Logical Consistency and Greater Descriptive Power for Facial Hair Attribute Learning

Haiyu Wu[1], Grace Bezold[1], Aman Bhatta[1], Kevin W. Bowyer[1]
[1]University of Notre Dame

Abstract

Face attribute research has so far used only simple binary attributes for facial hair; e.g., beard / no beard. We have created a new, more descriptive facial hair annotation scheme and applied it to create a new facial hair attribute dataset, FH37K. Face attribute research also so far has not dealt with logical consistency and completeness. For example, in prior research, an image might be classified as both having no beard and also having a goatee (a type of beard). We show that the test accuracy of previous classification methods on facial hair attribute classification drops significantly if logical consistency of classifications is enforced. We propose a logically consistent prediction loss, LCPLoss, to aid learning of logical consistency across attributes, and also a label compensation training strategy to eliminate the problem of no positive prediction across a set of related attributes. Using an attribute classifier trained on FH37K, we investigate how facial hair affects face recognition accuracy, including variation across demographics. Results show that similarity and difference in facial hairstyle have important effects on the impostor and genuine score distributions in face recognition. The code is at https://github.com/HaiyuWu/LogicalConsistency.

1. Introduction

Facial attributes have been widely used in face matching/recognition [8, 14, 30, 31, 37, 43], face image retrieval [34, 39], re-identification [42, 44, 45], training GANs [15, 16, 25, 33] for generation of synthetic images, and other areas. As an important feature of the face, facial hairstyle does not attract enough attention as a research area. One reason is that current datasets have only simple binary attributes to describe facial hair, and this does not support deeper investigation. This paper introduces a more descriptive set of facial hair attributes, representing dimensions of the area of face covered, the length of the hair, and connectedness of beard/mustache/sideburns. We also propose a logically consistent predictions loss function, LCPLoss, and label compensation strategy to enhance the logical consistency of the predictions. We illustrate the use of this new, richer set of facial hair annotations by investigating the effect of beard area on face recognition accuracy across demographic groups. Contributions of this work include:

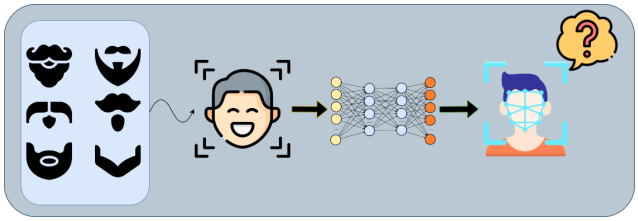

Figure 1. (1) *What is the best way to define the facial hair styles?* (2) *How does the facial hair classifier perform in the real-world cases?* (3) *How does the face matcher treat the same (different) person with different (same) beard styles?* This paper presents our approaches and answers for these questions.

- A richer scheme of facial hair attributes is defined and annotations are created for the FH37K dataset. The attributes describe facial hair features along dimensions of area of the face covered, length of hair and connectedness of elements of the hair (See Sec. 2 and 4.1).
- The logical consistency of classifications of the facial hair attribute classifier is analyzed. We show that the proposed LCPLoss and label compensation strategy can significantly reduce the number of logically inconsistent predictions (See Section 5 and Section 6.1).
- We analyze the effect of the beard area on face recognition accuracy. Larger difference in beard area between a pair of images matched for recognition decreases the similarity value of both impostor and genuine image pairs. Interestingly, the face matchers perform differently across demographic groups when image pairs have the same beard area. (See Section 6.2)

2. Facial Hair In Face Attribute Datasets

For a broad discussion of face attribute classification research, see the recent survey by Zheng et al [55]. Here, we

	# of images	# of ids	# of facial hair attributes	Area	Length	CNDN	E_c
Berkeley Human Attributes [10]⋆	8,053	-	0	0	0	0	✗
Attributes 25K [54]	24,963	24,963	0	0	0	0	✗
FaceTracer [29]⋆	15,000	15,000	1 (Mustache)	0	0	0	✗
Ego-Humans [49]	2,714	-	1 (Facial hair)	0	0	0	✗
CelebA [36]⋆	202,599	10,177	5 (5 o'Clock, Goatee, ...)	1	1	0	✗
LFWA [36]⋆	13,233	5,749	5 (5 o'Clock, Goatee, ...)	1	1	0	✗
PubFig [32]⋆	58,797	200	5 (5 o'Clock, Goatee, ...)	1	1	0	✗
LFW [26]⋆	13,233	5,749	5 (5 o'Clock, Goatee, ...)	1	1	0	✗
UMD-AED [22]	2,800	-	5 (5 o'Clock, Goatee, ...)	1	1	0	✗
YouTube Faces Dataset (with attribute labels [23])	3,425	1,595	5 (5 o'Clock, Goatee, ...)	1	1	0	✗
CelebV-HQ [56]⋆	35,666 video clips	15,653	5 (5 o'Clock, Goatee, ...)	1	1	0	✗
MAAD-Face [47]⋆	3.3M	9,131	5 (5 o'Clock, Goatee, ...)	1	1	0	✓
FH37K (this paper)	37,565	5,216	**17 (Chin area, Short...)**	**4**	**4**	**4**	✓

Table 1. Comparison of facial hair descriptions in face attribute datasets. CNDN and E_c stand for connectedness and estimating the consistency rate of the annotations. Datasets with ⋆ are available online. FH37K has richer annotations that can cover the area, length, and connectedness of the facial hair.

briefly summarize selected details of existing facial attribute datasets, focusing on attributes describing facial hair.

