Deep Face Recognition: a Survey

Iacopo Masi,[†] Yue Wu,[†] Tal Hassner,[§] and Prem Natarajan[†]
[†]Information Sciences Institute (ISI)
University of Southern California (USC), Marina del Rey, CA, USA
iacopo,yue_wu,pnataraj@isi.edu
[§]The Open University of Israel, Raanana, Israel
hassner@openu.ac.il

***Abstract*—Face recognition made tremendous leaps in the last five years with a myriad of systems proposing novel techniques substantially backed by deep convolutional neural networks (DCNN). Although face recognition performance sky-rocketed using deep-learning in classic datasets like LFW, leading to the belief that this technique reached human performance, it still remains an open problem in unconstrained environments as demonstrated by the newly released IJB datasets.**

This survey aims to summarize the main advances in deep face recognition and, more in general, in learning face representations for verification and identification. The survey provides a clear, structured presentation of the principal, state-of-the-art (SOTA) face recognition techniques appearing within the past five years in top computer vision venues.

The survey is broken down into multiple parts that follow a standard face recognition pipeline: (a) how SOTA systems are trained and which public data sets have they used; (b) face preprocessing part (detection, alignment, etc.); (c) architecture and loss functions used for transfer learning (d) face recognition for verification and identification. The survey concludes with an overview of the SOTA results at a glance along with some open issues currently overlooked by the community.

I. Introduction and Motivation

Recognizing people is one of the most important topics in computer vision and pattern recognition. Among various biometrics used for person recognition, the face is one of the most popular, since this ubiquitous biometric can be acquired in unconstrained environments while providing strong discriminative features for recognition. For this reason, face recognition became an extremely important tool that is used for augmenting the automatic capabilities of video-surveillance and security systems, video-analytics software, and thousands of applications in our daily lives like entertainment, smart shopping, and automatic face tagging in photo collections.

Unlike generic object recognition, face recognition—and more in general face analysis—is of great interest since it supplies machines with an automatic way to interpret humans and their interactions, along with their expressions and feelings.

Face recognition has been a mature discipline in computer vision for many years; the first effort is dated back to 1966 [1], [2]. It is interesting to revisit the words that Bledsoe used to describe the challenges met in developing automatic facial recognition tools:

> *This recognition problem is made difficult by the great variability in head rotation and tilt, lighting intensity and angle, facial expression, aging, etc. Some other attempts at facial recognition by machine have allowed for little or no variability in these quantities. Yet the method of correlation (or pattern matching) of unprocessed optical data, which is often used by some researchers, is certain to fail in cases where the variability is great. In particular, the correlation is very low between two pictures of the same person with two different head rotations.*

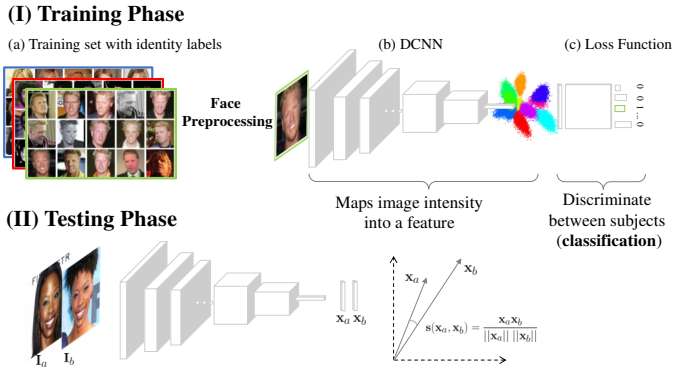

Fig. 1. **Face recognition pipeline.** Top: At training time, a huge labeled face set (a) is used to constrain the weights of a DCNN (b) optimizing a loss function (c) for the classification task. Bottom: At test time, the classification layer is often discarded, and the DCNN is used as a feature extractor for comparing face descriptors.

It is somewhat surprising that today, the main challenges for automatic recognition remain the same with those identified 52 years ago by Bledsoe and coincide with the problem of PIE (pose-illumination-expression) in face recognition. In order to obtain ideal recognition, a machine vision system needs to be able to fully correlate the identity signals given two face images, without being "distracted" by other confounding factors present in the media itself—such as pose variations, illuminations changes, and non-rigid expressions.

Recently, a big leap toward solving the PIE problem for face recognition was made with the introduction of massive training sets useful for training very deep convolutional neural networks (DCNN) [3]–[6]. This approach led to human-like performance (99% accuracy) in face verification on the set of Labeled Faces in the Wild (LFW) [7].

The powerful aspect of DCNN is the fact that, given an ample training set, these large-scale pattern recognition machines can be optimized end-to-end in order to develop features

that amplify the identity signal, while being robust to other PIE variations. Although the generalization ability of face recognition systems improved drastically, yet unconstrained face recognition in video-surveillance settings is far from being solved, and recently, a flurry of new methods have been proposed in top-tier conferences. In particular, the main effort was in pushing methods to perform well not only on celebrity faces (LFW) but also on surveillance footage and still images (IJB [8]).

The survey will review in detail methods in the last five years that addressed specific issues of different parts of a modern deep face recognition pipe. Fig. 1 depicts the structure of a common pipeline: given a training set, the approach usually performs some form of face preprocessing and trains a DCNN in order to classify a pool of subjects. Then, the network is used as a feature extractor to obtain descriptors that are matched together to perform face recognition.

The remainder of the survey is organized as follows: Sec. II briefly goes over face detection, since this is a necessary step although not the focus of this work; Sec. III reviews important topics of face preprocessing and the training collections; Sec. IV explains how to utilize training data through supervised learning of deep models; Sec. V offers common practices performed at test time to further improve face recognition accuracy, in the particular case that a subject is described by multiple media [8]; Sec. VI shows the progress of SOTA performance in the new challenging IJB-A set; and finally, we discuss possible open problems and conclude the survey in Sec. VII.

