Our Deep CNN Face Matchers Have Developed Achromatopsia

Aman Bhatta[1] Domingo Mery[2*] Haiyu Wu[1] Joyce Annan[3] Michael C. King[3]
Kevin W. Bowyer[1†]

[1]University of Notre Dame
[2]Pontificia Universidad Católica de Chile
[3]Florida Insitute of Technology

{abhatta,hwu6,kwb}@nd.edu, domingo.mery@uc.cl, michaelking@fit.edu, jannan2021@my.fit.edu

Abstract

Modern deep CNN face matchers are trained on datasets containing "color" images. We show that such matchers achieve essentially the same accuracy on color images when trained using only grayscale images. We then consider possible causes for deep CNN face matchers "not using color". Popular web-scraped face datasets actually have 30 to 60% of their identities with one or more grayscale images. We analyze whether this grayscale element in the training set impacts the accuracy achieved, and conclude that it does not. Comparable accuracy for **color test images** *using only* **grayscale images** *implies that the inclusion of "color" may not necessarily add any significant information to the recognition of individuals. This also implies the use of computing resources can be optimized to make the training process more efficient using only grayscale images. Utilizing grayscale images for training reduces the memory footprint of the training data, thereby decreasing system processing time during training. Additionally, our findings emphasize that the adoption of grayscale images not only makes face recognition training more efficient but also offers the opportunity to include more training data, which could result in more accurate face recognition models.*

1. Introduction

Achromatopsia is a condition characterized by a partial or total absence of color vision. People with complete achromatopsia cannot perceive any colors; they see only black, white, and shades of gray. [1]

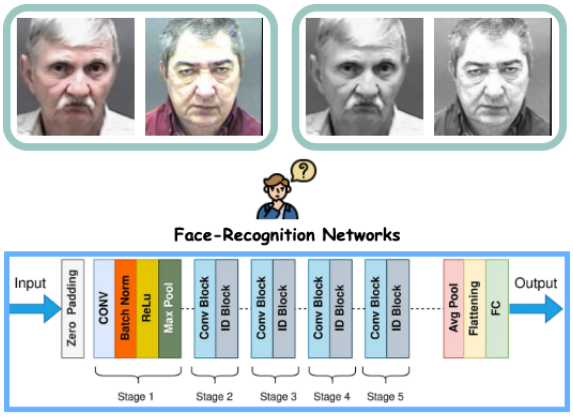


Figure 1. **Do Deep CNN Face Matchers "Use" Color?** Deep CNN face matchers are typically trained and tested on RGB color images (above left). We show that the networks achieve equivalent overall accuracy processing grayscale images (above right), taking 1/3 of the memory and reducing the complexity of the early layer of the network.

Deep convolutional neural networks (CNNs) have powered impressive accuracy gains in many areas of computer vision. Some of the widely used face matchers [17,31,36,45,57,60] use different loss functions in training a ResNet [28] deep CNN backbone. The size of the training set, in terms of the number of identities and the number of images, is a crucial factor in determining the accuracy of a deep CNN face matcher. Web-scraped, in-the-wild face datasets were popularized by Labelled Faces in the Wild (LFW) [29]. Numerous web-scraped, in-the-wild face datasets have been introduced since LFW, increasing in size seemingly every year. While the MS1Mv2 dataset [27] continues to be a widely-used training set, more accurate versions of matchers can be trained using newer, larger datasets

*Dr. Mery thanks National Center for Artificial Intelligence CENIA FB210017, Basal ANID, Chile

†Dr. Bowyer is a member of the FaceTec (facetec.com) Advisory Board. Results in this paper do not necessarily relate to FaceTec products.

such as Glint-360k [7] and WebFace [71]. All of these web-scraped, in-the-wild training sets contain images in RGB color format.

There is no question that color is essential for some general computer vision tasks [16, 22, 55]. *But do current deep CNN face matchers actually use color to achieve better accuracy than they could with grayscale?* The answer is important because training deep CNNs is notoriously memory- and compute-intensive. Color images require 3 times the memory of native grayscale, and 3 times the weights in the early layer of the CNN. By using native grayscale instead of color, it's possible to train the same deep CNN on a larger quantity of images while using only one-third of the weights in the initial layer.

The remainder of this paper is organized as follows. Section 2 gives a brief literature review. Sections 3 and 5 together analyze whether there is any accuracy difference between using grayscale or color for the training data or the test data; all results show that color does not give consistent accuracy gain over grayscale. Section 4 details the network implementation. Section 6 shows that using grayscale could improve the training effiency and saved disk-space allows opportunity to use additional data to improve models. Section 7 presents qualitative analysis on model performance with RGB and grayscale training sets. Finally, Section 8 summarizes and discusses the results.

2. Literature Review

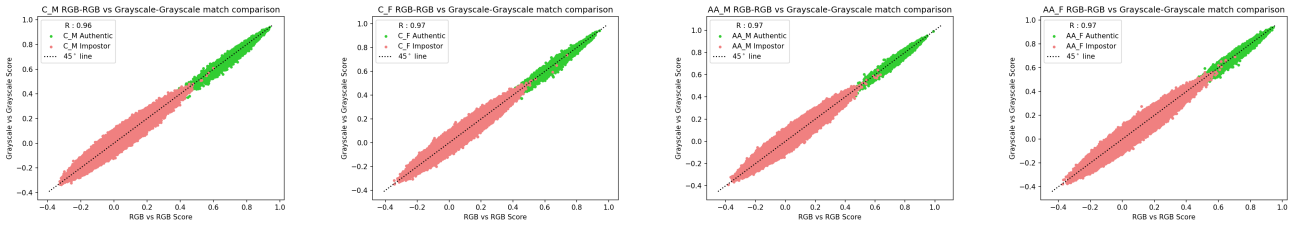
Impact of color on CNNs for general object classification/detection. Researchers have investigated how factors such as noise, blur, jitter, compression, and others affect accuracy of general object classification by deep networks [19, 20, 24, 49, 52, 70]. However, the impact of color has received comparatively less attention for general object classification. One early study by Engilberge et al. [22] analyzed the learned network to detect and characterize color-related features. They found color-specific units in CNNs and demonstrated that the depth of the layers affects color sensitivity. Buhrmester et al. [13] investigated the impact of several color augmentation techniques on the deeper layers of the network and found that luminance is the most robust against changes in color systems. This finding suggests that the intensity value in color images contains the most useful content. De et al. [16] showed that color information has significant impact on the inference of deep neural networks. Singh et al. [55] showed that CNNs often rely heavily on color information, but that this varies between datasets. Several researchers have found one color space better than another for general object classification [13, 18, 25, 53, 64]. Additionally, in their study, Buhrmester et al. [13] investigated the effects of using RGB on a model trained with grayscale data. Their experiments revealed minimal impact on accuracy. They speculated that essential visual cues such as edges and brightness are effectively learned and utilized by the model for object recognition tasks. For a short review of the impact of the color space on classification accuracy, see Velastegui et al. [59]