Bourdev et al [10] assembled 8,053 images from the H3D dataset [11] and the PASCAL VOC 2010 dataset [50] to create the Berkeley Human Attributes (BHA) dataset. They use Mechanical Turk to create 9 attributes, merging values from 5 independent annotators. Zhang et al [54] collect the Attribute 25K dataset, which contains 24,963 images from 24,963 people on Facebook. They provide 8 attributes for each image. This work, unlike most previous work in face attributes, acknowledges that some attributes may not be able to be inferred from some images. However, how they use the "uncertain" label is not mentioned in the original paper and the dataset is not available for use. **We have an attribute called "Info Not Vis" and this attribute is used in our training and testing.** Neither of [10, 54] includes any attribute to describe facial hair.

Kumar et al [29] collected 15,000 in-the-wild face images to build the FaceTracer dataset. The images have 10 groups of attributes including gender, age, race, environment, etc. The only attribute related to facial hair is mustache / no mustache. Similarly, Wang et al [49] collect five million images from videos by using the OpenCV frontal face detector to create the Ego-Humans dataset. There are annotations for 17 face attributes, including facial hair / no facial hair. These two works [29,49] each have only a single binary attribute related to facial hair.

The Labeled Faces in the Wild [26] (LFW) dataset has 13,233 images of cropped, aligned faces. There are 1,680 identities in LFW that have two or more images. Kumar et al [32] collected 65 attributes through Mechanical Turk [1] and added 8 more [30] for a total of 73 attributes. Kumar et al [32] also collect 58,797 images from 200 people to build the PubFig dataset. All the images are from the internet with varied pose, lighting, expression, etc. This dataset provides 73 facial attributes. Liu et al [36] collect the largest facial attribute dataset to date, CelebA, which has 202,599 images from 10,177 identities. It has 40 facial attributes and all the annotations are generated by a professional labeling company. They also provide the annotations of the same attributes on the LFW dataset. The University of Maryland Attribute Evaluation Dataset (UMD-AED) [22] serves as an evaluation dataset. It consists of 2,800 images and each attribute has 50 positive and 50 negative samples. They use the same 40 facial attributes as the LFWA and the CelebA datasets. Hand et al [23] collect 3,425 frames from the original YouTube Faces Dataset. They also use the same 40 attributes. Terhörst et al [47] created MAADFace by training a network to apply the 47 attributes across LFW and CelebA to the images from VGGFace2 [13]. An interesting element of this work is that the network estimates its confidence in assigning attribute values, and about *20%* of MAADFace attribute values are left unassigned due to uncertainty. A recent facial attributes related dataset [56] contains 35,666 high quality video clips. There are 83 manually labeled facial attributes covering appearance, action, and emotion.

The facial hair attributes used in existing datasets are summarized in Table 1. The same five attributes have been used is nearly all previous work. A richer description of facial hairstyles is needed to enable research into how facial hairstyle affects face recognition accuracy. Our FH37K dataset has attributes to describe dimensions of area of the face covered by facial hair, length of facial hair and connectedness of parts of facial hair. No previous work has this

level of descriptive power for facial hair, or considers the logical consistency of the set of facial hair attributes.

3. Overview of the FH37K Dataset

3.1. Dataset statistics

FH37K contains 37,565 images, coming from a subset of CelebA [36] and a subset of WebFace260M [57]. There are 5,216 identities (3,318 identities from CelebA and 1,898 from WebFace260M). The 3,318 identities of FH37K coming from the CelebA dataset are split into train/val/test as they were in CelebA. The identities from WebFace260M were randomly split 40%/30%/30% to train/val/test. The resulting FH37K has 28,485 images for training, 4,829 for validation, and 4,251 for testing.

All the images are manually annotated with respect to a detailed definition for each annotation, and examples and strategies for marking challenging images. The annotations of each image follow the logical relationship among the attributes. Because subjectivity and ambiguity in assigning annotation values can only be controlled and not eliminated, we also estimate the level of consistency expected between a new annotator re-annotating the FH37K images and the annotations distributed as part of FH37K.

3.2. Dimensions of facial hair properties

FH37K has a larger and richer set of facial hair attributes that can be grouped into three dimensions: facial hair area, length, connectedness.

- **Beard Area**: Three levels of beard area are *Clean Shaven* (no beard), *Chin Area* (beard limited to chin area) and *Side to Side* (extending to sides of face).
- **Beard Length**: The five levels of length are *Clean Shaven*, *5 O'clock Shadow*, *Short*, *Medium* and *Long*. The *Clean Shaven* attribute can be seen as an element of description for both area and length.
- **Mustache**: Mustache-related values are *Mustache-None*, *Mustache Isolated* (meaning not connected to beard) and *Mustache Connected to Beard*.
- **Sideburns**: Sideburns-related attribute values are *Sideburns-None*, *Sideburns-Present* (not connected to beard) and *Sideburns Connected to Beard*.
- **Bald**: Bald describes scalp hair rather than facial hair, but is included in FH37K to support possible future research without needing to annotate images again. Values include *Bald False*, *Bald Top Only*, *Bald Sides Only* and *Bald Top and Sides*.
- **Information is not visible**: With in-the-wild imagery, it is common that information is not visible in the image to assign a value for some attribute [53]. Most previous face attribute datasets ignore this issue. In FH37K, we use attribute values (*Beard Area Info Not Vis, Beard Length Info Not Vis, Mustache Info Not Vis, Sideburns Info Not Vis, Bald Info Not Vis*).

More details of these 22 attributes are in Sec. 4.1 and the number of positive samples for each attribute is in Table 1 of the Supplementary Material. Examples of each attribute can be found in Figure 1 to 6 of Supplementary material.