II. Face Detection

As one of the most famous problems in computer vision, face detection has been attracting attention for more than two decades. Yang *et al.* [9], and Zhao *et al.* [10] review face detection work which often focuses on developing discriminative hand-crafted features, and robust and efficient learning algorithms. With the powerful DCNNs, face detection performance has greatly improved in terms of both speed and accuracy. Detecting faces, at least frontal faces, is no longer a challenging task. Recent efforts have been made to improve face detection in challenging conditions, e.g. partial/occluded faces [11], faces in violent settings [12], and faces captured from depth sensors [13]. For a recent survey the reader can study [14].

III. Data Collections and Face Preprocessing

In this section we first review the publicly available data collections in the face recognition community used for training DCNNs, discussing the characteristics of each set with their advantages and disadvantages. Since face preprocessing is also very related to the input data, we also offer a review of the current techniques commonly used to preprocess a face.

A. Data Collections

With the rise of deep-learning [6] in recent years, a key aspect in developing face recognition systems is the training data used to learn face representations. Data collections are usually overlooked, but they are extremely important. Although some companies have internally labeled private face sets that scale to millions of images (Facebook [15]) or even millions of subjects (Google [17]), the situation is very different for publicly available collections. Currently, the main training sets available for the community are represented by the following: (i) CASIA WebFace [16]; (ii) VGGFace [18]; (iii) UMDFaces [19]; (iv) VGGFace2 [21]; (v) MS-Celeb-1M [20]; and (vi) IMDb-Face.

TABLE I
THE TABLE PRESENTS DATA COLLECTIONS USED FOR TRAINING AND TESTING. * DENOTES PRIVATE SETS.

Dataset	Subjects	(Img/Video)	Img. per sub.	Annotations	Curated	Year
Training Collections						
Facebook* [15]	4,030	4.4M/–	1,000	✗	✗	2014
CASIA [16]	10,575	494,414/–	46	✗	✓	2014
Google* [17]	8M	200M/–	25	✗	✗	2015
VGGFace [18]	2,622	2.6M/–	1,000	box, pose	✗	2015
UMDFaces [19]	8,277	367K/22K	45	keypts, pose, gender	✓	2016
MS-Celeb-1M [20]	100K	10M/–	100	box	✗	2016
VGGFace2 [21]	9,131	3.31M/–	362.6	pose, age	✓	2017
IMDb-Face [22]	1.7M	59K/–	28.8	pose, age	✓	2018
Benchmarks						
LFW [7]	5,749	13,233/–	2.3	✗	–	2007
YTF [23]	1,595	–/3,425	–	✗	–	2011
IJB-A [8]	500	5.7K/2K	11.4	box, pose, attributes	–	2015
MegaFace [24]	690,572	1.02M/–	1.5	✗	–	2016
IJB-B [25]	1,845	11.7K/7.0K	36.2	box	–	2017
IJB-C [26]	3,531	31K/11.8K	6	box	–	2018

CASIA WebFace is a dataset comprising around 500K images of 10K subjects. It was automatically collected by the CASIA group [16] and then manually refined. As is common for sets that are collected by looking at celebrities or famous people, this set presents a long tail distribution [27] in terms of the images that are associated to a subject. This means that there are some frequent and usually more famous subjects that comprise most of the images, while others are only described by a few images.

VGGFace is a dataset proposed by the Oxford group [18] for training deep models that comprises around 2.6M faces of 2,622 individuals. Unlike CASIA, the set has a flat distribution, i.e., each subject is described by about one thousand samples, most of which are of high-quality frontal faces since the images are scraped from web engines. Moreover, despite the effort of cleaning the set, the intra-class variance for a subject is also affected by noise (outliers), while the CASIA WebFace contains less severe errors.

UMDFaces is a face set released by [19]. Bansal *et al.* [19] used a mix of human annotators via Amazon Mechanical Turk (AMT) and already trained deep-based face analysis tools to build medium-sized sets that are much tougher than the already available sets—such as [16], [18]. Another UMDFaces peculiarity is the fact that, unlike CASIA and VGGFace, the set contains *both* still images (usually high quality) and video frames (often affected by motion blur). The set provides annotations of facial keypoints, face pose angles, gender information. These annotations are extracted automatically using [28]. The set consists of 367,888 face annotations in still

images for 8,277 subjects, and also 3.7 million annotated video frames from about 22K videos of 3,100 subjects. Although the UMDFaces numbers are smaller than the other sets, as mentioned by [19], it presents a wider pose distribution than CASIA and VGGFace.

MS-Celeb-1M was first released in the multimedia community, and it later spread throughout the computer vision community [20]. It has around 10 million images of 100K celebrities. Each celebrity has 100 images retrieved by the Bing search engine using the celebrity's name without any filtering in the retrieved results. All of the previous sets pale in comparison with the size of MS-Celeb-1M (10M images); however, the quality of the dataset is severely biased by label noise, duplicated images, non-face images present in the set, all of which makes it hard to use directly. After all, MS-Celeb-1M was released to learn from noisy labels and was never curated.

VGGFace2 is an improved version of VGGFace created in order to mitigate the deficiency of its predecessor. VGGFace2 [21] contains 3.31 million images of 9,131 subjects collected among celebrities, but also famous people such as professors or politicians. Compared to its predecessor, the average number of images decreased: on average, an individual is described by 362.6 media. Though this number dropped, VGGFace2 is designed to cover a large range of pose, age and ethnicity, and to reduce label noise as much as possible. The reduction in the label noise was achieved with the interplay of manual and automatic processes.

