On the Preservation of Soft-Biometric Information in Biometric Face Embeddings

[1] ANANYA CHANDRAN , [2] MAMATHA RANI B M
[1]Assistant Professor, [2]Assiatant Professor
[1] Computer Science & Engineering, [2] Computer Science & Engineering
[1]Brindavan College Of Engineering, Bangalore, India
[1]ananya8625@gmail.com, [2]mamatha.ranibm47@gmail.com

Abstract : The development of deeply-learned features is the foundation for the success of contemporary face recognition systems. These embeddings are designed to encode a person's identity so that they can be used for identification. Recent studies, however, have demonstrated that these embeddings also store information on demographics, image qualities, and social traits in addition to the user's identity. This brings up issues with prejudice and privacy in facial recognition. We examine the predictive power of 73 various soft-biometric features on three well-liked face embeddings with various learning philosophies. The tests were run on two databases that were accessible to the general public. We developed a huge attribute classifier that can accurately express the confidence in its predictions as part of the evaluation process. As a result, we are able to construct more complex statements concerning the property predictability. The findings show that most of the attributes under investigation are encoded in face embeddings. For instance, a robust encoding for accessories, accessories, and accessories was discovered. We discovered that these characteristics are particularly easy to anticipate from face embeddings, despite the fact that face recognition embeddings are taught to be resilient against non-permanent elements. Our research is intended to inform future efforts to create better bias-reducing and privacy-preserving face recognition technology.

IndexTerms **- Face recognition, bias, fairness, soft-biometrics, analysis, privacy, biometrics**

I.INTRODUCTION

Current face recognition systems show strong recognition capabilities enabled by the advances in learning deep neural feature embeddings [13]. This leads to a worldwide spreading of these systems and also increasingly affect everyone's daily life [8]. Although face recognition models are trained with the aim of extracting deeply-learned features. This work was partially funded by the National Research Centre for Applied Cybersecurity (ATHENE), the Hessen State Ministry for Higher Education, Research, and the Arts (HMWK), the German Federal Ministry of Education and Research (BMBF), and the German Federal Ministry of Education and Research (BMBF) through the Software Campus Project. After considering the feedback from the reviewers, Associate Editor J. Phillips suggested that this article be published. (Philipp Terhörst is the corresponding author.)

Identification Number for Digital Resources 10.1109/TBIOM.2021.3093920. Recent research revealed that the information included in these embeddings goes beyond identification recognition. These studies shown that various face embeddings hold data on social qualities [38], head pose [37], demographics [9], viewpoint [19], lighting [35], and picture attributes (such as quality [4], [18], and illumination [19]). But this raises moral questions about privacy and fairness in face recognition. For starters, many programmes only allow users to access information relevant to recognition [32], and extracting further information without a person's agreement is deemed a violation of their privacy [24]. This is referred to as soft-biometric privacy [32], and solutions are either image-[30], [31], or embedding-level [5, 42, 45, 51]. Second, the qualities recorded in biometric face embeddings may suggest biassed performances associated with these attributes, which may result in unfair performance discrepancies.

This is referred to as facial recognition bias, and solutions to this problem have primarily focused on demographic-bias [12], [28], [48], [52], and [55]. Knowledge of encoded properties in face embeddings is essential to construct more advanced bias-mitigating systems [43]. We extend the work of [43] by performing a prediction analysis on 73 different soft-biometric variables derived from face embeddings. In [43], an attribute predictability statement is created by taking into account an attribute classifier's attribute prediction performance at two difficulty levels. These difficulty levels reflect how well this classifier predicts an attribute and hence simulates, for example, various capture scenarios. This work, in contrast to [43].

1) Analyses an attribute's predictability throughout a continuous range of difficulty levels. This enables more fine-grained predictability claims regarding specific attributes to be derived.
2) Analyses the predictability of numerous attributes (attribute categories) at the same time to produce precise, concise, and simply understandable results.
3) Analyses the predictability of numerous attributes (attribute categories) at the same time to produce precise, concise, and simply understandable results.
4) The studies are expanded to include three alternative facial recognition models. This enables researchers to investigate the impact of embedding dimensionalities and underlying training losses on the qualities stored in face embeddings.
5) Furthermore, it analyses the implications of our findings for future works.

The inquiry methodology is built on a massive attribute classifier (MAC) that is trained on numerous attributes at the same time in order to take advantage of a common feature space. The MAC is designed in such a way that it can accurately state its prediction confidence [47]. This enables us to make more specific claims regarding the predictability of attributes in face embeddings. The trials were carried out using two public datasets, CelebA [29] and LFW [20], as well as three popular face embeddings, FaceNet [40], CosFace [53], and ArcFace [10]. We classified each attribute into one of three predictability classes in order to provide intelligible comments regarding the recorded attribute information: easy-to-predict, predictable, and difficult-to-predict. The findings show that many qualities are encoded in face embeddings. 39 of the 113 qualities examined are classified as easily predictable, while the remaining 74 are predictable. We discovered disparities in attribute predictability based on the underlying training principles of face recognition networks. Age, hairstyles, haircolors, beards, and accessories, on the other hand, are substantially encoded in all embedding types, including FaceNet, CosFace, and ArcFace. Despite the notion that face embeddings are learnt to be resistant to non-permanent elements, the results show that these features, in particular, are easily foreseeable.

II. RELATED WORK:

The development of deep neural network representations helps face recognition [13]. However, because these representations are developed from black-box models, there is a need to better understand what kind of information is stored in them. Parde et al. [37] proved in 2017 that the examined representations contain precise information about the head position (i.e., the yaw and pitch of a face) and the image source (i.e., whether the input-face originates from a still image or a video frame). They speculated that information about image quality might be available in these facial renderings as well. This has been demonstrated to be valid because the quality of a facial picture has been effectively predicted using face embeddings [4], [18], and [49]. Parde et al. investigated how effectively information regarding social qualities is maintained in face representations in [38]. In their experiments, they used linear classifiers to predict human-assigned social trait profiles. They proved that 11 social qualities such as talkativeness, assertiveness, shyness, quietness, warmth, artisticity, efficiency, carelessness, impulsiveness, anxiety, and laziness can be inferred to a high degree from face embeddings. The attributes that were most accurately predicted were impulsive, warm, and nervous. Hill et al. [19] investigated caricature face representations. Their research involved categorising viewpoint (0, 20, 30, 45, 60), illumination (ambient vs spotlight), gender (male vs female), and identity in embedding space.

Their findings show that information regarding face identification and imaging properties coexist in a highly organised and hierarchical structure established by the face recognition algorithm used. O'Toole et al. provide an overview of their findings as well as a review of known aspects of the face space in the context of previous-generation face recognition algorithms. [35]. Zhong et al. conducted facial attribute estimate experiments utilising various mid-level representations from face recognition networks in [56], [57]. They got highly accurate facial attribute estimation findings by employing a variety of mid-level representations. This suggests that high-level representations, such as face recognition templates, may also contain a substantial amount of facial attribute information.

The ability to derive demographic variables such as gender, age, and race from face templates is proven in [6, 9], [36], and [47].

Previous research has shown that face templates can be used to derive information about demographic attributes (e.g., gender, age, race), social traits (e.g., impulsive, warm, and anxious), as well as head pose and image characteristics (e.g., quality, source of the image, viewpoint, illumination). These publications concentrated on the examination of distinct characteristics. Terhörst et al. [43] conducted a more in-depth examination into the predictability of over 100 features in face templates. They classified each attribute into one of three predictability groups based on prediction performance at two different dependability levels. Their findings show that face templates can reliably predict up to 74 variables. We extend the analysis of [43] in this paper by doing a more in-depth investigation of attribute predictabilities.

While the study in [43] is based on two difficulty levels, we extend the studies to three alternative embedding types and present experiments on a continuous difficulty range. This enables more precise predictability declarations for each attribute. We extend the analysis and discussion to higher-level attribute categories in order to get more precise but also more concise and accessible findings. Furthermore, we especially examine the significance of our findings for future works.

III. INVESTIGATION METHODOLOGY

The purpose of this research is to determine what attributes are stored in biometric face embeddings. This analysis is carried out by jointly training a classifier to accurately predict these features. If the classifier predicts an attribute correctly given the face embeddings, we can conclude that the attribute is encoded within the embedding. This inquiry methodology, however, only permits detecting which qualities are stored in embeddings. It does not allow us to deduce what traits are not encoded because a logical reverse conclusion is not always possible. If an estimator is unable to learn an attribute's pattern, this does not indicate that the pattern does not exist. The estimator may just be unable of dealing with the attribute pattern's complexity or the diversity and representation of the data may be low.

The three subsections that follow outline the various steps of our investigation approach.

1) We describe our classifier's training technique. Training the classifier in a multi-task manner enables the utilisation of shared embedding space, resulting in overall performance improvements.
2) We describe the methods utilised, which enables the trained classifier to accurately state its prediction confidence (reliability).
3) This idea of prediction reliability is used to introduce predictability classes. These make it easier to analyse the observations.

