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 facial hair attribute dataset, FH37K, with more descriptive facial hair annotations. 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 binary cross-entropy 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, using "in-the-wild" and "controlled" datasets. Results show that similarity and difference in facial hairstyle have important effects on the impostor and genuine score distributions in face recognition. Our trained facial hair attribute model and the new FH37K dataset will be made available after acceptance.

1. Introduction

Facial attributes have been widely used in face matching/recognition [8, 13, 26, 27, 32, 41], face image retrieval [30, 34], re-identification [40, 42, 43], training GANs [14, 15, 22, 29] 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, which do 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:

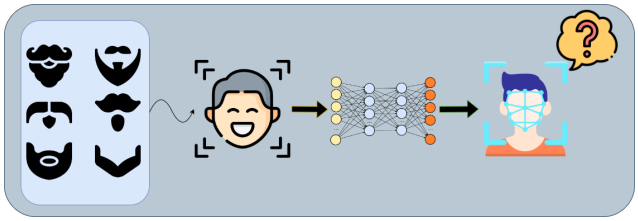


Figure 1. (1) *What is the best way to define the facial hair styles?* (2) *How does the facial hair classifier perform in the real-world cases?* (3) *How does the face matcher treat the same (different) person with different (same) beard styles?* This paper presents our approaches and answers for these 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 dimenisons of the area of the face covered, the length of the hiar and the connectedness of elements of the hair (See Section 2 and Section 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 oppositely across demographic groups when image pairs have the same beard area. (See Section 6.2)

	# of images	# of ids	# of facial hair attributes	Area	Length	CNDN	E_{in}
Berkeley Human Attributes [10]⋆	8,053	-	0	0	0	0	✗
Attributes 25K [48]	24,963	24,963	0	0	0	0	✗
FaceTracer [25]⋆	15,000	15,000	1 (Mustache)	0	0	0	✗
Ego-Humans [44]	2,714	-	1 (Facial hair)	0	0	0	✗
CelebA [31]⋆	202,599	10,177	5 (5 o'Clock, Goatee, ...)	1	1	0	✗
LFWA [31]⋆	13,233	5,749	5 (5 o'Clock, Goatee, ...)	1	1	0	✗
PubFig [28]⋆	58,797	200	5 (5 o'Clock, Goatee, ...)	1	1	0	✗
LFW [23]⋆	13,233	5,749	5 (5 o'Clock, Goatee, ...)	1	1	0	✗
UMD-AED [19]	2,800	-	5 (5 o'Clock, Goatee, ...)	1	1	0	✗
YouTube Faces Dataset (with attribute labels [20])	3,425	1,595	5 (5 o'Clock, Goatee, ...)	1	1	0	✗
CelebV-HQ [50]⋆	35,666 video clips	15,653	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_{in} stand for connectedness and estimating the inconsistency 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.

2. Facial Hair In Face Attribute Datasets

For a broad discussion of face attribute classification research, see the recent survey by Zheng et al [49]. Here, we 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 [45] 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 [48] 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, 48] includes any attribute to describe facial hair.

Kumar et al [25] 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 [44] 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 [25,44] each have only a single binary attribute related to facial hair.

The Labeled Faces in the Wild [23] (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 [28] collected 65 attributes through Mechanical Turk [1] and added 8 more [26] for a total of 73 attributes. Kumar et al [28] 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 [31] 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) [19] 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 [20] collect 3,425 frames from the original YouTube Faces Dataset. They also use the same 40 attributes. A recent facial attributes related dataset [50] contains 35,666 high quality video clips. There are 83 manually labeled facial attributes covering appearance, action, and emotion. All of these datasets have the same five binary attributes related to facial hair: No_beard, Mustache, Goatee, Sideburns and 5 O'clock Shadow. Therefore a richer description of facial hairstyle is urgently needed.

Our FH37K dataset used in this paper includes attributes to describe dimensions of facial hair related to area of the face, length of the hair and connectedness of the parts of facial hair. Some previous face attribute datasets ([19], [20], [23], [31], [50]) include one area-specific attribute (Goatee) and one length-specific attribute (5 o'clock shadow). PubFig [28] has one area-specific attribute (Goatee). None

of [25], [44], [10], and [48] provide information on area or length. No previous dataset has any attribute describing the connectedness among elements of facial hair. And no previous dataset provides an estimate of rater consistency in assignment of the annotations. This comparison of facial hair attributes across datasets is summarized in Table 1.

