IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 19, 2021, accepted March 1, 2021, date of publication March 4, 2021, date of current version March 18, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3063819

Tackling Age-Invariant Face Recognition With Non-Linear PLDA and Pairwise SVM

PABLO NEGRI[1,2], SANDRO CUMANI[3], AND ANDREA BOTTINO[3], (Member, IEEE)

[1]Computer Department, University of Buenos Aires (UBA-FCEN), Capital Federal 1428, Argentina
[2]Institute of Research in Computer Sciences (ICC), CONICET-UBA, Capital Federal 1428, Argentina
[3]DAUIN, Politecnico di Torino, 10129 Torino, Italy

Corresponding author: Pablo Negri (pnegri@dc.uba.ar)

This work was supported by SIFACE through the Argentina's National Security Ministry and National Scientific and Technical Research Council (CONICET), Argentina, under Grant 2069/16 and Grant 1431/17.

ABSTRACT Face recognition approaches, especially those based on deep learning models, are becoming increasingly attractive for missing person identification, due to their effectiveness and the relative simplicity of obtaining information available for comparison. However, these methods still suffer from large accuracy drops when they have to tackle cross-age recognition, which is the most common condition to face in this specific task. To address these challenges, in this paper we investigate the contribution of different generative and discriminative models that extend the Probabilistic Linear Discriminant Analysis (PLDA) approach. These models aim at disentangling identity from other facial variations (including those due to age effects). As such, they can improve the age invariance characteristics of state-of-the-art deep facial embeddings. In this work, we experiment with a standard PLDA, a non-linear version of PLDA, the Pairwise Support Vector Machine (PSVM), and introduce a nonlinear version of PSVM (NL–PSVM) as a novelty. We thoroughly analyze the proposed models' performance when addressing cross-age recognition in a large and challenging experimental dataset containing around 2.5 million images of 790,000 individuals. Results on this testbed confirm the challenges in age invariant face recognition, showing significant differences in the effects of aging across embedding models, genders, age ranges, and age gaps. Our experiments show as well the effectiveness of both PLDA and its proposed extensions in reducing the age sensitivity of the facial features, especially when there are significant age differences (more than ten years) between the compared images or when age-related facial changes are more pronounced, such as during the transition from childhood to adolescence or from adolescence to adulthood. Further experiments on three standard cross-age benchmarks (MORPH2, CACD-VS, and FG-NET) confirm the proposed models' effectiveness.

INDEX TERMS Face recognition, age-invariant face recognition, aging datasets, non-linear PLDA, PSVM.

I. INTRODUCTION

The ''Collegno amnesiac'' case is a notorious judicial affair that was discussed in the Italian media for more than 40 years [1]. In 1926, a man was arrested and later taken to the asylum in Collegno, a small town near Turin, because of his alleged mental illness. The man did not remember his name and was later identified as several missing persons before concluding a lengthy investigation. However, his true identity was never indisputably proven.

This case is an example of the daily situation that medical and law enforcement personnel have to face with people unable to tell their own identity. Several diseases, like Amnesia and Alzheimer, can cause temporary or permanent memory losses. Patients may arrive at hospitals with serious injuries, or in a coma, without the possibility to provide any clue about themselves, and some decedents remain unidentified. In addition, every year, hundreds of thousands of individuals go missing, and their fate often remains unknown. The number of missing persons in 2017 is approximately 660,000 in USA [2] and 38,000 in Australia [3] with about 190,000 missing children reported in Europe in the same year [4].

Therefore, practitioners, researchers, and law enforcement personnel strive to develop effective techniques for identifying unknown persons and, possibly, solving missing

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar.

person cases. Historically, the standard recognition methods use DNA or fingerprint analysis, which, however, requires the availability of suitable information for comparison. Due to this issue, the use of face recognition (FR) as a biometric identification system is becoming more and more common. Its advantage is twofold. First, facial images can be easily collected from digital scans of documents, live pictures, or social networks. Second, recent advances in FR algorithms show impressive detection accuracy [5]–[9]. However, FR requires to face several challenges related to varying poses, image resolution, and lighting conditions. In particular, the various models are not fully capable of disentangling age-related from identity components, with a detrimental effect on performances in case of significant age gaps between the compared pictures [10]. Moreover, the performances of these FR systems decrease when the size of the dataset of known individuals (the one you are checking against) increases [11].

Our interest in age invariant facial recognition stems from the development of SIFACE, an FR-based software application that aims to aid and improve law enforcement agencies' response to cases of unknown and missing persons. This application represents a complicated scenario for age invariant facial recognition (AIFR) methods due to the large number of individuals in the gallery and the large age differences between the images stored for each individual.

Given these challenges, this work's contribution is to analyze the effectiveness of different component analysis (CA) approaches in improving the age invariant properties of an FR system. CA consists of statistical techniques that aim to decompose data into latent variables relevant to the task at hand. In particular, we will focus on approaches derived from (and extending) Probabilistic Linear Discriminant Analysis (PLDA), a generative model that has demonstrated its effectiveness in various tasks, including FR [12]–[17]. However, in our opinion, this approach has not yet been fully exploited in the AIFR context.

PLDA describes data in terms of two components. The first is a latent variable that depends only on the label (i.e., the identity) and not on the particular sample (i.e., the submitted image) and, thus, it models the across-class variations. The second component is different for each sample and models the within-class variations. This feature is particularly relevant in our problem, where the first latent variable focuses on the person's identity and the latter on image changes due to illumination, pose, and (most of all) aging.

The main contribution of our work is the introduction in the field of AIFR of novel extensions of the basic PLDA model. These extensions, which were developed in recent years by the speaker recognition community and in particular by the authors of this work, show substantial performance improvements in different recognition tasks [18]–[24]. These models are the non–linear PLDA (NL–PLDA) classifier [23], [25] and the Pairwise Support Vector Machine (PSVM) approach described in [19], [26]. Then, we introduce as a novel approach a non–linear version of PSVM (NL–PSVM).

Experiments on both standard cross-age benchmarks and a challenging testbed, which comprises more than 2.5 million images of about 790,000 individuals, showing significant age gaps between compared images, demonstrates the effectiveness of the proposed models in reducing the age sensitivity of several standard deep facial feature extractors. In particular, our results show that the novel NL–PSVM model introduced in this work outperforms all other approaches.

The rest of the paper is organized as follows. In Section II we review past works. Section III outlines the proposed approach. Then, we describe the dataset and the experimental protocol of the performed tests in Section IV, discussing the results in Section V, and, finally, concluding the paper in Section VI.

II. BACKGROUND

A. FACE RECOGNITION

Face recognition (FR) is a well-studied problem in Computer Vision and Machine Learning. FR is a complex task since it is largely affected by the image variability due to differences in size, resolution, pose, expression, and illumination. The recent advances in deep learning technologies allowed a giant leap forward in FR. Since [5] surpassed human accuracy on the LFW dataset benchmark [27], novel Convolutional Neural Networks (CNN) architectures, lighter and more accurate, or using innovative objective functions, are proposed continuously in the literature.

As soon as the accuracy on the LFW benchmark reached 99.7% and there was no room for significant improvements on simple sets, researchers started focusing on more challenging datasets and specific tasks, like improving the robustness to challenging image variations (e.g., cross-pose or cross-age FR). For a comprehensive review of the recent literature, we refer the interested readers to these two recent surveys [28], [29].

B. CROSS-AGE RECOGNITION

AIFR deals with several challenges related to the face transformations induced by biological aging, which sums up to the other appearance variations already included in the images. AIFR has been tackled using both generative and discriminative methods. The generative approaches rely on aging models, which try to infer the age-related transformation to be applied to a gallery to match (an estimate of) the age of the subject in the probe image. An example is Age-cGAN [30], a deep age-invariant model capable of synthesizing both younger and older faces. Since the generated images do not adequately preserve the original identities, this approach has been subsequently combined with a Local Manifold Adaptation method [31] to improve its verification accuracy.

On the contrary, the discriminative approaches try to extract age-invariant features before solving the classification problem. First attempts leveraged standard textural features, using either individual [16], [32] or multiple [33]

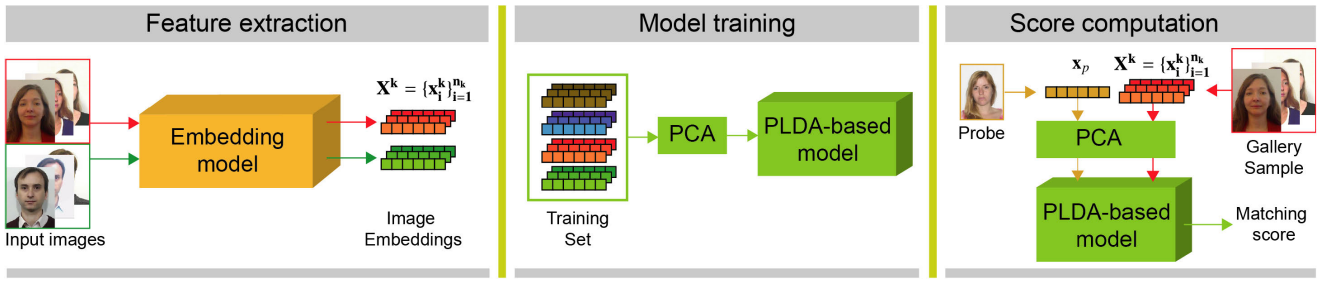

FIGURE 1. Overview of the proposed PLDA-based AIFR method. First, image embeddings are computed for all pictures of each individual (left). Then, after compacting the embedding space with PCA, the selected PLDA-based classifier is trained (center). Finally, the trained classifier is used to compute a matching score between the features of a probe and any gallery sample (right).

local descriptors. Since the characteristics of the extracted features have a strong influence on the recognition process, researchers started developing novel descriptors expressly tailored to the AIFR task [34]–[36]. Instead of using hand-crafted features, several works tried to learn age-invariant features from deep learning approaches. In [37], the authors proposed an injection scheme to integrate external features into the deepest layers of the network, while in [38] age-invariant features are computed from the joint learning of a CNN and a Latent Identity Module (LF-CNN). A different method converts the deep features extracted from a standard backbone into a discriminant codeword that reduces the intra-personal variations caused by aging [39]. The work [40] exploits auto-encoders to learn latent variables that can separate facial embeddings into identity, age-related, and noise features. Then, only the identity features are used for recognition. Wang *et al.* [41] introduce a Decorrelated Adversarial Learning (DAL) and use a Batch Canonical Correlation Analysis (BCCA) for optimization. The methodology trains a linear factorization module that decomposes the identity and age information. A different approach is presented in [42], where authors propose a meta-learning method to improve the generalization properties of general FR approaches that can effectively tackle AIFR as well.

Finally, some works try to combine discriminative and generative approaches. An example is [43], where authors develop a deep, unified architecture capable of jointly performing cross-age face synthesis and recognition in a mutual boosting way. Similar approaches are [44], [45], which propose various methods for synthesizing aging faces by leveraging personalized age progression features, and [46], which leverage an age estimation task to infer age-invariant features.

C. PROBABILISTIC GENERATIVE MODELS FOR FR

PLDA is a probabilistic approach for component analysis that leverages the knowledge about the class labels of the samples to create class–specific and sample–specific latent variables, thus modeling across and within–class variability as separate components.

PLDA and its derivations have been largely investigated in the fields of speaker recognition and verification [18], [47]–[49]. Despite their success in these challenging scenarios, their application to Computer Vision tasks has been relatively limited. In FR, this technique showed its effectiveness when applied to both gray-scale images and feature vectors computed from local textural descriptors [12], [14], [14]. Several variants, aimed at improving the scalability and the computational burden of basic PLDA for FR, were also presented [15], [50]. An interesting extension, taking into account the dynamic identification of individuals in video sequences, was proposed in [51]. In the field of AIFR, the contribution of PLDA techniques has been analyzed in [16], where the Histogram Of Gradient (HOG) features of different facial patches are first concatenated and then compressed with Principal Component Analysis (PCA) before extracting the latent identity variables. A similar approach, discussed in [40], makes use of deep learning methods to learn latent representations from inputs. In order to improve accuracies, [17] proposes to combine different features (namely weighted LBP and HOG) by first learning with PLDA an independent discriminative aging subspace for each feature and then combining their latent representations using a fusion mechanism based on Canonical Correlation Analysis.