Impact of color on human ability in face perception. The role of color in face perception by humans has been studied in Psychology. Early work by Bruce et al. [12] and Kemp et al. [35] largely dismissed the impact of color on face recognition. However, more recent work by Sinha et al. [56, 65] demonstrated the influence of color on human ability in face detection and recognition. Brosseau et al. [11] reported that color-blind individuals performed significantly poorer on face recognition tasks, underscoring the importance of color. Researchers have also explored the effects of both face and background color on the perception of facial expressions [46]. Bindemann et al. [10] examined face detection performance in the absence of color and found that performance declines when color information is removed from faces, regardless of whether the surrounding scene context is rendered in color.

Impact of color on deep CNN face perception. Researchers have extensively examined the effects of quality factors, including blur, noise, occlusion, distortion, and more, on the accuracy of deep CNN face recognition [26, 34, 43]. However, the influence of color has received relatively little attention. To our knowledge, Grm et al. [26] is the only previous work on this specific topic. They evaluated several pre-ArcFace matchers that were trained on RGB images, with RGB and grayscale versions of test images. They found that the matchers trained on RGB achieved similar accuracy on grayscale test images as on RGB test images. Our study differs from [26] in four significant ways. One, we utilize deeper networks and employ more advanced loss functions that are considered SOTA for face recognition. Two, we train the entire network using grayscale images and evaluate its performance on RGB images, aiming to assess the significance of color information for modern face recognition networks. Three, we examine the degree to which the network learns color-oriented features in the layer that analyzes the RGB input. Four, we examine the degree to which the RGB of the skin region of an individual varies across their training images.

3. Accuracy on Grayscale vs. RGB Test Sets

This section evaluates one of the widely used pre-trained face matchers to determine if it achieves better accuracy for RGB vs grayscale images. We present the results for combined-margin model based on ArcFace loss [17], trained on Glint-360K(R100) [7] with weights from [3]. The network processes 112×112 aligned faces to create 512-d feature vectors matched using cosine similarity. Test images are sourced from the MORPH dataset [4, 51], the same version as in [5]. MORPH includes images of Cau-



(a) Caucasian Male (b) Caucasian Female (c) African-American Male (d) African-American Female

Figure 2. **Model Trained with RGB Images Exhibits Similar Performance When Applied To Grayscale Images From Diverse Demographics**. This suggests that using grayscale images do not disproportionately influence any specific demographic group. Each image pair presented in the plot has similarity score for original RGB version and grayscale version. For each demographic, throughout the range of similarity, the cloud of points trends on the 45-degree line. If grayscale gave consistently lower similarity score, the cloud should trend below the 45-degree line. ResNet backbone, ArcFace loss, glint training set, MORPH test images.

casian and African-American males and females, with identities distinct from the web-scraped celebrity training sets.

Figure 2 presents a test of whether training on web-scraped color images has any consistent accuracy difference for grayscale versus RGB versions of test images. The test sets in this experiment are the four main demographic groups in MORPH. RGB-RGB image pairs are formed from the original MORPH color images. Corresponding gray-gray image pairs are formed by using OpenCV to create grayscale versions of the RGB images. To ensure compatibility with the same pre-trained model used for both grayscale and color images, grayscale images are loaded in a three-channel format where the values in all color planes are identical, i.e., R=G=B.

Each pair of images produces a point in the scatter plot in Figure 2, where the cosine similarity for the RGB version of the image pair is the horizontal axis and the similarity of the grayscale version of the same image pair is the vertical axis. If accuracy is generally higher when matching RGB pairs, the cloud of points would trend below the 45-degree line. If the accuracy is generally higher for grayscale, the points would trend above the 45-degree line. The actual result is that the cloud clusters along the 45-degree line. This indicates no consistent accuracy difference when matching grayscale versus color image pairs. This is true for each of the four demographic groups. Figure 2 illustrates that, on average, the Pearson product-moment correlation coefficient (R) between similarity scores computed for color image pairs and grayscale image pairs is ≈ 0.97. This quantitatively demonstrates that facial similarity remains consistent whether computed from color or grayscale images.

4. Implementation Details

In this section, we provide a concise overview of the configuration employed. To train ArcFace and AdaFace from scratch with color and grayscale images, the configurations remain consistent across different training instances of the same model, and the details are included to ensure that others could reproduce the results if desired. Both loss functions, ArcFace and AdaFace, are trained using ResNet-50 backbone [17, 28]. For ArcFace, we employ a combined margin with margin combination values of (1.0, 0, 0.4). The model is trained for 20 epochs using SGD as the optimizer, with momentum of 0.9 and initial learning rate of 0.1. On the other hand, for AdaFace, we set the initial margin to 0.4. The model is trained for 24 epochs using SGD as the optimizer, with momentum of 0.9 and initial learning rate of 0.1. The learning rate is reduced by a gamma factor of 0.1 at the 12th, 20th, and 24th epochs. All the mentioned configuration parameters align with the ones utilized for training WebFace4M on the ResNet-50 backbone, as mentioned in insightface [3, 7] and AdaFace [2] repositories.

5. Grayscale Images In "Color" Image Sets

The importance of "cleaner" training data, in the sense of accurate identity labels, is widely acknowledged as enabling training more accurate matchers. In this section, we show that (assumed) color training sets actually contain a fraction of grayscale images, and consider whether "cleaning" grayscale images out of an assumed color training set might help to explain the results in the previous section.

