Computer Vision and Image Understanding
journal homepage: www.elsevier.com

A survey on deep learning based face recognition

Guodong Guo[a,b,**], Na Zhang[b]

[a] School of Information and Communication Engineering, North University of China, Taiyuan, China
[b] LCSEE, West Virginia University, Morgantown, WV 26506, USA

ABSTRACT

Deep learning, in particular the deep convolutional neural networks, has received increasing interests in face recognition recently, and a number of deep learning methods have been proposed. This paper summarizes about 330 contributions in this area. It reviews major deep learning concepts pertinent to face image analysis and face recognition, and provides a concise overview of studies on specific face recognition problems, such as handling variations in pose, age, illumination, expression, and heterogeneous face matching. A summary of databases used for deep face recognition is given as well. Finally, some open challenges and directions are discussed for future research.

1. Introduction

Face is the most common characteristic used by humans for recognition. Face recognition (FR) is a classical problem and is still very active in computer vision and image understanding. Fig.1 shows the whole pipeline of an automatic face recognition system. A face image is fed into the system, and face detection and face alignment are processed. And then a feature extractor is used to extract features. Finally, the system compares the extracted features with the gallery faces to do face matching. In face matching, there are two different tasks: face verification (FV) and face identification (FI). FV is to determine whether a given pair of face images or videos belongs to the same subject. FI is a one-to-many matching, recognizing the person from a set of gallery face images or videos of different subjects. Face identification usually assumes the query person has already enrolled in the gallery, which is a closed-set problem. Watch-list is similar to face identification but it does not guarantee all query subjects are registered in the gallery, which is an open-set problem. In the real world, it is normal to treat FI as an open-set problem.

With the development of computer hardware and imaging technology, FR has been applied widely to daily lives, such as access control, video surveillance, etc. The demands of FR are also growing quickly in recent years. In practice, however, face recognition is affected by many factors, for example, in unconstrained face recognition, the face images may have many variations, such as low resolution, pose variation, complex illumination and motion blur, as shown in Fig.2, resulting in low recognition accuracies. Traditional algorithms, such as the Eigenfaces (Turk and Pentland, 1991), Fisherfaces (Belhumeur et al., 1997), Bayesian face (Moghaddam et al., 2000), Metaface (Yang et al., 2010), support vector machine (SVM) based (Guo et al., 2000), Boosting (Guo and Zhang, 2001), etc., may not do well for unconstrained face matching.

Artificial Neural network (ANN) has advantages in terms of learning ability, generalization, and robustness (Lawrence et al., 1997; Lin et al., 1997). Recently, there is a surge of increasing interests in neural network (Krizhevsky et al., 2012), especially the deep neural network (DNN). Deep and large networks have exhibited impressive results when there are large training data sets and computation resources, such as many CPU cores and/or GPUs. Deep learning (DL), through neural networks with multi hidden layers and massive training data, aims to learn the essential feature representation of data by constructing high-level features from the low-level pixels.

Due to deep learning techniques, there have been significant advances in face recognition. In early time, research interests mainly concentrated on face recognition with deep networks on visible light face images and/or video faces. Stephen (2015) provided a short review of deep learning techniques and representation learning in face recognition and compared several popular convolutional neural networks (CNNs) based deep models. Fu et al. (2014) analyzed the architecture of typical deep networks used in FR, such as deep belief network (DBN),

**Corresponding author: Tel.: +1-304-293-9143; fax: +1-304-293-8602;
e-mail: Guodong.Guo@mail.wvu.edu (Guodong Guo)

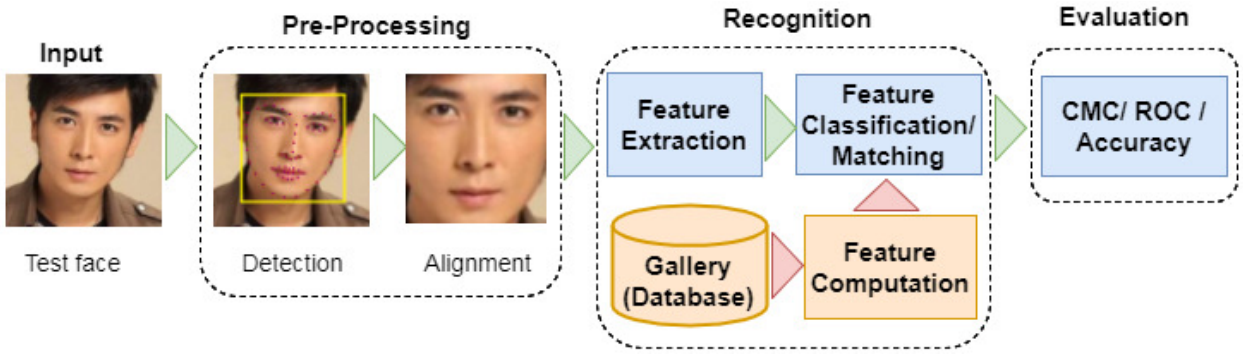

Fig. 1. The pipeline of a typical automated face recognition system.

Fig. 2. Illustrate different variations of face images in unconstrained face recognition: (a) low resolution, (b) pose variation, (c) complex illumination, and (d) motion blur.

convolutional neural network (CNN, or ConvNet), autoencoder (AE), etc. Mandal et al. (2016) reviewed a substantial amount of deep learning methods for FR. Sepas-Moghaddam et al. (2019) provided a survey of face recognition solutions based on a new, more encompassing and richer multi-level taxonomy. Learned-Miller et al. (2016) reviewed a remarkably wide variety of innovative methods on Labeled Faces in the Wild (LFW) database (Huang et al., 2007).

With the collection of various types of face data, research concentrations have also focused on some specific tasks, such as robustness to changes of pose, illumination, expression, age, or improving performance of video, 3D, and heterogeneous (e.g., NIR-VIS, Sketch-Photo, Still-to-Video) face recognition. Although some related surveys overviewed methods on handling pose (Ding and Tao, 2016), illumination (Chan et al., 2014), expression (Murtaza et al., 2013), occlusion (Lahasan et al., 2017), infrared (Ghiass et al., 2014), single-modal and multimodal (Zhou et al., 2014), video (Barr et al., 2012), 3D (Patil et al., 2015), heterogeneous face matching (Guo, 2014; Ouyang et al., 2016a), etc., most of them focused on the traditional methods, and few of them has been related to deep learning methods.

We present a complete, comprehensive overview of face recognition works using deep learning, considering both the deep architectures and specific recognition problems. We also give a review of related face databases. Fig.3 (a) shows some statistics of the related papers. Applying deep learning to face image analysis started several years ago, while the number of papers have been growing rapidly, as shown in Fig.3 (b). This survey includes about 330 papers, and most of them are within the recent five years. It is expected to cover most, if not all, of the works incorporating deep learning methods for face recognition.

By this survey, we show that:

- deep learning methods have been fully applied to face recognition and played important roles;
- many specific issues or challenges to address in FR by DL, such as pose, illumination, expression, 3D, heterogenous matching, etc.;
- various face datasets collected in recent years, including still images, videos, and heterogeneous data that raises the issue of cross-modal face matching.

The rest of this survey is organized as follows. Section 2 introduces main deep learning techniques that have been developed and used for face recognition. Section 3 describes the contributions of deep learning in some specific FR issues. Section 4 discusses face databases in recent years. Section 5 discusses some challenges and outlook for future research directions in deep learning based FR problems. Finally, Section 6 gives the conclusion.

2. Deep Learning Methods

Artificial Neural Network (ANN) (Haykin, 2009) is a computational nonlinear model inspired by the biological systems in information processing. It consists of artificial neurons or processing elements and is typically organized in three types of interconnected layers. Data are presented to the network via the input layer, which communicates to one or more hidden layers where the actual processing is done via weighted connections. The hidden layers then link to an output layer to give the output. It is possible to make the neural network more flexible and more powerful by using additional hidden layers. Artificial neural networks with many hidden layers between the input and output layers are called Deep Neural Networks (DNNs), and they can model complex relationships between the input and output.

There are various deep neural networks used in face recognition. Convolutional Neural Network (CNN) is the most popular. It shows outstanding results in image and speech applications. Autoencoder (AE) and its variants also gained much attention. They process data without using class labels and the

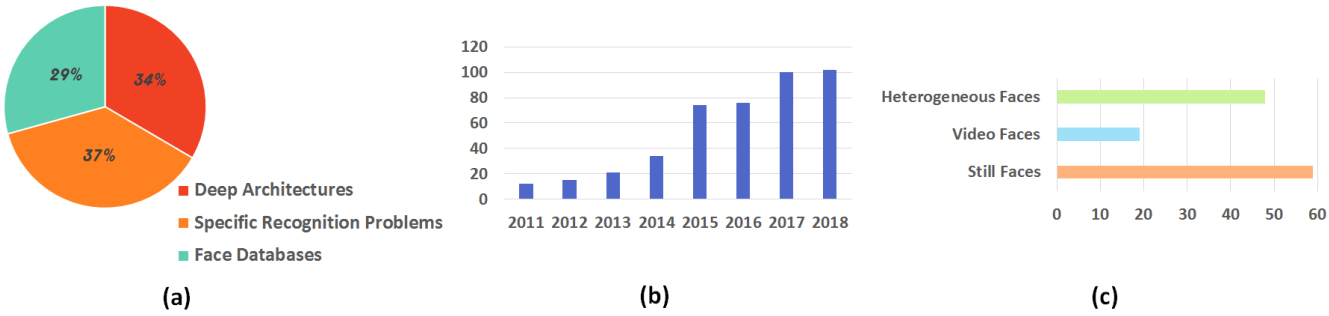

Fig. 3. Statistical figures (best view in color) of the related papers: (a) distribution by category, (b) distribution by year, (c) distribution of databases by face type, e.g., still, video, heterogeneous faces.

purpose is to find patterns, such as latent subspaces. Generative Adversarial Network (GAN) has increased rapidly recently. It usually contains two nets, putting one against the other, thus called adversarial. It can learn to mimic any distribution of data. Deep Belief Network (DBN), Deep Boltzmann Machine (DBM) are also used in FR. However, the Recurrent Neural Network (RNN), Self-Organizing Map (SOM), Radial Basis Function Network (RBFN), are not used very often in FR. Fig. 4 (a) gives the distribution of research works for different neural network architectures.

2.1. Convolutional Neural Networks (CNNs)

In the last decade, Convolutional Neural Network (CNN) (LeCun et al., 1998) has become one of the most popular techniques in computer vision, such as image classification (He et al., 2016a), object detection (Redmon et al., 2016), face recognition (Yi et al., 2014). Many vision tasks have benefited from the robust and discriminative representation learned via CNN and the performance has been improved significantly. Probably the first successful real world application of CNN is LeNet (LeCun et al., 1998) for hand-written digit recognition. AlexNet (Krizhevsky et al., 2012) is considered one of the most influential research work for DL in Computer Vision, having spurred many works employing CNNs and GPUs to accelerate deep learning research and development (Simonyan and Zisserman, 2014; Szegedy et al., 2015). In face recognition, deep CNN has now become the technique of choice.

CNN typically consists of convolutional layers, pooling layers and fully connected layers. Convolutional layers are core building blocks of CNN. The objective of a convolutional layer is to extract features from the input data. The parameters of each layer consist of a set of learnable filters $W = W_1, W_2, ..., W_k$ and added biases $B = b_1, ..., b_k$. Each layer applies a convolution operation to generate a feature map X_k and pass the result to next layer. These features are subject to an element-wise nonlinear transform $\sigma(\cdot)$ and the same process is repeated for each convolutional layer t,

$$X_k^t = \sigma(W_k^{t-1} * X^{t-1} + b_k^{t-1}). \quad (1)$$

Pooling is a form of nonlinear down-sampling. There are different nonlinear functions to implement pooling, such as average pooling, L_2-norm pooling and max pooling. Max pooling is the most effective one and superior to subsampling (Scherer et al., 2010).

In the following, we discuss some CNN based deep methods for face recognition, including single CNN, multi CNNs, some variants of CNN, etc. Fig. 4 (b) shows the paper distribution. Most are still image-based face recognition (SIFR). Other types of face recognition will be discussed in later sections in details.

2.1.1. Single CNN

Typical deep face recognition methods adopt one single CNN, which is usually trained in a supervised fashion. Table 1 gives an overview of FR methods based on a single deep CNN. The earlier DeepFace (Taigman et al., 2014) is a 9-layer CNN in which the input is RGB face images preprocessed with 3D-alignment. Several locally connected convolutional layers are adopted without weight sharing, and every location in feature maps of these layers learns a different set of filters. As an extension of DeepFace, Web-Scale (Taigman et al., 2015) used a bootstrapping process to select more efficient training set by replacing the naive random subsampling.

As the extensive usage of single CNN in face recognition community, various strategies are designed to improve the performance for face recognition. Common strategies used in CNN include: (1) learning more discriminative deep features, (2) fusing different types of face features, (3) utilizing efficient metric learning algorithms, (4) designing more powerful loss functions, (5) adopting proper activation functions, and (6) other strategies.

Learn more discriminative features. One major is to learn more discriminative deep features. Zheng et al. (2016) learned an improved discriminative representation called Vector of Locally Aggregated Descriptor encoded DCNN feature (VLAD-DCNN), in which the spatial and appearance information are processed simultaneously. Doppelganger mining (Smirnov et al., 2017) improved the discriminative power of features by inserting into the learning process with joint prototype-based and exemplar-based supervision. Ding et al. (2017) learned a noise-robust deep feature representation which can increase inter-class variations and reduce intra-class variations simultaneously. Hsieh et al. (2017) learned more semantic and discriminative face representations by incorporating identity and high-level human attributes (e.g., gender, age) in a multi-task learning framework. FV-DCNN (Chen et al., 2016b) combined

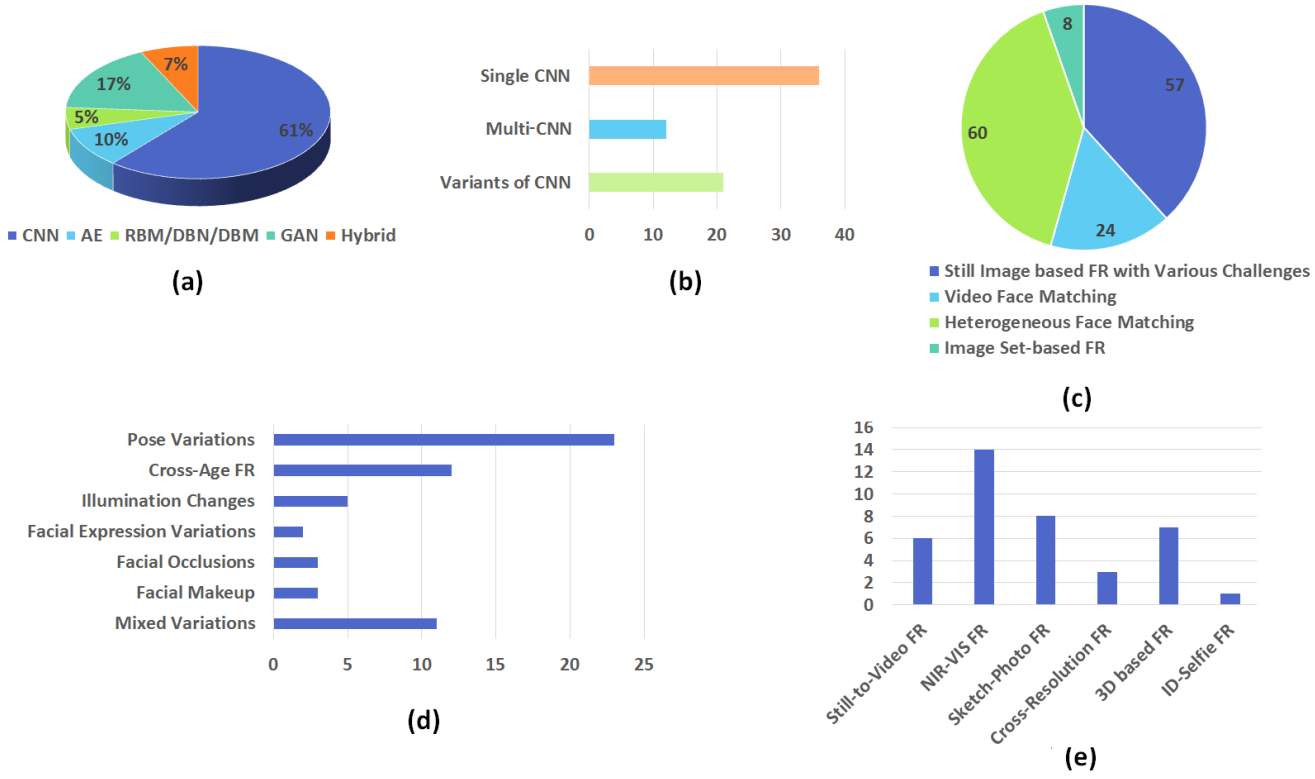

Fig. 4. The numbers of papers belong to (a) various neural network architectures, (b) CNN type, (c) specific recognition problems, (d) different challenges, (e) specific heterogeneous face recognition problems.

the deeply learned feature by CNN and Fisher vector representation to generate Fisher vector encoded DCNN features, which can capture both local and global variations. Park et al. (2017) proposed a residual learning method to learn face representations that can be used to directly determine whether two input images belong to the same identity. Wang et al. (2017c) proposed a Discriminative Covariance oriented Representation Learning (DCRL) framework for face recognition with image sets by learning deep representations which can match the subsequent image set modeling and classification.

Fuse different types of face features. Some methods attempt to fuse various features to enhance FR performance. For example, Hu et al. (2017a) introduced a facial attribute feature (FAF) and fused it with face recognition features (FRF) to enhance performance in various challenging scenarios. Lumini et al. (2016) combined deep features from CNN and hand-crafted features. Grundström (2015) utilized two distinct feature types, local feature representations around landmark points and deep representations extracted from CNN trained for generic object detection and fine-tuned on face image data, to evaluate face verification.

Utilize metric learning algorithms. Many metric learning algorithms have been proposed to enhance the discriminative power of learnt face representations (Cai et al., 2012; Cui et al., 2013; Koestinger et al., 2012). The common objective of these methods is to learn a good distance metric so that the distance between positive face pairs is reduced and that of negative pairs is enlarged as much as possible. For example, FaceNet (Schroff et al., 2015) learned a mapping from face images to a compact Euclidean space and optimizes the embedding itself by a triplet loss. VGGFace (Parkhi et al., 2015) fine-tuned the model via a triplet-based metric learning method like FaceNet. Sankaranarayanan et al. (2016a) coupled a deep CNN-based approach with a low-dimensional discriminative embedding, and adopted a triplet probability embedding learning method to improve the performance of deep features. Chen et al. (2016a) used a joint Bayesian metric learning to assess the similarity between two face representations. DDML (Hu et al., 2014) learned a set of hierarchical nonlinear transformations to project face pairs into one feature space in a deep architecture, where the nonlinear mappings are obtained explicitly.

Design powerful loss functions. Adopting new loss functions in deep neural networks is another effective approach. Recently, quantities of loss functions were proposed. Center Loss (Wen et al., 2016b) was used to learn a center for deep features in each class and penalized the distances between the deep features and their corresponding class centers. Yeung et al. (2017) introduced a constrained triplet loss layer (CTLL) to help the deep model to specify further distinguishable clusters between different subjects by placing extra constraints on images of the same person while putting margins on images of different persons. DeepVisage (Hasnat et al., 2017) incorporated residual learning framework (He et al., 2016a) and used normalized features to compute Softmax loss. NormFace (Wang et al., 2017a) used a modification of Softmax loss to optimize the cosine similarity instead of inner-product. Derived from the Softmax loss, Jones and Kobori (2017) proposed a similarity function called hyperplane similarity to train the CNN, which is more appro-

Table 1. Overview of deep learning methods based on a single deep CNN.

Algorithm	Description/Remark
DeepFace (Taigman et al., 2014)	Employ 3D face modeling to apply a piecewise affine tranformation to derive feature
Web-Scale (Taigman et al., 2015)	Use a bootstrapping process to select an efficient training set from a large dataset to alleviate performance saturation
Gruber et al. (2017)	Use a 50-layer deep residual network ResNet to face recognition task
VLAD-DCNN (Zheng et al., 2016)	Combine VLAD feature encoding with DCNN features
Smirnov et al. (2017)	Insert sampling method into feature learning process
NR-Network (Ding et al., 2017)	Learn noise-robust deep feature representation
Hsieh et al. (2017)	A multi-task learning model; Incorporate identity and high-level human attributes
FV-DCNN (Chen et al., 2016b)	Combine deep feature and Fisher vector representation
Park et al. (2017)	Get features directly used to determine if two input images are identical
Wang et al. (2017c)	Apply a Discriminative Covariance oriented Representation Learning framework
Hu et al. (2017a)	Fuse facial attribute feature with face recognition features
Lumini et al. (2016)	Combine deep features and hand-crafted features
Grundström (2015)	An algorithm suitable for real time use in an embedded environment with limited space and restricted computational resources
FaceNet (Schroff et al., 2015)	Directly learn a mapping from images to a compact Euclidean space; Great representational efficiency
VGGFace (Parkhi et al., 2015)	Combine very deep convolution neural network and the triplet embedding
Sankaranarayanan et al. (2016a)	Deep CNN based approach combined with a low-dimensional discriminative embedding which are learned by triplet probability constraints
Chen et al. (2016a)	Use a joint Bayesian metric learning to assess the similarity
DDML (Hu et al., 2014)	Learn a set of hierarchical nonlinear transformations
Center Loss (Wen et al., 2016b)	With the joint supervision of softmax loss and center loss
Yeung et al. (2017)	A constrained triplet loss layer to be replaced at the bottom of neural network
DeepVisage (Hasnat et al., 2017)	Incorporate residual learning framework; Normalized features used for softmax loss
NormFace (Wang et al., 2017a)	Use normalized features to train DCNN
Jones and Kobori (2017)	Use hyperplane similarity to train CNN
SphereFace (Liu et al., 2017b)	Learn features with angular margin; Discriminative on hypersphere manifold
Light CNN (Wu et al., 2015)	Light frameworks with reduced parameters and time to learn a 256-D compact embedding on the large scale face data with massive noisy labels
Wu (2015)	Use MFM activation function
Yang et al. (2017b)	A fully convolutional structure with higher speed and less computational cost; Use max-feature-map as activation function
Han et al. (2018)	Propose contrastive convolution
Kang et al. (2018)	To investigate the effective features for face recognition
Sparse ConvNets (Sun et al., 2016)	With sparse neural connections in an iterative way from the previously learned denser models with a neural correlation based weight selection criterion
He et al. (2015b)	A predictable hash code algorithm; Map face samples to Hamming space
Cui et al. (2018a)	A discriminative face depth estimation approach to improve 2D face recognition
Li et al. (2015b)	Batch learning strategy; Mahalanobis metric and distance threshold for optimization
Seo et al. (2015)	A multi-task learning; Use two-stage learning strategy to minimize error functions
Grm et al. (2016)	A two-structural parts network; Convolutional layers try to capture the joint characteristics of input image pair; Fully-connected layers produce a similarity index
Hayat et al. (2017)	A data-driven approach which can jointly learn registration with representation

priate than L_2 distance. A deep hypersphere embedding approach SphereFace (Liu et al., 2017b) used angular softmax (A-Softmax) as the loss function. More loss functions can be seen from Table 8 in the subsection 'Loss Functions' which gives an overview of definitions and a thorough comparison of various loss functions used in deep learning based FR.

Adopt proper activation functions. Choosing a proper activation function is also important. Take Max-Feature-Map (MFM) (Wu et al., 2015) for example. MFM is a variation of maxout activation. However, it does not use feature maps to linearly approximate an arbitrary convex activation function, but suppresses neurons by a competitive relationship. A pretty good performance has been achieved in Wu (2015) and Yang et al. (2017b). More typical activation functions, such as Sigmoid, Tanh, Rectified Linear Units (ReLU), Leaky Rectified Linear Units (LReLU) (Maas et al., 2013), Parametric Rectified Linear Units (PReLU) (He et al., 2015a), etc., can be seen in the subsection 'Activation Functions' which gives an overview of definitions and descriptions of various activation functions used in deep learning based FR.

Other strategies. Besides, quite a lot methods use other strategies not belonging to those mentioned above. Inspired by the observation that humans generally focus on varied characteristics of a face when comparing distinct persons, Han et al. (2018) designed a CNN architecture with a contrastive convolution, which specifically focuses on the contrastive characteris-

tics between two faces. PRN (Kang et al., 2018) tried to capture unique and discriminative pairwise relations among different identities by obtaining local appearance patches around landmark points on the feature map. Sparse ConvNets (Sun et al., 2016) learned an effective model with sparse neural connections, which can get a good initialization and avoid bad local minima. He et al. (2015b) proposed a predictable hash code algorithm to map face samples to Hamming space to further enhance the predictability of binary codes. The 3D or multi-modality RGB-D data can be helpful to achieve robustness against the challenges in unconstrained scenarios, such as large pose, bad illumination, and partial occlusion. Based on this, Cui et al. (2018a) proposed a discriminative face depth estimation approach to improve 2D face recognition accuracy using a cascaded FCN and CNN architecture.

CNN is not only used in general face recognition but also adopted to handle some specific issues, like large pose, poor illumination, HFR. We describe them in later sections.