3. Overview of the FH37K Dataset

3.1. Dataset statistics

Our FH37K dataset contains 37,565 images, coming from a subset of CelebA [31] and a subset of the WebFace260M [51]. There are 5,216 identities spanning a wide range of demographics (3,318 identities from CelebA and 1,898 from WebFace260M).

All the images are manually annotated with respect to a detailed definition for each annotation, and examples and strategies for marking challenging images. 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.

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.

3.2. Dimensions of facial hair properties

Previous face attribute datasets generally have simple binary annotations for facial hair. FH37K has a larger and richer set of facial hair attributes that can be grouped into three dimensions.

- **Beard Area**: The three levels of beard area are *Clean Shaven* (no beard), *Chin Area* (beard limited to chin area) and *Side to Side* (beard extending to sides of face). (Examples in Figure 3 of Supplementary Material.)
- **Beard Length**: The five levels of length are *Clean Shaven*, *5 O'clock Shadow*, *Short*, *Medium* and *Long*. The No beard is the same as Clean Shaven, so this option is not in the 22 attributes. (The *Clean Shaven* attribute can be seen as an element of description for both area and length.) (Examples in Figure 4 of Supplementary Material.)
- **Mustache**: Mustache-related values are *Mustache-None*, *Mustache Isolated* (mustache not connected to beard) and *Mustache Connected to Beard*. Note that mustache can be considered as an area of facial hair separate from or connected to beard. Examples can be found in Figure 5 of the Supplementary Material.
- **Sideburns**: Similar to mustache, sideburns-related attribute values are *Sideburns-None*, *Sideburns-Present* (not connected to beard) and*Sideburns Connected to Beard*. (Examples in Figure 6 of Supplementary Material.)
- **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*. (Examples in Figure 7 of Supplementary Material.)
- **Information is not visible**: With in-the-wild imagery, it is common that information is simply not visible in the image to assign a value for some attribute. 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*).

These 22 attributes can cover 29 real-world cases. More details are in the Section 4.1 and the number of positive samples for each attribute can be found in Table 1 of the Supplementary Material.

4. FH37K Data collection

Images in FH37K are cropped and aligned versions from the CelebA dataset and WebFace260M dataset. Images were selected from CelebA as follows. Images with distributed CelebA annotations of No_beard=false were manually reviewed for possible inclusion in FH37K and new annotations. There are 253 images with a No_beard=false annotation actually did not have face and were dropped from FH37K. CelebA images kept for FH37K were manually annotated by one of multiple annotators. Annotators read a document containing definitions and examples of the FH37K annotations before annotating and were encouraged to refer to the document as needed during annotation. At the end of this step, the dataset 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 number of initially under-represented positive examples. The 4,274 images selected from the WebFace260M dataset in this way resulted in addition of a enough images that all attributes except bald only on sides and long beard are represented by at least 1,000 images. Selected images from WebFace260M were then manually assigned attribute values in the same way as for CelebA. The final result is the FH37K collection of 37,565 images with an aggregate total of 0.8M annotations.

(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 8 of the Supplemental Material.

4.1. Complications for Consistent Annotation

Consistent annotations that align the content of the images with the concept to be learned is an important element of any machine learning datset. 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. The subjective term "chin area" can have slightly different meaning for different annotators, which hurts the consistency of the Beard Area annotations. We gave the annotators the specific definition that the chin area is within parallel vertical lines extending from the outer eye corners, as shown in Figure 2a. The images in Figure 2b show that the beard length is not a constant on one face and can vary over the area of the beard. We defined this for the annotators as picking the longest length when there are varying lengths over the area of the beard. Head pose, occlusion, and lighting angle varies broadly in any in-the-wild dataset, which causes the complications illustrated in Figure 2c and Figure 2d. A single attribute is not sufficient to describe these circumstances, and so we use the visible part plus Info Not Vis attribute to describe these images.

To evaluate the robustness of the our annotations, a fresh annotator independently annotated a random set of 1,000 images from FH37K. This annotator had not done any of the FH37K annotations, but had the same training documentation as the original annotators. This annotator's results were compared to the FH37K annotations to get an estimate of how consistently a new annotator would agree with the FH37K annotations. The estimated inconsistency rate is 5.95%. (Analysis is in Supplementary Material).

4.2. Backward-Compatible FH37K Annotations

The more descriptive facial hair annotations in FH37K can be mapped back to values for the simpler set of five binary attributes used in most previous research (as in Table 1). For example, the FH37K "Chin_Area=True" can be converted to the "Goatee=True" and "No_Beard=False" in CelebA and other datasets. This backward-compatible version of the FH37K annotations is also made available, for researchers who may want a larger and cleaner version of facial hair attributes than is currently available.