3.1 Massive Attribute Classifier (MAC)

A classification model is at the heart of our attribute predictability study of face embeddings. If this model predicts an attribute properly given face embeddings, we can conclude that the attribute is encoded in the embeddings. On the basis of face embeddings, we built a neural network model to jointly predict several features that could be kept within. Because of the enormous number of attributes that are simultaneously learned, we call this model a massive attribute classifier (MAC). Multiple random network architectures with 1-3 initial layers and 1-3 branch layers connecting the last initial layer with the softmax layers of each attribute were evaluated. Sizes of 128, 256, and 512 were considered for each layer. During this evaluation, we found that the expected performance per characteristic varied by only 1-2%. As a result, we chose the most straightforward network structure.

The chosen MAC-architecture is comprised of two initial layers: an input layer of size nin (representing the size of the employed face embedding) and a second dense layer of size 512. The architecture employs a common layer to boost the efficacy of associated factors such as age, gender, and race [14], [17]. While the multi-task MAC technique is well-suited for linked characteristics, training single classifiers to predict each of them separately may result in superior attribute prediction performances in some circumstances, as demonstrated in [29]. Beginning with the second layer, each attribute a gets its own branch composed of two further layers of size 512 and n(a) out, where n(a) out is the number of classes per attribute. A ReLU activation was employed for each layer. Only the output-layers that use softmax activations are exempt. Additionally, Batch-Normalization [21] and dropout [41] are applied to each layer. Using a dropout method allows for more generalised performance and, more crucially, allows us to generate confidence assertions about the prediction (discussed in Section III-B). A dependability statement's quality is resistant to varying degrees of dropout. As a result, we used the default dropout probability of pdrop = 0.5 [41]. The MAC-training was performed in a multi-task learning way, with a categorical cross-entropy loss applied to each attribute branch and an equal weighting between each of these attribute-related losses. The training was performed using an Adam optimizer [25] over e = 200 epochs with an initial learning rate of = 103 and a learning-rate decay of = /e.

These parameters were chosen based on the experiment setup of [47]. The batch size b was determined based on the amount of data available for training as b = 1024 for CelebA and b = 16 for LFW.

3.2 Prediction Reliability

To produce reliable predictions regarding attribute predictability in face embeddings, we employ prediction reliabilities to simulate classifier situations of varying difficulty. We train the MAC with dropout using the methods described in [46, 47]. This allows us to express the forecast confidence (reliability) of the MAC. In addition to an attribute prediction, we perform m = 100 stochastic forward passes to produce a reliability statement. Each forward pass employs a separate dropout-pattern, resulting in m distinct softmax outputs v(a) i for each attribute a. The dependability metric is given as x(a) = v(a) i,c given the outputs of the m stochastic forward passes of the anticipated class c.

$$rel(x^{(a)}) = \frac{1-\alpha}{m}\sum_{i=1}^{m} x_i^{(a)} - \frac{\alpha}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}|x_i^{(a)} - x_j^{(a)}|,$$

With = 0.5, as per the recommendation in [47]. The first portion of the equation is a centrality measure that employs the probability interpretation of the softmax output. A greater score indicates a high likelihood that the forecast is right. The second portion of the equation is a dispersion measure that quantifies the agreement of the stochastic outputs x. [47] shown that this is an accurate dependability metric. with = 0.5, as per the recommendation in [47]. The first portion of the equation is a centrality measure that employs the probability interpretation of the softmax output. A greater score indicates a high likelihood that the forecast is right. The second portion of the equation is a dispersion measure that quantifies the agreement of the stochastic outputs x. [47] shown that this is an accurate dependability metric.

Lower RCP-levels will reject more low-confidence predictions that may contain variance elements (such as blur and non-frontal head positions) that contribute to unstable, and thus erroneous, attribute assessments. As a result, a low RCP-level relates to the MAC prediction performance under more ideal classifier conditions. Please keep in mind that alternative predictability metrics can also be utilised for the suggested research. Alain and Bengio employed linear separability to assess the predictability of a categorical attribute in [1]. If a binary attribute is perfectly encoded in the face space, the amount of information about that attribute remains constant regardless of whether the decision boundary in the embedding space is linear or curved. Dahr et al. measured predictability using mutual information estimation in [11]. While this method does not rely on linear separability, it does necessitate the training of extra networks to evaluate predictability. For these reasons, we chose to quantify predictability in our studies using correct prediction reliabilities [47].

3.3 Predictability Classes

We categorise each attribute into one of three predictability classes to extract more intelligible assertions about which attribute information is stored in a face embedding. These are based on prediction performance at 50% and 100% RCP.

- **Easily-predictable (++)**: a property is considered easily-predictable if and only if its balanced accuracy at 100% RCP is greater than 90%. This means that highly accurate predictions are attainable even in less-than-ideal conditions such as poor lighting and non-frontal head postures.

- **Predictable (+):** an attribute is classified as predictable if and only if its balanced accuracy at 100% RCP is less than 90% but its balanced accuracy at 50% RCP is more than 90%. Because it simply takes into account 50% of the most confident MAC forecasts, this shows that highly accurate predictions are attainable under near-optimal conditions.
- **Hardly-predictable (0):** an attribute is classified as hardly-predictable if its balanced accuracy is less than 90% at both 100% and 50% RCP. Even under near-optimal conditions, the MAC cannot achieve high accuracies.cAs a result, the attribute patterns may be too complex for the MAC to manage, or there may not be a valid pattern for this attribute.

Because attribute categories are easily predictable and predictable, confident statements about the amount of attribute information stored in face embeddings may be made. This does not apply to the unpredictability. If an attribute is classified as hardly-predictable, the MAC is unable to learn the pattern accurately. This could be for a variety of reasons. First and foremost, the pattern does not exist. Second, while the pattern exists, it is too complex for the model to learn. Or, third, the pattern exists, but the amount of data and its representation are insufficient for the classifier to learn. As a result, we cannot identify whether a similar attribute pattern exists for qualities classified as hardly-predictable.

IV. EXPERIMENTAL SETUP

4.1 Databases

The Labelled Faces in the Wild (LFW) [20] and CelebFaces Attributes (CelebA) [29] datasets contain a significant number of attribute annotations and are thus well suited for our face space predictability study. An in-depth analysis of which of these features are encoded in face embeddings is undertaken using a variety of soft-biometric labels. Figure 1 depicts photos from both datasets. The CelebA dataset [29] has over 200k photos from over 10k different celebrities. Each image has 40 binary attributes added to it. There are also numerous variations in stance and surroundings. The LFW [20] dataset contains 13k photos from nearly 5k different individuals. Each image includes annotations for 73 binary attributes. Furthermore, the photographs vary greatly in terms of position, lighting, focus, resolution, facial expression, age, gender, race, accessories, make-up, occlusions, background, and photographic quality. The attribute labels of both databases [20], [29] contain a wide variety of features (for example, a person's demographics, complexion, hair, beard, facial geometry, periocular area, mouth, nose, accessories, and environment).

4.2 Cleaning Attribute Annotations of LFW

CelebA attribute annotations are binary in nature [29]. In contrast to CelebA, the LFW dataset's attribute annotations are continuous and measure the degree of the attribute present in the image [20], [26], [27]. A strong positive label score for the attribute beard, for example, should imply a notable beard, whereas a negative annotation value indicates that no beard is displayed. As a result, binary labels can be obtained by assigning true labels to positive label qualities and false labels to negative label attributes. A number around zero, on the other hand, implies that the presence of the property cannot be reliably confirmed.

We manually transformed the continuous attribute labels to binary labels to guarantee that the MAC performed well when trained on LFW. We awarded true labels to photographs with scores above the upper threshold and false labels to images with scores below the lower threshold, using an upper and lower score threshold for each attribute. Undefined attributes have scores that fall between the upper and lower score threshold limitations. The upper and lower criteria for a specific property are set manually by pushing potential thresholds away from zero. Ten photos with the closest attribute scores are investigated at each candidate level. By doing so, the photos' original LFW annotations are manually checked for accuracy. If just eight or fewer photographs show the presence of a specific attribute, the potential threshold is pushed away from the beginning point until a sufficient score threshold is found. If a prospective threshold produces 9 or more correctly identified photos, that threshold is used for that attribute. The lower and upper thresholds for each of the qualities are established by repeating this approach. The scores are then binarized using the upper and lower thresholds to ensure that the MAC's data is error-free. Training and testing can then be performed on meaningful and correctly tagged data.

Because LFW labels are often of low quality, our label-cleaning method reduces the quantity of used labels by 51,7%. This could lead to a prejudice in our evaluation. We assess another binary labelled database to avoid biassed findings that may result from this procedure.

TABLE 1
SAMPLE DISTRIBUTION ON LFW FOR Selected Attributes FOUND INSUFFICIENT FOR MEANINGFUL ATTRIBUTE ANALYSIS AFTER LABEL CLEANING. THE NUMBER OF POSITIVELY AND NEGATIVELY LABELLED SAMPLES FOR THE TRAIN AND TEST SET IS REFERRED TO BY POS AND NEG. DUE TO A LOW NUMBER OF SAMPLES IN EITHER THE POSITIVE OR NEGATIVE CLASS, THE LISTED 15 ATTRIBUTES WERE FOUND TO BE INSIGNIFICANT FOR THE ANALYSIS.