2.1.2. Multi-CNN

Choosing more than one CNNs to extract different deep features and concatenate them in some way as the final face representation is commonly investigated too. It mainly contains two types: (1) extracting deep features of different regions of a face, and (2) extracting features of different aspects of faces. These deep models (as shown in Table 2) usually require additional training data to train each CNN. It is necessary to explore particular modalities that can contribute to enhance their performance, which requires significant efforts in terms of data preparation or selection and computing resources.

Extract features in different regions of faces. A typical strategy is to extract deep features in different regions of the face. SIAMESE (Wang et al., 2014) used a layer-wise training method to learn features for different parts and scales of faces. MFRS (Zhou et al., 2015) cropped four face regions for representation extraction. Baidu (Liu et al., 2015) extracted low dimensional but very discriminative features of overlapped image patches centered at different landmarks of faces. The DeepID series methods (DeepID, DeepID2, DeepID2+, DeepID3) extracted robust features of different local face patches too. DeepID (Sun et al., 2014b) extracted features from 60 face patches with ten regions, three scales, and RGB or gray channels and forms the complementary and over-complete representations. DeepID2 (Sun et al., 2014a) took similar structures as in DeepID and got a 160-dimensional DeepID2 feature vector at its DeepID2 layer. Inherited from DeepID2, DeepID2+ (Sun et al., 2015b) is larger, and the final feature representation is increased to 512 dimensions. Besides, the training data is enlarged too. DeepID3 (Sun et al., 2015a) inherited a few characteristics of DeepID2+, such as unshared neural weights in the last few feature extraction layers, the way of adding supervisory signals to early layers. But it is significantly deeper due to stacking multiple convolution/inception layers before each pooling layer. Continuous convolution/inception helps to form features with larger receptive fields and more complex nonlinearity while restricting the number of parameters.

Extract features of different aspects of faces. Another common strategy is to extract features of different aspects of faces. Kang et al. (2017) designed a Multi-scale Convolution Layer Blocks (MCLBs) based face recognition system to extract low dimensional but discriminative feature and high-level abstracted feature by stacking MCLBs to present multi-scale abstraction. Xiong et al. (2017) used an unified learning framework to explore the complementarity of two distinct deep convolutional neural networks by training them with two different large datasets, and then fused the two types of deep features for classification. Bodla et al. (2017) constructed a deep heterogeneous feature fusion network to exploit the complementary information presented in features generated by different DCNNs for template-based face recognition. Lu et al. (2017c) concatenated different features of two deep CNNs after PCA reduction for FR. Each type of feature is a combination of multi-scale representations through the use of auxiliary classifiers. FR+FCN (Zhu et al., 2014b) used five CNNs to extract features of the pairs of whole faces or facial components to directly recover the canonical views of 2D face images.

2.1.3. Variants of CNN

Except the traditional CNN based architectures, some research works (in Table 3) have been done to investigate unconventional CNN based framework by (1) designing totally different layout of CNNs, (2) modifying kernel learning components or the way of kernel activation, (3) fusing CNNs with other types of modules, (4) adopting weakly-supervised or unsupervised learning, and (5) others.

Design different layout of multiple CNNs. Bilinear Convolutional Neural Network (BCNN) was introduced by Lin et al. (2015) which consists of two CNNs whose convolutional-layer outputs are multiplied (using outer product) at each location of the image. The resulting bilinear feature is pooled across the image resulting in an orderless descriptor for the entire image. Chowdhury et al. (2015) fine-tuned a trained base model of a symmetric BCNN to extract face features, and used subject-based SVM classifiers to identify individuals. Pyramid CNN (Fan et al., 2014) presented a pyramid-like structure of multiple CNNs. For each CNN, two images are fed into it and SIAMESE network is used to train it. Outputs are compared by the output neurons which predict whether the two face images are distinct. The pyramid CNN is trained in a greedy manner, after the first layer is well-trained, it trains the next layer. Output is a landmark-based multi-scale feature with a highly compact characteristic. Chowdhury et al. (2016) applied BCNN to the challenging face recognition benchmark, the IARPA Janus Benchmark A (IJB-A) (Klare et al., 2015), and achieved 89.5% rank-1 recall. Guided-CNN (Fu et al., 2017) used parallel sub-CNN models as the guide and learners. Li et al. (2015a) proposed a tree-structured convolutional architecture.

Modify the way of learning kernels. PCANet (Chan et al., 2015) combined principle component analysis (PCA) with deep neural networks to learn kernels. SPCANet (Tian et al., 2015a) employed PCA instead of the stochastic gradient descent (SGD) to learn filter kernels too. Spectral Regression Discriminant Analysis Network (SRDANet) (Tian et al., 2015b) is similar to SPCANet, but it uses eigenvectors as filter kernels. Weighted-PCANet (Huang and Yuan, 2015) learned features by combining Linear Regression Classification (LRC) model and PCANet

Table 2. Overview of deep learning methods based on Multi-CNN.

Algorithm	Description/Remark
SIAMESE (Wang et al., 2014)	Trained on different parts and scales of a face using a layer-wise training method; All face representations are concatenated as feature
MFRS (Zhou et al., 2015)	4 face regions are cropped for feature extraction and PCA for feature reduction
Baidu (Liu et al., 2015)	A two-stage approach combining multi-path deep CNN and deep metric learning; Extract overlapped image patches centered at different landmarks on face region; Concatenate representation together forming a high dimensional feature
DeepID (Sun et al., 2014b)	Each CNN takes a face region as input; Features are concatenated; All identities are classified simultaneously
DeepID2 (Sun et al., 2014a)	An ensemble of 25 CNNs trained on different local patches; Apply Joint Bayesian to obtain robust embedding space; Use identification and verification signals as supervision
DeepID2+ (Sun et al., 2015b)	Based on DeepID2, further combine verification and identification loss
DeepID3 (Sun et al., 2015a)	Joint identification-verification supervision added in final and a few intermediate layers
Kang et al. (2017)	Based on Multi-scale Convolution Layer Blocks (MCLBs); Stack MCLBs to present multi-scale abstraction; Use a deep ensemble; Extract two types of features from each DCNN and combine them to do FR
Xiong et al. (2017)	Explore complementarity of 2 DCNNs by training with two different large datasets
Bodla et al. (2017)	A deep heterogeneous feature fusion network for template-based face recognition
Lu et al. (2017c)	2 CNNs; Concatenate features of each CNN after PCA reduction
FR+FCN (Zhu et al., 2014b)	Contain five CNNs; Each takes a pair of whole faces or facial components (forehead, eye, nose and mouth) as input; Five CNNs are concatenated by fully connected layer to learn feature representation; Use a logistic regression layer to predict whether the two face images belong to the same identity

Table 3. Overview on some variants of CNN based framework.

Algorithm	Description/Remark
BCNN (Lin et al., 2015)	To bridge the gap between the texture models and part-based CNN models
Chowdhury et al. (2015)	Fine-tune a trained base-model of a symmetric BCNN to extract feature
Pyramid CNN (Fan et al., 2014)	Contain a group of CNNs divided into several levels with different depth and size, and they share some of the layers
Chowdhury et al. (2016)	Apply BCNN on IJB-A dataset
Guided-CNN (Fu et al., 2017)	Parallel sub-CNN models as guide and learners
Li et al. (2015a)	A tree-structure kernel adaptive CNN;Heirarchically fuse multiple local adaptive CNN subnets
PCANet (Chan et al., 2015)	PCA is employed to learn multistage filter banks
SPCANet (Tian et al., 2015a)	Stack multiple output features learned through each stage of the CNN as the input of nonlinear processing layer with hashing method-activation
SRDANet (Tian et al., 2015b)	Use leading eigenvectors from patches in facial image as filter kernels
Weighted-PCANet (Huang and Yuan, 2015)	Combine Linear Regression Classification model and PCANet construction to extract feature
MS-PCANet (Tian et al., 2016)	Multiscaled PCA Network
c-CNN (Xiong et al., 2015)	The samples in c-CNN are processed with dynamically activated sets of kernels; Kernels are only sparsely activated when a sample is passed through the network
LBPNet (Xi et al., 2016)	An unsupervised learning; Trainable kernels are replaced by LBP
Simón et al. (2016)	Fuse CNN and WNNC
NAN (Yang et al., 2017a)	Two modules: CNN based feature embedding and neural aggregation
ABTA (Dong et al., 2017a)	Two modules: attention based neural network, template adaptation module
Ranjan et al. (2016)	Employ a multi-task learning (MTL) framework to do multi-purpose task
Wu et al. (2017a)	ReST is introduced into CNN to do face alignment and recognition
SL-DCNN (Chen and Deng, 2016)	Weakly-supervised self-learning DCNN
JFL (Lu et al., 2015a)	Stack an unsupervised feature learning method into a deep CNN
Chen et al. (2015b)	An automatic end-to-end FR system: face detection, alignment and verification

construction, and shared the main construction characteristics with classical CNNs as a cascaded neural network. MS-PCANet (Tian et al., 2016) contains two convolutional layers to extract features hierarchically, followed by a nonlinear processing layer with a simple binary hashing and feature pooling, and it uses PCA to get the prefixed filter kernels. In the model proposed by Li et al. (2015a), the convolutional kernels are dynamically determined according to the spatial distribution of facial landmarks. The activations of kernels in Conditional CNN (c-CNN) (Xiong et al., 2015) model for each layer are conditioned on the present intermediate representation and the activation status in lower layers. Local Binary Pattern Network (LBPNet) (Xi et al., 2016) is a simplified deep network with handcrafted filters. It keeps the same topology of CNN whereas the trainable kernels are replaced by Local Binary Pattern (LBP). Two layers, using LBP and PCA filters respectively, are connected

hierarchically in the deep network to extract high-level overcomplete representations for face images.

Fuse CNNs with other types of modules. Weighted Nearest Neighbor Classifier (WNNC) (Simón et al., 2016) can be fused with a CNN classifier. It used RGB, depth and thermal captures of faces to train CNNs for a binary classification. Then the results were fused with Histograms of Gabor Ordinal Measures (HOGOMs) (Chai et al., 2014). Neural Aggregation Network (NAN) (Yang et al., 2017a) contains feature embedding module and neural aggregation module. The second module is composed of two content-based attention blocks. These attention blocks are driven by the memory storing all features extracted from the first module. Improved from NAN by combining transfer learning, Attention-Based Template Adaptation (ABTA) (Dong et al., 2017a) contains two modules also. The attention based neural network module (feature extractor) is used to integrate the template features of various lengths to a single fixed length feature representation, according to the attention mechanism, and template adaptation module is to transfer the knowledge from a hold-out dataset to the test templates to improve the performance via transfer learning. Ranjan et al. (2016) proposed a multi-purpose CNN architecture by employing a multi-task learning (MTL) framework to regularize the shared parameters of the network to simultaneously perform face detection, landmarks localization, face identification and verification, pose estimation, gender recognition, smile detection, and age estimation. Inspired by the spatial transformer, Wu et al. (2017a) introduced a Recursive Spatial Transformer (ReST) module into CNN to optimize face alignment and recognition jointly in an end-to-end fashion. The ReST can align faces to the canonical view in a progressive way, which can be considered as an alignment-free face recognition system.

Adopt weakly-supervised or unsupervised learning. SL-DCNN (Chen and Deng, 2016) is a weakly-supervised self-learning DCNN for face recognition. LBPNet (Xi et al., 2016) is a simplified deep network with handcrafted filters for unsupervised learning. JFL (Lu et al., 2015a) stacked an unsupervised feature learning module into a DCNN to learn a hierarchical feature representation. It used different feature dictionaries to represent the physical characteristics of various face regions, and learned multiple related feature projection matrices for these regions.

Others. Bayesian DCNN (B-DCNN) (Zafar et al., 2019) aims to improve the efficacy of face recognition by dealing with false positives through employing model uncertainty for robust surveillance systems. It gets the posterior distribution of class probabilities by employing dropout at both training and testing phases, and uses mean and variance of the samples as confidence and uncertainty for each class. In final classification stage, it uses a simple heuristic function to decide if the sample belong to the class.

2.2. Autoencoder (AE) and its Variants

Autoencoder (Bengio et al., 2009) is usually treated as one type of unsupervised networks, which are gaining more attentions in recent years. Similar to the multilayer perceptron (MLP), AE is a feedforward, non-recurrent neural network. It consists of two parts, encoder and decoder, with an input layer, an output layer and one or more hidden layers. Hidden layers have the purpose of reconstructing their own inputs (instead of predicting the target value Y, given inputs X), which forces hidden layers to try to learn good representations of the inputs. So AE is often used for efficient coding (Liou et al., 2014). The encoder maps the input to a code, latent variable, or latent representation. Decoder maps the code, latent variable, or latent representation to the reconstruction of input with the same shape. It usually applies the backpropagation technique to train the model layer-by-layer by setting the target values to be equal to the inputs, in which the goal is to minimize the reconstruction errors.

AE has a lot of variations. To further boost the ability of AE for image representation, Vincent et al. (2010) proposed a denoising autoencoder (DAE), which enhances the generalization by training with locally corrupted inputs. A DAE does two things: encode the input, and undo the effect of a corruption process. AE can be stacked to form a deep network, called stacked autoencoder (SAE), by feeding the latent representation of an autoencoder as input to the next autoencoder. Contractive autoencoder (CAE) and Variational autoencoder (VAE) are also the variants of AE.

AE is one of the commonly used building blocks in deep neural networks. A number of AE based deep methods (shown in Table 4) have been proposed recently. The extensive usage of AE is to learn common latent features between different domains for cross-domain FR. Besides, reducing the dimension of learned features or reconstructing images are common uses as well. For example, Huang et al. (2016b) proposed an Adaptive Deep Supervised Network Template (ADSNT) with a supervised autoencoder which is trained to extract characteristic features from corrupted/clean facial images and reconstruct the corresponding similar facial images.

Learn common latent features between different domains. This strategy is universally used in heterogeneous face recognition (HFR), such as cross-age, -large pose, -various expressions, etc. Riggan et al. (2015) proposed a coupled autoencoder, CpAEs, to learn a cross-modal transformation for HFR by forcing the hidden units (latent features) of two NNs to be as similar as possible, meanwhile preserving information from the input. Shao et al. (2015) used multiple AEs, where each AE generates input by randomly sampling data from another modality and the auxiliary database, and enforced the output to lie in a common feature space through Robust PCA. Coupled Autoencoder Networks (CAN) (Xu et al., 2017a) used AEs to handle the cross-age face recognition and retrieval problem. Deep Discriminant Analysis (DDA) Nets (Pathirage et al., 2016) adopted AEs to learn dynamic data adaptive features. Each shallow AE is trained to achieve simple but tractable goals required to address the global non-linear objective. This model can be used in various domains such as head pose and face expressions. Random faces guided sparse many-to-one encoder (RF-SME) (Zhang et al., 2013) is an AE-like high-level feature learning scheme to extract pose-invariant identity feature. It builds the encoder with a sparse constraint to extract pose-invariant feature in a supervised way, and uses multiple random

Table 4. Overview of deep methods based on AE and its variants.

Algorithm	Description/Remark
ADSNT (Huang et al., 2016b)	Extract characteristic features from corrupted/clean facial images and reconstruct the corresponding similar facial images
CpAEs (Riggan et al., 2015)	Coupled autoencoder for learning a target-to-source image representation for HFR
Shao et al. (2015)	Integrate multiple deep AEs with bagging strategy to deal with classification with missing modality problem
CAN (Xu et al., 2017a)	Coupled AE networks to handle age-invariant FR and retrieval problem
DDA (Pathirage et al., 2016)	Deep autoencoder for pose, expression
RF-SME (Zhang et al., 2013)	Extract pose-invariant identity feature
SPAE (Kan et al., 2014)	Stacked progressive AE; Learn pose-robust features
SFDAE (Pathirage et al., 2015)	Stacked face DAEs; A multiple-encoder single-decoder color fusion model
Liu et al. (2016a)	Fused 2D images of a face and motion history images with expressions to do FR
Gao et al. (2015)	Stack the supervised autoencoders (SSAE) to form deep architecture to extract features
D^2AE (Liu et al., 2018e)	Learn the identity-distilled features that discriminatively focus on inter-personal differences for identity verification with a minimal supervision by face identities

faces as the target values for the encoder to enhance discriminative capability of the feature. Kan et al. (2014) proposed a stacked progressive autoencoder (SPAE) to learn pose-robust features by modeling a complex non-linear transform from non-frontal face images to the frontal in a progressive way. SPAE contains multiple progressive AEs, and each of them maps faces at large poses to a virtual view at smaller pose angles. The output contains very small pose variations. Inspired by SPAE, stacked face denoising autoencoders (SFDAE) (Pathirage et al., 2015) was proposed for expression-robust feature acquisition. The model exploits contributions of different color components in different local face regions by recovering the neutral expression from various other expressions and denoises the face with dynamic expressions in a progressive way. Liu et al. (2016a) fused 2D images of a face and motion history images (MHIs), which are generated from the same face's image sequences with expressions to do face recognition. Motivated by DAE, Gao et al. (2015) proposed a supervised autoencoder to learn a robust image representation for the single training sample per person (SSPP) problem. It enforces faces with variations mapped to the canonical face and enforces features of the same person to be similar, and then it stacks the supervised autoencoders (SSAE) to form a deep architecture to extract features. Liu et al. (2018e) constructed an identity Distilling and Dispelling Autoencoder (D^2AE) framework with a minimal supervision by face identities to adversarially learn the identity-distilled features which can not only produce identity-distilled features that discriminatively focus on inter-personal differences with identity supervision, but also effectively extract the hidden identity-dispelled features to capture complementary knowledge including intra-personal variances and even background clutters.

2.3. Restricted Boltzmann Machine (RBM), Deep Belief Networks (DBNs) and Deep Boltzmann Machines (DBMs)

Boltzmann Machine (BM) is a particular form of log-linear Markov Random Field (MRF). The RBM is a variant of BM with the restriction that its neurons must form a bipartite graph, which means that a pair of nodes from each of the two groups of units (visible, hidden units) may have a symmetric connection between them and there are no connections between nodes within a group. RBM is a shallow, two-layer neural net. The first layer is called visible or input, and the second is the hidden layer. DBM (Salakhutdinov and Hinton, 2009) gained significant attentions in learning of higher-level and more complex representation of data and the distribution of observations. Nonlinear latent variables in DBM are organized in multiple connected layers in a way that variables in one layer can simultaneously contribute to the probabilities or states of variables in the next layer. Each layer learns a different factor to represent the variations in the given data. DBN can be formed by stacking Restricted Boltzmann Machines (RBMs) and optionally fine-tuning the resulting deep network with gradient descent and backpropagation.

In FR, there still exists some methods (see Table 5) using DBN, DBM and/or RBM. Chen et al. (2013c) proposed a feature learning method that first trains RBM networks for each modular region in the images, separately, and then stacks the RBM networks into a deep architecture to obtain high-level, hierarchical representations. Yi et al. (2015) adopted a 3-layer RBM to learn the relationship of face images between different modalities. Middle layer represents the shared properties of heterogeneous data. Both Huang et al. (2012b) and Jhuang et al. (2016) built a DBN based network to learn features. Wu et al. (2013) adopted DBM to obtain features under different poses and expressions. Deep Appearance Models (DAMs) (Duong et al., 2015) used two DBMs to robustly capture variations of facial shapes and appearances, respectively, and then construct one more high-level layer to interpret the connections between these two DBMs. Finally, a compact representation is generated for face classification.

2.4. Generative Adversarial Networks (GANs)

Another typical neural network, Generative Adversarial Network (GAN) (Goodfellow et al., 2014), gained much attention in recent years. It contains two independent networks, shown in Fig.5, which work separately and act as adversaries. The general idea is to build two competing neural network models. The generative model (generator) takes noise as input and generates samples. The discrimintive model (discriminator) receives samples from both generator and training data, and has to be able

Table 5. Overview of deep methods based on DBN, DBM, RBM.

Algorithm	Description/Remark
Chen et al. (2013c)	A feature learning method by stacking the RBM networks
Yi et al. (2015)	A local to global learning framework based on RBM for heterogeneous face recognition
CDBN (Huang et al., 2012b)	Convolutional deep belief networks to learn features in high-resolution face images
Jhuang et al. (2016)	Use DBN to train identification model using features with depth information of 3D data
Wu et al. (2013)	Use DBM to track facial feature under varying expressions and poses
DAMs (Duong et al., 2015)	2 DBMs capture variations of facial shapes and appearances respectively

Table 6. Overview of deep methods based on GAN.

Algorithm	Description/Remark
DA-GAN (Zhao et al., 2017)	For photo-realistic and identity preserving profile face synthesis under extreme poses by projecting a 3D face into the 2D image space
Age-cGAN (Antipov et al., 2017b)	A aging/rejuvenation method to synthesize more plausible and realistic faces
AgecGAN+LMA (Antipov et al., 2017a)	A generative aging/rejuvenation method
GAN-VFS (Zhang et al., 2017a)	Visible Face Synthesis method to synthesize photo realistic visible face images
DR-GAN (Tran et al., 2017)	GAN based framework for pose-invariant face recognition and face synthesis
UV-GAN (Deng et al., 2018a)	Pose-invariant face verification via adversarial facial UV map completion
DAN (Rao et al., 2017a)	A discriminative aggregation network for video face recognition
BLAN (Li et al., 2018b)	A bi-level adversarial network for makeup-invariant face verification
Song et al. (2017)	An adversarial discriminative feature learning framework for HFR
Cao et al. (2018a)	An asymmetric joint learning (AJL) method for HFR
Song et al. (2018)	Introduce adversarial learning for NIR-VIS face recognition

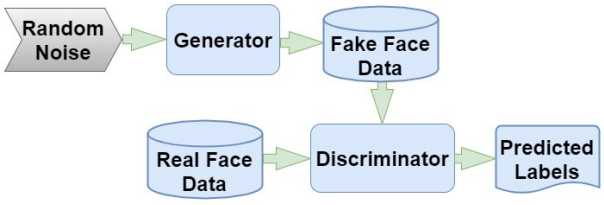

Fig. 5. Architecture of the Generative Adversarial Network.

to distinguish between the two sources. The two models play a continuous game, where the generator is learning to produce more and more realistic samples, and the discriminator is learning to get better and better in distinguishing the generated data from real data. The two models are trained simultaneously, and the goal is that the competition will drive the generated samples to be indistinguishable from real data.

GAN can be viewed as an architecture that is able to achieve far better performance compared to the traditional networks. This network has been universally applied to handle more complicated tasks in face recognition community (as shown in Table 6), such as some specific FR problem mentioned in Section 3. The common problem contains face synthesis, pose-invariant FR, cross-age FR, video-based FR, makeup-invariant FR, and so on. More GAN based methods will be described in Section 3.

Face synthesis. For example, DA-GAN (Zhao et al., 2017) was proposed for photorealistic and identity preserving profile face synthesis even under extreme poses. It combines prior knowledge from data distribution (adversarial training) and domain knowledge of faces (pose and identity perception losses) to exactly recover the lost information inherent in projecting a 3D face into the 2D image space. Antipov et al. (2017b) proposed Age-cGAN aging/rejuvenation method. Because of the slightly imperfect preservation of original identities in aged/rejuvenated faces, this model cannot be directly used to improve face verification. So, a generative AgecGAN+LMA aging/rejuvenation method (Antipov et al., 2017a) was proposed. It adopted a Local Manifold Adaptation (LMA) approach to resolve the stated issue of Age-cGAN. Zhang et al. (2017a) provided a GAN based Visible Face Synthesis (GAN-VFS) method to synthesize photo realistic visible face images from their corresponding polarimetric images.

Domain-invariant feature learning. Tran et al. (2017) developed a Disentangled Representation Learning GAN (DR-GAN) for pose-invariant face recognition and face synthesis. They designed an encoder-decoder structured generator and do pose classification in the discriminator. UV-GAN (Deng et al., 2018a) is a meticulously designed architecture that combines local and global adversarial DCNNs to learn an identity-preserving facial UV completion model for pose-invariant face recognition. Rao et al. (2017a) proposed a discriminative aggregation network (DAN) for video FR by combining the idea of adversarial learning with metric learning to aggregate the useful information of an input video into one or few more discriminative images in the feature space. Li et al. (2018b) proposed a bi-level adversarial network (BLAN) for makeup-invariant face verification. Two adversarial networks are combined in an end-to-end deep network, with the one on pixel level for reconstructing appealing facial images and the other on feature level for preserving identity information. To handle HFR problem, Song et al. (2017) proposed an adversarial discriminative

Table 7. Overview of deep methods using hybrid architectures.

Algorithm	Description/Remark
Gan et al. (2014)	Multi-layer network architecture; graph embedding framework
Nagpal et al. (2015)	SDAE; DBM; Learn weight invariant facial representations
Goswami et al. (2017)	SDAE; DBM; For crossmodality learning
Convnet-RBM (Sun et al., 2013)	CNN: characterize face similarities; RBM: perform inference
MM-DFR (Ding and Tao, 2015)	CNNs: extracts complementary facial features; SAE: compress dimension
MDLFace (Goswami et al., 2014)	SDAE: robust to noise; RBM: learn internal complex representation; DNN
McDFR (Chen et al., 2015c)	AE: extract generic feature of each facial regions; DNN: get discriminative feature; DNN:classification
Zhang et al. (2017b)	GAN: generative capacity; CNN: discriminative feature extraction

feature learning framework to close the gap between sensing patterns of different face modalities on both raw-pixel space and compact feature space. Cao et al. (2018a) proposed an asymmetric joint learning (AJL) method to transform the cross-modality differences mutually by incorporating the synthesized images into the learning process. Song et al. (2018) also introduced adversarial learning in NIR-VIS face hallucination and domain-invariant feature learning to close the sensing gap of heterogeneous data in pixel space and feature space simultaneously.