MS1MV2 [17] is a cleaned version of MS1-Celeb [27], and is widely used in training face matchers. Glint-360k [7] is newer than MS1MV2, and contains about 4x the number of identities and 3x the number of images. WebFace4M/12M [71] is newer still, and increasingly prominent, and its 4M subset has more than 2x the identities of MS1MV2 but fewer total images. All three of these datasets contain web-scraped, in-the-wild images in RGB format.

We ran a test to detect images that are effectively grayscale even though stored as RGB. These are images with R=G=B across all pixels, so that they are stored in three-channel format but contain no color content. *The presence of three channels in grayscale images allows the convolutional layers designed for RGB inputs to effortlessly handle them.* Results of this test are summarized in Ta-

Model Loss	Training Dataset: Type of WebFace4M	Training Dataset: Colorspace (Channels)	Test dataset: LFW [%]	CFP-FP [%]	AGEDB-30 [%]	CALFW [%]	CPLFW [%]	Mean [%]
ArcFace	Entire	RGB (3)	99.82	99.10	97.95	96.15	94.13	97.43
		Grayscale (3)	**99.82**	98.94	97.77	96.02	94.10	97.33
		Error* →	0.00	0.16	0.18	0.13	0.03	0.10
	Color Cleaned	RGB (3)	99.80	99.04	97.88	96.02	94.18	97.38
		Grayscale (3)	**99.82**	98.98	97.70	96.01	**94.20**	97.34
		Error* →	-0.02	0.06	0.18	0.01	-0.02	0.04
		HSV (3)	99.82	98.97	97.70	96.02	94.20	97.34
AdaFace	Entire	RGB (3)	99.76	98.94	97.63	95.95	94.30	97.32
		Grayscale (3)	**99.76**	98.82	97.51	95.83	93.78	97.14
		Error* →	0.00	0.12	0.12	0.12	0.52	0.17
	Color Cleaned	RGB (3)	99.80	98.81	97.66	95.96	94.30	97.31
		Grayscale (3)	99.68	98.58	97.36	**96.00**	93.88	97.10
		Error* →	0.12	0.23	0.30	-0.04	0.42	0.20
		HSV (3)	99.78	98.60	97.30	96.02	94.02	97.14

* Relative error of the accuracies computed as (RGB - Grayscale)

Table 1. **Training on Grayscale Images and Testing on RGB Images Yields Comparable Performance to RGB Training and Testing**. For example, the top three rows show that training ArcFace on grayscale and testing on RGB results in a drop in mean accuracy across the five datasets of (only) 0.10 compared to training and testing on RGB. *Training on grayscale (3-channel) and testing on RGB test set actually beats training and testing on RGB in a few instances*. Moving from RGB to an alternate color space (HSV) that separates color and brightness does not appreciably change the pattern of results. The results in this table motivate consideration of training and testing on single-channel grayscale in order to reduce disk space needed to store image and enable larger training datasets. [Key: **Gray Better or equal to RGB**]

ble 2. Approximately 6-8% of the images in each of the datasets are effectively grayscale. Even more importantly, 34% (WebFace4M) to 60% (MS1MV2) of the identities in each dataset have at least one grayscale image. This is relevant because the goal of the deep CNN training is to get all images of each identity to classify as that identity. Speculatively, having one or more grayscale images of an identity could lead the deep CNN to learn to ignore color to have unified classification for all identity-related images.

Dataset	Original: Total Identities	Original: Total Images	Grayscale Subset: Identities w/ at least one GrayScale image	Grayscale Subset: Total GrayScale Images
MS1MV2	85.7K	5.8M	51.7K (∼**60%**)	444K (∼**7.6%**)
Glint360k	360K	17.1M	154K (∼**43%**)	919K (∼**6%**)
WebFace4M	205K	4.2M	70K (∼**34%**)	246K (∼**6%**)

Table 2. **Grayscale Image / Identities in Popular Training Sets.** Similar to cleaning identity labels, cleaning is needed to avoid web-scraped RGB datasets having a large fraction of identities with one or more grayscale images.

To investigate how a fraction of the training data being effectively grayscale affects the accuracy of a trained network, we create three additional versions of WebFace4M. From the original WebFace4M, we create a "color cleaned" subset by dropping the images found to have R=G=B. After excluding these effectively grayscale images, we were left with around 205K identities and 3.9M color images. This subset of WebFace4M is the "**color cleaned**" version. We then train a network from scratch using each of (1) the original (94% color / 6% grayscale) WebFace4M, (2) a version of the original WebFace4M with all images converted to grayscale, (3) the color-cleaned (100% color) subset of WebFace4M, and (4) the color-cleaned version of WebFace4M with all images converted to grayscale. Training instances of ArcFace and AdaFace from scratch on these four datasets resulted in a total of eight trained models. The eight trained models fall into four pairs that give direct comparison of accuracy for a model trained on a color dataset and the grayscale version of that color dataset. *Note that all models, both in the grayscale and color training sets, have been trained with a three-channel input, and their accuracy is evaluated using the color version of the benchmark datasets.*

Table 1 summarizes these comparisons of accuracy for training on grayscale versus color. Accuracy is listed for each of the benchmark datasets LFW [29], CFP-FP [54], AgeDB30 [47], CALFW [69] and CPLFW [68]. Note that all of these benchmark datasets contain primarily color images. From Table 1, it is clear that (1) training on the color-cleaned version of the training set does not result in consistently worse or better accuracy across the validation sets than the original WebFace4M, and (2) the model trained with the grayscale version of the training set does not consistently exhibit either better or worse accuracy across the

validation set when compared to a model trained with the color version of the same training set. Considering that the models are trained on grayscale images, *with no exposure to color during training* and performing comparably on average and even better in certain cases (e.g., LFW and CPLFW for ArcFace ; CALFW for AdaFace) when tested with the model trained on color and gray version of "color-cleaned" WebFace4M, it implies that training on color images does not seem to extract any significant additional information beyond what is present in grayscale.

