MAAD-Face: A Massively Annotated Attribute Dataset for Face Images

Philipp Terhörst, Daniel Fährmann, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper

***Abstract*—Soft-biometrics play an important role in face biometrics and related fields since these might lead to biased performances, threaten the user's privacy, or are valuable for commercial aspects. Current face databases are specifically constructed for the development of face recognition applications. Consequently, these databases contain a large number of face images but lack in the number of attribute annotations and the overall annotation correctness. In this work, we propose a novel annotation-transfer pipeline that allows to accurately transfer attribute annotations from multiple source datasets to a target dataset. The transfer is based on a massive attribute classifier that can accurately state its prediction confidence. Using these prediction confidences, a high correctness of the transferred annotations is ensured. Applying this pipeline to the VGGFace2 database, we propose the MAAD-Face annotation database. It consists of 3.3M faces of over 9k individuals and provides 123.9M attribute annotations of 47 different binary attributes. Consequently, it provides 15 and 137 times more attribute annotations than CelebA and LFW. Our investigation on the annotation quality by three human evaluators demonstrated the superiority of the MAAD-Face annotations over existing databases. Additionally, we make use of the large number of high-quality annotations from MAAD-Face to study the viability of soft-biometrics for recognition, providing insights into which attributes support genuine and imposter decisions. The MAAD-Face annotations dataset is publicly available.**

***Index Terms*—Face recognition, Database, Facial Attributes, Soft-biometrics, Annotation-transfer, Human evaluation, Biometrics**

I. Introduction

Soft-biometric characteristics play a major role in face recognition research and applications [7]. Recently, there is a high interest in studying these attributes and mitigating their effects on recognition performances for fair face recognition systems [9]. Soft-biometrics are also a key factor for privacy-enhancing face recognition technologies, either by recognizing individuals based on soft-biometrics only [25] or by suppressing privacy-sensitive characteristics to prevent function creep [50]. However, most of these research efforts focus on demographic aspects only. One possible reason can be the lack of annotated data. Recent face databases are specifically constructed for the development of face recognition systems. Consequently, these contain large numbers of faces under diverse settings but lack annotations.

This work closes this gap by proposing the MAAD-Face dataset. MAAD-Face is a novel face annotations database that is characterized by its large number of high-quality face annotations. Utilizing our novel annotation-transfer pipeline, we transfer the attribute annotations from two source databases (LFW [22] and CelebA [30]) to the target database VGGFace2 [4]. The pipeline trains a massive attribute classifier (MAC) per source database to accurately predict the attributes of the source. Since the MAC makes use of prediction reliabilities [48], the pipeline neglects annotations origin from less-confident predictions and thus, ensures a high correctness of the transferred annotations. MAAD-Face consists of 3.3M faces of over 9k individuals, which is significantly higher than related annotated datasets such as CelebA (0.2M faces of 10k individuals with 40 different attributes) and LFW (13.2k faces of 5.7k individuals with 74 different attributes). With 123.9M attribute annotations of 47 different binary attributes, MAAD-Face provides 15 and 137 times more attribute annotations than CelebA and LFW. To analyse the quality of the attribute annotations, three human evaluators investigated the correctness of the annotations of CelebA, LFW, and MAAD-Face. The results demonstrate the superiority of the MAAD-Face annotations over the other databases. Finally, we investigated the viability of using soft-biometrics attributes for recognition using MAAD-Face. We show the relevance of each attribute for genuine and imposter decisions and analyse how many of the most important attributes are necessary to achieve a certain recognition performance. The MAAD-Face dataset is publicly available under the following link[1].

To summarize, this work presents *four main contributions*:

1) A novel annotation-transfer pipeline is proposed that can transfer attribute annotations from multiple source databases to a target database while ensuring a high correctness of the transferred annotations. We use this pipeline to create MAAD-Face.
2) We propose the MAAD-Face annotations dataset based on VGGFace2 [4]. MAAD-Face is a new face annotations database consisting of 123.9M attribute annotations of 47 different binary attributes. It provides 15 and 137 times more annotations than CelebA and LFW, while the attribute annotations are of higher quality.
3) The third contribution is a human evaluation of the annotation correctness of three large-scale annotation face databases, LFW, CelebA, and MAAD-Face. These demonstrate the superiority of the MAAD-Face annotations over the other investigated databases.
4) The last contribution is a study on how well these facial attributes can be used for verification and identification based on soft-biometrics only.

All authors are with the Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany and with the Technical University of Darmstadt, Darmstadt, Germany. e-mail: {forename.lastname@igd.fraunhofer.de}

[1]https://github.com/pterhoer/MAAD-Face

The rest of the paper is structured as follows. Section II provides an overview of annotated face datasets and a human evaluation of the annotation-correctness of three highly-annotated datasets. In Section III, the annotation-transfer pipeline is explained and how this is used to create MAAD-Face. Section IV-A provides statistical properties of MAAD-Face and in Section V, the soft-biometric annotations of MAAD-Face are used to evaluate how well these attributes can be utilized to recognize individuals.

II. Related Works

A. Review of Annotated Face Datasets

In recent years, many face databases have been released. These mainly aimed at providing a large dataset for developing face recognition solutions. With the use of deep-learning techniques in face recognition, the required data for training these solutions has grown strongly and thus, the sizes of face databases. However, less attention was given to the estimation of facial attributes. These soft-biometric characteristics can be of high importance in applications such as access control [7], human-computer interaction [48], and law enforcement [12]. Current face databases only provide insufficient numbers of training annotations for training accurate solutions. Moreover, these annotations often lack in their correctness and thus, prevent the development of soft-biometric solutions. In the following, we discuss popular face databases that also contain attribute information.

ColorFeret [35] consists of 14.1k images of 1.2k different individuals with different poses under controlled conditions. The dataset includes a variety of face poses, facial expressions, and lighting conditions. Each image contains annotations of the individual's gender, ethnicity, head pose, age, glasses, and beard. In total, ColorFeret provides around 183k soft-biometric annotations.

The Adience dataset [10] consists of over 26.5k images of over 2.2k different individuals in unconstrained environments. In total, the dataset provides around 263k annotations for gender and age. These images were manually annotated.

The Morph dataset [37] contains 55.1k frontal face images of more than 13.6k individuals. For each image, it provides information about the person's gender, ethnicity, age, beard, and glasses. 80.4% of the faces belong to the ethnicity black, 19.2% to white, and 0.4% to others. The individuals' age varies from 16-77 years. 79.4% of the faces are within an age-range of $[20, 50]$. In total, the Morph database provides over 0.5M annotations for soft-biometric attributes.

VGGFace [34] and VGGFace2 [4] are two databases from the University of Oxford. VGGFace [34] contains 2.6M images from 2.6k individuals and provides information about the head pose (frontal, profile). VGGFace2 [4] contains faces from over 9k subjects with over 3M images. The dataset contains a large variety of pose, age, and ethnicity. Over 40% of the face are frontal and over 50% are half-frontal. Most images belong to individuals over 18 years old and around 40% belong to the age group of $[25, 34]$. For each image, gender annotations are available. A subset of 30k images of celebrities was additionally annotated with 10 further attributes about the individual's hair, beard, glasses, and hat. In total, VGGFace2 provides 3.6M annotations about the person's face.

Labelled Faces in the Wild (LFW) [22] contains 13.2k images of 5.7k different identities from unconstrained environments. It contains variability in pose, lighting, expression, and demographics. With 74 binary attributes, it provides a large diversity on binary attribute annotations, such as attributes belonging to demographics, hair, skin, accessories, and capture environment. However, as we will show in Section IV-C, the correctness of these annotations are often weak (72% accuracy compared to human annotations). In total, LFW provides over 0.9M attribute annotations. Moreover, it should be mentioned that LFW and VGGFace2 have some overlapping subjects since both databases contain many images of celebrities [43].

The CelebFaces Attributes Dataset (CelebA) [30] contains over 202k images of 10.0k different subjects. It covers large pose variations and background clutter and provides rich annotations for 40 binary attributes. In total, CelebA provides over 8M annotations for soft-biometric attributes. These include attributes belonging to demographics, hair, face geometry, and accessories.

A summary of related face annotation databases are shown in Table I. There, the number of subjects and face images are shown along with the number of attributes and the total number of annotations.

In this work, we propose the MAAD-Face annotation database. Using our novel annotation-transfer technique we are able to create highly accurate face annotations building upon VGGFace2. Consequently, it contains over 3.3M face images from over 9.1k different subjects with a large variety of poses, ages, and ethnicities. MAAD-Face provides annotations for 47 binary attributes. In total, it consists of over 123.9M attribute annotations, which is over 15 times higher than the second-largest face annotation dataset.

B. Soft-Biometrics from Faces

Recently, research on the estimation of soft-biometric attributes from face images has shifted from the use of hand-crafted features to the use of deep convolutional neural networks [7]. These kinds of approaches often surpass human-level performance, e.g. for age [20], gender [14], or race estimation [21]. Due to the high performance of automatic approaches, several works demonstrated the benefits of soft-biometrics for recognition. Recent works demonstrated that soft-biometric information only, extracted from face images, can be successfully used to verify [40], [41] and identify [13], [1], [36] individuals. This is especially useful when facing low-quality capture, such as images from great distances [51]. Moreover, soft-biometric information can also support primary biometric modalities such as face recognition [15], [16], [55].

The basis for these successes is the developments of accurate soft-biometric estimators. The solutions cover a wide range of mechanisms such as domain-adaption [39], [18], [38], [48], [47], cascade CNN's [5], [19], [32], autoencoders [54], stacked model [52], and deep regression trees [44].

Taking into account the current pandemic conditions, Alonso-Fernandez et al. [2] demonstrated that soft-biometric

TABLE I: Statistics of related face annotation databases. Distinctive attributes refers to the number of different attributes that are annotated while the number of annotations refers to the total number of (attribute) annotations in the database. To make the number of attributes between different databases comparable, categorical attributes are transformed to binary via one-hot encoding. These are marked a (*). Compared to related databases, MAAD-Face provides the highest number of attribute annotations.