2.5. Hybrid Architectures

Some hybrid deep architectures were proposed for face recognition by combining two or more types of neural networks, e.g., AE+DBM, AE+CNN, GAN+CNN. An overview of existing hybrid deep architectures is shown in Table 7. Actually, AE can be treated as an efficient network for dimension reduction. Some hybrid deep architectures often adopt it to compress the high-dimensional feature vectors. Compared with the traditional PCA approach, AE has the advantage in learning non-linear feature transformations. For example, MM-DFR (Ding and Tao, 2015) integrated a set of elaborately designed CNNs and a three-layer SAE. The CNNs extract complementary facial features from multimodal data and the extracted features are concatenated to form a high-dimensional feature vector, whose dimension is compressed into a compact face signature by the SAE.

Most methods showed improvements on face recognition performance. Some of them combine AEs and DBMs. Nagpal et al. (2015) proposed a regularizer-based approach to learn weight invariant facial representations using sparse-stacked denoising AEs and DBMs. The experimental results showed an improvement on the identification accuracy. Goswami et al. (2017) built a deep learning framework for video face recognition with a combination of stacked denoising sparse autoencoder (SDAE) and DBM too. From the results, it is evident that both SDAE and DBM are required in the proposed architecture to extract a meaningful representation for face recognition and the joint representation can further improve the recognition performance. MDLFace (Goswami et al., 2014) presented an efficient memorability based frame selection algorithm using SDAE and DBM as well.

Some methods are the ensemble of CNNs and RBMs. Sun et al. (2013) proposed a hybrid CNN-RBM network. To characterize face similarities from different aspects, they concatenated the features extracted from different face region pairs by different deep CNNs. A RBM is used for face verification. The result showed that the entire hybrid network can further improve the accuracy. Several methods are designed using AEs and a supervised deep neural network. McDFR (Chen et al., 2015c) adopted unsupervised and supervised learning in a cascaded fashion to produce a generically descriptive yet class-specific deep multi-channel representation. It performs deep autoencoder to extract generic feature of each facial region (right eye, left eye, nose, mouth), then features are fed into a supervised learning (DNN) to get discriminative representations for different classes.

GAN is also a choice used in hybrid deep architecture. Zhang et al. (2017b) combined the generative capacity of conditional GAN and the discriminative feature extraction of DCNN for cross-modality learning. They demonstrated an outstanding performance on heterogeneous face recognition.

2.6. Comparison of Different Network Architectures

Unlike conventional machine learning algorithms that require users to tell the computer what to do, break big problems down into many small ones, and precisely define tasks that the computer can easily perform, neural networks directly learn from observational data, figure out their own solutions to the problem at hand. Today, deep neural networks or deep learning can work well for many difficult learning tasks, e.g., face recognition.

This paper investigated several neural networks commonly used in face recognition. CNN is the mostly used. A typical use case for CNN is that users feed the network images and the network classifies the data. The filters consisting of trainable parameters in CNN can convolve in a given image spatially to detect spatial features like edges and shapes. Stacked layers of filters can be used to detect complex spatial shapes from the spatial features at every subsequent level. Hence CNN can successfully boil down a given image into a highly abstracted representation for predicting.

Another neural network commonly used in FR is the GAN, which consists of any two networks (although often a combination of Feed Forwards and CNNs), with one tasked to generate content (generative) and the other has to judge content (discriminative). Recently it has been largely used to deal with specific challenging FR problems, e.g., cross-age, pose, HFR, and come across some very impressive results. GAN is one of the few successful techniques in unsupervised machine learning, and is

quickly revolutionizing our ability to perform generative learning.

However, some neural networks, e.g., AE, DBM, DBN, have become less popular in FR due to certain reasons. Take AE for example. It turned out to be very difficult to optimize deep autoencoders using back propagation. With small initial weights, the back propagated gradient dies. Nowadays they are rarely used in practical applications.

2.7. Loss Functions

Loss function plays an important role in deep feature learning. In various deep neural networks, usually there is a loss layer, normally the final layer, which specifies how to penalize the deviation between the predicted and true labels in training. An effective loss function is the one that can improve the discriminative power of the deeply learned features. Intuitively, the learning should minimize the intra-class variations and maximize the inter-class differences. With the development of deep neural networks, various loss functions (as shown in Table 8) have been proposed.

Most loss functions can be generally divided into two groups: (1) sample-based loss and (2) set-based loss. Sample-based supervision processes each sample individually. Set-based supervision considers a set of images as a unified entity. An image set is a collection of instances of the same object/person from varying viewpoints, illuminations and poses, and exhibits different characteristics. A set contains richer information of the target than a single image and is potentially more useful for problems like face recognition. As Wen et al. (2016b) illustrated, set-based supervision can learn more discriminative features than just separable features with sample-based approaches.

2.7.1. Sample-based Supervision

Softmax Loss is a traditional sample-based supervision which is often used to predict a single class of K mutually exclusive classes. Ranjan et al. (2017) designed an L_2-Softmax Loss by adding an L_2-constraint to the softmax loss, which restricts the features to lie on a hypersphere of a fixed radius.

Contrastive Loss (Hadsell et al., 2006) runs over pairs of samples. It is one such approach where the features are learned with supervision of a loss computed with (positive or negative) pairs of samples. Triplet Loss (Schroff et al., 2015) aims at ensuring a face image of a specific person (anchor) is closer to other images of the same person (positive) than to images of any other persons (negative). Both loss functions share the goal to minimize the distances between the samples from the same class and to maximize the distances between the samples from different classes. Since contrastive loss and triplet loss often lead to a slow convergence, Sohn (2016) proposed a multi-class N-pair Loss to address this issue. This loss function can improve upon the triplet loss by pushing away multiple negative examples jointly at each update.

Marginal Loss (Deng et al., 2017a) was proposed to minimize intra-class differences and maximize inter-class distances by focusing on the marginal samples. Congenerous Cosine Loss (COCO) (Liu et al., 2017d) considers both feature discrimination and polymerization by directly optimizing and comparing the cosine distance (similarity) between features. It has the softmax property to make features discriminative and keeps the class centroid. Deep Correlation Feature Learning (DCFL) method (Deng et al., 2017b) brought in Correlation Loss, which can encourage a large correlation between the deep feature vectors and their corresponding weight vectors in softmax loss. In correlation loss, it applies a weight vector in softmax loss as the prototype of each class.

Chen et al. (2017a) designed the Noisy Softmax Loss to mitigate the early saturation issue that softmax will impede the exploration of SGD and lead the model to converge at a bad local-minima by injecting an annealed noise in softmax during each iteration. Ring Loss (Zheng et al., 2018) is a simple and elegant approach to normalize all sample features through a convex augmentation of the primary loss function (such as Softmax). It applies soft normalization, where it gradually learns to constrain the norm to the scaled unit circle while preserving convexity leading to more robust features. Ring loss can be used along with any other loss functions such as the softmax or large-margin softmax.

It is possible to encourage intra-class variance minimization when a large-margin strategy is introduced into the classification model. Wan et al. (2018) proposed a Large-Margin Gaussian Mixture Loss (Large-margin GM or L-GM) established on the assumption that the deep features of the training set follow a Gaussian Mixture distribution. Based on the features likelihood to the training feature distribution, L-GM loss is superior to softmax loss and its major variants in the sense that it can be readily used to distinguish abnormal inputs. Large-margin Softmax Loss (L-Softmax) (Liu et al., 2016b) was proposed to explicitly encourage intra-class compactness and inter-class separability for the learned features. It can not only adjust the desired margin but also avoid overfitting.

Angular Softmax Loss (A-Softmax) (Liu et al., 2017b) adds an angular margin to the softmax loss. It renders a geometric interpretation by constraining learned features to be discriminative on a hypersphere manifold, which intrinsically matches the prior that faces also lie on a non-linear manifold. Angular margin was introduced in L-Softmax and A-Softmax to make the classification boundary more compact. It was proved to be an effective way to further improve the face recognition performance. However, the angular margin function introduced in A-Softmax or L-Softmax is difficult to train and is sensitive to parameters. To ease this issue, Qi and Zhang (2018) proposed a simple Adaptive Angular Margin Loss (AAM). Wang et al. (2018b) introduced a kind of margin to the softmax loss function, i.e., Additive Margin Softmax Loss (AM-Softmax), which is more intuitive and interpretable. ArcFace (Deng et al., 2018b) adopted an Additive Angular Margin Loss to obtain more discriminative features for face recognition. It utilizes the arc-cosine function to calculate the angle between the current feature and the target weight, add an additive angular margin to the target angle, and get the target logit back again by the cosine function. To improve the effectiveness of discrimination, Wang et al. (2018c) reformulated the softmax loss as a cosine loss called Large Margin Cosine Loss (LMCL) by L_2 normalizing both features and weight vectors to remove radial variations,

Table 8. Definitions of different loss functions used in deep networks for face recognition.

Loss function	Definition
Softmax Loss	$\bullet \mathcal{L}_s = -\sum_{i=1}^{m} log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}}$; m: classes; W: weights; b: bias
A-Softmax Loss (Liu et al., 2017b)	$\bullet \mathcal{L}_{ang} = \frac{1}{N}\sum_i -log\frac{e^{\|x_i\|\varphi(\theta_{y_i,i})}}{e^{\|x_i\|\varphi(\theta_{y_i,i})}+\sum_{j\neq y_i} e^{\|x_i\|\cos(\theta_{j,i})}}$; m($\geq 1$): an integer controling the size of angular margin; $\varphi(\theta_{y_i,i}) = (-1)^k\cos(m\theta_{y_i,i}) - 2k$; $\theta_{y_i,i} \in [\frac{k\pi}{m}; \frac{(k+1)\pi}{m}]$; $k \in [0, m-1]$
L-Softmax Loss (Liu et al., 2016b)	$\bullet \mathcal{L}_i = -log(\frac{e^{\|W_{y_i}\|\|x_i\|\varphi(\theta_{y_i})}}{e^{\|W_{y_i}\|\|x_i\|\varphi(\theta_{y_i})}+\sum_{j\neq y_i} e^{\|W_j\|\|x_i\|\cos(\theta_j)}})$; $\varphi(\theta) = \begin{cases}\cos(m\theta), & \text{if } 0 \leq \theta \leq \frac{\pi}{m} \\ \mathcal{D}(\theta), & \text{if } \frac{\pi}{m} \leq \theta \leq \pi\end{cases}$; m: integer closely related to classification margin; $\mathcal{D}(\theta)$: monotonically decrease; $\mathcal{D}(\frac{\pi}{m})$ should equal $\cos(\frac{\pi}{m})$
L_2-Softmax Loss (Ranjan et al., 2017)	$\bullet \mathcal{L}_{L_2} = -\frac{1}{M}\sum_{i=1}^{M} log\frac{e^{W_{y_i}^T f(X_i)+b_{y_i}}}{\sum_{j=1}^{C} e^{W_j^T f(X_i)+b_j}}$; minimizes $\mathcal{L}_{L_2}$ subject to $\|f(X_i)\|_2 = \alpha, \forall$ i=1,2,...,M; X_i: input in a mini-batch of size M; y_i: class label; $f(X_i)$: feature descriptor obtained from the penultimate layer; C:# classes; W,b: weights, bias for the last layer which acts as a classifier
Contrastive Loss (Hadsell et al., 2006)	$\bullet \mathcal{L}(W, Y, \overrightarrow{X_1}, \overrightarrow{X_2}) = (1-Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}max(0, m - D_W)^2$; m (> 0): a margin
Triplet Loss (Schroff et al., 2015)	$\bullet \|x_i^a - x_i^p\|_2^2 + \alpha < \|x_i^a - x_i^n\|_2^2, \forall (x_i^a, x_i^p, x_i^n) \in \tau$; $\mathcal{L} = \sum_i^N [\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha]_+$; α: a margin; τ: set of all possible triplets
N-pair Loss (Sohn, 2016)	$\bullet \mathcal{L}_{N-pair-mc}(\{(x_i, x_i^+)\}_{i=1}^N; f = \frac{1}{N}\sum_{i=1}^{N} log(1 + \sum_{j\neq i} exp(f_i^\top f_i^+ - f_i^\top f_i^+))$; x: input, x^+ and x^-: positive and negative examples of x; f: kernal taking x and generating an embedding vector f(x)
Marginal Loss (Deng et al., 2017a)	$\bullet \mathcal{L}_m = \frac{1}{m^2-m}\sum_{i,j,i\neq j}^{m}(\xi - y_{ij}(\theta - \|\frac{x_i}{\|x_i\|} - \frac{x_j}{\|x_j\|}\|_2^2))$; x_i, x_j: face samples; θ: threshold of distance; ξ: error margin besides the classification hyperplane; $y_{ij} \in \pm$: shows whether faces x_i and x_j are from same or different classes;
Correlation Loss (Deng et al., 2017b)	$\bullet \mathcal{L}_C = -\sum_i \cos(\theta_{y_i}) = -\sum_i \frac{W_{y_i}^T x_i}{\|W_{y_i}\|\|x_i\|}$; W_{y_i}: weight vector
Noisy Softmax (Chen et al., 2017a)	$\bullet \mathcal{L} = -\frac{1}{N}\sum_i log\frac{e^{f_{y_i} - \alpha\|W_{y_i}\|\|X_i\|(1-cos\theta_{y_i})+\epsilon|}}{\sum_{j\neq y_i} e^{f_j} + e^{f_{y_i} - \alpha\|W_{y_i}\|\|X_i\|(1-cos\theta_{y_i})+\epsilon|}}$; N:#training images; θ_{y_i}:the angle between vector W_{y_i} and X_i
Large-Margin GM Loss (Wan et al., 2018)	$\bullet \mathcal{L}_{GM} = \mathcal{L}_{cls} + \lambda\mathcal{L}_{lkd}$; $\mathcal{L}_{cls}$: related to the discriminative capability; $\mathcal{L}_{lkd}$: likelihood regularization, related to its probabilistic distribution; Both share all the parameters
COCO Loss (Liu et al., 2017d)	$\bullet \mathcal{L}_{COCO}(f^{(i)}, c_k) = -\sum_{i\in\mathcal{B},k} t_k^{(i)} log p_k^{(i)} = -\sum_{i\in\mathcal{B}} log p_{l_i}^{(i)}$; $f^{(i)}$: feature vector of i-th sample; $\mathcal{B}$: mini-batch; c_k: centroid of class k; k: index along the class dimension in $\mathcal{R}^K$; $t_k^{(i)} \in \{0, 1\}$: binary mapping of sample i based on its label l_i
Ring Loss (Zheng et al., 2018)	$\bullet \mathcal{L}_R = \frac{\lambda}{2m}\sum_{i=1}^{m}(\|\mathcal{F}(X_i)\|_2 - R)^2$; $\mathcal{F}(x_i)$: deep feature for sample X_i; R: learnt target norm value; λ: loss weight; m: batch size
Large Margin Cosine Loss (Wang et al., 2018c)	$\bullet \mathcal{L}_{lmc} = \frac{1}{N}\sum_i -log\frac{e^{s(cos(\theta_{y_i},i)-m)}}{e^{s(cos(\theta_{y_i},i)-m)}+\sum_{j\neq y_i} e^{scos(\theta_j,i)}}$ subject to $W = \frac{W^*}{\|W^*\|}, x = \frac{x^*}{\|x^*\|}, cos(\theta_j, i) = W_j^T x_i$; N: #training samples; x_i: $i-$th feature vector corresponding to groundtruth class of y_i; W_j: weight vector of the $j-$th class; θ_j: the angle between W_j and x_i
AM-Softmax (Wang et al., 2018b)	$\bullet \mathcal{L}_{AMS} = -\frac{1}{n}\sum_{i=1}^{n} log\frac{e^{s(cos\theta_{y_i}-m)}}{e^{s(cos\theta_{y_i}-m)}+\sum_{j=1,j\neq y_i}^{c} e^{scos\theta_j}}$
Adaptive Angular Margin Loss (Qi and Zhang, 2018)	$\bullet \mathcal{L}_{AAM} = \sum_i^N -log(p_{y_i}^{AAM})$; $p_{y_i}^{AAM} = \frac{exp(\|\Phi(X_i)\|cos(\eta\theta_{y_i,i}))}{exp(\|\Phi(X_i)\|cos(\eta\theta_{y_i,i}))+\sum_{k\neq y_i} exp(\|\Phi(X_i)\|cos(\theta_{k,i}))}$ η: an adaptive parameter,set based on the value of $\theta_{y_i,i}$
Additive Angular Margin Loss (Deng et al., 2018b)	$\bullet \mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s(cos(\theta_{y_i}+m))}}{e^{s(cos(\theta_{y_i}+m))}+\sum_{j=1,j\neq y_i}^{n} e^{scos\theta_j}}$; m:additive angular margin penalty; N: # batch size; n:class number; θ_j: the angle between the weight and feature
Center Loss (Wen et al., 2016b)	$\bullet \mathcal{L}_C = \frac{1}{2}\sum_{i=1}^{m}\|x_i - c_{y_i}\|_2^2$; $c_{y_i} \in \mathcal{R}^d$: y_i-th class center; x_i: input vector; m: # classes
Contrastive-Center Loss (Qi and Su, 2017)	$\bullet \mathcal{L}_{ct-c} = \frac{1}{2}\sum_{i=1}^{m}\frac{\|x_i - c_{y_i}\|_2^2}{(\sum_{j=1,j\neq y_i}^{k}\|x_i - c_j\|_2^2)+\delta}$; δ: constant for preventing denominator equal to 0
Range Loss (Zhang et al., 2017c)	$\bullet \mathcal{L}_R = \alpha\mathcal{L}_{R_{intra}} + \beta\mathcal{L}_{R_{inter}}$; α, β: weights; $\mathcal{L}_{R_{intra}}$: intra-class loss; $\mathcal{L}_{R_{inter}}$: inter-class loss
Git Loss (Calefati et al., 2018)	$\bullet \mathcal{L} = \mathcal{L}_S + \lambda_C\mathcal{L}_C + \lambda_G\mathcal{L}_G = -\sum_{i=1}^{m} log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} + \frac{\lambda_C}{2}\sum_{i=1}^{n}\|x_i - c_{y_i}\|_2^2 + \lambda_G\sum_{i,j=1,i\neq j}^{m}\frac{1}{1+\|x_i - c_{y_j}\|_2^2}$ $\mathcal{L}_S$: Softmax; $\mathcal{L}_C$:Center Loss; $\mathcal{L}_G$:Git Loss, to maximize distance between divergent identities
Max-Margin Loss (Gecer et al., 2017)	$\bullet \mathcal{L}_M = \lambda_M\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1-\delta_{(y_i=j)}}{m-1} e^{-\frac{\delta_{(y_i=j)}(\omega_j^T x_i + b_j)}{\|\omega_j\|_2}}$; n: #samples of a mini-batch; m: # classes; ω_j, b_j: parameters

based on which a cosine margin term is introduced to further maximize the decision margin in the angular space.

2.7.2. Set-based Supervision

Yet, recent set-based deep embedding studies show a good performance. Wen et al. (2016b) attempted to combine sample-based loss functions (e.g., softmax, contrastive, triplets) with a set-based term called Center Loss which simultaneously learns a center for deep features in each class and penalizes the distances between the deep features and their corresponding class centers. While the center loss only considers intra-class compactness, Contrastive-Center Loss (Qi and Su, 2017) improved it and considered both intra-class compactness and inter-class separability by penalizing two contrastive values, i.e., distances of input to its corresponding class centers and the sum of the distances of input to its non-corresponding class centers. Inspired by the contrastive loss, Range Loss (Zhang et al., 2017c) was proposed to utilize the tailed data in training, which can reduce the overall intra-personal variations and enlarge inter-personal differences simultaneously. However, unlike the contrastive loss defined on individual positive and negative pairs, range loss is defined on the overall distances between all sample pairs within one mini-batch. Git Loss (Calefati et al., 2018) is a joint supervision signal to leverage softmax and center loss functions aiming at minimizing the intra-class variations and maximizing the inter-class distances. Gecer et al. (2017) proposed Max-Margin Loss that benefits from set-based information by drawing inter-set (inter-class) margins. It improves the separability of learned features by maximizing the maximum possible inter-class margin that is calculated by a support vector machine and address the shortcomings of the existing set-based methods.

2.7.3. Other Loss Functions

There exists some specific loss functions that are not used extensively, e.g., Verification Loss and Classification Loss used in DeepID2 (Sun et al., 2014a), DeepID2+ (Sun et al., 2015b). Zhang et al. (2015a) provided a Sigmoid Cross-entropy loss for predicting K independent probability values in [0,1]. Kazemi et al. (2018) designed an Attribute-Centered Loss for soft-biometrics guided face sketch-photo recognition. In unsupervised deep learning, there are also some loss functions, such as Reconstruction Error used in AE and its variants (Zhu et al., 2013), Square-Loss function (Chen et al., 2015c; Pathirage et al., 2015), Coupling Error (Zhou et al., 2015), etc.

2.7.4. Comparison of Different Loss Functions

A thorough comparison of different loss functions used in deep learning based face recognition is presented in Table 9. Most methods adopts CNN as their basic network, e.g., ResNet1 (Wen et al., 2016b), ResNet2 (He et al., 2016a), Inception-ResNet (Schroff et al., 2015) and VGG-net (Simonyan and Zisserman, 2014). CASIA-WebFace (Yi et al., 2014) and MS-Celeb-1M (Guo et al., 2016) are two common public training datasets. CASIA-WebFace contains 0.49M face images belonging to 10K different individuals. MS-Celeb-1M contains 10M face images of 100K subjects. MS1MV2 is a semi-automatic refined version of MS-Celeb-1M. VGG Face (Parkhi et al., 2015) consists of around 1M face images of 2,558 individuals.

Four typical testing datasets, LFW (Huang et al., 2007), YTF (Wolf et al., 2011), IJB-A (Klare et al., 2015) and MegaFace (Kemelmacher-Shlizerman et al., 2016), are adopted. LFW includes 13,233 face images from 5,749 different identities, and provides 6,000 face pairs for verification protocol under unrestricted conditions. YTF includes 3,425 videos from 1,595 different individuals, with an average length of 181.3 frames per video. Both datasets contains faces with large variations in pose, expression and illuminations. IJB-A contains 500 subjects with a total of 25,813 images including 5,399 still images and 20,414 video frames. It contains faces with extreme viewpoints, resolution and illumination which makes it more challenging than the commonly used LFW dataset. MegaFace is a very challenging testing benchmark for large-scale (million scale) face identification and verification under two protocols (large or small training set), which contains a gallery set and a probe set. The gallery set in Megaface is composed of more than 1 million face images from 690K different individuals. The training set is defined as large if it contains more than 0.5M images and 20K subjects and vice versa. The probe set has two existing databases: Facescrub (Ng and Winkler, 2014) and FGNET (FG-NET, 2007). Facescrub contains 100K photos of 530 unique individuals. It is a common probe set used for evaluating Megaface performance. FGNet is a face ageing dataset with 1,002 images from 82 identities.

Most loss functions perform well on LFW. FV accuracy of more than half loss functions surpass 99.0%. Additive Angular Margin Loss gained the highest accuracy on YTF which is 98.02. The overall performance on IJB-A and MegaFace is much worse than LFW which proved that both datasets are more challenging than LFW.

2.8. Activation Functions

Sigmoid. The sigmoid activation function takes a real-valued number and squashes it into the range between 0 and 1. However, the sigmoid is rarely used in deep networks because of two drawbacks: when the activation of a neuron saturates at either tail of 0 or 1, the gradient there is almost zero, resulting in almost no signal flowing through the neuron to its weights, and recursively to its data; and the sigmoid outputs are not zero-centered.

Tanh. Tanh squashes a real-valued number to the range of [-1, 1]. Like sigmoid neuron, its activations saturate, but unlike the sigmoid neuron, its output is zero-centered. Therefore, in practice the tanh nonlinearity is preferable than the sigmoid.

ReLU. Rectified Linear Units (ReLU) has been used largely in the last few years. It increases the nonlinear properties of the decision function and overall network without affecting the receptive fields of the convolution layer. ReLU is preferable to other functions, because it trains the neural network several times faster (Krizhevsky et al., 2012) without a significant penalty to generalization capability. Compared to tanh and sigmoid neurons that involve expensive operations, ReLU can be implemented by simply thresholding a matrix of activations at

Table 9. Performance (%) of different loss functions on LFW, YTF, IJB-A and MegaFace datasets in face recognition community. * denotes the images are not publicly available. + denotes data expansion. Ver. indicates verification TAR for 10^{-6} FAR. TAR and FAR denote True Accept Rate and False Accept Rate, respectively. "Rank-1" indicates rank-1 identification accuracy with 1M distractors.