6. GrayScale Can Improve Accuracy and Efficiency

In the analyses in earlier sections, the grayscale images were in a 3-channel format with R=G=B across the pixels. The results points that color have no consistent advantage over grayscale for SOTA face recognition. If this is the case, then training and testing a network using only single-channel grayscale images should give essentially the same accuracy as training and testing on RGB color images. Furthermore, grayscale images use less disk space than native RGB images. If we utilize the freed-up disk space with additional images, can it enhance the model's accuracy? Additionally, does employing grayscale make the training process more efficient? This section explores these questions.

We modify the Resnet50 backbone so that the first convolutional layer processes a single-channel image rather than a three-channel image. The size of the first convolution block is changed from $64 \times 3 \times 3 \times 3$ to $64 \times 1 \times 3 \times 3$. Following that, we train the adjusted backbone using: a) single-channel grayscale images from the color-cleaned WebFace4M subset, and b) the dataset from (a) combined with additional data from WebFace12M to utilize the emptied disk space previously occupied by RGB images, now converted to grayscale. Subsequently, both trained networks are evaluated on single-channel grayscale images. The accuracy of this fully grayscale face recognition and training is compared to the accuracy of the corresponding fully color network.

Accuracy and Efficiency with single-channel GrayScale Data. The results of this experiment using a ResNet-50 backbone, with ArcFace and with AdaFace loss, appear in Table 3. For ArcFace, the fully grayscale results are marginally higher on CFP-FP and CALFW, and marginally lower on LFW, AGEDB-30 and CPLFW. For AdaFace, the fully grayscale results are marginally higher on CFP-FP, the same on CALFW, and marginally lower on the rest. Networks trained and tested on color offer no consistent improvement over the accuracy achieved by networks trained and tested on single-channel grayscale of the same images.

But, color images require more storage space in disk. For instance, the RGB version of "color-cleaned" WebFace4M training set, occupies 96 GB of disk storage, while the grayscale version takes up 67 GB. The storage of images in disk thus, is reduced to about 2/3 on average in going to single channel grayscale[1]. This reduction in disk space could be beneficial for training deep CNN face networks. During training, GPU memory comprises of the model itself, mini-batch for training with some additional overheads. Large training datasets exceed GPU memory capacity, requiring repeated transfer of mini-batches from the disk to the GPU for training. Grayscale images, with their smaller memory footprint (requiring only 1/3 of the bytes to represent the tensor) compared to RGB, result in reduced data transfer between the CPU and GPU during training. While the improvement in computational and training efficiency for a single forward pass may not be significant, the overall training process benefits from the reduced data transfer volume between the CPU and GPU, resulting in improved training efficiency. For example, with a cluster of 4 Titan-Xp GPUs, we observed approximately 1.2K GBs of data transferred during RGB training and about 0.48K GBs for single-channel grayscale training. This led to an approximately 20% improvement in system CPU time used. These numbers will vary based on the system and its configuration, but generally, training in single-channel grayscale should reduce data transfer volume and system CPU usage.

Improved Accuracy by Utilizing Freed-Up Disk Space. Opting for grayscale images to train and test face recognition networks delivers comparable performance to color images, while consuming less disk space, which frees up additional disk storage. To make the most of the available freed-up disk space, we use additional data from WebFace12M, which shares similar characteristics with WebFace4M. The added data is selected randomly to maintain a consistent number of images per identity as "color-cleaned" WebFace4M. As a result, the total disk space now equals approximately 96 GB, with 282K identities and 5.5M images. Results for this experiment are presented in Table 3. In the case of ArcFace, using extra data during training led to slightly better performance on four datasets: LFW, CFP-FP, AGEDB-30, and CALFW, and slightly worse performance only on CPLFW as compared to RGB training. As for AdaFace, additional data resulted in slightly improved performance on CFP-FP and CALFW, while showing marginally worse performance decrease on the other datasets as compared to RGB training. Thus, leveraging the additional data from the freed-up disk space can enhance the performance of grayscale training, either outperforming color training or reducing the performance gap.

[1] Additional disk space is allocated to store header files that hold essential image metadata. As a result, the occupied disk space is not precisely one-third of the total disk space for RGB images.

Loss →	ArcFace					AdaFace				
Train → / Test ↓	RGB (3)	Gray (1)	Δ ↓	Gray+ (1)	Δ ↓	RGB (3)	Gray (1)	Δ ↓	Gray+ (1)	Δ ↓
LFW	99.80	99.78	0.02%	**99.83**	-0.03%	99.80	99.71	0.09%	99.77	0.03%
CFP-FP	99.04	**99.11**	-0.07%	**99.21**	-0.17%	98.81	**98.84**	-0.03%	**98.96**	-0.17%
AGEDB-30	97.88	97.85	0.03%	**97.88**	0.00%	97.66	97.41	0.25%	97.33	0.33%
CALFW	96.02	**96.10**	-0.08%	**96.08**	-0.06%	95.96	**95.96**	0%	**96.05**	-0.09%
CPLFW	94.18	93.85	0.35%	94.10	0.08%	94.30	93.90	0.4%	94.13	0.17%
Average	97.38	97.34	0.04%	**97.42**	-0.04%	97.31	97.16	0.14%	97.25	0.05%

Δ Relative error of the accuracies computed as RGB (3) - Gray/+ (1)

Table 3. **Training with Additional Data using the Freed Disk Space can improve Grayscale accuracy compared to RGB**. 1:1 Verification accuracy (%) when trained and evaluated all in single-channel grayscale versus trained and evaluated in RGB is essentially the same. Improving accuracy is possible by using additional data in the freed-up disk space. RGB (3) and Gray (1) represents the color-cleaned WebFace4M in RGB and one-channel Grayscale format. **Gray+ (1)** represents the data in Gray (1) + additional one-channel grayscale data pooled from WebFace12M to make use of emptied diskspace. [Key: **Gray better or equal to RGB**]

7. Importance of Color for Face Recognition

We have quantitatively analyzed model performance using RGB and grayscale training sets in the previous sections. Now, we will qualitatively explore color's role in face recognition and address these key questions:

- Can altering color spaces enhance the extraction of color-related details?
- Does the feature extractor learn distinct features from different color planes?
- Is color consistent across an identity's training images?