			Attribute annotations	
Database	Num. of subjects	Num. of images	Distinctive attributes	Total number of annotations
ColorFeret [35]	1.2k	14.1k	13*	0.2M
Adience [10]	2.3k	26.6k	10*	0.3M
Morph [37]	13.6k	55.1k	10*	0.6M
VGGFace [34]	2.6k	2.6M	1	2.6M
VGGFace2 [4]	9.1k	3.3M	11	3.6M
LFW [22]	5.7k	13.2k	**74**	0.9M
CelebA [30]	10.0k	0.2M	40	8.0M
MAAD-Face (this paper)	9.1k	3.3M	47	**123.9M**

attributes can still be accurately states from faces wearing masks. In [46], Terhrst et al. showed that 74 out of 113 analysed soft-biometric attributes are encoded in face templates demonstrating a strong need for privacy-enhancing methods. In [49], it was shown that soft-biometric attributes, in general, have a strong influence on the face recognition performance demonstrating the need for bias mitigating solutions beyond demographics.

Building on the reliability measure developed in [48], this work proposes a novel annotation transfer pipeline that is able to do the transfer task with a high annotation correctness. This pipeline is used to create the MAAD-Face annotations dataset. This dataset provides the required data for the developing and evaluating of privacy-enhancing and bias-mitigating face recognition solutions to mitigate the privacy and bias issues mentioned in [46], [49]. For developing these solutions, it provides the required training data. For evaluating privacy-enhancing and bias-mitigating face recognition solutions, it provides a solid basis of data that can serve as the test set.

III. ANNOTATION-TRANSFER PIPELINE

In this section, we will present one of the main contributions of this work, a novel annotation-transfer pipeline that can create highly reliable and accurate attribute annotations. We will explain this pipeline based on the example of the MAAD-Face annotations database. The MAAD-Face database was created by transferring the attribute annotations of CelebA and LFW on the images of VGGFace2.

An overview of the proposed annotation-transfer pipeline is shown in Figure 1. The pipeline consists of five steps that aim to transfer the annotations of source databases to the target database.

1) A massive attribute classifier (MAC) is trained on the training part of each source dataset. Besides making predictions about the estimated annotations of a given image, the MAC is able to additionally providing a reliability statement that states the model's prediction confidence for each annotation.
2) The MAC predicts the annotations on the test-parts of the source datasets including the prediction reliabilities.
3) Based on this performance, the reliability threshold for each attribute is determined. Moreover, a performance-reliability mapping is calculated that allows assigning an attribute-reliability with its expected correctness (performance).
4) The MAC predicts the attribute annotations as well as the corresponding reliabilities for each image in the target dataset. Predicted annotations below the attribute threshold will be rejected to guarantee a high quality of the transferred source annotations.
5) Finally, the source annotations (with their reliabilities) are aggregated using the corresponding performance-reliability mapping. If the source annotations for an image produces different annotations, the annotation is used as the target annotation that has the higher expected correctness.

In the following sections, we describe how (a) the MAC training procedure is conducted on the source datasets, (b) the prediction reliability statements of the MAC are calculated, and (c) how this results in the final annotations for the target database.

A. The Massive Attribute Classifier (MAC)

To transfer the annotations for each attribute from source databases to a target database, we (a) train a MAC jointly on all attributes of a source database to make use of a shared embedding space and (b) construct the MAC such that it can produce accurate reliability measures for each attribute-annotation prediction.

The MAC is a neural network that is trained to predict the attributes of the source dataset. The network architecture is chosen to maximize the prediction accuracy. As it will be demonstrated in Section III-B, the only requirement for the MAC is to train with at least one dropout-layer [45]. We will need this layer to determine the reliability of a prediction. Each source database is subject-exclusively divided into a 80% training set and a 20% test set. A separate MAC is trained for each source training set. To construct MAAD-Face, we use VGGFace2 as the target database and CelebA and LFW as source databases for training two MACs.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2021.3096120, IEEE Transactions on Information Forensics and Security

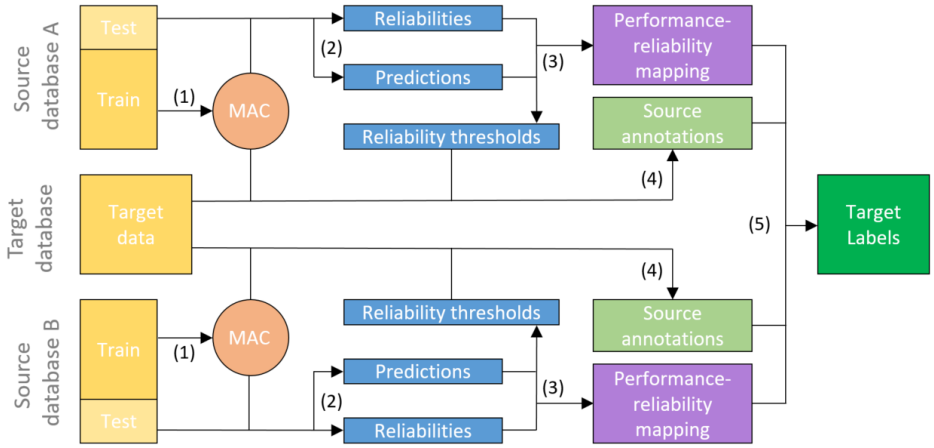


Fig. 1: Overview of the proposed annotation-transfer pipeline. (1) A MAC is trained on the training part of each source dataset. (2) The MAC produces predictions and prediction reliabilities on the test set. (3) These are used to determine the reliability thresholds per attribute and to calculate the performance-reliability mapping. (4) The MAC and the reliability thresholds are used to create (source) attribute annotations for the target dataset. Finally, (5) the source annotations from each source dataset are aggregated using the corresponding performance-reliability mappings to construct the final target annotations for the target dataset.

In the following, we describe the structure and the training details of the MAC, as well as the data cleaning process used. As we demonstrated in Section II-A, many annotations of LFW are wrongly assigned. To prevent confusion of the MAC trained on these annotations, we filter out annotations that are wrongly assigned with a high probability.

1) MAC training: Generally, the training of the MAC can vary and should be task and data-dependent. In order to prepare the MAC for our annotation-transfer pipeline, it needs to be trained with at least one dropout-layer [45] and consists of a soft-max layer as the output.

For the construction of MAAD-Face, we build the MAC on the templates of face images. As shown in the work of Terhörst et al. [46], one can easily and accurately predict many attributes from such templates. Based on these results, we trained a neural network model that takes FaceNet [42] embeddings as an input to jointly predict multiple attributes of the source database. However, a MAC can also be trained end-to-end or by fine-tuning an existing network. The utilized network structure follows the one used by Terhörst et al. [46]. It consists of two initial layers, the input layer of size n_{in} and the second dense layer of size 512. The size of the utilized face embedding is denoted by n_{in} and for our FaceNet model[2] refers to 128 dimensions. Starting from the second layer, each attribute a has an own branch consisting of two additional dense layers of size 512 and $n_{out}^{(a)}$, where $n_{out}^{(a)}$ refers to the number of attributes per class. Each layer has a ReLU activation, except for the output-layers. These have softmax activations. Moreover, Batch-Normalization [23] and dropout [45] ($p_{drop} = 0.5$) is applied to every layer. The dropout allows achieve a generalized performance and also enables us to derive reliability statements about the predictions as we will describe in Section III-B. The training of the MAC was done in a multi-task learning fashion by applying a categorical cross-entropy loss for each attribute branch and use an equal weighting between each of these attribute-related losses. For the training, an Adam optimizer [26] was used with $e = 200$ epochs, an initial learning rate $\alpha = 10^{-3}$, and a learning-rate decay of $\beta = \alpha/e$. The parameter choices followed [46]. The batch size b was chosen according to the amount of available training data, $b = 1024$ for CelebA and $b = 16$ for LFW.

2) Cleaning training attribute annotations: For the annotation-transfer pipeline, this step is only necessary if a source database consists of attribute annotations of low quality. As we demonstrated in Section II-A, this is the case for LFW. However, the quality of the input data of a model is important for the quality of its output data as demonstrated by Geiger et al. [11]. Therefore, in this section we will describe an annotation-cleaning process that was used on the LFW dataset.

While in CelebA the attributes are of binary nature, the annotations in LFW originate from the prediction probabilities of a binary classifier [22]. Therefore, these annotations are continuous and measure the degree of the attribute [28], [29]. Positive values represent "true" annotations and negative values represent "false" annotations. A positive annotation for an attribute a of an image means that the face in the image has the attribute a. For instance, a face with a positive annotation for *Beard* represents a face with a beard. In contrast, a negative annotation for an attribute a of an image means that the face in the image has not the attribute a. However, using the prediction probabilities of a binary classifier does not necessarily reflect the correctness of the prediction as shown in recent works [17], [27], [33]. Consequently, a wide range of the LFW annotations centred around a value of zero is unreliable.

To ensure that our MAC learns on meaningful LFW-annotations, we manually removed these centred annotations as described in the work of Terhörst et al. [46]. Therefore, we assigned an upper and lower score threshold for each attribute.

[2]https://github.com/davidsandberg/facenet

Images with a score over the upper threshold are assigned as true, images with a score under the lower threshold are assigned as false, images with scores within the range are rejected. The upper and lower thresholds for one attribute are manually determined by moving potential thresholds away from zero. At each potential threshold, ten images with the closest attribute scores are investigated. Here, the original LFW annotations of the images are manually investigated for correctness. If only eight or fewer attributes are investigated as correct, the potential threshold is further moved away from the starting point and the procedure is repeated. If a potential threshold returns images with 9 or more correct annotations, it is chosen as the limit. Repeating this over all attributes will result in a lower and an upper threshold for each of these attributes. By binaryzing the scores with these upper and lower thresholds, reduces the number of annotations by 51,7%. However, it also ensures an error-minimizing data basis of the MAC. Thus, it allows us to train the MAC on meaningful and mostly correctly annotated data.

B. Deriving reliability statements

To ensure that the target database will only get annotations of high quality, the prediction reliability is additionally estimated for each prediction (target annotation). Therefore, we follow the methodology described in [48] to enable our MAC to accurately state its own prediction confidence (reliability). To derive the reliability statement additionally to an attribute prediction, $m = 100$ stochastic forward passes are performed. In each forward pass, a different dropout-pattern is applied, resulting in m different softmax outputs $v_i^{(a)}$ for each attribute a. Given the outputs of the m stochastic forward passes of the predicted class $\hat{c}$ denoted as $x^{(a)} = v_{i,\hat{c}}^{(a)}$, the reliability measure is given as

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

with $\alpha = 0.5$, following the recommendation in [48]. The first part of the equation is a measure of centrality and utilizes the probability interpretation of the softmax output. A higher value can be interpreted as a high probability that the prediction is correct. The second part of the equation is the measure of dispersion and quantifies the agreement of the stochastic outputs x. In [48], this was shown to be an accurate reliability measure.