Loss function	Models	Training Data	LFW	YTF	IJB-A	MegaFace		
						Ver.	Rank-1	Protocol
Softmax	VGG-net	WebFace	96.53					
	VGG-net	WebFace+	98.83	94.22				
	ResNet1	MS-Celeb-1M	98.87	94.16	73.00			
	ResNet2	MS-Celeb-1M			83.10			
	ResNet2	WebFace	97.88	93.10		65.925	54.855	small
	ResNet2	WebFace+Celeb1M	98.27	93.10				
	Inception-ResNet	VGGFace2(3.31M)	98.40	93.60				
	Liu et al. (2017d)	half MS-1M	99.75				71.17	small
A-Softmax	ResNet2	WebFace	99.42	95.00		89.142	75.766	small
	ResNet2	MS-Celeb-1M	99.47		**93.24**	90.045	75.766	small
L-Softmax	VGG-net	WebFace	98.71					
	ResNet2	WebFace	99.1	94.00		80.423	67.128	small
L_2-Softmax	ResNet1	MS-Celeb-1M	99.78	96.08	90.90			
Contrastive Loss	ResNet2	WebFace	98.78	93.5		78.865	65.219	small
+ Softmax	VGG-net	WebFace	97.31					
Triplet Loss	FaceNet (Schroff et al., 2015)	200M*	99.63	95.12				
	ResNet2	WebFace	98.7	93.4		78.322	64.797	small
	Liu et al. (2017d)	half MS-1M	98.85				69.13	small
Triplet Loss +Softmax	Liu et al. (2017d)	half MS-1M	99.68				70.22	small
N-pair Loss	CasiaNet (Yi et al., 2014)	WebFace	98.33					
Marginal Loss	ResNet1	MS-Celeb-1M	99.48	95.98		92.640	80.278	large
	Inception-ResNet	WebFace	98.95					
Correlation Loss	ResNet2	WebFace	99.55	96.06				
Noisy Softmax	VGG-net	WebFace+	99.18	94.88				
L-GM Loss	ResNet2	WebFace	99.20					
COCO Loss	Liu et al. (2017d)	half MS-1M	**99.86**				76.57	small
Ring Loss +Softmax	ResNet2	MS-Celeb 1M	99.52	93.70	91.5			
Ring Loss +A-Softmax	ResNet2	MS-Celeb 1M	99.50		93.22			
Large Margin	ResNet2	WebFace	99.33	96.1		92.22	79.54	small
Cosine Loss	ResNet2-100	MS1MV2 (5.8M)				96.56	80.56	
AM-Softmax	ResNet2	WebFace	99.17			84.44	72.47	
AAM Loss	Inception-ResNet	WebFace	99.583	95.28			73.743	small
Additive Angular	ResNet2	WebFace	99.53			92.34	77.50	
Margin Loss	ResNet2-100	MS1MV2 (5.8M)	99.83	**98.02**		**96.98**	**81.03**	
Center Loss	ResNet1	0.7M*	99.28	94.9		76.516	65.234	small
	ResNet1	WebFace	99.05					
	ResNet2	MS-Celeb-1M	99.17					
	ResNet2	WebFace	99.00	94.4		75.68	63.38	
	Inception-ResNet	VGGFace2(3.31M)	99.20	95.10				
Center Loss +Softmax	ResNet2	WebFace	99.05	94.4		80.146	65.494	small
	Liu et al. (2017d)	half MS-1M	99.78				75.79	small
	ResNet2	MS-Celeb-1M			88.32			
Center Loss +A-Softmax	ResNet2	MS-Celeb-1M	99.52		89.26			
Contrastive-Center Loss	ResNet1	WebFace	98.68					
Range Loss	ResNet2	WebFace+Celeb1M	99.52	93.70				
Git Loss	Inception-ResNet	VGGFace2(3.31M)	99.30	95.30				
Max-Margin Loss	Inception-ResNet	VGG Face(0.83M)	96.03	92.44				

zero. ReLU activation offers a way to separate noisy data from informative signals. It uses a threshold (or bias) to determine

Table 10. Description of common activation functions.

Activation function	Definition
Sigmoid	$f(x) = (1 + e^{-x})^{-1}$
Tanh	$f(x) = tanh(x) = \frac{2}{1+e^{-2x}} - 1$
ReLU	$f(x) = \begin{cases} x, & \text{if } x > 0 \\ 0, & \text{if } x \leq 0 \end{cases}$
LReLU	$f(x) = \begin{cases} x, & \text{if } x > 0 \\ 0.01 * x, & \text{if } x \leq 0 \end{cases}$
PReLU	$f(a, x) = \begin{cases} x, & \text{if } x > 0 \\ a * x, & \text{if } x \leq 0 \end{cases}$
RReLU	$f(a_i, y_i) = \begin{cases} y_i, & \text{if } y_i > 0 \\ a_i * y_i, & \text{if } y_i \leq 0 \end{cases}$
ELU	$f(a, x) = \begin{cases} x, & \text{if } x > 0 \\ a(e^x - 1), & \text{if } x \leq 0 \end{cases}$
Maxout	$max(w_1^T x + b_1, w_2^T x + b_2)$
Gaussian	$\Phi(z) = e^{-\frac{z^2}{2\sigma^2}}$
Thin Plate Spline	$\Phi(z) = z^2 logz$
Quadratic	$\Phi(z) = (z^2 + r^2)^{1/2}$
Inverse Quadratic	$\Phi(z) = \frac{1}{(z^2+r^2)^{1/2}}, z = \|\|x - c_j\|\|$

the activation of each neuron. If a neuron is not activated, its output value will be 0. However, this thresholding might lead to the loss of some information, especially for the first several convolution layers, because these layers are similar to Gabor filters (i.e., both positive and negative responses are respected). To alleviate this problem, the Leaky Rectified Linear Units (LReLU) (Maas et al., 2013), Parametric Rectified Linear Units (PReLU) (He et al., 2015a) and Exponential Linear Units (ELU) (Clevert et al., 2015) were proposed.

LReLU. The motivation of LReLU is to avoid zero gradients. Experiments in (Maas et al., 2013) showed that the LReLU has negligible impact on the accuracy compared with ReLU. Instead of the function being zero when x<0, a LReLU will instead have a small negative slope (of 0.01, or so). Some researchers reported success with this form of activation function, but the results are not always consistent. The slope in the negative region can also be made into a parameter for each neuron, as seen in PReLU neurons. However, the consistency of the benefit across tasks is unclear presently.

PReLU. PReLU was proposed by He et al. (2015a). RReLU is the Randomized Leaky Rectified Linear Unit (Xu et al., 2015). The negative slope can be set to different values. In Table 10, a is a coefficient controlling the slope of the negative part. When $a = 0$, it becomes ReLU; when a is a learnable parameter, it is referred to PReLU. The PReLU is equivalent to $f(x) = max(0, x) + a \cdot min(0, x)$. If a is small and fixed, PReLU becomes LReLU ($a = 0.01$). PReLU can be trained using back-propagation and optimized simultaneously with other layers.

Maxout. Maxout (Goodfellow et al., 2013) generalizes ReLU and its leaky version. It has the benefits of a ReLU unit (linear regime of operation, no saturation), while does not have its drawbacks. However, unlike ReLU, it doubles the number of parameters for every single neuron, leading to a higher number of parameters in total. Max-Feature-Map was proposed with the Light CNN (Wu et al., 2015). It can be treated as an extension of Maxout activation. Different from Maxout activation that uses enough hidden neurons to approximate an arbitrary convex function, MFM suppresses only a small number of neurons to make the CNN models light and robust.

Gaussian radial function, thin plate spline (Duchon, 1977), quadratic, and inverse quadratic are often used in the hidden units of RBFN. Although RBFN exhibits several advantages, e.g. global optimal approximation and classification capabilities and has been found to be very attractive for many engineering problems, including face recognition (Oh et al., 2013; Balasubramanian et al., 2009; Park et al., 2008), it are not considered as deep learning methods for face recognition.

3. Some Specific Face Recognition Problems

Face recognition in visible domain has received a considerable amount of attention. Deep learning has significantly improved the performance of conventional face recognition to near human-levels. Besides, there exists some specific face recognition problems including challenges in Still Image-based Face Recognition (SIFR) (e.g., pose variations, cross-age, illumination changes), Video-based Face Recognition (VFR), Heterogeneous Face Recognition (HFR) (e.g., still-to-video, 3D-based, NIR-VIS, sketch-photo face matching), Image Set-based Face Recognition (ISFR), and Closed-set vs. Open-set Face Recognition. Fig.4 (c) shows the numbers of publications for each specific face recognition issues. Nearly half of the papers focused on the challenges in SIFR problem.

3.1. Challenges in Still Image-based Face Recognition (SIFR)

In the past decade, face recognition has made a significant progress in controlled scenarios, e.g., mugshot. Recently, researchers focus more on unconstrained face recognition, containing various poses, illuminations, expressions, blur, ages and occlusions. In developing deep learning techniques, there are deep methods that focus on some specific face recognition problems, using CNN, AE, GAN, etc. Fig.4 (d) gives a paper distribution for each challenge. Pose variations has drawn the greatest attention to researchers.

3.1.1. Pose Variations

Pose variation (as shown in Fig.6 (a)) in face images is still a challenge for FR. Pose-Invariant Face Recognition (PIFR) is far from being solved. A recent study (Sengupta et al., 2016) shows that the performance of most algorithms including deep learning methods degrades over 10% from frontal-frontal to frontal-profile face verification, while the human performance only drops slightly. This indicates that the pose variation remains a significant challenge, even with deep learning. Table 11 presents an overview of the related methods.

Existing PIFR methods can be grouped into four categories: (1) employing face frontalization to synthesize a frontal-view face image before feature extraction, (2) directly extracting pose-invariant features from non-frontal face images, (3) performing both strategies jointly, and (4) other strategies.

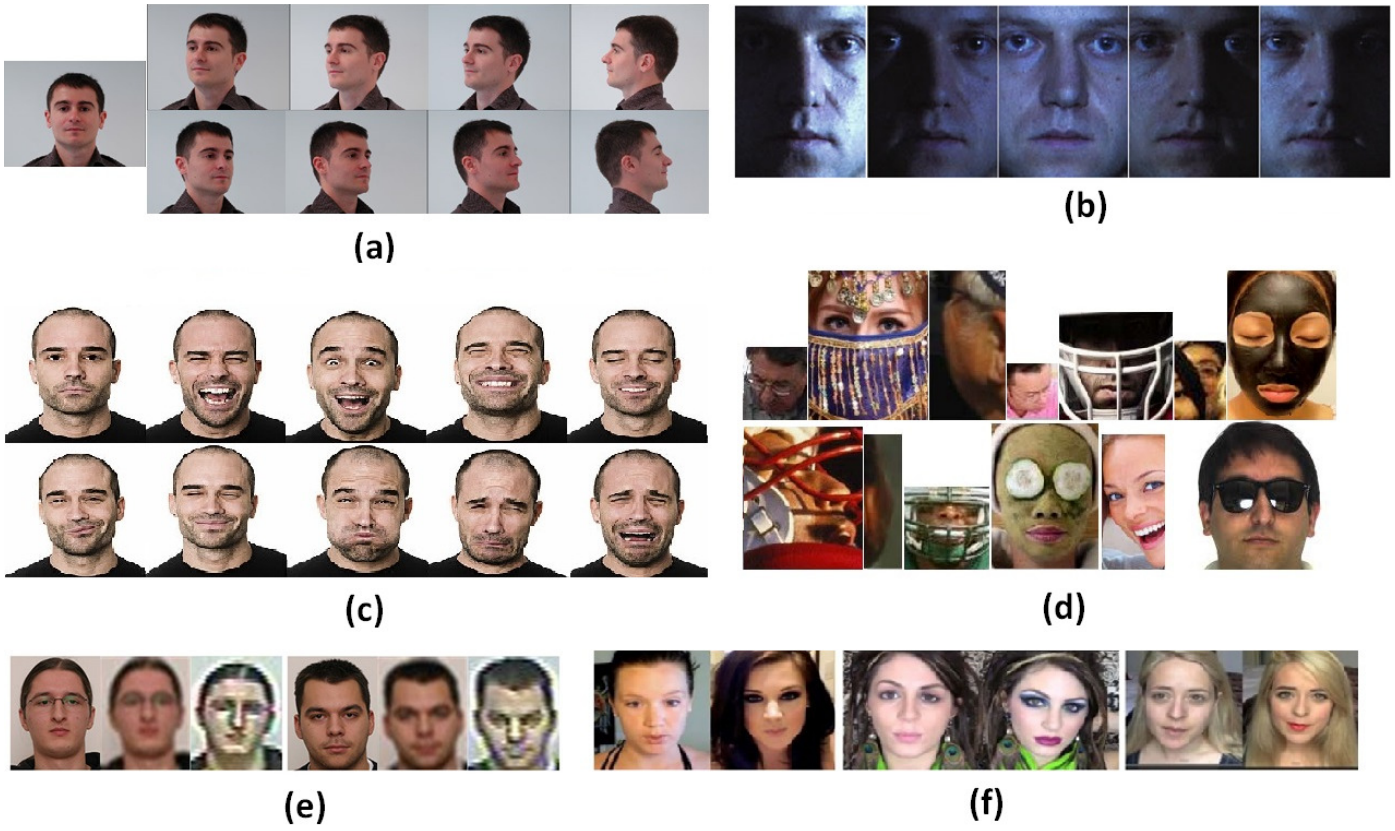

Fig. 6. Examples of (a) face pose variations, (b) illumination changes, (c) expression variations, (d) facial occlusions, (e) low resolution, (f) facial makeup.

Normalize profile face images to frontal view before feature extraction. The first strategy is to employ face frontalization to synthesize a frontal view face image. The ability of generating a realistic frontal face can be beneficial to deal with pose to some extent. For example, Kan et al. (2014) adopted multiple progressive autoencoders to do face frontalization. Hu et al. (2017b) proposed an end-to-end deep neural network to transform a non-frontal face image into a frontal view by learning the displacement field, which reflects the shifting relationship of pixels from the non-frontal face image and the transformed frontal view. Face Identity-Preserving (FIP) (Zhu et al., 2013), Multi-View Perceptron (MVP) (Zhu et al., 2014a), Controlled Pose Feature (CPF) (Yim et al., 2015) did face frontalization too. Besides pose variation, they can also handle other variations, e.g., illumination. Cao et al. (2018c) assumed that there is an inherent mapping between frontal and profile faces, and consequently, their discrepancy in the deep representation space can be bridged by an equivariant mapping. To exploit this mapping, they formulated a Deep Residual EquivAriant Mapping (DREAM) block, which is capable of adaptively adding residuals to the input deep representation to transform a profile face representation to a canonical pose that simplifies recognition.

The face-in-the-wild conditions show more challenges on illumination, head pose variation, self-occlusion and so on. It is a challenging task to frontalize faces in the wild under various head poses, including extreme profile views. Yin et al. (2017) proposed a deep 3D Morphable Model (3DMM) conditioned Face Frontalization Generative Adversarial Network (FF-GAN), to generate neutral head pose face images from a single input image that can be a profile view up to 90. Huang et al. (2017) proposed a deep architecture, Two-Pathway Generative Adversarial Network (TP-GAN), for photorealistic frontal view synthesis via considering the inference of global structure and the transformation of local texture, respectively.

Directly extract pose-invariant features from non-frontal face images. The second group focuses on learning a pose-invariant representation directly from non-frontal face images through either one joint model or multiple pose-specific models. There are a lot of algorithms using one joint model. For example, Random faces guided sparse many-to-one encoder (RF-SME) (Zhang et al., 2013) used a sparse many-to-one encoder to extract discriminative features. Xu et al. (2017b) proposed a 3D-aided 2D face recognition system. Peng et al. (2017) designed a reconstruction loss to regularize identity feature learning and adopted a data driven synthesis strategy to enrich the diversity of poses. Lu et al. (2017a) used a joint model to make the identity metrics more pose-robust for face verification by mitigating the information contained in the pose verification task. Amounts of methods with more than one pose-specific models are also designed. Almageed et al. (2016) and Masi et al. (2016a) built multiple DCNN models to deal with pose variations. Zhao et al. (2018c) incorporated a simulator (3D Morphable Model) to obtain shape and appearance prior and leveraged a global local GAN to enhance the realism of both global structures and local details of the face simulators output, while preserving the identity information.

Table 11. Overview of deep learning methods for handling pose variations.

Algorithm	Model	Description
Kan et al. (2014)	SAE	Stacked progressive AE; Transform faces from non-frontal to frontal progressively
Hu et al. (2017b)	DNN	Transform non-frontal faces into frontal by learning displacement field
FIP (Zhu et al., 2013)	AE-like	Reconstruct corresponding face under frontal-view
MVP (Zhu et al., 2014a)	CNN	Rotate a face with any pose to a target pose
CPF Yim et al. (2015)	DNN	Rotate the arbitrary pose face into several target pose faces
Cao et al. (2018c)	CNN	Formulate a Deep Residual EquivAriant Mapping block to transform a profile face representation to a canonical pose that simplifies recognition
FF-GAN (Yin et al., 2017)	GAN	Handle pose variations at extreme poses by incorporating elements from deep 3DMM and FR CNNs to achieve high-quality and identity-preserving frontalization
TP-GAN (Huang et al., 2017)	GAN	Present a global and local perception GAN framework for frontal view synthesis
RF-SME (Zhang et al., 2013)	SME	Extract pose-invariant feature using an sparse many-to-one encoder framework
Seo et al. (2015)	CNN	4 tasks; Two is used to minimize intra-pose variation and preserve pose continuity
Xu et al. (2017b)	DNN	3D-aided 2D FR system; Robust to pose variations as large as 90°
Peng et al. (2017)	DNN	Learn reconstruction-based pose-invariant feature without extensive pose coverage in training data
Lu et al. (2017a)	CNN	A joint model for face and pose verification tasks; Explicitly discourage the information sharing between pose and identity verification metrics
Almageed et al. (2016)	CNN	Multiple pose-aware DCNN models reducing sensitivity to pose variations
Masi et al. (2016a)	CNN	Use multiple pose-specific models and render face images to handle pose variation
Zhao et al. (2018c)		A 3D-aided model which automatically recovers realistic frontal faces from arbitrary poses
Yin and Liu (2018)	CNN	A pose-directed multi-task CNN; Group poses to learn pose-specific identity feature
DR-GAN (Tran et al., 2017)	GAN	Jointly merge face frontalization and pose-invariant identity representation learning
PIM (Zhao et al., 2018b)	GAN+CNN	Jointly learn face frontalization and pose invariant representations end-to-end to allow them to mutually boost each other
PAM (Masi et al., 2019a)	CNN	Trains multiple Pose-Aware Models and effectively exploits these models when matching images with faces appearing in different poses
Lin and Fan (2011)	DBN	Deal with the non-linearity caused by pose variations
Grm et al. (2016)	CNN	PISI; Use a DPSL strategy to handle large pose variations
UV-GAN Deng et al. (2018a)	GAN	Increase pose variations for training deep FV models; Minimizes pose discrepancy during testing by attaching the completed UV to the fitted mesh and generating instances of arbitrary poses

Perform both strategies jointly. Actually, it is more desirable to perform both tasks jointly to allow them to benefit from each other. Tran et al. (2017) merged and leveraged these two categories through a Disentangled Representation learning-GAN (DR-GAN) to handle the pose challenge. Zhao et al. (2018b) proposed a Pose Invariant Model (PIM) to jointly learn face frontalization and pose invariant representations end-to-end to allow them to mutually boost each other. Pose-Aware Models (PAM)(Masi et al., 2019a) was designed to explicitly tackle pose variations via processing a face image using several pose-specific, deep CNNs. And 3D rendering is used to synthesize multiple face poses from input images to both train these models and to provide additional robustness to pose variations in testing.

Other strategies. Contrary to these three categories, several methods adopted different strategies. Lin and Fan (2011) used a DBN to deal with the nonlinearity caused by pose variations or low resolution by learning the relationship between high resolution (HR) manifold and low resolution (LR) manifold. Pose-Invariant Similarity Index (PISI) model (Grm et al., 2016) adopted a deep pair-wise similarity learning strategy (DPSL). It takes two grayscale facial images with different poses as input and outputs a similarity index. A value of index close to 1 indicates that the input image pair represents the same subject, while a value close to 0 indicates different subjects. Deng et al. (2018a) proposed a UV-GAN. First, it trains DCNN to complete the facial UV map extracted from in-the-wild images. To this end, it gathers complete UV maps by fitting a 3D Morphable Model (3DMM) to various multiview image and video datasets. And then it combines local and global adversarial DCNNs to learn an identity-preserving facial UV completion model.

3.1.2. Cross-Age

Cross-age face recognition has remained a popular research topic as most regular facial recognition systems could fail in dealing with facial changes through aging, yet it still lacks sufficiently reliable solutions. Since facial appearance is subject to significant intra-class variations caused by the sophisticated aging process which poses different nonlinear effects on different individuals over time, age-invariant face recognition (AIFR) remains a major challenge in face recognition community (Guo et al., 2010). The appearance of a human face changes substantially over time, resulting in significant intra-class variations. The appearance changes can be different in different age groups.

Recently, deep learning has been applied to cross-age face recognition problem (as shown in Table 12). Existing methods can be grouped into four categories: (1) directly extracting age-

Table 12. Overview of deep learning methods for cross-age face recognition.

Algorithm	Model	Description
Li et al. (2015b)	CNN	Deep joint metric learning framework to learn age-invariant features
Wen et al. (2016a)	CNN	A latent factor guided CNN; Construct latent identity analysis module to help extract age-invariant feature
Zheng et al. (2017a)	CNN	An age estimation task guided CNN; Learn age-invariant features on training data with age label and identity label
Xu et al. (2017a)	AE	Coupled AE networks to handle age-invariant FR and retrieval problem
Wang et al. (2017d)	CNN	Cross-age FV by setting FV as primary learning task and age estimation as auxiliary learning task
Bianco (2017)	CNN	A feature injection layer; Further improve the discriminative power
Li et al. (2018c)	CNN	Present a distance metric optimization driven learning approach integrating traditional steps via DCNN
Wang et al. (2018d)	CNN	Propose an Orthogonal Embedding CNNs (OE-CNNs) to learn age-invariant deep face features
Li et al. (2018a)	CNN	Propose an age-related factor guided joint task modeling convolutional neural networks
Antipov et al. (2017b)	GAN	Age-cGAN; Synthesize aging/rejuvenation of the input face images to some predefined age categories to handle age variant
Antipov et al. (2017a)	GAN	Resolve the issue that Age-cGAN cannot be directly used for improving face verification
Zhao et al. (2018a)	GAN	Propose a deep Age-Invariant Model (AIM) for face recognition in the wild

invariant features for recognition, (2) synthesizing a face that matches target age before feature extraction, and (3) performing both tasks jointly.

Directly extract age-invariant features. A lot of methods (Li et al., 2015b; Wen et al., 2016a; Zheng et al., 2017a; Xu et al., 2017a) tried to directly learn age-invariant features using various deep neural networks, e.g., CNN and AE. Li et al. (2015b) designed a deep joint metric learning framework to learn age-invariant features. A latent factor guided CNN (Wen et al., 2016a) constructed latent identity analysis module to help extract age-invariant features. Zheng et al. (2017a) used an age estimation task guided CNN to learn age-invariant features on training data with the age labels and identity labels. Xu et al. (2017a) adopted a coupled AE network to handle age-invariant FR problem. Wang et al. (2017d) proposed a multi-task deep neural network architecture for cross-age face verification, which can effectively balance feature sharing and feature exclusion between face verification and age estimation, by exploiting an intrinsic, shared low-dimensional representation. Bianco (2017) proposed a deep CNN architecture to handle large age-gap face verification by adding a feature injection layer which can injects externally computed features into the deepest layers in the network. The discriminative power of the network is further improved. Li et al. (2018c) proposed a distance metric optimization driven learning approach for age invariant face recognition. Wang et al. (2018d) designed a deep learning method to learn age-invariant components from features by decomposing face features into age-related and identity related components, where the identity-related component is used for FR. Li et al. (2018a) proposed an age-related factor guided joint task modeling CNNs, which combines an identity discrimination network with an age discrimination network that shares the same feature layers. By alternatively training the fusion networks and the combined factor model, the cross-age identity features and cross-identity age features can be effectively separated with high inter-class distension and intra-class compactness.

Synthesize a face that matches target age before feature extraction. Some methods tried to handle the cross-age problem by synthesizing a target-matching age face before feature extraction. For example, Antipov et al. (2017b) proposed an Age-cGAN aging/ rejuvenation method, allowing to synthesize more plausible and realistic faces than alternative non-generative methods. Based on Age-cGAN, Local Manifold Adaptation (LMA) approach (Antipov et al., 2017a) was then proposed to address the problem when the Age-cGAN cannot be directly used.

Perform both strategies jointly. There are some methods that perform both tasks jointly. Zhao et al. (2018a) used an unified deep architecture jointly learning disentangled identity representations that are invariant to age and performing photorealistic cross-age face image synthesis that can highlight an important latent representation.

