Neurocomputing 429 (2021) 215–244

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Deep face recognition: A survey

Mei Wang [a], Weihong Deng [a,*]

[a] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China

ARTICLE INFO

Article history:
Received 10 May 2020
Revised 1 August 2020
Accepted 25 October 2020
Available online 10 November 2020
Communicated by Zidong Wang

Keywords:
Deep face recognition
Deep learning
Face processing
Face recognition database
Loss function
Deep network architecture

ABSTRACT

Deep learning applies multiple processing layers to learn representations of data with multiple levels of feature extraction. This emerging technique has reshaped the research landscape of face recognition (FR) since 2014, launched by the breakthroughs of DeepFace and DeepID. Since then, deep learning technique, characterized by the hierarchical architecture to stitch together pixels into invariant face representation, has dramatically improved the state-of-the-art performance and fostered successful real-world applications. In this survey, we provide a comprehensive review of the recent developments on deep FR, covering broad topics on algorithm designs, databases, protocols, and application scenes. First, we summarize different network architectures and loss functions proposed in the rapid evolution of the deep FR methods. Second, the related face processing methods are categorized into two classes: "one-to-many augmentation" and "many-to-one normalization". Then, we summarize and compare the commonly used databases for both model training and evaluation. Third, we review miscellaneous scenes in deep FR, such as cross-factor, heterogenous, multiple-media and industrial scenes. Finally, the technical challenges and several promising directions are highlighted.

1. Introduction

Face recognition (FR) has been the prominent biometric technique for identity authentication and has been widely used in many areas, such as military, finance, public security and daily life. FR has been a long-standing research topic in the CVPR community. In the early 1990s, the study of FR became popular following the introduction of the historical Eigenface approach [1]. The milestones of feature-based FR over the past years are presented in Fig. 1, in which the times of four major technical streams are highlighted. The holistic approaches derive the low-dimensional representation through certain distribution assumptions, such as linear subspace [2–4], manifold [5–7], and sparse representation [8–11]. This idea dominated the FR community in the 1990s and 2000s. However, a well-known problem is that these theoretically plausible holistic methods fail to address the uncontrolled facial changes that deviate from their prior assumptions. In the early 2000s, this problem gave rise to local-feature-based FR. Gabor [12] and LBP [13], as well as their multilevel and high-dimensional extensions [14–16], achieved robust performance through some invariant properties of local filtering. Unfortunately, handcrafted features suffered from a lack of distinctiveness and compactness. In the early 2010s, learning-based local descriptors were introduced to the FR community [17–19], in which local filters are learned for better distinctiveness and the encoding codebook is learned for better compactness. However, these shallow representations still have an inevitable limitation on robustness against the complex nonlinear facial appearance variations.

In general, traditional methods attempted to recognize human face by one or two layer representations, such as filtering responses, histogram of the feature codes, or distribution of the dictionary atoms. The research community studied intensively to separately improve the preprocessing, local descriptors, and feature transformation, but these approaches improved FR accuracy slowly. What's worse, most methods aimed to address one aspect of unconstrained facial changes only, such as lighting, pose, expression, or disguise. There was no any integrated technique to address these unconstrained challenges integrally. As a result, with continuous efforts of more than a decade, "shallow" methods only improved the accuracy of the LFW benchmark to about 95% [15], which indicates that "shallow" methods are insufficient to extract stable identity feature invariant to real-world changes. Due to the insufficiency of this technical, facial recognition systems were often reported with unstable performance or failures with countless false alarms in real-world applications.

But all that changed in 2012 when AlexNet won the ImageNet competition by a large margin using a technique called deep learning [22]. Deep learning methods, such as convolutional neural networks, use a cascade of multiple layers of processing units for

* Corresponding author.
E-mail address: whdeng@bupt.edu.cn (W. Deng).

https://doi.org/10.1016/j.neucom.2020.10.081
0925-2312/

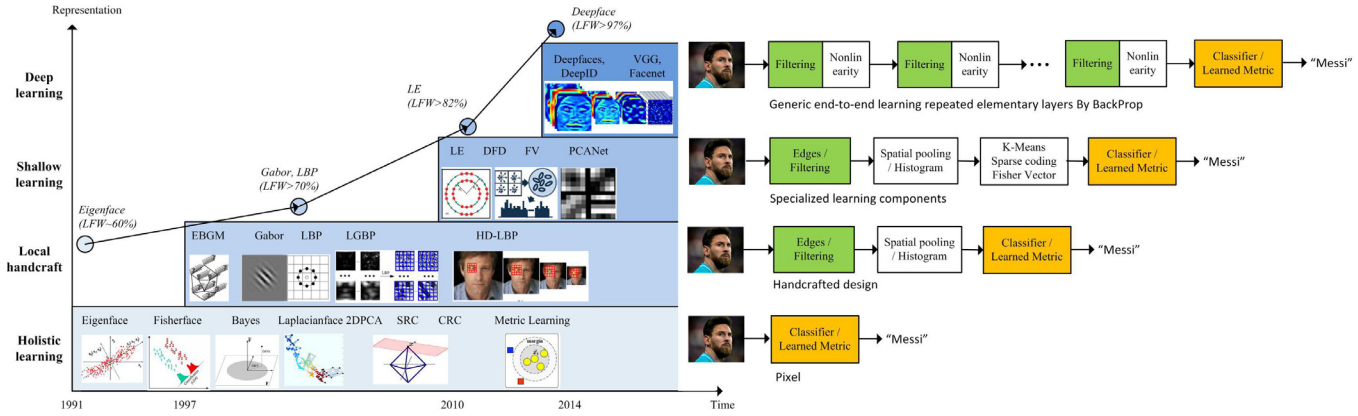

Fig. 1. Milestones of face representation for recognition. The holistic approaches dominated the face recognition community in the 1990s. In the early 2000s, handcrafted local descriptors became popular, and the local feature learning approaches were introduced in the late 2000s. In 2014, DeepFace [20] and DeepID [21] achieved a breakthrough on state-of-the-art (SOTA) performance, and research focus has shifted to deep-learning-based approaches. As the representation pipeline becomes deeper and deeper, the LFW (Labeled Face in-the-Wild) performance steadily improves from around 60% to above 90%, while deep learning boosts the performance to 99.80% in just three years.

feature extraction and transformation. They learn multiple levels of representations that correspond to different levels of abstraction. The levels form a hierarchy of concepts, showing strong invariance to the face pose, lighting, and expression changes, as shown in Fig. 2. It can be seen from the figure that the first layer of the deep neural network is somewhat similar to the Gabor feature found by human scientists with years of experience. The second layer learns more complex texture features. The features of the third layer are more complex, and some simple structures have begun to appear, such as high-bridged nose and big eyes. In the fourth, the network output is enough to explain a certain facial attribute, which can make a special response to some clear abstract concepts such as smile, roar, and even blue eye. In conclusion, in deep convolutional neural networks (CNN), the lower layers automatically learn the features similar to Gabor and SIFT designed for years or even decades (such as initial layers in Fig. 2), and the higher layers further learn higher level abstraction. Finally, the combination of these higher level abstraction represents facial identity with unprecedented stability.

In 2014, DeepFace [20] achieved the SOTA accuracy on the famous LFW benchmark [23], approaching human performance on the unconstrained condition for the first time (DeepFace: 97.35% vs. Human: 97.53%), by training a 9-layer model on 4 million facial images. Inspired by this work, research focus has shifted to deep-learning-based approaches, and the accuracy was dramatically boosted to above 99.80% in just three years. Deep learning technique has reshaped the research landscape of FR in almost all aspects such as algorithm designs, training/test datasets, application scenarios and even the evaluation protocols. Therefore, it is of great significance to review the breakthrough and rapid development process in recent years. There have been several surveys on FR [24–28] and its subdomains, and they mostly summarized and compared a diverse set of techniques related to a specific FR scene, such as illumination-invariant FR [29], 3D FR [28], pose-invariant FR [30,31]. Unfortunately, due to their earlier publication dates, none of them covered the deep learning methodology that is most successful nowadays. This survey focuses only on recognition problem, and one can refer to Ranjan et al. [32] for a brief review of a full deep FR pipeline with detection and alignment, or refer to Jin et al. [33] for a survey of face alignment. Specifically, the major contributions of this survey are as follows:

- A systematic review on the evolution of the network architectures and loss functions for deep FR is provided. Various loss functions are categorized into Euclidean-distance-based loss, angular/cosine-margin-based loss and softmax loss and its variations. Both the mainstream network architectures, such as Deepface [20], DeepID series [34,35,21,36], VGGFace [37], FaceNet [38], and VGGFace2 [39], and other architectures designed for FR are covered.

- We categorize the new face processing methods based on deep learning, such as those used to handle recognition difficulty on pose changes, into two classes: "one-to-many augmentation" and "many-to-one normalization", and discuss how emerging generative adversarial network (GAN) [40] facilitates deep FR.

- We present a comparison and analysis on public available databases that are of vital importance for both model training and testing. Major FR benchmarks, such as LFW [23], IJB-A/B/C [41–43], Megaface [44], and MS-Celeb-1 M [45], are reviewed and compared, in term of the four aspects: training methodology, evaluation tasks and metrics, and recognition scenes, which provides an useful reference for training and testing deep FR.

- Besides the *general purpose* tasks defined by the major databases, we summarize a dozen *scenario-specific* databases and solutions that are still challenging for deep learning, such as anti-attack, cross-pose FR, and cross-age FR. By reviewing specially designed methods for these unsolved problems, we attempt to reveal the important issues for future research on deep FR, such as adversarial samples, algorithm/data biases, and model interpretability.

The remainder of this survey is structured as follows. In Section 2, we introduce some background concepts and terminologies, and then we briefly introduce each component of FR. In Section 3, different network architectures and loss functions are presented. Then, we summarize the face processing algorithms and the datasets. In Section 5, we briefly introduce several methods of deep FR used for different scenes. Finally, the conclusion of this paper and discussion of future works are presented in Section 6.

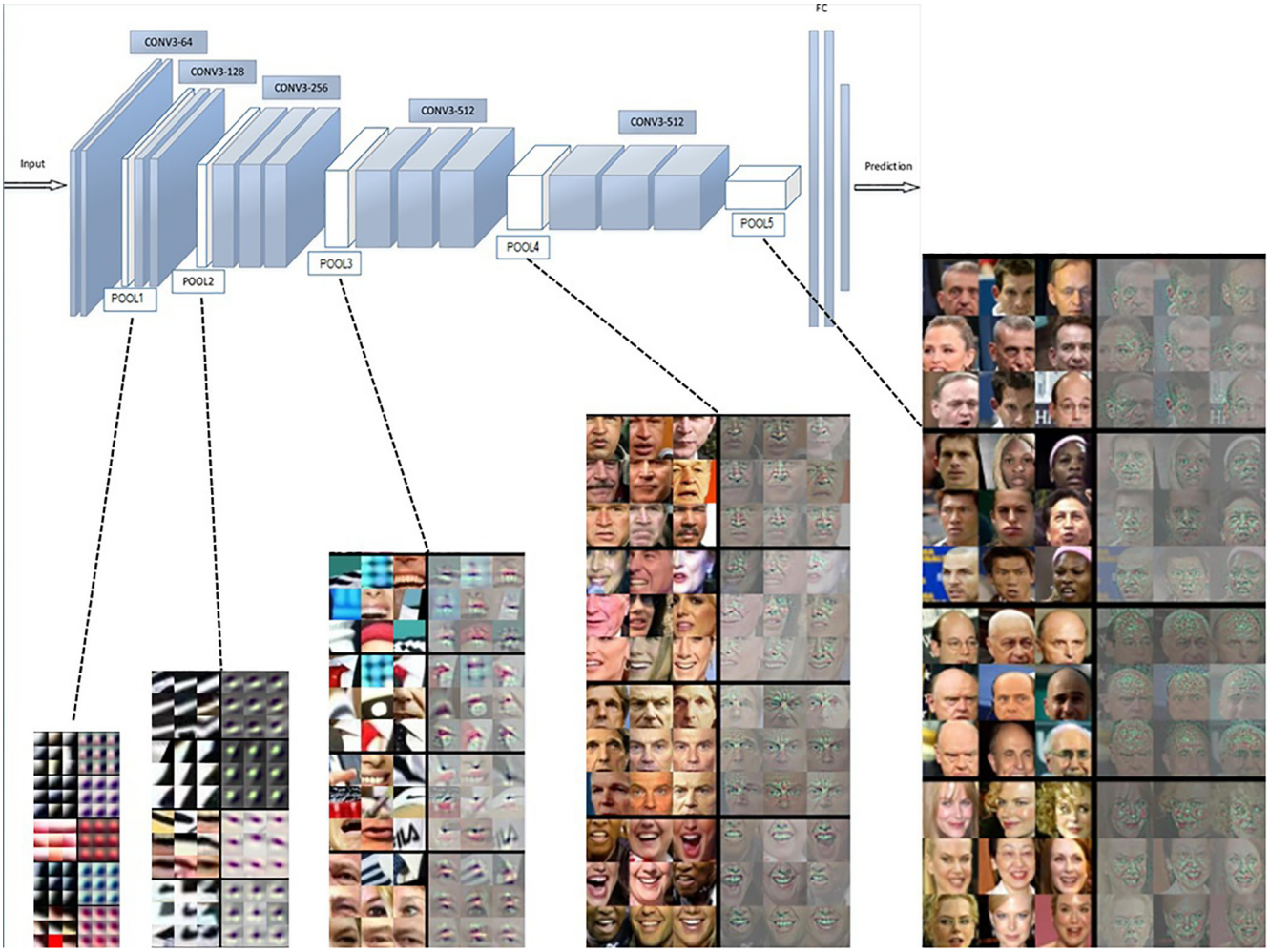

Fig. 2. The hierarchical architecture that stitches together pixels into invariant face representation. Deep model consists of multiple layers of simulated neurons that convolute and pool input, during which the receptive-field size of simulated neurons are continually enlarged to integrate the low-level primary elements into multifarious facial attributes, finally feeding the data forward to one or more fully connected layer at the top of the network. The output is a compressed feature vector that represent the face. Such deep representation is widely considered as the SOTA technique for face recognition.

2. Overview

2.1. Components of Face Recognition

As mentioned in [32], there are three modules needed for FR system, as shown in Fig. 3. First, a face detector is used to localize faces in images or videos. Second, with the facial landmark detector, the faces are aligned to normalized canonical coordinates. Third, the FR module is implemented with these aligned face images. We only focus on the FR module throughout the remainder of this paper.

Before a face image is fed to an FR module, face anti-spoofing, which recognizes whether the face is live or spoofed, is applied to avoid different types of attacks. Then, recognition can be performed. As shown in Fig. 3(c), an FR module consists of face processing, deep feature extraction and face matching, and it can be described as follows:

$$M\left[F(P_i(I_i)), F(P_j(I_j))\right] \tag{1}$$

where I_i and I_j are two face images, respectively. P stands for face processing to handle intra-personal variations before training and testing, such as poses, illuminations, expressions and occlusions. F denotes feature extraction, which encodes the identity information. The feature extractor is learned by loss functions when training, and is utilized to extract features of faces when testing. M means a face matching algorithm used to compute similarity scores of features to determine the specific identity of faces. Different from object classification, the testing identities are usually disjoint from the training data in FR, which makes the learned classifier cannot be used to recognize testing faces. Therefore, face matching algorithm is an essential part in FR.

2.1.1. Face processing

Although deep-learning-based approaches have been widely used, Mehdipour et al. [46] proved that various conditions, such as poses, illuminations, expressions and occlusions, still affect the performance of deep FR. Accordingly, face processing is introduced to address this problem. The face processing methods are categorized as "one-to-many augmentation" and "many-to-one normalization", as shown in Table 1.

- "One-to-many augmentation". These methods generate many patches or images of the pose variability from a single image to enable deep networks to learn pose-invariant representations.

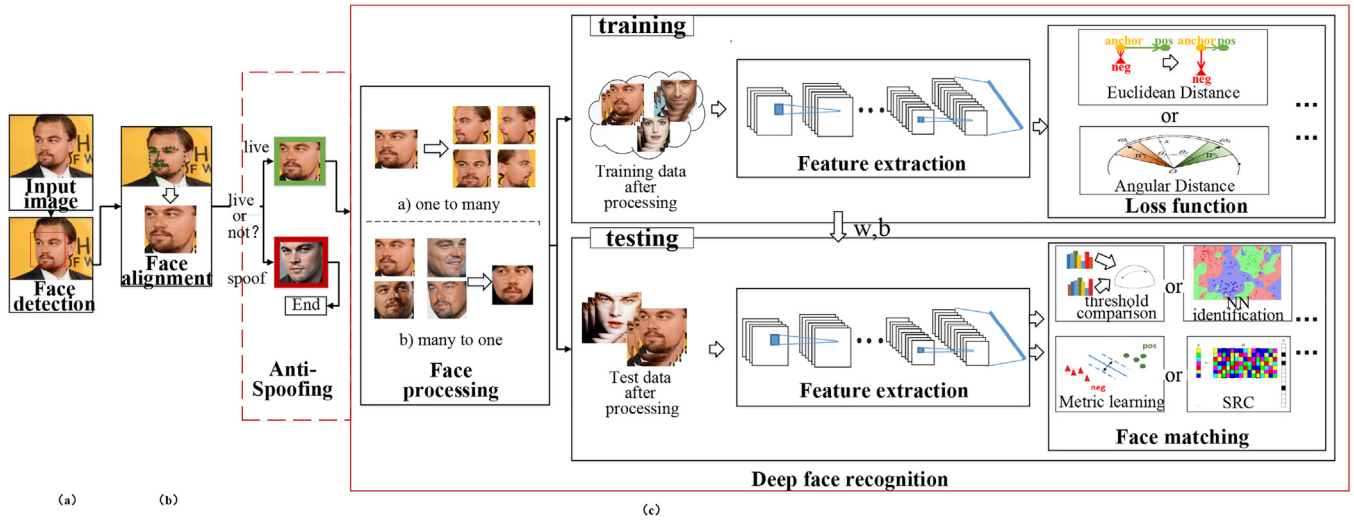

Fig. 3. Deep FR system with face detector and alignment. First, a face detector is used to localize faces. Second, the faces are aligned to normalized canonical coordinates. Third, the FR module is implemented. In FR module, face anti-spoofing recognizes whether the face is live or spoofed; face processing is used to handle variations before training and testing, e.g. poses, ages; different architectures and loss functions are used to extract discriminative deep feature when training; face matching methods are used to do feature classification after the deep features of testing data are extracted.

Table 1
Different data preprocessing approaches.

Data processing	Brief Description	Subsettings
one to many	These methods generate many patches or images of the pose variability from a single image	3D model [47–54] 2D deep model [55–57] data augmentation [58–60,35,21,36,61,62]
many to one	These methods recover the canonical view of face images from one or many images of nonfrontal view	Antoencoder [63–67] CNN [68,69] GAN [70–73]

Table 2
Different network architectures of FR.

Network Architectures	Subsettings
backbone network	mainstream architectures: AlexNet [80,81,38], VGGNet [37,47,82], GoogleNet [83,38], ResNet [84,82], SENet [39] light-weight architectures [85,86,61,87] adaptive architectures [88–90] joint alignment-recognition architectures [91–94]
assembled networks	multipose [95–98], multipatch [58–60,99,34,21,35], multitask [100]

- "Many-to-one normalization". These methods recover the canonical view of face images from one or many images of a nonfrontal view; then, FR can be performed as if it were under controlled conditions.

Note that we mainly focus on deep face processing method designed for pose variations in this paper, since pose is widely regarded as a major challenge in automatic FR applications and other variations can be solved by the similar methods.

2.1.2. Deep feature extraction

Network Architecture. The architectures can be categorized as backbone and assembled networks, as shown in Table 2. Inspired by the extraordinary success on the ImageNet [74] challenge, the typical CNN architectures, e.g. AlexNet, VGGNet, GoogleNet, ResNet and SENet [22,75–78], are introduced and widely used as the baseline models in FR (directly or slightly modified). In addition to the mainstream, some assembled networks, e.g. multi-task networks and multi-input networks, are utilized in FR. Hu et al. [79] shows that accumulating the results of assembled networks provides an increase in performance compared with an individual network.

Loss Function. The softmax loss is commonly used as the supervision signal in object recognition, and it encourages the separability of features. However, the softmax loss is not sufficiently effective for FR because intra-variations could be larger than inter-differences and more discriminative features are required when recognizing different people. Many works focus on creating novel loss functions to make features not only more separable but also discriminative, as shown in Table 3.

2.1.3. Face matching by deep features

FR can be categorized as face verification and face identification. In either scenario, a set of known subjects is initially enrolled in the system (the gallery), and during testing, a new subject (the probe) is presented. After the deep networks are trained on massive data with the supervision of an appropriate loss function, each of the test images is passed through the networks to obtain a deep feature representation. Using cosine distance or L2 distance, face verification computes one-to-one similarity between the gallery and probe to determine whether the two images are of the same subject, whereas face identification computes one-to-many similarity to determine the specific identity of a probe face. In addition to these, other methods are introduced to postprocess the deep features such that the face matching is performed efficiently and

Table 3
Different loss functions for FR.

Loss Functions	Brief Description
Euclidean-distance-based loss	These methods reduce intra-variance and enlarge inter-variance based on Euclidean distance. [21,35,36,101,102,82,38,37,80,81,58,103]
angular/cosine-margin-based loss	These methods make learned features potentially separable with larger angular/cosine distance. [104,84,105–108]
softmax loss and its variations	These methods modify the softmax loss to improve performance, e.g. features or weights normalization. [109–115]

accurately, such as metric learning, sparse-representation-based classifier (SRC), and so forth.

To sum up, we present FR modules and their commonly-used methods in Fig. 4 to help readers to get a view of the whole FR. In deep FR, various training and testing face databases are constructed, and different architectures and losses of deep FR always follow those of deep object classification and are modified according to unique characteristics of FR. Moreover, in order to address unconstrained facial changes, face processing methods are further designed to handle poses, expressions and occlusions variations. Benefiting from these strategies, deep FR system significantly improves the SOTA and surpasses human performance. When the applications of FR becomes more and more mature in general scenario, recently, different solutions are driven for more difficult specific scenarios, such as cross-pose FR, cross-age FR, video FR.

3. Network architecture and training loss

For most applications, it is difficult to include the candidate faces during the training stage, which makes FR become a "zero-shot" learning task. Fortunately, since all human faces share a similar shape and texture, the representation learned from a small proportion of faces can generalize well to the rest. Based on this theory, a straightforward way to improve generalized performance is to include as many IDs as possible in the training set. For example, Internet giants such as Facebook and Google have reported their deep FR system trained by $10^6 - 10^7$ IDs [38,20].

Unfortunately, these personal datasets, as well as prerequisite GPU clusters for distributed model training, are not accessible for academic community. Currently, public available training databases for academic research consist of only $10^3 - 10^5$ IDs. Instead, academic community makes effort to design effective loss functions and adopts efficient architectures to make deep features more discriminative using the relatively small training data sets. For instance, the accuracy of most popular LFW benchmark has been boosted from 97% to above 99.8% in the pasting four years, as enumerated in Table 4. In this section, we survey the research efforts on different loss functions and network architectures that have significantly improved deep FR methods.

3.1. Evolution of discriminative loss functions

Inheriting from the object classification network such as AlexNet, the initial Deepface [20] and DeepID [34] adopted cross-entropy based softmax loss for feature learning. After that, people realized that the softmax loss is not sufficient by itself to learn discriminative features, and more researchers began to explore novel loss functions for enhanced generalization ability. This becomes the hottest research topic in deep FR research, as illustrated in Fig. 5. Before 2017, Euclidean-distance-based loss played an important role; In 2017, angular/cosine-margin-based loss as well as feature and weight normalization became popular. It should be noted that, although some loss functions share the similar basic idea, the new one is usually designed to facilitate the training procedure by easier parameter or sample selection.