Compared to these works, the novelty of our paper is that we analyze the contribution of various approaches that extend the basic PLDA formulation (i.e., NL–PLDA, PSVM, and NL–PSVM) in improving the age-invariant properties of latent variables extracted from deep embeddings. We underline that, to the best of our knowledge, these models are applied for the first time in the field of FR (in general) and of cross-age facial recognition (in particular).

III. PLDA–BASED CLASSIFIERS FOR AIFR

This section describes the main elements of our AIFR framework, the structure of which is outlined in Figure 1. This approach starts by extracting image features. Let $\mathbf{x}$ be a face image embedding and $\mathbf{X}_k = \{\mathbf{x}_i^k\}_{i=1}^{n_k}$ the set of embeddings of the n_k images associated to individual k. Since CNN architectures are the state-of-the-art in FR, they are a natural choice for picking the most suited embedding models for our problem. In general, these architectures are trained by minimizing an appropriate loss function between the real and estimated labels. When these models are used to identify people that are not present in the training set, image embeddings are used to compute a matching score between a probe p and any gallery sample k, usually leveraging a suitable distance in the face space such as Euclidean or cosine distance.

However, despite the effectiveness in addressing FR demonstrated by recent deep architectures, results on cross-age datasets (such as CALFW [10]) show significant drops in the accuracy. A possible explanation of this behavior is that, while these methods can robustly cope with changes in illumination, pose, and expression, they cannot fully disentangle the age-related changes from the relevant facial traits.

Therefore, we propose to tackle this problem with PLDA-based models since they can separate, in the embeddings space, the identities from other possible sources of variability, thus making these models particularly attractive for the AIFR task. Once trained, PLDA-based models can be used as classifiers to compute the matching score between an input probe and any sample in the gallery.

In the following sections, we recall the standard Gaussian PLDA model [14], [48], and we detail the other considered techniques, i.e. a non–linear PLDA (NL–PLDA) classifier [23], [25], the Pairwise Support Vector Machine (PSVM) approach [19], [26], and its non–linear version, presented for the first time in this work.

A. PROBABILISTIC LINEAR DISCRIMINANT ANALYSIS

Vanilla PLDA is a generative model that represents $\mathbf{x}_i^k$, the embedding of the i–th image of individual k, as the sum of an identity factor $\mathbf{y}_k$ and two terms capturing any other relevant image variation not related to identity. They correspond to an ''intersession'' factor $\mathbf{w}_i^k$ and a residual noise $\mathbf{e}_i^k$:

$$\mathbf{x}_i^k = \mathbf{m} + \mathbf{U}\mathbf{y}_k + \mathbf{V}\mathbf{w}_i^k + \mathbf{e}_i^k \tag{1}$$

where $\mathbf{U}$ and $\mathbf{V}$ are rectangular matrices that constrain the identity and intersession factor to be of lower dimensionality than the embedding and $\mathbf{m}$ is the mean of the dataset. In equation (1), $\mathbf{y}_k$ is a latent variable representing an identity, and its value is assumed to be the same for all embeddings of the same person (i.e., it models the across class variations). The term $\mathbf{w}_i^k$ represents intra-individual variability (e.g., the differences in images of the same individual due to illumination, pose and age, thus modeling the within–class variations) and $\mathbf{e}_i^k$ is a stochastic noise. All these terms are assumed independent and normal distributed as:

$$\mathbf{y}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \mathbf{w}_i^k \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \mathbf{e}_i^k \sim \mathcal{N}(\mathbf{0}, \mathbf{\Lambda}_D^{-1})$$

where $\mathbf{\Lambda}_D^{-1}$ is a diagonal precision matrix and D is the dimension of the embedding space.

In case matrix $\mathbf{V}$ has full rank, the model can be simplified [18] as:

$$\begin{aligned} \mathbf{x}_i^k &= \mathbf{m} + \mathbf{U}\mathbf{y}_k + \boldsymbol{\epsilon}_i^k \\ \mathbf{y}_k &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\epsilon}_i^k \sim \mathcal{N}(\mathbf{0}, \mathbf{\Lambda}^{-1}) \end{aligned} \tag{2}$$

where, in contrast with (1), the covariance matrix $\mathbf{\Lambda}^{-1}$ of the residual term in (2) is not diagonal. In the following, we will refer to (2) as the PLDA model. PLDA allows expressing the probability density for a given embedding in terms of conditional and prior densities as:

$$\begin{aligned} p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}_i^k|\mathbf{y}_k) &= \mathcal{N}(\mathbf{x}_i^k|\mathbf{m} + \mathbf{U}\mathbf{y}_k; \mathbf{\Lambda}^{-1}) \\ p_{\mathbf{Y}}(\mathbf{y}_k) &= \mathcal{N}(\mathbf{y}_k|\mathbf{0}, \mathbf{I}) \end{aligned} \tag{3}$$

The model parameters ($\mathbf{m}$, $\mathbf{U}$ and $\mathbf{\Lambda}$) can be estimated on a training set through the Expectation–Maximization algorithm [18]. Then, given a trained PLDA model, a verification score for a probe ($\mathbf{x}_1$) and a gallery embedding ($\mathbf{x}_2$) can be computed as the log–likelihood ratio between the same–identity $\mathcal{H}_s$ and different–identity $\mathcal{H}_d$ hypotheses:

$$\begin{aligned} s(\mathbf{x}_1, \mathbf{x}_2) &= \log \frac{P(\mathbf{x}_1, \mathbf{x}_2|\mathcal{H}_s)}{P(\mathbf{x}_1, \mathbf{x}_2|\mathcal{H}_d)} \\ &= \log \frac{\int_{\mathbf{y}} p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}_1|\mathbf{y}) p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}_2|\mathbf{y}) p_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y}}{\int_{\mathbf{y}} p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}_1|\mathbf{y}) p_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y} \int_{\mathbf{y}} p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}_2|\mathbf{y}) p_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y}} \end{aligned} \tag{4}$$

Closed form solutions are available to compute the integrals in (4). In particular, the log–likelihood ratio corresponds to the following quadratic form:

$$s(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^T \mathbf{A}\mathbf{x}_1 + \mathbf{x}_2^T \mathbf{A}\mathbf{x}_2 + \mathbf{x}_1^T \mathbf{B}\mathbf{x}_2 + \mathbf{x}_1^T \mathbf{c} + \mathbf{x}_2^T \mathbf{c} + k \tag{5}$$

where $\mathbf{A}$ and $\mathbf{B}$ are symmetric matrices, $\mathbf{c}$ is a D–dimensional vector and k is a constant term. The relation between ($\mathbf{m}$, $\mathbf{U}$, $\mathbf{\Lambda}$) and ($\mathbf{A}$, $\mathbf{B}$, $\mathbf{c}$, k) is given in [19].

B. NON-LINEAR PLDA

Although PLDA provides good results for both face and speaker verification, the Gaussian assumption of the model is often a crude approximation. Different approaches have been proposed in the literature to relax the Gaussianity assumptions [18], [24], [52]. In this work, we follow the approach described in [23], [25]. The Non–Linear PLDA (NL–PLDA) assumes that embeddings have been generated by a PLDA model, but have further been transformed by the inverse of a non–linear, invertible function $\mathbf{g}$ as:

$$\begin{aligned} \mathbf{z}_i^k &= \mathbf{m} + \mathbf{U}\mathbf{y}_k + \boldsymbol{\epsilon}_i^k \\ \mathbf{x}_i^k &= \mathbf{g}^{-1}(\mathbf{z}_i^k) \end{aligned} \tag{6}$$

where $\mathbf{m}$, $\mathbf{U}$, $\mathbf{y}_k$ and $\boldsymbol{\epsilon}_i^k$ have the same meaning (and the same prior distributions) as in PLDA, and $\mathbf{z}_i^k$ is the PLDA–generated sample which is transformed to obtain the observed embedding $\mathbf{x}_i^k = \mathbf{g}^{-1}(\mathbf{z}_i^k)$.

As for PLDA, the model allows expressing the probability density for an embedding as

$$\begin{aligned} p_{\mathbf{Y}}(\mathbf{y}_k) &= \mathcal{N}(\mathbf{y}_k|\mathbf{0}, \mathbf{I}) \\ p_{\mathbf{Z}|\mathbf{Y}}(\mathbf{z}_i^k|\mathbf{y}) &= \mathcal{N}(\mathbf{z}_i^k|\mathbf{m} + \mathbf{U}\mathbf{y}_k; \mathbf{\Lambda}^{-1}) \\ p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}_i^k|\mathbf{y}) &= p_{\mathbf{Z}|\mathbf{Y}}(\mathbf{g}(\mathbf{x}_i^k)|\mathbf{y}_k) \left|\mathbf{J}_{\mathbf{g}}(\mathbf{x}_i^k)\right| \\ &= \mathcal{N}(\mathbf{g}(\mathbf{x}_i^k); \mathbf{m} + \mathbf{U}\mathbf{y}_k, \mathbf{\Lambda}^{-1}) \left|\mathbf{J}_{\mathbf{g}}(\mathbf{x}_i^k)\right| \end{aligned} \tag{7}$$

where $\left|\mathbf{J}_{\mathbf{g}}(\mathbf{x})\right|$ denotes the absolute value of the determinant of the Jacobian of $\mathbf{g}$.

In order to learn an appropriate non–linearity, we express $\mathbf{g}$ as the composition of parametric linear and non–linear functions. In particular, we adopt the same architecture of [23], alternating five linear layers with four sinh–arcsinh layers. As for PLDA, the EM algorithm can be used to learn the model parameters.

Then, combining equations (4) and (7), the verification score can be again computed as the log–likelihood ratio between the same–identity and different–identity hypotheses:

$$
\begin{aligned}
&s(\mathbf{x}_1, \mathbf{x}_2) \\
&= \log \frac{\int_{\mathbf{y}} p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}_1|\mathbf{y}) p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}_2|\mathbf{y}) p_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y}}{\int_{\mathbf{y}} p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}_1|\mathbf{y}) p_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y} \int_{\mathbf{y}} p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}_2|\mathbf{y}) p_{\mathbf{Y}}(\mathbf{Y}) d\mathbf{y}} \\
&= \log \frac{\int_{\mathbf{y}} p_{\mathbf{Z}|\mathbf{Y}}(g(\mathbf{x}_1)|\mathbf{y}) p_{\mathbf{Z}|\mathbf{Y}}(g(\mathbf{x}_2)|\mathbf{y}) p_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y}}{\int_{\mathbf{y}} p_{\mathbf{Z}|\mathbf{Y}}(g(\mathbf{x}_1)|\mathbf{y}) p_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y} \int_{\mathbf{y}} p_{\mathbf{Z}|\mathbf{Y}}(g(\mathbf{x}_2)|\mathbf{y}) p_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y}}
\end{aligned}
\tag{8}
$$

which does not depend on the determinant of the Jacobian of $\mathbf{g}$. It is worth noting that once the transformation $\mathbf{g}$ has been estimated, the NL–PLDA model becomes equivalent to a PLDA model in the feature space defined by $\mathbf{g}$. This can be verified by comparing (3) with (7) and (4) with (8).

Further details, together with an analysis of the benefits of the NL–PLDA approach concerning other non–linear variants, can be found in [23] and [53].

C. PAIRWISE SUPPORT VECTOR MACHINE

The third model we analyze is the Pairwise Support Vector Machine [19], [26], which can be interpreted as a discriminatively trained flavor of PLDA. As a matter of facts, while the PLDA log–likelihood ratio can be represented as a quadratic form of the embeddings $(\mathbf{x}_1, \mathbf{x}_2)$ as in Equation (5), it is also *linear* in the model parameters $(\mathbf{A}, \mathbf{B}, \mathbf{c}, k)$. Thus, it can be interpreted as a linear separation rule in an expanded *pair* space $\boldsymbol{\phi}\left([\mathbf{x}_1^T\ \mathbf{x}_2^T]^T\right)$. It is also worth noting that the feature space expansion corresponds to a quadratic polynomial kernel for embedding pairs.