C. Attribute annotation generation

In this section, we combine the MAC models of the source datasets and the reliability measure to create high-quality target annotations. First, we will describe how to set the reliability thresholds for each attribute and MAC. Then, we will describe how this can be used to create the annotations on the target dataset.

1) Defining reliability thresholds: For each source database, a MAC model $\mathcal{M}$ was already trained on the training part as described in Section III-A. Now, the MAC predicts the source annotations on the test-part including the prediction reliabilities. Moreover, the MAC repeats this step on the target database. For each attribute a of the source database, the reliability threshold $thr_{Source}^{(a)}$ is chosen such that the (balanced) prediction accuracy of a is over $acc_{min}\%$ and at least $d_{min}\%$ of the target samples are over this threshold. Consequently, acc_{min} defines the quality of the target annotations while d_{min} define the amount of the annotations in the target database. If an attribute does not accomplish this requirement, the attribute is discarded.

For the creation of MAAD-Face, we set $acc_{min} = 90\%$ and $d_{min} = 50\%$ to receive a large number of high-quality annotations. This results in manually chosen reliability thresholds $thr_{CelebA}^{(a)}$ and $thr_{LFW}^{(a)}$ for each attribute $a \in \mathcal{A}$.

2) Creating target annotations: After defining the reliability thresholds for each MAC and attribute $a \in \mathcal{A}$, we can create the target annotations. Therefore, each MAC computes its predictions p_{Source} and prediction reliabilities r_{Source} on the target dataset. The prediction *True* is defined as 1, the prediction *False* is defined as -1. If an attribute-prediction $p_{Source}^{(}a, i)$ for an image i has a prediction-reliability below the threshold $r_{Source}^{(a,i)} < thr_{Source}^{(a)}$, the annotation is set to 0 (*undefined*). In this case, the MAC is not confident enough about its prediction and rejecting these predictions guarantee high-quality remaining annotations. For each source dataset, this procedure results in a set of annotations l_{Source} for the target dataset images. Finally, this set of annotations have to be combined to create the target annotations. If an attribute just appears in one of the source datasets, the source annotations l_{Source} are directly used for the target dataset. If an attribute appears in multiple source datasets, we have to decide which annotation to use as the target annotation. In this case, the reliability r_{Source} is mapped back to the performance of the test set $acc(r_{Source})$ and the annotation assigned with the highest map-back performance is used for the target annotation. Please note that such a decision can not be made based on the reliability-level only since the range of the reliability values vary between each MAC. Mapping back the reliability values to the test-set performances allow an aligned comparison of the annotation-quality.

Algorithm 1 summarizes the annotation generation procedure. The inputs are the predictions $\{p_{Source}\}$, the corresponding reliabilities $\{r_{Source}\}$, the reliability thresholds $\{thr_{Source}\}$, as well as a set of all attribute $\mathcal{A}$. The output of the algorithm is the annotations l_{Target} of the target dataset. The *transfer* function transforms the predictions p_{Source} into the source annotations l_{Source} based on the prediction reliabilities r_{Source} and the corresponding attribute reliability thresholds thr_{Source}. If an attribute appears in multiple source databases, the *highest* function maps back the reliability to the test-set performance $acc(r_{Source}^{(a,i)}$ and returns the annotation $l_{Source}^{(a,i)}$ with the highest map-back performance.

The last step (*obtainPlausability*) performs a plausibility check including required corrections, given the target annotations l_{Target}, the attribute classes $\mathcal{A}$, and the corresponding attributes. For each class, at maximum one attribute can be true. For instance, for the class gender, either the attribute male or female can be true. A list of the classes with the

corresponding attributes is shown in Table III. Due to this restriction, we set all attribute annotations for an image i to undefined (0) if more than one attribute showed true before. This aims at maintaining high-quality annotations.

Algorithm 1 - Annotation Generation

Input: $\{p_{Source}\}, \{r_{Source}\}, \{thr_{Source}\}, \mathcal{A}$
Output: target dataset annotations l_{Target}

for $a \in \mathcal{A}$ **do**
  for each source dataset **do**
    $l^{(a)}_{Source} \leftarrow transfer(p^{(a)}_{Source}, r^{(a)}_{Source}, thr^{(a)}_{Source})$
  end for
end for
$l_{MAAD} = zeros(|\mathcal{A}|, |\mathcal{I}|)$
for $a \in \mathcal{A}$ **do**
  for $i \in \mathcal{I}$ **do**
    $l^{(a,i)}_{Target} \leftarrow highest(\{l^{(a,i)}_{Source}\}, \{acc(r^{(a,i)}_{Source})\})$
  end for
end for
$l_{Target} \leftarrow obtainPlausibility(l_{Target}, \mathcal{A})$
return l_{Target}

D. Discussion

The proposed annotation transfer pipeline is related to homogenous discrepancy-based domain adaptation methods [31], [53]. Since the feature spaces between the source and target domains are identical and only differ in terms of data distribution, it is similar to homogeneous domain adaptations [53]. Since the reliability measure of the transferred labels can be interpreted as the distance between the source and the target domain, the proposed pipeline is also similar to discrepancy-based domain adaption methods [31]. In contrast to classical domain adaptation methods that utilize labeled data in the source domains to execute new tasks in a target domain [53], the proposed approach solves the same task in the target domain. However, it measures the discrepancy (reliability) in the target domain to prevent false decision (annotations). This might only lead to decisions when the target and source domain share a specific similarity but also ensures a high correctness of the decisions (annotations).

IV. MAAD-Face

A. MAAD-Face Statistics

The biggest advantage of MAAD-Face is its large number of high-quality attribute annotations. Since it builds on the VGGFace2 database, it consists of over 9.1k identities with over 3.3M face images of various poses, ages, and illuminations. MAAD-Face has annotations for 47 distinctive attributes with a total of 38.3M annotations. On average 37.5 ± 3.7 annotations are defined per image. Figure 3 shows the annotation distribution of MAAD-Face for all 47 attributes. For each attribute, green indicates the percentage of positive annotations, red indicates the percentage of negatively annotated images, and grey represents the percentage of images with undefined annotations. Some attributes have a low number of positive annotations, such as *Mustache* (16.6k) or *Goatee* (9.2k) and instead, a higher number of undefined annotations. This way, we can ensure the high correctness of the annotations as explained in Section III-C (accuracy *Mustache* 98%, accuracy *Goatee* 95%). In total, this leads to MAAD-Face having 23.1% positive, 56.6% negative, and 20.3% undefined annotations. A list of all attributes with the correctness analysis was already discussed with Table III in Section IV-C. The high quality of the attribute annotations is also observable in Figure 4. There, five random sample images are shown with their corresponding attribute annotations.

TABLE II: Analysing the MAC prediction performance in comparison to Microsoft Azure under different head poses and lighting conditions. The values reported are the balanced accuracies based on ColorFeret under frontal and (artificial) lighting from the side. Moreover, the performance under different headpose from frontal ($0°$) to profile ($90°$). While the MAC classifier can predict 47 attributes, in this evaluation only shared attributes between ColorFeret and Azure are considered. The rejected rate describes the ratio of images that could not be processed by the algorithm. In general, both approaches perform well. However, for non-frontal images and images with side lighting, Azure rejects most of the images without any predictions.

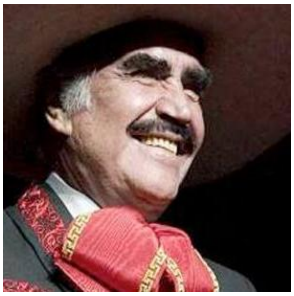
		MAC (ours)				Microsoft Azure			
		$0°$	$45°$	$90°$	avg	$0°$	$45°$	$90°$	avg
Frontal lighting	Young	0.75	0.81	0.77	0.78	0.55	0.51	0.49	0.52
	Middle_Aged	0.54	0.53	0.51	0.53	0.73	0.72	0.66	0.70
	Senior	0.87	0.86	0.84	0.85	0.84	0.82	0.83	0.83
	Male	0.94	0.93	0.83	0.90	0.99	0.99	1.00	0.99
	Eyeglasses	0.97	0.99	0.96	0.97	0.99	0.99	1.00	0.99
	Beard	0.89	0.88	0.91	0.90	0.86	0.85	0.84	0.85
	Mustache	0.98	0.97	0.90	0.95	0.56	0.55	0.61	0.58
	Average (avg)	0.85	0.85	0.82	0.84	0.79	0.78	0.78	0.78
	Rejected images	0.00	0.00	0.02	0.01	0.04	0.13	0.85	0.34
Side lighting	Young	0.75	0.80	0.73	0.76	0.56	0.52	0.92	0.67
	Middle_Aged	0.54	0.52	0.49	0.52	0.69	0.72	0.88	0.76
	Senior	0.87	0.84	0.79	0.83	0.77	0.81	0.96	0.85
	Male	0.89	0.92	0.86	0.89	0.99	0.99	1.00	0.99
	Eyeglasses	0.97	0.98	0.91	0.95	0.99	0.99	1.00	0.99
	Beard	0.89	0.89	0.96	0.91	0.87	0.91	1.00	0.93
	Mustache	0.98	0.95	0.67	0.87	0.58	0.60	0.48	0.55
	Average (avg)	0.84	0.84	0.77	0.82	0.78	0.79	0.89	0.82
	Rejected images	0.00	0.03	0.32	0.12	0.13	0.50	0.99	0.54

B. Evaluating MAC Performance under Pose and Lighting

In this section, we evaluate the performance of the MAC against the industry product Microsoft Azure[3] under different headposes and lighting conditions. The experiment was conducted on the ColorFeret database [35] consisting of over 11k images with different poses such as frontal (2.7k), profile (2.7k) and head poses in between (5.9k). Figure 4 shows some sample images visualizing the head poses and lighting conditions. While our MAC classifier is able to accurately state 47 different attributes, we focus on the attributes that