7.1. Is HSV Color Space Better Than RGB?

RGB is the universal format for color images in face recognition pipelines. However, results in the previous sections may motivate the question of whether a different color space might enable the network to learn more from the color images. RGB can be viewed as having the disadvantage of not explicitly separating chromaticity and luminosity. This limitation could be problematic when dealing with web-scraped images that are captured under varying lighting conditions. To investigate whether a color space that separates chroma and luma can enable to network to learn more from color, we ran a parallel set of experiments with color images converted to HSV (Hue, Saturation, Value) color space for training and testing. Since RGB and HSV both consist of three channels, no change is needed in the CNN architecture. The advantage of HSV for this experiment is that luma information is isolated in one plane (value), and chroma (color) in the other two planes (hue and saturation). Potentially, this could enable the network to better exploit color, if it is indeed relevant for the task.

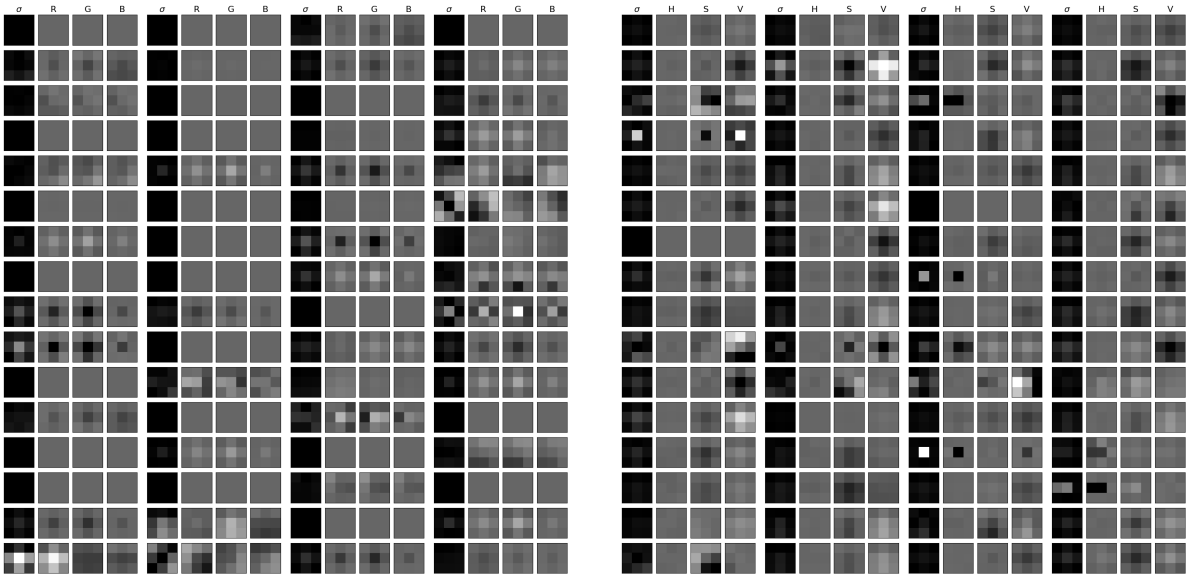
We trained and tested ArcFace and AdaFace using an HSV version of the "color-cleaned" subset of WebFace4M. Results of this experiment are in the last row of the ArcFace and AdaFace sections of Table 1, which shows that *representing color in HSV does not result in consistently better or worse accuracy than RGB*.

For example, in the case of ArcFace, the accuracy is slightly higher for the HSV color space compared to the RGB in LFW and CPLFW, and it remains the same for CALFW when using the model trained with the "color-cleaned" WebFace4M RGB and its corresponding HSV version, as indicated in Table 1. In other cases, HSV accuracy is slightly lower than the corresponding RGB accuracy. A more detailed analysis of which HSV planes the HSV-trained model relied on is presented in the next section.

7.2. Do the First-layer Filters "See" Color?

In this section, we analyze what the network learns about using color by visualizing the pattern of weights learned in the first convolutional layer for ArcFace. For the ResNet backbone, the color image is input to the first convolutional layer, and the first layer learns 64 different $3 \times 3 \times 3$ convolution filters, each of which can extract a different feature image from the original color image. Each convolution is $3 \times 3 \times 3$ because it is a 3×3 kernel applied to each of the 3 (R, G, B) color planes. After the first layer, the learned weights are no longer directly tied to the color planes.

We visualize the $3 \times 3 \times 3$ learned weights for a given one of the 64 convolutions through a set of four 3×3 grayscale grids. The first grid represents the standard deviation of the values across the R, G, and B weights at each pixel position. An all-black 3x3 grid in the first column shows that the weights are the same across R, G and B; in effect, the learned filter is extracting grayscale information. White in a 3×3 grid in the first column represents the maximum standard deviation across the R, G, and B weights among all 64 sets of $3 \times 3 \times 3$ weights. The second, third and fourth grids represent the weights for the R, G and B planes, respectively. The weights are linearly scaled for better visualization. In these three columns, negative weights are depicted as black, zero weights as gray, and positive weights as white. This scheme allows us visualize characteristics of



(a) Trained on "color-cleaned" WebFace4M (on RGB)

(b) Trained on "color-cleaned" WebFace4M (on HSV)

Figure 3. **Visualization of 64 Convolution Filter Weight Values for the First Convolution Block in row-major order.** Approximately one-third of the convolution blocks dedicated to RGB data reached nearly zero values, while the remaining blocks exhibited a strikingly similar pattern across the RGB planes. However, only a few convolution blocks displayed significantly different values for the RGB planes. In contrast, when considering the HSV plane, it appears that the most active block primarily derives its information from the V plane, indicating that the network learns to extract more valuable data from this particular plane compared to the others.

the learned filter weights.