The PSVM approach consists, therefore, in training a Support Vector Machine (SVM) classifier to discriminate between pairs of embeddings belonging to the same individual and those belonging to different individuals. The SVM is trained with the same classification rule defined in equation (5). Despite the huge amount of pairs and the fact that the SVM kernel is quadratic, we have devised effective approaches to train the model in such conditions, exploiting both the correlations between the different pairs and the fact that most of the training pairs are actually irrelevant for the classification rule. The details can be found in [26], [54].

D. NON–LINEAR PSVM

In Section III-B we have shown that, once the transformation $\mathbf{g}$ has been estimated, the NL–PLDA model can be interpreted as a PLDA model in the transformed feature space. Furthermore, we have shown that the PSVM approach has formally the same classification rules as PLDA. Therefore, a straightforward approach to combine the benefits of the non–linearity of NL–PLDA with the discriminative training of PSVM consists in estimating a PSVM model on the transformed embeddings. In practice, we first train a NL–PLDA model to estimate the transformation $\mathbf{g}$, and then we proceed by estimating a PSVM model from the transformed embeddings $\mathbf{g}(\mathbf{x}_i^k)$. The resulting scoring function is thus given by

$$
s(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{z}_1^T \mathbf{A} \mathbf{z}_1 + \mathbf{z}_2^T \mathbf{A} \mathbf{z}_2 + \mathbf{z}_1^T \mathbf{B} \mathbf{z}_2 + \mathbf{z}_1^T \mathbf{c} + \mathbf{z}_2^T \mathbf{c} + k \tag{9}
$$

where $\mathbf{z}_1 = \mathbf{g}(\mathbf{x}_1)$, $\mathbf{z}_2 = \mathbf{g}(\mathbf{x}_2)$. The parameters $(\mathbf{A}, \mathbf{B}, \mathbf{c}, k)$ are estimated as in the PSVM approach. In the following, we refer to this approach as NL–PSVM.

FIGURE 2. SIFACE User Interface with an example of a query on the left side and the first page of matching identities. Images were masked for privacy purposes.

IV. EXPERIMENTS

The following sections describe the main context of our research (section IV-A) and the dataset used in our experiments (section IV-B). Then, we provide some implementation details (Section IV-C). Finally, we introduce the experimental protocol of the tests aimed at assessing the accuracy, in the specific context of this work, of the various classifiers under analysis (Section IV-D).

A. THE SIFACE PLATFORM

The AIFR approach presented in this work has been developed as the core of SIFACE, a software platform implemented to support the Argentinian police response to missing person's cases. In Argentina, these procedures are managed by a division of the Government's Ministry of Security. When a new case is reported to the police or justice departments, it is first incorporated into a national database named SIFCOP. Then, with SIFACE, the operator can submit a picture of the unknown person to query a dataset that contains several images (collected at different ages) for each known identity. The operator can eventually specify additional pieces of information, such as gender and estimated age range of the depicted subject, to reduce the dimension of the search space. As a result, individuals in the gallery of known identities are presented in descending order of similarity with the query image, by showing to the operator four hypotheses per page (Fig. 2). In order to avoid clutter in the User Interface (UI), the application displays only one picture per candidate. The user can then navigate forward and backward between different pages and eventually access the full list of images and the judicial dossier of a candidate.

According to this description, it is clear that the accuracy of the FR module of SIFACE is vital in helping the operator

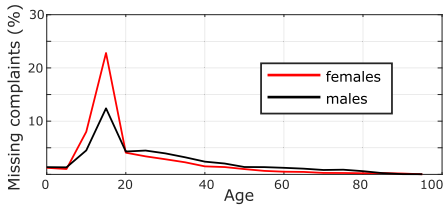

FIGURE 3. Current percentage of missing person complaints recorded in SIFCOP by age and gender.

FIGURE 4. Examples of the images available for four individuals in the ID-DATASET.

in his/her work. In particular, this module should provide a recognition that is robust against facial changes due to aging. This feature is of paramount relevance since 46% of missing person complaints recorded in SIFCOP correspond to persons in the age range between 10 and 20 years (Fig. 3), an interval where such facial changes are more pronounced.

B. ID-DATASET

In both SIFACE and the main tests of this paper, we employ a gallery of images, referred to in the following as the ID-DATASET, which contains the pictures provided to the Argentine Federal Police (PFA) by people requesting an identity document (ID or passport) between the year 2000 and 2017. The birth year of the persons depicted in the images ranges between 1948 and 2000. Each individual in the dataset is associated with a set of pictures taken every time she or he renewed an identity document. The number of pictures per individual is unevenly distributed in the range [2, 12], with 93.4% of persons having three or four images and only the 0.3% having more than five pictures.

Overall, the dataset contains $2,527,079$ images of $793,280$ individuals, almost equally distributed between males and females (respectively, 51.2% and 48.8%). All images have a resolution of 387×300 pixels and show varying face orientations since, due to a change in the official standards, three-quarter profile views were requested until the year 2005 and frontal views after this date (Figure 4). We conclude by highlighting that, due to its collection method and purpose, unfortunately, the ID-DATASET cannot be made public for privacy and security limitations.

C. IMPLEMENTATION DETAILS

1) FACE EMBEDDING MODELS

As shown in Figure 1, the extraction of facial embeddings is the first step of our approach. Since we have to cope with a challenging test suite that includes the identification of persons in an extensive dataset and across ages (spanning different age intervals and considering age gaps higher than ten years), the choice of the most effective facial embedding is vital in our framework. Given the plethora of different deep facial feature extractors available in the literature, we selected some attractive models that (i) leverage various state-of-the-art architectural solutions, and (ii) show good results on LFW [27] and CALFW [10] benchmarks, which are currently considered as the de-facto standard testbeds in FR. We considered the following embedding models:

- **ResNet-29 (DLIB)**. This model, available with the DLIB library [55], leverages residual blocks and is essentially a compact version of the ResNet-34 network [56] that reduces the number of layers (to 29) and the number of filters per layer (by a factor two). The DLIB embedding space has size $D = 128$, and the model achieves an accuracy of 99.10% on LFW and 89.52% on CALFW.
- **SENet-50** is based on the *Squeeze and Excitation* block proposed in [8]. The dimensionality of its feature space is $D = 2024$ and its accuracies on LFW and CALFW are 99.61% and 89.84%, respectively.
- **FaceNet**, an architecture based on the Inception module. We used the pre-trained model available at [57] that simplifies the original architecture in [7] to make training easier and faster. The resulting embedding has a dimension $D = 512$, and the model accuracy is 99.45% on LFW and 86.27% on CALFW dataset.
- **ArcFace** [9] introduces as loss a geodesic distance in the face manifold of the feature space. We used ArcFace code and pre-trained models available in [58], which uses a ResNet-50 as backbone. The embedding size is $D = 512$. Accuracy is 99.63% on LFW and 95.33% on CALFW dataset.

2) FEATURE EXTRACTION

Image embeddings are computed as follows. First, we delimit the same area for all frontal and three-quarter faces using the framework for joint face detection and alignment described in [59]. The output of this algorithm is a rectangular region of interest (ROI). Since this ROI focuses on the main facial traits (eyes, nose, and mouth), it is further processed by first enlarging the ROI around its center by a factor $1 + padding$ and then applying a vertical translation that places the eye line at the normalized ROI height vp (see Figure 5). Parameters $padding$ and vp are fine-tuned for each CNN model on a validation set.

To cope with varying image orientations and noise in the face alignment we average the features extracted from different image patches obtained by applying the ROI to randomly translated (in the interval $[-5, +5]$ pixels along each direction) and rotated ($[-5, +5]$ degrees) versions of the original image and then randomly applying a vertical mirroring. We found that, for all the models, a number of 20 patches was saturating the accuracy on a validation set (see Section V-F) and, thus, we used this number in all our tests.

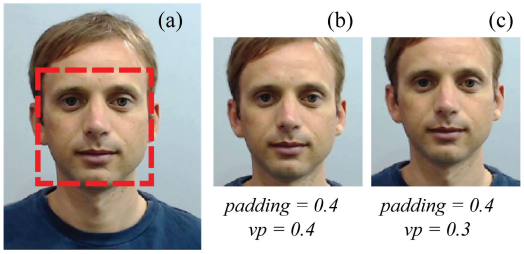

FIGURE 5. Example of automatic face detection (a) and of the final ROI obtained with the same *padding* and different *vp* values (b and c).

3) TRAINING OF PLDA MODELS

The classifiers presented in Section III require a sufficiently large training set including few thousand individuals and few tens of thousands of samples. This issue is particularly relevant for PSVM, which is a discriminative method more prone to overfitting than PLDA. A comparative analysis of the impact of the size of the training set on PLDA and PSVM accuracy can be found in [26]. Non-linear approaches need even more training data than standard PLDA since they require estimating a larger set of parameters.

We considered both gender-dependent (GD) and gender-independent (GI) models. Since we consistently obtained slightly better results with GD models for all classifiers and all feature extractors,[1] we report only the accuracy of GD models. Taking into account the previous consideration, each GD model has been trained on a held-out set of 25,000 individuals spanning uniformly different ages. The same trained models were then used for all experiments described in Section IV-D.

Finally, since in the embedding space human faces belong to a lower dimensionality manifold (usually referred to as the face space), we applied PCA to compact the embedding representations in order to (i) reduce the number of classifier parameters and (ii) prevent numerical instability due to dimensions with very low variance. In all cases, we selected the PCA dimensionality that preserves the 99.9% of the embedding variance. As an indication, the final PCA size is 75 for DLIB, 400 for SENet, 49 for FaceNet, and 240 for ArcFace. These numbers confirm the findings of the intrinsic dimensionality of the face space reported in [61].

D. EXPERIMENTAL PROTOCOL

We designed two experiments on the ID-DATASET to analyze how the accuracy of various combinations of embedding models and classifiers is affected by two main parameters, namely (i) the gallery size (**Test1**, Section IV-D2), and (ii) the age-related changes in facial appearance (**Test2**, Section IV-D3). In order to describe as clearly as possible the experimental protocol, we start by detailing the problem formulation.

[1]For PSVM models this contrasts with our previous findings [60], but can be explained by the larger amount of training individuals, which are enough to reliably estimate gender-dependent models

1) PROBLEM FORMULATION

We recall that the ID-DATASET (which will be referred in the following as $\mathcal{I}$ for simplicity) contains N identities, each associated with a picture set.

Let $I_k = \{H_k, b_k, g_k\}$ be an individual in $\mathcal{I}$, where b_k and g_k are, respectively, her/his birth year and gender. H_k is the set of n_k labeled images $\{(h_i^k, a_i^k)\}_{i=1}^{n_k}$ available for subject k, where $n_k \in [2, 12]$. Each image h_i^k is associated with the age a_i^k of the subject when the image was captured. The elements in H_k are sorted in ascending order of age (i.e., $h_{n_k}^k$ is the most recent picture). Let also $\mathbf{x} = f_M(h)$ be the embedding of image h computed by model M.

Then, let us define as s_p^k the similarity score between the probe image h_p (i.e., the sample submitted to the system for recognition) and an individual k in the gallery $\mathcal{G}$ (i.e., the set of known identities). Since each individual in the gallery is associated with a set of different images, there can be different ways to define the score. In our experiments, we computed this value as the similarity between the feature vector $\mathbf{x}_p$ and that of the most recent image of individual k:

$$s_p^k = s(\mathbf{x}_{n_k}^k, \mathbf{x}_p) \tag{10}$$

where s is the matching scores defined in Section III for each PLDA–based approach. The rationale behind this definition of the similarity score stems from the observation that the cross-age accuracy recognition depends primarily on the age difference between the probe and the gallery image [62]. It is also clear that when age information is missing, other similarity functions should be used (e.g., the minimal or the average score between h_p and all images in H_k). A detailed analysis of the effect of these different scoring functions is presented in Section V-F.