5. Logically Consistent Prediction

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 $X \in R^{D\times H\times W}$ as the D-dimension batch input and $Y \in R^{D\times K}$ 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:

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

However, due to the sparse ground truth labels on the multi-label classification tasks, the BCELoss guides the model to be biased towards to predict negative labels, which increases the accuracy on the benchmarks but reduces real-world usability. Our approach is to force the model considering the logic relationships - *mutually exclusive*, *collectively exhaustive*, *dependency* - among attributes. Examples of these in FH37K are as follows:

- *mutually exclusive*: Only one of a group of attribute values can be true, otherwise the predictions are impossible. For example, beard area = clean shaven and beard length = short cannot be both true for the same image.
- *dependency*: If attribute A is true, the attribute B must be true, otherwise the predictions are impossible. The sideburns is connected to the beard means the beard area must be side-to-side.
- *collectively exhaustive*: One of a group of attributes must be true, otherwise the predictions are incomplete. One of the beard area values (Clean Shaven, Chin Area, Side to Side, Info Not Vis) must be true.

Based on these three relationships, we propose the LCPLoss to force the model considering the logical consistency when doing the predictions.

For the *mutually exclusive* relation, 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 happen at the same time is:

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

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.

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

Since $\mathcal{P}_{ex} = \mathcal{P}(L_{ex}|A_{ex})P(A_{ex})$, we can formulate the calculation of $\mathcal{P}_{ex}$ and $\mathcal{P}_d$ as the follows:

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

Where l_i and $attr_i$ are from the binary predicted results. 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:

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

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.

5.1. 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 rise 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 the model's accuracy and logical consistency. We propose LCPLoss, and combine it with a label compensation strategy to improve the performance on the subset of WebFace260M. We analyze accuracy of ArcFace [16, 18] and MagFace [33] across demographics in both "in-the-wild" and "controlled" datasets.

6.1. Facial hair attribute classifier

We train facial hair attribute classifiers with a ResNet50 [21] backbone, both from scratch and with pretrained ImageNet [39] weights for transfer learning. 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. The top half of the Table 2 shows that directly comparing the predicted labels with the ground truth labels, the models trained only with BCELoss have slightly higher average accuracy than the models trained using LCPLoss and label compensation, 90.23 compared to 89.74. However, to reduce disagreement between predicted labels and ground truth labels, BCELoss guides the model to predict more to the negative side than the positive. The lower half of the Table 2 shows that when the logically inconsistent predictions are considered as failed/wrong, accuracy of the model trained only with the BCELoss drops significantly from 90.23 to 53.39 for the transfer learning model, and from 88.83 to 45.16 for the model trained from the scratch. With LCPLoss and the label compensation strategy, accuracy decreases by 7.74 for the transfer learning model, and by 12.58 for the model trained from the scratch. The results prove that using LCPLoss + label compensation strategy produces annotations that are overall more logically consistent. To further investigate the effect of LCPLoss, we use the label compensation strategy to complete those incomplete portions of the predictions of the BCE-only models, and LCPLoss still gives a 5% improvement on average (bottom two rows of Table 2). The performance of our model on each attribute is in Table 2 of the Supplementary Material.

To show the importance of logically consistent prediction of the model, we use all the 608,184 images in the subfolder 0 from the WebFace260M dataset as a test set. Table 3 shows that there is a huge number of logically inconsistent predictions generated by the BCE-only models. By adding the LCPLoss and label compensation strategy, the failed rate decreases from 54.94% to 0.45% for the model trained from the scratch and from 40.73% to 0.47% for the transfer learning model. Note that, more incomplete predictions will reduce the number of the impossible predictions, thus the comparison should consider these two numbers to-

	ACC_{avg}	ACC_n	ACC_p
BCE	88.83	93.73	54.98
BCE + LCPLoss + LC	88.51	92.22	59.93
BCE*	**90.23**	**94.72**	63.73
BCE + LCPLoss + LC*	89.74	92.68	**68.66**
BCE$^{\lozenge}$	45.16	46.09	32.67
BCE + LC$^{\lozenge}$	75.58	78.02	52.63
BCE + LCPLoss + LC$^{\lozenge}$	75.93	78.55	52.94
BCE*$^{\lozenge}$	53.39	54.68	42.48
BCE + LC*$^{\lozenge}$	77.23	79.71	59.32
BCE + LCPLoss + LC*$^{\lozenge}$	**82.00**	**84.49**	**62.90**

Table 2. Performances of the models trained and tested with different strategies. ACC_{avg} is the average accuracy of positive rate and negative rate. ACC_p is the accuracy on the positive samples. ACC_n is the accuracy on the negative samples. LC is the label compensation strategy. $*$ means using the pretrained model. $\lozenge$ means considering the logical consistency of the predictions.