	Train		Test	
Attribute	Pos	Neg	Pos	Neg
Color Photo	8806	29	3772	24
Mouth Slightly Open	674	109	315	57
Round Face	9	588	3	250
Goatee	20	3346	10	1557
Baby	23	9137	15	3913
Bangs	89	5238	44	2080
Bald	114	4413	47	1953
Big Lips	101	751	48	318
Sunglasses	74	8583	50	3631
Partially Visible F.	124	1501	55	601
Mouth Wide Open	107	6593	56	2925
Double Chin	154	172	57	136
Harsh Lighting	113	914	62	487
Outdoor	173	510	63	243
Teeth Not Visible	125	2209	66	1089

4.3 Evaluation Matrix

Our face space predictability analysis is based on the MAC's prediction performance. The accuracy metric is commonly used to calculate the prediction performance of facial features. However, because this metric is defined by the ratio of correct predictions to total number of predictions [33], it is heavily influenced by unbalanced label distribution. As a result, in order to be robust to attribute imbalances, we report prediction performance in terms of balanced accuracy. The standard accuracy with class-balanced sample weights is referred to as balanced accuracy [23].

The datasets are subject-exclusively partitioned into train and test data in a 70%/30% split.1 We decided against using a cross-database evaluation technique because both databases have more than 30 non-overlapping properties. The loss of this crucial characteristic information would arise from training on one database and evaluating on the other. The prediction performance of a facial attribute estimator is examined under a variety of demanding conditions. This is accomplished, as mentioned in Section III-B, by testing prediction performance at various RCP-levels. While high RCP-levels model more realistic scenarios, low RCP-levels focus on reliable predictions and so simulate more idealistic conditions. This is used to provide finer-grained comments concerning attribute predictability.

4.4 Face Template Extraction

For the studies, we employ three widely used face recognition models based on the losses of FaceNet [40], CosFace [53], and ArcFace [10]. FaceNet, CosFace,3 and ArcFace are pre-trained models that we use in this work.4 FaceNet and ArcFace are based on a ResNet100 model that was trained on the MS1M database [16]. CosFace is made up of a ResNet50 model that was trained on CASIA-WebFace [54]. The facial photos are preprocessed (aligned, resized, and cropped) before being used as input for the models. [22] describes the preprocessing for FaceNet, [53] describes the preparation for CosFace, and [15] describes the preprocessing for ArcFace. Face recognition models are used to extract the embeddings from the preprocessed facial photos.

4.5 Investigations

The goal of this research is to determine what information is preserved in biometric face embeddings. To accomplish this, we give an in-depth investigation organised into the following sections.

1) We investigate the relationships between attribute annotations. The results of an attribute may be highly predictable, which is due to associated annotations in the testing database rather than attribute information stored within an embedding.
2) We study which attributes are stored in face embeddings in two steps by analysing attribute prediction performances.

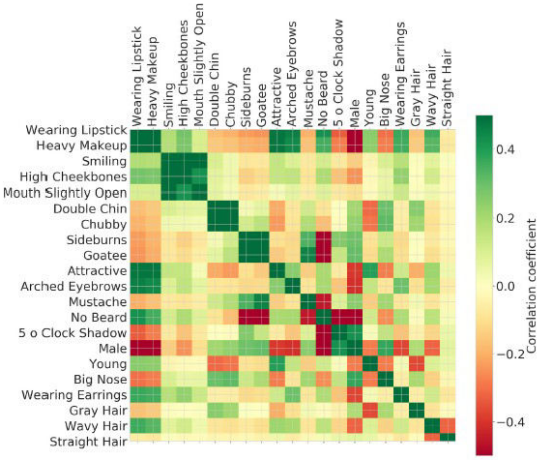
1Please note that attributes affected by imbalanced data training will be associated with a poorer prediction performance due to the use of balanced accuracies. Consequently, the imbalanced data training might lead to underestimating the amount of information stored in face embeddings for some attributes.

To begin, we examine the prediction performance of each attribute on two distinct confidence levels of the MAC to gain an understanding of the situation. Second, we study the prediction performance of each characteristic throughout a wide and continuous range of confidence levels in order to conduct a more in-depth investigation of the stored data.

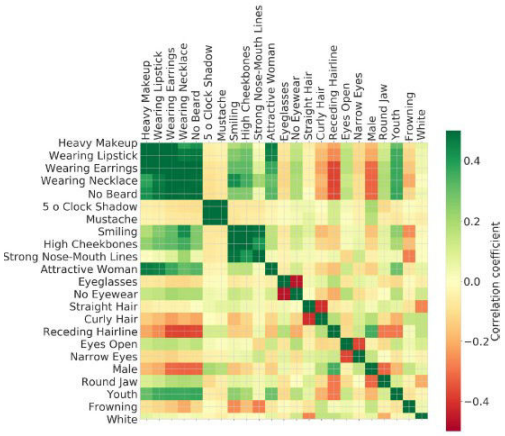
3) We compromise the extensive investigations in order to gain a clear picture of what kind of information is encoded in face embeddings. First, based on two-level prediction performance, we classify each attribute into one of three predictability classes. Second, we visualise the predictability of each group of attributes to give the reader a visual representation of which qualities are stored in face embeddings and how easily these may be anticipated.

V. RESULT

This section works on the defined investigation points in accordance with the investigation strategy from Section IV-E. Section V-A examines the attribute correlations of the used face datasets, Section V-B investigates attribute predictability in depth, and Section V-C summarises the findings qualitatively and quantitatively. Section V-D concludes by discussing the implications of our findings for future work.



(a) CelebA



(b) LFW

Fig. 2. Correlation of the attribute-annotations for CelebA and LFW. The attributes are chosen to show the 15 most positive and negative pairwise correlations.The attribute-correlation for LFW is shown after the label-cleaning process. Green indicate positive correlations, while red indicate a negative correlation. The correlation is based on the Pearson coefficient

5.1 Attribute-Correlation Analysis

To prevent making inaccurate statements about which attributes are stored in face embeddings, we first examine the attribute annotation correlation. This seeks to determine general label quality while avoiding any biases in attribute annotations. Figure 2 depicts some attribute-label correlations for CelebA and LFW. The traits have been chosen to demonstrate the 15 most positive and negative pairwise correlations. Figure 2(a) depicts the attribute correlations for CelebA annotations. It is clear that masculine faces do not correlate well with wearing lipstick, earrings, or heavy make-up. These characteristics are nearly entirely associated with female faces. Furthermore, a high proportion of male faces have a beard, whereas women do not.

Attractive faces are generally young and female faces sporting accessories and heavy make-up. Furthermore, the figure endorses the quality of several labels. For example, the trait No Beard has a negative link with all sorts of beards, including goatees, moustaches, and sideburns. Figure 2(b) depicts the pairwise correlations of LFW attribute labels. Heavy Makeup, Wearing Lipsticks, Wearing Earrings, and Wearing Necklace are traits that go well with Youth, Attractive Woman, Smiling, and High Cheekbones. These characteristics, on the other hand, are unrelated to Receding Hairline and Male. The correlation matrix in Figure 2(b), like the one in Figure 2(a), can be utilised to validate the label quality of specific antagonistic qualities. For example, No Eyewear correlates poorly with Eyeglasses, and Curly Hair correlates negatively with Straight Hair. Because these attribute correlations can influence the predictability investigation in Section V-B, we examined the annotation correlation and the relevant attribute prediction performance as well. Table II displays the results of an analysis of the ten highest-correlating attribute pairings for CelebA and LFW. Given the properties a and b, $\rho(a, b)$ is the Pearson correlation coefficient $a \rightarrow b$ refers to the balanced accuracy while utilising attribute a's label as the prediction for attribute b and vice versa. The accessories Wearing Lipstick, Wearing Earrings, and Heavy Makeup have the highest associations. These characteristics also have the best prediction accuracy. If an attribute an is highly predictable from face embeddings and has an accurate correlation to attribute b (a b > 90%), it cannot be properly discriminated whether both or only one of them is encoded in the face embedding. As a result, these relationships must be taken into account in the subsequent assessment.

5.2 Attribute-Analysis of the Face Space

The attribute prediction performance of the MAC is used to determine which attributes are encoded in face embeddings. This is accomplished in two levels of detail. To begin, the prediction performance of the qualities is determined at two difficulty levels to provide context. 100% RCP (hard) refers to using all samples under the stated conditions. The term 50% RCP (easy) refers to the 50% of predictions about which the classifier is most confident. Second, the performance of each attribute in terms of prediction is examined throughout a large and continuous range of confidence levels. Table III displays the two-level prediction performance of CelebA, including the assigned predictability classes.

Figures 3, 4, and 5 illustrate the prediction performance of all analysed facial embeddings, FaceNet (FN), CosFace (CF), and ArcFace (AF), for the continuous RCP range of [0.5, 1].

Two observations can be made in general.

First, lower RCP-level prediction performance is generally better than higher RCP-level prediction performance.

TABLE II

ANNOTATION CORRELATION AND CORRESPONDING ATTRIBUTE PREDICTION PERFORMANCE ANALYSIS: $a \rightarrow b$ REFERS TO THE BALANCED ACCURACY WHEN USING ATTRIBUTE a'S LABEL AS THE PREDICTION FOR ATTRIBUTE b AND VICE VERSA.