Once scores are computed, individuals in the gallery can be ranked in descending order of similarity with the test image. Since the SIFACE interface presents four candidates per page, we analyzed the results in terms of rank 1 and 4, which, respectively, represent the probabilities that the real identity of the probe is the first candidate (R1) or it is shown in the first page (R4) of the application UI.

2) TEST1: ACCURACY VS GALLERY SIZE

This test aims at assessing the robustness of the various combinations of embedding models and PLDA-based classifiers to the size of the gallery of known identities.

To this end, for each identity $k \in \mathcal{I}$, in **Test1** we first take the most recent picture ($h_{n_k}^k$) as probe image and add the remaining images into the gallery. Then, given the gender g_k and birthdate b_k of the probe, we sort all individuals $j \in \mathcal{I}$ of the same gender according to their absolute age difference $|b_j - b_k|$. Finally, we include in the gallery $\mathcal{G}_k$ the first g elements in this ordered list and compute the model accuracies at increasing values of g from 1,000 to the full gallery of individuals left in $\mathcal{I}$ after removing the training sets used by the classifiers, (see Section IV-C), that is about 350,000 subjects for each gender.

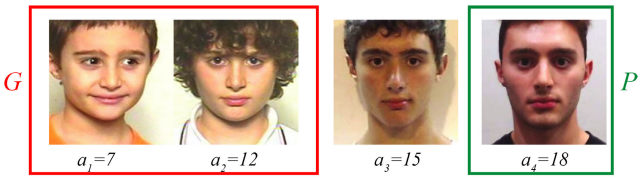

FIGURE 6. Example of subject corresponding to case ($a_g = 12$, $a_p = 18$). There is one probe available in *P* (with age greater or equal to a_p and highlighted in green); the corresponding gallery samples *G* (i.e., images at age lower or equal to a_g) are highlighted in red.

3) TEST2: ACCURACY VS AGING FACTORS

Test2 focuses the analysis on how aging affects the accuracy. Concerning this problem, the SIFACE context is particularly challenging, since, in real missing complaints, there can be a significant age gap between a probe image and the images possibly available for the same individual. This gap can result in large changes in facial appearance, which might severely hamper the identification process [32].

To this end, we created different sets of individuals guaranteeing a particular age gap between their probe and gallery images. These sets (detailed in Section V-B and summarized in Table 2) have been designed to analyze, on the different genders, the aging effect on different age intervals and the transitions they represent (i.e., from childhood to adolescence, adulthood, and old age).

We defined these test sets as follows. For each test set, we define a tuple (a_g, a_p). Then, we include in the set only the individuals in $\mathcal{I}$ having at least one image with age greater than or equal to a_p and at least one image with age lower than or equal to a_g. For an individual k belonging to this set, let P_k be the set of images with an age $\geq a_p$ and G_k those with an age $\leq a_g$. For the tests, we use each image in P_k as an individual probe sample, and we include the whole set G_k in the test gallery.

Consider for example the case where $a_g = 12$ and $a_p = 18$, and take an individual characterized by four images shot at age 7, 12, 15 and 18 (Fig. 6). This individual belongs to this test set, since it is associated with both images at or below age 12 and images at or above age 18. The only probe available in *P* is the image at age 18, the two pictures at age 7 and 12 (set *G*) eventually go into the gallery and the image at age 15 is discarded.

As for the selection of the other identities that should complete the gallery, we recall that, in the specific context of this work, this task can leverage the knowledge about gender and estimated age of the test sample p. In particular, we can use these pieces of information to narrow down the research of the missing-person to a sample-specific gallery $\mathcal{G}_p$ of size $g_p \ll N$, thus possibly helping improve the accuracy.

A possible choice is to include in $\mathcal{G}_p$ all the individuals of the same gender of p and an absolute (estimated) age difference with p lower than or equal to d_{age}. However, this choice might result in excluding the correct identity from the generated gallery since, in a real case, the estimated age of p can be affected by a large error. Thus, it is necessary to define a suitable heuristic minimizing the trade-off between the gallery size and the probability of excluding p in the generation of $\mathcal{G}_p$. As we will show in Section V, the definition of this heuristic can be based on the results of **Test1**, i.e. on the analysis of how the gallery size affects the robustness of our AIFR framework.

V. RESULTS AND DISCUSSION

In the following, we first discuss the results obtained for **Test1** and **Test2**. Then, we assess the proposed models on standard cross-age benchmarks. Finally, we investigate the effect of the parameters choice on accuracy.

A. TEST1

Results are summarized in Fig. 7, where the various diagrams show (for each combination of a classifier, embedding model, and gender) the R1 and R4 accuracies as a function of the gallery size, and in Table 1, which details the results obtained using the full gallery (i.e., about 350,000 samples for each gender). In both the figure and the table, embedding models are sorted in terms of (ascending) overall average performances (i.e., FaceNet, DLIB, SENet, and ArcFace). To assess the mutual contribution of embedding models and PLDA–based models, we selected as baselines the vanilla embedding models described in Section IV-C1. The similarity scores used for these models are their optimal ''natural'' distance in the face space (i.e., Euclidean distance for FaceNet, DLIB and ArcFace, cosine distance for SENet).

The following comments stem from these results.

1) CLASSIFIERS

Results in Table 1 and Fig. 7 show that all the proposed classifiers provide substantial benefits across embedding models and gallery sizes, largely improving the R1 baselines. For the full gallery (Table 1), the relative accuracy improvement ranges, for females, between 30.03% (SENet + NL–PSVM) and 58.02% (ArcFace + NL–PSVM), while for males, it is between 32.29% (SENet + PSVM) and 44.19% (ArcFace + NL–PSVM). Even higher improvements are obtained on R4 (females: from 39.65%, FaceNet + NL–PSVM, to 63.46%, ArcFace + NL–PSVM; males: from 43.62%, SENet + NL–PSVM, to 54.90%, DLIB + NL–PSVM).

It can be also seen that NL–PSVM is the optimal classifier in most cases. A possible explanation of this behavior is that NL–PSVM combines the benefits of discriminative training of PLDA (the PSVM approach) with a non–linear transformation of the embeddings in a latent space where their distribution better matches the PLDA assumptions (thus, helping to better discriminate the different identities). As another comment, the higher performance of the non–linear classifiers (NL–PLDA and NL–PSVM) with respect to their linear counterparts (PLDA and PSVM) hints to a non–linear shape of the face manifold.

TABLE 1. Test1 R1 / *R4* percentage accuracy, on the full gallery, for different combinations of embedding models and PLDA–based classifiers. Best results for each embedding model are displayed in bold and runner-up are underlined. Overall best R1/*R4* results are denoted with '*'.

	Baseline	PLDA	NL–PLDA	PSVM	NL–PSVM
			Males		
FaceNet	80.72 / *88.45*	86.08 / *92.63*	87.77 / *93.82*	87.11 / *93.46*	**87.83** / ***93.95***
DLIB	88.76 / *94.20*	92.41 / *96.39*	**93.40** / ***97.01***	92.68 / *96.69*	93.39 / ***97.01***
SENet	92.42 / *96.26*	93.43 / *96.87*	94.17 / *97.50*	94.82 / *97.82*	**94.87** / ***97.89***
ArcFace	99.86 / *99.91*	99.87 / *99.93*	99.90 / *99.94*	99.91 / ***99.96****	**99.92*** / ***99.96****
			Females		
FaceNet	69.28 / *80.11*	75.86 / *85.72*	78.45 / *87.75*	77.40 / *87.28*	**78.52** / ***87.99***
DLIB	80.19 / *88.07*	86.95 / *93.17*	89.11 / *94.57*	87.97 / *93.96*	**89.36** / ***94.66***
SENet	84.32 / *91.13*	86.42 / *92.71*	87.47 / *93.68*	**89.03** / *94.68*	88.97 / ***94.75***
ArcFace	99.35 / *99.75*	99.36 / *99.78*	99.56 / *99.83*	99.65 / *99.89*	**99.73*** / ***99.91****

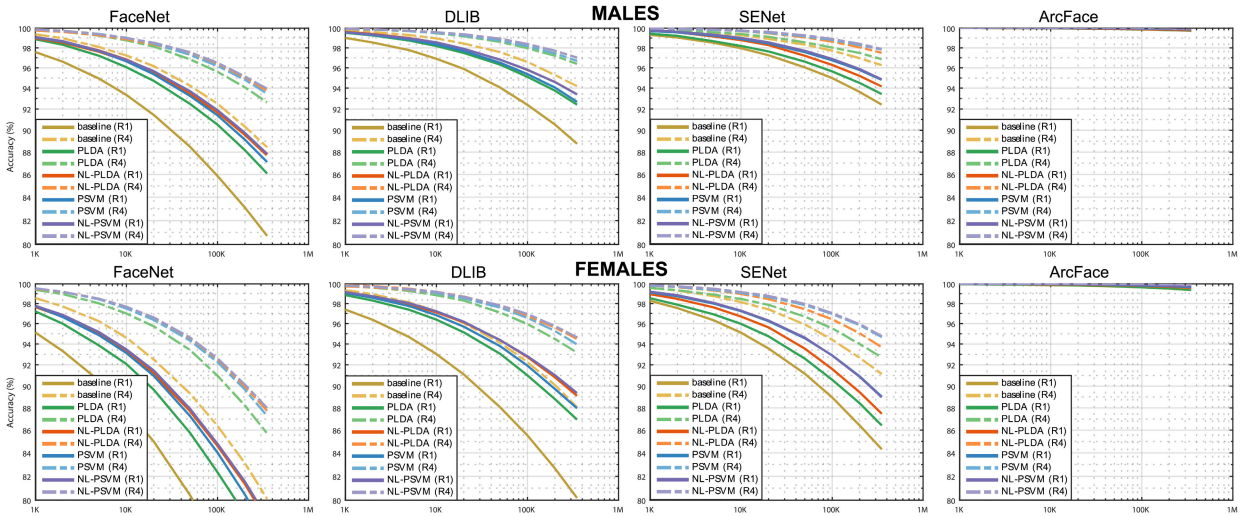

FIGURE 7. Recognition performances on Test1 for varying gallery size.

2) GALLERY SIZE

Fig 7 shows that the increase of the gallery size negatively affects the accuracies but to different extents. As a matter of fact, when the number of distractors increases from 1,000 to about 350,000, the R1 accuracy of the best classifier drops significantly for all embedding models (FaceNet: −11.27% for males and −19, 25% for females; DLIB: −6.20% and −9.81%, SENet: −4.88% and −10.09%) except ArcFace (−0.16% for males and −0.24% for females). A similar behaviour can be observed for R4 (FaceNet: −5.89% and −11.55%; DLIB: −2.93% and −5.18%; SENet: −2.07% and −5.18%; ArcFace: −0.06% and −0.07%). As for the classifiers, NL–PSVM is again the optimal one since it is the most robust to the gallery size variations across genders and embedding models.

3) EMBEDDING MODELS

In general, we can observe significant differences in the accuracies of the various models. ArcFace is consistently the best performer across all classification methods and gallery sizes. This result is consistent with the tests on both LFW and CALFW datasets. As suggested in [58], a possible explanation is that the angular margin defined by ArcFace is indeed robust in discriminating between a large number of identities.

The results in Table 1 show as well the effectiveness of the combination ArcFace + NL–PSVM, since (i) it achieves, on the full galleries, impressive R1 and R4 accuracies (respectively, 99.92% and 99.96% for males and 99.73% and 99.91% for females) and (ii) it significantly improves the full gallery recognition error of other combinations of embedding model/classifier (e.g., R1 accuracy increases of up to 10.37% for females and 4.95% for males).

From Table 1 and Fig. 7 we can also observe the differences between the recognition accuracy for the two genders. Since the distribution of the age difference is similar in the two subsets, these results seem to suggest that men better preserve their appearance across ages. However, this can also be due to additional external factors, such as cross-age variations of makeup or hairstyle.

Concluding, the major takeaways from these results are the following:

- all the PLDA–based classifier provide substantial improvements of the classification accuracy compared to natural metrics in the embedding space;
- among the various approaches under analysis, NL–PSVM is the most effective in our experimental settings;
- ArcFace + NL–PSVM is the most effective combination of deep embedding model and classifier among the ones

TABLE 2. Probe sets for Test2. For each set, we show the characterizing (a_g, a_p) values (Age gap), and the number of testing probes (Probes), unique individuals (Ids), and average gallery size with $d_{age} = 12$ (Gallery) for each gender.