	N_{inp}	N_{imp}	R_{failed}
BCE	333,773	586	54.98
BCE + LC	0	1,871	0.31
BCE + LCPLoss + LC	0	2,735	0.45
BCE*	247,702	5,423	40.73
BCE + LC*	0	8,407	1.38
BCE + LCPLoss + LC*	0	2,869	0.47

Table 3. Results of logically consistent prediction test on a subset of WebFace260M which has 608,184 images. LC is the label compensation strategy. $*$ means using the pretrained model. 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.

gether rather than separately.

These experiment results show that adding LCPLoss and label compensation strategy can significantly increase the usability of the model in the real-world cases while having a better effect on accuracy.

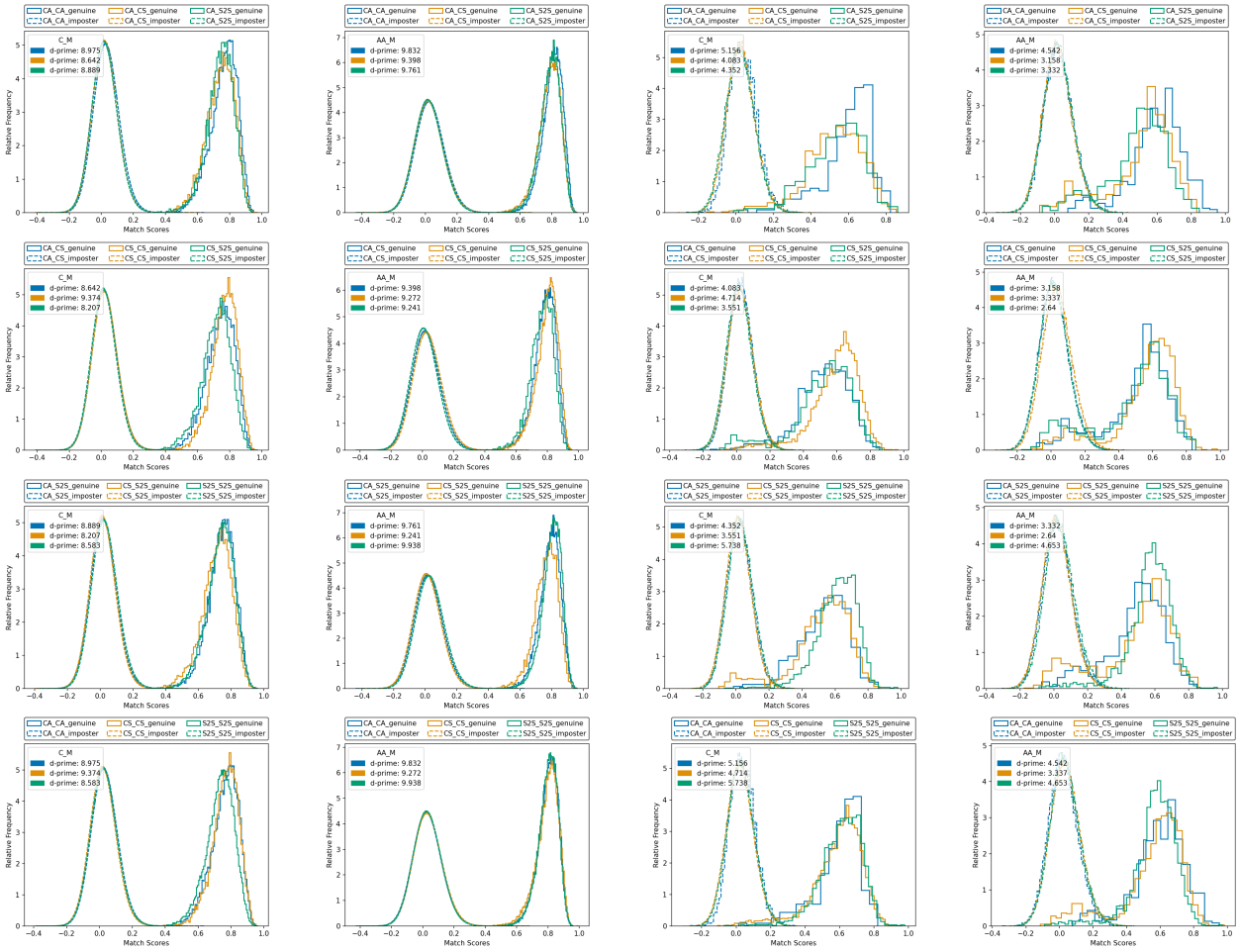
6.2. Annotations and Recognition Accuracy