4. FH37K Data Collection

Images in FH37K are cropped and aligned. For CelebA, images with distributed CelebA annotations of No_beard=false were reviewed for possible inclusion in FH37K and new annotations. A large fraction of CelebA images with a No_beard=false annotation actually did not have facial hair and were dropped from FH37K, and 253 images with a No_beard=false annotation actually did not contain a face and were dropped. CelebA images kept for FH37K were manually annotated. Annotators read a document containing definitions and examples of the FH37K annotations before annotating and were encouraged to refer to the document as needed. Images from CelebA had a low number of positive examples of some FH37K attributes. A classifier was trained using this data and run on WebFace260M to generate images of additional identities, with a focus on increasing the initially under-represented positive examples. The 4,274 images selected from WebFace260M resulted in all attributes except bald only on sides and long beard having at least 1,000 images. The images from WebFace260M were then manually assigned attribute values in the same way as for CelebA. The result is FH37K, 37,565 images with an aggregate total of 0.8M annotations.

4.1. Complications for Consistent Annotation

Consistent annotations that align the content of the images and the concept to be learned is an important element of any machine learning dataset. To ensure that each annotator is oriented to the same concept for each attribute, we provided a document with detailed definition and examples for each attribute. However, there are still difficulties to mark annotations consistently on these in-the-wild images.

Figure 2 shows four main complications: ambiguous definition of "chin area", varying beard length, beard area information partially visible, and beard length information partially visible. Without an explicit definition, "chin area" is subjective and can be interpreted differently by different annotators. To address this, we gave annotators the specific definition that the chin area is within parallel vertical lines extending from the outer eye corners, as shown in Figure 2a. Images in Figure 2b show that the beard length can vary over the area of the beard. To address this, annotators were asked to select the length value representing the

(a) Ambiguity on Chin Area (b) Multiple Beard Length (c) Area Info Partially Visible (d) Length Info partially visible

Figure 2. Example complications for marking images consistently. More examples are in Figure 7 of Supplementary Material.

longest length of the beard. Head pose, occlusion, and lighting angle varies broadly in any in-the-wild dataset, giving rise to complications illustrated in Figures 2c and 2d. A single attribute is not sufficient to describe these circumstances, and so we use the visible part plus the Info Not Vis attribute to describe these images.

To evaluate the consistency of our annotations, a fresh annotator independently annotated a random set of 1,000 images from FH37K. This annotator had the same training documentation as the original annotators, but did not know the initial annotation values. This annotator's results were compared to the FH37K annotations to estimate the level of agreement that a different annotator would have with the FH37K annotations. The estimated consistency is 94.05%. (Analysis is in Table 2 of Supplementary Material).

5. Logically Consistent Prediction

For facial attribute classification tasks, some papers group attributes based on position [12, 18] or correlation [21,46] to improve the accuracy on benchmarks. However, to our best knowledge, no previous work considers the logical relationship between attributes on predictions. For example, in CelebA, (no-beard=true) and (goatee=true) would be logically inconsistent; so would (bald=true) and any of the hairstyles=true or hair colors=true, male=false and any of beard related attributes=true. In FH37K, where groups of facial hair attributes are defined to cover the range of possibilities, there is another logical constraint, where the model should give exactly one positive prediction across a set of related attributes. We formulate these issues into three categories and introduce a solution in this section.

5.1. Logically Consistent Prediction Loss

Consider a set of N 2D image $X = \{x_1, x_2, ..., x_N\}$ and their ground truth labels $Y = \{y_1, y_2, ..., y_N\}$, where $X \in \mathbb{R}^{D\times H\times W}$ as the D-dimension batch input and $Y \in \mathbb{R}^{D\times K}$ as the D-dimension batch output with K predicted labels for each dimension. To train a multi-label classifier $f(X, W)$, Binary Cross Entropy Loss (BCELoss) is used:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N} y_i log(p(y_i)) + (1 - y_i) log(1 - p(y_i)) \tag{1}$$

The relative sparsity of positive labels in the multi-label classification tasks means that BCELoss guides the model to over-predict negative labels, which increases the accuracy on the benchmarks but reduces real-world utilty. Our approach is to force the model to consider the logical relationships - *mutually exclusive*, *collectively exhaustive*, *dependency* - among groups of attributes:

- *mutually exclusive*: For some attribute groups, logical consistency requites that at most one can be positive.
- *dependency*: If attribute A is true, the attribute B must be true, otherwise the predictions are impossible.
- *collectively exhaustive*: For some attribute groups, logical consistency requites exactly one must be positive.

Failure cases of these three relationships are in Algorithm 1 of Supplementary Material. Based on these three relationships, we propose the LCPLoss to force the model to make logically consistent predictions.

For *mutually exclusive*, we formulate the sets $A_{ex} = \{attr_1, attr_2, ..., attr_N\}$ and $L_{ex} = \{l_1, l_2, ..., l_N\}$, where l_N is the list of attributes that are mutually exclusive to the $attr_N$. Then, the probability of the mutually exclusive attributes happening at the same time is:

$$\mathcal{P}_{ex} = \mathcal{P}(L_{ex}|A_{ex})P(A_{ex}) \tag{2}$$

For the *dependency* relation, we formulate the set $A_d = \{attr_1, attr_2, ..., attr_N\}$ and $L_d = \{l_1, l_2, ..., l_N\}$, where $attr_N$ is the sufficient condition to the attributes in l_N. The function is:

$$\mathcal{P}_d = \mathcal{P}(L_d|A_d) \tag{3}$$

We formulate the calculation of $\mathcal{P}_{ex}$ and $\mathcal{P}_d$ as:

$$\mathcal{P} = \frac{1}{N}\sum_{i=0}^{N} \mathcal{P}(\sum l_i > 0 | attr_i == 1) \tag{4}$$

Where l_i and $attr_i$ are from the binary predicted results after thresholding. Since $\mathcal{P}_{ex} \in [0, 1]$ and $\mathcal{P}_d \in [0, 1]$, in order to minimize $\mathcal{P}_{ex}$ and maximize $\mathcal{P}_d$, the LCPLoss is as:

$$\mathcal{L}_{LCP} = ||\alpha\mathcal{P}_{ex} + \beta(1 - \mathcal{P}_d)||^2 \tag{5}$$

Where α and β are the coefficients to balance the ratio of $\mathcal{P}_{ex}$ and $\mathcal{P}_d$, we choose $\alpha = 1$ and $\beta = 24$. The final loss function is the combination of the BCELoss 1 and the LCPLoss 5:

$$\mathcal{L}_{total} = (1 - \lambda)\mathcal{L}_{BCE} + \lambda\mathcal{L}_{LCP} \tag{6}$$

Where λ is the coefficient to adjust the weights of the loss, and $\lambda = 0.5$ is our choice.