3.1.1. Euclidean-distance-based loss

Euclidean-distance-based loss is a metric learning method [118,119] that embeds images into Euclidean space in which intra-variance is reduced and inter-variance is enlarged. The contrastive loss and the triplet loss are the commonly used loss functions. The contrastive loss [35,21,36,61,120] requires face image pairs, and then pulls together positive pairs and pushes apart negative pairs.

$$\mathcal{L} = y_{ij} max\left(0, \|f(x_i) - f(x_j)\|_2 - \epsilon^+\right) + (1 - y_{ij}) max\left(0, \epsilon^- - \|f(x_i) - f(x_j)\|_2\right) \quad (2)$$

where $y_{ij} = 1$ means x_i and x_j are matching samples and $y_{ij} = 0$ means non-matching samples. $f(\cdot)$ is the feature embedding, ϵ^+ and ϵ^- control the margins of the matching and non-matching pairs respectively. DeepID2 [21] combined the face identification (softmax) and verification (contrastive loss) supervisory signals to learn a discriminative representation, and joint Bayesian (JB) was applied to obtain a robust embedding space. Extending from DeepID2 [21],

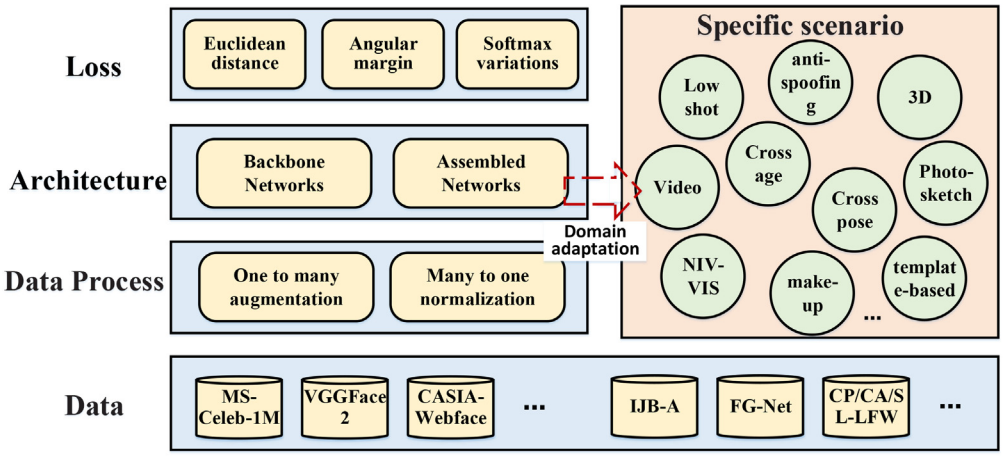

Fig. 4. FR studies have begun with general scenario, then gradually get close to more realistic applications and drive different solutions for specific scenarios, such as cross-pose FR, cross-age FR, video FR. In specific scenarios, targeted training and testing database are constructed, and face processing, architectures and loss functions are modified based on the special requirements.

Table 4
The accuracy of different methods evaluated on the LFW dataset.

Method	Public. Time	Loss	Architecture	Number of Networks	Training Set	Accuracy ± Std (%)
DeepFace [20]	2014	softmax	Alexnet	3	Facebook (4.4 M,4 K)	97.35 ± 0.25
DeepID2 [21]	2014	contrastive loss	Alexnet	25	CelebFaces+ (0.2 M,10 K)	99.15 ± 0.13
DeepID3 [36]	2015	contrastive loss	VGGNet-10	50	CelebFaces+ (0.2 M,10 K)	99.53 ± 0.10
FaceNet [38]	2015	triplet loss	GoogleNet-24	1	Google (500 M,10 M)	99.63 ± 0.09
Baidu [58]	2015	triplet loss	CNN-9	10	Baidu (1.2 M,18 K)	99.77
VGGface [37]	2015	triplet loss	VGGNet-16	1	VGGface (2.6 M,2.6 K)	98.95
light-CNN [85]	2015	softmax	light CNN	1	MS-Celeb-1 M (8.4 M,100 K)	98.8
Center Loss [101]	2016	center loss	Lenet+-7	1	CASIA-WebFace, CACD2000, Celebrity+ (0.7 M,17 K)	99.28
L-softmax [104]	2016	L-softmax	VGGNet-18	1	CASIA-WebFace (0.49 M,10 K)	98.71
Range Loss [82]	2016	range loss	VGGNet-16	1	MS-Celeb-1 M, CASIA-WebFace (5 M,100 K)	99.52
L2-softmax [109]	2017	L2-softmax	ResNet-101	1	MS-Celeb-1 M (3.7 M,58 K)	99.78
Normface [110]	2017	contrastive loss	ResNet-28	1	CASIA-WebFace (0.49 M,10 K)	99.19
CoCo loss [112]	2017	CoCo loss	-	1	MS-Celeb-1 M (3 M,80 K)	99.86
vMF loss [115]	2017	vMF loss	ResNet-27	1	MS-Celeb-1 M (4.6 M,60 K)	99.58
Marginal Loss [116]	2017	marginal loss	ResNet-27	1	MS-Celeb-1 M (4 M,80 K)	99.48
SphereFace [84]	2017	A-softmax	ResNet-64	1	CASIA-WebFace (0.49 M,10 K)	99.42
CCL [113]	2018	center invariant loss	ResNet-27	1	CASIA-WebFace (0.49 M,10 K)	99.12
AMS loss [105]	2018	AMS loss	ResNet-20	1	CASIA-WebFace (0.49 M,10 K)	99.12
Cosface [107]	2018	cosface	ResNet-64	1	CASIA-WebFace (0.49 M,10 K)	99.33
Arcface [106]	2018	arcface	ResNet-100	1	MS-Celeb-1 M (3.8 M,85 K)	99.83
Ring loss [117]	2018	Ring loss	ResNet-64	1	MS-Celeb-1 M (3.5 M,31 K)	99.50

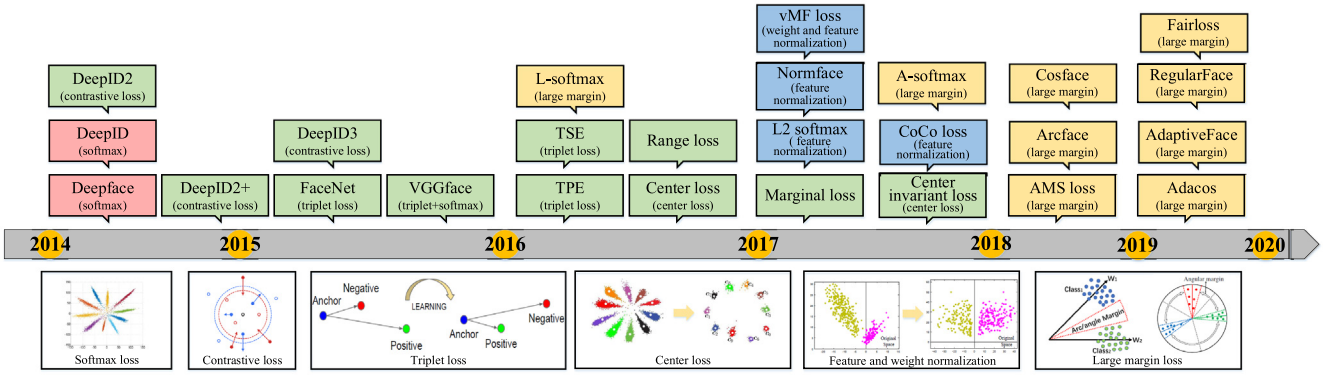

Fig. 5. The development of loss functions. It marks the beginning of deep FR that Deepface [20] and DeepID [34] were introduced in 2014. After that, Euclidean-distance-based loss always played the important role in loss function, such as contractive loss, triplet loss and center loss. In 2016 and 2017, L-softmax [104] and A-softmax [84] further promoted the development of the large-margin feature learning. In 2017, feature and weight normalization also begun to show excellent performance, which leads to the study on variations of softmax. Red, green, blue and yellow rectangles represent deep methods using softmax, Euclidean-distance-based loss, angular/cosine-margin-based loss and variations of softmax, respectively.

DeepID2+ [35] increased the dimension of hidden representations and added supervision to early convolutional layers. DeepID3 [36] further introduced VGGNet and GoogleNet to their work. However, the main problem with the contrastive loss is that the margin parameters are often difficult to choose.

Contrary to contrastive loss that considers the absolute distances of the matching pairs and non-matching pairs, triplet loss considers the relative difference of the distances between them. Along with FaceNet [38] proposed by Google, Triplet loss [38,37,81,80,58,60] was introduced into FR. It requires the face triplets, and then it minimizes the distance between an anchor and a positive sample of the same identity and maximizes the distance between the anchor and a negative sample of a different identity. FaceNet made $\left\|f(x_i^a)-f(x_i^p)\right\|_2^2+\alpha<-\left\|f(x_i^a)-f(x_i^n)\right\|_2^2$using hard triplet face samples, where x_i^a, x_i^p and x_i^n are the anchor, positive and negative samples, respectively, α is a margin and $f(\cdot)$ represents a nonlinear transformation embedding an image into a feature space. Inspired by FaceNet [38], TPE [81] and TSE [80] learned a linear projection W to construct triplet loss. Other methods optimize deep models using both triplet loss and softmax loss [59,58,60,121]. They first train networks with softmax and then fine-tune them with triplet loss.

However, the contrastive loss and triplet loss occasionally encounter training instability due to the selection of effective training samples, some paper begun to explore simple alternatives. Center loss [101] and its variants [82,116,102] are good choices for reducing intra-variance. The center loss [101] learned a center for each class and penalized the distances between the deep features and their corresponding class centers. This loss can be defined as follows:

$$\mathcal{L}_C=\frac{1}{2}\sum_{i=1}^{m}\left\|x_i-c_{y_i}\right\|_2^2 \tag{3}$$

where x_i denotes the i-th deep feature belonging to the y_i-th class and c_{y_i} denotes the y_i-th class center of deep features. To handle the long-tailed data, a range loss [82], which is a variant of center

loss, is used to minimize k greatest range's harmonic mean values in one class and maximize the shortest inter-class distance within one batch. Wu et al. [102] proposed a center-invariant loss that penalizes the difference between each center of classes. Deng et al. [116] selected the farthest intra-class samples and the nearest inter-class samples to compute a margin loss. However, the center loss and its variants suffer from massive GPU memory consumption on the classification layer, and prefer balanced and sufficient training data for each identity.

3.1.2. Angular/cosine-margin-based loss

In 2017, people had a deeper understanding of loss function in deep FR and thought that samples should be separated more strictly to avoid misclassifying the difficult samples. Angular/cosine-margin-based loss [104,84,105,106,108] is proposed to make learned features potentially separable with a larger angular/-cosine distance. The decision boundary in softmax loss is $(W_1 - W_2)x + b_1 - b_2 = 0$, where x is feature vector, W_i and b_i are weights and bias in softmax loss, respectively. Liu et al. [104] reformulated the original softmax loss into a large-margin softmax (L-Softmax) loss. They constrain $b_1 = b_2 = 0$, so the decision boundaries for class 1 and class 2 become $\|x\|(\|W_1\|cos(m\theta_1) - \|W_2\|cos(\theta_2)) = 0$ and $\|x\|(\|W_1\|\|W_2\|cos(\theta_1) - cos(m\theta_2)) = 0$, respectively, where m is a positive integer introducing an angular margin, and θ_i is the angle between W_i and x. Due to the non-monotonicity of the cosine function, a piece-wise function is applied in L-softmax to guarantee the monotonicity. The loss function is defined as follows:

$$\mathcal{L}_i = -log\left(\frac{e^{\|W_{yi}\|\|x_i\|\varphi(\theta_{yi})}}{\|W_{yi}\|\|x_i\|\varphi(\theta_{yi}) + \sum_{j\neq y_i} e^{\|W_{yi}\|\|x_i\|cos(\theta_j)}}\right) \tag{4}$$

where

$$\varphi(\theta) = (-1)^k cos(m\theta) - 2k, \theta \in \left[\frac{k\pi}{m}, \frac{(k+1)\pi}{m}\right] \tag{5}$$

Considering that L-Softmax is difficult to converge, it is always combined with softmax loss to facilitate and ensure the convergence. Therefore, the loss function is changed into: $f_{y_i} = \frac{\lambda\|W_{y_i}\|\|x_i\|cos(\theta_{y_i}) + \|W_{y_i}\|\|x_i\|\varphi(\theta_{y_i})}{1+\lambda}$, where λ is a dynamic hyper-parameter. Based on L-Softmax, A-Softmax loss [84] further normalized the weight W by L2 norm ($\|W\| = 1$) such that the normalized vector will lie on a hypersphere, and then the discriminative face features can be learned on a hypersphere manifold with an angular margin (Fig. 6). Liu et al. [108] introduced a deep hyperspherical convolution network (SphereNet) that adopts hyperspherical convolution as its basic convolution operator and is supervised by angular-margin-based loss. To overcome the optimization difficulty of L-Softmax and A-Softmax, which incorporate the angular margin in a multiplicative manner, ArcFace [106] and CosFace [105], AMS loss [107] respectively introduced an additive angular/cosine margin $cos(\theta + m)$ and $cos\theta - m$. They are extremely easy to implement without tricky hyper-parameters λ, and are more clear and able to converge without the softmax supervision. The decision boundaries under the binary classification case are given in Table 5. Based on large margin, FairLoss [122] and AdaptiveFace [123] further proposed to adjust the margins for different classes adaptively to address the problem of unbalanced data. Compared to Euclidean-distance-based loss, angular/cosine-margin-based loss explicitly adds discriminative constraints on a hypershpere manifold, which intrinsically matches the prior that human face lies on a manifold. However, Wang et al. [124] showed that angular/cosine-margin-based loss can achieve better results on a clean dataset, but is vulnerable to noise and becomes worse than center loss and softmax in the high-noise region as shown in Fig. 7.

3.1.3. Softmax loss and its variations

In 2017, in addition to reformulating softmax loss into an angular/cosine-margin-based loss as mentioned above, some works tries to normalize the features and weights in loss functions to improve the model performance, which can be written as follows:

$$\widehat{W} = \frac{W}{\|W\|}, \hat{x} = \alpha\frac{x}{\|x\|} \tag{6}$$

where α is a scaling parameter, x is the learned feature vector, W is weight of last fully connected layer. Scaling x to a fixed radius α is important, as Wang et al. [110] proved that normalizing both features and weights to 1 will make the softmax loss become trapped at a very high value on the training set. After that, the loss function, e.g. softmax, can be performed using the normalized features and weights.

Some papers [84,108] first normalized the weights only and then added angular/cosine margin into loss functions to make the learned features be discriminative. In contrast, some works, such as [109,111], adopted feature normalization only to overcome the bias to the sample distribution of the softmax. Based on the observation of [125] that the L2-norm of features learned using the softmax loss is informative of the quality of the face, L2-softmax [109] enforced all the features to have the same L2-norm by feature normalization such that similar attention is given to good quality frontal faces and blurry faces with extreme pose. Rather than scaling x to the parameter α, Hasnat et al. [111] normalized features with $\hat{x} = \frac{x-\mu}{\sqrt{\sigma^2}}$, where μ and σ^2 are the mean and variance. Ring loss [117] encouraged the norm of samples being value R (a learned parameter) rather than explicit enforcing through a hard normalization operation. Moreover, normalizing both features and weights [110,112,115,105,106] has become a common strategy. Wang et al. [110] explained the necessity of this normalization operation from both analytic and geometric perspectives. After normalizing features and weights, CoCo loss [112] optimized the cosine distance among data features, and Hasnat et al. [115] used the von Mises-Fisher (vMF) mixture model as the theoretical basis to develop a novel vMF mixture loss and its corresponding vMF deep features.

3.2. Evolution of network architecture

3.2.1. Backbone network

Mainstream architectures. The commonly used network architectures of deep FR have always followed those of deep object classification and evolved from AlexNet to SENet rapidly. We present the most influential architectures of deep object classification and deep face recognition in chronological order [1] in Fig. 8.

In 2012, AlexNet [22] was reported to achieve the SOTA recognition accuracy in the ImageNet large-scale visual recognition competition (ILSVRC) 2012, exceeding the previous best results by a large margin. AlexNet consists of five convolutional layers and three fully connected layers, and it also integrates various techniques, such as rectified linear unit (ReLU), dropout, data augmentation, and so forth. ReLU was widely regarded as the most essential component for making deep learning possible. Then, in 2014, VGGNet [75] proposed a standard network architecture that used very small 3×3 convolutional filters throughout and doubled the number of feature maps after the 2×2 pooling. It increased the

[1] The time we present is when the paper was published.

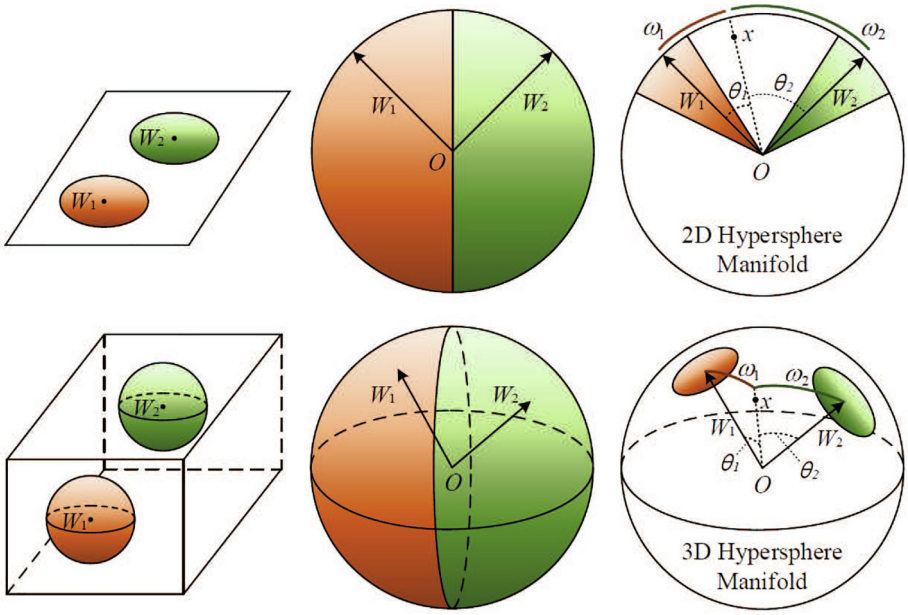

Fig. 6. Geometry interpretation of A-Softmax loss. [84].

Table 5
Decision boundaries for class 1 under binary classification case, where $\hat{x}$ is the normalized feature. [106]

Loss Functions	Decision Boundaries
Softmax	$(W_1 - W_2)x + b_1 - b_2 = 0$
L-Softmax [104]	$\|x\|(\|W_1\|cos(m\theta_1) - \|W_2\|cos(\theta_2)) > 0$
A-Softmax [84]	$\|x\|(cosm\theta_1 - cos\theta_2) = 0$
CosFace [105]	$\hat{x}(cos\theta_1 - m - cos\theta_2) = 0$
ArcFace [106]	$\hat{x}(cos(\theta_1 + m) - cos\theta_2) = 0$

depth of the network to 16–19 weight layers, which further enhanced the flexibility to learn progressive nonlinear mappings by deep architectures. In 2015, the 22-layer GoogleNet [76] introduced an "inception module" with the concatenation of hybrid feature maps, as well as two additional intermediate softmax supervised signals. It performs several convolutions with different receptive fields ($1 \times 1, 3 \times 3$ and 5×5) in parallel, and concatenates all feature maps to merge the multi-resolution information. In 2016, ResNet [77] proposed to make layers learn a residual mapping with reference to the layer inputs $\mathcal{F}(x) := \mathcal{H}(x) - x$ rather than directly learning a desired underlying mapping $\mathcal{H}(x)$ to ease the training of very deep networks (up to 152 layers). The original mapping is recast into $\mathcal{F}(x) + x$ and can be realized by "shortcut connections". As the champion of ILSVRC 2017, SENet [78] introduced a "Squeeze-and-Excitation" (SE) block, that adaptively recalibrates channel-wise feature responses by explicitly modelling interdependencies between channels. These blocks can be integrated with modern architectures, such as ResNet, and improves their representational power.

With the evolved architectures and advanced training techniques, such as batch normalization (BN), the network becomes deeper and the training becomes more controllable. Following these architectures in object classification, the networks in deep FR are also developed step by step, and the performance of deep FR is continually improving. We present these mainstream architectures of deep FR in Fig. 9. In 2014, DeepFace [20] was the first to use a nine-layer CNN with several locally connected layers. With 3D alignment for face processing, it reaches an accuracy of 97.35% on LFW. In 2015, FaceNet [38] used a large private dataset to train a GoogleNet. It adopted a triplet loss function based on triplets of roughly aligned matching/nonmatching face patches generated

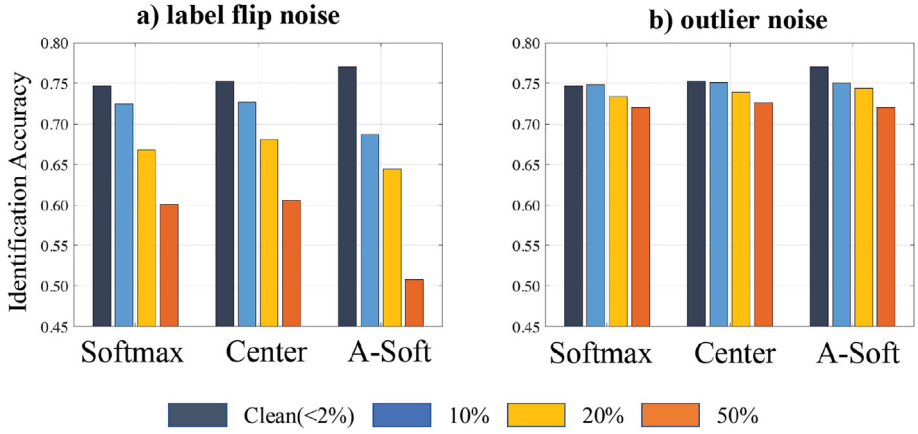

Fig. 7. 1:1 M rank-1 identification results on MegaFace benchmark: (a) introducing label flips to training data, (b) introducing outliers to training data. [124].

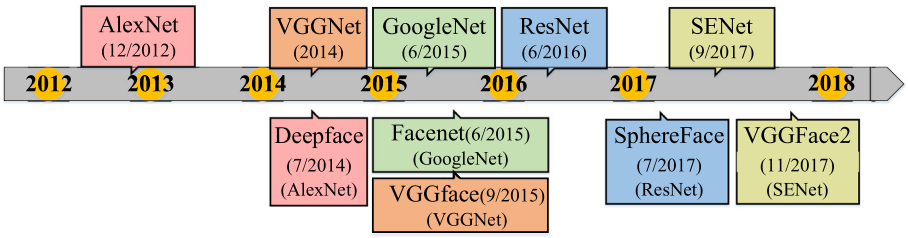

Fig. 8. The top row presents the typical network architectures in object classification, and the bottom row describes the well-known FR algorithms that use the typical architectures. We use the same color rectangles to represent the algorithms using the same architecture. It is easy to find that the architectures of deep FR have always followed those of deep object classification and evolved from AlexNet to SENet rapidly.