Set	Age gap (a_g, a_p)	Females			Males		
		Probes	Ids	Gallery	Probes	Ids	Gallery
A	(12, 18)	12,908	10,550	126,572	12,896	10,483	131,269
B	(16, 25)	9,691	8,738	177,225	8,747	7,833	185,987
C	(20, 30)	7,684	7,124	193,784	9,291	8,523	204,067
D	(30, 40)	7,672	7,165	194,464	9,487	8,775	215,322
E	(40, 50)	4,609	4,336	157,603	5,822	5,426	169,306
F	(50, 60)	5,790	5,473	108,863	5,400	5,086	105,743
	Total	48,354	43,386		51,643	46,126	

under analysis, and it is capable of maintaining high accuracies even when the number of distractors is high (about 350,000).

B. TEST2

Table 2 details the six probe sets aimed at analyzing the aging effects across different age intervals in **Test2**. These sets are characterized by a different (a_g, a_p) tuple and (for each of them and each gender) we list the number of probes and unique individuals belonging to the set. Since age-related changes are unevenly distributed across ages, we set the age gap to ten years for all groups exception made for the first two (A and B), which represent, respectively, the transition from childhood to adolescence and from adolescence to adulthood, both periods where facial changes are more pronounced. We also recall that the majority of missing person complaints submitted to SIFACE correspond to persons in these age ranges (see Fig. 3).

As for the gallery used in these tests, results from **Test1** confirm the need for finding a suitable trade-off between gallery size and accuracy loss. In Section IV-D2, we suggested using as size selection criterion the age gap d_{age} between the probe and the gallery individuals. However, since in a real scenario the probe age can only be roughly estimated, this threshold should minimize the risk of excluding the real identity from the gallery due to a significant error in such an estimate.

One possible way to deal with this issue starts from the analysis of the average gallery size spanned in $\mathcal{I}$ by different values of d_{age}. Combining this information (Fig. 8) with the results of **Test1**, we arbitrarily selected $d_{age} = 12$ for all experiments in **Test2**. This threshold guarantees both a suitable age margin and a contained gallery size (about 170,000 individuals for females, and about 180,000 for males, i.e. about 50% of the same-gender individuals in $\mathcal{I}$), which, as shown in Fig. 7, does not affect the accuracy in a critical manner, at least for the optimal feature/classifier combination. The average gallery size for each **Test2** set and gender when $d_{age} = 12$ is listed in Table 2.

Test2 results are detailed in Table 3. As expected, these results show a moderate to large accuracy drop with respect to **Test1**. These results are mainly caused by the larger age gap between probes and gallery samples (in **Test1** their average

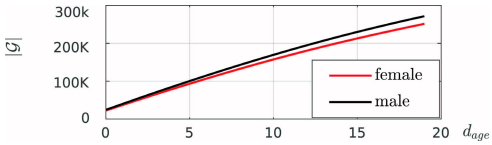

FIGURE 8. Average gallery size as a function of d_{age} parameter.

difference is 3.9 years for both males and females, here it can be higher than ten years). This effect is particularly relevant for set A (childhood to adolescence), where the best R1 results drop to 91.53% for females and a mere 80.08% for males (both results obtained by the combination of ArcFace and NL–PSVM). For other sets and combinations, accuracy reduction is less dramatic but still sensible.

At the same time, **Test2** results stress the effectiveness of all PLDA–based approaches in softening the aging effects on the facial embeddings. This contribution is particularly evident when these effects are more pronounced (e.g., groups A to C). If we analyze, for instance, set A, we can notice a dramatic accuracy loss for all baselines. The worst-case scenario is FaceNet (whose R1 accuracy falls to 3.69% for males and 17.07% for females), but also ArcFace is affected in a significant way (59.71% for males and 79.15% for females).

These are the cases where the PLDA–based classifiers show their strength. For instance, baseline improvements for group A range from 24.71% (DLIB) to 12.38% (ArcFace) for females and from 25.18% (DLIB) to 13.01% (ArcFace). For group B, improvements are between 22.64% and 3.51% for females and between 21.50% and 8.97% for males. Again, these improvements were obtained with DLIB and ArcFace. As another comment, **Test2** results confirm the most significant findings of **Test1**. The possibility of disentangling identity and age-related effects is beneficial for all test sets, particularly for those most affected by changes due to facial aging, confirming that the non–linear versions of the classifiers outperform the linear ones and NL–PSVM provides the most significant benefits. We found again differences in the classification of males and females, although in **Test2** we have somewhat mixed results (younger females are easier to recognize than younger males, the opposite happens for the elders).

Concluding, while these results confirm the difficulties in tackling cross-age recognition, we think that they also highlight the effectiveness of all PLDA–based approaches in softening these issues.

C. ASSESSMENT

In the following, we compare our approach with different state-of-the-art methods in two different settings, i.e., first comparing their accuracy on $\mathcal{I}$ and then on standard AIFR benchmarks. For the sake of brevity, in the discussion we will only consider the results of our optimal solution, i.e. the combination of ArcFace + NL–PSVM.

D. COMPARISONS ON ID-DATASET

First of all, we underline that a fair comparison with most of the state-of-the-art AIFR methods on $\mathcal{I}$ is difficult since their

TABLE 3. R1/*R4* results for Test2. For each age group (A to F), numbers in bold (underlined) represent the best result (the runner–up) for each embedding model and '*' denote the overall optimal value.

Model	Classifier	A	B	C	D	E	F
				Males			
FaceNet	Baseline	3.69 / *7.10*	14.98 / *24.10*	30.60 / *45.39*	53.81 / *69.23*	57.80 / *72.05*	54.48 / *69.07*
	PLDA	10.49 / *19.89*	27.56 / *42.20*	47.89 / *65.33*	66.70 / *80.12*	65.34 / *78.17*	58.78 / *73.61*
	NL–PLDA	16.67 / ***29.68***	33.61 / *50.02*	54.73 / *71.76*	73.55 / *85.93*	74.73 / *85.71*	70.81 / *83.35*
	PSVM	14.05 / *26.01*	31.87 / *47.47*	53.69 / *70.60*	70.70 / *83.56*	70.87 / *83.24*	65.57 / *79.80*
	NL–PSVM	**16.70** / *29.64*	**34.35** / ***50.73***	**55.60** / ***73.24***	**74.36** / ***86.58***	**75.54** / ***86.50***	**71.26** / ***83.89***
DLIB	Baseline	12.01 / *20.28*	35.90 / *50.09*	57.42 / *72.51*	71.89 / *84.33*	70.47 / *83.06*	63.59 / *76.74*
	PLDA	26.43 / *41.55*	49.31 / *64.89*	70.47 / *82.89*	83.86 / *92.47*	81.64 / *90.57*	72.28 / *83.91*
	NL–PLDA	**37.89** / ***54.79***	57.86 / ***72.96***	75.76 / ***87.28***	87.08 / *94.34*	86.40 / ***93.83***	82.07 / *90.72*
	PSVM	31.61 / *48.56*	53.38 / *69.06*	72.02 / *84.51*	84.55 / *92.99*	83.73 / *91.91*	77.63 / *87.74*
	NL–PSVM	35.84 / *52.82*	**57.87** / *72.87*	**76.01** / *87.22*	**87.27** / ***94.49***	**86.41** / *93.80*	**82.80** / ***91.20***
SENet	Baseline	7.66 / *14.28*	32.39 / *47.22*	55.26 / *70.83*	78.11 / *88.65*	82.60 / *91.21*	79.37 / *89.48*
	PLDA	11.33 / *19.05*	38.10 / *52.53*	63.23 / *77.62*	81.41 / *90.75*	83.01 / *91.79*	78.74 / *89.46*
	NL–PLDA	21.41 / *36.09*	47.38 / *64.80*	68.27 / *82.67*	84.15 / *92.86*	85.57 / *93.63*	81.37 / *91.26*
	PSVM	**29.89** / ***43.22***	**51.46** / *67.41*	**72.56** / *85.21*	85.54 / *93.59*	86.93 / *94.56*	84.26 / *92.74*
	NL–PSVM	23.98 / *40.28*	49.54 / ***67.57***	72.33 / ***85.95***	**86.59** / ***94.37***	**87.17** / ***94.83***	**84.37** / ***93.13***
ArcFace	Baseline	59.71 / *73.99*	86.04 / *92.74*	96.64 / *98.63*	99.47 / *99.85*	99.69 / *99.93*	99.80 / ***99.94****
	PLDA	62.44 / *76.73*	88.86 / *94.67*	97.56 / *98.99*	99.53 / *99.81*	99.69 / *99.86*	99.67 / ***99.94****
	NL–PLDA	73.77 / *86.10*	92.63 / *96.80*	98.35 / *99.37*	99.77 / *99.89*	99.76 / *99.95*	99.67 / *99.85*
	PSVM	75.67 / *87.76*	94.06 / *97.31*	98.70 / *99.50*	99.72 / ***99.91****	**99.86*** / *99.93*	**99.81*** / ***99.94****
	NL–PSVM	**80.08*** / ***90.64****	**95.02*** / ***98.23****	**99.06*** / ***99.66****	**99.87*** / *99.89*	99.81 / ***99.97****	**99.81*** / *99.89*
DAL [41]		43.08 / *60.29*	80.23 / *88.72*	95.25 / *98.12*	99.19 / *99.68*	99.38 / *99.73*	99.02 / *99.69*
				Females			
FaceNet	Baseline	17.07 / *27.12*	30.78 / *45.09*	38.98 / *53.57*	44.98 / *59.91*	42.74 / *56.52*	39.67 / *54.78*
	PLDA	25.24 / *38.97*	40.62 / *56.50*	48.87 / *65.24*	58.88 / *73.98*	53.89 / *68.74*	44.82 / *61.30*
	NL–PLDA	**29.47** / ***44.39***	43.84 / *60.74*	54.10 / *69.42*	63.99 / *78.32*	62.57 / *76.39*	58.01 / *73.44*
	PSVM	27.34 / *41.82*	43.82 / *59.80*	51.11 / *67.71*	62.64 / *77.23*	60.30 / *74.66*	54.61 / *70.29*
	NL–PSVM	29.16 / *44.01*	**44.54** / ***61.60***	**54.32** / ***69.92***	**64.85** / ***78.95***	**63.09** / ***77.28***	**59.02** / ***74.37***
DLIB	Baseline	26.39 / *38.17*	44.78 / *59.68*	56.27 / *69.60*	63.11 / *75.93*	55.83 / *70.04*	47.89 / *61.93*
	PLDA	43.85 / *59.63*	61.86 / *75.78*	70.12 / *82.47*	77.41 / *87.41*	74.09 / *84.25*	62.76 / *76.15*
	NL–PLDA	**51.09** / ***66.65***	67.12 / *80.28*	74.45 / *85.78*	81.03 / *90.28*	79.52 / *89.06*	73.87 / *84.96*
	PSVM	46.11 / *61.81*	63.76 / *77.67*	71.58 / *83.93*	78.92 / *88.65*	77.54 / *87.18*	69.24 / *81.61*
	NL–PSVM	49.94 / *65.79*	**67.42** / ***80.63***	**74.79** / ***86.01***	**81.39** / ***90.47***	**80.08** / ***89.93***	**74.25** / ***85.58***
SENet	Baseline	21.43 / *32.51*	45.57 / *60.97*	57.14 / *71.04*	67.65 / *80.85*	67.76 / *80.10*	61.28 / *74.96*
	PLDA	26.41 / *37.50*	50.64 / *65.02*	61.32 / *75.22*	72.26 / *83.90*	71.79 / *83.23*	66.58 / *79.38*
	NL–PLDA	37.29 / *53.40*	56.96 / *72.79*	65.80 / *79.57*	75.40 / *86.82*	73.99 / *84.94*	68.22 / *80.90*
	PSVM	**44.85** / ***58.89***	**61.06** / *75.49*	68.34 / *81.69*	**78.36** / *88.66*	**78.24** / *88.33*	**73.40** / ***85.34***
	NL–PSVM	39.87 / *56.27*	59.92 / ***75.84***	**68.79** / ***82.29***	78.30 / ***89.04***	77.67 / ***88.35***	73.02 / *84.97*
ArcFace	Baseline	79.15 / *87.74*	93.82 / *97.24*	96.76 / *98.76*	98.38 / *99.40*	98.07 / *99.15*	97.34 / *98.98*
	PLDA	80.28 / *89.16*	93.70 / *97.37*	96.67 / *98.74*	98.23 / *99.39*	98.07 / *99.39*	97.75 / *98.98*
	NL–PLDA	87.38 / *93.91*	95.87 / *98.34*	97.89 / *99.35*	99.06 / *99.67*	98.89 / *99.54*	97.86 / *99.29*
	PSVM	88.49 / *94.42*	96.63 / *98.70*	98.20 / *99.49*	99.34 / *99.70*	98.92 / *99.72*	**98.76*** / *99.57*
	NL–PSVM	**91.53*** / ***96.40****	**97.33*** / ***99.14****	**98.66*** / ***99.64****	**99.37*** / ***99.71****	**99.26*** / ***99.83****	**98.76*** / ***99.67****
DAL [41]		73.78 / *84.55*	90.41 / *95.59*	94.55 / *97.55*	97.17 / *98.85*	96.77 / *98.57*	96.18 / *98.46*