5.2. Label Compensation

The proposed LCPLoss is a solution to the impossible predictions, but it cannot handle the incomplete predictions. Hence, we propose the label compensation strategy which chooses the attribute that has the maximum confidence value in the incomplete portion as the positive prediction. For example, if none of the attributes that are related to beard area $[Clean\_Shaven = 0.3, Chin\_Area = -2, Side\_to\_Side = 0.1, Beard\_Area\_Info\_Not\_Vis = -1.5]$ has the confidence higher than the threshold value 0.5, then the attribute that has the highest confidence value among these attributes $Clean\_Shaven$ is the positive prediction. This strategy can eliminate all the incomplete predictions but increases the number of impossible predictions. In order to reduce this negative effect, we implement the label compensation strategy during both training and testing process. Code 1 and Code 2 in the Supplementary Material show the part of the training and testing code.

6. Experiments

In this section, we train a facial hair attribute classifier with FH37K, and evaluate accuracy and logical consistency. We propose LCPLoss, and combine it with a label compensation strategy to improve the performance of logically consistent predictions on a subset of WebFace260M. We analyze accuracy of ArcFace [17, 20] and MagFace [38] across demographics in two in-the-wild datasets.

6.1. Facial hair attribute classifier

We train facial hair attribute classifiers with the ResNet50 [24] backbone, both from scratch (BCELoss and LCPLoss only) and with pretrained ImageNet [41] weights for transfer learning (all methods). We resize images to 224×224 and use random horizontal flip for augmentation. Batch size is 256 and the learning rate is 0.001.

We evaluate model performance both without considering the logical consistency, as traditionally done in face attribute research, and also with logical consistency. We compare the baseline BCEloss, two loss functions - Binary Focal (BF) Loss [35], BCE-MOON [40] - that handle the imbalanced dataset problem, and the proposed LCPLoss.

Table 2 shows that, **before considering the logical consistency on predictions**, BCE-MOON outperforms the other methods on predicting positive labels. The proposed method has the best overall accuracy 90.78% on average. However, **after considering the logical consistency on predictions**, the accuracy of previous methods drops significantly, 43.26% decrease on average. The accuracy of the proposed method decreases from 90.90% to 89.89% for transfer learning training strategy, and from 89.11% to 88.30% for training from scratch.

To further investigate the effect of LCPLoss, we use the label compensation strategy to complete those incomplete portions of the predictions of BCELoss scratch, BCELoss transfer learning, BCE-MOON, and BF, the accuracy increases to 87.47%, 88.83%, 49.39%, 88.10% respectively. It reflects that labeling images in a logically consistent way can guide the model learning to a consistent pattern on-the-fly. In addition, the methods for handling the imbalanced data could make a high-accuracy illusion without considering the logical consistency on prediction, e.g. the accuracy

model training	ACC_{avg}	ACC^n_{avg}	ACC^p_{avg}
Not considering logical consistency ...			
BCE	88.82	93.72	54.97
BCE*	90.22	94.72	63.73
BCE-MOON*	88.96	90.67	**81.75**
BF*	89.84	95.43	58.41
BCE + LCP	88.90	95.55	46.13
BCE + LCP*	90.63	95.87	58.15
BCE + LCP + LC	89.11	95.06	52.17
BCE + LCP + LC*	**90.90**	**95.98**	63.30
Considering logical consistency ...			
BCE	45.10	46.02	32.62
BCE*	53.29	54.59	42.40
BCE-MOON*	46.46	47.54	32.95
BF*	39.96	40.95	31.45
BCE + LCP	27.66	28.19	18.80
BCE + LCP*	42.86	43.70	33.67
Label compensation on test ...			
BCE + LC	87.47	90.08	61.55
BCE + LC*	88.83	91.49	68.78
BCE-MOON + LC*	49.39	50.55	34.62
BF + LC*	88.10	90.91	66.05
BCE + LCP + LC	87.82	90.37	59.05
BCE + LCP + LC*	89.46	92.02	66.71
Label compensation on train and test ...			
BCE + LCP + LC	88.30	91.10	62.44
BCE + LCP + LC*	**89.89**	**92.65**	**70.23**

Table 2. Accuracy of models trained with different strategies. ACC_{avg} is the average accuracy for all attributes, ACC^p_{avg} on positive samples, ACC^n_{avg} on the negative samples. LC is the label compensation strategy. $*$ means using the transfer learning.

model training	N_{inp}	N_{imp}	R_{failed}
BCE	331,870	1,038	55.13
BCE*	240,761	6,001	40.86
BCE-MOON*	31,512	313,044	57.05
BF*	339,136	1,295	56.37
BCE + LCP	470,806	117	77.98
BCE + LCP*	307,576	300	50.98
Label compensation on test ...			
BCE + LC	0	10,215	1.69
BCE + LC*	0	11,134	1.84
BCE-MOON + LC*	0	330,115	54.66
BF + LC*	0	14,007	2.32
BCE + LCP + LC	0	14,097	2.33
BCE + LCP + LC*	0	6,083	1.01
Label compensation on train and test ...			
BCE + LCP + LC	0	7,693	1.27
BCE + LCP + LC*	0	5,595	**0.93**

Table 3. Results of logically consistent prediction test on a subset of WebFace260M which has 603,910 images. LC is the label compensation strategy. $*$ means using transfer learning. N_{inp} is the number of the incomplete predictions. N_{imp} is the number of the impossible predictions. R_{failed} is the ratio of the failed cases.

of BCE-MOON decreases from 81.75% to 34.62% on positive side. The performance of our model on each attribute is in Table 3 of the Supplementary Material.