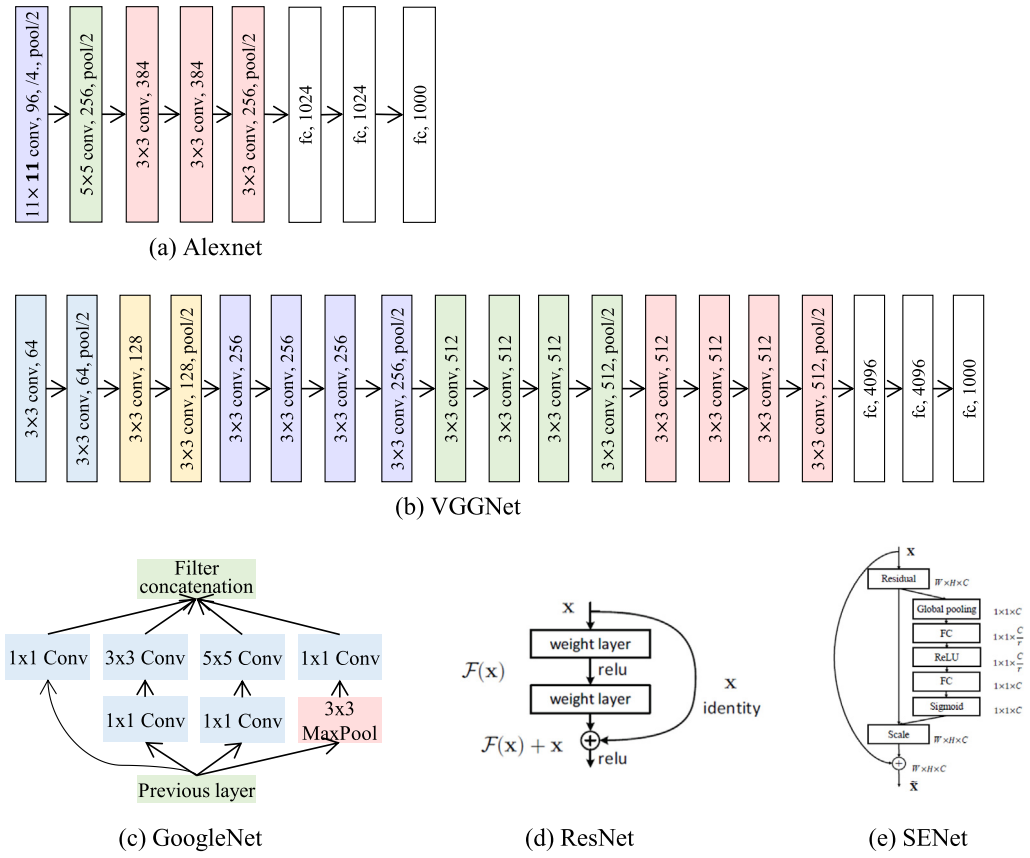

Fig. 9. The architecture of Alexnet [22], VGGNet [75], GoogleNet [76], ResNet [77], SENet [78].

by a novel online triplet mining method and achieved good performance of 99.63%. In the same year, VGGface [37] designed a procedure to collect a large-scale dataset from the Internet. It trained the VGGNet on this dataset and then fine-tuned the networks via a triplet loss function similar to FaceNet. VGGface obtains an accuracy of 98.95%. In 2017, SphereFace [84] used a 64-layer ResNet architecture and proposed the angular softmax (A-Softmax) loss to learn discriminative face features with angular margin. It boosts the achieves to 99.42% on LFW. In the end of 2017, a new large-scale face dataset, namely VGGface2 [39], was introduced, which consists of large variations in pose, age, illumination, ethnicity and profession. Cao et al. first trained a SENet with MS-celeb-1 M dataset [45] and then fine-tuned the model with VGGface2 [39], and achieved the SOTA performance on the IJB-A [41] and IJB-B [42].

Light-weight networks. Using deeper neural network with hundreds of layers and millions of parameters to achieve higher accuracy comes at cost. Powerful GPUs with larger memory size are needed, which makes the applications on many mobiles and embedded devices impractical. To address this problem, light-weight networks are proposed. Light CNN [85,86] proposed a max-feature-map (MFM) activation function that introduces the concept of maxout in the fully connected layer to CNN. The MFM obtains a compact representation and reduces the computational cost. Sun et al. [61] proposed to sparsify deep networks iteratively from the previously learned denser models based on a weight selection criterion. MobiFace [87] adopted fast downsampling and bottleneck residual block with the expansion layers and achieved high performance with 99.7% on LFW database. Although some other light-weight CNNs, such as SqueezeNet, MobileNet, ShuffleNet and Xception [126–129], are still not widely used in FR, they deserve more attention.

Adaptive-architecture networks. Considering that designing architectures manually by human experts are time-consuming and error-prone processes, there is growing interest in adaptive-

architecture networks which can find well-performing architectures, e.g. the type of operation every layer executes (pooling, convolution, etc) and hyper-parameters associated with the operation (number of filters, kernel size and strides for a convolutional layer, etc), according to the specific requirements of training and testing data. Currently, neural architecture search (NAS) [130] is one of the promising methodologies, which has outperformed manually designed architectures on some tasks such as image classification [131] or semantic segmentation [132]. Zhu et al. [88] integrated NAS technology into face recognition. They used reinforcement learning [133] algorithm (policy gradient) to guide the controller network to train the optimal child architecture. Besides NAS, there are some other explorations to learn optimal architectures adaptively. For example, conditional convolutional neural network (c–CNN) [89] dynamically activated sets of kernels according to modalities of samples; Han et al. [90] proposed a novel contrastive convolution consisted of a trunk CNN and a kernel generator, which is beneficial owing to its dynamistic generation of contrastive kernels based on the pair of faces being compared.

Joint alignment-recognition networks. Recently, an end-to-end system [91–94] was proposed to jointly train FR with several modules (face detection, alignment, and so forth) together. Compared to the existing methods in which each module is generally optimized separately according to different objectives, this end-to-end system optimizes each module according to the recognition objective, leading to more adequate and robust inputs for the recognition model. For example, inspired by spatial transformer [134], Hayat et al. [91] proposed a CNN-based data-driven approach that learns to simultaneously register and represent faces (Fig. 10), while Wu et al. [92] designed a novel recursive spatial transformer (ReST) module for CNN allowing face alignment and recognition to be jointly optimized.

3.2.2. Assembled networks

Multi-input networks. In "one-to-many augmentation", multiple images with variety are generated from one image in order to augment training data. Taken these multiple images as input, multiple networks are also assembled together to extract and combine features of different type of inputs, which can outperform an individual network. In [58–60,99,34,21,35], assembled networks are built after different face patches are cropped, and then different types of patches are fed into different sub-networks for representation extraction. By combining the results of sub-networks, the performance can be improved. Other papers [96,95,98] used assembled networks to recognize images with different poses. For example, Masi et al. [96] adjusted the pose to frontal (0°), half-profile (40°) and full-profile views (75°) and then addressed pose variation by assembled pose networks. A multi-view deep network (MvDN) [95] consists of view-specific subnetworks and common subnetworks; the former removes view-specific variations, and the latter obtains common representations.

Multi-task networks. FR is intertwined with various factors, such as pose, illumination, and age. To solve this problem, multi-task learning is introduced to transfer knowledge from other relevant tasks and to disentangle nuisance factors. In multi-task networks, identity classification is the main task and the side tasks are pose, illumination, and expression estimations, among others. The lower layers are shared among all the tasks, and the higher layers are disentangled into different sub-networks to generate the task-specific outputs. In [100], the task-specific sub-networks are branched out to learn face detection, face alignment, pose estimation, gender recognition, smile detection, age estimation and FR. Yin et al. [97] proposed to automatically assign the dynamic loss weights for each side task. Peng et al. [135] used a feature reconstruction metric learning to disentangle a CNN into sub-networks for jointly learning the identity and non-identity features as shown in Fig. 11.

3.3. Face matching by deep features

During testing, the cosine distance and L2 distance are generally employed to measure the similarity between the deep features x_1 and x_2; then, threshold comparison and the nearest neighbor (NN) classifier are used to make decision for verification and identification. In addition to these common methods, there are some other explorations.

3.3.1. Face verification

Metric learning, which aims to find a new metric to make two classes more separable, can also be used for face matching based on extracted deep features. The JB [136] model is a well-known metric learning method [35,21,36,34,120], and Hu et al. [79] proved that it can improve the performance greatly. In the JB model, a face feature x is modeled as $x = \mu + \varepsilon$, where μ and ε are identity and intra-personal variations, respectively. The similarity score $r(x_1, x_2)$ can be represented as follows:

$$r(x_1, x_2) = log \frac{P(x_1, x_2|H_I)}{P(x_1, x_2|H_E)} \tag{7}$$

where $P(x_1, x_2|H_I)$ is the probability that two faces belong to the same identity and $P(x_1, x_2|H_E)$ is the probability that two faces belong to different identities.

3.3.2. Face identification

After cosine distance was computed, Cheng et al. [137] proposed a heuristic voting strategy at the similarity score level to combine the results of multiple CNN models and won first place in Challenge 2 of MS-celeb-1 M 2017. Yang et al. [138] extracted the local adaptive convolution features from the local regions of the face image and used the extended SRC for FR with a single sample per person. Guo et al. [139] combined deep features and the SVM classifier to perform recognition. Wang et al. [62] first used product quantization (PQ) [140] to directly retrieve the top-k most similar faces and re-ranked these faces by combining similarities from deep features and the COTS matcher [141]. In addition, Softmax can be also used in face matching when the identities of training set and test set overlap. For example, in Challenge 2 of MS-celeb-1 M, Ding et al. [142] trained a 21,000-class softmax classifier to directly recognize faces of one-shot classes and normal classes after augmenting feature by a conditional GAN; Guo et al. [143] trained the softmax classifier combined with underrepresented-classes promotion (UP) loss term to enhance the performance on one-shot classes.

When the distributions of training data and testing data are the same, the face matching methods mentioned above are effective. However, there is always a distribution change or domain shift between two data domains that can degrade the performance on test data. Transfer learning [144,145] has recently been introduced into deep FR to address the problem of domain shift. It learns transferable features using a labeled source domain (training data) and an unlabeled target domain (testing data) such that domain discrepancy is reduced and models trained on source domain will also perform well on target domain. Sometimes, this technology is applied to face matching. For example, Crosswhite et al. [121] and Xiong et al. [146] adopted template adaptation to the set of media in a template by combining CNN features with template-specific linear SVMs. But most of the time, it is not enough to do transfer learning only at face matching stage. Transfer learning should be embedded in deep models to learn more transferable representations. Kan et al. [147] proposed a bi-shifting autoencoder network

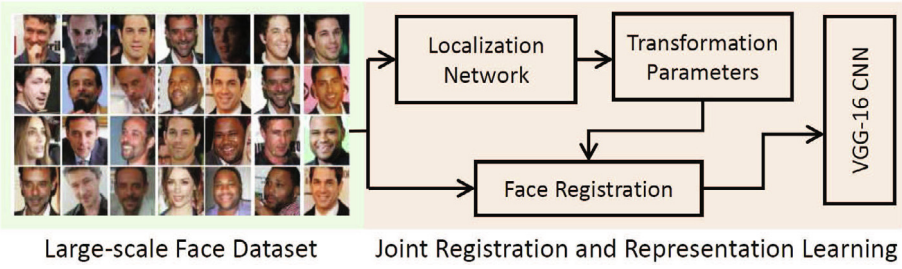

Fig. 10. Joint face registration and representation learning. [91].

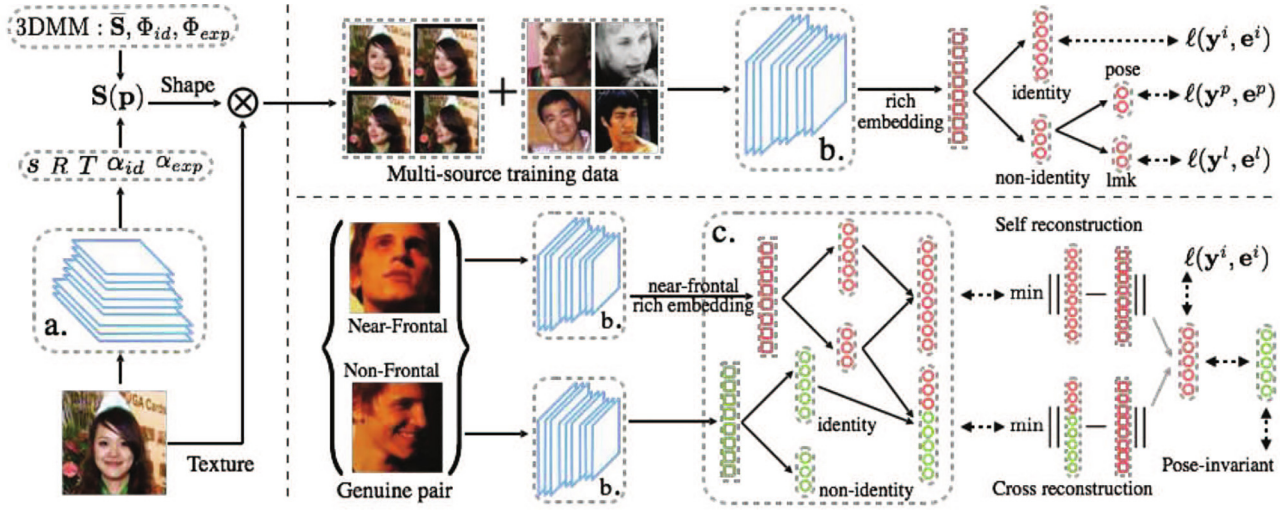

Fig. 11. Reconstruction-based disentanglement for pose-invariant FR. [135].

(BAE) for domain adaptation across view angle, ethnicity, and imaging sensor; while Luo et al. [148] utilized the multi-kernels maximum mean discrepancy (MMD) to reduce domain discrepancies. Sohn et al. [149] used adversarial learning [150] to transfer knowledge from still image FR to video FR. Moreover, fine-tuning the CNN parameters from a prelearned model using a target training dataset is a particular type of transfer learning, and is commonly employed by numerous methods [151,152,103].

4. Face processing for training and recognition

We present the development of face processing methods in chronological order in Fig. 12. As we can see from the figure, most papers attempted to perform face processing by autoencoder model in 2014 and 2015; while 3D model played an important role in 2016. GAN [40] has drawn substantial attention from the deep learning and computer vision community since it was first proposed by Goodfellow et al. It can be used in different fields and was also introduced into face processing in 2017. GAN can be used to perform "one-to-many augmentation" and "many-to-one normalization", and it broke the limit that face synthesis should be done under supervised way. Although GAN has not been widely used in face processing for training and recognition, it has great latent capacity for preprocessing, for example, Dual-Agent GANs (DA-GAN) [56] won the 1st places on verification and identification tracks in the NIST IJB-A 2017 FR competitions.

4.1. One-to-many augmentation

Collecting a large database is extremely expensive and time consuming. The methods of "one-to-many augmentation" can mitigate the challenges of data collection, and they can be used to augment not only training data but also the gallery of test data. we categorized them into four classes: data augmentation, 3D model, autoencoder model and GAN model.

Data augmentation. Common data augmentation methods consist of photometric transformations [75,22] and geometric transformations, such as oversampling (multiple patches obtained by cropping at different scales) [22], mirroring [153], and rotating [154] the images. Recently, data augmentation has been widely used in deep FR algorithms [58–60,35,21,36,61,62]. for example, Sun et al. [21] cropped 400 face patches varying in positions, scales, and color channels and mirrored the images. Liu et al. [58] generated seven overlapped image patches centered at different landmarks on the face region and trained them with seven CNNs with the same structure.

3D model. 3D face reconstruction is also a way to enrich the diversity of training data. They utilize 3D structure information to model the transformation between poses. 3D models first use 3D face data to obtain morphable displacement fields and then apply them to obtain 2D face data in different pose angles. There is a large number of papers about this domain, but we only focus on the 3D face reconstruction using deep methods or used for deep FR. In [47], Masi et al. generated face images with new intra-class facial appearance variations, including pose, shape and expression, and then trained a 19-layer VGGNet with both real and augmented data. Masi et al. [48] used generic 3D faces and rendered fixed views to reduce much of the computational effort. Richardson et al. [49] employed an iterative 3D CNN by using a secondary input channel to represent the previous network's output as an image for reconstructing a 3D face as shown in Fig. 13. Dou et al. [51] used a multi-task CNN to divide 3D face reconstruction into neutral 3D reconstruction and expressive 3D reconstruction. Tran et al. [53] directly regressed 3D morphable face model (3DMM)

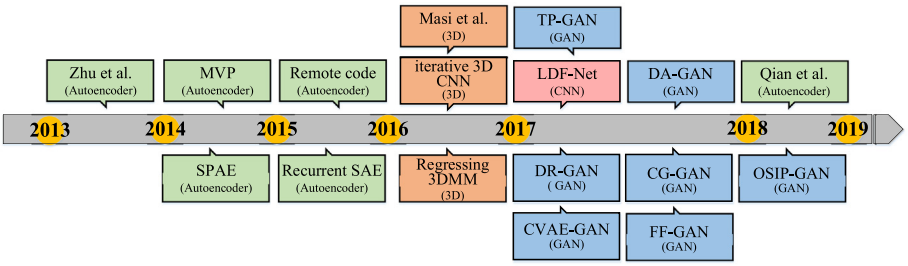

Fig. 12. The development of deep face processing methods. Red, green, orange and blue rectangles represent CNN model, autoencoder model, 3D model and GAN model, respectively.

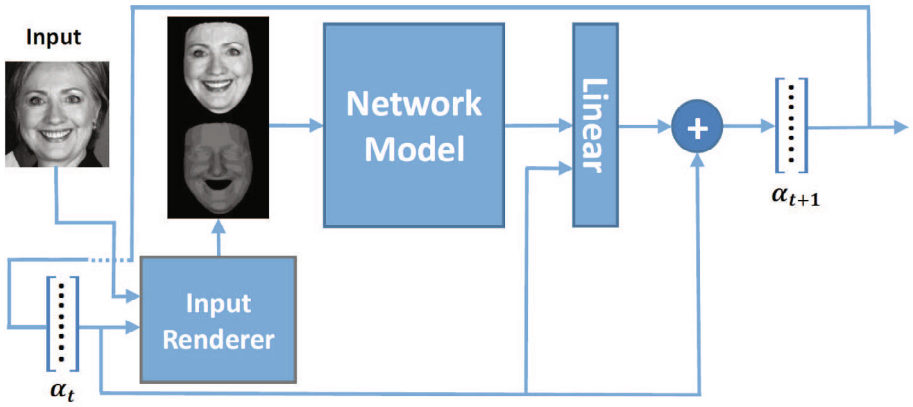

Fig. 13. Iterative CNN network for reconstructing a 3D face. [49].

[155] parameters from an input photo by a very deep CNN architecture. An et al. [156] synthesized face images with various poses and expressions using the 3DMM method, then reduced the gap between synthesized data and real data with the help of MMD.

Autoencoder model. Rather than reconstructing 3D models from a 2D image and projecting it back into 2D images of different poses, autoencoder models can generate 2D target images directly. Taken a face image and a pose code encoding a target pose as input, an encoder first learns pose-invariant face representation, and then a decoder generates a face image with the same identity viewed at the target pose by using the pose-invariant representation and the pose code. For example, given the target pose codes, multi-view perceptron (MVP) [55] trained some deterministic hidden neurons to learn pose-invariant face representations, and simultaneously trained some random hidden neurons to capture pose features, then a decoder generated the target images by combining pose-invariant representations with pose features. As shown in Fig. 14, Yim et al. [157] and Qian et al. [158] introduced an auxiliary CNN to generate better images viewed at the target poses. First, an autoencoder generated the desired pose image, then the auxiliary CNN reconstructed the original input image back from the generated target image, which guarantees that the generated image is identity-preserving. In [65], two groups of units are embedded between encoder and decoder. The identity units remain unchanged and the rotation of images is achieved by taking actions to pose units at each time step.

GAN model. In GAN models, a generator aims to fool a discriminator through generating images that resemble the real images, while the discriminator aims to discriminate the generated samples from the real ones. By this minimax game between generator and discriminator, GAN can successfully generate photo-realistic images with different poses. After using a 3D model to generate profile face images, DA-GAN [56] refined the images by a GAN, which combines prior knowledge of the data distribution and knowledge of faces (pose and identity perception loss). CVAE-GAN [159] combined a variational auto-encoder with a GAN for augmenting data, and took advantages of both statistic and pairwise feature matching to make the training process converge faster and more stably. In addition to synthesizing diverse faces from noise, some papers also explore to disentangle the identity and variation, and synthesize new faces by exchanging identity and variation from different people. In CG-GAN [160], a generator directly resolves each representation of input image into a variation code and an identity code and regroups these codes for cross-generating, simultaneously, a discriminator ensures the reality of generated images. Bao et al. [161] extracted identity representation of one input image and attribute representation of any other input face image, then synthesized new faces by recombining these representations. This work shows superior performance in generating realistic and identity preserving face images, even for identities outside the training dataset. Unlike previous methods that treat classifier as a spectator, FaceID-GAN [162] proposed a three-player GAN where the classifier cooperates together with the discriminator to compete with the generator from two different aspects, i.e. facial identity and image quality respectively.

4.2. Many-to-one normalization

In contrast to "one-to-many augmentation", the methods of "many-to-one normalization" produce frontal faces and reduce appearance variability of test data to make faces align and compare easily. It can be categorized as autoencoder model, CNN model and GAN model.

Autoencoder model. Autoencoder can also be applied to "many-to-one normalization". Different from the autoencoder model in "one-to-many augmentation" which generates the

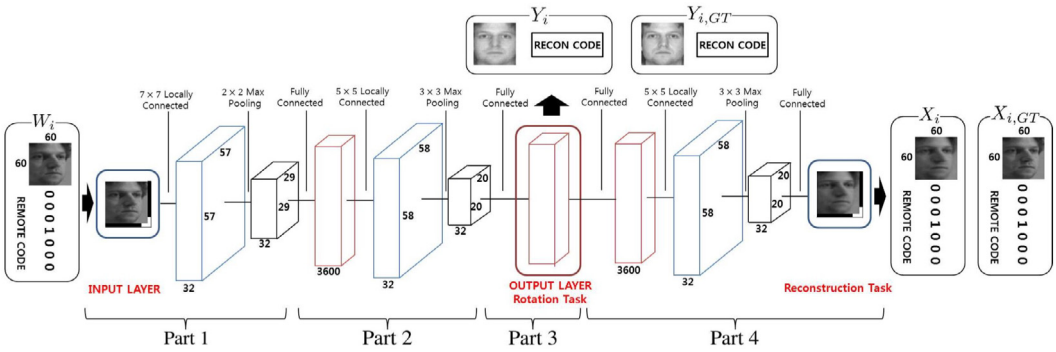

Fig. 14. Autoencoder model of "one-to-many augmentation" proposed by [157]. The first part extracts feature from an input image, then the second and third part generate a target image with the same identity viewed at the target pose. The forth part is an auxiliary task which reconstructs the original input image back from the generated image to guarantee that the generated image is identity-preserving.

desired pose images with the help of pose codes, autoencoder model here learns pose-invariant face representation by an encoder and directly normalizes faces by a decoder without pose codes. Zhu et al. [66,67] selected canonical-view images according to the face images' symmetry and sharpness and then adopted an autoencoder to recover the frontal view images by minimizing the reconstruction loss error. The proposed stacked progressive autoencoders (SPAE) [63] progressively map the nonfrontal face to the frontal face through a stack of several autoencoders. Each shallow autoencoders of SPAE is designed to convert the input face images at large poses to a virtual view at a smaller pose, so the pose variations are narrowed down gradually layer by layer along the pose manifold. Zhang et al. [64] built a sparse many-to-one encoder to enhance the discriminant of the pose free feature by using multiple random faces as the target values for multiple encoders.