Figure 3 contains the visualization of the weights for training ArcFace on the RGB color-cleaned subset of WebFace4M on the left, and the weights for training ArcFace on the HSV version of the same images on the right. Each set of 64 convolution weights is shown as 4 columns of 16. Consider the visualization of the convolution weights for ArcFace trained on RGB, as shown in Fig 3a. The visualization of convolution weights in the upper left corner shows that the standard deviation of the weights is zero (black in the first 3×3 grid) and that the weights in the 3×3 convolution are zero for each of the R, G and B color planes (grey in the other 3×3 grids). This is an example of a filter that converged to a convolution that just produces the value zero. About one third of the 64 learned convolutions converged to a similar result, consistent with the observations documented in [8]. Most of the remaining convolutions have a weight pattern that is very similar for R, G and B. *A similar pattern of weights across R, G and B would produce a result very similar to applying the average weights on a grayscale version of the image.* Only a few of the convolutions have a pattern of weights that appears substantially different across R, G and B. For example, the convolution in the lower left in Fig 3a appears to be a spot or line detector using primarily the R plane.

Contrast the visualization of the weights for the RGB-trained ArcFace with those for the HSV-trained ArcFace, shown in Figure 3b. In the visualization of these 64 convolutions, there is much more variation in the pattern of weights across H, S and V. The weights are generally near zero across the 3×3 grid for H. There is generally more variation in the pattern of weights for S. And by far the greatest overall variation is in V. Thus, the HSV-learned weights suggest that the *network learns to extract more information from the V plane, which is effectively just the grayscale*, than from the other two planes. Our results are consistent with Albiol et al's finding that there is an equivalent optimal skin detector for every color space [6]. These visualizations together do not prove that there are no useful information in the color content of the images, but they do show that the network learns to extract mostly grayscale information. The next section presents an analysis of the color of the skin region for selected identities in the training set that suggests why this might be the case.

7.3. Color Variation Within an Identity's Images

The training of a deep CNN face matcher aims to classify all images of a given identity as that identity, for all identities in the training set. The network has the potential to extract features from color information, provided it is beneficial, by identifying shared color elements that are unique to each identity. We analyze selected training set images to illustrate how useful, or not, the color information in the training images could be.

Fifty of the most frontal images for each of four distinct identities are chosen from the WebFace4M training

data, on the basis of having strongly different skin tone from each other, and each identity having a large number of images in the training set. To select suitable quality images for analysis of the face skin region in the image, we use a BiSeNet semantic segmentation [66] to filter out faces where less than 30% of the face area is classified as skin. This gives us mostly frontal faces, without too much occlusion by glasses or scalp hair. To avoid inconsistencies stemming from mouth open/closed and facial expression, we focus on the part of the face above the upper lip, and on the pixels classified as skin (omitting eyebrows and eyes), and calculate the average RGB of skin pixels. For each of the four identities, we select the 50 images that have the largest number of such skin pixels and plot the average RGB for each of their images in a 3D RGB space. If color is useful to separate images belonging to different identities in the training set, then the 50 images of each identity should form a compact cluster that is well separated from the other identities.

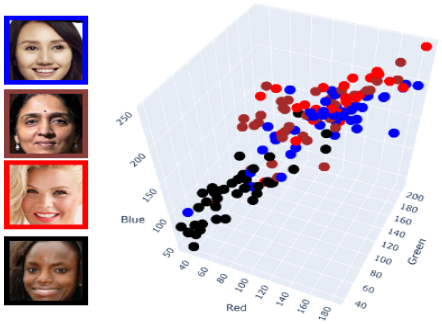


Figure 4. **Identities with different skin tones tend to cluster in 3D RGB space**. It is evident that the RGB representation of visible skin pixels for the same identity is not tightly clustered. This variation reflects the diverse nature of the training set, where individuals of the same identity can exhibit different skin tones depending on lighting conditions. This observation suggests that color may not be crucial information for the network to utilize, given the variation in skintones within the images of the same identity.

From Figure 4, it is clear that none of the four identities have images where the average RGB of their skin region forms a compact cluster. Visually, it is clear that the cluster of points for each of the identities is broad, and that the four clusters are highly overlapped. As a quantitative indication of this, we can consider the fraction of points whose nearest neighbor is from the same identity. If color was highly useful in classifying images of these four identities, then each point's nearest neighbor would be from the same identity. If color was a purely random clue, then with four identities, a point's nearest neighbor would be from the same identity about 25% of the time. Across the 200 points from the four strongly different identities in Figure 4, a point's nearest neighbor is from the same identity about two thirds of the time, roughly halfway between random and highly useful. Analyzing a larger number of identities, and identities not chosen to have clearly different skin color, would drive the result closer to random. This analysis suggests that the color information in a set of web-scraped, in-the-wild face images simply does not contain very much useful identity-related information.

8. Results and Discussion

Computing resources used unproductively processing color. A deep CNN face matcher trained on single-channel grayscale images, and matching single-channel grayscale images, achieves essentially the same accuracy as a network working with RGB color images. But the single-channel grayscale images use 1/3 the memory of the RGB images. And the early convolutional layer of the grayscale network has 1/3 the weights of the color network. Furthermore, opting for grayscale images can save disk space in comparison to color images, enabling the use of additional data for training a superior face recognition model.

SOTA deep CNN face matchers do not "use" color. When the deep CNN is trained on only grayscale training images (stored in three-channel format) and evaluated on RGB color images, essentially the same accuracy is achieved as when the deep CNN is trained on RGB color images – the network learns as much from grayscale training data as from color training data, for the purpose of matching color face images.

Conditions specific to deep CNN face matching. It is known that the color is important for some general object detection tasks solved by deep CNNs [16, 55]. Our results here are not in conflict with these studies. Deep CNN face matching is a specialized task. Matchers are trained to recognize (categorize) persons from in-the-wild, web-scraped images, and color is not consistent across images of a person in this context. It is possible that a very tightly controlled face matching application, with all images always acquired in the same lighting and with consistent background, could result in color being more useful.

The role of color in demographic accuracy differences. It is acknowledged that face recognition accuracy varies across demographic groups [9, 14, 15, 21, 23, 30, 32, 33, 37–42, 44, 48, 50, 58, 61–63, 67]. Discussion of this topic often mentions skin tone or skin color. Our results suggest that the accuracy differences are not specific to using images with color content. The same differences can be observed when processing grayscale images rather than color.