[3]https://azure.microsoft.com/en-us/services/cognitive-services/face

Male	1	Bangs	-1	Round Face	-1	Big Lips	0
Young	0	Sideburns	-1	Double Chin	-1	Big Nose	0
Middle Aged	1	Black Hair	1	High Cheekbones	-1	Pointy Nose	1
Senior	-1	Blond Hair	-1	Chubby	-1	Heavy Makeup	-1
Asian	-1	Brown Hair	0	Obstructed Forehead	0	Wearing Hat	-1
White	1	Gray Hair	-1	Fully Visible Forehead	0	Wearing Earrings	-1
Black	-1	No Beard	0	Brown Eyes	0	Wearing Necktie	1
Rosy Cheeks	-1	Mustache	0	Bags Under Eyes	1	Wearing Lipstick	-1
Shiny Skin	0	5 o Clock Shadow	0	Bushy Eyebrows	1	No Eyewear	1
Bald	-1	Goatee	0	Arched Eyebrows	-1	Eyeglasses	-1
Wavy Hair	-1	Oval Face	0	Mouth Closed	-1	Attractive	-1
Receding Hairline	0	Square Face	1	Smiling	1		



Male	1	Bangs	-1	Round Face	0	Big Lips	0
Young	-1	Sideburns	1	Double Chin	1	Big Nose	1
Middle Aged	-1	Black Hair	0	High Cheekbones	0	Pointy Nose	-1
Senior	1	Blond Hair	-1	Chubby	1	Heavy Makeup	-1
Asian	-1	Brown Hair	-1	Obstructed Forehead	1	Wearing Hat	1
White	0	Gray Hair	1	Fully Visible Forehead	-1	Wearing Earrings	-1
Black	-1	No Beard	-1	Brown Eyes	0	Wearing Necktie	-1
Rosy Cheeks	0	Mustache	1	Bags Under Eyes	0	Wearing Lipstick	-1
Shiny Skin	1	5 o Clock Shadow	-1	Bushy Eyebrows	1	No Eyewear	1
Bald	-1	Goatee	-1	Arched Eyebrows	-1	Eyeglasses	-1
Wavy Hair	-1	Oval Face	-1	Mouth Closed	0	Attractive	-1
Receding Hairline	0	Square Face	1	Smiling	0		

Male	-1	Bangs	1	Round Face	0	Big Lips	0
Young	1	Sideburns	-1	Double Chin	-1	Big Nose	-1
Middle Aged	-1	Black Hair	-1	High Cheekbones	1	Pointy Nose	1
Senior	-1	Blond Hair	-1	Chubby	-1	Heavy Makeup	1
Asian	-1	Brown Hair	1	Obstructed Forehead	0	Wearing Hat	-1
White	1	Gray Hair	-1	Fully Visible Forehead	-1	Wearing Earrings	1
Black	-1	No Beard	1	Brown Eyes	0	Wearing Necktie	-1
Rosy Cheeks	0	Mustache	-1	Bags Under Eyes	-1	Wearing Lipstick	1
Shiny Skin	0	5 o Clock Shadow	-1	Bushy Eyebrows	-1	No Eyewear	1
Bald	-1	Goatee	-1	Arched Eyebrows	-1	Eyeglasses	-1
Wavy Hair	1	Oval Face	0	Mouth Closed	0	Attractive	1
Receding Hairline	-1	Square Face	-1	Smiling	0		

Male	-1	Bangs	-1	Round Face	-1	Big Lips	-1
Young	1	Sideburns	-1	Double Chin	-1	Big Nose	-1
Middle Aged	-1	Black Hair	0	High Cheekbones	1	Pointy Nose	1
Senior	-1	Blond Hair	0	Chubby	-1	Heavy Makeup	1
Asian	-1	Brown Hair	0	Obstructed Forehead	-1	Wearing Hat	-1
White	1	Gray Hair	0	Fully Visible Forehead	1	Wearing Earrings	1
Black	-1	No Beard	1	Brown Eyes	-1	Wearing Necktie	-1
Rosy Cheeks	0	Mustache	-1	Bags Under Eyes	-1	Wearing Lipstick	1
Shiny Skin	0	5 o Clock Shadow	-1	Bushy Eyebrows	-1	No Eyewear	1
Bald	-1	Goatee	-1	Arched Eyebrows	0	Eyeglasses	-1
Wavy Hair	1	Oval Face	1	Mouth Closed	0	Attractive	1
Receding Hairline	-1	Square Face	-1	Smiling	1		

Male	-1	Bangs	-1	Round Face	0	Big Lips	1
Young	0	Sideburns	-1	Double Chin	-1	Big Nose	-1
Middle Aged	1	Black Hair	1	High Cheekbones	1	Pointy Nose	0
Senior	-1	Blond Hair	-1	Chubby	-1	Heavy Makeup	1
Asian	1	Brown Hair	0	Obstructed Forehead	-1	Wearing Hat	-1
White	-1	Gray Hair	-1	Fully Visible Forehead	1	Wearing Earrings	1
Black	-1	No Beard	1	Brown Eyes	1	Wearing Necktie	-1
Rosy Cheeks	-1	Mustache	-1	Bags Under Eyes	-1	Wearing Lipstick	1
Shiny Skin	1	5 o Clock Shadow	-1	Bushy Eyebrows	-1	No Eyewear	0
Bald	-1	Goatee	-1	Arched Eyebrows	1	Eyeglasses	-1
Wavy Hair	1	Oval Face	0	Mouth Closed	-1	Attractive	1
Receding Hairline	-1	Square Face	-1	Smiling	-1		

Fig. 2: Samples images from MAAD-Face with the corresponding 47 attribute-annotations.

TABLE III: Attribute annotation analysis of MAAD-Face based on the ground truth of three human evaluators. The annotation quality is reported in terms of accuracy, precision, and recall. Main source describe from which dataset most of the annotations are transferred from.

Main source	Category	Class	Attribute	Accuracy	Precision	Recall
CelebA	Demographics	Gender	Male	0.99	0.98	1.00
CelebA		Age	Young	0.99	1.00	0.98
LFW			Middle Aged	0.93	0.98	0.89
LFW			Senior	0.97	0.96	0.98
LFW		Race	Asian	0.90	0.88	0.92
LFW			White	0.89	1.00	0.82
LFW			Black	0.94	0.90	0.98
CelebA	Skin	Rosy Cheeks	Rosy Cheeks	0.99	0.98	1.00
LFW		Shiny Skin	Shiny Skin	0.77	0.84	0.74
CelebA	Hair	Hairstyle	Bald	0.96	0.92	1.00
CelebA			Wavy Hair	0.99	1.00	0.98
CelebA		Receding Hairline	Receding Hairline	0.77	0.54	1.00
CelebA		Bangs	Bangs	0.98	0.96	1.00
CelebA		Sideburns	Sideburns	0.93	0.88	0.98
CelebA		Haircolor	Black Hair	0.98	0.96	1.00
CelebA			Blond Hair	1.00	1.00	1.00
CelebA			Brown Hair	0.97	0.94	1.00
CelebA			Gray Hair	0.95	0.90	1.00
CelebA	Beard	Beard	No Beard	0.98	1.00	0.96
CelebA			Mustache	0.98	0.98	0.98
CelebA			5 o Clock Shadow	0.97	0.94	1.00
CelebA			Goatee	0.95	0.90	1.00
LFW	Face Geometry	Face Shape	Oval Face	0.81	0.90	0.76
LFW			Square Face	0.80	0.78	0.81
LFW			Round Face	0.69	0.56	0.76
CelebA		Double Chin	Double Chin	0.94	0.88	1.00
CelebA		High Cheekbones	High Cheekbones	0.92	0.92	0.92
CelebA		Chubby	Chubby	0.94	0.88	1.00
LFW		Forehead visibility	Obstructed Forehead	0.91	0.94	0.89
LFW			Fully Visible Forehead	0.80	0.75	1.00
LFW	Periocular	Brown Eyes	Brown Eyes	0.68	0.44	0.85
LFW		Bags Under Eyes	Bags Under Eyes	0.68	0.40	0.91
CelebA		Bushy Eyebrows	Bushy Eyebrows	0.95	0.94	0.96
CelebA		Arched Eyebrows	Arched Eyebrows	1.00	1.00	1.00
LFW	Mouth	Mouth Closed	Mouth Closed	0.84	0.80	0.87
CelebA		Smiling	Smiling	0.95	1.00	0.91
LFW		Big Lips	Big Lips	0.70	0.50	0.83
CelebA	Nose	Nose type	Big Nose	0.97	0.98	0.96
LFW			Pointy Nose	0.88	0.88	0.88
CelebA	Accessories	Heavy Makeup	Heavy Makeup	0.98	0.98	0.98
CelebA		Wearing Hat	Wearing Hat	0.92	0.84	1.00
CelebA		Wearing Earrings	Wearing Earrings	0.83	0.70	0.95
LFW		Wearing Necktie	Wearing Necktie	0.91	0.84	0.98
CelebA		Wearing Lipstick	Wearing Lipstick	0.95	0.90	1.00
LFW		Eyeglasses	No Eyewear	0.98	0.98	0.98
CelebA			Eyeglasses	0.90	0.80	1.00
CelebA	Other	Attractive	Attractive	1.00	1.00	1.00
		Total		0.91	0.87	0.94

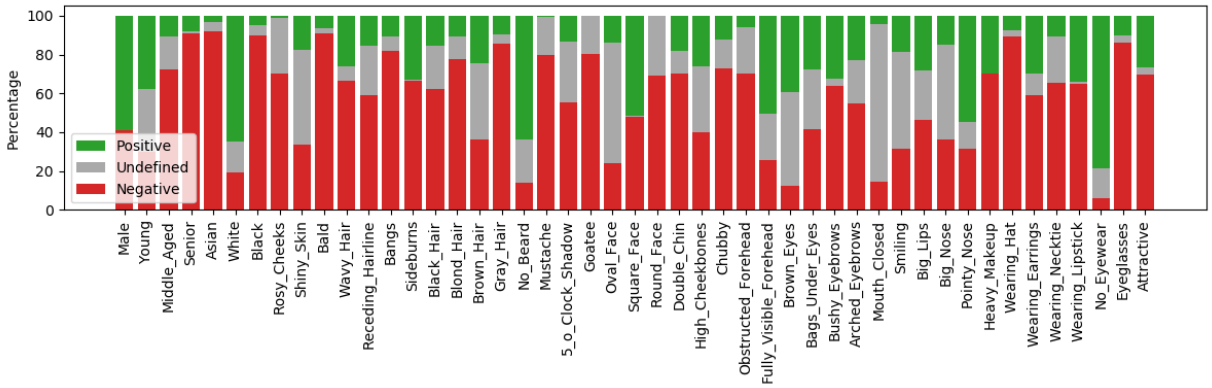


Fig. 3: Annotation distribution of the proposed MAAD-Face database. For each of the 47 attributes, green indicates the percentage of positive annotations, red indicates the percentage of negatively annotated images, and grey represents the percentage of images that have an undefined annotation for the attribute.