CNN model. CNN models usually directly learn the 2D mappings between non-frontal face images and frontal images, and utilize these mapping to normalize images in pixel space. The pixels in normalized images are either directly the pixels or the combinations of the pixels in non-frontal images. In LDF-Net [68], the displacement field network learns the shifting relationship of two pixels, and the translation layer transforms the input non-frontal face image into a frontal one with this displacement field. In GridFace [69] shown in Fig. 15, first, the rectification network normalizes the images by warping pixels from the original image to the canonical one according to the computed homography matrix, then the normalized output is regularized by an implicit canonical view face prior, finally, with the normalized faces as input, the recognition network learns discriminative face representation via metric learning.

GAN model. Huang et al. [70] proposed a two-pathway generative adversarial network (TP-GAN) that contains four landmark-located patch networks and a global encoder-decoder network. Through combining adversarial loss, symmetry loss and identity-preserving loss, TP-GAN generates a frontal view and simultaneously preserves global structures and local details as shown in Fig. 16. In a disentangled representation learning generative adversarial network (DR-GAN) [71], the generator serves as a face rotator, in which an encoder produces an identity representation, and a decoder synthesizes a face at the specified pose using this representation and a pose code. And the discriminator is trained to not only distinguish real vs. synthetic images, but also predict the identity and pose of a face. Yin et al. [73] incorporated 3DMM into the GAN structure to provide shape and appearance priors to guide the generator to frontalization.

5. Face databases and evaluation protocols

In the past three decades, many face databases have been constructed with a clear tendency from small-scale to large-scale, from single-source to diverse-sources, and from lab-controlled to real-world unconstrained condition, as shown in Fig. 17. As the performance of some simple databases become saturated, e.g. LFW [23], more and more complex databases were continually developed to facilitate the FR research. It can be said without exaggeration that the development process of the face databases largely leads the direction of FR research. In this section, we review the development of major training and testing academic databases for the deep FR.

5.1. Large-scale training data sets

The prerequisite of effective deep FR is a sufficiently large training dataset. Zhou et al. [59] suggested that large amounts of data with deep learning improve the performance of FR. The results of Megaface Challenge also revealed that premier deep FR methods

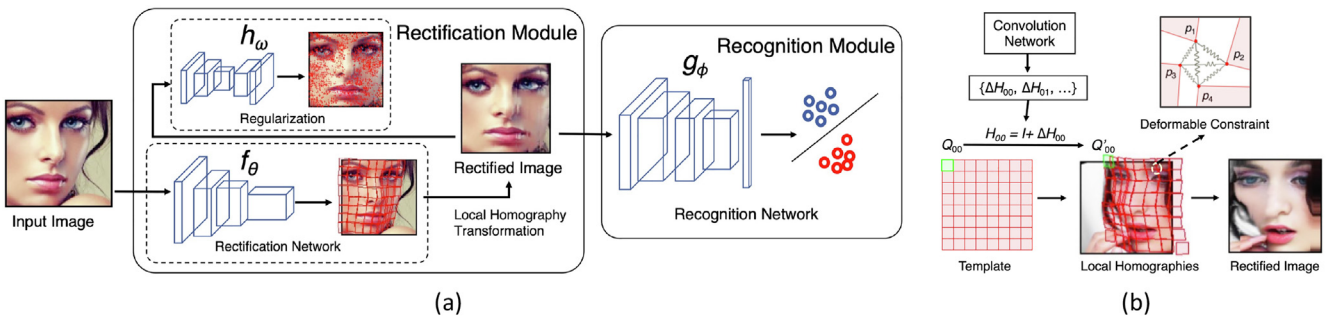

Fig. 15. (a) System overview and (b) local homography transformation of GridFace [69]. The rectification network normalizes the images by warping pixels from the original image to the canonical one according to the computed homography matrix.

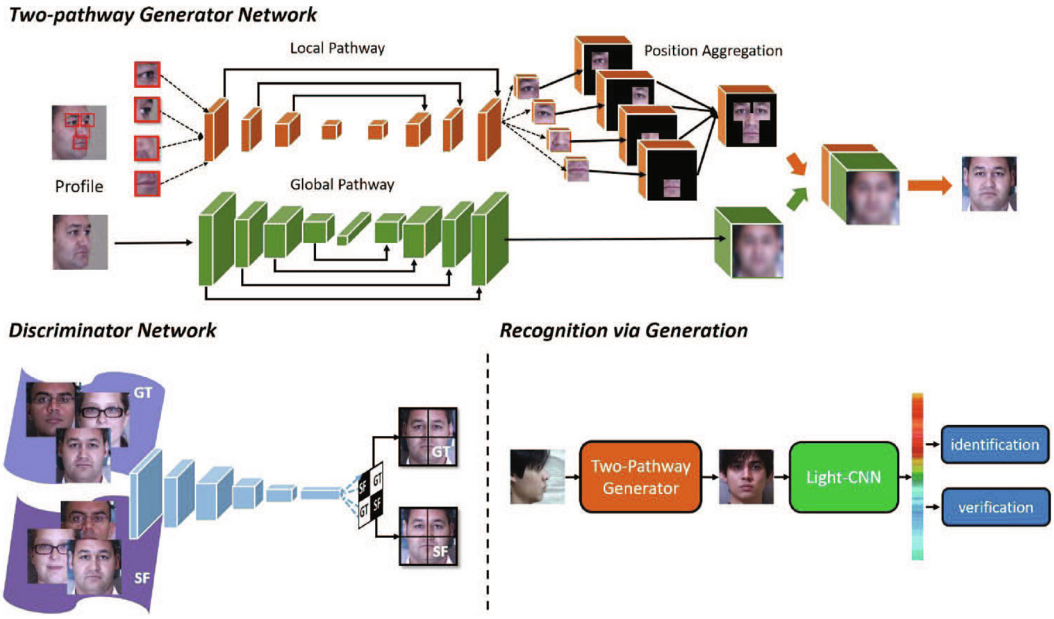

Fig. 16. General framework of TP-GAN [70]. The generator contains two pathways with each processing global or local transformations. The discriminator distinguishes between synthesized frontal views and ground-truth frontal views.

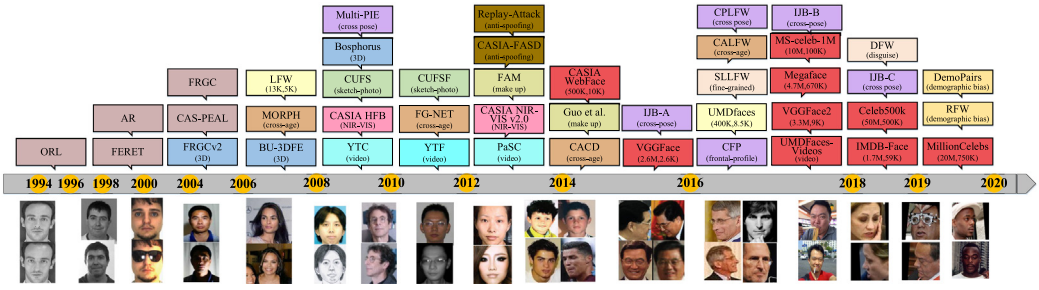

Fig. 17. The evolution of FR datasets. Before 2007, early works in FR focused on controlled and small-scale datasets. In 2007, LFW [23] dataset was introduced which marks the beginning of FR under unconstrained conditions. Since then, more testing databases designed for different tasks and scenes are constructed. And in 2014, CASIA-Webface [120] provided the first widely-used public training dataset, large-scale training datasets begun to be hot topic. Red rectangles represent training datasets, and other color rectangles represent different testing datasets.

were typically trained on data larger than 0.5 M images and 20 K people. The early works of deep FR were usually trained on private training datasets. Facebook's Deepface [20] model was trained on 4 M images of 4 K people; Google's FaceNet [38] was trained on 200 M images of 3 M people; DeepID serial models [34,35,21,36] were trained on 0.2 M images of 10 K people. Although they reported ground-breaking performance at this stage, researchers cannot accurately reproduce or compare their models without public training datasets.

To address this issue, CASIA-Webface [120] provided the first widely-used public training dataset for the deep model training purpose, which consists of 0.5 M images of 10 K celebrities collected from the web. Given its moderate size and easy usage, it has become a great resource for fair comparisons for academic deep models. However, its relatively small data and ID size may not be sufficient to reflect the power of many advanced deep learning methods. Currently, there have been more databases providing public available large-scale training dataset (Table 6), especially three databases with over 1 M images, namely MS-Celeb-1 M [45], VGGface2 [39], and Megaface [44,164], and we summary some interesting findings about these training sets, as shown in Fig. 18.

Depth v.s. breadth. These large training sets are expanded from depth or breadth. VGGface2 provides a large-scale training dataset of depth, which have limited number of subjects but many images for each subjects. The depth of dataset enforces the trained model to address a wide range intra-class variations, such as lighting, age, and pose. In contrast, MS-Celeb-1 M and Mageface (Challenge 2) offers large-scale training datasets of breadth, which contains many subject but limited images for each subjects. The breadth of dataset ensures the trained model to cover the sufficiently variable appearance of various people. Cao et al. [39] conducted a systematic studies on model training using VGGface2 and MS-Celeb-1 M, and found an optimal model by first training on MS-Celeb-1 M (breadth) and then fine-tuning on VGGface2 (depth).

Long tail distribution. The utilization of long tail distribution is different among datasets. For example, in Challenge 2 of MS-Celeb-1 M, the novel set specially uses the tailed data to study low-shot learning; central part of the long tail distribution is used by the Challenge 1 of MS-Celeb-1 M and images' number is approximately limited to 100 for each celebrity; VGGface and VGGface2 only use the head part to construct deep databases; Megaface utilizes the whole distribution to contain as many images as possible, the minimal number of images is 3 per person and the maximum is 2469.

Table 6
The commonly used FR datasets for training

Datasets	Publish Time	#photos	#subjects	# of photos per subject [1]	Key Features
MS-Celeb-1 M (Challenge 1)[45]	2016	10 M 3.8 M(clean)	100,000 85 K(clean)	100	breadth; central part of long tail; celebrity; knowledge base
MS-Celeb-1 M (Challenge 2)[45]	2016	1.5 M(base set) 1 K (novel set)	20 K(base set) 1 K(novel set)	1/-/100	low-shot learning; tailed data; celebrity
MS-Celeb-1 M (Challenge 3) [163]	2018	4 M(MSv1c) 2.8 M (Asian-Celeb)	80 K(MSv1c) 100 K (Asian-Celeb)	-	breadth;central part of long tail; celebrity
MegaFace [44,164]	2016	4.7 M	672,057	3/7/2469	breadth; the whole long tail;commonalty
VGGFace2 [39]	2017	3.31 M	9,131	87/362.6/843	depth; head part of long tail; cross pose, age and ethnicity; celebrity
CASIA WebFace [120]	2014	494,414	10,575	2/46.8/804	celebrity
MillionCelebs [165]	2020	18.8 M	636 K	29.5	celebrity
IMDB-Face [124]	2018	1.7 M	59 K	28.8	celebrity
UMDFaces-Videos [166]	2017	22,075	3,107	–	video
VGGFace [37]	2015	2.6 M	2,622	1,000	depth; celebrity; annotation with bounding boxesand coarse pose
CelebFaces+ [21]	2014	202,599	10,177	19.9	private
Google [38]	2015	>500 M	>10 M	50	private
Facebook [20]	2014	4.4 M	4 K	800/1100/1200	private

The min/average/max numbers of photos or frames per subject.

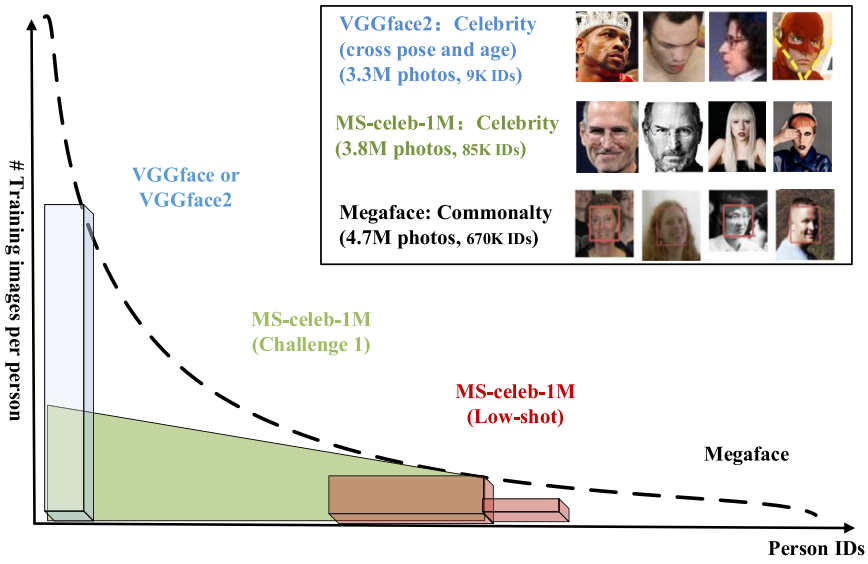

Fig. 18. The distribution of three new large-scale databases suitable for training deep models. They have larger scale than the widely-used CAISA-Web database. The vertical axis displays number of images per person, and the horizontal axis shows person IDs.

Data engineering. Several popular benchmarks, such as LFW unrestricted protocol, Megaface Challenge 1, MS-Celeb-1 M Challenge 1&2, explicitly encourage researchers to collect and clean a large-scale data set for enhancing the capability of deep neural network. Although data engineering is a valuable problem to computer vision researchers, this protocol is more incline to the industry participants. As evidence, the leaderboards of these experiments are mostly occupied by the companies holding invincible hardwares and data scales. This phenomenon may not be beneficial for developments of new models in academic community.

Data noise. Owing to data source and collecting strategies, existing large-scale datasets invariably contain label noises. Wang et al. [124] profiled the noise distribution in existing datasets in Fig. 19 and showed that the noise percentage increases dramatically along the scale of data. Moreover, they found that noise is more lethal on a 10,000-class problem of FR than on a 10-class problem of object classification and that label flip noise severely deteriorates the performance of a model, especially the model using A-softmax [84]. Therefore, building a sufficiently large and clean dataset for academic research is very meaningful. Deng et al. [106] found there are serious label noise in MS-Celeb-1 M [45], and they cleaned the noise of MS-Celeb-1 M, and made the refined dataset public available. Microsoft and Deepglint jointly released the largest public data set [163] with cleaned labels, which includes 4 M images cleaned from MS-Celeb-1 M dataset

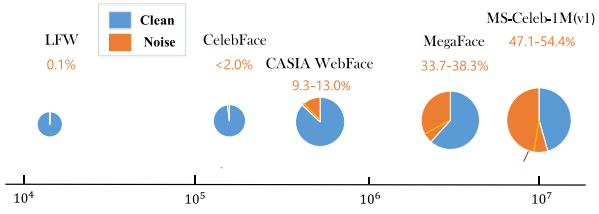

Fig. 19. A visualization of the size and estimated noise percentage of datasets. [124].

and 2.8 M aligned images of 100 K Asian celebrities. Moreover, Zhan et al. [167] shifted the focus from cleaning the datasets to leveraging more unlabeled data. Through automatically assigning pseudo labels to unlabeled data with the help of relational graphs, they obtained competitive or even better results over the fully-supervised counterpart.

Data bias. Large-scale training datasets, such as CASIA-WebFace [120], VGGFace2 [39] and MS-Celeb-1 M [45], are typically constructed by scraping websites like Google Images, and consist of celebrities on formal occasions: smiling, make-up, young, and beautiful. They are largely different from databases captured in the daily life (e.g. Megaface). The biases can be attributed to many exogenous factors in data collection, such as cameras, lightings, preferences over certain types of backgrounds, or annotator tendencies. Dataset biases adversely affect cross-dataset generalization; that is, the performance of the model trained on one dataset drops significantly when applied to another one. One persuasive evidence is presented by P.J. Phillips' study [168] which conducted a cross benchmark assessment of VGGFace model [37] for face recognition. The VGGFace model achieves 98.95% on LFW [23] and 97.30% on YTF [169], but only obtains 26%, 52% and 85% on Ugly, Bad and Good partition of GBU database [170].

Demographic bias (e.g., race/ethnicity, gender, age) in datasets is a universal but urgent issue to be solved in data bias field. In existing training and testing datasets, the male, White, and middle-aged cohorts always appear more frequently, as shown in Table 7, which inevitably causes deep learning models to replicate and even amplify these biases resulting in significantly different accuracies when deep models are applied to different demographic groups. Some researches [145,171,172] showed that the female, Black, and younger cohorts are usually more difficult to recognize in FR systems trained with commonly-used datasets. For example, Wang et al. [173] proposed a Racial Faces in-the-Wild (RFW) database and proved that existing commercial APIs and the SOTA algorithms indeed work unequally for different races and the maximum difference in error rate between the best and worst groups is 12%, as shown in Table 8. Hupont et al. [171] showed that SphereFace has a TAR of 0.87 for White males which drops to 0.28 for Asian females, at a FAR of $1e-4$. Such bias can result in mistreatment of certain demographic groups, by either exposing them to a higher risk of fraud, or by making access to services more difficult. Therefore, addressing data bias and enhancing fairness of FR systems in real life are urgent and necessary tasks. Collecting balanced data to train a fair model or designing some debiasing algorithms are effective way.

5.2. Training protocols

In terms of training protocol, FR can be categorized into subject-dependent and subject-independent settings, as illustrated in Fig. 20.

Subject-dependent protocol. For subject-dependent protocol, all testing identities are predefined in training set, it is natural to classify testing face images to the given identities. Therefore, subject-dependent FR can be well addressed as a classification problem, where features are expected to be separable. The protocol is mostly adopted by the early-stage (before 2010) FR studies on FERET [177], AR [178], and is suitable only for some small-scale applications. The Challenge 2 of MS-Celeb-1 M is the only large-scale database using subject-dependent training protocol.

Subject-independent protocol. For subject-independent protocol, the testing identities are usually disjoint from the training set, which makes FR more challenging yet close to practice. Because it is impossible to classify faces to known identities in training set, generalized representation is essential. Due to the fact that human faces exhibit similar intra-subject variations, deep models can display transcendental generalization ability when training with a sufficiently large set of generic subjects, where the key is to learn discriminative large-margin deep features. This generalization ability makes subject-independent FR possible. Almost all major face-recognition benchmarks, such as LFW [23], PaSC [179], IJB-A/B/C [41–43] and Megaface [44,164], require the tested models to be trained under subject-independent protocol.

5.3. Evaluation tasks and performance metrics

In order to evaluate whether our deep models can solve the different problems of FR in real life, many testing datasets are designed to evaluate the models in different tasks, i.e. face verification, close-set face identification and open-set face identification. In either task, a set of known subjects is initially enrolled in the system (the gallery), and during testing, a new subject (the probe) is presented. Face verification computes one-to-one similarity between the gallery and probe to determine whether the two images are of the same subject, whereas face identification computes one-to-many similarity to determine the specific identity of a probe face. When the probe appears in the gallery identities, this is referred to as closed-set identification; when the probes include those who are not in the gallery, this is open-set identification.

Face verification is relevant to access control systems, re-identification, and application independent evaluations of FR algorithms. It is classically measured using the receiver operating characteristic (ROC) and estimated mean accuracy (Acc). At a given threshold (the independent variable), ROC analysis measures the true accept rate (TAR), which is the fraction of genuine comparisons that correctly exceed the threshold, and the false accept rate (FAR), which is the fraction of impostor comparisons that incorrectly exceed the threshold. And Acc is a simplified metric introduced by LFW [23], which represents the percentage of correct classifications. With the development of deep FR, more accurate recognitions are required. Customers concern more about the TAR when FAR is kept in a very low rate in most security certification scenario. PaSC [179] reports TAR at a FAR of 10^{-2}; IJB-A [41] evaluates TAR at a FAR of 10^{-3}; Megaface [44,164] focuses on TAR@10^{-6}FAR; especially, in MS-celeb-1 M challenge 3 [163], TAR@10^{-9}FAR is reported.

Close-set face identification is relevant to user driven searches (e.g., forensic identification), rank-N and cumulative match characteristic (CMC) is commonly used metrics in this scenario. Rank-N is based on what percentage of probe searches return the probe's gallery mate within the top k rank-ordered results. The CMC curve reports the percentage of probes identified within a given rank (the independent variable). IJB-A/B/C [41–43] concern on the rank-1 and rank-5 recognition rate. The MegaFace challenge [44,164] systematically evaluates rank-1 recognition rate function of increasing number of gallery distractors (going from 10 to 1 Million), the results of the SOTA evaluated on MegaFace challenge are listed in Table 9. Rather than rank-N and CMC, MS-Celeb-1 M [45] further applies a precision-coverage curve to measure identification performance under a variable threshold t. The probe is rejected when its confidence score is lower than t. The algorithms are compared in term of what fraction of passed probes, i.e. coverage, with a high recognition precision, e.g. 95% or 99%, the results of the SOTA evaluated on MS-Celeb-1 M challenge are listed in Table 10.

Open-set face identification is relevant to high throughput face search systems (e.g., de-duplication, watch list identification), where the recognition system should reject unknown/unseen subjects (probes who do not present in gallery) at test time. At present,

Table 7
Statistical demographic information of commonly-used training and testing databases. [173,171]

Train/ Test	Database	Race (%)				Gender (%)	
		Caucasian	Asian	Indian	African	Female	Male
train	CASIA-WebFace [120]	84.5	2.6	1.6	11.3	41.1	58.9
	VGGFace2 [39]	74.2	6.0	4.0	15.8	40.7	59.3
	MS-Celeb-1 M [45]	76.3	6.6	2.6	14.5	-	-
test	LFW [23]	69.9	13.2	2.9	14.0	25.8	74.2
	IJB-A [41]	66.0	9.8	7.2	17.0	-	-

Table 8
Racial bias in existing commercial recognition APIs and face recognition algorithms. Face verification accuracies (%) on RFW database are given [173].

Model	LFW	RFW			
		Caucasian	Indian	Asian	African
Microsoft	98.22	87.60	82.83	79.67	75.83
Face++	97.03	93.90	88.55	92.47	87.50
Baidu	98.67	89.13	86.53	90.27	77.97
Amazon	98.50	90.45	87.20	84.87	86.27
mean	98.11	90.27	86.28	86.82	81.89
Center-loss [101]	98.75	87.18	81.92	79.32	78.00
Sphereface [84]	99.27	90.80	87.02	82.95	82.28
Arcface [106]	99.40	92.15	88.00	83.98	84.93
VGGface2 [39]	99.30	89.90	86.13	84.93	83.38
mean	99.18	90.01	85.77	82.80	82.15

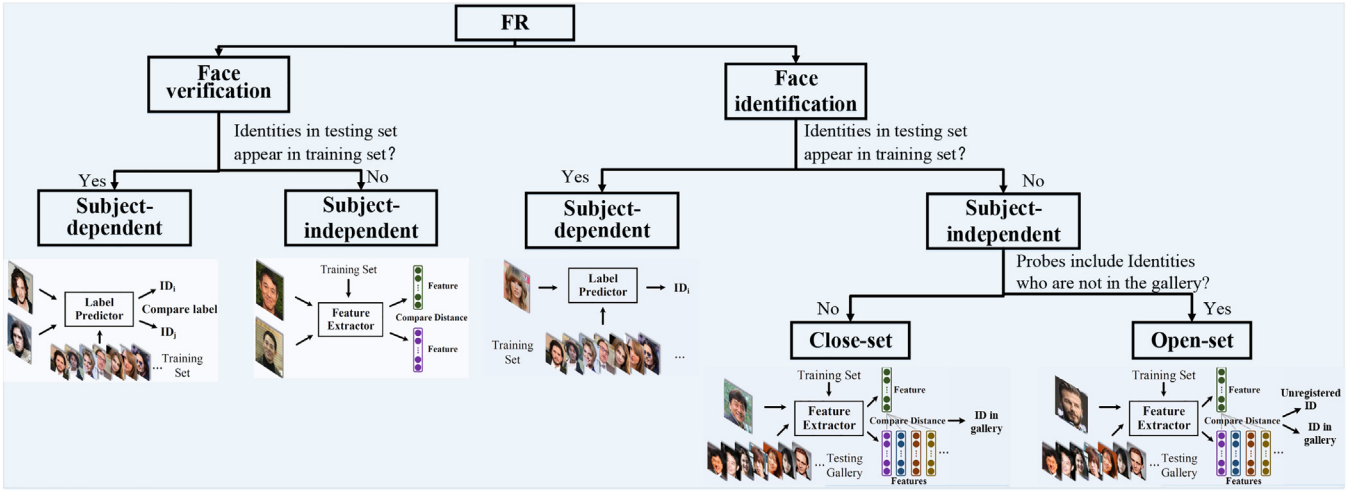

Fig. 20. The comparison of different training protocol and evaluation tasks in FR. In terms of training protocol, FR can be classified into subject-dependent or subject-independent settings according to whether testing identities appear in training set. In terms of testing tasks, FR can be classified into face verification, close-set face identification, open-set face identification.