To show the importance of logically consistent prediction of the model, we use the images of the first 30,000 identities in the sub-folder 0 from the WebFace260M dataset as a test set. Table 3 shows that, on average, 52.35% of the predictions generated by the BCE, BCE-MOON, and BF methods are logically inconsistent. After adding the label compensation strategy, the failure rates decrease dramatically. The proposed LCPLoss has the lowest fail rate 0.93%. Note that, more incomplete predictions will reduce the number of the impossible predictions, so comparison should consider these two numbers together rather than separately.

These results show that adding LCPLoss and label compensation strategy can significantly increase the usability of the model in real-world cases while improving accuracy.

6.2. Annotations and Recognition Accuracy

Experiments presented in this section show the potential value of accurate facial hair annotations in adaptive thresholding for recognition accuracy. ArcFace and MagFace are used to extract the feature vectors. Previous work [2,3,5–7,9,28,48,52] shows that biases exist across gender, age, and race. In order to reduce the impact of these factors, the BUPT-Balancedface (BUPT-B) [51] and BA-test datasets are used. BUPT-B has 1.3M images from Asian (A), Black (B), Indian (I), White (W). Each ethnicity has

Demographic	CS/%	CA/%	S2S/%	Total
Asian Male	21,374 / 75.17	6,336 / 22.28	726 / 2.55	28,436
	15,378 / 84.38	1,922 / 10.54	925 / 5.08	18,225
Black Male	5,529 / 21.63	11,784 / 46.10	8,247 / 32.27	25,560
	3,539 / 33.64	3,322 / 31.58	3,658 / 34.78	10,519
Indian Male	12,697 / 51.07	3,368 / 13.55	8,795 / 35.38	24,860
	7,654 / 48.47	1,781 / 11.28	6,356 / 40.25	15,791
White Male	13,265 / 54.7	4,823 / 19.89	6,162 / 25.41	24,250
	25,980 / 63.47	3,668 / 8.96	11,287 / 27.57	40,935

Table 4. High-confidence ($\geq$ 0.9) beard area predictions for BUPT-B (top number) and BA-test (bottom number) images, broken out by prediction and demographic.

7,000 identities. Since it does not have gender information, we use FairFace [27] to predict the gender for each identity. BA-test is a bias-aware test set we assembled on VGGFace2 [13], which has 665,562 face images from 8,870 identities. It groups the people into A, B, I, W, and by gender (M,F). Images from BA-test are samples from VGGFace2 [13] with the head pose, image quality, brightness balanced. The gender and ethnicity labels are predicted by FairFace. The images in these two datasets are cropped and aligned by using img2pose [4].

For each matcher, we compute the impostor distribution separately for each demographic, and then select the threshold for a 1-in-10,000 FMR for the Caucasian male demographic as the threshold for all demographics. This follows the NIST report on demographic effects in face recognition accuracy [19]. Also, this method makes the cross demographic differences in FMR more readily apparent.

Facial hair is a male characteristic in general. To investigate how beard area affects accuracy across demographic groups, we first select images with Clean Shaven (CS), Chin Area (CA), or Side to Side (S2S) beard area, using 0.9 as the threshold to pick the high-confidence samples. There are six categories of image pairs based on beard area: (CA,CA), (CA,CS), (CA,S2S), (CS,CS), (CS,S2S), and (S2S,S2S). The number of image pairs varies greatly across facial hair categories and demographic. The number of images selected from each demographic group is in Table 4.

Figure 3 (and Figures 1, 2, 3 of Supplementary Material) shows the impostor and genuine distributions of WM, BM, IM, and AM from BA-test and BUPT-B. As a general conclusion for both matchers and both datasets, beard