IMDb-Face was very recently announced in [22] as a novel set derived by cleaning the label noise in MS-Celeb-1M and MegaFace [24]. This new set claims to be the largest noise-controlled face collection; besides releasing the set, the paper [22] proposes rigorous ablation studies of the impact of label noise in training, both in terms of outliers or swapped labels.

In this section we reviewed only collections used for training. Nevertheless, it is worthwhile to also list benchmark datasets. Tab. I shows an overview of the discussed training collections along with evaluation datasets. Considering new benchmarks—besides the novel IJB sets— [8], [25], [26]—it is worth mentioning the recent effort called MegaFace [24] for benchmarking the performance of face algorithms with millions of "distractors": briefly, distractors correspond to a large number of individuals added to the gallery yet not present in the probe with the scope of confusing the recognition algorithm. Another effort worth to mention uses the *same, fixed* set for training [29] in order to better isolate the performance of the learning algorithm from the data size.

B. Face Preprocessing and Alignment

Although face recognition shares similarities with generic object recognition, there is a particular aspect which is proper of a face: faces have a well-structured shape that can be modeled very well. On the other hand, this is hard to do for generic objects since diverse classes could present very different shapes. Face preprocessing uses strong domain knowledge in order to properly modify the input face to ease the learning of face representations. Domain knowledge is represented by facial landmark detection, pose estimation, rendering and data augmentation. For a survey on this topic please refer to [30].

Generally speaking, face preprocessing corresponds to face alignment. Also, the word "face alignment" is often used interchangeably in the community to point to face landmark detectors [31]–[33]. In this survey we do not review landmark detection systems since these are used to constrain the alignment through their output; we instead are more interested in the types of transformations applied for face alignment.

More broadly, we will review common methods to achieve face alignment or to compensate for spatial changes of the face. After face detection, a face is always localized in an image with a bounding box, but there is no guarantee that faces will be aligned; moreover, it is often the case that the box is subject to jitter. A common approach to making the DCNN robust to spatial changes is data augmentation [34]: [17], [18], [35] employed strong in-plane data augmentation when training the network in order to make it robust to possible misalignments that the system encounters at test time. An advantage of this approach is that the method is kept very simple with no extra preprocessing tools. In order to improve performance at test-time, these methods often perform 2D data augmentation (perturbing the sample to create multiple, different copies of it) and average pooling in the feature space. It is interesting that, although they used strong in-plane augmentation at training time, some of them report [17], [18] improved recognition at test-time if face imagery is aligned with a 2D similarity transformation using detected face landmarks. This leads to an easy alternative way for alignment: the use of a simple 2D similarity transformation that compensates for scale, in-plane rotation and translation, and makes roughly all the salient parts of faces overlapping each other. This approach is used by [28], [36]–[39] with moderate augmentation to train the network. Ablation studies performed in [40] brought us to the conclusion that this form of alignment[1] constantly improves face recognition performance if applied in the same manner for both training and testing. Note that both [41] and [38] pointed out that in order to get a performance boost the 2D similarity needs to be tuned differently for frontal faces and for profile faces. Frontal faces are usually aligned with a few fiducial points (eye corners, tip of the nose, mouth corners, etc.); since out-of-plane rotation of the head causes a drastic change of the face image, in this case, faces are aligned using only fiducial visible on the profile contour.

With the intent of unifying the preprocessing step, Ranjan *et al.* [28] created an all-in-one neural network for face detection and face preprocessing trained with multi-task learning for different face analysis tasks, while other researchers developed more advanced face alignment methods [15], [42]–[44].

A 2D similarity is a transformation that is defined up to 4 four DoF (scale, rotation angle, and 2D translation).

[1]The authors of [40] refer to this step as "face thumbnail creation."

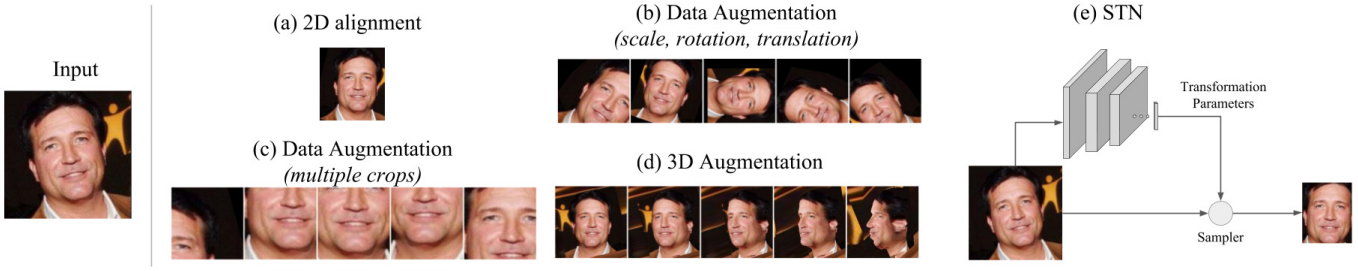

Fig. 2. **Different face alignment methods.** An input face (left) can be aligned with: (a) 2D alignment with a similarity transformation (b) data augmentation for scale, rotation and translation (c) multi-crops (d) 3D face-specific augmentation to synthesize novel poses (e) STN does not use any other external tools for alignment yet a subnetwork learns transformation parameters that are then applied to the RGB input tensor (or further, even to feature maps).