Table 9
Performance of state of the arts on Megaface dataset.

Method	Megaface challenge1				Method	Megaface challenge2			
	FaceScrub		FGNet			FaceScrub		FGNet	
	Rank1 @10^6	TPR @10^{-6}FPR	Rank1 @10^6	TPR @10^{-6}FPR		Rank1 @10^6	TPR @10^{-6}FPR	Rank1 @10^6	TPR @10^{-6}FPR
Arcface [106]	0.9836	0.9848	-	-	Cosface [105]	0.7707	0.9030	0.6118	0.6350
Cosface [105]	0.9833	0.9841	-	-					
A-softmax [84]	0.9743	0.9766	-	-					
Marginal loss [116]	0.8028	0.9264	0.6643	0.4370					

there are very few databases covering the task of open-set FR. IJB-A/B/C [41–43] benchmarks introduce a decision error tradeoff (DET) curve to characterize the false negative identification rate (FNIR) as function of the false positive identification rate (FPIR). FPIR measures what fraction of comparisons between probe templates and non-mate gallery templates result in a match score exceeding T. At the same time, FNIR measures what fraction of probe searches will fail to match a mated gallery template above

Table 10
Performance of state of the arts on MS-celeb-1 M dataset.

Method	MS-celeb-1 M challenge1			Method	MS-celeb-1 M challenge2		
	External Data	C@P = 0.95 random set	C@P = 0.95 hard set		External Data	Top 1 Accuracy base set	C@P = 0.99 novel set
MCSM [174]	w	0.8750	0.7910	Cheng et al. [137]	w	0.9974	0.9901
Wang et al. [175]	w/o	0.7500	0.6060	Ding et al. [142]	w/o	-	0.9484
				Hybrid Classifiers [176]	w/o	0.9959	0.9264
				UP loss [143]	w/o	0.9980	0.7748

Table 11
Face Identification and Verification Evaluation of state of the arts on IJB-A dataset

Method	IJB-A Verification (TAR@FAR)			IJB-A Identification			
	0.001	0.01	0.1	FPIR = 0.01	FPIR = 0.1	Rank = 1	Rank = 10
TDFF [146]	0.979 ± 0.004	0.991 ± 0.002	0.996 ± 0.001	0.946 ± 0.047	0.987 ± 0.003	0.992 ± 0.001	0.998 ± 0.001
L2-softmax [109]	0.943 ± 0.005	0.970 ± 0.004	0.984 ± 0.002	0.915 ± 0.041	0.956 ± 0.006	0.973 ± 0.005	0.988 ± 0.003
DA-GAN [56]	0.930 ± 0.005	0.976 ± 0.007	0.991 ± 0.003	0.890 ± 0.039	0.949 ± 0.009	0.971 ± 0.007	0.989 ± 0.003
VGGface2 [39]	0.921 ± 0.014	0.968 ± 0.006	0.990 ± 0.002	0.883 ± 0.038	0.946 ± 0.004	0.982 ± 0.004	0.994 ± 0.001
TDFF [146]	0.919 ± 0.006	0.961 ± 0.007	0.988 ± 0.003	0.878 ± 0.035	0.941 ± 0.010	0.964 ± 0.006	0.992 ± 0.002
NAN [83]	0.881 ± 0.011	0.941 ± 0.008	0.979 ± 0.004	0.817 ± 0.041	0.917 ± 0.009	0.958 ± 0.005	0.986 ± 0.003
All-In-One Face [100]	0.823 ± 0.020	0.922 ± 0.010	0.976 ± 0.004	0.792 ± 0.020	0.887 ± 0.014	0.947 ± 0.008	0.988 ± 0.003
Template Adaptation [121]	0.836 ± 0.027	0.939 ± 0.013	0.979 ± 0.004	0.774 ± 0.049	0.882 ± 0.016	0.928 ± 0.010	0.986 ± 0.003
TPE [81]	0.813 ± 0.020	0.900 ± 0.010	0.964 ± 0.005	0.753 ± 0.030	0.863 ± 0.014	0.932 ± 0.010	0.977 ± 0.005

a score of *T*. The algorithms are compared in term of the FNIR at a low FPIR, e.g. 1% or 10%, the results of the SOTA evaluated on IJB-A dataset as listed in Table 11.

5.4. Evaluation scenes and data

Public available training databases are mostly collected from the photos of celebrities due to privacy issue, it is far from images captured in the daily life with diverse scenes. In order to study different specific scenarios, more difficult and realistic datasets are constructed accordingly, as shown in Table 12. According to their characteristics, we divide these scenes into four categories: cross-factor FR, heterogenous FR, multiple (or single) media FR and FR in industry (Fig. 21).

- Cross-factor FR. Due to the complex nonlinear facial appearance, some variations will be caused by people themselves, such as cross-pose, cross-age, make-up, and disguise. For example, CALFW [188], MORPH [189], CACD [191] and FG-NET [194] are commonly used datasets with different age range; CFP [182] only focuses on frontal and profile face, CPLFW [181] is extended from LFW and contains different poses. Disguised faces in the wild (DFW) [214] evaluates face recognition across disguise.
- Heterogenous FR. It refers to the problem of matching faces across different visual domains. The domain gap is mainly caused by sensory devices and cameras settings, e.g. visual light vs. near-infrared and photo vs. sketch. For example, CUFSF [201] and CUFS [199] are commonly used photo-sketch datasets and CUFSF dataset is harder due to lighting variation and shape exaggeration.
- Multiple (or single) media FR. Ideally, in FR, many images of each subject are provided in training datasets and image-to-image recognitions are performed when testing. But the situation will be different in reality. Sometimes, the number of images per person in training set could be very small, such as MS-Celeb-1 M challenge 2 [45]. This challenge is often called low- shot or few-shot FR. Moreover, each subject face in test set may be enrolled with a set of images and videos and set-to-set recognition should be performed, such as IJB-A [41] and PaSC [179].
- FR in industry. Although deep FR has achieved beyond human performance on some standard benchmarks, but some other factors should be given more attention rather than accuracy when deep FR is adopted in industry, e.g. anti-attack (CASIA-FASD [210]) and 3D FR (Bosphorus [203], BU-3DFE [205] and FRGCv2[206]). Compared to publicly available 2D face databases, 3D scans are hard to acquire, and the number of scans and subjects in public 3D face databases is still limited, which hinders the development of 3D deep FR.

6. Diverse recognition scenes of deep learning

Despite the high accuracy in the LFW [23] and Megaface [44,164] benchmarks, the performance of FR models still hardly meets the requirements in real-world application. A conjecture in industry is made that results of generic deep models can be improved simply by collecting big datasets of the target scene. However, this holds only to a certain degree. More and more concerns on privacy may make the collection and human-annotation of face data become illegal in the future. Therefore, significant efforts have been paid to design excellent algorithms to address the specific problems with limited data in these realistic scenes. In this section, we present several special algorithms of FR.

6.1. Cross-factor face recognition

6.1.1. Cross-pose face recognition

As [182] shows that many existing algorithms suffer a decrease of over 10% from frontal-frontal to frontal-profile verification, cross-pose FR is still an extremely challenging scene. In addition to the aforementioned methods, including "one-to-many augmentation", "many-to-one normalization" and assembled networks (Section 4 and 3.2.2), there are some other algorithms designed for cross-pose FR. Considering the extra burden of above methods, Cao et al. [215] attempted to perform frontalization in the deep

Table 12
The commonly used FR datasets for testing.

Datasets	Publish Time	#photos	#subjects	# of photos per subject [1]	Metrics	Typical Methods & Accuracy [2]	Key Features (Section)
LFW [23]	2007	13 K	5 K	1/2.3/530	1:1: Acc, TAR vs. FAR (ROC); 1:N: Rank-N, DIR vs. FAR (CMC)	99.78% Acc [109]; 99.63% Acc [38]	annotation with several attribute
MS-Celeb-1 M Challenge 1 [45]	2016	2 K	1 K	2	Coverage@P = 0.95	random set: 87.50%@P = 0.95; hard set: 79.10%@P = 0.95 [174];	large-scale
MS-Celeb-1 M Challenge 2 [45]	2016	100 K(base set) 0 K (novel set)	20 K(base set) 1 K (novel set)	5/-/20	Coverage@P = 0.99	99.01%@P = 0.99 [137]	low-shot learning (6.3.1)
MS-Celeb-1 M Challenge 3 [163]	2018	274 K (ELFW) 1 M (DELFW)	5.7 K (ELFW) 1.58 M (DELFW)	-	1:1: TPR@FPR = 1e−9; 1:N: TPR@FPR = 1e−3	1:1: 46.15% [106]; 1:N: 43.88% [106]	trillion pairs; large distractors
MegaFace [44,164]	2016	1 M	690,572	1.4	1:1: TPR vs. FPR (ROC); 1:N: Rank-N (CMC)	1:1: 86.47%@10^{-6}FPR [38]; 1:N: 70.50% Rank-1 [38]	large-scale; 1 million distractors
IJB-A [41]	2015	25,809	500	51.6	1:1: TAR vs. FAR (ROC); 1:N: Rank-N, TPIR vs. FPIR (CMC, DET)	1:1: 92.10%@10^{-3}FAR [39]; 1:N: 98.20% Rank-1 [39]	cross-pose; template-based (6.1.1 and 6.3.2)
IJB-B [42]	2017	11,754 images 7,011 videos	1,845	41.6	1:1: TAR vs. FAR (ROC); 1:N: Rank-N, TPIR vs. FPIR (CMC, DET)	1:1: 52.12%@10^{-6}FAR [180]; 1:N: 90.20% Rank-1 [39]	cross-pose; template-based (6.1.1 and 6.3.2)
IJB-C [43]	2018	31.3 K images 11,779 videos	3,531	42.1	1:1: TAR vs. FAR (ROC); 1:N: Rank-N, TPIR vs. FPIR (CMC, DET)	1:1: 90.53%@10^{-6}FAR [180]; 1:N: 74.5% Rank-1 [71]	cross-pose; template-based (6.1.1 and 6.3.2)
RFW [173]	2018	40607	11429	3.6	1:1: Acc, TAR vs. FAR (ROC)	Caucasian: 92.15% Acc; Indian: 88.00% Acc; Asian: 83.98% Acc; African: 84.93% Acc [84]	evaluating race bias
DemogPairs [171]	2019	10.8 K	800	18	1:1: TAR vs. FAR (ROC)	White male: 88%; White female: 87% @10^{-4}FAR; Black male: 55%; Black female: 65%@10^{-4}FAR [84]	evaluating race and gender bias
CPLFW [181]	2017	11652	3968	2/2.9/3	1:1: Acc, TAR vs. FAR (ROC)	77.90% Acc [37]	cross-pose (6.1.1)
CFP [182]	2016	7,000	500	14	1:1: Acc, EER, AUC, TAR vs. FAR (ROC)	Frontal-Frontal: 98.67% Acc [135]; Frontal-Profile: 94:39% Acc [97]	frontal-profile (6.1.1)
SLLFW [183]	2017	13 K	5 K	2.3	1:1: Acc, TAR vs. FAR (ROC)	85.78% Acc [37]; 78.78% Acc [20]	fine-grained
UMDFaces [184]	2016	367,920	8,501	43.3	1:1: Acc, TPR vs. FPR (ROC)	69.30%@10^{-2}FAR [22]	annotation with bounding boxes, 21 keypoints, gender and 3D pose
YTF [169]	2011	3,425	1,595	48/181.3/ 6,070	1:1: Acc	97.30% Acc [37]; 96.52% Acc [185]	video (6.3.3)
PaSC [179]	2013	2,802	265	–	1:1: VR vs. FAR (ROC)	95.67%@10^{-2}FAR [185]	video (6.3.3)
YTC [186]	2008	1,910	47	–	1:N: Rank-N (CMC)	97.82% Rank-1 [185]; 97.32% Rank-1 [187]	video (6.3.3)
CALFW [188]	2017	12174	4025	2/3/4	1:1: Acc, TAR vs. FAR (ROC)	86.50% Acc [37]; 82.52% Acc [114]	cross-age; 12 to 81 years old (6.1.2)
MORPH [189]	2006	55,134	13,618	4.1	1:N: Rank-N (CMC)	94.4% Rank-1 [190]	cross-age, 16 to 77 years old (6.1.2)
CACD [191]	2014	163,446	2000	81.7	1:1 (CACD-VS): Acc, TAR vs. FAR (ROC); 1:N: MAP	1:1 (CACD-VS): 98.50% Acc [192]; 1:N: 69.96% MAP (2004–2006)[193]	cross-age, 14 to 62 years old (6.1.2)
FG-NET [194]	2010	1,002	82	12.2	1:N: Rank-N (CMC)	88.1% Rank-1 [192]	cross-age, 0 to 69 years old (6.1.2)
CASIA NIR-VIS v2.0 [195]	2013	17,580	725	24.2	1:1: Acc, VR vs. FAR (ROC)	98.62% Acc, 98.32%@10^{-3}FAR [196]	NIR-VIS; with eyeglasses, pose and expression variation (6.2.1)
CASIA-HFB [197]	2009	5097	202	25.5	1:1: Acc, VR vs. FAR (ROC)	97.58% Acc, 85.00%@10^{-3}FAR [198]	NIR-VIS; with eyeglasses and expression variation (6.2.1)
CUFS [199]	2009	1,212	606	2	1:N: Rank-N (CMC)	100% Rank-1 [200]	sketch-photo (6.2.3)
CUFSF [201]	2011	2,388	1,194	2	1:N: Rank-N (CMC)	51.00% Rank-1 [202]	sketch-photo; lighting variation; shape exaggeration (6.2.3)
Bosphorus [203]	2008	4,652	105	31/44.3/ 54	1:1: TAR vs. FAR (ROC); 1:N: Rank-N (CMC)	1:N: 99.20% Rank-1 [204]	3D; 34 expressions, 4 occlusions and different poses (6.4.1)
BU-3DFE [205]	2006	2,500	100	25	1:1: TAR vs. FAR (ROC); 1:N: Rank-N (CMC)	1:N: 95.00% Rank-1 [204]	3D; different expressions (6.4.1)
FRGCv2 [206]	2005	4,007	466	1/8.6/22	1:1: TAR vs. FAR (ROC); 1:N: Rank-N (CMC)	1:N: 94.80% Rank-1 [204]	3D; different expressions (6.4.1)

(*continued on next page*)

Table 12 (*continued*)

Datasets	Publish Time	#photos	#subjects	# of photos per subject [1]	Metrics	Typical Methods & Accuracy [2]	Key Features (Section)
Guo et al. [207]	2014	1,002	501	2	1:1: Acc, TAR vs. FAR (ROC)	94.8% Rank-1, 65.9%@10^{-3}FAR [208]	make-up; female (6.1.3)
FAM [209]	2013	1,038	519	2	1:1: Acc, TAR vs. FAR (ROC)	88.1% Rank-1, 52.6%@10^{-3}FAR [208]	make-up; female and male (6.1.3)
CASIA-FASD [210]	2012	600	50	12	EER, HTER	2.67% EER, 2.27% HTER [211]	anti-spoofing (6.4.4)
Replay-Attack [212]	2012	1,300	50	–	EER, HTER	0.79% EER, 0.72% HTER [211]	anti-spoofing (6.4.4)
WebCaricature [213]	2017	12,016	252	–	1:1: TAR vs. FAR (ROC); 1:N: Rank-N (CMC)	1:1: 34.94%@10^{-1}FAR [213]; 1:N: 55.41% Rank-1 [213]	Caricature (6.2.3)

The min/average/max numbers of photos or frames per subject.
We only present the typical methods that are published in a paper, and the accuracies of the most challenging scenarios are given.

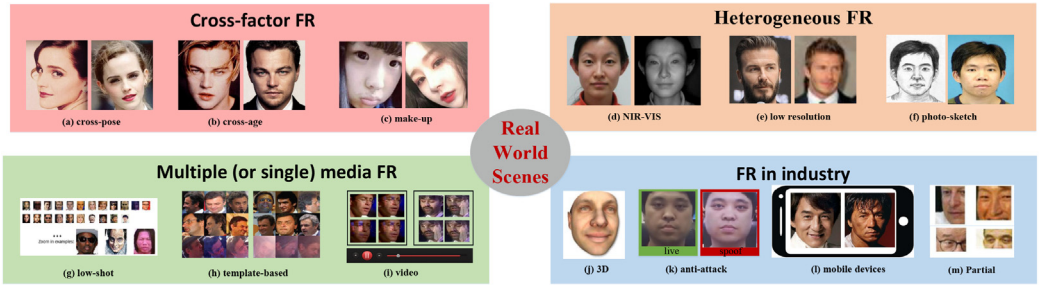

Fig. 21. The different scenes of FR. We divide FR scenes into four categories: cross-factor FR, heterogenous FR, multiple (or single) media FR and FR in industry. There are many testing datasets and special FR methods designed for each scene.

feature space rather than the image space. A deep residual equivariant mapping (DREAM) block dynamically added residuals to an input representation to transform a profile face to a frontal image. Chen et al. [216] proposed to combine feature extraction with multi-view subspace learning to simultaneously make features be more pose-robust and discriminative. Pose Invariant Model (PIM) [217] jointly performed face frontalization and learned pose invariant representations end-to-end to allow them to mutually boost each other, and further introduced unsupervised cross-domain adversarial training and a learning to learn strategy to provide high-fidelity frontal reference face images.

6.1.2. Cross-age face recognition

Cross-age FR is extremely challenging due to the changes in facial appearance by the aging process over time. One direct approach is to synthesize the desired image with target age such that the recognition can be performed in the same age group. A generative probabilistic model was used by [218] to model the facial aging process at each short-term stage. The identity-preserved conditional generative adversarial networks (IPCGANs) [219] framework utilized a conditional-GAN to generate a face in which an identity-preserved module preserved the identity information and an age classifier forced the generated face with the target age. Antipov et al. [220] proposed to age faces by GAN, but the synthetic faces cannot be directly used for face verification due to its imperfect preservation of identities. Then, they used a local manifold adaptation (LMA) approach [221] to solve the problem of [220]. In [222], high-level age-specific features conveyed by the synthesized face are estimated by a pyramidal adversarial discriminator at multiple scales to generate more lifelike facial details. An alternative to address the cross-age problem is to decompose aging and identity components separately and extract age-invariant representations. Wen et al. [192] developed a latent identity analysis (LIA) layer to separate these two components, as shown in Fig. 22. In [193], age-invariant features were obtained by subtracting age-specific factors from the representations with the help of the age estimation task. In [124], face features are decomposed in the spherical coordinate system, in which the identity-related components are represented with angular coordinates and the age-related information is encoded with radial coordinate. Additionally, there are other methods designed for cross-age FR. For example, Bianco ett al. [223] and El et al. [224] fine-tuned the CNN to transfer knowledge across age. Wang et al. [225] proposed a siamese deep network to perform multi-task learning of FR and age estimation. Li et al. [226] integrated feature extraction and metric learning via a deep CNN.

6.1.3. Makeup face recognition

Makeup is widely used by the public today, but it also brings challenges for FR due to significant facial appearance changes. The research on matching makeup and nonmakeup face images is receiving increasing attention. Li et al. [208] generated nonmakeup images from makeup ones by a bi-level adversarial network (BLAN) and then used the synthesized nonmakeup images for verification as shown in Fig. 23. Sun et al. [227] pretrained a triplet network on videos and fine-tuned it on a small makeup datasets. Specially, facial disguise [214,228,229] is a challenging research topic in makeup face recognition. By using disguise accessories such as wigs, beard, hats, mustache, and heavy makeup, disguise introduces two variations: (i) when a person wants to obfuscate his/her own identity, and (ii) another individual impersonates someone else's identity. Obfuscation increases intra-class

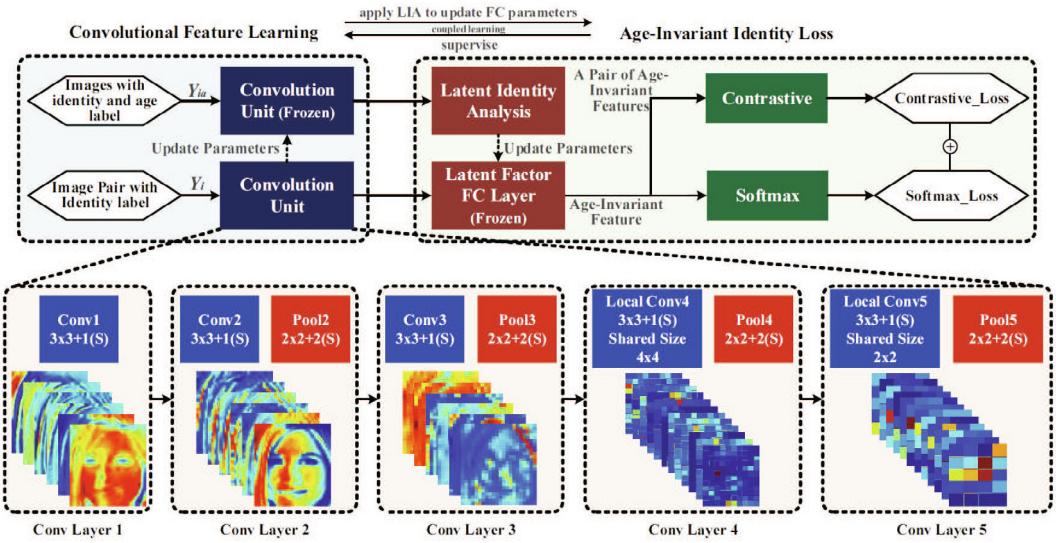

Fig. 22. The architecture of the cross-age FR with LIA. [192].