3.1.3. Illumination Changes

Illumination changes (as shown in Fig.6 (b)) may cause huge differences of facial shading or shadow from varying directions or energy distributions of the ambient lighting, together with the 3D structure of faces. Lighting condition is one of the big factors for facial appearance changes and recognition performance degradation. It is possible that the difference between two images of the same person taken under varying illuminations to be greater than the difference between images of two different persons under the same illumination. Table 13 gives a brief overview of deep learning methods for handling illumination changes. Thakare and Thakare (2011) used a fuzzy-neural network to deal with depth information of face images for feature matching. Face Identity-Preserving (Zhu et al., 2013), Multi-View Perceptron (Zhu et al., 2014a) and Controlled Pose Feature (Yim et al., 2015) are three methods that can be used to handle both pose variations and illumination changes. Choi et al. (2016) used a DCNN model to eliminate the illumination effect and maximize the discriminative power for feature representation.

3.1.4. Partial Face Images

The unavailability of the whole faces (as shown in Fig. 6 (d)) is another challenge in an unconstrained environment. Par-

Table 13. Overview of deep learning methods for dealing with illumination changes.

Algorithm	Model	Description
Thakare and Thakare (2011)	FNN	Use the normalized depth map of 3D face data to handle illumination changes
FIP (Zhu et al., 2013)	AE-like	Reconstruct corresponding face under neutral light
MVP (Zhu et al., 2014a)	CNN	Rotate a face with any pose and illumination to a target pose
CPF (Yim et al., 2015)	DNN	Rotate the arbitrary pose, illumination face into several target pose faces
Choi et al. (2016)	CNN	Illumination-reduced feature learning method to eliminate illumination effect

tial face images occur when a face is: (1) occluded by objects such as faces of other individuals, sunglasses, hats, beard, masks or scarves; (2) captured in various poses without user awareness; 3) positioned partially out of the camera's filed of view. Partial face recognition (PFR) has become an emerging problem with increasing requirements for identification from CCTV cameras and embedded vision systems in mobile devices, robots and smart home facilities. However, PFR is challenging, without a solution from traditional face recognition approaches. Trigueros et al. (2017) proposed a method to find out which parts of the face are more important to achieve a high recognition rate, and used that information during training to force the CNN to learn discriminative features from all face regions, including those that typical approaches tend to pay less attention to.

Most existing deep learning based face recognition algorithms require fixed-size face images as inputs. In order to match the size, most of them usually re-scale the original images to a fixed-size. However, the performance of these methods could be affected by the undesired geometric deformation. So several methods are proposed to directly handle arbitrary-size input images. He et al. (2016b) proposed a Multi-Scale Region-based Convolutional Neural Network (MR-CNN) model which extracts features of each sub-region of a partial face and does partial face recognition using region-to-region matching. He et al. (2018a) introduced a Dynamic Feature Matching (DFM) method. It applies a Fully Convolutional Network (FCN) to extract spatial feature maps of given gallery and probe faces, and then decomposes the gallery feature maps into several gallery sub-feature maps by setting up a sliding window with the same size as the probe feature maps. In the end, it does alignment-free dynamic feature matching via Sparse Representation Classification (SRC).

3.1.5. Facial Makeup

Nowadays people are more likely to enhance their facial attractiveness by using the makeup. They can easily smooth face skin, change the shape of eyebrows, accentuate eye regions, alter lip colour, and so on, with appropriate cosmetic products, to hide facial flaws and improve the perceived attractiveness. However, these operations bring about remarkable facial appearance changes as exhibited in Fig.6 (f), resulting in both global and local appearance discrepancies between makeup and non-makeup face images.

Most of the existing face verification methods rely much on the various cues and information captured by the effective appearance features. These methods inherently lack robustness over the application of makeup that is non-permanent as well as miscellaneous, and challenge the face recognition performance (Guo et al., 2014; Zheng and Guo, 2016). Recently, Li et al. (2018b) proposed a bi-level adversarial network (BLAN) to settle the makeup-invariant face verification problem via a learning from generation framework. This framework simultaneously considers makeup removal and face verification, and is implemented by an end-to-end two adversarial networks, with one in pixel level for reconstructing appealing facial images, and the other in feature level for preserving the identity information.

3.1.6. Facial Expression Variations

Facial expression changes (as shown in Fig. 6 (c)) may impose difficulties for face recognition too. Facial deformations with expressions can change the appearance. Researchers have used deep learning methods to address the expression problems. For example, Pathirage et al. (2015) proposed a stacked denoising autoencoder for expression-robust feature acquisition. It exploits contributions of different color components in different local face regions by recovering the neutral expression from various other expressions, and processes the faces with dynamic expressions progressively. Liu et al. (2016a) fused 2D images of a face and motion history images (MHIs), which are generated from the same subject's image sequences with expressions to do face recognition.

3.1.7. Mixed Variations

Deep learning methods are good at dealing with nonlinear variations in face images and making the extracted features more discriminative. Rather than focusing on one specific variation, there are a number of methods proposed, as shown in Table 14, to address multiple mixed challenges, e.g., pose, illumination, expressions, age.

Pose+Illumination. FIP (Zhu et al., 2013), MVP (Zhu et al., 2014a) and CPF (Yim et al., 2015) were proposed to deal with pose and illumination problems, by rotating a face with any pose and illumination to a canonical view. Wu and Deng (2016) used a simplified architecture of the one proposed in CPF. Unlike FIP which only has a normalization task, CPF introduced an auxiliary reconstruction task that reconstructs the original input image from the output of the normalization task, to improve the identity-preserving ability of the DNN. The idea is that the output of normalization task should be identity-preserving and contains sufficient information of the identity to reconstruct the input image.

Pose+Expression. Deep Discriminant Analysis (DDA) Nets (Pathirage et al., 2016), can learn dynamic data adaptive features used for various problems such as face pose and expres-

Table 14. Overview of deep learning methods for handling mixed variations.

Algorithm	Model	Description
FIP (Zhu et al., 2013)	AE-like	Reconstruct corresponding face under frontal-view and neural light
MVP (Zhu et al., 2014a)	CNN	Rotate a face with any pose, illumination to a target pose
CPF (Yim et al., 2015)	DNN	Rotate the arbitrary pose, illumination face into several target pose faces
Wu and Deng (2016)	DNN	Build a pose, illumination normalization NN with much less training data
DDA (Pathirage et al., 2016)	AE	Learn dynamic data adaptive features used for pose, expression domains
Li et al. (2015a)	CNN	Tree-structure Kernel Adaptive CNN to disentangle irrelevant non-rigid appearance variations of viewpoint and expression changes
Ding and Tao (2015)	CNNs+SAE	Jointly learn face representation with pose, illumination, expression issues
Yin and Liu (2018)	CNN	A multi-task CNN for pose, illumination, expression (PIE) estimations
Sun et al. (2014a)	CNN	Extract deep identification-verification features with various face regions and resolutions; Handle pose, illumination, expression, ages, occlusion challenges
Zhu et al. (2014b)	CNN	Directly transform original images to canonial view handling multiple challenges
Hu et al. (2017b)	DNN	Deal with pose and other variations by learning the displacement field

sions. Li et al. (2015a) proposed a tree-structure Kernel Adaptive CNN to disentangle such irrelevant non-rigid appearance variations of viewpoint and expression.

Pose+Illumination+Expression. Yin and Liu (2018) and Ding and Tao (2015) proposed methods to handle pose, illumination, and expression (PIE) changes. Ding and Tao (2015) used a comprehensive deep learning framework to jointly learn a face representation with pose, illumination and expression issues.

Multiple Challenges There are also some deep models used for overcoming multiple challenges. DeepID2 (Sun et al., 2014a) can extract deep identification-verification features from images with various face regions and resolutions to deal with challenges including pose, illumination, expression, ages, occlusions. Zhu et al. (2014b) proposed a deep learning framework that can transform original images to a canonical view, which can also deal with other challenges. Hu et al. (2017b) used a deep network to deal with pose and other variations by learning the displacement field.

3.2. Video-based Face Recognition (VFR)

Compared to still image-based face recognition (SIFR), video-based face recognition (VFR) is significantly more challenging. Still-images are usually captured or framed under better conditions. Even through videos offer a myriad of data for face modeling, sampling, and recognition, the image quality of video frames tends to be significantly lower and faces exhibit much richer variations because the video acquisition may be much less constrained. For example, subjects in videos are usually mobile, resulting in serious motion blur, out-of-focus blur, and a large range of pose variations. Furthermore, surveillance and mobile cameras are often low-cost (and therefore low-quality) devices, which further exacerbates problems with video frames.

Recent advances in face recognition have tended to ignore the peculiarities of videos when extending techniques from SIFR to VFR (Schroff et al., 2015; Sun et al., 2015b; Parkhi et al., 2015; Li and Hua, 2015). On one hand, a major difficulty in VFR, such as severe image blur, is largely unsolved (Beveridge et al., 2015). One important reason is that large amounts of real-world video training data are still lacking, and existing still image databases are usually blur-free. On the other hand, although pose variations and occlusion are partially solved in SIFR by ensemble modelling (Sun et al., 2015b; Liu et al., 2015), the strategy may not be directly extended to VFR.

However, due to the increasing number of CCTV cameras installed and the easy availability of video recordings, an enormous quantity of videos are constantly being captured. Compared to still face images, videos usually contain more information, e.g., temporal and multi-view information. The ubiquity of videos offers society far-reaching benefits in terms of security and law enforcement. It is highly desirable to build surveillance systems coupled with face recognition techniques to automatically identify subjects of interest. VFR has emerged as a more and more important research topic. Unfortunately, the majority of existing face recognition literature focuses on matching of still images, and VFR research is still in its infancy. Even though, a few algorithms have been developed to utilize varying approaches, ranging from frame by frame matching to advanced deep learning architectures. The key issue is to build an appropriate visual representation of the video faces, such that it can effectively integrate the information across different frames together. Table 15 gives an overview of VFR models, which can be divided into two groups: (1) Methods that performed on images can be used on videos; (2) Methods that specially targeted for VFR.

Methods that performed on images can do videos also. While image-based face recognition has been studied extensively, many deep learning approaches can perform both image and video based face recognition, such as DDML (Hu et al., 2014), DeepFace (Taigman et al., 2014), DeepID2+ (Sun et al., 2015b), FaceNet (Schroff et al., 2015), Light CNN (Wu et al., 2015), VGGFace (Parkhi et al., 2015), etc. He et al. (2015b) proposed a predictable hash code algorithm to map face samples (images or video) in the original feature space to the Hamming space.

Methods that specially targeted for VFR. There exists a few methods specially targeted VFR. Zou et al. (2012) proposed an unsupervised learning method for learning invariant features from videos using a temporal slowness principle. Parchami

Table 15. Overview of deep learning methods for video based face recognition.

Algorithm	Model	Description
DDML (Hu et al., 2014)	DNN	Present a new discriminative deep metric learning (DDML) method
DeepFace (Taigman et al., 2014)	CNN	Use 3D face modeling to apply piecewise affine tranformation to get features
DeepID2+ (Sun et al., 2015b)	CNN	Combine verification+identification loss to get discriminative feature
FaceNet (Schroff et al., 2015)	CNN	An end-to-end system; Map face to a compact Euclidean space where distances directly correspond to a measure of face similarity
Light CNN (Wu et al., 2015)	CNN	Light CNN with reduced parameters & time to learn 256-D embedding
VGGFace (Parkhi et al., 2015)	CNN	Combine the very deep convolution neural network
He et al. (2015b)	CNN	A predictable hash code algorithm; map face samples in original space to Hamming space
Zou et al. (2012)	CNN	An unsupervised learning algorithm for learning invariant features from video using the temporal slowness principle
Parchami et al. (2017a)	CNN	Extract discriminative embedding of still ROI and compared with ROIs of video
Sohn et al. (2017)	CNN	Feature-level domain adaptation approach to learn domain-invariant features
ASML (Hu et al., 2017c)	CNN	Measure the statistical characteristics of image sets for VFR
Rao et al. (2017a)	GAN-like	Integrate information from video frames effectively and efficiently by combining metric learning and adversarial learning
Rao et al. (2017b)	CNN	An attention-aware deep reinforcement learning framework; Seek the focuses of attention in video
Goswami et al. (2014)	SDAE+DBM	Automatic memorability based frame selection algorithm for feature extraction
Goswami et al. (2017)	SDAE+DBM	Get feature-rich frames by discrete wavelet transform& entropy computation
Dong et al. (2016)	CNN	An input aggregated network; learn fixed-length representations for variable length videos
Yang et al. (2017a)	CNN	Build an attention based model to aggregate features of video frames
Wang et al. (2017b)	CNN	A framework with triplet loss to identify few suspects from the crowd in real time for public video surveillance
Wang et al. (2017e)	DNN	A method for face recognition in real-world surveillance videos
Grundström (2015)	CNN	Focus on real-time VFR using two feature types: local feature representations around landmark points and deep representations extracted from CNN
Ding and Tao (2018)	CNN	Use training data composed of both still images and artificially blurred data to learn blur-insensitive features
Liu et al. (2018c)	CNN	Resorts to actor-critic reinforcement learning for sequential attention decision of each image embedding
Kim et al. (2018)	CNN	Take advantage of face and body association (FBA) for VFR
Sharma et al. (2016)	DBN	Use Generalized mean Deep Learning Neural Network

et al. (2017a) proposed a CNN based method to extract discriminative embeddings of still regions of interest (ROI) and then compare with regions of interests (ROIs) in videos. Sohn et al. (2017) proposed an image to video feature-level domain adaptation approach to learn some domain-invariant discriminative representations for VFR. It uses a pre-trained face recognition engine on labeled still images to extract discriminative information, adapts them to video domain by synthetic data augmentation and then learns a domain-invariant feature through a domain adversarial discriminator. ASML (Hu et al., 2017c) is an Attention-Set based Metric Learning method proposed to measure the statistical characteristics of image sets for VFR.

Generally speaking, existing algorithms (Rao et al., 2017a,b; Goswami et al., 2014, 2017) either select a small number of frames from all available frames, or use all frames to extract information-rich features. Dong et al. (2016) and Yang et al. (2017a) proposed a representation with a compact, fixed-size visual representation for video faces, irrespective of the varied lengths of video clips. Wang et al. (2017b), Wang et al. (2017e) and Grundström (2015) proposed methods to handle real-world or real-time video surveillance. Wang et al. (2017b) built a DCNN framework with a triplet supervisory signal to identify few suspects from the crowd in real time for public video surveillance. Wang et al. (2017e) proposed a method for face recognition in real-world surveillance videos by constructing a face dataset and fine-tuning the VGGFace model (Parkhi et al., 2015). Grundström (2015) focused on real-time video face recognition using two distinct feature types: local feature representations around landmark points and deep representations extracted from CNN. Ding and Tao (2018) proposed a Trunk-Branch Ensemble CNN model (TBE-CNN) to efficiently extract blur-robust representations of the holistic face image and facial components by sharing the low- and middle-level layers of different CNNs. Liu et al. (2018c) presented a dependency-aware attention control (DAC) network, which resorts to actor-critic reinforcement learning for sequential attention decision of each image embedding to fully exploit the rich correlation cues among the unordered images. Kim et al. (2018) generated a more robust face representation to associate the annotated face in multiple shots and scenes by using jointly the information coming from new, detected faces along with the body appearance.

3.3. Heterogeneous Face Recognition

Heterogeneous face recognition (HFR), aka. cross-modality face recognition, is the problem of matching faces across dif-

Table 16. Overview of deep learning methods for generative heterogeneous face recognition.

Algorithm	Model	Description
Srivastava and Salakhutdinov (2012)	DBM	A generative model; Extract an unified representation with multiple modalities; Then fuse the features together
Ding and Tao (2015)	CNNs+SAE	Use CNNs to extract complementary facial features from multimodal data; Features are concatenated to form a high-dimensional feature
Yi et al. (2015)	RBMs	Extract Gabor features at localized facial points; Use RBMs to learn shared representations locally and connected together
Saxena and Verbeek (2016)	CNN	Explore different metric learning strategies to reduce discrepancies between different modalities
Kan et al. (2016)	Deep Net	A multi-view deep network including view-specific sub-network (removing view-specific variations) and common sub-network (finding common representation shared by all views)
Song et al. (2017)	GAN	An adversarial discriminative feature learning framework to close the gap between sensing patterns of different face modalities on both raw-pixel space and compact feature space
DA-JL (Cao et al., 2018b)	CNN	A data augmentation-based joint learning approach to mutually transform the cross-modality differences by incorporating synthesized images into learning process
Peng et al. (2019)	CNN	A deep local descriptor learning framework; Directly learn deep local descriptor from raw local facial patches
MC-CNN (Deng et al., 2019)	CNN	Mutual Component CNN; modal-invariant framework without massive data
Riggan et al. (2015)	AE	A coupled AEs for learning a target-to-source image representation
Zhang et al. (2017b)	GAN+CNN	Combine the generative capacity of conditional GAN and the discriminative feature extraction of DCNN for crossmodality learning
Cao et al. (2018a)	GAN	An asymmetric joint learning (AJL) method to transform the cross-modality differences mutually by incorporating the synthesized images into learning process
Wu et al. (2017b)	CNN	A coupled DL approach; Transform HFR problem into homogeneous face matching problem by seeking a shared feature space
Wu et al. (2018b)	CNN	Proposed a coupled deep learning (CDL) approach to seek a shared feature space in which the heterogeneous face matching problem can be approximately treated as a homogeneous face matching problem
Liu et al. (2018a)	CNN	A deep face attributes guided representation based method (DAG-HFR) to directly map face images in heterogeneous scenarios to a compact common space

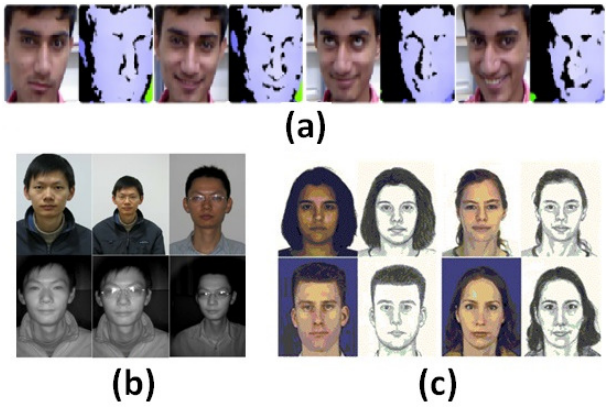

Fig. 7. Some face example: (a) 3D-VIS, (b) NIR-VIS, (c) Sketch-Photo.

ferent modalities such as between the visible and near-infrared images. Guo (2014) presented several specific HFR problems, including visible light image (VIS) vs. 3D (Fig.7 (a)), VIS vs. Near Infrared image (NIR) (Fig.7 (b)), VIS vs. Sketch (Fig.7 (c)), VIS vs. Video, Cross-resolution, ID vs. Selfie, etc.

HFR has become important due to its wider range of practical applications in surveillance, authentication, law enforcement, and forensic verification. Nevertheless, HFR poses a variety of serious challenges beyond conventional homogeneous face recognition. The main challenges lie in the large modality discrepancy, such as comparing single versus multi-channel imagery, linear and non-linear variations in intensity value due to different specular reflection properties, different coordinate systems, reduction of appearance detail, non-rigid distortion preventing alignment (Ouyang et al., 2016a), etc., and insufficient training samples.

To reduce the gap between sensing patterns of different face modalities, a wide variety of approaches have been proposed. More recently, deep learning based approaches have emerged as potentially viable techniques to tackle the cross-domain face recognition problem by learning a common latent embedding between the two modalities. The primary approaches are to extract common latent features between different modalities, so that a classifier trained on one modality may generalize to another. Table 16 gives an overview of generative HFR methods used for multiple scenarios of face matching between different modalities. In the following subsections, several typical deep learning based HFR problems are discussed, e,g., Still-to-Video, NIR-VIS, Sketch-Photo, Cross-resolution, 3D based, ID-Selfie. The paper distribution can be seen in Fig. 4 (e).

Existing generative HFR methods (shown in Table 16) could be classified into three categories: (1) Feature descriptor based

methods, (2) Synthesis based methods and (3) Common space projection based methods. Feature descriptor based methods aim to directly extract modality invariant features for recognition (Ding and Tao, 2015; Yi et al., 2015; Saxena and Verbeek, 2016; Kan et al., 2016; Song et al., 2017; Cao et al., 2018b; Peng et al., 2019; Deng et al., 2019). Synthesis based methods firstly transform images in one modality to another, which would make these images as homogeneous scenarios, and then conventional homogeneous face recognition methods could be directly utilized (Riggan et al., 2015; Zhang et al., 2017b; Cao et al., 2018a). Common space methods attempt to project heterogeneous face images into a latent common space where the probe image and the gallery images could be matched directly (Wu et al., 2017b, 2018b; Liu et al., 2018a).

3.3.1. Still-to-Video Face Recognition

Still-to-video (S2V) face recognition has real-world applications. Usually, the gallery set has higher resolution still images, while the probe is video clips with lower resolutions. Zhu et al. (2015b) addressed the S2V face recognition problem as a heterogeneous face matching and developed a domain adaptation method for S2V. Recently, some deep methods (see Table 17) have been proposed to bridge the gap between these two modalities. Existing still-to-video face recognition methods mainly contains four categories. (1) Feature descriptor based methods. Zhu and Guo (2016) did S2V face recognition with DCNN where face gallery is formed by a few still face images, and the query is video clips. Lin et al. (2017a) presented a pairwise similarity measure and unified it with feature representation learning via DCNN to handle S2V problem. (2) Synthesis based methods. For example, Parchami et al. (2017b) proposed an efficient Canonical Face Representation CNN (CFR-CNN) for S2V face recognition. It uses a supervised autoencoder network to generate canonical face representations from video regions of interest. (3) Common space projection based methods. Bao et al. (2017) transferred still and video face images to an Euclidean space, and adopted Euclidean metrics to measure the distance between still and video images. (4) Others. Savchenko and Belova (2017) addressed S2V face recognition for the small sample size problem using a statistical recognition method, which casts S2V into a Maximum A Posteriori estimation.

3.3.2. NIR-VIS Face Recognition

Infrared spectra have different regions: (1) reflection dominated region contains near infrared (NIR) and short-wave infrared (SWIR) bands; (2) emission dominated thermal region consists of mid-wave infrared (MWIR) and long-wave infra-red (LWIR) bands (Kong et al., 2005). The main advantage of thermal imaging is the acquisition in low light conditions where the visible light cameras cannot work. NIR images are close enough to the visible light spectrum to capture the structure of the face, while simultaneously being far enough to be invariant to visible light illumination changes. It offers the potential for face recognition where controlling the visible environment light is difficult or impossible, such as in night-time surveillance or automated gate control.

Recently, the problem of matching thermal probe face images against visible light image has attracted an increasing attention because of its much desired attribute of illumination invariance, and the decreasing cost of NIR acquisition devices. More and more deep learning based methods have been proposed to handle this problem as shown in Table 18. Traditional thermal to visible face verification methods first extract features from the visible and thermal images and then verify the identity based on the extracted features. Existing approaches attempt to tackle NIR-VIS face recognition using three strategies: (1) projecting heterogeneous data onto a common latent space for cross-modal matching, (2) extracting domain-invariant features from these modalities, and (3) synthesizing visible faces from NIR faces.

Projection based approaches. Reale et al. (2016) used coupled deep convolutional neural networks to map VIS and NIR faces into a domain independent, latent feature space in which two types of features can be compared directly. He et al. (2018b) mapped both NIR and VIS images to a compact Euclidean feature space and learned invariant features too. Wu et al. (2018a) explored an disentangled latent variable space to model NIR and VIS representations with intrinsic identity information and its within-person variations, which effectively reduces the NIR and VIS domain discrepancy and alleviates overfitting. Coupled Deep Convolutional Neural Network (CpD-CNN) (Iranmanesh et al., 2018) made full use of the polarimetric thermal information and found global discriminative features in a nonlinear embedding space to relate the polarimetric thermal faces to their corresponding visible faces. Ghosh et al. (2016) proposed an effective method to combine handcrafted and learned features for cross-resolution near infrared face recognition.

Feature based approaches. Riggan et al. (2016a) exploited the polarization state information of thermal emissions for polarimetric thermal-to-visible face recognition with a polarimetric thermal imaging technique. Liu et al. (2016c) applied the triplet loss to reduce intra-class variations among different modalities as well as augment the number of training sample pairs. Observation and results (Chen et al., 2012) demonstrated that the appearance of a face is composed of identity information and variation information (e.g., lighting, pose, and expression). Inspired by these observations, He et al. (2017) presented a deep convolutional network to learn modality Invariant Deep Representation (IDR) that contains the identity information of both NIR and VIS face images via mapping both NIR and VIS images to a compact Euclidean space. Lezama et al. (2017) adopted a cross-spectral hallucination and low-rank embedding to generate discriminative features for VIS and NIR face images. In order to close the sensing gap, Song et al. (2018) employed GAN with a two-path model to learn discriminative features.

Synthesis based approaches. Unlike the above mentioned traditional methods, synthesis-based thermal to visible face recognition algorithms leverage the synthesized visible faces for verification. Due to the success of CNNs and recently introduced GANs in synthesizing realistic images, various deep learning-based approaches have been proposed for thermal to

Table 17. Overview of deep learning methods for S2V face recognition.

Algorithm	Model	Description
Zhu and Guo (2016)	CNN	Study the choice of different similarity measures for face matching
Lin et al. (2017a)	CNN	Present a pairwise similarity measure unified with feature learning
Parchami et al. (2017b)	CNN+AE	Supervised AE to generate canonical representations from video ROIs
Bao et al. (2017)	CNN	Transfer still and video face images to an Euclidean space; Use Euclidean metrics to measure the distance between still and video images
Savchenko and Belova (2017)	CNN	Handle S2V for small sample size problem based on computation of distances between high-dimensional deep bottleneck features

Table 18. Overview of deep learning methods for NIR-VIS face recognition.