THE PEARSON COEFFICIENT PROVIDES THE CORRELATION. ON BOTH DATABASES, CELEBA AND LFW, THE 10 HIGHEST CORRELATED ATTRIBUTES ARE INVESTIGATED. ONLY A FEW ATTRIBUTE CORRELATIONS HAVE STRONG EFFECTS ON PREDICTION PERFORMANCE.

			$\rho(a,b)$	Accuracy	
	Attribute a	Attribute b		$a \mapsto b$	$b \mapsto a$
CelebA	Wearing Lipstick	Heavy Makeup	0.80	91.1%	89.1%
	Smiling	High Cheekbones	0.68	84.3%	84.0%
	Smiling	Mouth Open	0.54	76.8%	76.8%
	Double Chin	Chubby	0.53	74.2%	79.5%
	Sideburns	Goatee	0.51	74.4%	76.9%
	Wearing Lipstick	Attractive	0.48	74.0%	74.0%
	Wearing Lipstick	Arched Eyebrows	0.46	76.0%	70.4%
	Goatee	Mustache	0.45	77.4%	68.5%
	Wearing Lipstick	No Beard	0.42	78.2%	65.6%
	5 o Clock Shadow	Male	0.42	63.3%	82.8%
LFW	Heavy Makeup	Wearing Lipstick	0.64	77.7%	87.5%
	Wearing Lipstick	Wearing Earrings	0.60	71.1%	92.0%
	Wearing Earrings	Wearing Necklace	0.57	74.8%	83.2%
	5 o Clock Shadow	Mustache	0.55	85.0%	71.3%
	Smiling	High Cheekbones	0.54	86.5%	69.6%
	Heavy Makeup	Wearing Earrings	0.51	65.4%	91.4%
	Wearing Necklace	No Beard	0.49	75.4%	73.6%
	Strong No.-Mou. Lines	Smiling	0.48	76.8%	71.9%
	Heavy Makeup	Attractive Woman	0.46	79.8%	68.1%
	Wearing Lipstick	Attractive Woman	0.45	83.6%	65.1%

This proves that the MAC learned to anticipate the CelebA dataset reliably. Second, the prediction performance of FaceNet and CosFace is always slightly higher than that of ArcFace. The reason for this could be ArcFace's big angular margin approach, which distorts the feature space more incoherently and thus makes it more difficult for estimators to learn existing patterns. The embedding size, on the other hand, appears to have less of an effect on predictability, as the lowest and largest embeddings (FaceNet-128, CosFace-1024) both yield higher predictabilities than ArcFace (512). To summarise, numerous CelebA features obtain good prediction accuracy on all three face recognition models. This contains demographics, haircuts, haircolors, and beard kinds.

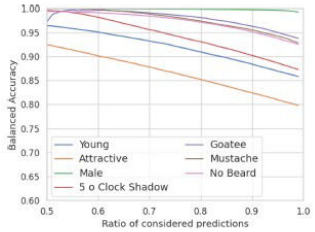
Furthermore, the person's accessories are encoded in great detail in the deeply-learned features. Table IV shows the two-level prediction performance of LFW, including the assigned predictability classes. The grey highlights denote results that have limited validity. Because the label-cleaning operation removed several samplesusing low-quality attribute annotations. The small number of the lack of prediction may be explained by the use of training and testing samples.

Baby, Sunglasses, and other qualities work well. Mouth. The prediction performance is depicted in Figures 3, 4, and 5. @ the [0.5, 1] continuous RCP range for all three embedding types. A lower RCP-level indicates greater confidence in the classifier and, as a result, greater balanced accuracy of the projected attribute.

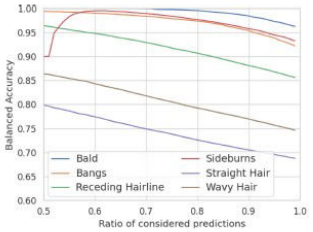
TABLE III
CELEBA PREDICTION PERFORMANCE IS BASED ON FACENET (FN), COSFACE (CN), AND ARCFACE (AF) EMBEDDINGS AND IS REPORTED IN TERMS OF BALANCED ACCURACIES AT TWO DIFFICULTY SCENARIOS: 100% RCP (HARD) AND 50% RCP (EASY). THE ASSIGNED PREDICTABILITY CLASS IS DEFINED BY ++,+, AND 0

		100% RCP			50% RCP		
	Attribute	FN	CF	AF	FN	CF	AF
Demo	Male^{++}	98.9%	97.1%	98.4%	99.9%	99.9%	99.9%
	Young^{+}	85.5%	82.7%	83.6%	96.4%	94.3%	94.5%
Skin	Pale Skin^{+}	76.0%	77.5%	71.9%	87.1%	90.0%	83.0%
	Rosy Cheeks^{+}	83.4%	85.8%	78.2%	96.3%	94.9%	81.7%
Hairstyle	Bald^{++}	95.7%	95.4%	94.0%	100.0%	100.0%	100.0%
	Bangs^{++}	91.7%	92.5%	89.3%	99.4%	99.6%	98.3%
	Receding Hairline^{+}	85.4%	84.6%	82.5%	96.4%	96.3%	94.2%
	Sideburns^{++}	92.8%	91.7%	92.1%	90.0%	96.2%	99.7%
	Straight Hair^{0}	68.6%	69.0%	70.7%	79.9%	80.0%	82.0%
	Wavy Hair^{0}	74.4%	74.5%	76.6%	86.4%	86.7%	89.4%
Haircolor	Black Hair^{+}	83.7%	84.5%	81.5%	96.6%	97.0%	94.3%
	Blond Hair^{++}	91.9%	91.8%	90.1%	99.3%	99.4%	98.3%
	Brown Hair^{+}	76.5%	78.2%	75.9%	90.1%	91.2%	88.3%
	Gray Hair^{++}	93.0%	92.9%	91.1%	99.6%	99.4%	98.8%
Beard	5 o Clock Shadow^{+}	86.9%	85.6%	85.8%	99.6%	99.2%	99.0%
	Goatee^{++}	93.4%	90.8%	91.8%	97.2%	100.0%	98.9%
	Moustache^{++}	92.2%	87.9%	89.7%	100.0%	94.4%	98.8%
	No Beard^{++}	92.1%	89.8%	90.8%	99.4%	99.4%	99.0%
Face Geo.	Chubby^{+}	86.5%	86.2%	83.1%	96.5%	97.4%	95.4%
	Double Chin^{+}	86.6%	87.4%	82.9%	96.9%	98.7%	95.4%
	High Cheekbones^{+}	78.5%	82.7%	72.2%	91.6%	95.0%	82.6%
	Oval Face^{0}	63.4%	64.6%	61.9%	70.8%	72.3%	68.1%
Periocular	Arched Eyebrows^{+}	79.8%	80.1%	77.0%	93.3%	93.6%	89.5%
	Bags Under Eyes^{0}	72.1%	74.6%	70.7%	80.6%	84.3%	80.7%
	Bushy Eyebrows^{+}	83.4%	83.1%	78.5%	95.9%	95.5%	91.9%
	Narrow Eyes^{0}	66.5%	70.2%	60.7%	75.4%	80.0%	66.7%
Mouth	Big Lips^{0}	74.6%	71.5%	68.8%	86.4%	83.7%	78.7%
	Mouth Slightly Open^{+}	74.5%	82.9%	67.5%	86.5%	95.5%	76.5%
	Smiling^{+}	80.1%	86.7%	71.7%	92.9%	97.7%	82.1%
Nose	Pointy Nose^{0}	71.7%	70.8%	69.3%	83.1%	83.1%	78.9%
	Big Nose^{0}	77.4%	76.7%	75.8%	88.1%	86.7%	87.1%
Accessories	Eyeglasses^{++}	97.3%	94.0%	90.6%	99.8%	99.7%	98.7%
	Heavy Makeup^{++}	90.1%	90.5%	88.7%	99.2%	99.5%	98.5%
	Wearing Earrings^{+}	79.2%	78.8%	77.0%	94.8%	93.6%	91.6%
	Wearing Hat^{++}	95.4%	95.1%	92.8%	99.4%	99.3%	99.0%
	Wearing Lipstick^{++}	92.8%	92.7%	91.4%	99.4%	99.7%	98.7%
	Wearing Necklace^{0}	71.8%	71.9%	71.4%	86.9%	86.5%	84.2%

A counteracting behaviour for low RCP-level is noticed for a few traits, such as Sideburns. These could be explained by the ground truth's low annotation quality [44]. When the results of LFW are compared to the results of CelebA, it is clear that similar prediction performances are obtained on attributes found in both datasets. As a result, our label-cleaning approach removed low-quality attribute labels while causing no substantial bias in the data.