original code is either unavailable or not available in a complete form. For instance, the implementation of [40] misses the second stage described in the paper, and the code of [43] does not include the identity module, which is a fundamental element for model training. In both cases, our implementation would have certainly been sub-optimal, thus advantaging our framework. In other cases, despite our efforts, we could not reach satisfactory results. As an example, we tested the Age Estimation Guided CNN (AE-CNN [46]) implementation available from [63]. Despite the good results on CACD and MORPH2 datasets, this model (retrained and fine-tuned with the same recipe detailed in the following) achieved very low accuracies on $\mathcal{I}$ (less than 25% on all test sets).

The only significant possibility of comparison on $\mathcal{I}$ that we have found is that with the Decorrelated Adversarial Learning (DAL, [41]). DAL extracts image embeddings that can be compared for recognition using cosine distance. In our experiments, we used the DAL implementation from [64] training this model from scratch using MS-Celeb-1M dataset and leveraging the DEX model [65] to estimate the age of each picture (a piece of mandatory information for training). Then, we fine-tuned DAL on the held-out sets of 50.000 individuals used to train our GD classifiers (Section IV-C).

The **Test2** results in Table 3 show that DAL substantially improves all combinations of embedding models and classifiers except those based on ArcFace. However, results also

TABLE 4. Comparison with state of the art on MORPH2, CACD-VS and FG-NET. Best results are highlighted in bold, and runner-ups are underlined.

MORPH2	
LF-CNN [38]	97.51
AE-CNN [46]	98.13
Shakeel et al. [39]	98.67
DAL [41]	98.97
Zhao et al. [43]	<u>99.65</u>
ArcFace + NL–PSVM	**99.96**

CACD-VS	
CARC [34]	94.2
AG-IIM [17]	95.62
DAL [41]	99.40
Zhao et al. [43]	<u>99.76</u>
MFR [42]	**99.78**
ArcFace + NL–PSVM	99.53

FG-NET	
LF-CNN [38]	88.1
AG-IIM [17]	88.23
Shakeel et al. [39]	92.33
Zhao et al. [43]	93.20
DAL [41]	**94.5**
ArcFace + NL–PSVM	<u>94.44</u>

show that DAL is still affected by large performance drops when the age effects are more pronounced (groups A to C) and that the optimal combination (ArcFace + NL–PSVM) is more effective in separating identity and age information across all age groups. In particular, our solution provides significant accuracy improvements over DAL in set A (R1: +37.40% for males and +17.75% for females) and set B (R1: +14.79% for males and +6.92% for females).

E. COMPARISONS ON CROSS-AGE BENCHMARKS

Table 4 shows a comparison between our optimal model (ARcFace + NL–PSVM) and different methods on three challenging face-aging benchmarks, namely MORPH2, CACD-VS and FG-NET.

The non-commercial version of MORPH2 (i.e., MORPH Album2) contains about 55,134 face images from 13,000 individuals, divided into a training set of 10,000 samples and a test set of 3,000 identities. MORPH2 is the only benchmark that provides a fair number of training samples for our model. In this case, we created a unique GI model for males and females. The probe set consists of each subject's oldest image and the gallery of their youngest image, with an average age-gap around 1.5 years.

FG-NET contains 1,002 images from 82 subjects with considerable variability in age, covering from child to elder, and large variations of expression, illumination and pose. We followed the experimental protocol used in [39], [41], [43] for a fair comparison with previous methods. However, the small number of training identities is insufficient for our model. Thus, we used as the training set the IMDb-Face dataset [66], which contains cross-age images from 59K identities, defining a unique GI model for NL–PSVM. Since the same dataset size issues affect CACD-VS, which comprises 4,000 images from 2,000 celebrities, we again trained a GI NL–PSVM model on IMDb-Face (whose identities have no overlap with that of CACD-VS). In our experiments, we followed the test protocol defined in [34], [38].

Results in Table 4 show that our approach can reach state-of-the-art results in all the benchmarks. In particular, we obtain impressive results on MORPH2, outperforming all other methods. As for the other datasets, our optimal model is the runner up in FG-NET, and it is very close to the state-of-the-art results in CACD-VS. One possible explanation of this sub-optimal behavior on these latter datasets is the domain shift between training and test set in FG-NET and CACD-VS, which is not present in the experiments with ID-DATASET and MORPH2. Indeed, the effectiveness of the PLDA–based classifiers is reduced when there is a mismatch between the training and test distributions [67]. Thus, results on CACD-VS and FG-NET highlight the need to extend the current approaches adapting them to handle possible changes in the domain between training and test phases, which is left as future work.

F. ABLATION STUDY

In the following, we briefly discuss the effect of various parameters involved in our approach.

1) PATCH PARAMETERS

As stated in Section IV-C2, facial patch extraction relies on two parameters, *padding* (resize factor) and *vp* (vertical translation). For each embedding model, these parameters were fine-tuned using a grid search approach on the 50K individual set used to train the classifiers. One relevant finding of this process was that, while most of the models show moderate sensibilities to these parameters, their sub–optimal choice result in dramatic accuracy losses for ArcFace.

As an example, Table 5 reports the R1 male accuracies of **Test2** for baseline and NL–PSVM of two versions of SENet and ArcFace embeddings. Both versions are computed with the same *padding* but with different values of *vp* (namely, 0.3 and 0.4). It can be seen that the accuracy drops are extremely different between ArcFace and SENet. Experiments on LFW and CALFW with the same features confirm this finding. Changing *vp* from 0.3 to 0.4, the accuracy drops by 0.20% on LFW and 1.47% on CALFW for SENet, while for ArcFace the differences are 7.75% on LFW and 25.08% on CALFW.

A possible explanation of this peculiar behavior of ArcFace is related to the way the model was trained. The combination of a small input size (112 × 112) and (apparently) no use of data augmentation (the original paper does not report evidence of that) might cause a high sensibility to translation. In particular, the vertical eye position seems to be a crucial parameter for recognition.

2) FACIAL PATCHES

In order to show how the number of random facial patches used to compute the embeddings (Section IV-C2) affects the accuracy, we ran the following tests. For each model, we computed (on the classifiers training set) the R1 and R4 accuracies of the embeddings obtained with a varying number of patches

TABLE 5. Test2 R1 male results of SENet and ArcFace using different vertical alignment parameters.

vp	Classifier	A	B	C	D	E	F
SENet							
0.3	Baseline	8.53	29.70	48.08	68.56	73.86	69.19
0.4	Baseline	7.66	32.39	55.26	78.11	82.60	79.37
0.3	NL-PSVM	16.70	38.86	60.25	75.00	78.19	72.67
0.4	NL-PSVM	23.98	49.54	72.33	86.59	87.17	84.37
ArcFace							
0.3	Baseline	1.37	2.52	5.65	14.48	17.31	21.11
0.4	Baseline	59.71	86.04	96.64	99.47	99.69	99.80
0.3	NL-PSVM	20.47	38.54	58.97	74.00	74.97	74.44
0.4	NL-PSVM	80.08	95.02	99.06	99.87	99.81	99.81

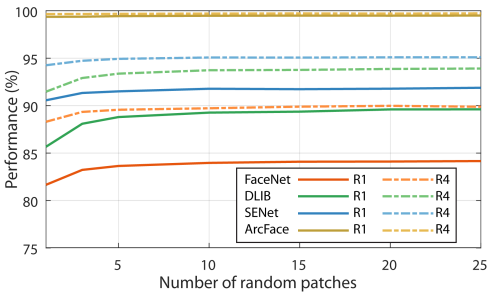

FIGURE 9. R1 and R4 recognition accuracy vs. patch number.

TABLE 6. Comparison of different scoring functions: (Test2 results for ArcFace + NL–PSVM).

Function	A	B	C	D	E	F
Males						
best	74.98	93.74	98.77	99.81	99.76	99.72
avg	76.85	94.38	99.04	99.82	**99.86**	99.72
oldest	**80.08**	**95.02**	**99.06**	**99.87**	99.81	**99.81**
Females						
best	89.03	96.53	98.23	99.23	99.09	98.53
avg	90.56	**97.33**	**98.67**	99.36	99.22	**98.93**
oldest	**91.53**	**97.33**	98.66	**99.37**	**99.26**	98.76

(from 1 to 25). We used as score the natural distance in the embedding space and, for each individual, the whole set was used as gallery. Results in Fig 9 clearly show that (i) a greater number of patches help better deal with image variations and (ii) the accuracy starts saturating for all models when the number of patches is higher than ten and reaches a plateau for 20.

3) SCORING FUNCTION

The scoring function defined in (10), referred to in the following as *oldest*, employs the most recent (i.e. oldest age) image, and thus requires the knowledge of the age attribute of all images. When this information is not available, other options should be chosen, such as computing the similarity as the minimal (*best*) or average (*avg*) score between a probe and all gallery samples of an individual.

To analyze the contribution on the accuracy of the scoring function, in Table 6 we detail the **Test2** results obtained with these three different scoring functions on **Test2** using the combination ArcFace + NL–PSVM (these results are again consistent with those obtained on other combinations of embedding models and classifiers). It can be seen that *oldest* scoring is the optimal choice in most of the cases and that when age information is not available, *avg* function provides excellent results as well and consistently outperforms *best* function. It can also be seen that the only (real) critical case is set A, where switching from *oldest* to *avg* causes an accuracy loss of 3.23% for males and 0.87% for females.

4) COMBINING FINE-TUNING AND PLDA–BASED CLASSIFIERS

Another question related to the analysis of our approach is the following. If we have enough data to train the PLDA–based classifiers, we also have enough data for fine-tuning the embedding networks. Then, which is the effect of this fine-tuning? And which is the effect of combining it with the proposed classifiers? To answer these questions, we run other experiments (on the **Test2** protocol) using as feature extractor ArcFace (the optimal embedding model among the ones analyzed) and DAL. As for DAL, we also combined these features (in both the standard and fine-tuned versions) with the proposed PLDA–based classifiers, thus verifying as well if these models allow improvements even with approaches already aimed at softening the effects of age-related information. For a fair comparison, the original ArcFace model (trained on a huge and undisclosed dataset) was re-trained on the same MS-Celeb-1M dataset used to train our DAL model. In the following, we refer to this re-trained version as ArcFace*. Then, we fine-tuned both DAL and ArcFace* on the held-out sets of 50.000 individuals used to train the GD classifiers (Section IV-C).