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. The MORPH [35, 36] dataset to represent the controlled imaging scenario. and Balanced Faces In The Wild (BFW) [37] dataset for the in-the-wild scenario. Images in MORPH are acquired in conditions typical of mugshot, passport or ID-card photos, including nominally frontal pose, neutral expression, consistent indoor lighting and plain gray background. MORPH was assembled from public records, and is widely used in face aging [38] and in study of demographic accuracy variation [2, 3, 5–7, 9, 24, 46]. The version of MORPH used contains 127,319 images: 56,245 images of 8,839 African-American males, 24,857 images of 5,929 African-American females, 35,276 images of 8,835 Caucasian males, and 10,941 images of 2,798 Caucasian females. Faces were detected and aligned using img2pose [4]. BFW has 20,000 face images from 800 identities. Each identity has 25 images. It groups the people into Asian (A), Black (AA), Indian (I), White (C) and by gender (M,F). Images in BFW are sampled from VGGFace2 [12] and the faces are detected by MTCNN [47], cropped based on the bounding box and aligned based on the predefined eye locations.

The pattern of recognition accuracy results across beard area attribute and demographics is similar for the both matchers on both datasets. For each matcher, we compute the impostor distribution separately for each demographic, and then select the threshold corresponding to a 1-in-10,000 FMR for the Caucasian male demographic as the threshold for all demographics. This follows the recent NIST report on demographic effects in face recognition accuracy [17]. 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 the recognition accuracy across demographic groups, we first pick the 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, especially for Indian male and Asian male in the BFW dataset. Therefore, our analysis is based on the results of African-American male (AAM) and Caucasian male (CM) in the main paper. The number of images picked from each demographic group can be found in Table 5.

Figure 3 shows the impostor and genuine distributions of CM and AAM. As a general conclusion for both matchers and for both datasets, beard area has more effect on the genuine distribution than the impostor. To ensure the conclusion is statistically meaningful, we mainly focus on the MORPH dataset for genuine analysis. 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 plot, (CS,CS) has highest similarity and (CS,S2S) has lowest similarity. The BFW dataset plots show a similar result. For image pairs that have the same beard area, the matchers performs differently for AAM and CM. For AAM, (S2S,S2S) has lowest similarity, (CA,CA) is in the middle but close to (CS,CS), and (CS,CS) has highest similarity. However, for CM, (CS,CS) has lowest similarity value, and (CA,CA) and (S2S,S2S) are the same. The same conclusions can be found in the MagFace results (Figure 1 of Supp. Material).



(a) MORPH dataset (b) BFW dataset

Figure 3. Facial hair attribute based genuine and imposter distributions for ArcFace. First row is CA focused plots, second is CS focused plots, third is S2S focused, and last row is same-beard-area focused plots.

On the impostor side, the difference is not visually obvious in Figure 3, so we compare false match rate (FMR) to study the effect. Categories with less than 100k image pairs are in gray in Table 4, to indicate lower statistical support. For AAM and CM in both BFW and MORPH datasets, the pattern of FMR is: (CS,CS) $>$ (CS,CA) $>$ (CS,S2S); (CA,CA) $>$ (CA,CS) $>$ (CA,S2S); (S2S,S2S) $>$ (S2S,CA) $>$ (S2S,CS). However, the pattern of impostor pairs that have the same beard area is opposite between AAM and CM. For CM, (S2S,S2S) $>$ (CA,CA) $>$ (CS,CS), but for AAM, the order is (CS,CS) $>$ (CA,CA) $>$ (S2S,S2S). Explaining this phenomenon is one of our future works.

General conclusions from this analysis are as follows. One, for both AAM and CM, image pairs with larger difference in beard area have lower similarity on both genuine and impostor distributions, and image pairs with the same beard area have higher similarity on both genuine and impostor distributions. Two, among the three cases of image pairs with the same beard area, the matchers perform oppositely for AAM and CM. (The analysis on AM and IM is in last paragraph of the Supplementary Material.)