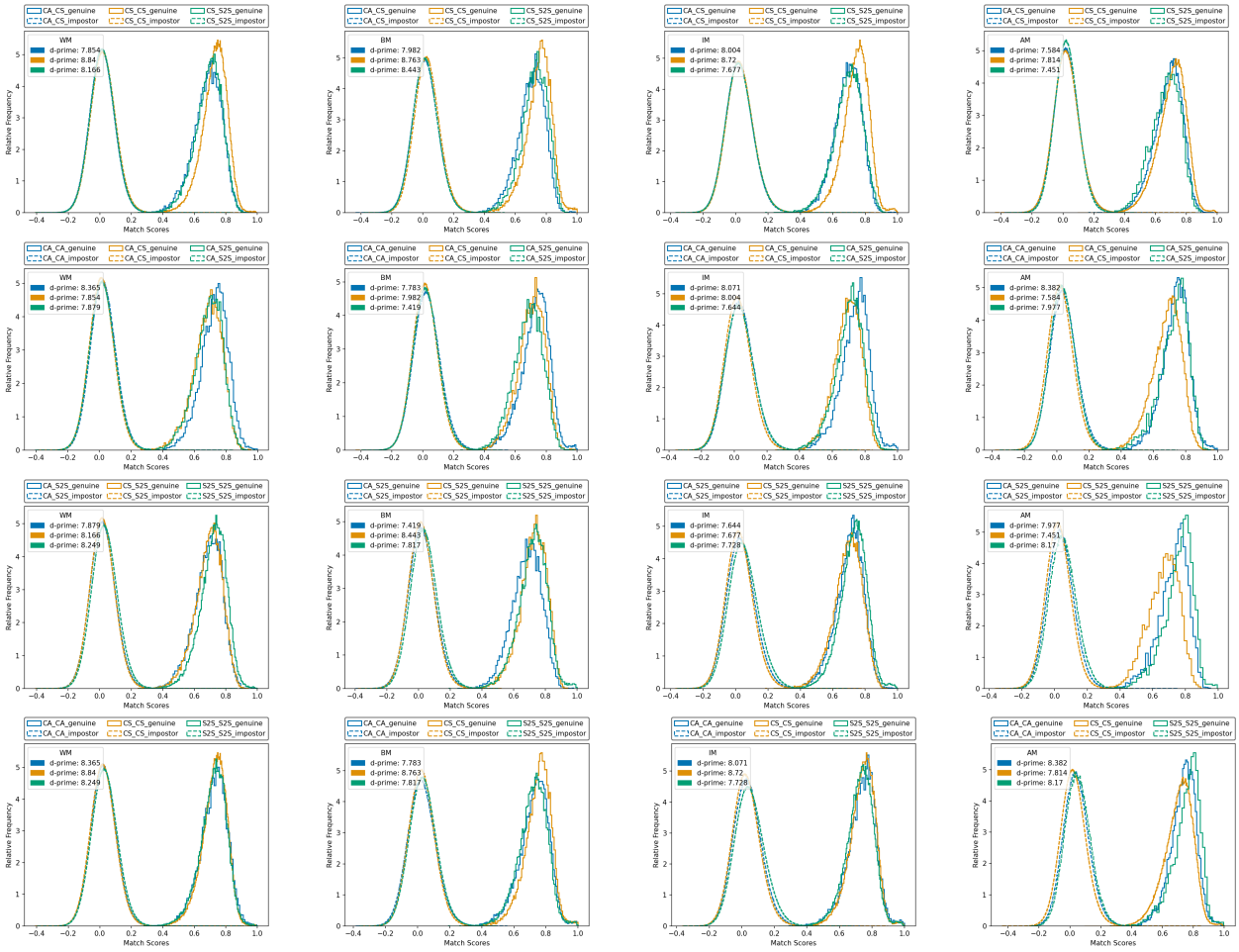

Figure 3. Facial hair attribute based genuine and impostor distributions for WM, BM, IM, AM in BA-test dataset. First row is CS focused plots, second is CA focused plots, third is S2S focused, and last row is same-beard-area focused plots. The feature extractor is MagFace.

area has more effect on the genuine distribution than the impostor. Image pairs with larger difference in beard area have lower similarity, and image pairs with the same beard area attribute have higher similarity. For instance, in the CS focused plots, (CS,CS) has highest similarity and (CS,S2S) has lowest similarity. For the image pairs that have the same beard area, the matchers perform differently for WM, BM, IM, and IM across the datasets. (CS,CS) has highest similarity in general. However, (CA,CA) and (S2S,S2S) have a larger difference on WM and AM than BM and IM in the BUPT-B dataset for both matchers. (S2S,S2S) has highest similarity and (CS,CS) has lowest similarity for AM in BA-test dataset for both matchers.

On the impostor side, the difference is not visually obvious in Figure 3 (and Figures 1, 2, 3 of Supplementary Material), so we compare false match rate (FMR) to study the effect, shown in Table 5 and Table 4 of Supplementary Material. In general, images pairs with the same beard area have the highest similarity, and image pairs in the having beard vs. clean shaven pattern have the lowest similarity. For instance, the pattern of FMR of WM is: (CA,CA) $>$ (CA,S2S) $>$ (CA,CS); (CS,CS) $>$ (CS,CA) $>$ (CS,S2S); (S2S,S2S) $>$ (CA,S2S) $>$ (CS,S2S). It is interesting that, for AM, (CA,CS) has higher FMR than (CA,S2S) and, for IM, (CA,S2S) has higher FMR than (CA,CA) in BA-test dataset. For the impostor pairs that have the same beard area, (CA,CA) $>$ (S2S, S2S) $>$ (CS,CS) is the pattern for AM, (S2S,S2S) $>$ (CS,CS) $>$ (CA,CA) is the pattern for IM, (S2S,S2S) $>$ (CA,CA) $>$ (CS,CS) is the pattern for both BM and WM. It is interesting that beard area causes different trends across demographics. Explaining these phenomena is one of our future works.

The analyses above indicates that: (1) the fraction of each type of facial hair area varies largely across demographics; for AM, over 75% of images are clean shaven and less than 6% of them have side to side beard area, (2) image pairs with larger difference in beard area have lower similarity, and image pairs with the same beard area have

BA-test	N_{pairs}	AM	N_{pairs}	IM	N_{pairs}	BM	N_{pairs}	WM
(CA,CA)	1,833,490	0.0558	1,575,614	0.0567	5,492,657	0.0571	6,716,047	0.0142
		0.0892		0.0849		0.0742		0.0176
(CA,CS)	29,508,374	0.0307	13,610,147	0.0368	11,743,810	0.0238	95,253,307	0.0073
		0.0435		0.0467		0.0298		0.0101
(CA,S2S)	1,773,388	0.0277	11,290,041	0.0745	12,133,642	0.0418	41,378,314	0.0116
		0.0447		0.1111		0.0557		0.0165
(CS,CS)	117,910,749	0.043	29,190,869	0.0603	6,238,358	0.0309	337,299,241	0.012
		0.0544		0.0705		0.0378		0.0153
(CS,S2S)	14,217,616	0.0143	48,596,378	0.0434	12,932,207	0.0224	293,188,100	0.0071
		0.0207		0.0609		0.0291		0.0111
(S2S,S2S)	424,161	0.0691	20,097,880	0.1455	6,663,160	0.0549	63,635,960	0.0219
		0.0861		0.2103		0.0749		0.0318

Table 5. False match rate and corresponding fraction of each beard area comparison group in BA-test. For the false match rate and fraction of each category, top number is ArcFace model, bottom is MagFace.

higher similarity, (3) matchers do not all have the same relative accuracy differences across demographics, and (4) the different demographics AM, IM, and WM do not follow the same relative accuracy differences across hairstyles. In particular, the fraction of images with facial hair varies greatly across demographics. We speculate the number of training samples of each facial hair area attribute are unbalanced and the beard length can cause this phenomenon.