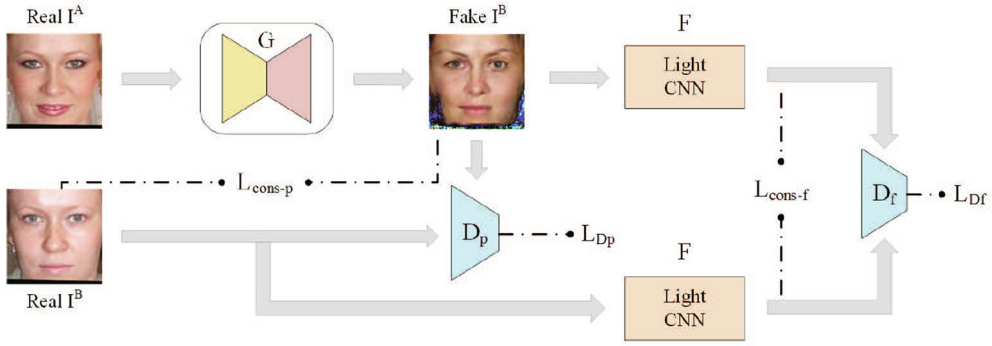

Fig. 23. The architecture of BLAN. [208].

variations whereas impersonation reduces the inter-class dissimilarity, thereby affecting face recognition/verification task. To address this issue, a variety of methods are proposed. Zhang et al. [230] first trained two DCNNs for generic face recognition and then used Principal Components Analysis (PCA) to find the transformation matrix for disguised face recognition adaptation. Kohli et al. [231] finetuned models using disguised faces. Smirnov et al. [232] proposed a hard example mining method benefitted from class-wise (Doppelganger Mining [233]) and example-wise mining to learn useful deep embeddings for disguised face recognition. Suri et al. [234] learned the representations of images in terms of colors, shapes, and textures (COST) using an unsupervised dictionary learning method, and utilized the combination of COST features and CNN features to perform recognition.

6.2. Heterogenous face recognition

.

6.2.1. NIR-VIS face recognition

Due to the excellent performance of the near-infrared spectrum (NIS) images under low-light scenarios, NIS images are widely applied in surveillance systems. Because most enrolled databases consist of visible light (VIS) spectrum images, how to recognize a NIR face from a gallery of VIS images has been a hot topic. Saxena et al. [235] and Liu et al. [236] transferred the VIS deep networks to the NIR domain by fine-tuning. Lezama et al. [237] used a VIS CNN to recognize NIR faces by transforming NIR images to VIS faces through cross-spectral hallucination and restoring a low-rank structure for features through low-rank embedding. Reale et al. [198] trained a VISNet (for visible images) and a NIRNet (for near-infrared images), and coupled their output features by creating a siamese network. He et al. [238,239] divided the high layer of the network into a NIR layer, a VIS layer and a NIR-VIS shared layer, then, a modality-invariant feature can be learned by the NIR-VIS shared layer. Song et al. [240] embedded cross-spectral face hallucination and discriminative feature learning into an end-to-end adversarial network. In [196], the low-rank relevance and cross-modal ranking were used to alleviate the semantic gap.

6.2.2. Low-resolution face recognition

Although deep networks are robust to low resolution to a great extent, there are still a few studies focused on promoting the performance of low-resolution FR. For example, Zangeneh et al. [241] proposed a CNN with a two-branch architecture (a super-resolution network and a feature extraction network) to map the high- and low-resolution face images into a common space where the intra-person distance is smaller than the inter-person distance. Shen et al. [242] exploited the face semantic information and local structural constraints to better restore the shape and detail of face images. In addition, they optimized the network with perceptual and adversarial losses to produce photo-realistic results.

6.2.3. Photo-sketch face recognition

The photo-sketch FR may help law enforcement to quickly identify suspects. The commonly used methods can be categorized as

two classes. One is to utilize transfer learning to directly match photos to sketches. Deep networks are first trained using a large face database of photos and are then fine-tuned using small sketch database [243,244]. The other is to use the image-to-image translation, where the photo can be transformed to a sketch or the sketch to a photo; then, FR can be performed in one domain. Zhang et al. [200] developed a fully convolutional network with generative loss and a discriminative regularizer to transform photos to sketches. Zhang et al. [245] utilized a branched fully convolutional neural network (BFCN) to generate a structure-preserved sketch and a texture-preserved sketch, and then they fused them together via a probabilistic method. Recently, GANs have achieved impressive results in image generation. Yi et al. [246], Kim et al. [247] and Zhu et al. [248] used two generators, G_A and G_B, to generate sketches from photos and photos from sketches, respectively (Fig. 24). Based on [248], Wang et al. [202] proposed a multi-adversarial network to avoid artifacts by leveraging the implicit presence of feature maps of different resolutions in the generator subnetwork. Similar to photo-sketch FR, photo-caricature FR is one kind of heterogenous FR scenes which is challenging and important to understanding of face perception. Huo et al. [213] built a large dataset of caricatures and photos, and provided several evaluation protocols and their baseline performances for comparison.

6.3. Multiple (or single) media face recognition

6.3.1. Low-shot face recognition

For many practical applications, such as surveillance and security, the FR system should recognize persons with a very limited number of training samples or even with only one sample. The methods of low-shot learning can be categorized as 1) synthesizing training data and 2) learning more powerful features. Hong et al. [249] generated images in various poses using a 3D face model and adopted deep domain adaptation to handle other variations, such as blur, occlusion, and expression (Fig. 25). Choe et al. [250] used data augmentation methods and a GAN for pose transition and attribute boosting to increase the size of the training dataset. Wu et al. [176] proposed a framework with hybrid classifiers using a CNN and a nearest neighbor (NN) model. Guo et al. [143] made the norms of the weight vectors of the one-shot classes and the normal classes aligned to address the data imbalance problem. Cheng et al. [137] proposed an enforced softmax that contains optimal dropout, selective attenuation, L2 normalization and model-level optimization. Yin et al. [251] augmented feature space of low-shot classes by transferring the principal components from regular to low-shot classes to encourage the variance of low-shot classes to mimic that of regular classes.

6.3.2. Set/template-based face recognition

Different from traditional image-to-image recognition, set-to-set recognition takes a set (heterogeneous contents containing both images and videos) as the smallest unit of representation. This kind of setting does reflect the real-world biometric scenarios, thereby attracting a lot of attention. After learning face representations of media in each set, two strategies are generally adopted to perform set-to-set matching. One is to use these representations to perform pair-wise similarity comparison of two sets and aggregate the results into a single and final score by max score pooling [96], average score pooling [252] and its variations [253,254]. The other strategy is feature pooling [96,103,81] which first aggregates face representations into a single representation for each set and then performs a comparison between two sets. In addition to the commonly used strategies, there are also some novel methods proposed for set/template-based FR. For example, Hayat et al. [255] proposed a deep heterogeneous feature fusion network to exploit the features' complementary information generated by different CNNs. Liu et al. [256] introduced the actor-critic reinforcement learning for set-based FR. They casted the inner-set dependency modeling to a Markov decision process in the latent space, and trained a dependency-aware attention control agent to make attention control for each image in each step.

6.3.3. Video face recognition

There are two key issues in video FR: one is to integrate the information across different frames together to build a representation of the video face, and the other is to handle video frames with severe blur, pose variations, and occlusions. For frame aggregation, Yang et al. [83] proposed a neural aggregation network (NAN) in which the aggregation module, consisting of two attention blocks driven by a memory, produces a 128-dimensional vector representation (Fig. 26). Rao et al. [187] aggregated raw video frames directly by combining the idea of metric learning and adversarial learning. For dealing with bad frames, Rao et al. [185] discarded the bad frames by treating this operation as a Markov decision process and trained the attention model through a deep reinforcement learning framework. Ding et al. [257] artificially blurred clear images for training to learn blur-robust face representations. Parchami et al. [258] used a CNN to reconstruct a lower-quality video into a high-quality face.

6.4. Face recognition in industry

6.4.1. 3D face recognition

3D FR has inherent advantages over 2D methods, but 3D deep FR is not well developed due to the lack of large annotated 3D data. To enlarge 3D training datasets, most works use the methods of "one-to-many augmentation" to synthesize 3D faces. However, the effective methods for extracting deep features of 3D faces remain to be explored. Kim et al. [204] fine-tuned a 2D CNN with a small amount of 3D scans for 3D FR. Zulqarnain et al. [259] used a three-channel (corresponding to depth, azimuth and elevation angles of the normal vector) image as input and minimized the average prediction log-loss. Zhang et al. [260] first selected 30 feature points from the Candide-3 face model to characterize faces, then conducted the unsupervised pretraining of face depth data, and finally performed the supervised fine-tuning.

6.4.2. Partial face recognition

Partial FR, in which only arbitrary-size face patches are presented, has become an emerging problem with increasing requirements of identification from CCTV cameras and embedded vision systems in mobile devices, robots and smart home facilities. He et al. [261] divided the aligned face image into several multi-scale patches, and the dissimilarity between two partial face images is calculated as the weighted L2 distance between corresponding patches. Dynamic feature matching (DFM) [262] utilized a sliding window of the same size as the probe feature maps to decompose the gallery feature maps into several gallery sub-feature maps, and the similarity-guided constraint imposed on sparse representation classification (SRC) provides an alignment-free matching.

6.4.3. Face recognition for mobile devices

With the emergence of mobile phones, tablets and augmented reality, FR has been applied in mobile devices. Due to computational limitations, the recognition tasks in these devices need to be carried out in a light but timely fashion. MobiFace [87] required efficient memory and low cost operators by adopting fast down-sampling and bottleneck residual block, and achieves 99.7% on LFW database and 91.3% on Megaface database. Tadmor et al. [263] proposed a multibatch method that first generates signatures

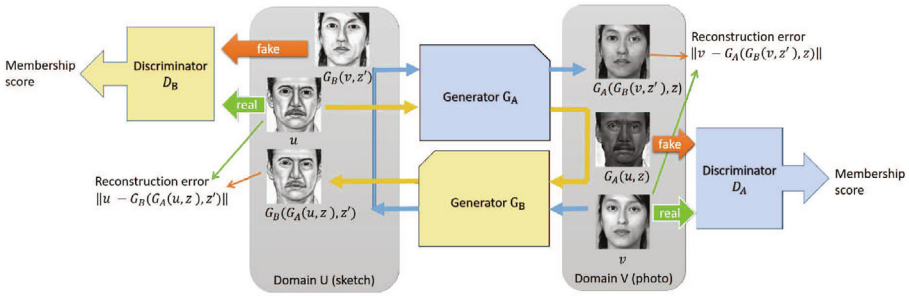

Fig. 24. The architecture of DualGAN. [246].

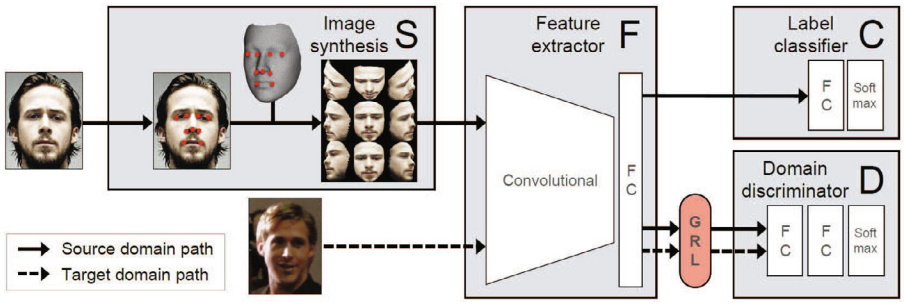

Fig. 25. The architecture of a single sample per person domain adaptation network (SSPP-DAN). [249].

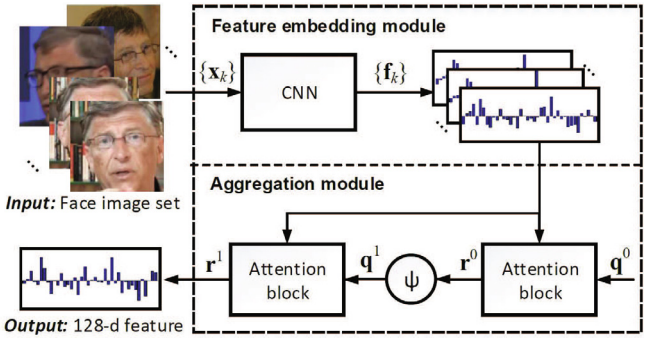

Fig. 26. The FR framework of NAN. [83].

for a minibatch of k face images and then constructs an unbiased estimate of the full gradient by relying on all $k^2 - k$ pairs from the minibatch. As mentioned in Section 3.2.1, light-weight deep networks [126–129] perform excellently in the fundamental tasks of image classification and deserve further attention in FR tasks. Moreover, some well-known compressed networks such as Pruning [264–266], BinaryNets [267–270], Mimic Networks [271,272], also have potential to be introduced into FR.

6.4.4. Face anti-attack

With the success of FR techniques, various types of attacks, such as face spoofing and adversarial perturbations, are becoming large threats. Face spoofing involves presenting a fake face to the biometric sensor using a printed photograph, worn mask, or even an image displayed on another electronic device. In order to defense this type of attack, several methods are proposed [211,273–279]. Atoum et al. [211] proposed a novel two-stream CNN in which the local features discriminate the spoof patches that are independent of the spatial face areas, and holistic depth maps ensure that the input live sample has a face-like depth. Yang et al. [273] trained a CNN using both a single frame and multiple frames with five scales as input, and using the live/spoof label as the output. Taken the sequence of video frames as input, Xu et al. [274] applied LSTM units on top of CNN to obtain end-to-end features to recognize spoofing faces which leveraged the local and dense property from convolution operation and learned the temporal structure using LSTM units. Li et al. [275] and Patel et al. [276] fine-tuned their networks from a pretrained model by training sets of real and fake images. Jourabloo et al. [277] proposed to inversely decompose a spoof face into the live face and the spoof noise pattern. Adversarial perturbation is the other type of attack which can be defined as the addition of a minimal vector r such that with addition of this vector into the input image x, i.e. $(x + r)$, the deep learning models misclassifies the input while people will not. Recently, more and more work has begun to focus on solving this perturbation of FR. Goswami et al. [280] proposed to detect adversarial samples by characterizing abnormal filter response behavior in the hidden layers and increase the network's robustness by removing the most problematic filters. Goel et al. [281] provided an open source implementation of adversarial detection and mitigation algorithms. Despite of progresses of anti-attack algorithms, attack methods are updated as well and remind us the need to further increase security and robustness in FR systems, for example, Mai et al. [282] proposed a neighborly de-convolutional neural network (NbNet) to reconstruct a fake face using the stolen deep templates.

6.4.5. Debiasing face recognition

As described in Section 5.1, existing datasets are highly biased in terms of the distribution of demographic cohorts, which may dramatically impact the fairness of deep models. To address this issue, there are some works that seek to introduce fairness into face recognition and mitigate demographic bias, e,g. unbalanced-training [283], attribute removal [284–286] and domain adaptation [173,287,147]. 1) Unbalanced-training methods mitigate the bias via model regularization, taking into consideration of the fairness goal in the overall model objective function. For example, RL-RBN [283] formulated the process of finding the optimal margins for non-Caucasians as a Markov decision process and employed

deep Q-learning to learn policies based on large margin loss. 2) Attribute removal methods confound or remove demographic information of faces to learn attribute-invariant representations. For example, Alvi et al. [284] applied a confusion loss to make a classifier fail to distinguish attributes of examples so that multiple spurious variations are removed from the feature representation. SensitiveNets [288] proposed to introduce sensitive information into triplet loss. They minimized the sensitive information, while maintaining distances between positive and negative embeddings. 3) Domain adaptation methods propose to investigate data bias problem from a domain adaptation point of view and attempt to design domain-invariant feature representations to mitigate bias across domains. IMAN [173] simultaneously aligned global distribution to decrease race gap at domain-level, and learned the discriminative target representations at cluster level. Kan [147] directly converted the Caucasian data to non-Caucasian domain in the image space with the help of sparse reconstruction coefficients learnt in the common subspace.

7. Technical challenges

In this paper, we provide a comprehensive survey of deep FR from both data and algorithm aspects. For algorithms, mainstream and special network architectures are presented. Meanwhile, we categorize loss functions into Euclidean-distance-based loss, angular/cosine-margin-based loss and variable softmax loss. For data, we summarize some commonly used datasets. Moreover, the methods of face processing are introduced and categorized as "one-to-many augmentation" and "many-to-one normalization". Finally, the special scenes of deep FR, including video FR, 3D FR and cross-age FR, are briefly introduced.

Taking advantage of big annotated data and revolutionary deep learning techniques, deep FR has dramatically improved the SOTA performance and fostered successful real-world applications. With the practical and commercial use of this technology, many ideal assumptions of academic research were broken, and more real-world issues are emerging. To the best our knowledge, major technical challenges include the following aspects.

- **Security issues.** Presentation attack [289], adversarial attack [280,281,290], template attack [291] and digital manipulation attack [292,293] are developing to threaten the security of deep face recognition systems. 1) Presentation attack with 3D silicone mask, which exhibits skin-like appearance and facial motion, challenges current anti-sproofing methods [294]. 2) Although adversarial perturbation detection and mitigation methods are recently proposed [280,281], the root cause of adversarial vulnerability is unclear and thus new types of adversarial attacks are still upgraded continuously [295,296]. 3) The stolen deep feature template can be used to recover its facial appearance, and how to generate cancelable template without loss of accuracy is another important issue. 4) Digital manipulation attack, made feasible by GANs, can generate entirely or partially modified photorealistic faces by expression swap, identity swap, attribute manipulation and entire face synthesis, which remains a main challenge for the security of deep FR.

- **Privacy-preserving face recognition.** With the leakage of biological data, privacy concerns are raising nowadays. Facial images can predict not only demographic information such as gender, age, or race, but even the genetic information [297]. Recently, the pioneer works such as Semi-Adversarial Networks [298,299,285] have explored to generate a recognizable biometric templates that can hidden some of the private information presented in the facial images. Further research on the principles of visual cryptography, signal mixing and image perturbation to protect users' privacy on stored face templates are essential for addressing public concern on privacy.

- **Understanding deep face recognition.** Deep face recognition systems are now believed to surpass human performance in most scenarios [300]. There are also some interesting attempts to apply deep models to assist human operators for face verification [183,300]. Despite this progress, many fundamental questions are still open, such as what is the "identity capacity" of a deep representation [301]? Why deep neural networks, rather than humans, are easily fooled by adversarial samples? While bigger and bigger training dataset by itself cannot solve this problem, deeper understanding on these questions may help us to build robust applications in real world. Recently, a new benchmark called TALFW has been proposed to explore this issue [93].

- **Remaining challenges defined by non-saturated benchmark datasets.** Three current major datasets, namely, MegaFace [44,164], MS-Celeb-1 M [45] and IJB-A/B/C [41–43], are corresponding to large-scale FR with a very large number of candidates, low/one-shot FR and large pose-variance FR which will be the focus of research in the future. Although the SOTA algorithms can be over 99.9 percent accurate on LFW [23] and Megaface [44,164] databases, fundamental challenges such as matching faces cross ages [181], poses [188], sensors, or styles still remain. For both datasets and algorithms, it is necessary to measure and address the racial/gender/age biases of deep FR in future research.

- **Ubiquitous face recognition across applications and scenes.** Deep face recognition has been successfully applied on many user-cooperated applications, but the ubiquitous recognition applications in everywhere are still an ambitious goal. In practice, it is difficult to collect and label sufficient samples for innumerable scenes in real world. One promising solution is to first learn a general model and then transfer it to an application-specific scene. While deep domain adaptation [145] has recently been applied to reduce the algorithm bias on different scenes [148], different races [173], general solution to transfer face recognition is largely open.

- **Pursuit of extreme accuracy and efficiency.** Many killer-applications, such as watch-list surveillance or financial identity verification, require high matching accuracy at very low alarm rate, e.g. 10^{-9}. It is still a big challenge even with deep learning on massive training data. Meanwhile, deploying deep face recognition on mobile devices pursues the minimum size of feature representation and compressed deep network. It is of great significance for both industry and academic to explore this extreme face-recognition performance beyond human imagination. It is also exciting to constantly push the performance limits of the algorithm after it has already surpassed human.

- **Fusion issues.** Face recognition by itself is far from sufficient to solve all biometric and forensic tasks, such as distinguishing identical twins and matching faces before and after surgery [302]. A reliable solution is to consolidate multiple sources of biometric evidence [303]. These sources of information may correspond to different biometric traits (e.g., face + hand [304]), sensors (e.g., 2D + 3D face cameras), feature extraction and matching techniques, or instances (e.g., a face sequence of

various poses). It is beneficial for face biometric and forensic applications to perform information fusion at the data level, feature level, score level, rank level, and decision level [305].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially supported by National Key R&D Program of China (2019YFB1406504) and BUPT Excellent Ph.D. Students Foundation CX2020207.

References

[1] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cognitive Neurosci. 3 (1) (1991) 71–86.
[2] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.
[3] B. Moghaddam, W. Wahid, and A. Pentland, "Beyond eigenfaces: probabilistic matching for face recognition," Automatic Face and Gesture Recognition, 1998. Proc. Third IEEE Int. Conf., pp. 30–35, Apr 1998.
[4] W. Deng, J. Hu, J. Lu, J. Guo, Transform-invariant pca: A unified approach to fully automatic facealignment, representation, and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 36 (6) (June 2014) 1275–1284.
[5] X. He, S. Yan, Y. Hu, P. Niyogi, H.-J. Zhang, Face recognition using laplacianfaces, IEEE Trans. Pattern Anal. Mach. Intell. 27 (3) (2005) 328–340.
[6] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, "Graph embedding: A general framework for dimensionality reduction," Computer Vision and Pattern Recognition, IEEE Computer Society Conference on, vol. 2, pp. 830–837, 2005.
[7] W. Deng, J. Hu, J. Guo, H. Zhang, C. Zhang, Comments on "globally maximizing, locally minimizing: Unsupervised discriminant projection with applications to face and palm biometrics", IEEE Trans. Pattern Anal. Mach. Intell. 30 (8) (2008) 1503–1504.
[8] J. Wright, A. Yang, A. Ganesh, S. Sastry, Y. Ma, Robust Face Recognition via Sparse Representation, IEEE Trans. Pattern Anal. Machine Intell. 31 (2) (2009) 210–227.
[9] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in 2011 International conference on computer vision. IEEE, 2011, pp. 471–478.
[10] W. Deng, J. Hu, J. Guo, Extended src: Undersampled face recognition via intraclass variant dictionary, IEEE Trans. Pattern Anal. Machine Intell. 34 (9) (2012) 1864–1870.
[11] W. Deng, J. Hu, and J. Guo, "Face recognition via collaborative representation: Its discriminant nature and superposed representation," IEEE Trans. Pattern Anal. Mach. Intell., vol. PP, no. 99, pp. 1–1, 2018.
[12] C. Liu, H. Wechsler, Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition, Image processing, IEEE Transactions on 11 (4) (2002) 467–476.
[13] T. Ahonen, A. Hadid, M. Pietikainen, Face description with local binary patterns: Application to face recognition, IEEE Trans. Pattern Anal. Machine Intell. 28 (12) (2006) 2037–2041.
[14] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition," in ICCV, vol. 1. IEEE, 2005, pp. 786–791.
[15] D. Chen, X. Cao, F. Wen, J. Sun, Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 3025–3032.
[16] W. Deng, J. Hu, and J. Guo, "Compressive binary patterns: Designing a robust binary face descriptor with random-field eigenfilters," IEEE Trans. Pattern Anal. Mach. Intell., vol. PP, no. 99, pp. 1–1, 2018.
[17] Z. Cao, Q. Yin, X. Tang, and J. Sun, "Face recognition with learning-based descriptor," in 2010 IEEE Computer society conference on computer vision and pattern recognition. IEEE, 2010, pp. 2707–2714.
[18] Z. Lei, M. Pietikainen, S.Z. Li, Learning discriminant face descriptor, IEEE Trans. Pattern Anal. Machine Intell. 36 (2) (2014) 289–302.
[19] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, Y. Ma, Pcanet: A simple deep learning baseline for image classification?, IEEE Trans Image Process. 24 (12) (2015) 5017–5032.
[20] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: Closing the gap to human-level performance in face verification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1701–1708.
[21] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in Advances in neural information processing systems, 2014, pp. 1988–1996.
[22] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.
[23] G.B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Technical Report 07-49, University of Massachusetts, Amherst, Tech. Rep., 2007.
[24] W. Zhao, R. Chellappa, P.J. Phillips, A. Rosenfeld, Face recognition: A literature survey, ACM Computing Surveys (CSUR) 35 (4) (2003) 399–458.
[25] K.W. Bowyer, K. Chang, P. Flynn, A survey of approaches and challenges in 3d and multi-modal 3d+ 2d face recognition, Computer Vision Image Understanding 101 (1) (2006) 1–15.
[26] A.F. Abate, M. Nappi, D. Riccio, G. Sabatino, 2d and 3d face recognition: A survey, Pattern Recogn. Lett. 28 (14) (2007) 1885–1906.
[27] R. Jafri, H.R. Arabnia, A survey of face recognition techniques, J. Inform. Process. Syst. 5 (2) (2009) 41–68.
[28] A. Scheenstra, A. Ruifrok, and R.C. Veltkamp, "A survey of 3d face recognition methods," in International Conference on Audio-and Video-based Biometric Person Authentication. Springer, 2005, pp. 891–899.
[29] X. Zou, J. Kittler, and K. Messer, "Illumination invariant face recognition: A survey," in 2007 first IEEE international conference on biometrics: theory, applications, and systems. IEEE, 2007, pp. 1–8.
[30] X. Zhang, Y. Gao, Face recognition across pose: A review, Pattern Recogn. 42 (11) (2009) 2876–2896.
[31] C. Ding, D. Tao, A comprehensive survey on pose-invariant face recognition, ACM Trans. Intelligent Systems Technol. 7 (3) (2015) 37.
[32] R. Ranjan, S. Sankaranarayanan, A. Bansal, N. Bodla, J.C. Chen, V.M. Patel, C.D. Castillo, R. Chellappa, Deep learning for understanding faces: Machines may be just as good, or better, than humans, IEEE Signal Process. Mag. 35 (1) (2018) 66–83.
[33] X. Jin, X. Tan, Face alignment in-the-wild: A survey, Comput. Vis. Image Underst. 162 (2017) 1–22.
[34] Y. Sun, X. Wang, X. Tang, Deep learning face representation from predicting 10,000 classes, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1891–1898.
[35] Y. Sun, X. Wang, X. Tang, Deeply learned face representations are sparse, selective, and robust, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2892–2900.
[36] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," arXiv preprint arXiv:1502.00873, 2015.
[37] O.M. Parkhi, A. Vedaldi, A. Zisserman et al., "Deep face recognition." in BMVC, vol. 1, no. 3, 2015, p. 6.
[38] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.
[39] Q. Cao, L. Shen, W. Xie, O.M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018, pp. 67–74.
[40] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in neural information processing systems, 2014, pp. 2672–2680.
[41] B.F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, A.K. Jain, Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1931–1939.
[42] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J.A. Duncan, K. Allen et al., "Iarpa janus benchmark-b face dataset," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 90–98.
[43] 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 2018 International Conference on Biometrics (ICB). IEEE, 2018, pp. 158–165.
[44] I. Kemelmacher-Shlizerman, S.M. Seitz, D. Miller, E. Brossard, The megaface benchmark: 1 million faces for recognition at scale, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4873–4882.
[45] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in European conference on computer vision. Springer, 2016, pp. 87–102.
[46] M. Mehdipour Ghazi, H. Kemal Ekenel, A comprehensive analysis of deep learning based representation for face recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2016, pp. 34–41.
[47] I. Masi, A.T. Tr?n, T. Hassner, J.T. Leksut, G. Medioni, "Do we really need to collect millions of faces for effective face recognition?" in ECCV. Springer, 2016, pp. 579–596.
[48] I. Masi, T. Hassner, A.T. Tran, G. Medioni, "Rapid synthesis of massive face sets for improved face recognition," in 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, 2017, pp. 604–611.