Algorithm	Model	Description
Reale et al. (2016)	CNN	Use coupled DCNN to map VIS & NIR faces into domain independent, latent feature space in which two types of features are compared
Ghosh et al. (2016)	SDAE+RBM	Cross-resolution near infrared face identification without preprocessing or enhancement
He et al. (2018b)	CNN	Map NIR, VIS images to a compact Euclidean feature space to learn invariant features
Wu et al. (2018a)	CNN	Used the Disentangled Variational Representation (DVR) for cross-modal matching
Iranmanesh et al. (2018)	CNN	Utilize both thermal and polarization state information to enhance the performance of a cross-spectrum face recognition system
Riggan et al. (2016a)	NN	A framework by exploiting the polarization state information of thermal emissions to facilitate training of a discriminant classifier
Liu et al. (2016c)	CNN	Apply triplet loss to reduce intra-class variations among different modalities as well as augment the number of training sample pairs
He et al. (2017)	DCNN	Learn an invariant deep representation by mapping both NIR and VIS images to a compact Euclidean space
Lezama et al. (2017)	CNN	Adopt a pre-trained VIS deep model (2 components: cross-spectral hallucination, low-rank embedding) to generate discriminative features for VIS and NIR face images
Song et al. (2018)	GAN	Uses cross-spectral face hallucination and discriminative feature learning to enhance domain-invariant feature learning and modality independent noise removing
Sarfraz and Stiefelhagen (2017)	DNN	Treat as a non-linear regression (perceptual mapping) directly between visible and thermal data on the features
Zhang et al. (2018)	GAN	Utilize image transformation techniques to thermal query images;do VIS domain FR
Di et al. (2018)	GAN+CNN	Uses the attributes extracted from visible image to synthesize attribute-preserved visible image from input thermal image for cross-modal matching
CFC (He et al., 2019)	GAN	Models HR heterogeneous face synthesis as a complementary combination

visible face synthesis (Riggan et al., 2018). Riggan et al. (2016b) and Zhang et al. (2017a) synthesized visible face images from thermal face images. One major advantage of these synthesis methods is that given the synthesized visible face images, any VIS face recognition method trained on VIS face data can be used to match the synthesized image to the enrolled VIS images.

Deep Perceptual Mapping (DPM) (Sarfraz and Stiefelhagen, 2017) directly learned a mapping from visible features to thermal or polarimetric features, or vice versa. Zhang et al. (2018) proposed a Thermal-to-Visible Generative Adversarial Network (TV-GAN) to transform thermal query face images into their corresponding visible light domain (VLD) images. The key point is that it can preserve sufficient identity information during the transformation. However, due to self-occlusion and sensing gap, NIR face images lose some visible lighting contents so that they are always incomplete compared to VIS face images. Di et al. (2018) proposed an Attribute Preserved Generative Adversarial Network (AP-GAN) to extend VIS face recognition methods to the NIR spectrum by synthesizing VIS images from thermal images guided by the extracted attributes from VIS image. He et al. (2019) proposed a GAN-based end-to-end deep framework, named Cross-spectral Face Completion (CFC), for generating a frontal VIS image of a person's face given an input NIR face image without assembling multiple image patches. It decomposes the unsupervised heterogeneous synthesis problem into two complementary problems, generating a texture inpainting component and a pose correction component.

The utilization of face synthesis methods from NIR to VIS poses opportunities as well as new challenges. Current synthesis results are less appealing in high-resolution due to two possible reasons. (1) Sensing gap. The sensory devices used to capture VIS and NIR face images are different and adopt different settings so that the visual appearances (e.g., geometric, textural) are significantly different, which leads to high intra-class variations and makes the high resolution synthesis of one spectrum from another very difficult. (2) Pose difference between VIS and NIR faces. Actually, VIS faces and NIR faces are often captured under different distances and environments, so it is hard to

simultaneously capture the VIS faces and NIR faces under the same pose. And pose variations often result in self-occlusion so that the texture of a NIR face image may be incomplete. Besides, the NIR-VIS datasets are usually in a small-scale, leading to over-fitting potentially. Therefore, cross-spectral face rotation is more challenging than face rotation in VIS domain.

3.3.3. Sketch Based Face Recognition (SBFR)

The problem of matching facial sketches to photos is commonly known as Sketch Based Face Recognition (SBFR). It typically involves a gallery dataset of visible light images and a probe dataset of facial sketches. According to the types of sketches available, existing sketch based face recognition methods can be grouped to four categories: (1) methods based on hand-drawn viewed sketch; (2) methods based on hand-drawn semi-forensic sketch; (3) methods based on hand-drawn forensic sketch; and (4) methods based on software-generated composite sketch. Face recognition from viewed sketch has been drawn much attention in the early time. Because the viewed sketches are drawn through viewing the mug shot photo directly, viewed sketches and photos are quite similar in terms of both the shape and texture.

Over the past decades, SBFR has been considered as an effective tool in law reinforcement to identify suspects by retrieving their photos automatically from existing police databases. In most cases, actual face photos of suspects are not available, only hand-drawn or computer generated sketches, based on the recollection of eyewitnesses as the clue. Therefore, an efficient automatic sketch-photo FR system is needed to identify possible suspects.

However, SBFR problem is more challenging than the classical face recognition from the deep learning point of view. The reasons behind this contains two aspects. (1) The heterogeneous nature of sketch and photo modalities. Most existing methods try to close the semantic gap between the two domains. (2) The lack of large databases in order to avoid overfitting and local minima. For example, most current publicly available sketch-photo datasets contain only a few number of sketch-photo pairs. More importantly, there is only one sketch per subject in most datasets making it difficult, and sometimes impossible for the network to learn robust latent features (Galea and Farrugia, 2017). As a result, many deep techniques utilize relatively shallow model or train the network only on the photo modality (Mittal et al., 2015).

Existing state-of-the-art approaches primarily focus on closing the semantic gap between the two domains by (1) transfer learning, (2) designing effective similarity measures and (3) using facial attributes in conjunction with sketch. Transfer learning is often adopted as an effective technique for SBFR. Mittal et al. (2015) presented a method to extract transfer learning based representation. Galea and Farrugia (2017) applied the transfer learning in a pre-trained face photo recognition system to tune for sketch-photo matching. Designing effective similarity measures is also an option. For example, Lin et al. (2017a) proposed a pairwise similarity measure and unified it with feature representation learning.

Besides, because some facial attributes do not exist in sketch and could be considered as the complementary information, using facial attributes in conjunction with sketch can be more advantageous. Moreover, some attributes such as wearing a hat or eyeglasses can be utilized as an auxiliary information to narrow down the suspect in the databases more accurately. For example, Iranmanesh et al. (2019) designed an Attribute-Assisted DCNN to exploit the facial attributes and leverage the loss functions from the facial attributes identification and face verification tasks to learn rich discriminative features in a common embedding subspace. Kazemi et al. (2018) designed a facial attribute-guided Deep Coupled Convolutional Neural Network (DCCNN) and adopted an attribute-centered loss to learn several distinct centers in a shared embedding space.

In real-world applications, it is more practical to obtain multiple stylistic sketches rather than a single sketch for recognizing the suspect, such as clue from multiple eyewitnesses, cooperation with multiple forensic artists, etc.. Peng et al. (2018) did a fundamental study on this challenging task. They designed three specific scenarios with corresponding datasets to mimic law enforcement investigation situations, carefully defined evaluation protocols of proposed scenarios, and demonstrated the benchmark performance, respectively, under the proposed protocols.

3.3.4. Cross-Resolution Face Recognition

Low resolution (LR) face images such as those shown in Fig.6 (e), captured by surveillance cameras, can degrade the face recognition performance significantly. Some research works have focused on the LR face recognition (LRFR) problem of matching LR probe face images to HR gallery images. Matching low-resolution against high-resolution face images has a clear importance under contemporary security considerations. In practice, face images with a high resolution such as mug-shots or passport photos in gallery need to be compared against probe images with a low resolution captured by surveillance cameras at a standoff distance. In this case, there is a dimension mismatch between them. The simplest solution is to up-scale the probe images, or down-sample the HR images, but it is possible to do better.

Herrmann et al. (2017) did a comparison among three types of high-resolution CNN frameworks, Microsoft's residual architecture (He et al., 2016a), Google's inception architecture (Schroff et al., 2015), and classical VGGFace architecture (Parkhi et al., 2015). They found that the VGGFace performs the best on low-resolution face image matching. In matching across resolutions, existing deep learning approaches can be categorized into synthesis based and projection-based. Synthesis based method is to reconstruct the HR probe image from the LR one by super-resolution (SR) techniques and use it for classification. Projection-based method is to simultaneously transform the LR probe and corresponding HR gallery images into a common feature subspace where the distance between them is minimized. For example, Lu et al. (2018) proposed a deep coupled ResNet (DCR) model for low resolution face recognition. The DCR model consists of one big trunk network and two small branch networks. The trunk network is trained to learn discriminant features shared by face images of different resolutions. Two branch networks are trained to learn resolution specific coupled-mappings (CMs) so that HR gallery im-

Table 19. Overview of deep learning methods for 3D face recognition.

Algorithm	Model	Description
Thakare and Thakare (2011)	FNN	An efficient hybrid fuzzy neural network using the depth map to extract features and to handle varying lighting effects
Lee et al. (2016)	CNN	Verify and identify a subject from the colour and depth face images; Show higher accuracy under harsh illumination environment or large head pose variation
Zulqarnain Gilani and Mian (2018)	CNN	Trained on 3.1M 3D facial scans of 100K identities; Use color and depth images to achieve more accurate recognition
Kim et al. (2017)	CNN	Only requires standard preprocessing; Does not involve complex feature extraction, matching
Jhuang et al. (2016)	DBN	Use PCL to estimate features and train a DBN model
Liu et al. (2017a)	CNN	Two deep CNNs based approach for Depth-to-RGB face recognition
Simón et al. (2016)	CNN	Tri-modal RGB-D-T based facial recognition

ages and LR probe images are projected to a space where their distances are minimized. Since point-pairs in high resolution (HR) manifold share the topology with the corresponding LR manifold, Lin and Fan (2011) used a DBN to learn the relationship between HR and LR manifolds by sending both HR and LR images to a deep architecture.

3.3.5. 3D based Face Recognition

Deep learning based 2D face recognition with conventional 2D images has shown remarkable performance on some benchmarks like LFW. Owning to the 2D projection nature of these faces, such systems often exhibit high sensitivity to illumination, scale and pose. Furthermore, facial texture is not always stable for identities as it can change with make up or other factors. The fast evolution of 3D sensors reveals a new path for face recognition that could overcome the fundamental limitations of 2D technologies since poses can be fully encoded and illumination can be modeled. 3D information represents more discriminative features by the virtue of increased dimensionality (Mohammadzade and Hatzinakos, 2013). On the other hand, 3D face recognition has the potential to address these shortcomings mentioned above. Many researchers have turned their focuses to 3D face recognition and made this research area a new trend.

Choi et al. (2013) provided different strategies to address the problem of face recognition from 3D data: (1) 1F-NF, explored earlier, is to match each individual frame to a set of reference frames. (2) 1F-3D, is to replace the set of reference frames by a 3D model resulting from the integration of individual frames. (3) 3D-3D, is to use a 3D face model inferred from multiple frames as the input probe. The first strategy can be treated as 2D-2D matching. The third one, 3D-3D face matching, has been of interest for some time. However, it is hampered in practice by the complication and cost of 3D compared to 2D equipment. Matching 3D models generally is more computational resource demanding and incurs a relatively higher cost (labor and hardware) in data acquisition.

The second one is a 2D-3D HFR problem by using 3D images for enrollment, and 2D images for probes. This is useful, for example, in access control where enrollment is centralized (and 3D images are easy to obtain), but the access gate can be deployed with simpler and cheaper 2D equipment. In this case, 2D probe images can potentially be matched more reliably against the 3D enrollment model than a 2D enrollment image if the cross-domain matching problem can be solved effectively. A second motivation for 2D-3D HFR indirectly arises in the situation where pose-invariant 2D-2D matching is desired. In this case the faces can be dramatically out of correspondence, so it may be beneficial to project one face to 3D in order to better reason about alignment, or synthesize a better aligned or lit image for a better matching.

The RGB-D cameras usually provide synchronized images of both color and depth. The color image characterizes the appearance and texture information of a face, while the depth image provides the distance of each pixel from the camera, representing the face geometry to a certain degree. With the advances of 3D sensors, e.g., Kinect, and point cloud library (PCL) (Rusu and Cousins, 2011), the information of geometric coordinates of real-world objects can be easily collected, and more three-dimensional volume data (as shown in Fig.7 (a)) can be processed to mitigate the problem associated with 2D images.

In recent years, some researches have focused on face recognition using 3D facial surface and shape. A brief overview of the methods is given in Table 19. Face recognition methods (Thakare and Thakare, 2011; Lee et al., 2016; Zulqarnain Gilani and Mian, 2018) with RGB-D images utilize two complementary types of image data, i.e., color and depth images, to achieve a more accurate recognition. Kim et al. (2017) proposed a 3D face recognition model which only requires standard preprocessing, including a nose tip detection and Iterative Closest Point (ICP) (Castellani and Bartoli, 2012), while does not involve complex feature extraction and matching. Jhuang et al. (2016) proposed a 3D face verification method via the features from depth information to train a generative model. Point cloud library was adopted to estimate features, which were then fed into a DBN to train the model. Liu et al. (2017a) proposed a two deep CNNs based approach for Depth-to-RGB face recognition. Simón et al. (2016) applied deep CNNs to the tri-modal RGB-D-T based facial recognition problem. The result shows that, in most cases, using such three modalities provides a better identification performance than an isolated or bi-modal approach.

Although 3D face recognition has advantages over its 2D counterpart, it has not yet been fully benefited from the recent developments in deep learning, due to the unavailability of large training sets as well as large test datasets. Besides, the high cost

of specialized 3D sensors limits their use in practical applications.

3.3.6. ID-Selfie Face Recognition

Identity verification plays an important role in our daily lives. Numerous activities (transactions, access to services and transportation, etc.) require to verify who we are by showing our ID documents containing face images, e.g., passports and driver licenses. DocFace (Shi and Jain, 2019) is a domain-specific network to match scanned or digital ID document photos to digital camera photos of live faces by employing a transfer learning technique. Experiments indicate that given more training data, a viable system for automatic ID document photo matching can be developed and deployed.

3.4. Image Set-based Face Recognition (ISFR)

In face recognition, it usually solves a recognition problem by using a single image. With the video cameras being widely used in our real life, it is a nature choice to solve FR problem by image sets. Compared with the single image based methods, the image set FR deals with severe changes of appearance and makes decisions by comparing the query set with gallery sets. So the image set recognition offers more promises and has therefore attracted significant research attentions in recent years. For set-based face recognition, the user supplies a set of images of the same unknown individual rather than supplying a single query image. In general, the gallery also contains a set of images for each known individual; therefore, the system must recover the individual whose gallery set is the best match for a given query set.

Methods based on image sets are expected to give a better performance than those based on single images, because an image set contains variation information that is unavailable to a single, isolated image, which helps improve classification performance under challenging conditions, e.g., large variations in pose, illumination, low resolution, etc., where the conventional face recognition systems based on single-shot images often fail to perform well. Video based recognition can be treated as a special case of image set classification where a temporal relationship between the consecutive frames is available. Generally, multiple face images of a person are available for training and testing. These images may come from multiple surveillance cameras, personal photo albums or online resources and correspond to different facial appearances under varying poses, illumination and expressions.

Within a set, the common semantic relationship is shared across individual face images since they all belong to the same person. These facial images complement the appearance variations of the person under different conditions. While image sets offer more opportunities for face recognition, they also pose new challenges to the classification task. Image sets contain more information that is useful for accurate classification. However, they introduce a challenge of effectively and efficiently measuring the similarity between image sets with high inter-class ambiguity and huge intra-class variability.

Existing set based recognition methods mainly differ in the ways in which they represent the image sets and compute the distances (or similarity) between them. Based on the set model representation types, methods can be divided into two categories: parametric (Arandjelovic et al., 2005) and non-parametric (Wang and Shi, 2009) methods. More recently, deep learning methods have been used for set-based face recognition. Hayat et al. (2015, 2014) proposed a deep learning framework to estimate the nonlinear geometric structure of the image sets. They trained an Adaptive Deep Network Template for each set to learn the class-specific models and then the query set is classified based on the minimum reconstruction error computed by using those pre-learned class-specific models. Lu et al. (2015b) also used deep networks to model nonlinear face manifolds and then they applied a learning algorithm to maximize the margin between different manifolds approximated by deep networks. Cevikalp and Serhan Yavuz (2017) proposed a fast and accurate deep method to approximate the distances from gallery images to the region spanned by the query set for large-scale applications. Sun et al. (2017) proposed building deep local match kernels upon the arc-cosine kernel (Cho and Saul, 2009) to leverage its great capability of measuring the similarity between images. By mimicking the computation in deep learning networks of infinite units, the arc-cosine kernel outperforms the widely used radius basis function (RBF) kernel. Lu et al. (2017b) proposed a method to learn discriminative features and dictionaries simultaneously from raw face image pixels so that discriminative information from facial image sets can be jointly exploited by a one-stage learning procedure. Duplex metric learning (DML) method (Cheng et al., 2018) consists of two progressive metric learning stages for feature learning and image set classification, respectively. The first stage, a discriminative stacked autoencoder (DSAE) is trained imposing a metric learning regularization term on the neurons in the hidden layers and meanwhile minimizing the reconstruction error to obtain new feature mappings in which similar samples are mapped closely to each other and dissimilar samples are mapped farther apart. Sankaran et al. (2018) presented an approach of using metadata to judge the relative quality of every feature vector in a template/set for aggregation and investigate its ability to outperform related approaches.

3.5. Hard Mining

In FR, it is important to use large training data to obtain a high performance, which requires the training procedure to be more discriminative. However, the training can be highly unbalanced because there are vastly more background objects than faces. This motivates a process of searching through the background data to find a relatively small number of potential false positives, or hard negative examples (false positives are regarded as “hard negatives”).

Hard mining, previously called bootstrapping, includes hard-positive mining and hard-negative mining. Here, positive samples are images of the object to detect, and negative samples are randomly extracted from scenes which do not contain the object to detect. Using “hard” samples can help to improve the decision boundary of the model. Intuitively, mining hard-positives enables the model to discover and expand sparsely sampled minority class boundaries, whilst mining hard-negatives aims to

improve the margins of minority class boundaries corrupted by visually very similar imposter classes, e.g., significantly overlapped outliers. Instead of general negative mining, the rational for mining hard negatives (unexpected) is that they are more informative than easy negatives (expected). Hard negative mining enables the model to improve itself quicker and more effectively with less data. Similarly, model learning can also benefit from mining hard positives (unexpected).

Hard mining is commonly used in object detection (Du Terrail and Jurie, 2017; Dai et al., 2016; Shrivastava et al., 2016; Canévet and Fleuret, 2014). There are also some methods in FR using a methodology of hard mining to improve training discrimination (Schroff et al., 2015; Parkhi et al., 2015; Zhang et al., 2016b). Take FaceNet (Schroff et al., 2015) for instance. It constructed a triplet loss to train the deep model. A triplet contains a query image, a positive image, and a negative image, where the positive image is more similar to the query image than the negative image. It explores hard-positive mining techniques which encourage spherical clusters for the embedding of a single person to improve the clustering accuracy.

3.6. Closed-Set vs. Open-Set Face Recognition

Although face verification or closed-set face identification has gained a good performance, the open-set face identification is still a challenge. In real systems, only a fraction of probe sample identities are enrolled in the gallery, which fails to make the closed-set assumption. Therefore, an open-set matching is met. It is shown that the open-set face recognition is a difficult problem, and simply thresholding the similarity scores is a weak solution (Gunther et al., 2017). Research works have been done to investigate Open-set Face Recognition. Gunther et al. (2017) formulated an open-set face identification protocol based on LFW dataset and evaluated different strategies for assessing the similarity, e.g., thresholded verification-like scores, linear discriminant analysis (LDA) scores, and extreme value machine (EVM) probabilities. Vareto et al. (2017) combined hashing functions and classification methods to estimate when probe samples are known (i.e., belong to the gallery set). They did experiments with partial least squares and neural networks, and showed how response value histograms tend to behave for known and unknown individuals whenever they test a probe. Günther et al. (2017) evaluated the challenges for unconstrained open-set face recognition, which is far from being solved. Wang et al. (2017b) built a DCNN framework with a triplet supervisory signal to identify few suspects from the crowd in real time for public video surveillance.

4. Databases

Data and algorithms are two essential components for face recognition. With the wider use of deep neural networks in face recognition, the requirement of a huge amount of training data becomes more urgent, and the deep learning methods are expected to learn a more complex data distribution from large-scale training datasets containing a huge number of identities. Experiments have demonstrated that the large amount of labeled data can help the network learn better deep models. In this section, we give an overview of face datasets (e.g., still faces, videos, heterogeneous faces), mainly related to deep learning, and show the performance of deep learning methods on several databases, e.g., LFW, IJB-A, YTF.

4.1. Still Image Face Databases

In some sense, the face recognition research is driven by face data. Early face datasets were often collected under pre-defined or controlled environments, such as the CMU PIE (Sim et al., 2002), FERET (Phillips et al., 2000). Along with the practical requirement, more attentions are paid to uncontrolled or unconstrained scenarios. i.e., face recognition in the wild. With the advent of Labeled Faces in the Wild (LFW) (Huang et al., 2007), research activity in unconstrained face recognition was accelerated rapidly.

Table 20 shows a thorough list of still face datasets. Most datasets are public with provided links for download. These still faces are mainly divided into five groups: faces used for handling poses, illumination, expression, occlusion (e.g. 300WLP, PIE); faces used for cross-age FR (e.g., CACD, FG-NET, AgeDB); faces used for makeup variations (e.g., YMU, MIFS); regular testing faces (e.g., LFW, MegaFace); usually training sets (e.g., CASIA-WebFace, MS-Celeb-1M, CelebFaces). Several large training sets are private, such as MSRA's WDRef, Facebook's SFC, MFC (Megvii Face Classification). GANFaces (i.e., GANFaces-500K, GANFaces-5M) is a synthetic dataset of face images with a wide range of expressions, poses, and illuminations. It is augmented with a real face dataset, i.e., VGG face (Parkhi et al., 2015). Extensive qualitative and quantitative experiments show that the generated images are realistic and identity preserving.

LFW can be viewed as a milestone dataset in which images are crawled from the Internet containing variations in pose, illumination, expression, resolution, etc. A large number of current face recognition methods, especially the CNN based obtained robust features and outperformed the traditional methods with handcrafted features and/or classifiers (Chen et al., 2013b; Cao et al., 2013). The results on the LFW benchmark keep climbing as more deep methods are introduced. For example, the accuracy has been improved from 96% with FR+FCN (Zhu et al., 2014b) to 99% with FaceNet (Schroff et al., 2015). COCO Loss (Liu et al., 2017d) reported an accuracy of 99.86%. Table 21 gives the accuracy report of face verification with deep learning based face recognition methods on LFW under the standard protocol.

Under real-world conditions, current face verification systems still have shortcomings even though very high accuracies are reported on LFW. Realizing that the face recognition problem is far from being solved, the IARPA Janus Benchmark-A was proposed by Klare et al. (2015). IJB-A was designed to encourage studies on novel methods for unconstrained face recognition. The release of IJB-A marked a new milestone in unconstrained face recognition research (Grother and Ngan, 2017). It contains 21,230 face images and 2,085 videos of 500 individuals as shown in Table 22 with extreme viewing conditions, variations in pose, expression, illumination, and more. Each subject in IJB-A is represented by a set containing images and/or video frames. The IJB-A evaluation protocol consists of face verification (1:1) and face identification (1:N). When IJB-A was

Table 20. Overview of still face image datasets used for face recognition. 'C' means controlled, and 'U' means unconstrained.