References

[1] Achromatopsia. https://medlineplus.gov/genetics/condition/achromatopsia/. 1

[2] Adaface: Quality adaptive margin for face recognition. https://github.com/mk-minchul/AdaFace. 3

[3] Insightface: 2d and 3d face analysis project. https://github.com/deepinsight/insightface/. 2, 3

[4] Morph dataset. https://uncw.edu/oic/tech/morph.html. 2

[5] Vítor Albiero, Kai Zhang, Michael C King, and Kevin W Bowyer. Gendered differences in face recognition accuracy explained by hairstyles, makeup, and facial morphology. In *Transactions on Information, Forensics and Security (TIFS)*, volume 17, pages 127–137, 2021. 2

[6] A. Albiol, L. Torres, and E.J. Delp. Optimum color spaces for skin detection. In *Interantional Conference on Image Processing (ICIP)*, 2001. 7

[7] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, et al. Partial fc: Training 10 million identities on a single machine. In *International Conference on Computer Vision (ICCV)*, pages 1445–1449, 2021. 2, 3

[8] Aman Bhatta, Domingo Mery, Haiyu Wu, and Kevin W. Bowyer. Craft: Contextual re-activation of filters for face recogntion training. *arXiv preprint arXiv:2312.00072*, 2023. 7

[9] Aman Bhatta, Gabriella Pangelinan, Micheal C. King, and Kevin W. Bowyer. Impact of blur and resolution on demographic disparities in 1-to-many facial identification. In *Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2024. 8

[10] Markus Bindemann and A Mike Burton. The role of color in human face detection. In *Cognitive Science*, volume 33, pages 1144–1156. Wiley Online Library, 2009. 2

[11] Patricia Brosseau, Adrian Nestor, and Marlene Behrmann. Colour blindness adversely impacts face recognition. In *Visual Cognition*, volume 28, pages 279–284. Taylor & Francis, 2020. 2

[12] Vicki Bruce and Andy Young. *In the eye of the beholder: The science of face perception.* Oxford university press, 1998. 2

[13] Vanessa Buhrmester, David Münch, Dimitri Bulatov, and Michael Arens. Evaluating the impact of color information in deep neural networks. In *Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA)*, pages 302–316, 2019. 2

[14] Cynthia M Cook, John J Howard, Yevgeniy B Sirotin, Jerry L Tipton, and Arun R Vemury. Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems. In *Transactions on Biometrics, Behavior, and Identity Science (T-BIOM)*, volume 1, pages 32–41, 2019. 8

[15] Cynthia M Cook, John J Howard, Yevgeniy B Sirotin, Jerry L Tipton, and Arun R Vemury. Demographic effects across 158 facial recognition systems. Technical report, DHS, Aug. 2023. 8

[16] Kanjar De and Marius Pedersen. Impact of colour on robustness of deep neural networks. In *International Conference on Computer Vision (ICCV)*, pages 21–30, 2021. 2, 8

[17] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3

[18] Javier Diaz-Cely, Carlos Arce-Lopera, Juan Cardona Mena, and Lina Quintero. The effect of color channel representations on the transferability of convolutional neural networks. In *Computer Vision Conference (CVC)*, pages 27–38, 2020. 2

[19] Samuel Dodge and Lina Karam. Understanding how image quality affects deep neural networks. In *International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2016. 2

[20] Samuel Dodge and Lina Karam. A study and comparison of human and deep learning recognition performance under visual distortions. In *International Conference on Computer Communication and Networks (ICCCN)*, pages 1–7, 2017. 2

[21] Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. Demographic bias in biometrics: A survey on an emerging challenge. In *Transactions on Technology and Society (TTS)*, volume 1, pages 89–103, 2020. 8

[22] Martin Engilberge, Edo Collins, and Sabine Süsstrunk. Color representation in deep neural networks. In *Interantional Conference on Image Processing (ICIP)*, pages 2786–2790, 2017. 2

[23] Biying Fu and Naser Damer. Towards explaining demographic bias through the eyes of face recognition models. In *International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2022. 8

[24] Sanjukta Ghosh, Rohan Shet, Peter Amon, Andreas Hutter, and André Kaup. Robustness of deep convolutional neural networks for image degradations. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2916–2920, 2018. 2

[25] Shreyank N Gowda and Chun Yuan. Colornet: Investigating the importance of color spaces for image classification. In *Asian Conference on Computer Vision (ACCV)*, pages 581–596, 2019. 2

[26] 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. In *IET Biometrics*, volume 7, pages 81–89, 2018. 2

[27] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision (ECCV)*, pages 87–102, 2016. 1, 3

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1, 3

[29] 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. 1, 4

[30] Linzhi Huang, Mei Wang, Jiahao Liang, Weihong Deng, Hongzhi Shi, Dongchao Wen, Yingjie Zhang, and Jian Zhao. Gradient attention balance network: Mitigating face recognition racial bias via gradient attention. In *Computer Vision*

and Pattern Recognition Workshops (CVPRW), pages 38–47, 2023. 8

[31] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5901–5910, 2020. 1

[32] Marco Huber, Anh Thi Luu, Fadi Boutros, Arjan Kuijper, and Naser Damer. Bias and diversity in synthetic-based face recognition. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 6215–6226, 2024. 8

[33] Anubhav Jain, Nasir Memon, and Julian Togelius. Zero-shot racially balanced dataset generation using an existing biased stylegan2. In *International Joint Conference on Biometrics (IJCB)*, pages 1–18, 2023. 8

[34] 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 *International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5, 2016. 2

[35] Richard Kemp, Graham Pike, Peter White, and Alex Musselman. Perception and recognition of normal and negative faces: the role of shape from shading and pigmentation cues. In *Perception*, volume 25, pages 37–52. SAGE Publications Sage UK: London, England, 1996. 2

[36] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 18750–18759, 2022. 1

[37] Manideep Kolla and Aravinth Savadamuthu. The impact of racial distribution in training data on face recognition bias: A closer look. In *Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 313–322, 2023. 8

[38] Ketan Kotwal and Sébastien Marcel. Mitigating demographic bias in face recognition via regularized score calibration. In *Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2024. 8

[39] KS Krishnapriya, Vítor Albiero, Kushal Vangara, Michael C King, and Kevin W Bowyer. Issues related to face recognition accuracy varying based on race and skin tone. In *Transactions on Information, Forensics and Security (TIFS)*, volume 1, pages 8–20, 2020. 8