7. Conclusions and Discussion

We introduce a more detailed scheme of facial hair description and create a dataset, FH37K, with these annotations. FH37K contains a threshold number of positive examples of as many of our new attributes as possible. The introduction of a fundamentally better dataset for exploring facial hair attributes is one contribution of this work.

We illustrate that the classifiers trained with the baseline BCELoss and the methods that handle the imbalance data have difficulty predicting logically consistent labels. As a novel approach to logical consistency in attribute learning, we introduce LCPLoss and a label compensation strategy to cause models to learn more logically consistent predictions and enforce consistency on predictions. To our best knowledge, this is the first work investigating the logical consistency on predictions in facial attribute area. Highlighting the issues of logical consistency across attributes and introducing an approach to solve them is another contribution of this work. Our approach is not specific to facial hair, and should be generally applicable in attribute prediction.

Using our attribute model trained on FH37K, we classify images from BUPT-B and BA-test datasets, and explore how recognition accuracy is affected by facial hair. One general conclusion is that image pairs with the same beard area attribute have, on average, a higher similarity score, for both impostor pairs and genuine pairs. (Two different persons look more alike to the face matcher when they have a similar beard area.) Similarly, image pairs with a larger difference in the beard area attribute have a lower similarity score. Interestingly, the pattern of change in similarity score for image pairs that are both clean-shaven, both chin-only or both side-to-side beards shows a different trend between Asian males, Indian males, Black males, and White males. This suggests that facial hairstyle plays a subtle causal role in the widely-commented-on demographic differences in face recognition accuracy. Additional factors beyond the effects of facial hair may be needed to better understand demographic accuracy differences.

Possibilities for future research include improving both accuracy and logical consistency of predictions, extending experiments on logical consistency of predictions to other multi-label classification tasks, investigating the effects of the other attributes of the facial hair on the face recognition accuracy, and exploring the explanation of demographic differences in face recognition accuracy.

8. Acknowledgement

Thanks to Dr. Terrance Boult, Dr. Manuel Günther, Dr. Emily M Hand, Dr. Wes Robbins, and teachers from RET program - Cara Storer, Nur Islam, John Gensic, Jonathan Woodard, Jill McNabnay, and Rebekah Spencer. This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via *2022-21102100003*. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] Amazon: Amazon mechanical turk. https://www.mturk.com/. 2

[2] Salem Hamed Abdurrahim, Salina Abdul Samad, and Aqilah Baseri Huddin. Review on the effects of age, gender, and race demographics on automatic face recognition. *Visual Computer*, 34(11):1617–1630, 2018. 6

[3] Vítor Albiero, Kevin Bowyer, Kushal Vangara, and Michael King. Does face recognition accuracy get better with age? deep face matchers say no. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 261–269, 2020. 6

[4] Vítor Albiero, Xingyu Chen, Xi Yin, Guan Pang, and Tal Hassner. img2pose: Face alignment and detection via 6dof, face pose estimation. In *CVPR*, 2021. 6

[5] Vítor Albiero, Kai Zhang, and Kevin W Bowyer. How does gender balance in training data affect face recognition accuracy? In *IJCB*, pages 1–10. IEEE, 2020. 6

[6] Vítor Albiero, Kai Zhang, Michael C. King, and Kevin W. Bowyer. Gendered differences in face recognition accuracy explained by hairstyles, makeup, and facial morphology. *IEEE Transactions on Information Forensics and Security*, 17:127–137, 2022. 6

[7] Vítor Albiero, Kai Zhang, Michael C. King, and Kevin W. Bowyer. Gendered differences in face recognition accuracy explained by hairstyles, makeup, and facial morphology. *IEEE Transactions on Information Forensics and Security*, 17:127–137, 2022. 6

[8] Thomas Berg and Peter N Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *CVPR*, pages 955–962, 2013. 1

[9] Aman Bhatta, Vítor Albiero, Kevin W Bowyer, and Michael C King. The gender gap in face recognition accuracy is a hairy problem. *arXiv preprint arXiv:2206.04867*, 2022. 6

[10] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Describing people: A poselet-based approach to attribute classification. In *ICCV*, pages 1543–1550. IEEE, 2011. 2

[11] Lubomir Bourdev and Jitendra Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, pages 1365–1372. IEEE, 2009. 2

[12] Jiajiong Cao, Yingming Li, and Zhongfei Zhang. Partially shared multi-task convolutional neural network with local constraint for face attribute learning. In *CVPR*, pages 4290–4299, 2018. 4

[13] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 2, 6

[14] Jui-Shan Chan, Gee-Sern Jison Hsu, Hung-Cheng Shie, and Yan-Xiang Chen. Face recognition by facial attribute assisted network. In *ICIP*, pages 3825–3829. IEEE, 2017. 1

[15] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, pages 8789–8797, 2018. 1

[16] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, pages 8188–8197, 2020. 1

[17] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 5

[18] Hui Ding, Hao Zhou, Shaohua Zhou, and Rama Chellappa. A deep cascade network for unaligned face attribute classification. In *AAAI*, volume 32, 2018. 4