Dataset	#Identities	#Images	C/U	Description
Yale (Belhumeur et al., 1997)	15	160	C	expressions, lighting changes
YaleB (Georghiades et al., 2001)	38	2,414	C	illumination changes
AT&T Face (Cambridge, 2018)	40	400	C	ORL; grayscale; multiple facial variations
CFP-FP (Sengupta et al., 2016)	500	7,000	U	with both frontal and profile poses
300WLP (Zhu et al., 2016)	3,837	122,430	U	ideal for pose evaluation
AR (Martinez and Benavente, 2007)	100	2,600	C	expression, illumination, and occlusion
CMU PIE (Sim et al., 2002)	68	41,368	C	pose, illumination, expressions
Multi PIE (Gross et al., 2010)	337	754,204	C	pose, illumination, facial expression
CAS-PEAL (Gao et al., 2008)	1,040	99,594	C	pose, expression, accessory, lighting
MORPH Album 1(Ricanek and Tesafaye, 2006)	515	1,690	C	age in [15,68]; different races
MORPH Album 2 (Ricanek and Tesafaye, 2006)	20,569	78,207	C	age in [16,99]; different races
FG-NET (FG-NET, 2007)	82	1,002	-	age in [0,69]
WhoIsIt (Singh et al., 2014)	110	1,109	U	age in [1,81]; three weight groups
CACD (Chen et al., 2015a)	2,000	163,446	U	age in [16,62]
IMDB-Wiki (Rothe et al., 2015)	20,284	523,051	U	age in [0,100]; from IMDB and Wikipedia
AgeDB (Moschoglou et al., 2017)	568	16,488	U	age in [1,101]; pose,expression,illumination
Large Age-Gap (LAG) (Bianco, 2017)	1010	3828	U	spanning large age gaps, e.g., 0 to 80
CALFW (Zheng et al., 2017b)	5,749	12,174	U	large age difference; same identities with LFW
CAF (Wang et al., 2018d)	4,668	313,986	U	age in [0,80]; includes lots Asian individuals
CAFR (Zhao et al., 2018a)	25,000	1,446,500	U	age in [0,99]; labels(gender,race,landmarks)
YMU (Dantcheva et al., 2012)	151	604	U	2 before+2 after makeup per subject
VMU (Dantcheva et al., 2012)	51	204	C+U	add makeup to FRGC (Phillips, 2010)
MIW (Chen et al., 2013a)	125	154	U	77 with makeup+77 without makeup
MIFS (Chen et al., 2017b)	214	642	U	2 before+2 after makeup+2 target per subject
FERET (Phillips et al., 2000)	1,199	14,126	C	standard dataset used for FR evaluation
PubFig (Kumar et al., 2009)	200	58,797	U	public figures from web
PubFig83 (Pinto et al., 2011)	83	13,002	U	modified PubFig
MSRA-CFW (Zhang et al., 2012)	421,436	2.45M	U	celebrities on the web
Essex (Anggraini, 2014)	395	7,900	C	various racial origins; glasses, beards
Social Face (Fan et al., 2014)		48,927	U	realistic face images on social network
FaceScrub (Ng and Winkler, 2014)	530	107,818	U	balanced with respect to gender
Web Images (Lu and Tang, 2015)	3,261	40,000	U	pose, expression, illumination
LFW (Huang et al., 2007)	5,749	13,233	U	pose, illumination, expression, etc.
CPLFW (Zheng and Deng, 2018)	5,749	11,652	U	add pose difference; same identities from LFW
FGLFW (Deng et al., 2017c)	-	-	U	a derivative of LFW; challenging face pairs
MegaFace (Kemelmacher-Shlizerman et al., 2016)	690k	1M	U	used as gallery; million-scale
Trillion-Pairs (DeepGlint, 2019)	5.7k	274k	U	testing set; two parts:ELFW, DELFW
WDRef (Chen et al., 2012)	2,995	99,773	U	MSRA; usually as training set
CelebFaces (Sun et al., 2013)	5,436	87,628	U	from web; usually as training set
CelebFaces+ (Sun et al., 2014b)	10,177	202,599	U	extended CelebFaces
SFC (Taigman et al., 2014)	4,030	4.4M	U	Facebook; usually as training set
CASIA-WebFace (Yi et al., 2014)	10,575	494,414	U	usually as training set
VGG Face (Parkhi et al., 2015)	2,622	2.6M	U	usually as training set
VGGFace2 (Cao et al., 2017)	9,131	3.31M	U	pose,age,illumination,ethnicity,profession
MFC (Zhou et al., 2015)	20,000	5M	U	from web; usually as training set
MS-Celeb-1M (Guo et al., 2016)	100K	10M	U	usually as training set
MS1MV2	85K	5.8M	U	semi-automatic refined version of MS-Celeb-1M
UMDFaces (Bansal et al., 2016)	8,277	367,888	U	annotated faces
Megaface 2(Nech and Kemelmacher-Shlizerman, 2016)	672,057	4.7M	U	large dataset; usually as training set
IMDb-Face (Wang et al., 2018a)	59K	1.7M	U	a noise-controlled database
MS1M-DeepGlint (DeepGlint, 2019)	87K	3.9M	U	large-scale training database
Asian-DeepGlint (DeepGlint, 2019)	94K	2.83M	U	large-scale Asian training dataset
GANFaces-500K (Gecer et al., 2018)	10K	500K	U	synthetic data; usually as training set
GANFaces-5M (Gecer et al., 2018)	10K	5M	U	synthetic data; usually as training set

released, results from multiple submissions to the challenge showed significantly worse recognition performance compared to the results on previously mentioned datasets. The performance of the state-of-the-art face recognition systems are far less than satisfactory. This benchmark is considered more challenging than LFW. Guo and Zhang (2018) investigated whether

Table 21. Accuracy (%) report of deep learning based face verification on LFW.

Algorithm	Training Data	LFW
Human	-	97.53
Max-Margin Loss (Gecer et al., 2017)	0.83M	96.03
FF-GAN (Yin et al., 2017)	0.49M	96.42
FR+FCN (Zhu et al., 2014b)	87K	96.45
Convnet-RBM (Sun et al., 2013)	87K	97.08
Pyramid CNN (Fan et al., 2014)	-	97.3
DeepFace (Taigman et al., 2014)	4.4M	97.35
DeepID (Sun et al., 2014b)	0.2M	97.47
WebFace (Yi et al., 2014)	0.49M	97.73
Aug (Masi et al., 2016b)	0.5M	98.06
FastSearch (Wang et al., 2016)	0.49M	98.2
p-CNN (Yin and Liu, 2017)	0.49M	98.27
N-pair Loss (Sohn, 2016)	0.49M	98.33
Web-Scale (Taigman et al., 2015)	500M	98.37
MM-DFR (Ding and Tao, 2015)	0.494M	98.43
VIPLFaceNet (Liu et al., 2017c)	0.5M	98.60
Contrastive-Center (Qi and Su, 2017)	0.49M	98.68
L-Softmax (Liu et al., 2016b)	0.494M	98.71
Multibatch (Tadmor et al., 2016)	2.6M	98.80
VGGFace (Parkhi et al., 2015)	2.62M	98.95
Contrastive CNN (Han et al., 2018)	0.49M	99.12
DeepID2 (Sun et al., 2014a)	160K	99.15
AM-Softmax (Wang et al., 2018b)	0.49M	99.17
Noisy Softmax (Chen et al., 2017a)	0.49M	99.18
NormFace (Wang et al., 2017a)	0.49M	99.19
L-GM Loss (Wan et al., 2018)	0.49M	99.20
CenterFace (Wen et al., 2016b)	700K	99.28
Sparse (Sun et al., 2016)	300k*	99.30
Git Loss (Calefati et al., 2018)	3.31M	99.30
Light CNN (Wu et al., 2015)	0.49M	99.33
Yin et al. (2018)	10M	99.37
SphereFace (Liu et al., 2017b)	0.49M	99.42
SphereFace+ (Liu et al., 2018b)	0.5M	99.47
DeepID2+ (Sun et al., 2015b)	290K	99.47
Marginal Loss (Deng et al., 2017a)	10M	99.48
MFRS (Zhou et al., 2015)	5M	99.50
LF-CNNs (Wen et al., 2016a)	700K	99.50
Range Loss (Zhang et al., 2017c)	10.49M	99.52
DeepID3 (Sun et al., 2015a)	290K	99.53
Correlation Loss (Deng et al., 2017b)	0.49M	99.55
AAM Loss (Qi and Zhang, 2018)	0.49M	99.583
FaceNet (Schroff et al., 2015)	200M*	99.63
CCS face (Guo and Zhang, 2017)	10M	99.71
CosFace (Wang et al., 2018c)	5M	99.73
PRN (Kang et al., 2018)	2.8M	99.76
Deep Embedding (Liu et al., 2015)	1.3M	99.77
L_2-Softmax (Ranjan et al., 2017)	10M	99.78
ArcFace (Deng et al., 2018b)	5.8M	99.83
COCO Loss (Liu et al., 2017d)	5M	**99.86**

the face image quality is a big challenge for deep learning, especially in unconstrained face recognition, even though the deep methods have been trained on a large dataset with face images of different qualities. The evaluation of the recognition performance on IJB-A was performed using several representative deep networks. Table 23 gives the performance report of deep model based face verification and identification methods on IJB-A in terms of TAR (%) at FAR = 0.1, 0.01, 0.001 and Top-1/Top-5 accuracy of face identification.

As of 2017, the performance on IJB-A is approaching saturation, with a top true accept rate of 97.6% at a 1.0% false accept rate (Zhao et al., 2017). The successive dataset, IARPA Janus BenchmarkB (IJB-B) (Whitelam et al., 2017), released in 2017, continued to push the state of the art in unconstrained face recognition. It includes 11,754 images and 7,011 videos of 1,845 subjects. The protocols support face detection, verification, recognition, and clustering, and allow for evaluation at more operationally relevant points at low ends of the ROC curve (e.g., FAR at 0.01% and 0.001%). The IARPA Janus BenchmarkC (IJB-C) (Maze et al., 2018) was proposed to address the problem of allowing the evaluation of an end-to-end system.

4.2. Video Face Databases

Video based face recognition has also gained much attention, and several video face datasets have been released. Table 22 lists serveral datasets with both still and video faces (e.g., COX Face, PaSC, IJB-A, IJB-B, IJB-C). Table 24 shows a list of video face datasets. Most of them are public available. YTF and PaSC are often used to test the recognition performance of various deep models.

YouTube Face (YTF) (Wolf et al., 2011) dataset contains 3,425 videos of 1,595 different people. The clip durations vary from 48 frames to 6, 070 frames, with an average length of 181.3 frames. It contains 10 folds of 500 video pairs. Point and Shoot Challenge (PaSC) (Beveridge et al., 2013) includes 9,376 images and 2,802 videos of 293 subjects. It is collected with some controls for different locations, poses, distances from the camera. Performance on YTF and PaSC is reported in Table 25. As shown in the table, existing algorithms have attained a high performance on YTF (97.7%). However, Face recognition in videos presents unique challenges due to the variations which can degrade the frame quality. Furthermore, as videos usually contain many frames, it brings considerable computational burdens too.

iQIYI-VID (Liu et al., 2018d) is a newly released video dataset. It is the largest video dataset for multi-modal person identification so far, including face, cloth, voice, gait and subtitles, for character identification. It is composed of 565,372 video clips (training set 219,677, validation set 172,860, and test set 172,835) of 4,934 celebrities. These video clips are extracted from 400K hours of online videos of various types, ranging from movies, variety shows, TV series, to news broadcasting. All video clips pass through a careful human annotation process. The length of the videos ranges from 1 to 30 seconds.

4.3. Heterogeneous Face Databases

For heterogeneous face recognition, multi-modal data are needed, e.g., visible, thermal, sketch, RGB-D. Table 26 shows a list of heterogeneous face datasets. These sets are divided into six groups: (1) Still-to-Video faces, such as COX-S2V (Huang et al., 2012c), (2) NIR-VIS faces, (3) Sketch-Photo

Table 22. Overview of still+video datasets used for face recognition. 'C' means controlled, and 'U' means unconstrained.

Dataset	#Identities	#Images	#video	C/U	Description
PaSC (Beveridge et al., 2013)	293	9,376	2,802	C	still+video; collected at different locations, poses, distances
COX Face (Huang et al., 2015)	1,000	1,000	1,000	C	still +surveillance-like videos; still images with seated subjects; surveillance-like videos captured with walking subjects
IJB-A (Klare et al., 2015)	500	21,230	2,085	U	still+video; near complete variations
IJB-B (Whitelam et al., 2017)	1,845	11,754	7,011	U	still+video
IJB-C (Maze et al., 2018)	3,531	31.3K	11,779	U	still+video;a further extension of IJB-B

Table 23. Performance report of deep learning based face recognition on IJB-A. Symbol "-" indicates that the metric is not available for that protocol.

Algorithm	Face Verification(TAR)			Face Identification(Rec. Rate)	
	@FAR=0.1	@FAR=0.01	@FAR=0.001	@Rank-1	@Rank-5
B-CNN (Chowdhury et al., 2016)	-	-	-	0.588 ± 0.020	0.796 ± 0.017
Pooling faces (Hassner et al., 2016)	0.631	0.309	-	0.846	0.933
GOTS (Klare et al., 2015)	0.627 ± .012	0.406 ± 0.014	0.198 ± 0.008	0.433 ± 0.021	0.595 ± 0.020
ReST (Wu et al., 2017a)	-	0.630	0.548	-	-
FastSearch (Wang et al., 2016)	0.893	0.729	0.510		
LSFS (Wang et al., 2016)	0.895 ± 0.013	0.733 ± 0.034	0.514 ± 0.060	0.820 ± 0.024	0.929 ± 0.013
DR-GAN (Tran et al., 2017)		0.774 ± 0.027	0.539 ± 0.043	0.855 ± 0.015	0.947 ± 0.011
p-CNN (Yin and Liu, 2017)	-	0.775 ± 0.025	0.539 ± 0.042	0.855 ± 0.014	0.938 ± 0.009
$DCNN_{manual}$ (Chen et al., 2015b)	0.947 ± 0.011	0.787 ± 0.043	-	0.852 ± 0.018	0.937 ± 0.010
Deep Multi-pose (Almageed et al., 2016)	0.911	0.787	-	0.846	0.927
Triplet Similarity (Sankaranarayanan et al., 2016b)	0.945 ± 0.002	0.790 ± 0.030	0.590 ± 0.050	0.880 ± 0.015	0.950 ± 0.007
VGGFace (Parkhi et al., 2015)	0.937 ± 0.01	0.805 ± 0.030	0.604 ± 0.06	0.913 ± 0.011	-
Joint Bayesian (Chen et al., 2016a)	0.961	0.818	-	-	-
$PAM_{frontal}$ (Masi et al., 2016a)	-	0.826 ± 0.018	0.652 ± 0.037	0.840 ± 0.012	0.925 ± 0.008
PAMs (Masi et al., 2016a)	0.652 ± 0.037	0.826 ± 0.018	-	0.840 ± 0.012	0.925 ± 0.008
PAM (Masi et al., 2019a)	-	0.847 ± 0.016	0.711 ± 0.037	0.862 ± 0.013	0.943 ± 0.009
$DCNN_{fusion}$ (Chen et al., 2016a)	0.967 ± 0.009	0.838 ± 0.042	-	0.903 ± 0.012	0.965 ± 0.008
Contrastive CNN (Han et al., 2018)	0.9531	0.8401	0.6391		
FF-GAN (Yin et al., 2017)	-	0.852 ± 0.010	0.663 ± 0.033	0.902 ± 0.006	0.954 ± 0.005
Aug (Masi et al., 2016b)	-	0.88	0.725	0.906	0.962
Triplet Embedding (Sankaranarayanan et al., 2016a)	0.964 ± 0.005	0.900 ± 0.010	0.813 ± 0.020	0.932 ± 0.010	-
MTL (Ranjan et al., 2016)	0.976 ± 0.004	0.922 ± 0.010	0.823 ± 0.020	0.947 ± 0.008	-
Yin et al. (2018)	-	0.931	0.873	0.939	0.966
Template Adaptation (Crosswhite et al., 2016)	0.979 ± 0.004	0.939 ± 0.013	0.836 ± 0.027	0.928 ± 0.010	0.977 ± 0.004
NAN (Yang et al., 2017a)	0.978 ± 0.003	0.941 ± 0.008	0.881 ± 0.011	0.958 ± 0.005	0.980 ± 0.005
QAN (Liu et al., 2017e)	0.980 ± 0.006	0.942 ± 0.015	0.893 ± 0.039	-	-
DREAM (Cao et al., 2018c)	-	0.944 ± 0.009	0.868 ± 0.015	0.946 ± 0.011	0.968 ± 0.010
DAC (Liu et al., 2018c)	0.981 ± 0.008	0.954 ± 0.01	-	0.973 ± 0.011	-
TDFF (Xiong et al., 2017)	0.988 ± 0.003	0.961 ± 0.007	0.919 ± 0.006	0.964 ± 0.006	-
TDFF (Xiong et al., 2017)+ TPE (Sankaranarayanan et al., 2016a)	0.989 ± 0.003	0.961 ± 0.007	0.921 ± 0.005	0.964 ± 0.007	-
PRN (Kang et al., 2018)	0.988 ± 0.002	0.965 ± 0.004	0.919 ± 0.013	0.982 ± 0.004	0.992 ± 0.002
M-FAN (Sankaran et al., 2018)	0.980 ± 0.003	0.966 ± 0.004	0.944 ± 0.005	-	-
VGGFace2 (Cao et al., 2017)	0.990 ± 0.002	0.968 ± 0.006	0.921 ± 0.014	0.982 ± 0.004	**0.993 ± 0.002**
L_2-Softmax (Ranjan et al., 2017)	0.984 ± 0.002	0.970 ± 0.004	0.943 ± 0.005	0.973 ± 0.005	-
DA-GAN (Zhao et al., 2017)	**0.991 ± 0.003**	0.976 ± 0.007	0.930 ± 0.005	0.971 ± 0.007	0.989 ± 0.003
3D-PIM (Zhao et al., 2018c)	-	**0.989 ±0.002**	**0.977 ± 0.004**	**0.990 ± 0.002**	-

faces, (4) 3D/RGB-D faces, (5) Cross-Resilution faces, like NJU-ID (Huo et al., 2016), and (6) ID-Selfie faces.

NIR-VIS faces. There are 17 main heterogeneous datasets covering the NIR-VIS condition. CASIA NIR-VIS 2.0 (Li et al., 2013b) is a widely used NIR dataset. CASIA HFB (Li et al., 2009), composed of visual (VIS), near infrared (NIR) and 3D faces, is widely used too. The Cross Spectral Dataset (Goswami et al., 2011) consists of 430 subjects from various

Table 24. Overview of video face datasets used for face recognition.

Dataset	#Identities	#Videos	Description
Honda (Lee et al., 2003)	20	59	large pose/expression variations; 400 frame/video
FIA (Goh et al., 2005)	180	6,470	captured by 6 synchronized cameras from 3 different angles
Faces96 (Essex, 2015)	152	152	significant head variations
YTC (Kim et al., 2008)	47	1,910	high compression rate; large variations; from YouTube
ChokePoint (Wong et al., 2011)	29	48	video surveillance dataset; 64,204 still images
YTF (Wolf et al., 2011)	1,595	3,425	low resolution, motion blur; from YouTube
Celebrities-1000 (Liu et al., 2014)	1,000	7,021	covering illuminations, poses, etc.
SN-Flip (Barr et al., 2014)	190	28	multiple subjects in frame; less motion
McGillFaces (Demirkus et al., 2014)	60	60	Real-world Face Video
ACVF-2014 (Dhamecha et al., 2015)	133	201	multiple subjects in frame; use handheld cameras
ESOGU-285(Yalcin et al., 2015)	285	2,280	764K frames; set-based FR
CSCRV (Singh et al., 2016)	160	193	video; with open-set protocol
UMDFaces-Videos (Bansal et al., 2017)	3,107	22,075	video; from YouTube
iQIYI-VID (Liu et al., 2018d)	5,000	600K	from 400K hours of online videos

Table 25. Accuracy (%) report of deep learning based face verification on YTF and PaSC.

Algorithm	Training Data	YTF	PaSC	
			handled	control
DeepFace (Taigman et al., 2014)	4.4M	91.40		
DeepID (Sun et al., 2014b)	0.2M	93.20		
WebFace (Yi et al., 2014)	0.49M	92.20		
VGGFace (Parkhi et al., 2015)	2.62M	97.30	87.03	91.25
CenterFace (Wen et al., 2016b)	700K	94.9		
Sparse (Sun et al., 2016)	300k*	92.70		
Light CNN (Wu et al., 2015)	0.49M	91.6		
SphereFace (Liu et al., 2017b)	0.49M	95.0		
DeepID2+ (Sun et al., 2015b)	290K	93.2		
FaceNet (Schroff et al., 2015)	200M*	95.12		
CosFace (Wang et al., 2018c)	5M	97.6		
PRN (Kang et al., 2018)	2.8M	96.3		
Sohn et al. (2017)	0.49M	95.38		
MDLFace (Goswami et al., 2014)	-	88.6	87.4	93.4
NAN (Yang et al., 2017a)	3M	95.7		
DAC (Liu et al., 2018c)	3M	96.01		
DAN (Rao et al., 2017a)	-	97.32	80.33	92.06
QAN (Liu et al., 2017e)	5M	96.17		
Goswami et al. (2017)	-	95.3	93.1	95.9
ADRL (Rao et al., 2017b)	-	95.96	93.78	95.67
TBE-CNN (Ding and Tao, 2018)	0.49M	94.96	**95.9**	**96.2**
ASML (Hu et al., 2017c)	10M	97.6		
DML (Cheng et al., 2018)	-	**97.7**		

ethnic backgrounds (more than 20% of non-European origin).

Sketch-Photo faces. There are 9 commonly used datasets for benchmarking SBFR systems. Each contains pairs of sketches and photos. They differ by size, whether sketches are viewed and if drawn by artist or composited by software. CUHK Face sketch dataset (CUFS) (Wang and Tang, 2009) includes 188 subjects from the Chinese University of Hong Kong (CUHK) student dataset, 123 faces from the AR dataset, and 295 faces from the XM2VTS dataset (Messer et al., 1999). The sketch is drawn by an artist based on the photo. CUHK Face Sketch FERET Dataset (CUFSF) (Zhang et al., 2011) has 1,194 subjects from FERET dataset (Phillips et al., 2000). Compared to CUFS, the photos in CUFSF are taken with illumination variations. Meanwhile, the sketches were drawn with shape exaggeration based on the photos. Hence, CUFSF is more challenging and closer to practical scenarios. Unlike CUFS and CUFSF, IIIT-Delhi Sketch Dataset (Bhatt et al., 2012) contains three types of sketches, namely IIIT-D viewed, IIIT-D semi-forensic and IIIT-D forensic sketch dataset. IIIT-D viewed sketches are drawn by a professional sketch artist based on photos collected from various sources. IIIT-D semi-forensic sketches are drawn based on an artist's memory. PRIP-VSGC (Han et al., 2013) is Pattern Recognition and Image Processing Viewed Software-Generated Composite database in which subjects are from AR database. PRIP-HDC (Klum et al., 2014) is Pattern Recognition and Image Processing Hand-Drawn Com-

Table 26. Overview of datasets for heterogeneous face recognition.

Dataset	#Identities	Description
COX-S2V (Huang et al., 2012c)	1,000	still+video; 3 video clips/subject; various illumination,poses,motion blurs
Notre Dame LWIR (Kevin and Bowyer, 2003)	159	LWIR+VIS; various lighting,expression and time lapse
CASIA HFB (Li et al., 2009)	202	2,095 VIS+3,002 NIR face images
USTC-NVIE (Wang et al., 2010)	215	VIS+infrared facial expression; with three types of illumination
Cross-Spectral (Goswami et al., 2011)	430	2,103 NIR+2,086 VIS; different pose angles in pitch, yaw directions
LDHF-DB (Maeng et al., 2012)	100	VIS+NIR; 1,600 images; Long distance to cameras
CASIA NIR-VIS 2.0 (Li et al., 2013b)	725	VIS+NIR; 17,580 images; multiple facial variations; more close to practical applications captured in constrained situation
WSRI (Riggan et al., 2015)	64	1,615 VIS+1,615 MWIR; 25 per subject, vary facial expression
UND Collection X1 (Sarfraz and Stiefelhagen, 2017)	241	2,451 VIS+2,451 LWIR
Night Vision (NVESD)	50	VIS,SWIR,MWIR,LWIR; collected by U.S. Army CERDEC-NVESD
BUAA-VisNir (Huang et al., 2012a)	150	NIR+VIS; vary in poses and expressions
Oulu-CASIA (Chen et al., 2009)	80	NIR+VIS; Videos; 6 expressions; 3 lighting conditions
PolyU NIR (Zhang et al., 2010)	335	NIR; 33,500 images; faces with expression, pose variations
SCface (Grgic et al., 2011)	130	IR+VIS; 4,160 static images
CUHK VIS-NIR (Gong et al., 2017)	2,800	NIR+VIS; each subject has one pair of near infrared image-visible image
NIR-PF (He et al., 2016b)	276	NIR; 5300 images; various scales,focus,occlusion,distance,view
Polarimethric Thermal (Hu et al., 2016)	60	Polrimetric LWIR+VIS; collected at 3 different distances: 2.5m,5m,7.5m
IRIS (UTK, 2012)	29	Thermal+VIS; 4,228 pairs of thermal/visible images; various poses
CUFS (Wang and Tang, 2009)	606	VIS+sketch; 1,216 images; frontal pose,normal lighting,neutral expression
CUFSF (Zhang et al., 2011)	1,194	VIS+sketch; 2,388 image pairs
IIIT-Delhi Sketch (Bhatt et al., 2012)	-	VIS+sketch; 238 viewed pairs,140 semi-forensic pairs,190 forensic pairs
PRIP-VSGC (Han et al., 2013)	123	VIS+sketch; composite sketch and digital image pairs
PRIP-HDC (Klum et al., 2014)	265	hand-drawn and composite sketches with corresponding mugshots
MGDB (Ouyang et al., 2016b)	100	VIS+4 facial sketches drawn at various time-delays: viewed sketch, 1 hour sketch, 24 hour sketch and unviewed sketches
e-PRIP (Mittal et al., 2017)	123	sketch; extended-PRIP dataset
UoM-SGFS (Galea and Farrugia, 2016)	300	600 software-generated sketches; containing sketches represented in color
Extended UoM-SGFS(Galea and Farrugia, 2018)	600	1200 sketches
FRGCv2 (Phillips et al., 2005)	446	3D; 4,007 images; with additional expression tags
BU-3DFE (Yin et al., 2006)	100	3D; 2500 scans; for expression-invariant FR
Bosphorus (Savran et al., 2008)	105	3D; 4666 scans; variations on expressions, poses, occlusion
ND-2006 (Faltemier et al., 2007)	888	3D; 13,450 scans; a superset of FRGCv2; 6 different expression tags
Texas 3DFRD (Gupta et al., 2010)	105	3D; 1149 pairs of face texture descriptions
BJUT-3D (Yin et al., 2006)	500	3D; Chinese Face 3D Face Dataset with high-resolution
CASIA (Xu et al., 2006)	123	3D; 4,624 scans; changes in pose, expression, lighting
3D-TEC (Vijayan et al., 2011)	214	3D; 428 scans; 107 pairs of twins; with a smile and a neutral expression
NPU3D (Yanning et al., 2012)	300	3D; 10,500 scans with VIS images; Chinese VIS+3D
UHDB11 (Toderici et al., 2013)	23	2D+3D; $> 1,600$ images; 2D(illumination,pose,etc.)+3D facial
UHDB31 (Wu et al., 2016)	77	2D+3D; 1,617 images; 2D with various poses+3D facial models
LS3DFace (Zulqarnain Gilani and Mian, 2018)	1,853	3D; 31,860 images; extreme variations: pose,occlusion,missing data,etc.
CurtinFaces (Li et al., 2013a)	52	RGB-D; 5,000 images; various poses,illumination,expression,occlusion
IIIT-D (Goswami et al., 2013)	106	RGB-D; 4,603 images;a few pose,expression variations
BUAA Lock3DFace (Zhang et al., 2016a)	509	5,711 RGB-D video sequences; various pose,expression,occlusion,time
RGB-D-T (Nikisins et al., 2014)	51	RGB-D; different rotations, illuminations, expressions
KinectFaceDB (Min et al., 2014)	52	RGB-D; 936 images; multimodal(2D/3D/video);multiple facial variations
RGB-D DB (Cui et al., 2018a)	747	845K RGB-D; continuous pose variations,a few illumination changes
NJU-ID (Huo et al., 2016)	256	13,056 images with different resolutions; 1 LR + 50 HR image per subject
ID-Selfie-A (Shi and Jain, 2019)	-	10,000 pairs of ID Cards photo and selfies; private
ID-Selfie-B (Shi and Jain, 2019)	547	10,844 images; ID document-selfie dataset; private

posite database in which the facial sketches are drawn based on the verbal description by the eyewitness or victim. Memory Gap Database (MGDB) (Ouyang et al., 2016b) not only contains viewed and unviewed sketch, but unique sketches rendered at different time-delays between viewing and sketching.