[49] E. Richardson, M. Sela, and R. Kimmel, "3d face reconstruction by learning from synthetic data," in 2016 fourth international conference on 3D vision (3DV). IEEE, 2016, pp. 460–469.
[50] E. Richardson, M. Sela, R. Or-El, R. Kimmel, Learning detailed face reconstruction from a single image, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1259–1268.
[51] P. Dou, S.K. Shah, I.A. Kakadiaris, End-to-end 3d face reconstruction with deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5908–5917.
[52] Y. Guo, J. Zhang, J. Cai, B. Jiang, and J. Zheng, "3dfacenet: Real-time dense face reconstruction via synthesizing photo-realistic face images," arXiv preprint arXiv:1708.00980, 2017.
[53] A. Tuan Tran, T. Hassner, I. Masi, G. Medioni, Regressing robust and discriminative 3d morphable models with a very deep neural network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5163–5172.
[54] A. Tewari, M. Zollhofer, H. Kim, P. Garrido, F. Bernard, P. Perez, C. Theobalt, Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 1274–1283.
[55] Z. Zhu, P. Luo, X. Wang, X. Tang, Multi-view perceptron: a deep model for learning face identity and view representations, Advances in Neural Information Processing Systems (2014) 217–225.
[56] J. Zhao, L. Xiong, P.K. Jayashree, J. Li, F. Zhao, Z. Wang, P.S. Pranata, P.S. Shen, S. Yan, and J. Feng, "Dual-agent gans for photorealistic and identity preserving profile face synthesis," in Advances in neural information processing systems, 2017, pp. 66–76.
[57] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, R. Webb, Learning from simulated and unsupervised images through adversarial training, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2107–2116.
[58] J. Liu, Y. Deng, T. Bai, Z. Wei, and C. Huang, "Targeting ultimate accuracy: Face recognition via deep embedding," arXiv preprint arXiv:1506.07310, 2015.
[59] E. Zhou, Z. Cao, and Q. Yin, "Naive-deep face recognition: Touching the limit of lfw benchmark or not?" arXiv preprint arXiv:1501.04690, 2015.
[60] C. Ding, D. Tao, Robust face recognition via multimodal deep face representation, IEEE Trans. Multimedia 17 (11) (2015) 2049–2058.
[61] Y. Sun, X. Wang, X. Tang, Sparsifying neural network connections for face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4856–4864.
[62] D. Wang, C. Otto, A.K. Jain, Face search at scale, IEEE Trans. Pattern Analysis Machine Intell. 39 (6) (2016) 1122–1136.
[63] M. Kan, S. Shan, H. Chang, X. Chen, Stacked progressive auto-encoders (spae) for face recognition across poses, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1883–1890.
[64] Y. Zhang, M. Shao, E.K. Wong, Y. Fu, Random faces guided sparse many-to-one encoder for pose-invariant face recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2416–2423.
[65] J. Yang, S.E. Reed, M.-H. Yang, H. Lee, Weakly-supervised disentangling with recurrent transformations for 3d view synthesis, Adv. Neural Inform. Process. Syst. (2015) 1099–1107.
[66] Z. Zhu, P. Luo, X. Wang, X. Tang, Deep learning identity-preserving face space, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 113–120.
[67] Z. Zhu, P. Luo, X. Wang, X. Tang, "Recover canonical-view faces in the wild with deep neural networks," arXiv preprint arXiv:1404.3543, 2014.
[68] L. Hu, M. Kan, S. Shan, X. Song, X. Chen, "Ldf-net: Learning a displacement field network for face recognition across pose," in 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, 2017, pp. 9–16.
[69] E. Zhou, Z. Cao, J. Sun, "Gridface: Face rectification via learning local homography transformations," in The European Conference on Computer Vision (ECCV), September 2018.
[70] R. Huang, S. Zhang, T. Li, R. He, Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2439–2448.
[71] L. Tran, X. Yin, X. Liu, Disentangled representation learning gan for pose-invariant face recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1415–1424.
[72] J. Deng, S. Cheng, N. Xue, Y. Zhou, S. Zafeiriou, Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7093–7102.
[73] X. Yin, X. Yu, K. Sohn, X. Liu, M. Chandraker, Towards large-pose face frontalization in the wild, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 3990–3999.
[74] 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, Int. J. Comput. Vision 115 (3) (2015) 211–252.
[75] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
[76] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.
[77] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[78] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
[79] G. Hu, Y. Yang, D. Yi, J. Kittler, W. Christmas, S.Z. Li, and T. Hospedales, "When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition," in ICCV workshops, 2015, pp. 142–150.
[80] S. Sankaranarayanan, A. Alavi, and R. Chellappa, "Triplet similarity embedding for face verification," arXiv preprint arXiv:1602.03418, 2016.
[81] S. Sankaranarayanan, A. Alavi, C.D. Castillo, and R. Chellappa, "Triplet probabilistic embedding for face verification and clustering," in 2016 IEEE 8th international conference on biometrics theory, applications and systems (BTAS). IEEE, 2016, pp. 1–8.
[82] X. Zhang, Z. Fang, Y. Wen, Z. Li, Y. Qiao, Range loss for deep face recognition with long-tailed training data, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5409–5418.
[83] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, G. Hua, Neural aggregation network for video face recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4362–4371.
[84] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, L. Song, Sphereface: Deep hypersphere embedding for face recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 212–220.
[85] X. Wu, R. He, Z. Sun, T. Tan, A light cnn for deep face representation with noisy labels, IEEE Trans. Inf. Forensics Secur. 13 (11) (2018) 2884–2896.
[86] X. Wu, R. He, and Z. Sun, "A lightened cnn for deep face representation," in CVPR, vol. 4, 2015.
[87] C.N. Duong, K.G. Quach, N. Le, N. Nguyen, K. Luu, Mobiface: A lightweight deep learning face recognition on mobile devices, 2018, arXiv preprint arXiv:1811.11080.
[88] N. Zhu, Z. Yu, and C. Kou, "A new deep neural architecture search pipeline for face recognition," IEEE Access, vol. 8, pp. 91 303–91 310, 2020.
[89] C. Xiong, X. Zhao, D. Tang, K. Jayashree, S. Yan, T.-K. Kim, Conditional convolutional neural network for modality-aware face recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3667–3675.
[90] C. Han, S. Shan, M. Kan, S. Wu, and X. Chen, "Face recognition with contrastive convolution," in The European Conference on Computer Vision (ECCV), September 2018.
[91] M. Hayat, S.H. Khan, N. Werghi, R. Goecke, Joint registration and representation learning for unconstrained face identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2767–2776.
[92] W. Wu, M. Kan, X. Liu, Y. Yang, S. Shan, X. Chen, Recursive spatial transformer (rest) for alignment-free face recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3772–3780.
[93] Y. Zhong, J. Chen, B. Huang, Toward end-to-end face recognition through alignment learning, IEEE signal processing letters 24 (8) (2017) 1213–1217.
[94] J.-C. Chen, R. Ranjan, A. Kumar, C.-H. Chen, V.M. Patel, R. Chellappa, An end-to-end system for unconstrained face verification with deep convolutional neural networks, in: Proceedings of the IEEE international conference on computer vision workshops, 2015, pp. 118–126.
[95] M. Kan, S. Shan, X. Chen, Multi-view deep network for cross-view classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4847–4855.
[96] I. Masi, S. Rawls, G. Medioni, P. Natarajan, Pose-aware face recognition in the wild, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4838–4846.
[97] X. Yin, X. Liu, Multi-task convolutional neural network for pose-invariant face recognition, IEEE Trans. Image Process. 27 (2) (2017) 964–975.
[98] W. Wang, Z. Cui, H. Chang, S. Shan, and X. Chen, "Deeply coupled auto-encoder networks for cross-view classification," arXiv preprint arXiv:1402.2031, 2014.
[99] Y. Sun, X. Wang, X. Tang, Hybrid deep learning for face verification, in: Proceedings of the IEEE international conference on computer vision, 2013, pp. 1489–1496.
[100] R. Ranjan, S. Sankaranarayanan, C.D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," in 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, 2017, pp. 17–24.
[101] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in European conference on computer vision. Springer, 2016, pp. 499–515.
[102] Y. Wu, H. Liu, J. Li, Y. Fu, Deep face recognition with center invariant loss, in: Proceedings of the on Thematic Workshops of ACM Multimedia 2017 ACM, 2017, pp. 408–414.
[103] J.-C. Chen, V.M. Patel, and R. Chellappa, "Unconstrained face verification using deep cnn features," in 2016 IEEE winter conference on applications of computer vision (WACV). IEEE, 2016, pp. 1–9.
[104] W. Liu, Y. Wen, Z. Yu, M. Yang, Large-margin softmax loss for convolutional neural networks, ICML 2 (3) (2016) 7.
[105] F. Wang, J. Cheng, W. Liu, H. Liu, Additive margin softmax for face verification, IEEE Signal Process. Lett. 25 (7) (2018) 926–930.

[106] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4690–4699.
[107] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, W. Liu, Cosface: Large margin cosine loss for deep face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5265–5274.
[108] W. Liu, Y.-M. Zhang, X. Li, Z. Yu, B. Dai, T. Zhao, and L. Song, "Deep hyperspherical learning," in Advances in neural information processing systems, 2017, pp. 3950–3960.
[109] R. Ranjan, C.D. Castillo, and R. Chellappa, "L2-constrained softmax loss for discriminative face verification," arXiv preprint arXiv:1703.09507, 2017.
[110] F. Wang, X. Xiang, J. Cheng, A.L. Yuille, Normface: L2 hypersphere embedding for face verification, in: Proceedings of the 25th ACM international conference on Multimedia, 2017, pp. 1041–1049.
[111] A. Hasnat, J. Bohné, J. Milgram, S. Gentric, L. Chen, Deepvisage: Making face recognition simple yet with powerful generalization skills, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 1682–1691.
[112] Y. Liu, H. Li, and X. Wang, "Rethinking feature discrimination and polymerization for large-scale recognition," arXiv preprint arXiv:1710.00870, 2017.
[113] X. Qi and L. Zhang, "Face recognition via centralized coordinate learning," arXiv preprint arXiv:1801.05678, 2018.
[114] B. Chen, W. Deng, J. Du, Noisy softmax: Improving the generalization ability of dcnn via postponing the early softmax saturation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5372–5381.
[115] M. Hasnat, J. Bohné, J. Milgram, S. Gentric, L. Chen et al., "von mises-fisher mixture model-based deep learning: Application to face verification," arXiv preprint arXiv:1706.04264, 2017.
[116] J. Deng, Y. Zhou, S. Zafeiriou, Marginal loss for deep face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 60–68.
[117] Y. Zheng, D.K. Pal, M. Savvides, Ring loss: Convex feature normalization for face recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5089–5097.
[118] E.P. Xing, M.I. Jordan, S.J. Russell, and A.Y. Ng, "Distance metric learning with application to clustering with side-information," in Advances in neural information processing systems, 2003, pp. 521–528.
[119] K.Q. Weinberger and L.K. Saul, "Distance metric learning for large margin nearest neighbor classification," Journal of Machine Learning Research, vol. 10, no. Feb, pp. 207–244, 2009.
[120] D. Yi, Z. Lei, S. Liao, and S.Z. Li, "Learning face representation from scratch," arXiv preprint arXiv:1411.7923, 2014.
[121] N. Crosswhite, J. Byrne, C. Stauffer, O. Parkhi, Q. Cao, A. Zisserman, Template adaptation for face verification and identification, Elsevier 79 (2018) 35–48.
[122] B. Liu, W. Deng, Y. Zhong, M. Wang, J. Hu, X. Tao, Y. Huang, Fair loss: Margin-aware reinforcement learning for deep face recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019.
[123] H. Liu, X. Zhu, Z. Lei, and S.Z. Li, "Adaptiveface: Adaptive margin and sampling for face recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.
[124] F. Wang, L. Chen, C. Li, S. Huang, Y. Chen, C. Qian, C. Change Loy, The devil of face recognition is in the noise, The European Conference on Computer Vision (ECCV) (September 2018).
[125] C.J. Parde, C. Castillo, M.Q. Hill, Y.I. Colon, S. Sankaranarayanan, J.-C. Chen, and A.J. O'Toole, "Deep convolutional neural network features and the original image," arXiv preprint arXiv:1611.01751, 2016.
[126] F.N. Iandola, S. Han, M.W. Moskewicz, K. Ashraf, W.J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and 0.5 mb model size," arXiv preprint arXiv:1602.07360, 2016.
[127] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017, arXiv preprint arXiv:1704.04861.
[128] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.
[129] X. Zhang, X. Zhou, M. Lin, J. Sun, Shufflenet: An extremely efficient convolutional neural network for mobile devices, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6848–6856.
[130] B. Zoph and Q.V. Le, "Neural architecture search with reinforcement learning," arXiv preprint arXiv:1611.01578, 2016.
[131] E. Real, A. Aggarwal, Y. Huang, and Q.V. Le, "Aging evolution for image classifier architecture search," in AAAI Conference on Artificial Intelligence, 2019.
[132] L.-C. Chen, M. Collins, Y. Zhu, G. Papandreou, B. Zoph, F. Schroff, H. Adam, and J. Shlens, "Searching for efficient multi-scale architectures for dense image prediction," in Advances in neural information processing systems, 2018, pp. 8699–8710.
[133] V. Mnih, K. Kavukcuoglu, D. Silver, A.A. Rusu, J. Veness, M.G. Bellemare, A. Graves, M. Riedmiller, A.K. Fidjeland, G. Ostrovski et al., "Human-level control through deep reinforcement learning," nature, vol. 518, no. 7540, pp. 529–533, 2015.
[134] M. Jaderberg, K. Simonyan, A. Zisserman et al., "Spatial transformer networks," in Advances in neural information processing systems, 2015, pp. 2017–2025.
[135] X. Peng, X. Yu, K. Sohn, D.N. Metaxas, M. Chandraker, Reconstruction-based disentanglement for pose-invariant face recognition, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 1623–1632.
[136] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun, "Bayesian face revisited: A joint formulation," in European conference on computer vision. Springer, 2012, pp. 566–579.
[137] Y. Cheng, J. Zhao, Z. Wang, Y. Xu, K. Jayashree, S. Shen, J. Feng, Know you at one glance: A compact vector representation for low-shot learning, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 1924–1932.
[138] M. Yang, X. Wang, G. Zeng, L. Shen, Joint and collaborative representation with local adaptive convolution feature for face recognition with single sample per person, Pattern Recogn. 66 (C) (2016) 117–128.
[139] S. Guo, S. Chen, Y. Li, Face recognition based on convolutional neural network and support vector machine, in: IEEE International Conference on Information and Automation, 2017, pp. 1787–1792.
[140] H. Jegou, M. Douze, C. Schmid, Product quantization for nearest neighbor search, IEEE Transactions on Pattern Analysis & Machine Intelligence 33 (1) (2011) 117.
[141] P.J. Grother and L.N. Mei, "Face recognition vendor test (frvt) performance of face identification algorithms nist ir 8009," NIST Interagency/Internal Report (NISTIR) - 8009, 2014.
[142] Z. Ding, Y. Guo, L. Zhang, and Y. Fu, "One-shot face recognition via generative learning," in 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018, pp. 1–7.
[143] Y. Guo and L. Zhang, "One-shot face recognition by promoting underrepresented classes," arXiv preprint arXiv:1707.05574, 2017.
[144] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359.
[145] M. Wang, W. Deng, Deep visual domain adaptation: A survey, Neurocomputing 312 (2018) 135–153.
[146] L. Xiong, J. Karlekar, J. Zhao, J. Feng, S. Pranata, and S. Shen, "A good practice towards top performance of face recognition: Transferred deep feature fusion," arXiv preprint arXiv:1704.00438, 2017.
[147] M. Kan, S. Shan, X. Chen, Bi-shifting auto-encoder for unsupervised domain adaptation, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 3846–3854.
[148] Z. Luo, J. Hu, W. Deng, and H. Shen, "Deep unsupervised domain adaptation for face recognition," in 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018, pp. 453–457.
[149] K. Sohn, S. Liu, G. Zhong, X. Yu, M.-H. Yang, M. Chandraker, Unsupervised domain adaptation for face recognition in unlabeled videos, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3210–3218.
[150] E. Tzeng, J. Hoffman, K. Saenko, T. Darrell, Adversarial discriminative domain adaptation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7167–7176.
[151] W. AbdAlmageed, Y. Wu, S. Rawls, S. Harel, T. Hassner, I. Masi, J. Choi, J. Lekust, J. Kim, P. Natarajan et al., "Face recognition using deep multi-pose representations," in 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2016, pp. 1–9.
[152] D. Wang, C. Otto, A.K. Jain, Face search at scale, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1122–1136.
[153] H. Yang and I. Patras, "Mirror, mirror on the wall, tell me, is the error small?" in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4685–4693.
[154] S. Xie, Z. Tu, Holistically-nested edge detection, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1395–1403.
[155] V. Blanz, T. Vetter, Face recognition based on fitting a 3d morphable model, IEEE Trans. Pattern Analysis Machine Intell. 25 (9) (2003) 1063–1074.
[156] Z. An, W. Deng, T. Yuan, and J. Hu, "Deep transfer network with 3d morphable models for face recognition," in 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018, pp. 416–422.
[157] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, J. Kim, Rotating your face using multi-task deep neural network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 676–684.
[158] Y. Qian, W. Deng, and J. Hu, "Task specific networks for identity and face variation," in 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018, pp. 271–277.
[159] J. Bao, D. Chen, F. Wen, H. Li, G. Hua, Cvae-gan: fine-grained image generation through asymmetric training, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2745–2754.
[160] W. Chai, W. Deng, and H. Shen, "Cross-generating gan for facial identity preserving," in 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018, pp. 130–134.
[161] J. Bao, D. Chen, F. Wen, H. Li, G. Hua, Towards open-set identity preserving face synthesis, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6713–6722.
[162] Y. Shen, P. Luo, J. Yan, X. Wang, X. Tang, Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 821–830.
[163] "Ms-celeb-1m challenge 3," http://trillionpairs.deepglint.com.