Therefore, the deformation power can be limited—e.g., a 2D similarity cannot compensate for out-of-plane variations. To overcome this limitation, the community developed more complex face alignment taking advantage of 3D face shapes either using a 3D generic model [42], [43] or a subject-specific 3D model [15]. The process of warping an image using a 3D model to make it artificially face a frontal view is referred to as "face frontalization"[2] [42], [45], [46]. Notably, frontalization improved the LFW accuracy of both hand-crafted features [42] and deeply learned [15], with the LFW benchmark quickly reaching saturation with DCNNs [47], [48].

With the introduction of more general evaluation benchmarks such as IJB-A [8] (IARPA JANUS Benchmark A) and IJB-B [25] that offer a wider pose distribution, the frontalization effort has been cast as a new way of performing data augmentation introducing novel, unseen poses in the training set [49]. Instead of warping the face to a frontal view, Masi *et al.* [44] precompute off-line multiple rendered views so that a face can be seen from any arbitrary frontal, half-profile and profile view. Face-specific augmentation has been shown [49] to greatly improve performance on IJB sets; furthermore, it was later on extended to augment for illumination variations in [50]. Efficient 3D face rendering [44] was also recently coupled with a dedicated network (FacePoseNet [51]), directly regressing 3D head pose and thereby bypassing fragile landmark detectors. Notably, all the above methods used a generic shape or a bank of generic models, nevertheless, the community is beginning to develop robust deep-based 3D face estimation methods [52], [53]. The advantages of face-specific augmentation are that it can be used to greatly enlarge the intra-class variations of pose-deficient training sets—yet hard alignment could cause artifacts, if not robust enough; to overcome this problem, recently the idea of "recognition by synthesis" has been augmented with an auxiliary step using a GAN-based network [54], in which synthetically generated images are refined by a dual-agent GAN [55].

All of the methods described above design the alignment beforehand using some prior knowledge; contrarily, there is a method proposed in [56], dubbed Spatial Transformer Network (STN), that performs alignment end-to-end inside the network while training. A specialized sub-net regresses the alignment parameters given an input sample, and a differentiable grid is used to deform the face; thereafter the training proceeds in the usual way. Note that, the transformation inside the STN can be arbitrarily defined (common choices are similarity or affine transformations), but, unlike previous methods, the parameters of the STN are learned online with the objective of easing the classification task without any other loss on the alignment. Although this method is very versatile, since it removes the need for other face preprocessing tools that can bog a pipeline down, currently it is not often used in the face recognition community which instead prefers to fall back to landmark detectors or pose estimation methods [51], [53]. Fig. 2 provides a summary of all the preprocessing methods described above. For more details on face preprocessing and its interactions with DCNN, one can read [57].

[2]In some cases, given that this method performs dense alignment guided by the manifold of the 3D face, it is also called hard alignment since the face is aligned on a 3D shape; other papers mention it as 3D rectification.

IV. Network Architectures, Loss Functions and Disentanglement

After the training data is preprocessed using methodologies presented in Sec. III-B, a standard pipeline proceeds in learning face representations by defining a DCNN architecture and a loss function for classification. Most of the work reviewed performs transfer learning: this means training the DCNN on a closed pool of subjects to then use the DCNN as a descriptor extractor on unseen faces.

A. Architectures and Loss Functions

We briefly summarize the different network architectures that are optimized for face recognition. The first recent paper that resumed the use of DCNNs for face recognition is DeepFace [15], in which convolutions and locally connected layers are used to learn representations from input images all registered together via 3D frontalization [42]. Although locally connected layers are in theory the preferable solution for the face domain since we can register the data very well, yet standard pipelines substantially use network building blocks developed for generic object recognition with convolutional layers, batch normalization (BN), non-linear activation (ReLu), and pooling layers. Examples of these architectures are VGG16 [18] and networks with residual connections, such as ResNet-50 [58]. Also, very often these DCNNs are pretrained for generic object

recognition on the ImageNet Large Scale Visual Recognition Competition (ILSVRC) dataset [59].

Researchers made attempts to modify DCNN architectures for better handling pose variations: Cao *et al.* [60] used the property of equivariant mapping to propose a novel block in the final stage of the encoder. The proposed DREAM block adds residuals to the feature representation to map back a profile face to a canonical pose to simplify the classification.

Besides developing novel architectures, the major part of the community made an effort to develop new loss functions on top of the SoftMax layer based on cross-entropy (Eq. (1)):

$$\ell_{cross-entropy} = -\log\left(\frac{e^{\boldsymbol{W}_i\boldsymbol{x}^T+\boldsymbol{b}_i}}{\sum_j e^{\boldsymbol{W}_j\boldsymbol{x}^T+\boldsymbol{b}_j}}\right) \quad (1)$$

where $\{\boldsymbol{W}, \boldsymbol{b}\}$ indicates the final classification layer in the network and $\boldsymbol{x}$ the feature embedding; i indicates the ground-truth subject and j runs over all the subjects.

The main disadvantage of this loss for transfer learning is that it separates training classes well, it but does not explicitly minimize the intra-class variation of each subject. In [61], Eq. (1) was augmented with a CenterLoss, such as $\frac{\lambda}{2}||\boldsymbol{x}-\boldsymbol{c}_i||_2^2$, to also reduce intra-class variance. The new loss minimizes the intra-class distance between the sample $\boldsymbol{x}$ and the centroid $\boldsymbol{c}_i$ of the class i; the centroid is updated online in the learning. Other efforts to improve cross-entropy to reduce the within-class variability are the L2-constrained SoftMax [62], in which features $\boldsymbol{x}$ are first normalized by their L2 norm to lie on a hyper-sphere and then scaled by a constant factor α. Though this approach reports remarkable performance on IJB-A, it is unclear how to select the α parameters effectively. Following the observation by [62] that the L2 norm of a deeply learned feature carries information about face image quality —the network tends to place ambiguous images close to the origin— Zheng *et al.* [63] proposed ring loss that adds regularization terms to Eq. (1) to require the network to produce feature embedding on a hypersphere. Unlike [62] they did not use a projection method such as $\frac{\boldsymbol{x}}{||\boldsymbol{x}||_2}$ to achieve this, but rather added $\frac{\lambda}{2}\left(||\boldsymbol{x}||_2 - r\right)^2$ as a second term in Eq. (1), showing improved results. This term punishes the network in case features do not lie on a hypersphere of radius r. Unlike [62], this approach jointly learns the weights of the DCNN and the parameter of the radius of the hypersphere.