The results, summarized in Table 7, are somewhat mixed. As for the baselines, the fine-tuned versions of both ArcFace* and DAL largely improve the accuracies of the non fine-tuned ones. Comparing the non fine-tuned versions of the two models, ArcFace* is still superior to DAL, with larger differences between the two in the sets A-C. In contrast, for their fine-tuned counterparts, results for female are substantially equivalent, while ArcFace* provides higher accuracies for male. PLDA classifiers applied to non fine-tuned DAL offer similar effects to fine-tuning. Surprisingly, the combination of fine-tuning and PLDA-based classifiers has a detrimental impact on DAL, with an average R1 accuracy decrease of -3.00% for female and -1.50% for male. On the contrary, there is a $+4.56\%$ and $+4.99\%$ increase for, respectively, female and male with ArcFace*, and much more significant improvements for the most critical A and B sets. One possible explanation for this behavior is that fine-tuned DAL features are more effective than the not fine-tuned ones in separating age and identity information in $\mathcal{I}$. As a result, PLDA–based

TABLE 7. Comparison on Test2 protocol of DAL and ArcFace* (i.e., the ArcFace model trained on MS-Celeb-1M dataset) in their standard and fine-tuned versions (indicated as DAL-FT and ArcFace*-FT) combined with different classifiers. For each age group (A to F), numbers in bold (underlined) represent the best result (the runner–up) for each embedding model and '*' denote the overall optimal value.

Model	Classifier	A	B	C	D	E	F
				Males			
DAL	Baseline	13.78 / *22.53*	50.76 / *64.43*	77.73 / *87.78*	94.48 / *97.71*	95.84 / *98.30*	95.50 / *98.39*
	PLDA	15.81 / *26.00*	52.34 / *66.27*	78.65 / *88.47*	92.81 / *97.27*	93.70 / *97.66*	92.70 / *97.17*
	NL–PLDA	24.96 / *40.22*	58.66 / *72.36*	81.47 / *90.54*	94.93 / *98.01*	95.45 / *98.11*	93.52 / *97.50*
	PSVM	24.14 / *39.13*	62.70 / *76.76*	85.04 / *93.17*	95.76 / *98.46*	96.39 / *98.85*	95.26 / *98.44*
	NL–PSVM	**25.97** / ***41.77***	**64.83** / ***77.73***	**86.38** / ***93.98***	**96.58** / ***98.80***	**97.06** / ***99.00***	**95.43** / ***98.56***
DAL–FT	Baseline	**43.08** / ***60.29***	**80.23** / ***88.72***	**95.25** / ***98.12***	**99.19** / ***99.68***	**99.38*** / ***99.73***	**99.02*** / ***99.69****
	PLDA	26.45 / *40.35*	66.65 / *78.94*	88.88 / *94.92*	96.99 / *99.06*	97.44 / *99.07*	96.85 / *98.93*
	NL–PLDA	32.90 / *48.90*	70.66 / *81.90*	89.93 / *95.34*	97.72 / *99.22*	97.48 / *99.14*	96.22 / *98.63*
	PSVM	42.14 / *59.61*	77.52 / *87.18*	93.17 / *97.21*	98.39 / *99.52*	98.42 / *99.45*	97.93 / *99.39*
	NL–PSVM	37.04 / *54.74*	76.08 / *86.41*	93.01 / *97.18*	98.67 / *99.57*	98.39 / *99.26*	97.37 / *99.26*
ArcFace*	Baseline	26.46 / *41.35*	69.62 / *80.95*	88.74 / *95.14*	97.96 / *99.41*	98.61 / *99.64*	98.83 / *99.68*
	PLDA	21.23 / *35.52*	64.58 / *77.12*	85.57 / *93.11*	95.71 / *98.45*	97.06 / *99.11*	97.41 / *99.04*
	NL–PLDA	30.46 / *48.80*	70.66 / *83.12*	90.52 / *96.14*	97.83 / *99.43*	97.85 / *99.35*	96.94 / *98.92*
	PSVM	33.28 / *52.15*	75.24 / *86.20*	91.67 / *97.09*	98.07 / *99.42*	98.28 / *99.52*	**98.15** / ***99.54***
	NL–PSVM	**34.66** / ***54.56***	**76.08** / ***87.50***	**92.98** / ***97.70***	**98.49** / ***99.60***	**98.56** / ***99.62***	98.04 / *99.43*
ArcFace*–FT	Baseline	71.63 / *82.73*	86.43 / *92.57*	94.98 / *97.49*	98.26 / *99.17*	98.25 / *99.16*	97.35 / *98.94*
	PLDA	78.77 / *87.87*	90.96 / *95.45*	96.88 / *98.56*	98.75 / *99.46*	98.71 / *99.48*	97.91 / *99.22*
	NL–PLDA	85.03 / *92.43*	94.11 / *97.43*	97.93 / *99.01*	99.43 / *99.83*	99.31 / ***99.76****	**98.89** / *99.61*
	PSVM	84.72 / *92.11*	94.12 / *97.62*	97.94 / *99.13*	99.40 / *99.84*	99.18 / *99.60*	98.57 / *99.56*
	NL–PSVM	**85.76*** / ***92.83****	**94.97*** / ***97.82****	**98.33*** / ***99.19****	**99.56*** / ***99.86****	**99.38*** / *99.71*	98.87 / ***99.69****
				Females			
DAL	Baseline	39.42 / *51.51*	66.67 / *78.61*	77.75 / *87.43*	85.90 / *92.56*	84.83 / *91.80*	81.62 / *89.84*
	PLDA	43.56 / *56.94*	68.26 / *80.65*	78.33 / *87.84*	85.81 / *92.67*	84.23 / *92.15*	82.36 / *90.46*
	NL–PLDA	50.10 / *64.93*	72.64 / *84.07*	82.05 / *90.04*	88.84 / *94.66*	87.57 / *93.95*	84.38 / *91.64*
	PSVM	52.90 / *67.61*	76.06 / *86.42*	84.16 / *91.54*	90.03 / *95.46*	89.63 / *95.05*	87.48 / *93.76*
	NL–PSVM	**54.53** / ***69.65***	**77.65** / ***87.98***	**85.84** / ***92.70***	**91.63** / ***96.14***	**90.80** / ***95.36***	**88.08** / ***94.45***
DAL–FT	Baseline	**73.78** / ***84.55***	**90.41** / ***95.59***	**94.55** / ***97.55***	**97.17** / ***98.85***	**96.77** / ***98.57***	**96.18** / ***98.46***
	PLDA	59.95 / *73.59*	81.08 / *90.18*	88.66 / *93.94*	93.30 / *97.00*	93.14 / *96.75*	91.52 / *95.98*
	NL–PLDA	62.00 / *74.84*	82.52 / *90.91*	89.54 / *94.65*	93.93 / *97.31*	93.14 / *96.70*	91.50 / *96.08*
	PSVM	69.03 / *81.23*	86.66 / *93.38*	91.92 / *96.07*	95.29 / *98.18*	94.64 / *97.54*	93.54 / *97.36*
	NL–PSVM	66.54 / *79.37*	86.05 / *93.26*	92.01 / *96.11*	95.41 / *98.07*	94.64 / *97.70*	93.25 / *97.34*
ArcFace*	Baseline	41.80 / *56.16*	71.10 / *82.24*	81.78 / *90.60*	89.39 / *95.03*	90.13 / *95.51*	89.86 / *95.42*
	PLDA	39.14 / *53.27*	66.96 / *79.31*	78.12 / *88.09*	86.71 / *93.22*	86.87 / *94.08*	87.51 / *93.54*
	NL–PLDA	48.86 / *64.70*	74.54 / *85.77*	84.99 / *92.29*	91.41 / *96.18*	90.82 / *95.79*	89.44 / *94.97*
	PSVM	**54.44** / ***70.24***	78.41 / *88.59*	86.46 / *93.63*	91.98 / *96.39*	91.89 / *96.66*	91.39 / *96.18*
	NL–PSVM	52.90 / *69.45*	**78.97** / ***89.24***	**87.84** / ***94.05***	**93.57** / ***97.24***	**93.04** / ***97.03***	**91.62** / ***96.68***
ArcFace*–FT	Baseline	78.44 / *87.27*	89.77 / *94.57*	94.01 / *97.06*	95.96 / *98.20*	95.94 / *98.18*	95.11 / *97.89*
	PLDA	85.66 / *92.34*	93.40 / *97.17*	96.55 / *98.52*	97.72 / *99.01*	97.46 / *99.13*	96.68 / *98.45*
	NL–PLDA	89.00 / *94.49*	94.68 / *97.73*	97.12 / *98.67*	98.31 / *99.15*	**98.13*** / *99.41*	97.53 / *98.93*
	PSVM	88.95 / *94.61*	95.05 / *98.02*	97.27 / *98.74*	98.15 / *99.18*	98.09 / *99.39*	97.36 / *98.74*
	NL–PSVM	**89.60*** / ***94.87****	**95.48*** / ***98.15****	**97.54*** / ***98.80****	**98.38*** / ***99.22****	98.07 / ***99.46****	**97.55*** / ***99.02****

approaches are forced to focus on other across-class variations that, when removed, hinder the AIFR task. Another possible explanation is an overfitting effect during DAL fine-tuning (the same images were used to fine-tune the network and to train the backend classifiers), which results in too large differences between the distributions of training and test data.

VI. CONCLUSION

This paper proposed a thorough analysis of the contribution of various models derived from basic PLDA in improving the robustness of AIFR. The approach combines image embeddings, extracted with effective state-of-the-art deep convolutional models, with generative and discriminative methods extending PLDA and aimed at better separating the across-class variations (i.e., the identities) from the within-class ones (i.e., those related to various image changes such as illumination, pose and aging). The proposed approach has been assessed on a challenging test suite, addressing the identification problem in an extensive dataset containing hundreds of thousands of different identities, spanning different age intervals, and showing significant age gaps among the same individual's pictures. Our results highlight the difficulties that all the analyzed embedding models have to face when there is a significant age difference between the

compared pictures, mainly when it corresponds to substantial changes in the facial appearance (e.g., passing from childhood to adolescence and from adolescence to adulthood). In this context, our results show that PLDA–based approaches are effective classifiers in the embedding space, with the non–linear versions outperforming the linear ones. In particular, among the different analyzed options, the combination of ArcFace as feature extractor and NL–PSVM as classifier obtains the best performances and state-of-the-art results on different standard cross-age benchmarks.

However, our results on CACD-VS and FG-NET show as well that the proposed PLDA–based classifiers are negatively affected by domain shift issues. To address this problem, as future work, we plan to extend the current approaches by integrating domain adaptation methods to deal with possible mismatches between the training and test distributions.

REFERENCES

[1] S. Zago, G. Sartori, and G. Scarlat, "Malingering and retrograde amnesia: The historic case of the collegno amnesic," *Cortex*, vol. 40, no. 3, pp. 519–532, 2004, doi: 10.1016/S0010-9452(08)70144-8.

[2] NCIC. (2017). *NCIC Missing Person and Unidentified Person Statistics*. Accessed: Feb. 15, 2021. [Online]. Available: https://bit.ly/2EJ3aRP

[3] NMPCC. *Australian National Missing Persons Coordination Centre Report*. Accessed: Feb. 15, 2021. [Online]. Available: https://bit.ly/2YUoFJQ

[4] MCE. (2017). *Missing Children Europe Annual Review*. Accessed: Feb. 15, 2021. [Online]. Available: https://missingchildreneurope.eu/?wpdmdl=926

[5] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1701–1708, doi: 10.1109/CVPR.2014.220.

[6] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. BMVC*, 2015, pp. 41.1–41.12.

[7] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823, doi: 10.1109/CVPR.2015.7298682.

[8] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.

[9] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699, doi: 10.1109/CVPR.2019.00482.

[10] T. Zheng, W. Deng, and J. Hu, "Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments," 2017, *arXiv:1708.08197*. [Online]. Available: http://arxiv.org/abs/1708.08197

[11] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The MegaFace benchmark: 1 million faces for recognition at scale," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4873–4882, doi: 10.1109/CVPR.2016.527.

[12] S. Ioffe, "Probabilistic linear discriminant analysis," in *Proc. 9th Eur. Conf. Comput. Vis.*, May 2006, pp. 531–542, doi: 10.1007/11744085.

[13] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8, doi: 10.1109/ICCV.2007.4409052.