(a) (b) (c) (d) (e) (f)

Fig. 4: Sample Images [35] for the pose and lighting evaluation. Three different head poses are considered from frontal to profile images. To also analyse the effect of lighting, artificial lighting was introduced (see Figures d-f).

Microsoft Azure and the ColorFeret dataset have in common. This allows a comparison of our approach with an industry product under several conditions.

Table II shows the described analysis. In general, our MAC classifier reaches high performances on most attributes and additionally turns out to be relatively robust against the changes in the head poses as well as the lighting. The weakest performance is observed for the only vaguely-defined age classes. Similar to our MAC approach, Microsoft Azure turns out to have accurate and robust predictions under both lighting conditions. However, Microsoft Azure tends to reject many images when it comes to non-frontal poses or difficult lighting conditions. Under optimal lighting conditions, 85% of the profile images are rejected and in combination with the more challenging lighting, over 99% of the images are rejected without any predictions. This short analysis showed that our MAC classifier used for the proposed annotation transfer pipeline is well suited and reaches comparable performances to the industry product Microsoft Azure.

C. Evaluating Annotation-Correctness

In this section, we evaluate the quality of attribute annotations from three face datasets, LFW, CelebA, and MAAD-Face. The quality refers to the correctness of the annotations compared to the annotations of human evaluators. The annotation-correctness of each attribute in LFW, CelebA, and MAAD-Face was manually evaluated by three human evaluators. For each attribute, the evaluators got 50 positively-annotated and 50 negatively-annotated images to prevent a bias evaluation due to unbalanced testing data. Since the choice of the images submitted to the evaluators affect the correctness evaluation, the selection of these images is based on a randomized process to prevent a human-bias in the choice of these images. Then, each evaluator was asked to carefully annotate these images for the given attribute. This led to over 16k manually created annotations[4]. The manually created annotations are used to compute the accuracy, precision, and recall for each attribute of the database. The accuracy refers to the percentage of correct annotations, where the ground truth is determined by the human evaluators. Precision is defined as the number of true positives over the number of true and false positives In our context, precision refers to "What proportion of positive annotated samples in the database is also positively annotated by the human evaluators?". Recall is defined as the number of true positives over the number of true positives and false negatives. In our context, recall refers to "What proportion of positive human annotations are identified correctly?". Tables V, VI, and III present the results for this analysis on LFW, CelebA, and MAAD-Face.

1) LFW: For LFW (Table V), many attributes show a very weak performance and thus, a low correlation with the annotations of the human evaluators. Young age group annotations (baby, child, youth) are close to a random accuracy and additionally often have a small precision. This is also observed e.g. for *frowning, chubby, curly hair, wavy hair, bangs, goatee, and square face*. Moreover, annotations for *attractive man* are mostly placed on female faces. In general, there is a big mismatch between the annotations of LFW and the annotations of the human evaluators. The accuracy for most attributes is below 80% and only 5 out of 76 attributes have an accuracy of over 90%. Over all attributes, this leads to an accuracy of 72%, a precision of 61%, and a recall of 84%.

[4]Please note that this only represents a small fraction of all annotations and additionally reflects the subjective opinion of the three evaluators. Therefore, the results should not be considered as absolute values but should rather be used as indicators.

TABLE IV: Analysis of the attribute annotation correctness. The correctness of the attribute annotations is shown for the most relevant databases, LFW, CelebA, and MAAD-Face, since these contain annotations for a high number of distinct attributes. The correctness was evaluated by three human evaluators. In total, MAAD-Face does not only provide the highest number of attribute annotations, it additionally provides annotations of much higher quality than related databases.

			Attribute annotations		Attribute annotation correctness		
Database	Num. of subjects	Num. of images	Distinctive attributes	Number of annotations	Accuracy	Precision	Recall
LFW	5.7k	13.2k	74	0.9M	0.72	0.61	0.84
CelebA	10.0k	0.2M	40	8.0M	0.85	0.83	0.89
MAAD-Face (this paper)	9.1k	3.3M	47	123.9M	**0.91**	**0.87**	**0.94**

The high gap between the low precision and the relatively high recall indicates that there are a lot of false-positive annotations in LFW.

CelebA: The attribute performance for CelebA is shown in Table VI. It has annotations for 40 binary attribute, which is a lower number than on LFW. However, these annotations are of much higher quality. Only 2 attributes have an accuracy of less than 70% and 14 attributes even reach over 90%. Over all attributes, the accuracy is 85%, the precision is 83%, and the recall is 89%. Similar to LFW, there is a tendency that most of the wrong annotations are within the positives.

MAAD-Face: Table III shows the attribute performance of MAAD-Face. MAAD-Face has 47 binary attributes. In the evaluation against the human annotations, 3 attributes reach a performance of below 70%. However, also 34 attributes reach over 90% accuracy with the majority of close to 100%. Over all attributes, this leads to an accuracy of 91%, a precision of 87%, and a recall of 94%.

Summary: Table IV shows the properties of the investigated databases including the overall performance of our annotation-correctness study. Even though LFW provides the highest number of binary attributes, it provides the lowest number of attribute annotations with the lowest annotation qualities. Only 72% of the investigated annotations match the annotations of the human evaluators. CelebA consists of 8.0M attribute annotations of 40 binary attributes. Moreover, with an accuracy of 85%, the quality of these annotations is significantly higher. In terms of numbers of annotations and annotation-quality, MAAD-Face exceeds the other databases. It provides 47 binary attributes with a total of 123.9M annotations. This is 15 times higher than CelebA and 137 times higher than LFW. Moreover, the annotations quality (in terms of accuracy, precision, and recall) is significantly higher than the other databases. 91% of the MAAD-Face annotations match the annotations of the human evaluators. Consequently, MAAD-Face provides significantly more and higher-quality attribute annotations.

V. Soft-Biometric based Identity Recognition

In this section, we evaluate the discrimination strength of soft-biometric attributes for identity verification and identification. The use of these attribute might be especially interesting for identity recognition applications with short time windows between the reference and probe images, such as person re-identification. In the following of this section, we describe the experimental setup. Afterwards, the results are presented, discussed, and summarized.

A. Experimental Setup

The high number of face annotations with sufficient quality in MAAD-Face allow us to investigate the usefulness of soft-biometrics for face recognition. For the experiments, the MAAD-Face database is divided into a 20% training and a 80% test set in a subject exclusive manner. As a result, the training set contains around 630k samples while the test set contains around 2.5M instances. This test/train division allows to make the use of a large test set while the training set size is still suitable for training a linear (logistic regression) model [3].

For the identification experiments, the test set is further divided into a reference and a probe set. From each identity, one sample with the most annotated attributes is placed in the reference. All the others are placed in the probe set. The identification performance is reported in terms of Cumulative Match Characteristic (CMC) curves [8]. It measures the identification performance based on the relative ordering of match scores corresponding to each biometric sample in a closed-set identification scenario. For the verification experiments, all samples pairs are considered. The face verification performance is reported in terms of false non-match rate (FNMR) at a fixed false match rate (FMR). These measures are specified for biometric verification evaluation in the international standard [24].

To ensure that the comparison of two samples will contain a sufficient number of attributes that jointly appear in both samples, we neglect comparisons with less than 10 overlapping attribute annotations (not valid). We set this constraint of at least 10 overlapping attributes since we see a direct relation between the number of considered attributes and the expected accuracy of the made decision. For a low number of overlapping attribute annotations, the decision is made based on less information and thus, a false decision is more likely. For a high number of overlaps, the decision takes into account more information and therefore, it is more likely that the decision will be correct. Figure 5 shows the probability that a comparison is valid depending on the number of used (most important) attributes. It can be seen that if a comparison is made using 20 or more of the most important attributes, the probability that the comparison is neglected is very low.

The comparison of two samples is made with a joint feature representation. Therefore, a joint (soft-biometric) feature representation

$$x(x_{ref}, x_{probe}) = [x_1^{a_1}, x_2^{a_1}, x_3^{a_1}, x_1^{a_2}, x_2^{a_2}, x_3^{a_2}, \dots] \quad (1)$$

TABLE V: Attribute annotation analysis of LFW based on the ground truth of three human evaluators. The annotation quality is reported in terms of accuracy, precision, and recall.

Attribute	Acc	Precision	Recall
Male	0.89	0.96	0.84
Asian	0.86	0.74	0.97
White	0.74	0.98	0.66
Black	0.91	0.84	0.98
Baby	0.54	0.08	1.00
Child	0.55	0.10	1.00
Youth	0.56	0.14	0.88
Middle Aged	0.67	0.90	0.62
Senior	0.87	0.94	0.82
Black Hair	0.78	0.88	0.73
Blond Hair	0.91	0.84	0.98
Brown Hair	0.70	0.60	0.75
Bald	0.74	0.50	0.96
No Eyewear	0.91	0.98	0.86
Eyeglasses	0.91	0.88	0.94
Sunglasses	0.86	0.72	1.00
Moustache	0.84	0.72	0.95
Smiling	0.87	0.80	0.93
Frowning	0.61	0.22	1.00
Chubby	0.53	0.16	0.62
Blurry	0.69	0.90	0.63
Harsh Lighting	0.64	0.92	0.59
Flash	0.73	0.66	0.77
Soft Lighting	0.75	0.66	0.80
Outdoor	0.83	0.82	0.84
Curly Hair	0.51	0.02	1.00
Wavy Hair	0.50	0.08	0.50
Straight Hair	0.60	0.78	0.54
Receding Hairline	0.75	0.62	0.84
Bangs	0.54	0.08	1.00
Sideburns	0.61	0.40	0.69
Fully Visible Forehead	0.79	1.00	0.70
Partially Visible Forehead	0.82	0.80	0.83
Obstructed Forehead	0.62	0.24	1.00
Bushy Eyebrows	0.64	0.42	0.75
Arched Eyebrows	0.79	0.80	0.78
Narrow Eyes	0.69	0.46	0.85