(a) Demographics and Beard



(b) Hairstyle

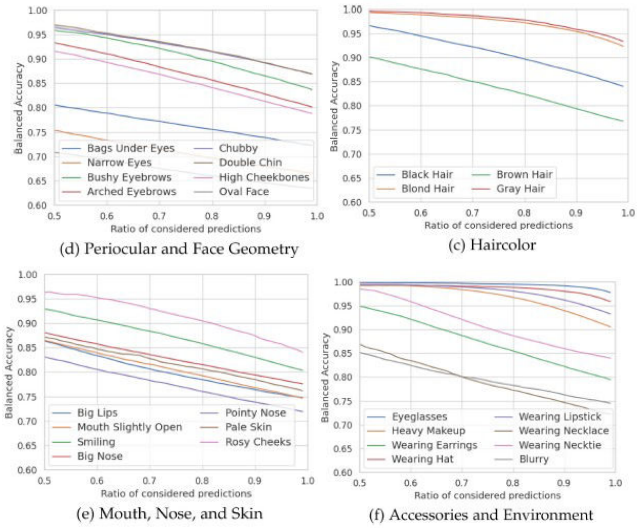


Fig. 3. Accuracy-Reliability plots for the CelebA database on FaceNet embeddings. The balanced accuracy of the MAC is shown for a continuous RCP range of [0.5, 1]. The MAC performance of the 40 attributes is divided into 6 categories represented by subfigures (a)-(f) to allow a simple category-base analysis.

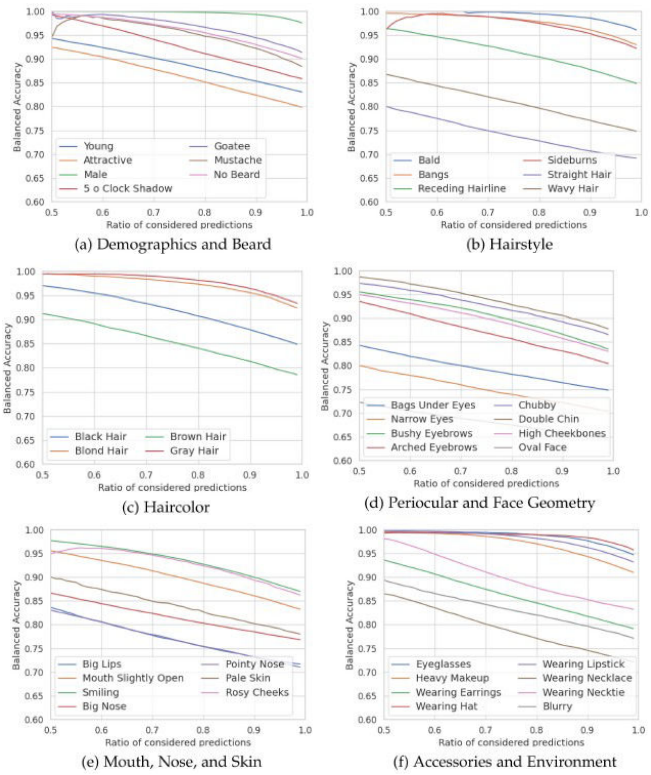
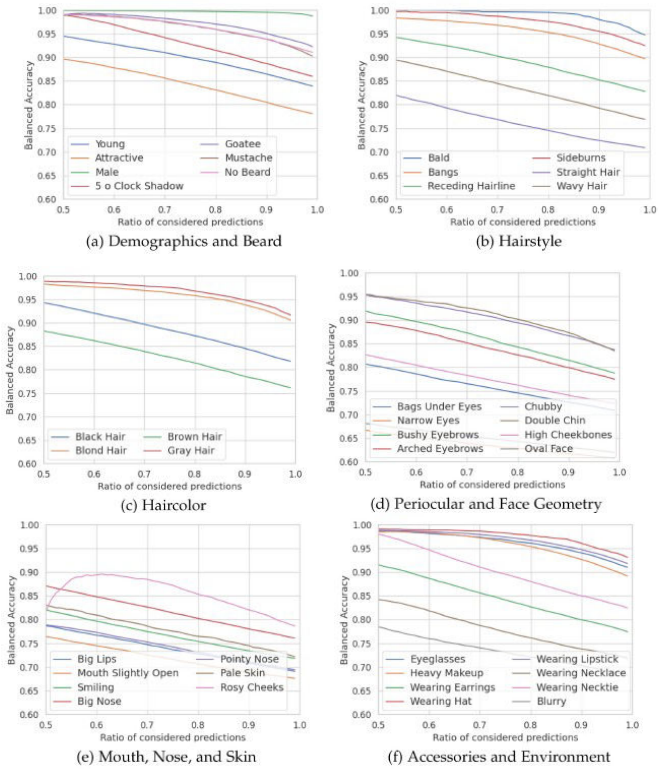


Fig. 4. Accuracy-Reliability plots for the CelebA database on CosFace embeddings. The balanced accuracy of the MAC is shown for a continuous RCP range of [0.5, 1]. The MAC performance of the 40 attributes is ivided into 6 categories represented by subfigures (a)-(f) to allow a simple category-based analysis.

Bold, Bangs, and Goatee are simple to master and so produce high results. This suggests that a single classifier with greater capacity trained on these attributes could outperform the MAC technique in terms of prediction performance. In general, FaceNet and CosFace outperform ArcFace in prediction performance.



ArcFace embeddings have more complex attribute patterns due to the big angle margin principle. Less data was available for training in the LFW experiments since we required to filter low-quality labels to ensure high validity of the results. As a result, it is reasonable to predict that performance on ArcFace will improve if more training data is available. Many soft-biometric properties are strongly encoded in CosFace embeddings, as they are in FaceNet. Furthermore, the predictability of attributes in the categories Mouth and Environment is substantially higher for CosFace than for FaceNet. On ArcFace embeddings, only a few attribute categories, such as Haircolor, Hairstyle, Accessories, and Beard, exhibit a high predictability. Both CosFace and ArcFace are margin-based losses. However, only the additive angle margin loss is considered.

Nonetheless, only the Loss of additive angular margin ArcFace has lower predictability results than triplet-loss and CosFace. This shows how the training loss affects the characteristic. Predictability. Furthermore, the training deficit may have an impact. Be greater than the potential effect of embedding size since the lowest- and highest-dimensional attribute predictability both embeddings (FaceNet 128, CosFace 1024) performed admirably in predicting soft-biometric features, whereas prediction performance on 512-dimensional ArcFace embeddings is more compact. Nonetheless, many features can be predicted with high accuracy from the embeddings. This is true for demographics, haircuts, haircolors, beard kinds, and accessories. However, as shown in Section V-A, there is a considerable link between the accessories Wearing Lipstick and Heavy Makeup, which has a significant impact on the MAC's prediction performance. As a result, it is impossible to tell if both of these attributes are encoded in the face embedding or only one of these. Face geometry factors such as face shape, the presence of a double chin, and forehead visibility can also be determined. The MAC

could not predict attributes that did not directly belong to the user, such as lighting circumstances or image blurriness.

5.3 Category-Wise Analysis of the Face Space

We addressed the results on the level of single qualities in the preceding section. We discovered that 39 of the 113 studied qualities belong to the class easily-predictable, 35 to the class predictable, and 39 to the class hardly-predictable. In this section, we discuss the findings from coarse to fine and, on a more abstract level, attribute categories. Table V summarises the attribute categories in these three predictability classes in order to provide a more broad overview of the encoded information in the face embeddings. This table also incorporates observations from similar investigations, such as discoveries about head posture [37] and image quality [4], to provide a more full picture of the situation. Despite the fact that facial recognition models are trained for recognition, features such as facial Geometry, Periocular Area, Nose, and Mouth are not easily foreseeable. Non-permanent factors, on the other hand, which modern face recognition algorithms strive to be robust at, turn out to be easily foreseeable. This covers, for example, hairstyles, haircolors, beards, accessories, head poses, and social traits.

TABLE VI
THE PERFORMANCE CENTRES ON FACENET (FN), COSFACE (CF), AND ARCFACE (AF) EMBEDDINGS AND IS REPORTED IN TERMS OF COMPLETE ACCURACIES AT TWO DIFFICULTY SCENARIOS: 100% RCP (HARD) AND 50% RCP (EASY). THE ASSIGNED PREDICTABILITY CLASS IS DEFINED BY ++,+, AND 0. GREY HIGHLIGHTING REFERS TO REDUCED EXPRESSIVENESS AFTER THE LABEL-CLEANING PROCESS DUE TO LIMITED DATA.