[40] KS Krishnapriya, Gabriella Pangelinan, Michael C King, and Kevin W Bowyer. Analysis of manual and automated skin tone assignments. In *Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 429–438, 2022. 8

[41] Hao Liang, Pietro Perona, and Guha Balakrishnan. Benchmarking algorithmic bias in face recognition: An experimental approach using synthetic faces and human evaluation. In *International Conference on Computer Vision (ICCV)*, pages 4977–4987, 2023. 8

[42] Pranay K Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R Varshney, and Ruchir Puri. Bias mitigation post-processing for individual and group fairness. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2847–2851, 2019. 8

[43] Puspita Majumdar, Surbhi Mittal, Richa Singh, and Mayank Vatsa. Unravelling the effect of image distortions for biased prediction of pre-trained face recognition models. In *International Conference on Computer Vision (ICCV)*, pages 3786–3795, 2021. 2

[44] Vongani H Maluleke, Neerja Thakkar, Tim Brooks, Ethan Weber, Trevor Darrell, Alexei A Efros, Angjoo Kanazawa, and Devin Guillory. Studying bias in gans through the lens of race. In *European Conference on Computer Vision (ECCV)*, pages 344–360, 2022. 8

[45] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Computer Vision and Pattern Recognition (CVPR)*, pages 14225–14234, 2021. 1

[46] Tetsuto Minami, Kae Nakajima, and Shigeki Nakauchi. Effects of face and background color on facial expression perception. In *Frontiers in psychology*, volume 9, page 1012. Frontiers Media SA, 2018. 2

[47] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, volume 2, page 5, 2017. 4

[48] Kagan Ozturk, Grace Bezold, Aman Bhatta, Haiyu Wu, and Kevin W. Bowyer. Beard segmentation and recognition bias. *arXiv preprint arXiv:2308.15740*, 2023. 8

[49] Yanting Pei, Yaping Huang, Qi Zou, Xingyuan Zhang, and Song Wang. Effects of image degradation and degradation removal to cnn-based image classification. In *Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, volume 43, pages 1239–1253, 2019. 2

[50] Malsha V Perera and Vishal M Patel. Analyzing bias in diffusion-based face generation models. In *International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2023. 8

[51] Karl Ricanek and Tamirat Tesafaye. MORPH: A longitudinal image database of normal adult age-progression. In *Automatic Face and Gesture Recognition (AFGR)*, 2006. 2

[52] Prasun Roy, Subhankar Ghosh, Saumik Bhattacharya, and Umapada Pal. Effects of degradations on deep neural network architectures. *arXiv preprint arXiv:1807.10108*, 2018. 2

[53] Rajan Sachin, V Sowmya, D Govind, and KP Soman. Dependency of various color and intensity planes on cnn based image classification. In *International Symposium on Signal Processing and Intelligent Recognition Systems (SIRS)*, pages 167–177, 2018. 2

[54] S. Sengupta, J.C. Cheng, C.D. Castillo, V.M. Patel, R. Chellappa, and D.W. Jacobs. Frontal to profile face verification in the wild. In *Winter Conference on Applications of Computer Vision (WACV)*, February 2016. 4

[55] Aditya Singh, Alessandro Bay, and Andrea Mirabile. Assessing the importance of colours for cnns in object recognition. In *Conference on Neural Information Processing Systems Workshops (NeurIPSW)*, 2020. 2, 8

[56] Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell. Face recognition by humans: Nineteen results all

computer vision researchers should know about. In *Proceedings of the IEEE*, volume 94, pages 1948–1962, 2006. 2

[57] Philipp Terhörst, Malte Ihlefeld, Marco Huber, Naser Damer, Florian Kirchbuchner, Kiran Raja, and Arjan Kuijper. Qmagface: Simple and accurate quality-aware face recognition. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 3484–3494, 2023. 1

[58] Philipp Terhörst, Jan Niklas Kolf, Marco Huber, Florian Kirchbuchner, Naser Damer, Aythami Morales Moreno, Julian Fierrez, and Arjan Kuijper. A comprehensive study on face recognition biases beyond demographics. In *Transactions on Technology and Society (TTS)*, volume 3, pages 16–30, 2021. 8

[59] Ronny Velastegui, Linna Yang, and Dong Han. The importance of color spaces for image classification using artificial neural networks: a review. In *Computational Science and Its Applications (ICCSA)*, pages 70–83, 2021. 2

[60] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5265–5274, 2018. 1

[61] Haiyu Wu, Vítor Albiero, KS Krishnapriya, Michael C King, and Kevin W Bowyer. Face recognition accuracy across demographics: Shining a light into the problem. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1041–1050, 2023. 8

[62] Haiyu Wu, Grace Bezold, Aman Bhatta, and Kevin W Bowyer. Logical consistency and greater descriptive power for facial hair attribute learning. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8588–8597, 2023. 8

[63] Haiyu Wu and Kevin W Bowyer. What should be balanced in a “balanced" face recognition dataset? In *British Machine Vision Conference (BMVC)*, 2023. 8

[64] Jiasong Wu, Shijie Qiu, Rui Zeng, Lotfi Senhadji, and Huazhong Shu. Pcanet for color image classification in various color spaces. In *International Conference on Cloud Computing and Security (ICCCS)*, pages 494–505, 2017. 2

[65] Andrew W Yip and Pawan Sinha. Contribution of color to face recognition. In *Perception*, volume 31, pages 995–1003. SAGE Publications Sage UK: London, England, 2002. 2

[66] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *European Conference on Computer Vision (ECCV)*, 2018. 8

[67] Seyma Yucer, Furkan Tektas, Noura Al Moubayed, and Toby P Breckon. Racial bias within face recognition: A survey. *arXiv preprint arXiv:2305.00817*, 2023. 8

[68] T. Zheng and W. Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. Technical Report 18-01, Beijing University of Posts and Telecommunications, February 2018. 4

[69] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017. 4

[70] Yiren Zhou, Sibo Song, and Ngai-Man Cheung. On classification of distorted images with deep convolutional neural networks. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1213–1217, 2017. 2

[71] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10492–10502, 2021. 2, 3