Attribute	Acc	Precision	Recall
Eyes Open	0.73	0.96	0.66
Big Nose	0.75	0.54	0.93
Pointy Nose	0.80	0.82	0.79
Big Lips	0.73	0.56	0.85
Mouth Closed	0.86	0.82	0.89
Mouth Slightly Open	0.79	0.88	0.75
Mouth Wide Open	0.93	0.88	0.98
Teeth Not Visible	0.86	0.78	0.93
No Beard	0.69	1.00	0.62
Goatee	0.62	0.24	1.00
Round Jaw	0.77	0.76	0.78
Double Chin	0.66	0.34	0.94
Wearing Hat	0.69	0.40	0.95
Oval Face	0.59	0.78	0.57
Square Face	0.55	0.12	0.86
Round Face	0.81	0.72	0.88
Color Photo	0.57	1.00	0.54
Posed Photo	0.64	0.32	0.89
Attractive Man	0.62	0.26	0.93
Attractive Woman	0.75	0.50	1.00
Indian	0.65	0.32	0.94
Gray Hair	0.89	0.94	0.85
Bags Under Eyes	0.75	0.76	0.75
Heavy Makeup	0.88	0.76	1.00
Rosy Cheeks	0.63	0.30	0.88
Shiny Skin	0.66	0.44	0.79
Pale Skin	0.82	0.90	0.78
5 o Clock Shadow	0.59	0.18	1.00
Strong Nose-Mouth Lines	0.86	0.88	0.85
Wearing Lipstick	0.81	0.64	0.97
Flushed Face	0.61	0.28	0.82
High Cheekbones	0.81	0.70	0.90
Brown Eyes	0.44	0.46	0.44
Wearing Earrings	0.79	0.58	1.00
Wearing Necktie	0.76	0.66	0.83
Wearing Necklace	0.61	0.22	1.00
Total	0.72	0.61	0.84

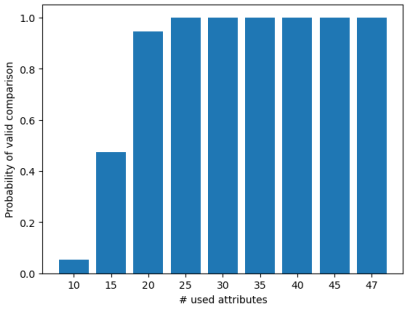


Fig. 5: Probability that a comparison is valid depending on the number of (most important) attributes used for the comparison. A comparison is considered as valid if at least 10 attributes are annotated in both, the probe and the reference sample.

is computed and a hamming-based and a logistic regression model is utilized for the comparison process itself. The joint feature representation for the attribute a

$$x_i^a(x_{ref}, x_{probe}) = \begin{cases} 1 & \text{if } i = 1 \ \& \ x_{ref}^a = x_{probe}^a = \textit{True} \\ 1 & \text{if } i = 2 \ \& \ x_{ref}^a = x_{probe}^a = \textit{False} \\ 1 & \text{if } i = 3 \ \& \ x_{ref}^a \neq x_{probe}^a \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

of a reference sample x_{ref} and a probe sample x_{probe} is defined binary by the relation of a depending if the attribute of both samples is annotated as both as True (x_1^a), both as False (x_2^a), or differently (x_3^a). We chose this kind of representation to ensure that the comparison models can additionally learn the relation between different attributes. Two simple comparison models are used for the experiments that exploit the joint feature representation. The first one is a hamming-based model that simply determines the number of equally-annotated attributes in a normalized manner. The comparison score of

TABLE VI: Attribute annotation analysis of CelebA based on the ground truth of three human evaluators. The annotation quality is reported in terms of accuracy, precision, and recall.

Attribute	Acc	Precision	Recall
5 o Clock Shadow	0.85	0.74	0.95
Arched Eyebrows	0.89	0.92	0.87
Attractive	0.81	0.74	0.86
Bags Under Eyes	0.80	0.80	0.80
Bald	0.84	0.68	1.00
Bangs	0.75	0.50	1.00
Big Lips	0.73	0.84	0.69
Big Nose	0.79	0.86	0.75
Black Hair	0.87	0.96	0.81
Blond Hair	0.94	0.94	0.94
Blurry	0.88	0.78	0.98
Brown Hair	0.90	0.88	0.92
Bushy Eyebrows	0.81	0.78	0.83
Chubby	0.83	0.66	1.00
Double Chin	0.76	0.58	0.91
Eyeglasses	0.96	0.92	1.00
Goatee	0.93	0.94	0.92
Gray Hair	0.98	0.98	0.98
Heavy Makeup	0.90	0.92	0.88
High Cheekbones	0.88	0.86	0.90
Male	1.00	1.00	1.00
Mouth Slightly Open	0.90	0.88	0.92
Mustache	0.95	0.94	0.96
Narrow Eyes	0.86	0.82	0.89
No Beard	0.91	1.00	0.85
Oval Face	0.62	0.92	0.58
Pale Skin	0.85	0.92	0.81
Pointy Nose	0.83	0.94	0.77
Receding Hairline	0.66	0.38	0.86
Rosy Cheeks	0.78	0.70	0.83
Sideburns	0.84	0.88	0.81
Smiling	0.94	0.92	0.96
Straight Hair	0.83	1.00	0.75
Wavy Hair	0.82	0.66	0.97
Wearing Earrings	0.93	0.88	0.98
Wearing Hat	1.00	1.00	1.00
Wearing Lipstick	0.91	0.90	0.92
Wearing Necklace	0.86	0.80	0.91
Wearing Necktie	0.85	0.72	0.97
Young	0.75	0.52	0.96
Total	0.85	0.83	0.89

this model is given by

$$s(x) = 1 - NHD(x), \tag{3}$$

where *NHD* counts the number of 1's in x and divides it by the number of attributes $|\mathcal{A}|$. The second one makes use of the training set and trains a logistic regression model on the joint feature representations. The choice of a simple linear model prevents overfitting and additionally allows to determine the importance of each soft-biometric attribute.

B. Results

To evaluate the discriminateness of soft-biometric attributes for recognition, in Section V-B1 the attribute importance for the recognition decision is presented. In Section V-B2 the verification and identification performance based on the attribute information is reported. Section V-B3 demonstrates how well these attributes can support hard face biometrics in verification and identification tasks. Finally, the findings are summarized in Section V-B4.

1) Attribute Importance: To get an understanding of which attributes support making accurate genuine and imposter decisions, Figure 6 shows the attribute importance derived from the logistic regression model. A green bar refers to the contribution of an attribute for genuine decisions while a red bar indicates the contribution of an attribute for imposter decisions. The top figure shows the feature importance for *True-True* annotations, the middle figure for *False-False* annotations, and the bottom figure demonstrates the feature importance if the attribute annotation for one attribute is different (*True-False*) for the probe and the reference sample.

In Figure 6 it can be clearly seen that the top two figures show mostly green bars while the figure on the bottom is mostly red. This indicates that if the probe and reference faces share the same soft-biometric attributes, it supports a genuine decision while not sharing an attribute strongly supports imposter decisions. It turns out that the attribute "gender" has the strongest discriminative strength of all investigated attributes. This is reasonable since (a) it is the most annotated attribute, (b) it helps to eliminate many potential candidates, and (c) it is hard to change this attribute. But also hairstyle, haircolor, and wearing a beard have a significant impact on the recognition decision as well as more permanent factors of the face such as bushy eyebrows, big lips, and a pointy nose. Surprisingly wearing eyeglasses strongly supports genuine decisions but not wearing eyeglasses is of no real significance. This might be explained by the fact that only a smaller percentage of the faces in the database have glasses but if people wear glasses they usually wear them permanently.

Please note that, although the database is of significant size, these results should only be interpreted as indications since (a) the underlying annotation distributions affects the results and (b) the utilized logistic regression model might lead to oversimplified (linear) conclusions.

2) Identity Recognition based on Soft-Biometrics only: To analyse how well soft-biometric attributes can be used for identity recognition, Figure 7 shows the face recognition performance using soft-biometric annotations only. In Figure 7a, the verification performance is reported as an ROC curve including area under the curve (AUC) and equal error rate (EER) values. Figures 7b and 7c shows the closed- and open-set identification performance in terms of CMC and DET curves. The plots follow the definitions of the international standards [24]. The performance is reported using the logistic regression model on all attributes and using the hamming-based model on different numbers of the most relevant attributes. Moreover, considering the current pandemic times, a special subset of soft-biometric attributes is considered that can be reliably extracted in the presence of a face mask. This excludes attributes related to beards, mouth, nose, and face geometry ... as well as the attributes *Wearing Lipstick*, *Rosy Cheeks*, *Bangs*, and *Heavy Makeup*. In total, 25 attributes are included in this subset.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2021.3096120, IEEE Transactions on Information Forensics and Security