		100% RCP			50% RCP		
	Attribute	FN	CN	AF	FN	CN	AF
Periocular	Eyes Open^{0}	60.4%	70.9%	54.4%	63.6%	71.7%	54.8%
	Brown Eyes^{+}	82.1%	84.8%	64.0%	92.8%	93.9%	66.8%
	Bags Under Eyes^{++}	87.2%	93.3%	73.7%	95.4%	98.3%	83.5%
	Narrow Eyes^{+}	77.1%	83.2%	66.2%	86.3%	92.3%	74.1%
	Bushy Eyebrows^{++}	96.3%	95.7%	83.8%	99.1%	98.8%	91.7%
	Arched Eyebrows^{+}	85.3%	86.6%	71.6%	94.5%	96.1%	76.8%
Mouth	Mouth Closed^{+}	73.2%	85.0%	64.0%	83.9%	95.9%	72.4%
	Mouth Slightly Open^{+}	73.8%	89.1%	61.8%	83.0%	96.6%	65.1%
	Mouth Wide Open^{+}	66.6%	85.5%	50.8%	59.9%	90.9%	50.0%
	Teeth Not Visible^{+}	70.0%	84.8%	65.2%	75.3%	99.8%	58.3%
	Smiling^{++}	72.0%	93.8%	67.9%	81.3%	99.7%	75.9%
	Big Lips^{++}	87.6%	92.5%	57.3%	98.0%	92.3%	57.8%
Nose	Big Nose^{+}	84.5%	88.8%	71.6%	93.6%	97.3%	81.5%
	Pointy Nose^{++}	96.5%	95.4%	71.5%	100.0%	100.0%	71.3%
	Str. No.-Mou. Lines^{++}	70.0%	94.2%	61.7%	80.7%	99.3%	71.6%
Accessories	Heavy Makeup^{++}	96.7%	96.3%	69.9%	99.0%	100.0%	57.1%
	Wearing Hat^{++}	87.2%	91.7%	67.9%	96.9%	98.3%	53.8%
	Wearing Earrings^{++}	91.7%	91.0%	73.3%	97.9%	97.8%	72.9%
	Wearing Necktie^{+}	84.6%	81.5%	72.8%	93.5%	91.1%	75.2%
	Wearing Necklace^{+}	83.7%	86.0%	74.1%	92.1%	95.1%	82.5%
	Wearing Lipstick^{++}	98.5%	99.1%	75.9%	99.5%	100.0%	74.0%
	No Eyewear^{++}	95.5%	90.4%	86.1%	98.2%	97.7%	90.3%
	Eyeglasses^{++}	96.1%	87.6%	90.0%	98.4%	97.3%	95.6%
	Sunglasses^{+}	71.6%	82.7%	50.8%	62.4%	100.0%	50.0%
Environment	Blurry	61.4%	78.6%	57.2%	66.3%	89.5%	58.6%
	Harsh Lighting^{+}	76.0%	87.3%	61.3%	89.1%	90.8%	57.9%
	Flash^{++}	78.3%	92.6%	58.3%	88.3%	98.8%	51.5%
	Soft Lighting	65.7%	73.8%	60.2%	72.3%	84.8%	66.1%
	Outdoor^{+}	77.2%	88.8%	60.8%	81.9%	97.0%	65.9%
Other	Frowning^{++}	78.3%	97.4%	72.4%	88.8%	99.9%	79.5%
	Color Photo^{0}	72.8%	70.6%	54.0%	75.0%	50.0%	60.0%
	Posed Photo^{+}	76.0%	88.3%	60.7%	80.9%	98.4%	63.0%
	Attractive Man^{0}	74.4%	75.0%	65.0%	85.1%	85.9%	74.2%
	Attractive Woman^{++}	95.3%	95.7%	75.1%	100.0%	98.6%	71.4%
Demographics	Male^{++}	98.3%	96.9%	83.9%	99.5%	99.6%	94.2%
	Baby^{0}	55.1%	49.9%	49.9%	50.0%	50.0%	50.0%
	Child^{0}	68.8%	73.1%	57.5%	75.8%	85.5%	52.4%
	Youth^{+}	79.9%	81.8%	70.5%	93.1%	94.4%	79.8%
	Middle Aged^{+}	88.4%	88.6%	74.0%	95.2%	97.9%	82.9%
	Senior^{++}	99.6%	97.8%	83.9%	100.0%	100.0%	88.4%
	Asian^{++}	95.5%	90.4%	66.2%	100.0%	97.3%	69.6%
	White^{++}	97.4%	94.4%	73.6%	99.4%	99.1%	81.4%
	Black^{++}	95.3%	92.3%	63.2%	98.3%	100.0%	53.6%
	Indian^{+}	85.2%	63.0%	50.2%	92.5%	50.0%	54.7%
Skin	Rosy Cheeks^{0}	67.2%	71.0%	58.8%	73.0%	77.1%	64.3%
	Shiny Skin^{+}	82.1%	89.4%	67.9%	89.7%	99.9%	75.6%
	Pale Skin^{0}	68.0%	73.1%	62.9%	79.9%	83.2%	67.2%
	Flushed Face^{0}	66.5%	73.9%	55.5%	77.5%	77.5%	52.3%
Hairstyle	Curly Hair^{0}	69.0%	72.6%	61.7%	77.8%	83.5%	68.7%
	Wavy Hair^{++}	95.0%	96.7%	80.5%	99.7%	99.7%	83.3%
	Straight Hair	67.5%	69.5%	59.8%	76.8%	80.0%	65.5%
	Receding Hairline^{+}	83.3%	83.9%	73.0%	93.5%	95.0%	84.9%
	Bangs^{++}	97.0%	94.9%	64.1%	100.0%	100.0%	50.0%
	Bald^{++}	93.6%	84.2%	75.8%	97.9%	96.4%	75.0%
	Sideburns^{++}	98.9%	98.5%	84.1%	99.7%	99.7%	89.2%
Haircolor	Black Hair^{++}	90.4%	89.0%	65.6%	96.5%	96.4%	61.5%
	Blond Hair^{++}	95.2%	94.6%	71.7%	98.8%	100.0%	55.6%
	Brown Hair^{+}	81.5%	84.1%	71.9%	91.9%	95.3%	82.7%
	Gray Hair^{++}	98.8%	96.5%	88.4%	100.0%	100.0%	93.9%
Beard	No Beard^{++}	98.1%	94.9%	83.9%	100.0%	100.0%	92.1%
	Moustache^{++}	98.5%	93.7%	79.7%	99.3%	96.8%	78.1%
	5 o Clock Shadow^{++}	96.5%	95.7%	83.8%	99.6%	99.7%	92.4%
	Goatee^{++}	94.5%	84.8%	70.0%	100.0%	100.0%	100.0%
Face Geometry	Oval Face^{++}	82.7%	90.4%	71.6%	95.4%	96.8%	75.8%
	Square Face^{++}	99.1%	96.3%	89.1%	100.0%	99.6%	96.3%
	Round Face^{+}	84.2%	71.4%	49.6%	100.0%	100.0%	50.0%
	Round Jaw^{+}	70.6%	84.7%	60.8%	81.1%	95.0%	58.4%
	Double Chin^{++}	91.5%	96.0%	81.1%	100.0%	100.0%	88.7%
	High Cheekbones^{++}	79.9%	96.9%	73.3%	90.4%	99.9%	81.8%
	Chubby^{+}	85.5%	86.1%	74.3%	98.0%	97.5%	79.4%
	Obstructed Forehead^{++}	85.9%	93.2%	65.0%	99.9%	98.3%	61.3%
	Partially Visible F.$^{+}$	85.2%	85.0%	65.9%	94.0%	95.7%	50.0%
	Fully Visible F.$^{+}$	85.9%	88.7%	71.8%	95.4%	98.2%	82.2%

TABLE V
A CATEGORISED SUMMARY OF THE PREDICTABILITY CLASSES, INCLUDING RELATED WORKS' FINDINGS

Easily-predictable	Predictable	Hardly-predictable
Demographics	Face Geometry	Skin
Hairstyle	Periocular	Mouth
Haircolor	Nose	Environment
Beard	Image Quality [4]	
Accessories		
Head Pose [37]		
Social Traits [38]		

Figure 6 depicts a more comprehensive predictability overview of the attribute categories. The investigation is separated into two face embeddings. The forecast performance of two RCP-levels is depicted on the axis. Each graphic is broken into three sections that reflect the three predictability classes. The grey area indicates the unpredictability class (0), the light green area the predictability class (+), and the dark green area the readily predictability class (++). Furthermore, each point represents the average performance of the attributes in the attribute-category. The (standard) deviation of individual performance of the associated qualities is indicated by the ellipse shaded region around each point. The darkened area's x-axis reflects

the standard deviation of performance under 100% RCP (more realistic circumstances), while the y-axis shows the deviation of performance at 50% RCP (more idealistic circumstances).

