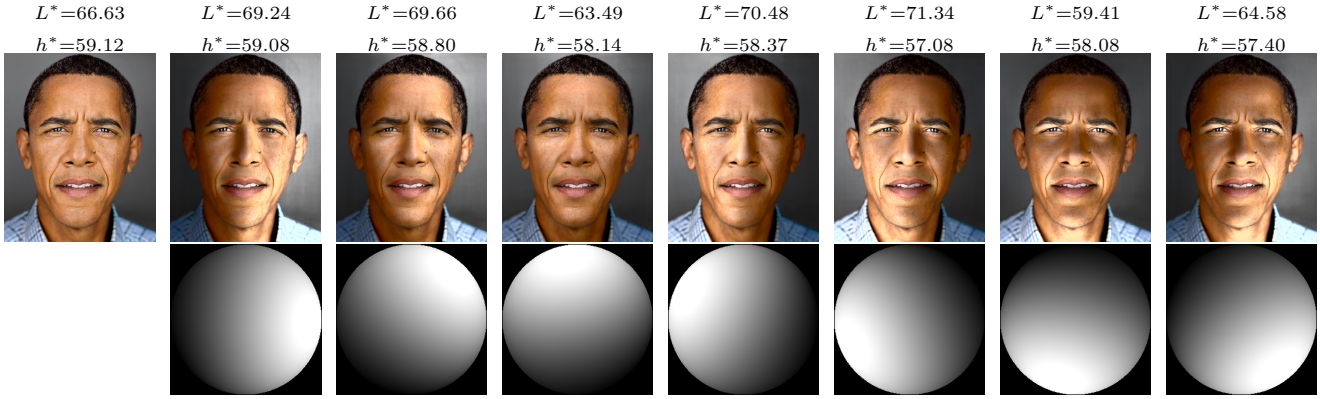


Figure S3: **Robustness of skin color scores** to face relighting. When changing the direction and intensity of the illumination with [79], the perceptual lightness L^* varies by up to 6 points w.r.t. the original image. while the hue angle h^* remains stable. Given that we strive to measure the "apparent" skin color, as seen by a computer vision model, rather than the "true" skin color of individuals, changes in perceptual lightness are expected.

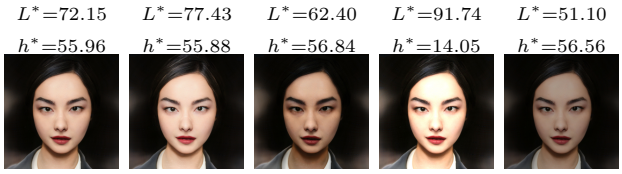
(b) As we are interested in skin color, we start by segmenting the skin pixels. Segmentation can be done manually by an annotator or predicted by a skin segmentation model.

(c) Once skin pixels have been identified, they are converted from the standard RGB space to the target space of the desired scoring values. We convert to the CIELAB space to extract the L^* component, and further use the a^* and b^* components to compute the hue angle h^*. This results in a point measurement of L^* and h^* for every skin pixel in the image.

(d) We then apply a clustering algorithm, such as K-Means [45], to group the skin pixels. For every cluster, we compute a histogram of distribution and set the number of bins with the Sturges formula [64]. The mode of the histogram is then used to assign a scalar value for each considered skin color score. In our case, this results in L^* and h^* scalar values for every cluster.

To obtain the final scalar scores for the image, we average the scalar values of every group normalized by their pixel size. However, as some parts of the face can skew the results towards darker values (*e.g.*, facial hair or shaded regions), we prefer to exclude some groups which yield a very low L^*. We cluster the face skin into five groups and keep the top-3 groups with the highest L^* to compute the final L^* and h^* scalar scores for the image. Such approach is inspired by how human artists would perform value grouping in five different groups when simplifying an image [56].

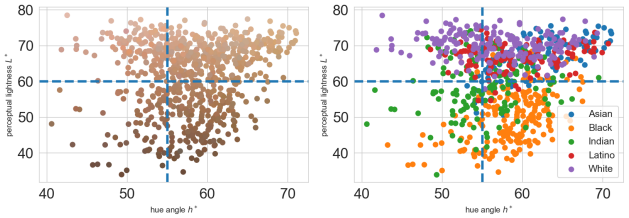
Similar to Merler *et al*. [48], we start from an image of a subject and require a segmentation of facial parts to obtain a skin mask. Such segmentation can be obtained via manual labeling or automatically via a model for skin segmentation. The difference lies mainly in steps (c) and (d). In step (c), we consider any skin color score and not only the individual typology angle. In step (d), we remove the need for facial landmarks by relying on skin pixel clustering. This better deals with atypical facial poses, as clustering can handle faces that are unaligned or from the side (*i.e.*, without visible landmarks). Moreover, clustering can identify shaded areas of the face or facial hair, which we remove to avoid contaminating the final skin color scores.