[14] P. Li, Y. Fu, U. Mohammed, J. H. Elder, and S. J. D. Prince, "Probabilistic models for inference about identity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 144–157, Jan. 2012, doi: 10.1109/TPAMI.2011.104.

[15] L. El Shafey, C. McCool, R. Wallace, and S. Marcel, "A scalable formulation of probabilistic linear discriminant analysis: Applied to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1788–1794, Jul. 2013, doi: 10.1109/TPAMI.2013.38.

[16] D. Gong, Z. Li, D. Lin, J. Liu, and X. Tang, "Hidden factor analysis for age invariant face recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2872–2879, doi: 10.1109/ICCV.2013.357.

[17] H. Zhou and K.-M. Lam, "Age-invariant face recognition based on identity inference from appearance age," *Pattern Recognit.*, vol. 76, pp. 191–202, Apr. 2018, doi: 10.1016/j.patcog.2017.10.036.

[18] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proc. Odyssey*, 2010, p. 14.

[19] S. Cumani, N. Brummer, L. Burget, P. Laface, O. Plchot, and V. Vasilakakis, "Pairwise discriminative speaker verification in the *I*-vector space," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 6, pp. 1217–1227, Jun. 2013, doi: 10.1109/TASL.2013.2245655.

[20] S. Cumani, O. Plchot, and P. Laface, "On the use of i–vector posterior distributions in probabilistic linear discriminant analysis," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 4, pp. 846–857, Apr. 2014, doi: 10.1109/TASLP.2014.2308473.

[21] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7649–7653, doi: 10.1109/ICASSP.2013.6639151.

[22] S. Cumani, "Fast scoring of full posterior PLDA models," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 11, pp. 2036–2045, Nov. 2015, doi: 10.1109/TASLP.2015.2464678.

[23] S. Cumani and P. Laface, "Joint estimation of PLDA and nonlinear transformations of speaker vectors," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 10, pp. 1890–1900, Oct. 2017, doi: 10.1109/TASLP.2017.2724198.

[24] N. Brummer, A. Silnova, L. Burget, and T. Stafylakis, "Gaussian meta-embeddings for efficient scoring of a heavy-tailed PLDA model," in *Proc. Odyssey Speaker Lang. Recognit. Workshop*, Jun. 2018, pp. 349–356, doi: 10.21437/Odyssey.2018-49.

[25] S. Cumani and P. Laface, "I–vector transformation and scaling for PLDA based speaker recognition," in *Proc. Odyssey*, Jun. 2016, pp. 39–46, doi: 10.21437/Odyssey.2016-6.

[26] S. Cumani and P. Laface, "Large-scale training of pairwise support vector machines for speaker recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 11, pp. 1590–1600, Nov. 2014, doi: 10.1109/TASLP.2014.2341914.

[27] 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, Tech. Rep. 07-49, 2007.

[28] I. Masi, Y. Wu, T. Hassner, and P. Natarajan, "Deep face recognition: A survey," in *Proc. 31st SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI)*, Nov. 2018, pp. 471–478, doi: 10.1109/SIBGRAPI.2018.00067.

[29] G. Guo and N. Zhang, "A survey on deep learning based face recognition," *Comp. Vis. Image Underst.*, vol. 189, Dec. 2019, Art. no. 102805, doi: 10.1016/j.cviu.2019.102805.

[30] G. Antipov, M. Baccouche, and J.-L. Dugelay, "Face aging with conditional generative adversarial networks," 2017, *arXiv:1702.01983*. [Online]. Available: http://arxiv.org/abs/1702.01983

[31] G. Antipov, M. Baccouche, and J.-L. Dugelay, "Boosting cross-age face verification via generative age normalization," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 191–199, doi: 10.1109/BTAS.2017.8272698.

[32] H. Ling, S. Soatto, N. Ramanathan, and D. W. Jacobs, "Face verification across age progression using discriminative methods," *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 1, pp. 82–91, Mar. 2010, doi: 10.1109/TIFS.2009.2038751.

[33] D. Sungatullina, J. Lu, G. Wang, and P. Moulin, "Multiview discriminative learning for age-invariant face recognition," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–6, doi: 10.1109/FG.2013.6553724.

[34] B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset," *IEEE Trans. Multimedia*, vol. 17, no. 6, pp. 804–815, Jun. 2015, doi: 10.1109/TMM.2015.2420374.

[35] D. Gong, Z. Li, D. Tao, J. Liu, and X. Li, "A maximum entropy feature descriptor for age invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5289–5297, doi: 10.1109/CVPR.2015.7299166.

[36] Z. Li, D. Gong, X. Li, and D. Tao, "Aging face recognition: A hierarchical learning model based on local patterns selection," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2146–2154, May 2016, doi: 10.1109/TIP.2016.2535284.

[37] S. Bianco, "Large age-gap face verification by feature injection in deep networks," *Pattern Recognit. Lett.*, vol. 90, pp. 36–42, Apr. 2017, doi: 10.1016/j.patrec.2017.03.006.
[38] Y. Wen, Z. Li, and Y. Qiao, "Latent factor guided convolutional neural networks for age-invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4893–4901, doi: 10.1109/CVPR.2016.529.
[39] M. S. Shakeel and K.-M. Lam, "Deep-feature encoding-based discriminative model for age-invariant face recognition," *Pattern Recognit.*, vol. 93, pp. 442–457, Sep. 2019, doi: 10.1016/j.patcog.2019.04.028.
[40] C. Xu, Q. Liu, and M. Ye, "Age invariant face recognition and retrieval by coupled auto-encoder networks," *Neurocomputing*, vol. 222, pp. 62–71, Jan. 2017, doi: 10.1016/j.neucom.2016.10.010.
[41] H. Wang, D. Gong, Z. Li, and W. Liu, "Decorrelated adversarial learning for age-invariant face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3527–3536, doi: 10.1109/CVPR.2019.00364.
[42] J. Guo, X. Zhu, C. Zhao, D. Cao, Z. Lei, and S. Z. Li, "Learning meta face recognition in unseen domains," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6163–6172, doi: 10.1109/CVPR42600.2020.00620.
[43] J. Zhao, Y. Cheng, Y. Cheng, Y. Yang, F. Zhao, J. Li, H. Liu, S. Yan, and J. Feng, "Look across elapse: Disentangled representation learning and photorealistic cross-age face synthesis for age-invariant face recognition," in *Proc. AAAI*, 2019, pp. 9251–9258, doi: 10.1609/aaai.v33i01.33019251.
[44] X. Shu, J. Tang, H. Lai, L. Liu, and S. Yan, "Personalized age progression with aging dictionary," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3970–3978, doi: 10.1109/ICCV.2015.452.
[45] X. Shu, J. Tang, Z. Li, H. Lai, L. Zhang, and S. Yan, "Personalized age progression with bi-level aging dictionary learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 905–917, Apr. 2018, doi: 10.1109/TPAMI.2017.2705122.
[46] T. Zheng, W. Deng, and J. Hu, "Age estimation guided convolutional neural network for age-invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 12–16, doi: 10.1109/CVPRW.2017.77.
[47] N. Brümmer and E. De Villiers, "The speaker partitioning problem," in *Proc. Odyssey*, Jun. 2010, pp. 194–201.
[48] D. Garcia-Romero and C. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. Interspeech*, 2011, pp. 249–252.
[49] J. Alam, N. Brummer, L. Burget, M. Diez, O. Glembek, P. Kenny, M. Klco, F. Landini, A. Lozano-Diez, P. Matejka, and G. Bhattacharya, "ABC NIST SRE 2018 system description," in *Proc. NIST SRE Workshop*, 2018, pp. 1–10.
[50] B. Vesnicer, J. Ž. Gros, N. Pavešić, and V. Štruc, "Face recognition using simplified probabilistic linear discriminant analysis," *Int. J. Adv. Robotic Syst.*, vol. 9, no. 5, p. 180, Nov. 2012, doi: 10.5772/52258.
[51] A. Fabris, M. A. Nicolaou, I. Kotsia, and S. Zafeiriou, "Dynamic probabilistic linear discriminant analysis for video classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2781–2785, doi: 10.1109/ICASSP.2017.7952663.
[52] S. Cumani and P. Laface, "Generative pairwise models for speaker recognition," in *Proc. Odyssey*, Jun. 2014, pp. 273–279.
[53] S. Cumani and P. Laface, "Nonlinear I-Vector transformations for PLDA-based speaker recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 4, pp. 908–919, Apr. 2017, doi: 10.1109/TASLP.2017.2674966.
[54] S. Cumani and P. Laface, "Training pairwise support vector machines with large scale datasets," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 1645–1649, doi: 10.1109/ICASSP.2014.6853877.
[55] D. E. King. (2017). *DLIB Face Recognition Model*. Accessed: Feb. 15, 2021. [Online]. Available: http://dlib.net/files/dlib_face_recognition_resnet_model_v1.dat.bz2
[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
[57] D. Sandberg. (2018). *Facenet Face Recognition Models*. Accessed: Feb. 15, 2021. [Online]. Available: https://github.com/davidsandberg/facenet/
[58] J. Deng, J. Guo, X. Niannan, and S. Zafeiriou. (2019). *ArcFace Pretrained Models*. Accessed: Feb. 15, 2021. [Online]. Available: https://github.com/deepinsight/insightface
[59] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016, doi: 10.1109/LSP.2016.2603342.
[60] S. Cumani, O. Glembek, N. Brummer, E. de Villiers, and P. Laface, "Gender independent discriminative speaker recognition in i-vector space," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 4361–4364, doi: 10.1109/ICASSP.2012.6288885.
[61] S. Gong, V. N. Boddeti, and A. K. Jain, "On the intrinsic dimensionality of image representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4893–4901, doi: 10.1109/CVPR.2019.00411.
[62] G. Guo, G. Mu, and K. Ricanek, "Cross-age face recognition on a very large database: The performance versus age intervals and improvement using soft biometric traits," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 3392–3395, doi: 10.1109/ICPR.2010.828.
[63] S. Rai. *AAE-CNN Models*. Accessed: Feb. 15, 2021. [Online]. Available: https://tinyurl.com/AE-CNN-code
[64] (2020). *DAL Face Recognition Model*. Accessed: Feb. 15, 2021. [Online]. Available: https://github.com/neverUseThisName/Decorrelated-Adversarial-Learning
[65] R. Rothe, R. Timofte, and L. V. Gool, "DEX: Deep expectation of apparent age from a single image," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 10–15, doi: 10.1109/ICCVW.2015.41.
[66] F. Wang, L. Chen, C. Li, S. Huang, Y. Chen, C. Qian, and C. Change Loy, "The devil of face recognition is in the noise," 2018, *arXiv:1807.11649*. [Online]. Available: http://arxiv.org/abs/1807.11649
[67] M. H. Rahman, A. Kanagasundaram, I. Himawan, D. Dean, and S. Sridharan, "Improving PLDA speaker verification performance using domain mismatch compensation techniques," *Comput. Speech Lang.*, vol. 47, pp. 240–258, Jan. 2018, doi: 10.1016/j.csl.2017.08.001.

PABLO NEGRI was born in La Plata, Argentine, in 1974. He received the degree in electronic engineering from the Universidad Nacional de La Plata, in 1998, and the Ph.D. degree in computer vision from the Université Pierre et Marie Curie-Paris VI, in 2008. Since 2010, he has been a Researcher with CONICET. He joined the Institute of Research in Computer Sciences (ICC), University of Buenos Aires, in 2017. His research interests include machine learning and pattern recognition applied to computer vision for intelligent traffic or biometric applications.

SANDRO CUMANI received the Ph.D. degree in computer and system engineering from the Politecnico di Torino, Italy, in 2012. He is currently an Assistant Professor with the Department of Control and Computer Engineering, Politecnico di Torino. His current research interests include machine learning, speech processing and biometrics, in particular speaker and language recognition.

ANDREA BOTTINO (Member, IEEE) is currently an Associate Professor with the Department of Control and Computer Engineering, Politecnico di Torino, where he is heading the Computer Graphics and Vision Research Group. His current research interests include computer vision, machine learning, human–computer interaction, serious games, and virtual and augmented reality.