[19] Patrick Grother, Mei Ngan, and Kayee Hanaoka. NISTIR 8280: Ongoing face recognition vendor test (frvt) part 3: Demographic effects. Technical report. 6

[20] Jia Guo. Insightface: 2D and 3D face analysis project. https://github.com/deepinsight/insightface, last accessed on February 2021. 5

[21] Hu Han, Anil K Jain, Fang Wang, Shiguang Shan, and Xilin Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *PAMI*, 40(11):2597–2609, 2017. 4

[22] Emily Hand, Carlos Castillo, and Rama Chellappa. Doing the best we can with what we have: Multi-label balancing with selective learning for attribute prediction. In *AAAI*, volume 32, 2018. 2

[23] Emily M Hand, Carlos D Castillo, and Rama Chellappa. Predicting facial attributes in video using temporal coherence and motion-attention. In *WACV*, pages 84–92. IEEE, 2018. 2

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[25] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE transactions on image processing*, 28(11):5464–5478, 2019. 1

[26] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 2

[27] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021. 6

[28] 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. *IEEE Transactions on Technology and Society*, 1(1):8–20, 2020. 6

[29] Neeraj Kumar, Peter Belhumeur, and Shree Nayar. Facetracer: A search engine for large collections of images with faces. In *ECCV*, pages 340–353. Springer, 2008. 2

[30] Neeraj Kumar, Alexander Berg, Peter N Belhumeur, and Shree Nayar. Describable visual attributes for face verification and image search. *PAMI*, 33(10):1962–1977, 2011. 1, 2

[31] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In *ICCV*, pages 365–372. IEEE, 2009. 1

[32] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In *ICCV*, pages 365–372. IEEE, 2009. 2

[33] Daiqing Li, Junlin Yang, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In *CVPR*, pages 8300–8311, 2021. 1

[34] Yan Li, Ruiping Wang, Haomiao Liu, Huajie Jiang, Shiguang Shan, and Xilin Chen. Two birds, one stone: Jointly learning binary code for large-scale face image retrieval and attributes prediction. In *ICCV*, pages 3819–3827, 2015. 1

[35] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 5

[36] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *CVPR*, pages 3730–3738, 2015. 2, 3

[37] Ohil K Manyam, Neeraj Kumar, Peter Belhumeur, and David Kriegman. Two faces are better than one: Face recognition in group photographs. In *IJCB*, pages 1–8. IEEE, 2011. 1

[38] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. MagFace: A universal representation for face recognition and quality assessment. In *CVPR*, 2021. 5

[39] Hung M Nguyen, Ngoc Q Ly, and Trang TT Phung. Large-scale face image retrieval system at attribute level based on facial attribute ontology and deep neuron network. In *Asian conference on intelligent information and database systems*, pages 539–549. Springer, 2018. 1

[40] Ethan M Rudd, Manuel Günther, and Terrance E Boult. Moon: A mixed objective optimization network for the recognition of facial attributes. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 19–35. Springer, 2016. 5

[41] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 5

[42] Zhiyuan Shi, Timothy M Hospedales, and Tao Xiang. Transferring a semantic representation for person re-identification and search. In *CVPR*, pages 4184–4193, 2015. 1

[43] Fengyi Song, Xiaoyang Tan, and Songcan Chen. Exploiting relationship between attributes for improved face verification. *Computer Vision and Image Understanding*, 122:143–154, 2014. 1

[44] Chi Su, Fan Yang, Shiliang Zhang, Qi Tian, Larry Steven Davis, and Wen Gao. Multi-task learning with low rank attribute embedding for multi-camera person re-identification. *PAMI*, 40(5):1167–1181, 2017. 1

[45] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. In *ECCV*, pages 475–491. Springer, 2016. 1

[46] Fariborz Taherkhani, Ali Dabouei, Sobhan Soleymani, Jeremy Dawson, and Nasser M Nasrabadi. Tasks structure regularization in multi-task learning for improving facial attribute prediction. *arXiv preprint arXiv:2108.04353*, 2021. 4

[47] Philipp Terhörst, Daniel Fährmann, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Maadface: a massively annotated attribute dataset for face images. *IEEE Transactions on Information Forensics and Security*, 16:3942–3957, 2021. 2

[48] 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. *IEEE Transactions on Technology and Society*, 3(1):16–30, 2021. 6

[49] Jing Wang, Yu Cheng, and Rogerio Schmidt Feris. Walk and learn: Facial attribute representation learning from egocentric video and contextual data. In *CVPR*, pages 2295–2304, 2016. 2

[50] Jing Wang, Yu Cheng, and Rogerio Schmidt Feris. Walk and learn: Facial attribute representation learning from egocentric video and contextual data. In *CVPR*, pages 2295–2304, 2016. 2

[51] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *ICCV*, October 2019. 6

[52] Haiyu Wu, Vítor Albiero, KS Krishnapriya, Michael C King, and Kevin W Bowyer. Face recognition accuracy across demographics: Shining a light into the problem. *arXiv preprint arXiv:2206.01881*, 2022. 6

[53] Haiyu Wu, Grace Bezold, Manuel Günther, Terrance Boult, Michael C King, and Kevin W Bowyer. Consistency and accuracy of celeba attribute values. *arXiv preprint arXiv:2210.07356*, 2022. 3

[54] Ning Zhang, Manohar Paluri, Marc'Aurelio Ranzato, Trevor Darrell, and Lubomir Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *CVPR*, pages 1637–1644, 2014. 2

[55] Xin Zheng, Yanqing Guo, Huaibo Huang, Yi Li, and Ran He. A survey of deep facial attribute analysis. *IJCV*, 128(8):2002–2034, 2020. 1

[56] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. *arXiv preprint arXiv:2207.12393*, 2022. 2

[57] 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 *CVPR*, pages 10492–10502, 2021. 3