(a) Reconstructed (b) +light (c) +dark (d) +contrast (e) -contrast

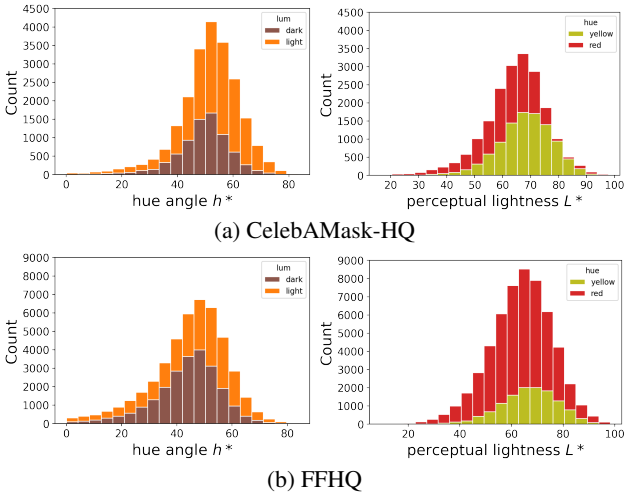
Figure S4: **Skin color *vs*. contrast manipulation** on CelebAMask-HQ. When modifying the image contrast (d-e), faces appear less realistic and the hue angle can change.

Robustness. To gain insights about the robustness of the proposed method, we extract skin color scores for a series of images in which the illumination changes in terms of direction and intensity for a given individual. To achieve this, we rely on samples coming from the work of Zhou *et al*. [79] where a deep neural network produces different face relighting images depending on the target lighting. Figure S3 reports for the perceptual lightness L^* and hue angle h^* for every sample. Face relighting affects L^*, which can differ from up to 6 points with respect to the original image. Interestingly, h^* remains stable and robust to face relighting, which confirms that it provides complementary and orthogonal information about skin color than the skin tone. Differences in L^* are expected as we measure the "apparent"



(a) Skin color representation (b) Ethnicity representation

Figure S5: **Skin color distribution** on CFD dataset w.r.t. perceptual lightness and hue angle. 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° (yellow *vs*. red).



(a) CelebAMask-HQ

(b) FFHQ

Figure S6: **Skin color distribution** on common face datasets. Histograms show a dominance of light skin tone and red skin hue in both CelebAMask-HQ and FFHQ.

skin color in images, which is affected by the illumination, rather than the "true" skin color of individuals.

Contrast. To gain additional insights on the relevance of the proposed method, we measure the root mean square contrast on the whole image using the perceptual lightness channel. To achieve this, we consider the manipulated images in Section 4.3. Reconstructed images have an average contrast of 0.644 while images manipulated to have a darker skin tone are at 0.630 and the ones to have a lighter skin tone at 0.652. Manipulating images has an effect on the overall contrast. Furthermore, we compare our skin color manipulation with a contrast manipulation by scaling the pixel values. Figure S4 shows that image contrast should not be conflated with the skin tone. Increasing or decreasing the image contrast does not result in the same visual modifications, as manipulated faces appear less realistic and the hue angle is not preserved.

A.3. Skin color distribution

Figure S5 depicts the distribution of the CFD dataset in terms of perceptual lightness and hue angle. Contrary to the distribution of the CelebAMask-HQ or FFHQ-Ageing datasets, the CFD dataset depicts less variation (Figure S5a *vs*. Figure 2). This is explained by the fact that images in CFD have been captured in a controlled setting, enabling fair comparisons among images.

Another interesting aspect of CFD is the available self-reported ethnic labels. When breaking down the distribution with the ethnic labels (Figure S5b), we observe some trends in the skin color for the subjects included in the dataset. When comparing White and Black skin color scores, the skin tone –expressed through the perceptual lightness L^*– is sufficient to distinguish both groups. This explains why the fairness literature (*e.g*., [8, 57]) has focused on skin tone to characterize skin color in images, as it serves as a proxy for White and Black skins. Nevertheless, when subjects from other ethnicities are included, boundaries become fuzzy and the skin tone is no longer enough to capture the variability in skin colors. Towards this goal, it is relevant to consider the hue angle h^* to assess the skin hue and reveal other skin color differences. For example, Indian and Black subjects have a darker skins than the other considered ethnicities, with Black subjects having a darker and more yellow skin color than Indian subjects. Another difference lies in Asian subjects, which are mostly in the yellow side of skin hue, as opposed to Latino and White subjects, which have the same skin tone but appear to be more spread in terms of skin hue.

In the scenario where a data collection process would only measure the skin tone for diversity, collecting data from White and Black subjects would be enough to cover the whole spectrum. This is an issue because this would ignore other types of skin color coming from Indian or Asian skin colors for example. Prior work has notably shown that computer vision systems produced in the West often exhibit lower performance for Asian individuals [54]. Including the hue angle, as proposed in the paper, would avoid such an effect where subgroups could be conflated with others despite different skin color characteristics because it gets collapsed into a single "light *vs*. dark" dimension. Overall, adding the hue offers a complementary perspective to assess skin color beyond the tone and reveal previously invisible biases.

A.4. Additional results on skin color bias in common face datasets

Figure S6 offers an alternative representation to highlight the skewed skin color distribution in common face datasets. Instead of a binary thresholding for both perceptual lightness L^* and hue angle h^*, we plot histograms of both scores with 20 bins. In both CelebAMask-HQ and FFHQ, distribu-

tions are unimodal with a bell curve shape. Individuals with a light skin tone and a red skin hue are over-represented with a much larger count. When considering the skin tone and varying the hue angle thresholding, we observe that the hue angle has a lower spread for dark skin tones than light skin tones. Conversely, when considering the skin hue and varying the perceptual lightness thresholding, the yellow skin hue tends to have a larger skewness towards light skin tones than the red skin hue. These representations confirm the relevance of a multidimensional measure for skin color, which could help increase the diversity when collecting a human-centric dataset.