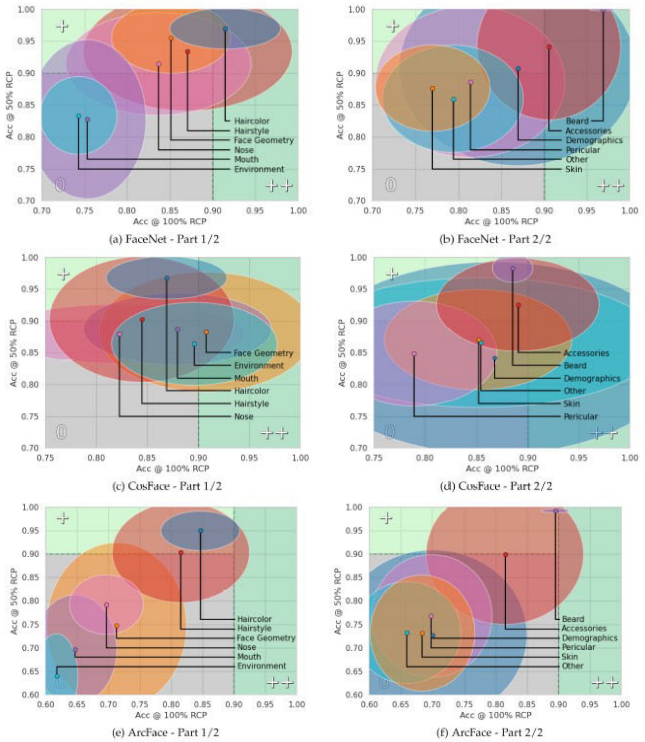
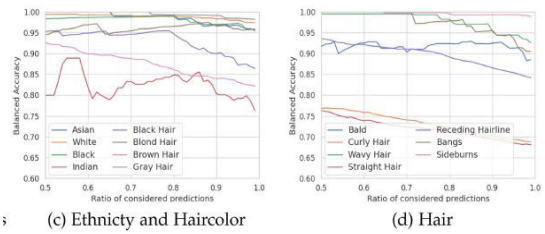
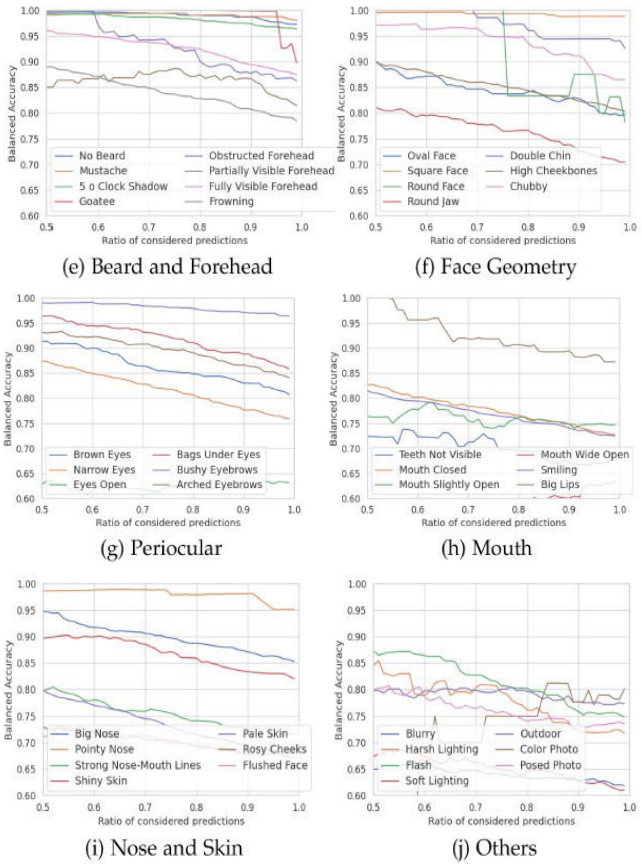


Figure 6 shows a visual representation of the categorised property predictability. The axes depict balanced forecast accuracy at two RCP levels. The figures are arranged into three sections that correspond to the three predictability classes. The dark green area represents the easily-predictable class (++), the light green area represents the predictable class (+), and the grey area represents the hardly-predictable class (0). Each point represents an attribute category's average performance. The grey area surrounding each point shows the (standard) deviation of the category's distinct attribute-performances. In the face embeddings, several attributes are highly encoded.

The predictable nature of the attribute categories for FaceNet is demonstrated in Figures 6(a) and 6(b). Figures 6(c) and 6(d) show the predictability of the attribute-categories for CosFace and ArcFace, respectively. Figures 6(e) and 6(f) demonstrate the predictability of the attribute-categories. Many attribute-categories can be found to be richly embedded in the FaceNet and CosFace embeddings. This contains various Haircolors, Hairstyles, Beards, Accessories, and Demographics, as well as features of the Face Geometry, Nose, and Periocular region. For ArcFace, it is clear that more attribute-categories fall into the grey (unpredictable) category.





Accuracy-reliability charts for the LFW databases on FaceNet embeddings (Fig. 7). The MAC's balanced accuracy is illustrated for a continuous RCP range of [0.5, 1]. To facilitate a simple category-based analysis, the MAC performance of the 73 qualities is separated into 10 categories shown in subfigures (a)-(j).

Face embeddings have more complex attribute patterns due to ArcFace's big angular margin idea. Many attribute categories may fall into the hardlypredictable category due to a lack of training data mixed with ArcFace's additive angular margin loss. Both the reduced amount of training data as a result of the label cleaning process and the more complex attribute pattern as a result of the ArcFace loss may make it more difficult for the MAC to accurately forecast attributes. However, the broad elliptic hues in the grey areas suggest that these groups also have some very predictable characteristics.

In an effort to simplify the relationships, the high-level overview on the attribute-categories results in some valuable information loss.

Our analysis approach only allows us to indicate what information is contained in biometric face embeddings and does not allow us to state which attributes are not encoded. As a result, we can only be certain of four attribute-categories. ArcFace embeddings firmly encode the characteristics Haircolor, Hairstyle, Beard, and Accessories.

The link between facial recognition networks and their users' identities may explain why they keep soft-biometric information. Recent studies [2, 39, 44] demonstrated that softbiometric features of a face provide adequate information to be used successfully in verification and recognition tasks.

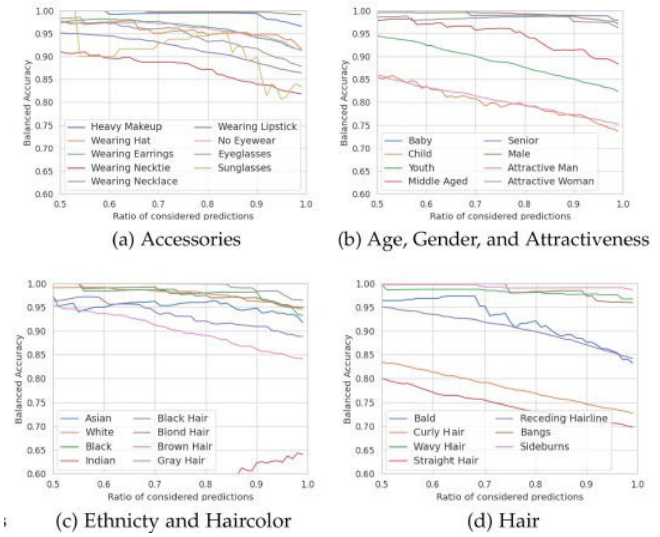
As a result, here is a strong connection between these characteristics, a person's appearance, and its identity. This

relationship could explain why deep neural networks trained for recognition retain these characteristics.

5.4 Implications of Our Findings

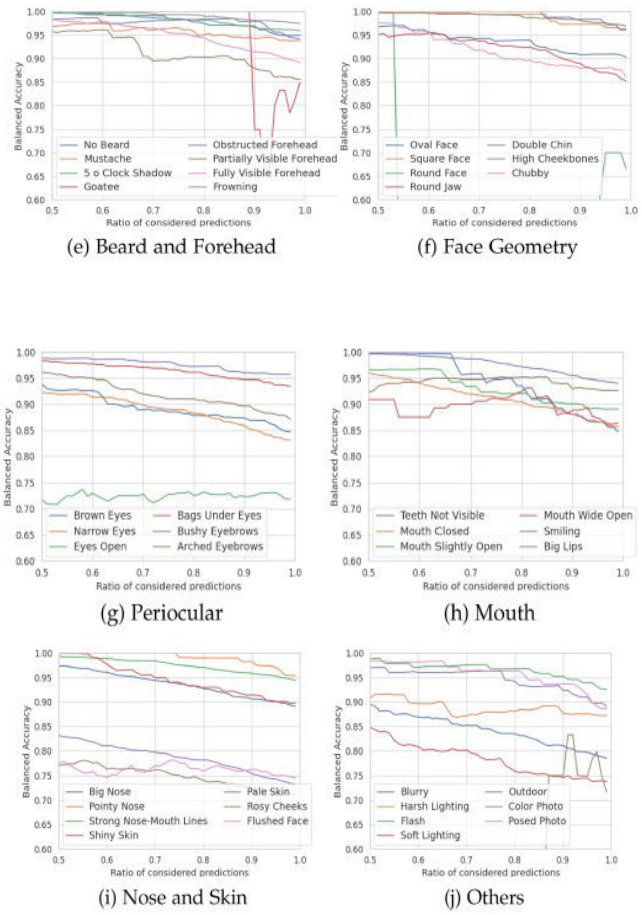
The findings of this study could have far-reaching implications for future research in privacy-preserving and bias-reducing face recognition.

1) Face Recognition Privacy: The trials revealed significant privacy risks in facial recognition technologies. Many applications require the user of a face recognition system to give biometric data alone for recognition. To avoid potential exploitation (function creep) of this private data, embeddings extracted from face recognition systems should only contain identity-related information. However, our investigations revealed that face embeddings also contain information about privacy-sensitive features, posing significant privacy issues. As a result, future works must address these privacy concerns, such as giving techniques to conceal attribute information in face embeddings.
2) Bias in Face Recognition: As our investigations have shown, many qualities are encoded in face embeddings. Despite the fact that face recognition embeddings are trained to be resilient against non-permanent features, the results show that certain attributes, in particular, are reliably anticipated from face templates. This covers information regarding ArcFace Hairstyles, Haircolors, Beards, and Accessories, as well as additional FaceNet features. The presence of these attribute-traits in face embeddings implies that existing face recognition algorithms are still vulnerable to these non-demographic aspects, as demonstrated in recent publications [3, 7, 50]. As a result, future research must suggest strategies to decrease non-demographic bias in face recognition.