While the above cited approaches aim to reduce intra-class variance, there is another trend in the community that achieves similar results by increasing the margin between classes while training: methods such as [64] and [65] proposed to increase the angular margin in the cross-entropy loss with SoftMax activations, offering a new way for optimizing such loss.

The above losses solve most of the problems for discriminative representation learning, the drawback is that they parameterize the training subjects inside the network. This can work for a training pool of subjects that range from a few to 60K, but it is hard to scale beyond this value, ending up to sacrifice the batch size. In order to train with a huge number of subjects, one possibility is to use Hierarchical SoftMax [66] or, alternatively, what can be done is to unroll all the comparisons made at once with the last inner product, outside of the loss. Deep metric learning such as contrastive loss [67], double margin contrastive loss [68] or triplet loss [17] allow the training to scale with a very large pool of subjects. Double-margin contrastive loss aims to change the feature space to pull all the positive pairs close together up to some margin, while repelling the negative up to some other margin. Formally it optimizes:

$$\ell_{double-margin} = y\ \left[d - m_p\right]_+^2 + (1-y)\ \left[m_n - d\right]_+^2, \quad (2)$$

where $y = \{0, 1\}$ corresponds to same and not-same identity, d is the Euclidean distance, and $[\cdot]_+ \doteq \max(0, \cdot)$. Note that m_n defines the margin for negative pairs. In the formulation of double-margin contrastive loss, since the $m_n - m_p \geq 0$ and that $m_p \geq 0$, then $0 \leq m_p \leq m_n$ has to hold. Although a margin-based contrastive loss has been reported to be successful enough for learning deep embeddings [69], recently the community has preferred triplet loss [17] that instead optimizes:

$$\ell_{triplet} = \left[d_{ap}^2 - d_{an}^2 + m\right]_+, \quad (3)$$

requiring only that the distance of a positive pair (anchor, positive) d_{ap} is lower than the distance of anchor and a negative d_{an}, up to a margin m. A final remark on deep metric learning losses is that one can avoid parameterizing the subjects in the network, but doing so can increase the complexity in sampling the pairs. For contrastive loss, the sampling scales quadratically; for triplet it scales cubically with respect to the training images; this is why it is important to select highly informative negative pairs [69], [70].

B. Disentanglement for Face Recognition

Another aspect to consider when training a face recognition system is that there is an alternative way to data augmentation. This alternative is disentanglement of signals in the learned feature space. While data augmentation injects on purpose other factors in the pixel space, disentanglement aims to decouple the factors in the feature space; this means that the *only* factor present in the learned representation regards identity with no other contaminations from confounding factors (pose, expressions, illumination).

Peng *et al.* [71] proposed a method for disentangling the identity signal, present in a face image, from all the other factors, such as pose and keypoint locations. Disentangling the identity from other signals means making sure that the learned representation $\boldsymbol{x}$ contains only factors that characterize the identity of an individual. Doing so, in principle, it should be impossible to predict pose or other factors from the learned embedding. Their method reaches disentanglement by branching the encoder into two parts: one develops identity features and the other is supervised with non-identity labels. The encoder is thus trained with multi-task learning. While multi-task learning can help decouple the signals, it does not ensure disentanglement, which is instead achieved with self-reconstruction and cross-reconstruction tasks. Following the

same objective, Liu *et al.* [72] used an autoencoder to distill the identity signal and dispel other factors via adversarial training. Unlike [71], this method does not need any other supervising signals but the identity. This approach shows nearly perfect accuracy on LFW and enables semantic face editing of input images, though the improvement over the baseline is marginal. Similar to [73], disentanglement is also achieved in [74] and paired with a GAN supervised with identity and pose labels; besides disentanglement, this approach is able to generate synthetic images given as input continuous pose codes.

V. Face Recognition

After the DCNN is trained following methods in Sec. IV, it can be used to extract face descriptors. A standard testing pipeline boils down to using the activations in the layer prior the classification layer as a descriptor to encode the input; these descriptors are then either L2 normalized and compared with Euclidean distance or cosine similarity (see Fig. 1).

While all the current SOTA systems loosely follow this procedure, each of them contains some unique aspects that differ from the others. We will focus on cases in which a probe sample is described by multiple, different heterogeneous media — [8], [25], [26]. This is a likely case when multiple shots of the same person can be obtained with a face tracker, a mining method or with human supervision in the loop.

Masi *et al.* [75] performed average score pooling across multiple Pose-Aware DCNNs to improve pose-invariance. The same method was improved in [41] with the early fusion of feature descriptors using the average for the different Pose-Aware DCNNs. In fact, other researchers also obtained improved performance by using some form of data aggregation technique to compress multiple information into a compact representation (called *template*) via media-averaging [35], [39], [44]; others, instead, proposed to employ an aggregation mechanism (NAN) directly inside the network [76]. Deep face descriptors can be refined with a post-processing step based on unsupervised PCA and non-linear square rooting [41] or on linear supervised embedding learning (TPE [38]).