Note that facial hairstyle has significant differences across demographics and across datasets. For example, clean-shaven facial hairstyle is 4 or 5 times more common in MORPH images than in BFW images for AAM and CM. Also, clean-shaven is 2 to 3 times more frequent in Caucasian images than in African-American images. Clean-shaven has the dominant fraction (85%) for AM than for the other demographic groups, but side-to-side barely occurs for AM. IM has close fractions of clean-shaven and side-to-side style, but chin-area is not frequent in this group. Since facial hairstyle can cause changes in face recognition accuracy, this presents issues for datasets that are supposed to be "balanced" and for cross-demographic accuracy comparisons.

	BFW				MORPH			
	N_{pairs} / %	AAM	N_{pairs} / %	CM	N_{pairs} / %	AAM	N_{pairs} / %	CM
(CA,CA)	21,844	0.0549	5,865	0.0171	284,230,811	0.0451	56,848,421	0.0148
	/ 3.1	0.1511	/ 0.4	0.0512	/ 25.3	0.0592	/ 12.2	0.0161
(CA,CS)	102,156	0.0343	130,163	0.0238	228,312,883	0.0387	93,950,415	0.0096
	/ 14.6	0.1273	/ 9.3	0.0446	/ 20.3	0.0444	/ 20.2	0.0111
(CA,S2S)	103,109	0.0456	42,078	0.0095	334,344,517	0.0282	117,589,903	0.0084
	/ 14.7	0.1009	/ 3.0	0.0238	/ 29.7	0.0405	/ 25.3	0.0103
(CS,CS)	116,398	0.1237	699,936	0.0121	45,822,278	0.0626	38,800,481	0.0120
	/ 16.7	0.2466	/49.7	0.0376	/ 4.1	0.0615	/ 8.3	0.0131
(CS,S2S)	238,170	0.0348	456,503	0.0090	134,263,355	0.0205	97,147,390	0.0060
	/ 34.0	0.1155	/ 32.5	0.0212	/ 11.9	0.0275	/ 20.9	0.0080
(S2S,S2S)	118,203	0.0973	71,469	0.0084	98,262,602	0.0411	60,759,155	0.0151
	/ 16.9	0.2073	/ 5.1	0.0350	/ 8.7	0.0559	/ 13.1	0.0167

Table 4. False match rate and corresponding fraction of each beard area comparison group. For the false match rate and fraction of each category, top number is ArcFace model, bottom number is MagFace.

Dataset	Groups	CA	CS	S2S
BFW	AAM	212	487	492
	AM	121	1,715	16
	CM	111	1,191	385
	IM	71	812	843
MORPH	AAM	23,846	9,576	14,023
	CM	10,665	8,811	11,028

Table 5. The number of picked images for each beard area attribute from the BFW and MORPH dataset.

7. 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. Our annotations can be back-converted to a more accurate version of the binary beard/no-beard annotations for CelebA. The introduction of a fundamentally better dataset for exploring facial hair attributes is one contribution of this work.

Models trained with common BCELoss can have higher accuracy with logically inconsistent predictions. As a novel approach to the problem of 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.

Using our facial hair attribute model trained on FH37K, we classify images from two popular 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 image 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 the opposite trend for African-American males and Caucasian males. This suggests that facial hairstyle plays a subtle causal role in the widely-commented-on demographic differences in face recognition accuracy.

Future research includes improving the performance of the classifier on 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.

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. 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, 3

[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] 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. 6

[13] 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

[14] 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

[15] 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

[16] 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

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

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

[19] 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

[20] 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

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

[22] 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

[23] 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

[24] 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

[25] 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, 3

[26] 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

[27] 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

[28] 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

[29] 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

[30] 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

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

[32] 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

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

[34] 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

[35] University of North Carolina at Wilmington. Morph dataset. https://www.faceaginggroup.com/?page_id=1414. 6

[36] Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 341–345. IEEE, 2006. 6

[37] Joseph Robinson. Balanced faces in the wild, 2022. 6

[38] R. Rothe, R. Timofte, and L. Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *IJCV*, 126:144–157, 2018. 6

[39] 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

[40] 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

[41] 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

[42] 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

[43] 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

[44] 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, 3

[45] 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

[46] 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

[47] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016. 6

[48] 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, 3

[49] Xin Zheng, Yanqing Guo, Huaibo Huang, Yi Li, and Ran He. A survey of deep facial attribute analysis. *IJCV*, 128(8):2002–2034, 2020. 2

[50] 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

[51] 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