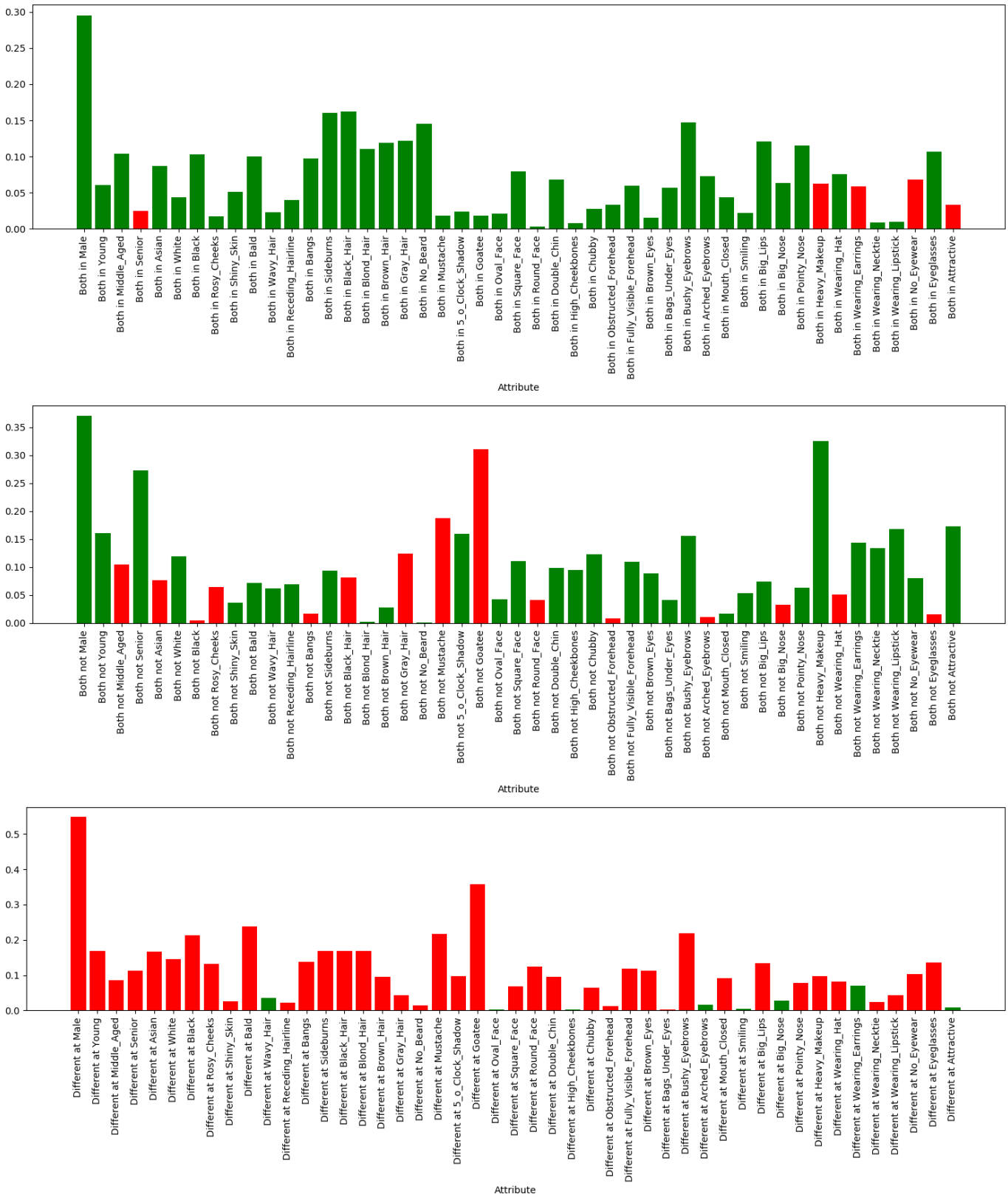


Fig. 6: Importance for each soft-biometric attribute derived from a logistic regression model. Green indicate the importance of genuine decisions while red indicates the importance of imposter decisions. The top figure shows the feature importance for *True-True* annotations, the middle figure for *False-False* annotations, and the bottom figure demonstrates the feature importance if the attribute annotation for one attribute is different (*True-False*) for the probe and the reference sample.

The results demonstrate that it is possible to use soft-biometric attributes for both, verification and identification. Previous works [13], [1], [36] on person identification based on soft-biometrics only reported higher performances than in our experiments. However, these works operate on data captured in strongly controlled conditions only. We fill this research gap by using data that were captured in strongly uncontrolled conditions and thus, posses large variations. Most annotations used in the experiments are non-permanent and the images used for the comparisons are captured with larger time differences. Even if such accuracy may not be enough for specific applications, we provide a starting point for further research that consider real life scenarios. In general, a higher number of considered attributes for the hamming-based model increases the performance in all three scenarios. Moreover, the strongest performances are observed with the logistic regression based model since it is able to weight the importance of each attribute.

3) Face Biometrics Supported by Soft-Biometrics: To analyse how well soft-biometrics can support hard face biometrics, Figure 8 shows the face verification and identification performance using face embeddings only and in combination with soft-biometrics. The verification performance is reported as an ROC curve with area under the curve (AUC) and equal error rate (EER) values. The identification performance is reported as a CMC curve for closed-set identification and as a DET curve for open-set identification. For the face embeddings, we utilized the widely-used FaceNet model[5] [42]. The performance using face embeddings only is reported as well as in combination with soft-biometrics. The combined approaches are based on a simple EER-based weighted-score fusion approach as described [6]. The results show that for open-set identification, the score-fusion approach with soft-biometric attributes does not show significant differences compared to using the face embeddings only. However, for the verification and closed-set identification scenarios, the additional soft-biometric information is able to enhance the recognition performance.

4) Recognition with Soft-Biometrics - Summary: In this section, we analysed how well soft-biometrics can be used for recognition. The annotations came from the proposed MAAD-Face annotations database. First, we determined the attribute importance for genuine and imposter decisions. Second, we investigated how many of these attributes are needed to achieve certain verification and identification performances given the soft-biometrics only. Despite that the data was collected over large time-windows and the annotations are mainly of non-permanent nature, a decent verification and identification performance observed using 25 and more attribute annotations. These results might be useful for re-identification scenarios or description-based identity search. Lastly, we demonstrated that soft-biometrics can support hard face biometrics in verification and closed-set identification scenarios.

[5]https://github.com/davidsandberg/facenet

VI. Conclusion

Soft-biometric attributes play a major role in the development of various face recognition topics, such as bias-mitigating, information fusion, and privacy-preserving face recognition solutions. To support the developments in these fields, in this work, we presented four contributions. (1) A novel annotation transfer pipeline is proposed that allows to transfer attribute annotations of high accuracy from multiple source datasets to a target dataset. This pipeline is used to create MAAD-Face. (2) MAAD-Face is a novel face annotations database that provides over 3.3M faces with 123.9M annotations of 47 different attributes. To the best of our knowledge, MAAD-Face is the publicly available database that provides the largest number of attribute annotations. (3) We analyse the correctness of the attribute annotations of three annotated face databases, CelebA, LFW, and MAAD-Face. The evaluation was performed manually by three human evaluators and demonstrated that the attribute annotations of MAAD-Face are of significantly higher quality than related databases. (4) Finally, the large number of high-quality annotations of MAAD-Face are used to study how well soft-biometrics can be used for identity recognition. The advantage of the proposed annotation-transfer pipeline is that it allows transferring arbitrary attributes from a database to images while it ensures a high correctness of the transferred annotations. This leads to attribute annotations of higher quality than related databases as the annotation correctness evaluation showed. The high correctness is ensured by the use of accurate prediction reliabilities. However, the use of this reliability might also lead to annotations correlating with specific situations in which the classifier is confident about its predictions. This has to be addressed by future work. We hope that this work will support the development of novel face recognition technologies.

Acknowledgment

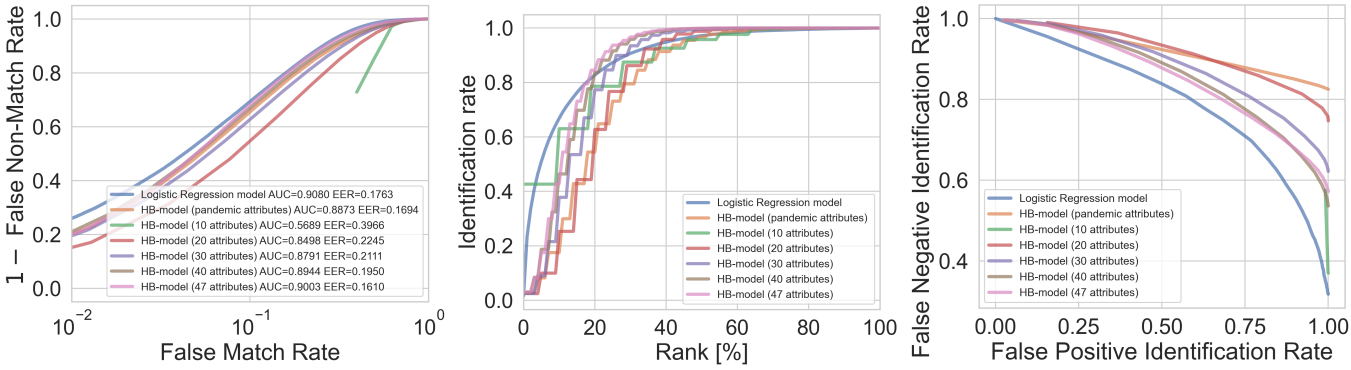
This work was supported by the German Federal Ministry of Education and Research (BMBF) as well as by the Hessen State Ministry for Higher Education, Research and the Arts (HMWK) within the National Research Center for Applied Cybersecurity (ATHENE), and in part by the German Federal Ministry of Education and Research (BMBF) through the Software Campus project. Portions of the research in this paper use the FERET database of facial images collected under the FERET program, sponsored by the DOD Counterdrug Technology Development Program Office.

References

[1] N. Almudhahka, M. S. Nixon, and J. S. Hare. Human face identification via comparative soft biometrics. In *IEEE International Conference on Identity, Security and Behavior Analysis, ISBA 2016, Sendai, Japan, February 29 - March 2, 2016*, pages 1–6. IEEE, 2016.

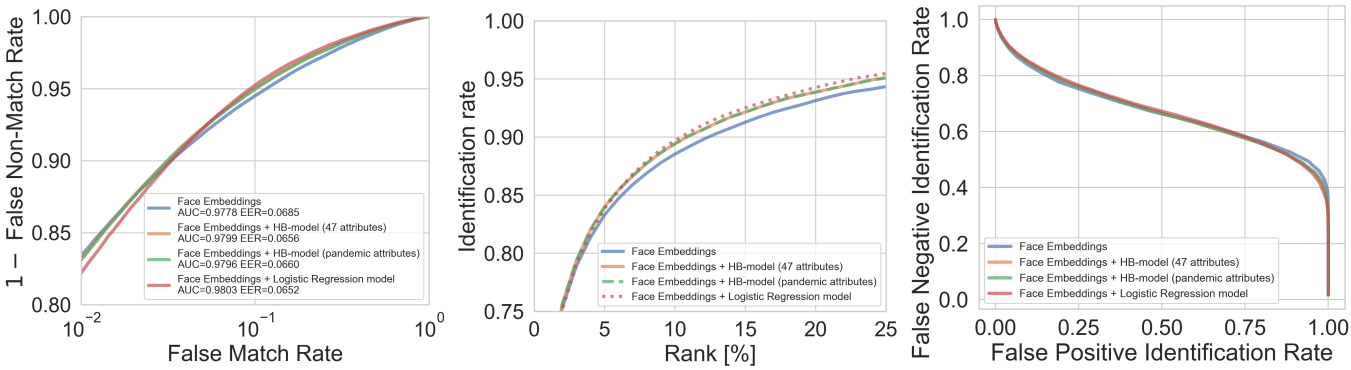
[2] F. Alonso-Fernandez, K. Hernandez-Diaz, S. Ramis, F. J. P. López, and J. Bigün. Soft-biometrics estimation in the era of facial masks. In A. Brömme, C. Busch, A. Dantcheva, K. B. Raja, C. Rathgeb, and A. Uhl, editors, *BIOSIG 2020 - Proceedings of the 19th International Conference of the Biometrics Special Interest Group, online, 16.-18. September 2020*, volume P-306 of *LNI*, pages 11–19. Gesellschaft für Informatik e.V., 2020.