Inspired by the one-shot similarity kernel [77], Crosswhite *et al.* [35] proposed template adaptation. Given an averaged feature of a template, a linear SVM is trained with lazy learning at test time against a collection of reference subjects (that act as a cohort of negative). This [35] showed huge improvement on top of deep features from the VGGFace network [18] on IJB-A, although it has the disadvantage that an SVM needs to be trained for every probe. Template adaptation can also be applied directly to identification. In this case, the reference cohort corresponds to the gallery set.

The community made an effort in improving preprocessing, training and verification steps in face recognition, while putting less emphasis on adapting the model for identification, either in case of a non-curated gallery ([24]) or when a gallery subject is described by a single image (one-shot learning).

VI. Overview of State-of-The-Art Results

Reviewing SOTA face recognition performance is a hard task, since multiple factors can affect a system and methods behave differently on diverse sets. Starting from 2014, face recognition made a big leap in performance through DCNNs [15] reaching saturation on LFW.

More interesting is to observe the impact of deep face recognition in the recent IJB-A set. IJB-A offers tests for 1:1 face recognition (verify) and 1:N face identification (search). It introduces probe samples that could not have a mate in the gallery, highlighting the problem of open set recognition. Evaluation metrics considered in IJB-A are a ROC (receiver operating characteristic) curve for verification, reporting the recall (True Acceptance Rate - TAR) at multiple cutoff points of the precision (False Alarm Rate - FAR). For identification, performers report the recognition rate at multiple ranks using a cumulative match characteristic (CMC) curve, counting only the probe samples that have a mate in the gallery; additionally, one can report the DET curve (detection error tradeoff), that, similar to the ROC, measures the quality of the identification—but up to a certain amount of k retrieved subjects. For a further in-depth understanding of the evaluation metrics, the reader can consult [78].

Following the above evaluation, Fig. 3 shows at a glance the progress in the IJB-A set in terms of verification (Fig. 3a) and identification (Fig. 3b) during the last four years. As can be seen from the figure, hand-crafted features used in OpenBR [79] and GOTS (government-off-the-shelf) [8], struggled to handle the tough variability present in the IJB data. The use of deep learning, starting from LSFS [80] and VGGFace [18], along with medium-sized dataset available for the community, reduced the gap in verification and identification; nevertheless it took a couple of years to reach saturation also in IJB-A. The contribution to this progress can be explained by the introduction of very deep models such as ResNet-101 along with large-sized sets [20]. Face preprocessing contributed to the progress with better methods for alignment and for data augmentation [49], [55]. The last source of improvement is provided by the use of more advanced loss functions for better transfer learning [62], [63].

VII. Conclusions and Open Issues

While performance figures on IJB-A seem reassuring—in four years, they jumped from 20% to 90% TAR at 1% FAR—there are still problems to solve. Currently, IJB-A is using templates (a mix of media defined a priori) to describe an unknown individual, but it is not clear how this "template" can be automatically generated in an accurate way: along this line, overlooked problems are video-based face recognition [81], in which a system is not given still images but it has to track a person throughout the entire video. Automatic video-processing can be an effective way to create templates automatically, yet a better synergy between multi-target tracking and recognition is needed. Furthermore, another related problem in face recognition is the automatic self-organization of a large corpus of unlabeled faces. This problem is related

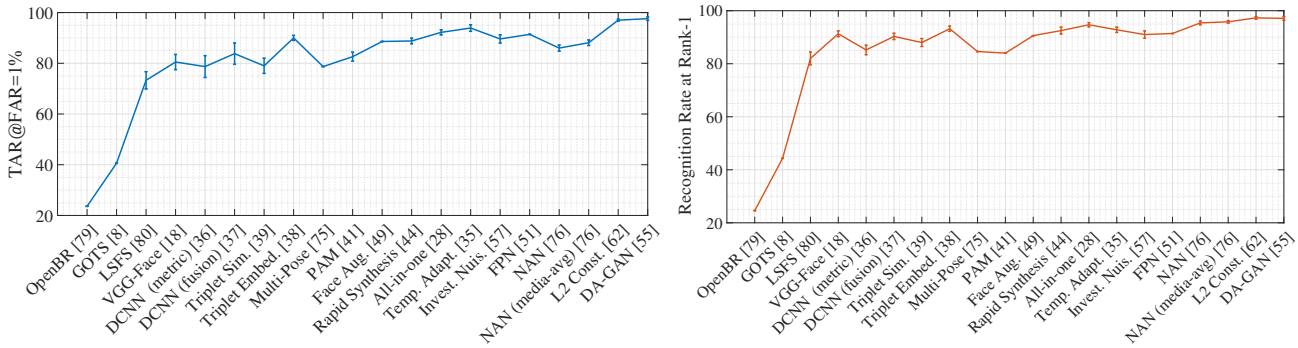

Fig. 3. **Progress on IJB-A in the last four years.** While deep face recognition methods can get near perfect result on the LFW benchmark, deep-learning along is not enough for IJB-A. This figure summarizes the methods developed during the last four years to obtain improvement on the IJB-A set. The figure shows two metrics along with their standard deviations (a) reports TAR at FAR=1% (b) presents the recognition rate at rank-1.

to clustering, and an open question is: *"If we are given an unlabeled corpus of data, such as videos and images, with the scope of clustering same identities, what is the best way to proceed? Simply train a DCNN offline and then use standard clustering methods to discover novel subjects? Or explore the myriad of unlabeled data we have?"* The two issues mentioned above (video-processing and clustering) are currently the next frontiers for face recognition and, not surprisingly, they have been recently introduced in benchmarks. IJB-B [25] and IJB-C [26] have specific novel protocols to measure the progress on these two. Other underdeveloped aspects are the adaptation of the model to a watch-list and how to tune the model in case new subjects are enrolled in the watch-list.