[164] A. Nech, I. Kemelmacher-Shlizerman, Level playing field for million scale face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7044–7053.
[165] Y. Zhang, W. Deng, M. Wang, J. Hu, X. Li, D. Zhao, D. Wen, Global-local gcn: Large-scale label noise cleansing for face recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7731–7740.
[166] A. Bansal, C. Castillo, R. Ranjan, R. Chellappa, The do's and don'ts for cnn-based face verification, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 2545–2554.
[167] X. Zhan, Z. Liu, J. Yan, D. Lin, C. Change Loy, Consensus-driven propagation in massive unlabeled data for face recognition, The European Conference on Computer Vision (ECCV) (September 2018).
[168] P.J. Phillips, "A cross benchmark assessment of a deep convolutional neural network for face recognition," in 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, 2017, pp. 705–710.
[169] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in CVPR. IEEE, 2011, pp. 529–534.
[170] P.J. Phillips, J.R. Beveridge, B.A. Draper, G. Givens, A.J. O'Toole, D. Bolme, J. Dunlop, Y.M. Lui, H. Sahibzada, S. Weimer, The good, the bad, and the ugly face challenge problem, Image Vis. Comput. 30 (3) (2012) 177–185.
[171] I. Hupont and C. Fernández, "Demogpairs: Quantifying the impact of demographic imbalance in deep face recognition," in 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). IEEE, 2019, pp. 1–7.
[172] I. Serna, A. Morales, J. Fierrez, M. Cebrian, N. Obradovich, and I. Rahwan, "Sensitiveloss: Improving accuracy and fairness of face representations with discrimination-aware deep learning," arXiv preprint arXiv:2004.11246, 2020.
[173] M. Wang, W. Deng, J. Hu, X. Tao, Y. Huang, Racial faces in the wild: Reducing racial bias by information maximization adaptation network, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 692–702.
[174] Y. Xu, Y. Cheng, J. Zhao, Z. Wang, L. Xiong, K. Jayashree, H. Tamura, T. Kagaya, S. Shen, S. Pranata et al., "High performance large scale face recognition with multi-cognition softmax and feature retrieval," in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 1898–1906.
[175] C. Wang, X. Zhang, X. Lan, How to train triplet networks with 100k identities?, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp 1907–1915.
[176] Y. Wu, H. Liu, Y. Fu, Low-shot face recognition with hybrid classifiers, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 1933–1939.
[177] P.J. Phillips, H. Wechsler, J. Huang, P.J. Rauss, The feret database and evaluation procedure for face-recognition algorithms, Image & Vision Computing J 16 (5) (1998) 295–306.
[178] A.M. Martinez, "The ar face database," CVC Technical Report 24, 1998.
[179] J.R. Beveridge, P.J. Phillips, D.S. Bolme, B.A. Draper, G.H. Givens, Y.M. Lui, M.N. Teli, H. Zhang, W.T. Scruggs, K.W. Bowyer et al., "The challenge of face recognition from digital point-and-shoot cameras," in 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS). IEEE, 2013, pp. 1–8.
[180] Y. Kim, W. Park, M.-C. Roh, and J. Shin, "Groupface: Learning latent groups and constructing group-based representations for face recognition," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.
[181] T. Zheng, W. Deng, "Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments," Beijing University of Posts and Telecommunications, Tech. Rep. 18–01 (February 2018).
[182] S. Sengupta, J.-C. Chen, C. Castillo, V.M. Patel, R. Chellappa, and D.W. Jacobs, "Frontal to profile face verification in the wild," in 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2016, pp. 1–9.
[183] W. Deng, J. Hu, N. Zhang, B. Chen, J. Guo, Fine-grained face verification: Fglfw database, baselines, and human-dcmn partnership, Pattern Recogn. 66 (2017) 63–73.
[184] A. Bansal, A. Nanduri, C.D. Castillo, R. Ranjan, and R. Chellappa, "Umdfaces: An annotated face dataset for training deep networks," in 2017 IEEE International Joint Conference on Biometrics (IJCB). IEEE, 2017, pp. 464–473.
[185] Y. Rao, J. Lu, J. Zhou, Attention-aware deep reinforcement learning for video face recognition, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 3931–3940.
[186] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley, "Face tracking and recognition with visual constraints in real-world videos," in 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2008, pp. 1–8.
[187] Y. Rao, J. Lin, J. Lu, J. Zhou, Learning discriminative aggregation network for video-based face recognition, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 3781–3790.
[188] T. Zheng, W. Deng, and J. Hu, "Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments," arXiv preprint arXiv:1708.08197, 2017.
[189] K. Ricanek and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in 7th International Conference on Automatic Face and Gesture Recognition (FGR06). IEEE, 2006, pp. 341–345.
[190] L. Lin, G. Wang, W. Zuo, X. Feng, L. Zhang, Cross-domain visual matching via generalized similarity measure and feature learning, IEEE Trans. Pattern Anal. Machine Intell. 39 (6) (2016) 1089–1102.
[191] B.-C. Chen, C.-S. Chen, and W.H. Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in European conference on computer vision. Springer, 2014, pp. 768–783.
[192] Y. Wen, Z. Li, Y. Qiao, Latent factor guided convolutional neural networks for age-invariant face recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4893–4901.
[193] T. Zheng, W. Deng, J. Hu, Age estimation guided convolutional neural network for age-invariant face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 1–9.
[194] "Fg-net aging database," http://www.fgnet.rsunit.com.
[195] S. Li, D. Yi, Z. Lei, S. Liao, The casia nir-vis 2.0 face database, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2013, pp. 348–353.
[196] X. Wu, L. Song, R. He, and T. Tan, "Coupled deep learning for heterogeneous face recognition," in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[197] S.Z. Li, Z. Lei, and M. Ao, "The hfb face database for heterogeneous face biometrics research," in 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE, 2009, pp. 1–8.
[198] C. Reale, N.M. Nasrabadi, H. Kwon, and R. Chellappa, "Seeing the forest from the trees: A holistic approach to near-infrared heterogeneous face recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 54–62.
[199] X. Wang, X. Tang, Face photo-sketch synthesis and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 31 (11) (2009) 1955–1967.
[200] L. Zhang, L. Lin, X. Wu, S. Ding, L. Zhang, End-to-end photo-sketch generation via fully convolutional representation learning, in: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval ACM, 2015, pp. 627–634.
[201] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in CVPR 2011. IEEE, 2011, pp. 513–520.
[202] L. Wang, V. Sindagi, and V. Patel, "High-quality facial photo-sketch synthesis using multi-adversarial networks," in 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). IEEE, 2018, pp. 83–90.
[203] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3d face analysis," in European Workshop on Biometrics and Identity Management. Springer, 2008, pp. 47–56.
[204] D. Kim, M. Hernandez, J. Choi, and G. Medioni, "Deep 3d face identification," in 2017 IEEE international joint conference on biometrics (IJCB). IEEE, 2017, pp. 133–142.
[205] L. Yin, X. Wei, Y. Sun, J. Wang, M.J. Rosato, "A 3d facial expression database for facial behavior research," in 7th international conference on automatic face and gesture recognition (FGR06), IEEE (2006) 211–216.
[206] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol. 1. IEEE, 2005, pp. 947–954.
[207] G. Guo, L. Wen, S. Yan, Face authentication with makeup changes, IEEE Trans. Circuits Syst. Video Technol. 24 (5) (2014) 814–825.
[208] Y. Li, L. Song, X. Wu, R. He, and T. Tan, "Anti-makeup: Learning a bi-level adversarial network for makeup-invariant face verification," in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[209] J. Hu, Y. Ge, J. Lu, and X. Feng, "Makeup-robust face verification," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013, pp. 2342–2346.
[210] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S.Z. Li, "A face antispoofing database with diverse attacks," in 2012 5th IAPR international conference on Biometrics (ICB). IEEE, 2012, pp. 26–31.
[211] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu, "Face anti-spoofing using patch and depth-based cnns," in 2017 IEEE International Joint Conference on Biometrics (IJCB). IEEE, 2017, pp. 319–328.
[212] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in 2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG). IEEE, 2012, pp. 1–7.
[213] J. Huo, W. Li, Y. Shi, Y. Gao, and H. Yin, "Webcaricature: a benchmark for caricature face recognition," arXiv preprint arXiv:1703.03230, 2017.
[214] V. Kushwaha, M. Singh, R. Singh, M. Vatsa, N. Ratha, and R. Chellappa, "Disguised faces in the wild," in IEEE Conference on Computer Vision and Pattern Recognition Workshops, vol. 8, 2018.
[215] K. Cao, Y. Rong, C. Li, X. Tang, C. Change Loy, Pose-robust face recognition via deep residual equivariant mapping, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5187–5196.
[216] G. Chen, Y. Shao, C. Tang, Z. Jin, and J. Zhang, "Deep transformation learning for face recognition in the unconstrained scene," Machine Vision and Applications, pp. 1–11, 2018.
[217] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing et al., "Towards pose invariant face recognition in the wild," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 2207–2216.
[218] C. Nhan Duong, K. Gia Quach, K. Luu, N. Le, M. Savvides, Temporal non-volume preserving approach to facial age-progression and age-invariant face recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3735–3743.

[219] Z. Wang, X. Tang, W. Luo, S. Gao, Face aging with identity-preserved conditional generative adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7939–7947.
[220] G. Antipov, M. Baccouche, and J.-L. Dugelay, "Face aging with conditional generative adversarial networks," in 2017 IEEE international conference on image processing (ICIP). IEEE, 2017, pp. 2089–2093.
[221] G. Antipov, M. Baccouche, and J.-L. Dugelay, "Boosting cross-age face verification via generative age normalization," in 2017 IEEE International Joint Conference on Biometrics (IJCB). IEEE, 2017, pp. 191–199.
[222] H. Yang, D. Huang, Y. Wang, A.K. Jain, Learning face age progression: A pyramid architecture of gans, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 31–39.
[223] S. Bianco, Large age-gap face verification by feature injection in deep networks, Pattern Recogn. Lett. 90 (2017) 36–42.
[224] H. El Khiyari, H. Wechsler, Age invariant face recognition using convolutional neural networks and set distances, J. Inform. Security 8 (03) (2017) 174.
[225] X. Wang, Y. Zhou, D. Kong, J. Currey, D. Li, and J. Zhou, "Unleash the black magic in age: a multi-task deep neural network approach for cross-age face verification," in 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, 2017, pp. 596–603.
[226] Y. Li, G. Wang, L. Nie, Q. Wang, W. Tan, Distance metric optimization driven convolutional neural network for age invariant face recognition, Pattern Recogn. 75 (2018) 51–62.
[227] Y. Sun, L. Ren, Z. Wei, B. Liu, Y. Zhai, S. Liu, A weakly supervised method for makeup-invariant face verification, Pattern Recogn. 66 (2017) 153–159.
[228] M. Singh, M. Chawla, R. Singh, M. Vatsa, R. Chellappa, Disguised faces in the wild 2019, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019.
[229] M. Singh, R. Singh, M. Vatsa, N.K. Ratha, R. Chellappa, Recognizing disguised faces in the wild, IEEE Trans. Biometrics, Behavior, Identity Sci. 1 (2) (2019) 97–108.
[230] K. Zhang, Y.-L. Chang, W. Hsu, Deep disguised faces recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 32–36.
[231] N. Kohli, D. Yadav, A. Noore, Face verification with disguise variations via deep disguise recognizer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 17–24.
[232] E. Smirnov, A. Melnikov, A. Oleinik, E. Ivanova, I. Kalinovskiy, E. Luckyanets, Hard example mining with auxiliary embeddings, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 37–46.
[233] E. Smirnov, A. Melnikov, S. Novoselov, E. Luckyanets, G. Lavrentyeva, Doppelganger mining for face representation learning, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 1916–1923.
[234] S. Suri, A. Sankaran, M. Vatsa, and R. Singh, "On matching faces with alterations due to plastic surgery and disguise," in 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS). IEEE, 2018, pp. 1–7.
[235] S. Saxena and J. Verbeek, "Heterogeneous face recognition with cnns," in European conference on computer vision. Springer, 2016, pp. 483–491.
[236] X. Liu, L. Song, X. Wu, and T. Tan, "Transferring deep representation for nir-vis heterogeneous face recognition," in 2016 International Conference on Biometrics (ICB). IEEE, 2016, pp. 1–8.
[237] J. Lezama, Q. Qiu, G. Sapiro, Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6628–6637.
[238] R. He, X. Wu, Z. Sun, and T. Tan, "Learning invariant deep representation for nir-vis face recognition," in Thirty-First AAAI Conference on Artificial Intelligence, 2017.
[239] R. He, X. Wu, Z. Sun, T. Tan, Wasserstein cnn: Learning invariant features for nir-vis face recognition, IEEE Trans. Pattern Analysis Machine Intell. 41 (7) (2018) 1761–1773.
[240] L. Song, M. Zhang, X. Wu, R. He, "Adversarial discriminative heterogeneous face recognition," in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[241] E. Zangeneh, M. Rahmati, Y. Mohsenzadeh, Low resolution face recognition using a two-branch deep convolutional neural network architecture, Expert Syst. Appl. 139 (2020) 112854.
[242] Z. Shen, W.-S. Lai, T. Xu, J. Kautz, M.-H. Yang, Deep semantic face deblurring, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8260–8269.
[243] P. Mittal, M. Vatsa, and R. Singh, "Composite sketch recognition via deep network-a transfer learning approach," in 2015 International Conference on Biometrics (ICB). IEEE, 2015, pp. 251–256.
[244] C. Galea, R.A. Farrugia, Forensic face photo-sketch recognition using a deep learning-based architecture, IEEE Signal Process. Lett. 24 (11) (2017) 1586–1590.
[245] D. Zhang, L. Lin, T. Chen, X. Wu, W. Tan, E. Izquierdo, Content-adaptive sketch portrait generation by decompositional representation learning, IEEE Trans. Image Process. 26 (1) (2017) 328–339.
[246] Z. Yi, H. Zhang, P. Tan, M. Gong, Dualgan: Unsupervised dual learning for image-to-image translation, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2849–2857.
[247] T. Kim, M. Cha, H. Kim, J. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," arXiv preprint arXiv:1703.05192, 2017.
[248] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223–2232.
[249] S. Hong, W. Im, J. Ryu, and H.S. Yang, "Sspp-dan: Deep domain adaptation network for face recognition with single sample per person," in 2017 IEEE International Conference on Image Processing (ICIP). IEEE, 2017, pp. 825–829.
[250] J. Choe, S. Park, K. Kim, J. Hyun Park, D. Kim, H. Shim, Face generation for low-shot learning using generative adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 1940–1948.
[251] X. Yin, X. Yu, K. Sohn, X. Liu, M. Chandraker, Feature transfer learning for face recognition with under-represented data, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5704–5713.
[252] J. Lu, G. Wang, W. Deng, P. Moulin, J. Zhou, Multi-manifold deep metric learning for image set classification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1137–1145.
[253] J. Zhao, J. Han, L. Shao, Unconstrained face recognition using a set-to-set distance measure on deep learned features, IEEE Trans. Circuits Syst. Video Technol. (2017).
[254] N. Bodla, J. Zheng, H. Xu, J.-C. Chen, C. Castillo, and R. Chellappa, "Deep heterogeneous feature fusion for template-based face recognition," in 2017 IEEE winter conference on applications of computer vision (WACV). IEEE, 2017, pp. 586–595.
[255] M. Hayat, M. Bennamoun, S. An, Learning non-linear reconstruction models for image set classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1907–1914.
[256] X. Liu, B. Vijaya Kumar, C. Yang, Q. Tang, J. You, Dependency-aware attention control for unconstrained face recognition with image sets, The European Conference on Computer Vision (ECCV) (September 2018).
[257] C. Ding, D. Tao, Trunk-branch ensemble convolutional neural networks for video-based face recognition, IEEE Trans. Pattern Analysis Machine Intell. (2017).
[258] M. Parchami, S. Bashbaghi, E. Granger, and S. Sayed, "Using deep autoencoders to learn robust domain-invariant representations for still-to-video face recognition," in 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2017, pp. 1–6.
[259] S. Zulqarnain Gilani, A. Mian, Learning from millions of 3d scans for large-scale 3d face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1896–1905.
[260] J. Zhang, Z. Hou, Z. Wu, Y. Chen, and W. Li, "Research of 3d face recognition algorithm based on deep learning stacked denoising autoencoder theory," in 2016 8th IEEE International Conference on Communication Software and Networks (ICCSN). IEEE, 2016, pp. 663–667.
[261] L. He, H. Li, Q. Zhang, Z. Sun, and Z. He, "Multiscale representation for partial face recognition under near infrared illumination," in IEEE International Conference on Biometrics Theory, Applications and Systems, 2016, pp. 1–7.
[262] L. He, H. Li, Q. Zhang, and Z. Sun, "Dynamic feature learning for partial face recognition," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.
[263] O. Tadmor, Y. Wexler, T. Rosenwein, S. Shalev-Shwartz, and A. Shashua, "Learning a metric embedding for face recognition using the multibatch method," arXiv preprint arXiv:1605.07270, 2016.
[264] S. Han, H. Mao, and W.J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," arXiv preprint arXiv:1510.00149, 2015.
[265] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," in Advances in neural information processing systems, 2015, pp. 1135–1143.
[266] Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang, "Learning efficient convolutional networks through network slimming," in Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, 2017, pp. 2755–2763.
[267] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1," arXiv preprint arXiv:1602.02830, 2016.
[268] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks," in Advances in neural information processing systems, 2016, pp. 4107–4115.
[269] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "Xnor-net: Imagenet classification using binary convolutional neural networks," in European Conference on Computer Vision. Springer, 2016, pp. 525–542.
[270] M. Courbariaux, Y. Bengio, and J.-P. David, "Binaryconnect: Training deep neural networks with binary weights during propagations," in Advances in neural information processing systems, 2015, pp. 3123–3131.
[271] Q. Li, S. Jin, and J. Yan, "Mimicking very efficient network for object detection," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017, pp. 7341–7349.
[272] Y. Wei, X. Pan, H. Qin, W. Ouyang, J. Yan, Quantization mimic: Towards very tiny cnn for object detection, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 267–283.
[273] J. Yang, Z. Lei, and S.Z. Li, "Learn convolutional neural network for face anti-spoofing," arXiv preprint arXiv:1408.5601, 2014.

[274] Z. Xu, S. Li, and W. Deng, "Learning temporal features using lstm-cnn architecture for face anti-spoofing," in 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR). IEEE, 2015, pp. 141–145.
[275] L. Li, X. Feng, Z. Boulkenafet, Z. Xia, M. Li, and A. Hadid, "An original face anti-spoofing approach using partial convolutional neural network," in 2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA). IEEE, 2016, pp. 1–6.
[276] K. Patel, H. Han, and A.K. Jain, "Cross-database face antispoofing with robust feature representation," in Chinese Conference on Biometric Recognition. Springer, 2016, pp. 611–619.
[277] A. Jourabloo, Y. Liu, and X. Liu, "Face de-spoofing: Anti-spoofing via noise modeling," in The European Conference on Computer Vision (ECCV), September 2018.
[278] R. Shao, X. Lan, P.C. Yuen, Joint discriminative learning of deep dynamic textures for 3d mask face anti-spoofing, IEEE Trans. Inf. Forensics Secur. 14 (4) (2019) 923–938.
[279] R. Shao, X. Lan, and P.C. Yuen, "Deep convolutional dynamic texture learning with adaptive channel-discriminability for 3d mask face anti-spoofing," in Biometrics (IJCB), 2017 IEEE International Joint Conference on. IEEE, 2017, pp. 748–755.
[280] G. Goswami, N. Ratha, A. Agarwal, R. Singh, and M. Vatsa, "Unravelling robustness of deep learning based face recognition against adversarial attacks," arXiv preprint arXiv:1803.00401, 2018.
[281] A. Goel, A. Singh, A. Agarwal, M. Vatsa, and R. Singh, "Smartbox: Benchmarking adversarial detection and mitigation algorithms for face recognition," IEEE BTAS, 2018.
[282] G. Mai, K. Cao, P.C. Yuen, A.K. Jain, On the reconstruction of face images from deep face templates, IEEE Trans. Pattern Analysis Machine Intell. 41 (5) (2018) 1188–1202.
[283] M. Wang, W. Deng, Mitigating bias in face recognition using skewness-aware reinforcement learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9322–9331.
[284] M. Alvi, A. Zisserman, C. Nellåker, Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018.
[285] V. Mirjalili, S. Raschka, and A. Ross, "Gender privacy: An ensemble of semi adversarial networks for confounding arbitrary gender classifiers," in 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS). IEEE, 2018, pp. 1–10.
[286] A. Othman and A. Ross, "Privacy of facial soft biometrics: Suppressing gender but retaining identity," in European Conference on Computer Vision. Springer, 2014, pp. 682–696.
[287] J. Guo, X. Zhu, C. Zhao, D. Cao, Z. Lei, S.Z. Li, Learning meta face recognition in unseen domains, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6163–6172.
[288] A. Morales, J. Fierrez, and R. Vera-Rodriguez, "Sensitivenets: Learning agnostic representations with application to face recognition," arXiv preprint arXiv:1902.00334, 2019.
[289] R. Ramachandra, C. Busch, Presentation attack detection methods for face recognition systems: a comprehensive survey, ACM Computing Surveys (CSUR) 50 (1) (2017) 8.
[290] Y. Zhong and W. Deng, "Towards transferable adversarial attack against deep face recognition," arXiv preprint arXiv:2004.05790, 2020.
[291] G. Mai, K. Cao, P.C. Yuen, and A.K. Jain, "On the reconstruction of deep face templates," arXiv preprint arXiv:1703.00832, 2017.
[292] H. Dang, F. Liu, J. Stehouwer, X. Liu, A.K. Jain, On the detection of digital face manipulation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5781–5790.
[293] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Niessner, Faceforensics++: Learning to detect manipulated facial images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019.
[294] I. Manjani, S. Tariyal, M. Vatsa, R. Singh, A. Majumdar, Detecting silicone mask-based presentation attack via deep dictionary learning, IEEE Trans. Inf. Forensics Secur. 12 (7) (2017) 1713–1723.
[295] M. Sharif, S. Bhagavatula, L. Bauer, and M.K. Reiter, "Adversarial generative nets: Neural network attacks on state-of-the-art face recognition," arXiv preprint arXiv:1801.00349, 2017.
[296] M. Sharif, S. Bhagavatula, L. Bauer, M.K. Reiter, Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition, in: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security ACM, 2016, pp. 1528–1540.
[297] Y. Gurovich, Y. Hanani, and e. a. Bar, Omri, "Identifying facial phenotypes of genetic disorders using deep learning," Nature Medicine, vol. 25, pp. 60 – 64, 2019.
[298] V. Mirjalili and A. Ross, "Soft biometric privacy: Retaining biometric utility of face images while perturbing gender," in Biometrics (IJCB), 2017 IEEE International Joint Conference on. IEEE, 2017, pp. 564–573.
[299] V. Mirjalili, S. Raschka, A. Namboodiri, and A. Ross, "Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images," in 2018 International Conference on Biometrics (ICB). IEEE, 2018, pp. 82–89.
[300] P.J. Phillips, A.N. Yates, Y. Hu, C.A. Hahn, E. Noyes, K. Jackson, J.G. Cavazos, G. Jeckeln, R. Ranjan, S. Sankaranarayanan et al., "Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms," Proceedings of the National Academy of Sciences, p. 201721355, 2018.
[301] S. Gong, V.N. Boddeti, and A.K. Jain, "On the capacity of face representation," arXiv preprint arXiv:1709.10433, 2017.
[302] R. Singh, M. Vatsa, H.S. Bhatt, S. Bharadwaj, A. Noore, S.S. Nooreyezdan, Plastic surgery: A new dimension to face recognition, IEEE Trans. Inf. Forensics Secur. 5 (3) (2010) 441–448.
[303] A. Ross and A.K. Jain, "Multimodal biometrics: An overview," in Signal Processing Conference, 2004 12th European. IEEE, 2004, pp. 1221–1224.
[304] A.A. Ross and R. Govindarajan, "Feature level fusion of hand and face biometrics," in Biometric Technology for Human Identification II, vol. 5779. International Society for Optics and Photonics, 2005, pp. 196–205.
[305] A. Ross, A. Jain, Information fusion in biometrics, Pattern Recogn. Lett. 24 (13) (2003) 2115–2125.

Mei Wang received the B.E. degree in information and communication engineering from the Dalian University of Technology (DUT), Dalian, China, in 2013 and received M.E. degree in communication engineering from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2016. From September 2018, she is a Ph.D. student in school of information and communication engineering of BUPT. Her research interests include pattern recognition and computer vision, with a particular emphasis in deep face recognition and transfer learning.

Weihong Deng received the B.E. degree in information engineering and the Ph.D. degree in signal and information processing from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2004 and 2009, respectively. From Oct. 2007 to Dec. 2008, he was a postgraduate exchange student in the School of Information Technologies, University of Sydney, Australia. He is currently an professor in School of Information and Telecommunications Engineering, BUPT. His research interests include statistical pattern recognition and computer vision, with a particular emphasis in face recognition. He has published over 100 technical papers in international journals and conferences, such as IEEE TPAMI and CVPR. He serves as associate editor for IEEE Access, and guest editor for Image and Vision Computing Journal and the reviewer for dozens of international journals, such as IEEE TPAMI / TIP / TIFS / TNNLS / TMM / TSMC, IJCV, PR / PRL. His Dissertation titled "Highly accurate face recognition algorithms" was awarded the Outstanding Doctoral Dissertation Award by Beijing Municipal Commission of Education in 2011. He has been supported by the program for New Century Excellent Talents by the Ministry of Education of China in 2013 and Beijing Nova Program in 2016.