[3] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.



(a) Verification (b) Closed-set identification (c) Open-set identification

Fig. 7: Analysis of the verification and identification performance based on different choices of soft-biometrics only. The verification performance is reported as an ROC curve with area under the curve (AUC) and equal error rate (EER) values. The identification performance is reported as a CMC curve for closed-set identification and as a DET curve for open-set identification. Pandemic attributes refer to attributes reliably detectable from a face in presence of a face mask. In general, the performance increases with a higher number of considered attributes.



(a) Verification (b) Closed-set identification (c) Open-set identification

Fig. 8: Analysis on how well soft-biometrics can support hard face biometrics. The verification performance is reported as an ROC curve with area under the curve (AUC) and equal error rate (EER) values. The identification performance is reported as a CMC curve for closed-set identification and as a DET curve for open-set identification. The performance using face embeddings only is reported as well as in combination with soft-biometrics. The combined approaches are achieved through a simple weighted-score fusion approach. For verification and closed-set identification, soft-biometric information enhances the performance.

[4] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018.

[5] J. Chen, A. Kumar, R. Ranjan, V. M. Patel, A. Alavi, and R. Chellappa. A cascaded convolutional neural network for age estimation of unconstrained faces. In *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8, 2016.

[6] N. Damer, A. Opel, and A. Nouak. Biometric source weighting in multi-biometric fusion: Towards a generalized and robust solution. In *22nd European Signal Processing Conference, EUSIPCO 2014, Lisbon, Portugal, September 1-5, 2014*, pages 1382–1386. IEEE, 2014.

[7] A. Dantcheva, P. Elia, and A. Ross. What else does your biometric data reveal? A survey on soft biometrics. *IEEE Trans. Inf. Forensics Secur.*, 11(3):441–467, 2016.

[8] B. DeCann and A. Ross. Relating roc and cmc curves via the biometric menagerie. In *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–8, 2013.

[9] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, 1(2):89–103, 2020.

[10] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. *Trans. Info. For. Sec.*, 9(12):2170–2179, Dec. 2014.

[11] R. S. Geiger, K. Yu, Y. Yang, M. Dai, J. Qiu, R. Tang, and J. Huang. Garbage in, garbage out?: do machine learning application papers in social computing report where human-labeled training data comes from? In M. Hildebrandt, C. Castillo, E. Celis, S. Ruggieri, L. Taylor, and G. Zanfir-Fortuna, editors, *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 325–336. ACM, 2020.

[12] X. Geng, Z. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(12):2234–2240, 2007.

[13] A. E. K. Ghalleb, R. B. Slamia, and N. E. B. Amara. Contribution to the fusion of soft facial and body biometrics for remote people identification. In *2nd International Conference on Advanced Technologies for Signal and Image Processing, ATSIP 2016, Monastir, Tunisia, March 21-23, 2016*, pages 252–257. IEEE, 2016.

[14] B. Golomb, D. Lawrence, and T. Sejnowski. Sexnet: A neural network identifies sex from human faces. In *Advances in Neural Information Processing Systems 3*. 1991.

[15] E. Gonzalez-Sosa, J. Fiérrez, R. Vera-Rodríguez, and F. Alonso-Fernandez. Facial soft biometrics for recognition in the wild: Recent

works, annotation, and COTS evaluation. *IEEE Trans. Inf. Forensics Secur.*, 13(8):2001–2014, 2018.

[16] E. Gonzalez-Sosa, R. Vera-Rodríguez, J. Hernandez-Ortega, and J. Fiérrez. Person recognition at a distance: Improving face recognition through body static information. In *24th International Conference on Pattern Recognition, ICPR 2018, Beijing, China, August 20-24, 2018*, pages 3439–3444. IEEE Computer Society, 2018.

[17] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. *CoRR*, abs/1706.04599, 2017.

[18] 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.*, 40(11):2597–2609, 2018.

[19] H. Han, A. K. Jain, F. Wang, S. Shan, and X. Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 40(11):2597–2609, 2018.

[20] H. Han, C. Otto, and A. K. Jain. Age estimation from face images: Human vs. machine performance. In *2013 International Conference on Biometrics (ICB)*, pages 1–8, 2013.

[21] H. Han, C. Otto, X. Liu, and A. K. Jain. Demographic estimation from face images: Human vs. machine performance. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 37(6):1148–1161, 2015.

[22] 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. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[23] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In F. R. Bach and D. M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 448–456. JMLR.org, 2015.

[24] ISO/IEC 19795-1:2006 Information technology Biometric performance testing and reporting. Standard, International Organization for Standardization, 2016.

[25] A. Jourabloo, X. Yin, and X. Liu. Attribute preserved face de-identification. In *2015 International Conference on Biometrics (ICB)*, pages 278–285, 2015.

[26] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[27] V. Kuleshov and P. S. Liang. Calibrated structured prediction. In *Advances in Neural Information Processing Systems 28*, pages 3474–3482. Curran Associates, Inc., 2015.

[28] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*, pages 365–372. IEEE Computer Society, 2009.

[29] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Describable visual attributes for face verification and image search. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(10):1962–1977, 2011.

[30] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 3730–3738. IEEE Computer Society, 2015.

[31] Y. Madadi, V. Seydi, K. Nasrollahi, R. Hosseini, and T. B. Moeslund. Deep visual unsupervised domain adaptation for classification tasks: a survey. *IET Image Process.*, 14(14):3283–3299, 2020.

[32] J. Mansanet, A. Albiol, and R. Paredes. Local deep neural networks for gender recognition. *Pattern Recognition Letters*, 70:80 – 86, 2016.

[33] A. M. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *CoRR*, abs/1412.1897, 2014.

[34] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In X. Xie, M. W. Jones, and G. K. L. Tam, editors, *Proceedings of the British Machine Vision Conference 2015, BMVC 2015, Swansea, UK, September 7-10, 2015*, pages 41.1–41.12. BMVA Press, 2015.

[35] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, Oct 2000.

[36] D. A. Reid, M. S. Nixon, and S. V. Stevenage. Soft biometrics; human identification using comparative descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(6):1216–1228, 2014.

[37] K. Ricanek and T. 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, 2006.

[38] P. Rodríguez, G. Cucurull, J. M. Gonfaus, F. X. Roca, and J. Gonzàlez. Age and gender recognition in the wild with deep attention. *Pattern Recognition*, 72:563–571, 2017.

[39] R. Rothe, R. Timofte, and L. Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2):144–157, 2018.

[40] E. M. Rudd, M. Günther, and T. E. Boult. MOON: A mixed objective optimization network for the recognition of facial attributes. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, volume 9909 of *Lecture Notes in Computer Science*, pages 19–35. Springer, 2016.

[41] P. Samangouei and R. Chellappa. Convolutional neural networks for attribute-based active authentication on mobile devices. In *8th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2016, Niagara Falls, NY, USA, September 6-9, 2016*, pages 1–8. IEEE, 2016.

[42] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 815–823. IEEE Computer Society, 2015.

[43] H. Security. Deep face recognition challenges and tips for real-life deployment. Report, GPU Tech. Conference, 2018.

[44] W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, and A. L. Yuille. Deep regression forests for age estimation. *CoRR*, abs/1712.07195, 2017.

[45] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, Jan. 2014.

[46] P. Terhörst, D. Fährmann, N. Damer, F. Kirchbuchner, and A. Kuijper. Beyond identity: What information is stored in biometric face templates? In *2020 IEEE International Joint Conference on Biometrics, IJCB 2020, Houston, TX, USA, September 28 - October 1, 2020*, pages 1–10. IEEE, 2020.

[47] P. Terhörst, M. Huber, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper. Multi-algorithmic fusion for reliable age and gender estimation from face images. In *22th International Conference on Information Fusion, FUSION 2019, Ottawa, ON, Canada, July 2-5, 2019*, pages 1–8. IEEE, 2019.

[48] P. Terhörst, M. Huber, J. N. Kolf, I. Zelch, N. Damer, F. Kirchbuchner, and A. Kuijper. Reliable age and gender estimation from face images: Stating the confidence of model predictions. In *10th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2019, Tampa, Florida, USA, September 23-26, 2019*. IEEE, 2019.

[49] P. Terhörst, J. N. Kolf, M. Huber, F. Kirchbuchner, N. Damer, A. Morales, J. Fiérrez, and A. Kuijper. A comprehensive study on face recognition biases beyond demographics. *CoRR*, abs/2103.01592, 2021.

[50] P. Terhörst, K. Riehl, N. Damer, P. Rot, B. Bortolato, F. Kirchbuchner, V. Struc, and A. Kuijper. PE-MIU: A training-free privacy-enhancing face recognition approach based on minimum information units. *IEEE Access*, 8:93635–93647, 2020.

[51] P. Tome-Gonzalez, J. Fiérrez, R. Vera-Rodríguez, and M. S. Nixon. Soft biometrics and their application in person recognition at a distance. *IEEE Trans. Inf. Forensics Secur.*, 9(3):464–475, 2014.

[52] J. v. d. Wolfshaar, M. F. Karaaba, and M. A. Wiering. Deep convolutional neural networks and support vector machines for gender recognition. In *2015 IEEE Symposium Series on Computational Intelligence*, pages 188–195, 2015.

[53] M. Wang and W. Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[54] S. Zaghbani, N. Boujneh, and M. S. Bouhlel. Age estimation using deep learning. *Computers & Electrical Engineering*, 68:337 – 347, 2018.

[55] H. Zhang, J. R. Beveridge, B. A. Draper, and P. J. Phillips. On the effectiveness of soft biometrics for increasing face verification rates. *Comput. Vis. Image Underst.*, 137:50–62, 2015.