Nonetheless, if the modern way of automatic machine-based recognition is heavily based on large training sets, it is natural that systems trained on those will contain biases that are inherent in the data. Therefore, it is still an open problem—how to design machines to make predictions without being biased by the ethnicity, gender, age or other factors that are inherently present in the training data. Is the answer to this question just *"we need to collect more uniform and better distributed data samples of the entire world population, or maybe there is something we can do from the side of the learning algorithm?"* These kinds of questions will keep face recognition researchers busy in the following years.

ACKNOWLEDGEMENTS

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA 2014-14071600011. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purpose notwithstanding any copyright annotation thereon.

REFERENCES

[1] W. W. Bledsoe, "The model method in facial recognition," *Panoramic Research Inc., Palo Alto, CA, Rep. PR1*, vol. 15, no. 47, p. 2, 1966.

[2] W. Bledsoe, "Man-machine facial recognition: Report on a large-scale experiment, panoramic research," *Inc, Palo Alto, CA*, 1966.

[3] K. Fukushima and S. Miyake, "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition," in *Competition and Cooperation in Neural Nets*. Springer, 1982, pp. 267–285.

[4] K. Fukushima, "Neocognitron: A hierarchical neural network capable of visual pattern recognition." *Neural networks*, vol. 1, no. 2, pp. 119–130, 1988.

[5] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[7] 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," UMass, Amherst, Tech. Rep. 07-49, October 2007.

[8] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark-A," in *CVPR*, 2015.

[9] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *TPAMI*, vol. 24, no. 1, pp. 34–58, 2002.

[10] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[11] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Faceness-net: Face detection through deep facial part responses," *TPAMI*, vol. 40, no. 8, pp. 1845–1859, 2018.

[12] M. K. Yucel, Y. C. Bilge, O. Oguz, N. Ikizler-Cinbis, P. Duygulu, and R. G. Cinbis, "Wildest faces: Face detection and recognition in violent settings," *arXiv preprint arXiv:1805.07566*, 2018.

[13] G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara, "Poseidon: Face-from-depth for driver pose estimation," in *CVPR*, 2017.

[14] S. Zafeiriou, C. Zhang, and Z. Zhang, "A survey on face detection in the wild: past, present and future," *CVIU*, vol. 138, pp. 1–24, 2015.

[15] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *CVPR*, 2014.

[16] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[17] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015.

[18] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC*, 2015.

[19] A. Bansal, A. Nanduri, C. D. Castillo, R. Ranjan, and R. Chellappa, "UMDFaces: An annotated face dataset for training deep networks," in *IJCB*, 2017.

[20] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *ECCV*, 2016.

[21] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *AFGR*, 2018.

[22] F. Wang, L. Chen, C. Li, S. Huang, Y. Chen, C. Qian, and C. C. Loy, "The devil of face recognition is in the noise," in *ECCV*, 2018.

[23] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *CVPR*, 2011.

[24] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The MegaFace benchmark: 1 million faces for recognition at scale," in *CVPR*, 2016.

[25] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, J. Cheney, and P. Grother, "Iarpa janus benchmark-b face dataset," in *CVPR Workshops*, July 2017.

[26] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney *et al.*, "Iarpa janus benchmark–c: Face dataset and protocol," in *IJCB*, 2018.

[27] E. Zhou, Z. Cao, and Q. Yin, "Naive-deep face recognition: Touching the limit of LFW benchmark or not?" *arXiv preprint*, vol. arXiv:1501.04690, 2015.

[28] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," in *AFGR*, 2017.

[29] A. Nech and I. Kemelmacher-Shlizerman, "Level playing field for million scale face recognition," in *CVPR*, 2017.

[30] F. H. de Bittencourt Zavan, N. Gasparin, J. C. Batista, L. P. e Silva, V. Albiero, O. R. P. Bellon, and L. Silva, "Face analysis in the wild," in *SIBGRAPI*, 2017.

[31] Y. Wu and Q. Ji, "Facial landmark detection: A literature survey," *IJCV*, pp. 1–28, 2017.

[32] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)," in *ICCV*, 2017.

[33] Y. Wu, T. Hassner, K. Kim, G. Medioni, and P. Natarajan, "Facial landmark detection with tweaked convolutional neural networks," *TPAMI*, 2018.

[34] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *BMVC*, 2014.

[35] N. Crosswhite, J. Byrne, C. Stauffer, O. Parkhi, Q. Cao, and A. Zisserman, "Template adaptation for face verification and identification," in *AFGR*, 2017.

[36] J.-C. Chen, R. Ranjan, A. Kumar, C.-H. Chen, V. M. Patel, and R. Chellappa, "An end-to-end system for unconstrained face verification with deep convolutional neural networks," in *CVPR Workshops*, 2015.

[37] J.-C. Chen, V. M. Patel, and R. Chellappa, "Unconstrained face verification using deep cnn features," in *WACV*, 2016.

[38] S. Sankaranarayanan, A. Alavi, C. Castillo, and R. Chellappa, "Triplet probabilistic embedding for face verification and clustering," in *BTAS*, 2016.

[39] S. Sankaranarayanan, A. Alavi, and R. Chellappa, "Triplet similarity embedding for face verification," *arxiv preprint*, vol. arXiv:1602.03418, 2016.