(a) Accessories (b) Age, Gender, and Attractiveness

(c) Ethnicty and Haircolor (d) Hair

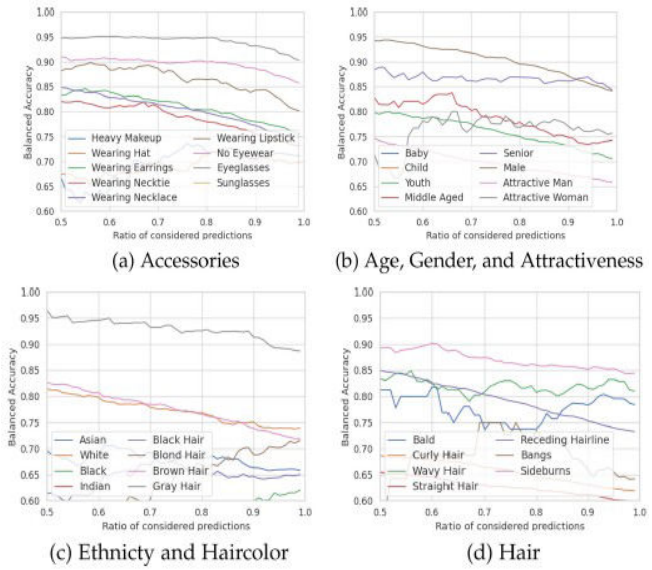


(e) Beard and Forehead (f) Face Geometry

(g) Periocular (h) Mouth

(i) Nose and Skin (j) Others

Fig. 8. Accuracy-Reliability plots for the LFW database on CosFace embeddings. The balanced accuracy of the MAC is shown for continuous RCP range of [0.5, 1]. The MAC performance of the 73 attributes is divided into 10 categories represented by subfigures (a)-(j) to allow a simple category-based analysis.



(a) Accessories (b) Age, Gender, and Attractiveness

(c) Ethnicty and Haircolor (d) Hair

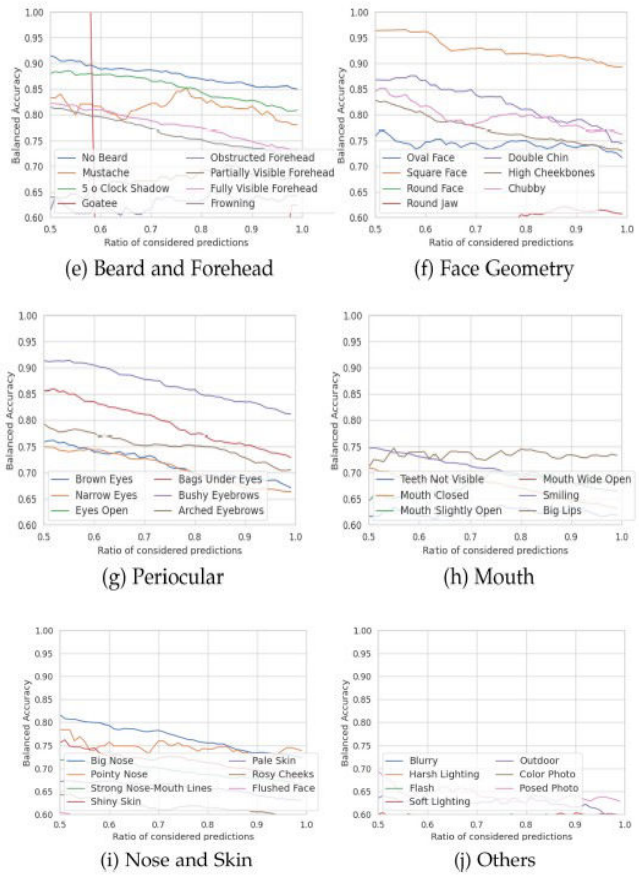


Fig. 9. Accuracy-Reliability plots for the LFW database on ArcFace embeddings. The balanced accuracy of the MAC is shown for continuous RCP range of [0.5, 1]. The MAC performance of the 73 attributes is divided into 10 categories represented by subfigures (a)-(j) to allow a simple category-based analysis

VI. CONCLUSION

Face recognition systems' present success is fueled by breakthroughs in deeply learned face embeddings. Recent research has proven that these embeddings include more information than just the person's identification. These embeddings, for example, encode demographics, picture features, and social factors. This could lead to biassed judgements in facial recognition algorithms and generate serious privacy concerns. To address these privacy and prejudice concerns, a thorough understanding of the encoded information in face embeddings is required. As a result, in this paper, we give a more in-depth examination of the information preserved in biometric face embeddings. We assessed the predictability of 73 different softbiometric features from three common face embeddings over a range of difficulty levels. We also explored the predictability of numerous types of variables to improve the understandability of the results. This was accomplished by categorising each group into one of three predictability classes and assessing predictability in a continuous range. The findings show that several attributes are stored in biometric face embeddings.

From face embeddings, around one-third of the examined qualities are easily predicted, another third are predictable, and one-third are just hardly predictable. We were able to demonstrate that haircolor, hairstyles, beards, and accessories are strongly stored in face embeddings. Despite the fact that face identification template are trained to be resilient against non-permanent elements, we proved that these characteristics are easily foreseeable from face embeddings. We expect that future research will build on our work's knowledge to develop accurate face recognition solutions that also address prejudice and privacy concerns of diverse origins.

REFERENCE

[1] G. Alain and Y. Bengio, “Understanding intermediate layers using linear classifier probes,” in *Proc. 5th Int. Conf. Learn. Represent.*, Apr. 2017, pp. 1–13

[2] G. Balakrishnan, Y. Xiong, W. Xia, and P. Perona, “Towards causal benchmarking of bias in face analysis algorithms,” in Proc. 16th Eur. Conf. Comput. Vis.,Glasgow, U.K., Aug. 2020, pp. 547–563

[3] B. Bortolato et al., “Learning privacy-enhancing face representations through feature disentanglement,” in Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit., 2020, pp. 495–502.

[4] N. Almudhahka, M. S. Nixon, and J. S. Hare, “Human face identification via comparative soft biometrics,” in Proc. IEEE Int. Conf. Identity Security Behav. Anal. 2016, pp. 1–6.

[5] L. Best-Rowden and A. K. Jain, “Learning face image quality from human assessments,” IEEE Trans. Inf. Forensics Security, vol. 13, pp. 3064–3077, 2018.

[6] F. Boutros, N. Damer, P. Terhörst, F. Kirchbuchner, and A. Kuijper, “Exploring the channels of multiple color spaces for age and gender estimation from face images,” in Proc. 22th Int. Conf. Inf. Fusion (FUSION), Ottawa, ON, Canada, Jul. 2019, pp. 1–8.

[7] C. M. Cook, J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury, “Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems,” IEEE Trans. Biom., Behav., Ident. Sci., vol. 1, no. 1, pp. 32–41, Jan. 2019.

[8] N. Damer, Y. Wainakh, V. Boller, S. von den Berken, P. Terhörst, A. Braun, and A. Kuijper. “Crazyfaces: Unassisted circumvention of watchlist face identification,” in Proc. 9th IEEE Int. Conf. Biometr. Theory Appl. Syst. (BTAS), Oct. 2018, pp. 1–9.

[9] A. Das, A. Dantcheva, and F. Bremond. “Mitigating bias in gender, age and ethnicity classification: A multi-task convolution neural network approach,” in Proc. Comput. Vis. Workshops, Munich, Germany, Sep. 2018, pp. 573–585.

[10] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “ArcFace: Additive angular margin loss for deep face recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 4690–4699.

[11] P. Dhar, A. Bansal, C. D. Castillo, J. Gleason, P. J. Phillips, and R. Chellappa, “How are attributes expressed in face DCNNs?” in Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit., 2020, pp. 85–92.

[12] S. Gong, X. Liu, and A. K. Jain, “DebFace: De-biasing face recognition,” 2019. [Online]. Available: arXiv:1911.08080v4.

[13] P. Grother, M. Ngan, and K. Hanaoka, “Ongoing face recognition vendor test (FRVT) part 2: Identification,”

Nat. Inst. Stand. Technol., Gaithersburg, MD, USA, Rep. NISTIR 8271, 2018.

[14] G. Guo and G. Mu, "Joint estimation of age, gender and ethnicity: CCA vs. PLS," in Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit., Shanghai, China, Apr. 2013, pp. 1–6.

[15] J. Guo, J. Deng, N. Xue, and S. Zafeiriou, "Stacked dense U-nets with dual transformers for robust face alignment," in Proc. Brit. Mach. Vis. Conf., 2018, p. 44.

[16] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in Proc. 14th Eur. Conf. Comput. Vis., Oct. 2016, pp. 87–102.

[17] H. Han, A. K. Jain, F. Wang, S. Shan, and X. Chen, "Heterogeneous face attribute estimation: A deep multi-task learning approach," IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 11, pp. 2597–2609, Nov. 2018.

[18] J. Hernandez-Ortega, J. Galbally, J. Fiérrez, R. Haraksim, and L. Beslay, "FaceQnet: Quality assessment for face recognition based on deep learning," in Proc. IEEE Int. Conf. Biometr., Crete, Greece, Jun. 2019, pp. 1–8.

[19] M. Q. Hill et al., "Deep convolutional neural networks in the faceof caricature: Identity and image revealed," 2018. [Online]. Available: arXiv:1812.10902.

[20] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. Massachusetts, Amherst, MA, USA, Rep. 07-49, Oct. 2007. [Online]. Available: http://viswww.cs.umass.edu/lfw/#reference

.