[40] A. Bansal, C. Castillo, R. Ranjan, and R. Chellappa, "The do's and don'ts for cnn-based face verification," in *ICCV Workshops*, 2017.

[41] I. Masi, F. J. Chang, J. Choi, S. Harel, J. Kim, K. Kim, J. Leksut, S. Rawls, Y. Wu, T. Hassner, W. AbdAlmageed, G. Medioni, L. P. Morency, P. Natarajan, and R. Nevatia, "Learning pose-aware models for pose-invariant face recognition in the wild," *TPAMI*, 2018.

[42] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *CVPR*, 2015.

[43] T. Hassner, I. Masi, J. Kim, J. Choi, S. Harel, P. Natarajan, and G. Medioni, "Pooling faces: Template based face recognition with pooled face images," in *CVPR Workshops*, June 2016.

[44] I. Masi, T. Hassner, A. T. Trần, , and G. Medioni, "Rapid synthesis of massive face sets for improved face recognition," in *AFGR*, 2017.

[45] I. Masi, C. Ferrari, A. Del Bimbo, and G. Medioni, "Pose independent face recognition by localizing local binary patterns via deformation components," in *ICPR*, 2014.

[46] C. Ferrari, G. Lisanti, S. Berretti, and A. Del Bimbo, "Dictionary learning based 3D morphable model construction for face recognition with varying expression and pose," in *3DV*, 2015.

[47] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," *arXiv preprint*, vol. arXiv:1502.00873, 2015.

[48] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *CVPR*. IEEE, 2014.

[49] I. Masi, A. T. Trần, T. Hassner, J. T. Leksut, and G. Medioni, "Do we really need to collect millions of faces for effective face recognition?" in *ECCV*, 2016.

[50] D. E. Crispell, O. Biris, N. Crosswhite, J. Byrne, and J. L. Mundy, "Dataset augmentation for pose and lighting invariant face recognition," in *AIPR*, 2016.

[51] F. Chang, A. Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni, "FacePoseNet: Making a case for landmark-free face alignment," in *ICCV Workshops*, 2017.

[52] A. T. Tran, T. Hassner, I. Masi, and G. Medioni, "Regressing robust and discriminative 3d morphable models with a very deep neural network," in *CVPR*, 2017.

[53] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Li, "Face alignment across large poses: A 3d solution," in *CVPR*, 2016.

[54] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[55] J. Zhao, L. Xiong, J. Li, J. Xing, S. Yan, and J. Feng, "3d-aided dual-agent gans for unconstrained face recognition," *TPAMI*, 2018.

[56] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *NIPS*, 2015.

[57] C. Ferrari, G. Lisanti, S. Berretti, and A. D. Bimbo, "Investigating nuisances in DCNN-based face recognition," *TIP*, 2018.

[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[59] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *IJCV*, pp. 1–42, 2014.

[60] C. L. X. T. Kaidi Cao, Yu Rong and C. C. Loy, "Pose-robust face recognition via deep residual equivariant mapping," in *CVPR*, 2018.

[61] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *ECCV*, 2016.

[62] R. Ranjan, C. D. Castillo, and R. Chellappa, "L2-constrained softmax loss for discriminative face verification," *arXiv preprint arXiv:1703.09507*, 2017.

[63] Y. Zheng, D. K. Pal, and M. Savvides, "Ring loss: Convex feature normalization for face recognition," in *CVPR*, 2018.

[64] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks." in *ICML*, 2016.

[65] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *CVPR*, vol. 1, 2017.

[66] A. Mnih and G. E. Hinton, "A scalable hierarchical distributed language model," in *NIPS*, 2009.

[67] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *CVPR*, 2006.

[68] J. Lin, O. Morere, A. Veillard, L.-Y. Duan, H. Goh, and V. Chandrasekhar, "Deephash for image instance retrieval: Getting regularization, depth and fine-tuning right," in *ACM ICMR*, 2017.

[69] C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krahenbuhl, "Sampling matters in deep embedding learning," in *ICCV*, Oct 2017.

[70] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *CVPR*, 2016.

[71] X. Peng, X. Yu, K. Sohn, D. N. Metaxas, and M. Chandraker, "Reconstruction-based disentanglement for pose-invariant face recognition," in *ICCV*, 2017.

[72] Y. Liu, F. Wei, J. Shao, L. Sheng, J. Yan, and X. Wang, "Exploring disentangled feature representation beyond face identification," in *CVPR*, 2018.

[73] J. T. Springenberg, "Unsupervised and semi-supervised learning with categorical generative adversarial networks," in *ICLR*, 2016.

[74] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *CVPR*, 2017.

[75] I. Masi, S. Rawls, G. Medioni, and P. Natarajan, "Pose-Aware Face Recognition in the Wild," in *CVPR*, 2016.

[76] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua, "Neural aggregation network for video face recognition," in *CVPR*, July 2017.

[77] L. Wolf, T. Hassner, and Y. Taigman, "The one-shot similarity kernel," in *CVPR*, 2009.

[78] P. J. Phillips, P. Grother, and R. Micheals, "Evaluation methods in face recognition," in *Handbook of Face Recognition*. Springer, 2011, pp. 551–574.

[79] J. Klontz, B. Klare, S. Klum, E. Taborsky, M. Burge, and A. K. Jain, "Open source biometric recognition," in *BTAS*, 2013.

[80] D. Wang, C. Otto, and A. K. Jain, "Face search at scale: 80 million gallery," *TPAMI*, vol. 39, no. 6, pp. 1122–1136, June 2017.

[81] K. Kim, Z. Yang, I. Masi, R. Nevatia, and G. Medioni, "Face and body association for video-based face recognition," in *WACV*, 2018.