3D faces. In contrast to 2D face images, 3D faces contain more geometry information and are insensitive to pose and illumination variations. Recently, many research institutes have established different kinds of 3D face databases to test and evaluate their methods for 3D face recognition. The face Recog-

nition Grand Challenge (FRGC) V2.0 (Phillips et al., 2005) database has tremendous influence on the development of 3D based face recognition. It is widely accepted as a standard reference database to evaluate the performance of 3D face recognition algorithms. BU-3DFE (Yin et al., 2006) is released specifically for 3D based expression-invariant face recognition, which contains 6 types of expressions: anger, happiness, sadness, surprise, disgust, and fear. Each type of expression is further tagged with four different levels. The most recent LS3DFace (Zulqarnain Gilani and Mian, 2018) dataset includes 3D models and 31,860 facial images from 1,853 subjects with extreme pose, occlusion, missing data, etc.

5. Discussion and Challenges

Even though a significant progress has been made in face recognition with deep learning, there are still challenges, e.g., network design, architecture optimization, alignment necessity, face database related issues, and cross-quality matching.

5.1. Network Design and Architecture Optimization

It seems that the deep learning requires no hand-designed feature extractors, everything is learned from data, and there is almost no human intervention. But that is not entirely true! It is clear that the features are learned from data, and a hierarchy of learned features can lead to a great representational power. But there is still a lot of human interventions in the model design and optimization, which requires a lot of expert knowledge and takes ample time. It is a challenge. One needs to decide which neural network to use, and how to set hyperparameters when building and training the selected network. Furthermore, model selection in deep networks is not just about choosing hyperparameters. Designing the architecture of a model also involves choosing the types of layers and the way they are arranged and connected to each other, as there are many possible ways one can consider to design a network.

Hyperparameters are settings used to control the behavior of a model. It contains the variables which determine the network structure (e.g., the number of hidden units, hidden layers, dropout, activation function, padding, loss function, kernel size, stride) and how the network is trained (e.g., learning rate, momentum, the number of epochs, batch size). Actually four methods (Manual Search, Grid Search, Random Search and Bayesian Optimization) have been developed to find out hyperparameters. However, designing a good model usually involves a lot of trial and error. There is no generic way to determine a priori given just a problem description. There is not even much guidance to determine good values as a starting point. The easiest way is to pick a model that has been proven to work for a similar problem. It is not necessary to train it from scratch. One can take a pre-trained model and fine-tune the weights to adapt to a new problem. Even if for a novel problem or the existing models do not meet the needs, one can always borrow ideas from successful models to design a new one.

Model optimization is a process of modifying the code and elements including the hyperparameters to minimize the testing error, improve generalization by reducing the redundancy in the parameter space, reduce the computational cost, or figure out how many features needed in each layer for the best performance. It is one of the tough challenges in the implementation of deep learning solutions. The weights in the neural networks are highly interdependent. Some weights change in one layer, and affect the weights in next layers. The size and depth of neural networks interact with other hyperparamaters, so that changing one thing elsewhere can affect where the best values are. So it is different to isolate a "best" size and depth for a network then continue to tune other parameters in isolation. There are several ways to do this: (1) Regularization modifies the objective function/learning problem so the optimization is likely to find a neural network with a small number of parameters. (2) Pruning takes a large network and deletes features or parameters that are redundant in some sense. (3) Or a less widely used approach, growing, can be applied by starting at a small network and incrementally adding new units by some growth criterion. Although several research works investigate the optimization problem, for example, Reale et al. (2017) presented a method to remove unnecessary hidden nodes from a deep neural network by using the group lasso penalty (Meier et al., 2008) to select the appropriate number of hidden nodes for each convolutional and fully connected layer, designing more effective and powerful networks and adopting useful optimization strategies are still a challenge.

5.2. Face Data Related Issues

Although a number of face datasets have been assembled for various face recognition applications, helping achieve an excellent performance in certain aspects, there still exists quantities of problems, e.g., generating sufficient and more useful face data.

5.2.1. Training Data Volume and Data Augmentation

It is not trivial to get a huge amount of labeled face data for learning. Many large-scale datasets were generated, especially still face images, which are automatically collected from the Internet. However, the requirement is urgent for specific FR problems, e.g., 3D based FR, HFR. It requires great efforts to collect a large dataset. For instance, the publicly available 3D face dataset, ND-2006 (Faltemier et al., 2007) has only 13,540 scans of 888 unique identities and took over two years to collect. Although it might be possible to design a smart modeling method to minimize the need of a huge amount of data (Peng et al., 2016), it is still time-consuming and needs lots of tricks. One common method nowadays is data augmentation. The usage of synthetic data as additional training data is shown to be helpful in some cases even if they are rendered images.

Data augmentation techniques are label-preserving transformations typically applied to training images. It is either done by geometric transformations or by manipulating the facial appearance of existing data with variations. (1) Geometric transformations contains oversampling (multiple image translations by cropping at different offsets) (Krizhevsky et al., 2012), mirroring (horizontal flipping) (Yang and Patras, 2015), rotation (Xie and Tu, 2015), and various photometric transformations (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; Eigen

and Fergus, 2015). (2) Introducing face variations means synthesizing new face images by changing the pose, shape, illuminations, occlusions, and expressions (Lv et al., 2017; Masi et al., 2016b; Gecer et al., 2018; Masi et al., 2019b; Zulqarnain Gilani and Mian, 2018). Instead of directly manipulating the input images, Leng et al. (2017) performed a virtual sample generation at the feature level.

Recently, GAN has been proposed as an effective method for data augmentation by synthesizing samples from the underlying appearance distribution through a generative model. This ability proved to be particularly effective for structured object classes such as faces. Although this approach has showed promising results, regular GANs, can generate novel images but not new subjects. It offers no explicit control over the identity of the generated samples.

Although data augmentation technique gives a chance for specific FR problems, and promote the development to some extent, acquiring a large amount of data from real world is still urgent for learning deep models.

5.2.2. Long-tail Distritubted Datasets

Deep learning has achieved impressive results in face recognition, however, most large-scale face datasets, e.g., CASIA-Webface where about 39% of the 10K subjects have less than 20 images, even with a significant number of identities, exhibit long-tailed distribution characteristics, which may result in biased classifiers in conventionally trained deep neural networks or insufficient data when long-tail classes are ignored. The real world data often have a long-tailed distribution. This means a few classes are predominant while others are rare. The classes with abundant training instances are referred as classes in the head, and unrepresented classes as classes in the tail.

With highly imbalanced numbers of images across categories as the training data, models can lead to unsatisfied performance on under-represented tail categories. This is because classifiers tend to generalize well for classes in the head, but lack the generalization for classes in the tail. Without any re-sampling of the training images or re-weighting of the loss, categories with more images in the head will pose a greater impact on the feature learning procedure (Zhou et al., 2015) and inversely cripple the model ability on the tailed part. The trained model leans to overfit the rich classes with large samples, and spare samples from poor classes tend to exhibit large intra-class dispension. Thus the learned models may not perform well in recognition.

A practical recognition system must classify majority and minority classes, generalize from a few known instances, and acknowledge novelty upon a never seen instance. So, how to deal with a long-tailed distribution becomes an important problem in real world data. However, it is a rather unexplored area in FR. While it has been established that recognition engines are data-hungry and keep improving with more volumes of data, mechanisms to derive benefits from the vast diversity of real data are relatively unexplored. There are only few previous works that discuss about learning from long-tail classes. Techniques to handle imbalanced datasets are typically split into two regimes: data solutions and algorithmic solutions.

Data Solutions. Most previous works handled this problem simply by removing the samples from poor classes to achieve the class balance, or designing data sampling rules or regularization on tail classes. According to Ouyang et al. (2016c), the performance can improve slightly if just 40% of positive samples are preserved to make the training samples more uniform. Besides, by simply abandoning the data partially, information contained in the data may also be omitted. Poor classes can include complementary knowledge to rich classes which can boost the performance of the final models. Another simple technique is class weights normalization by changing the sampling frequency (He and Garcia, 2008). For instance, random oversampling effectively repeats training instances from the classes in the tail, while random undersampling removes instances from the classes with abundant training instances. Unfortunately, significant imbalance still exists after weight norm regularization via data re-sampling (Yin et al., 2018). The low intra-class variance of the tail classes is not fully resolved, causing the decision boundary to be biased, which impacts the recognition performance.

Algorithmic Solutions. For example, Zhang et al. (2017c) proposed a feature regularization method by applying range loss to effectively utilize the entire long-tailed data in the training process. Guo and Zhang (2017) proposed tail class promotion loss to regularize the norm of weight vectors for tail classes. Yin et al. (2018) used a center-based feature transfer learning to augment tail classes by generating feature-level samples through transfer of intra-class variance from regular classes. Besides, there are a few methods from other fields, which can be borrowed to deal with the problem in FR, such as metric learning (Huang et al., 2016a; Oh Song et al., 2016; Fragoso and Ramanan, 2018), hard negative mining (Dong et al., 2017b; Lin et al., 2017b), feature transfer learning (Cui et al., 2018b), meta learning (Wang et al., 2017f), metric+meta learning (Liu et al., 2019).

5.2.3. Cross-Quality Face Matching

Cross-quality matching is still a big issue for deep learning based face recognition. In face recognition, pose variations, illumination changes, cross-age, facial expression variations, facial occlusions, low resolution, makeup, etc. are the main factors that can influence the face recognition performance. Guo and Zhang (2018) raised a question: does the face recognition problem have been solved or almost solved, given the great success of deep learning? They consider that the face image quality issue may still be one challenge for DL, and studied the matching across different face image qualities to better understand the performance of deep learning methods. The results showed that the face image quality variations are still a great challenge for deep learning methods, even though various training face images of different qualities have been fed into the networks during the trainings. Thus, it is needed to develop more robust deep neural networks to address the issue of significant face image quality changes.

5.2.4. More Subjects or More Images?

For training deep neural networks, there is no clear guideline to follow on choosing what kinds of dataset, a dataset with more subjects but less images per subject or one with less subjects but more images per subject? Bansal et al. (2017) made

an investigation on this. For two datasets with the same number of images, they call one wider than the other if on average it has less images per subject than the other. In fact, given enough images, both deep and wide datasets can contain a variety of face images. Deep datasets have more changes in pose, expression, illuminations, etc. Wide datasets contain large variations as well because of the large number of unique identities. They try to resolve the dilemma of choosing one kind of dataset over the other by a set of experiments on a wider and deeper datasets. The result shows that the choice of the dataset depends on the type of deep networks being trained. Deeper networks perform well with deeper datasets and shallower networks work well with wider datasets. This observation is important since it can guide researchers towards better practices to follow while collecting data or selecting data for training deep networks. Data acquisition is an expensive and time consuming process, and these experiments shed a light on how to obtain the maximum benefit from the investment in data.

5.3. Necessity and Kind of Face Alignment

In most face recognition pipelines, the detection, alignment and face cropping are the first step, and then train deep networks on the cropped faces. Face alignment is an important computer vision technology for identifying the geometric structure of human faces in digital images, an essential preprocess for face recognition, and face synthesis. It is usually done by detecting locations of facial keypoints in the face image and then using some kind of strategies to transform the face to a canonical view. If the alignment is not applied to the images, the relative position of the faces inside the bounding box can vary, with more pronounced variations for larger bounding boxes.

Although AAM-based approaches (Tzimiropoulos and Pantic, 2013; Saragih and Goecke, 2007) and regression-based approaches (Zhu et al., 2015a; Cao et al., 2014; Xiong and De la Torre, 2013; Ren et al., 2014) work well for face images with small poses, they usually cannot handle profile face images because of the visibility of landmarks. Existing deep learning based face alignment methods are either cascaded regression methods or end-to-end deep regression methods. (1) Cascaded regression methods (Kowalski et al., 2017; Trigeorgis et al., 2016; Dapogny et al., 2019) consist in learning a sequence of updates, starting from an initial guess, and refining the landmark localization in a coarse-to-fine manner. This allows to robustly learn rigid transformations, e.g., translation and rotation, in the first cascade stages, while learning non-rigid deformation, e.g., due to facial expression or non-planar rotation. (2) Deep end-to-end regression methods (Zhang et al., 2015b; Kumar and Chellappa, 2018; Dong et al., 2018b,a; Miao et al., 2018; Yue et al., 2018) aim at a regression on the landmark position from the original image directly. Because of the scarcity of the data, end-to-end approaches usually rely on learning an intermediate representation, such as edge detection to drive the alignment process.

The main challenge in face alignment arises from pervasive ambiguities in low-level image features. While the main face structures are present in the feature maps, the contours of face components are frequently disrupted by gaps or corrupted by spurious fragments. Strong gradient responses could be due to reflectance, occlusion, fine facial texture, pose, or background clutter. In contrast, the boundaries of face components such as nose and eyebrow are often obscure and incomplete. Searching face components separately is difficult and often yields noisy results.

The necessity of the alignment step is well founded for engineered computer vision methods based on hand-crafted features, so Bansal et al. (2017) investigated whether the performance of deep face recognition networks is affected by the face alignment process. They note that there is a clear dependence of performance on the type of face alignment used for training and testing. Using a good landmark detection method and alignment procedure for both training and testing is essential for achieving a good performance. As landmark detection and alignment methods continue to improve, we expect the face recognition performance could be less affected by the alignment.

6. Conclusion

We have presented a comprehensive survey of face recognition methods based on deep learning, mainly focusing on deep architectures and some specific recognition problems. Deep learning techniques have been fully applied to face recognition, and have played important roles in addressing or circumventing challenges in FR, including pose variations, illumination changes, facial expression, etc. Deep methods have also shown good performance in processing RGB-D, video, and heterogeneous face matching. A review of related face databases has been given as well, such as still images, videos, and heterogeneous face data for cross-modal FR. Although the face recognition accuracies have been improved for many still image based face matching, there are still some challenges in practice.

Acknowledgments

Thanks to Q. Wang and M. Nouyed for comments on the manuscript.

References

Almageed, W.A., Wu, Y., Rawls, S., Harel, S., Hassner, T., Masi, I., Choi, J., Lekust, J., Kim, J., Natarajan, P., et al., 2016. Face recognition using deep multi-pose representations, in: Applications of Computer Vision, Winter Conf. on, IEEE. pp. 1–9.

Anggraini, D.R., 2014. Face recognition using principal component analysis and self organizing maps, in: Student Project Conf. (ICT-ISPC), ICT Intl., IEEE. pp. 91–94.

Antipov, G., Baccouche, M., Dugelay, J.L., 2017a. Boosting cross-age face verification via generative age normalization, in: International Joint Conference on Biometrics.

Antipov, G., Baccouche, M., Dugelay, J.L., 2017b. Face aging with conditional generative adversarial networks. arXiv preprint arXiv:1702.01983 .

Arandjelovic, O., Shakhnarovich, G., Fisher, J., Cipolla, R., Darrell, T., 2005. Face recognition with image sets using manifold density divergence, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), IEEE. pp. 581–588.

Balasubramanian, M., Palanivel, S., Ramalingam, V., 2009. Real time face and mouth recognition using radial basis function neural networks. Expert Systems with Applications 36, 6879–6888.

Bansal, A., Castillo, C., Ranjan, R., Chellappa, R., 2017. The do's and don'ts for cnn-based face verification. arXiv preprint arXiv:1705.07426 .

Bansal, A., Nanduri, A., Castillo, C., Ranjan, R., Chellappa, R., 2016. Umdfaces: An annotated face dataset for training deep networks. arXiv preprint arXiv:1611.01484 .

Bao, T., Ding, C., Karmoshi, S., Zhu, M., 2017. Video-based face recognition via convolutional neural networks, in: Second Intl. Workshop on Pattern Recognition, Intl. Society for Optics and Photonics. p. 104430I.

Barr, J.R., Bowyer, K.W., Flynn, P.J., Biswas, S., 2012. Face recognition from video: A review. Intl. Journal of Pattern Recognition and Artificial Intelligence 26, 1266002.

Barr, J.R., Cament, L.A., Bowyer, K.W., Flynn, P.J., 2014. Active clustering with ensembles for social structure extraction, in: Applications of Computer Vision, Winter Conf. on, IEEE. pp. 969–976.

Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J., 1997. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE trans. on pattern analysis and machine intelligence 19, 711–720.

Bengio, Y., et al., 2009. Learning deep architectures for ai. Foundations and trends in Machine Learning 2, 1–127.

Beveridge, J.R., Phillips, P.J., Bolme, D.S., Draper, B.A., Givens, G.H., Lui, Y.M., Teli, M.N., Zhang, H., Scruggs, W.T., Bowyer, K.W., et al., 2013. The challenge of face recognition from digital point-and-shoot cameras, in: Biometrics: Theory, Applications and Systems, Intl. Conf. on, IEEE. pp. 1–8.

Beveridge, J.R., Zhang, H., Draper, B.A., Flynn, P.J., Feng, Z., Huber, P., Kittler, J., Huang, Z., Li, S., Li, Y., et al., 2015. Report on the fg 2015 video person recognition evaluation, in: 2015 11th IEEE international conference and workshops on Automatic Face and Gesture Recognition (FG), IEEE. pp. 1–8.

Bhatt, H.S., Bharadwaj, S., Singh, R., Vatsa, M., 2012. Memetic approach for matching sketches with digital face images. Technical Report.

Bianco, S., 2017. Large age-gap face verification by feature injection in deep networks. Pattern Recognition Letters 90, 36–42.

Bodla, N., Zheng, J., Xu, H., Chen, J.C., Castillo, C., Chellappa, R., 2017. Deep heterogeneous feature fusion for template-based face recognition, in: Applications of Computer Vision, Winter Conf. on, IEEE. pp. 586–595.

Cai, X., Wang, C., Xiao, B., Chen, X., Zhou, J., 2012. Deep nonlinear metric learning with independent subspace analysis for face verification, in: Proceedings of ACM Intl. Conf. on Multimedia, ACM. pp. 749–752.

Calefati, A., Janjua, M.K., Nawaz, S., Gallo, I., 2018. Git loss for deep face recognition. arXiv preprint arXiv:1807.08512 .

Cambridge, A.L., 2018. The at&t database of faces. http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html .

Canévet, O., Fleuret, F., 2014. Efficient sample mining for object detection. Technical Report.

Cao, B., Wang, N., Gao, X., Li, J., 2018a. Asymmetric joint learning for heterogeneous face recognition, in: Thirty-Second AAAI Conference on Artificial Intelligence.

Cao, B., Wang, N., Li, J., Gao, X., 2018b. Data augmentation-based joint learning for heterogeneous face recognition. IEEE transactions on neural networks and learning systems .

Cao, K., Rong, Y., Li, C., Tang, X., Change Loy, C., 2018c. Pose-robust face recognition via deep residual equivariant mapping, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5187–5196.

Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A., 2017. Vggface2: A dataset for recognising faces across pose and age. arXiv preprint arXiv:1710.08092 .

Cao, X., Wei, Y., Wen, F., Sun, J., 2014. Face alignment by explicit shape regression. International Journal of Computer Vision 107, 177–190.

Cao, X., Wipf, D., Wen, F., Duan, G., Sun, J., 2013. A practical transfer learning algorithm for face verification, in: Proceedings of the IEEE Intl. Conf. on Computer Vision, pp. 3208–3215.

Castellani, U., Bartoli, A., 2012. 3d shape registration, in: 3D Imaging, Analysis and Applications. Springer, pp. 221–264.

Cevikalp, H., Serhan Yavuz, H., 2017. Fast and accurate face recognition with image sets, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1564–1572.

Chai, Z., Sun, Z., Mendez-Vazquez, H., He, R., Tan, T., 2014. Gabor ordinal measures for face recognition. IEEE trans. on information forensics and security 9, 14–26.

Chan, C.H., Zou, X., Poh, N., Kittler, J., 2014. Illumination invariant face recognition: a survey. Face Recognition in Adverse Conditions , 147–166.

Chan, T.H., Jia, K., Gao, S., Lu, J., Zeng, Z., Ma, Y., 2015. Pcanet: A simple deep learning baseline for image classification? IEEE trans. on Image Processing 24, 5017–5032.

Chen, B., Deng, W., 2016. Weakly-supervised deep self-learning for face recognition, in: Multimedia and Expo, Intl. Conf. on, IEEE. pp. 1–6.

Chen, B., Deng, W., Du, J., 2017a. 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, pp. 5372–5381.

Chen, B.C., Chen, C.S., Hsu, W.H., 2015a. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. IEEE trans. on Multimedia 17, 804–815.

Chen, C., Dantcheva, A., Ross, A., 2013a. Automatic facial makeup detection with application in face recognition, in: 2013 international conference on biometrics (ICB), IEEE. pp. 1–8.

Chen, C., Dantcheva, A., Swearingen, T., Ross, A., 2017b. Spoofing faces using makeup: An investigative study, in: 2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA), IEEE. pp. 1–8.

Chen, D., Cao, X., Wang, L., Wen, F., Sun, J., 2012. Bayesian face revisited: A joint formulation, in: European Conference on Computer Vision, Springer. pp. 566–579.

Chen, D., Cao, X., Wen, F., Sun, J., 2013b. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 3025–3032.

Chen, J., Yi, D., Yang, J., Zhao, G., Li, S.Z., Pietikainen, M., 2009. Learning mappings for face synthesis from near infrared to visual light images, in: Computer Vision and Pattern Recognition, Conf. on, IEEE. pp. 156–163.

Chen, J.C., Patel, V.M., Chellappa, R., 2016a. Unconstrained face verification using deep cnn features, in: Applications of Computer Vision, Winter Conf. on, IEEE. pp. 1–9.

Chen, J.C., Ranjan, R., Kumar, A., Chen, C.H., Patel, V.M., Chellappa, R., 2015b. An end-to-end system for unconstrained face verification with deep convolutional neural networks, in: Proceedings of IEEE Intl. Conf. on Computer Vision Workshops, pp. 118–126.

Chen, J.C., Zheng, J., Patel, V.M., Chellappa, R., 2016b. Fisher vector encoded deep convolutional features for unconstrained face verification, in: Image Processing, Intl. Conf. on, IEEE. pp. 2981–2985.

Chen, X., Xiao, B., Wang, C., Cai, X., Lv, Z., Shi, Y., 2013c. Modular hierarchical feature learning with deep neural networks for face verification, in: Image Processing, Intl. Conf. on, IEEE. pp. 3690–3694.

Chen, X.w., Aslan, M., Zhang, K., Huang, T., 2015c. Learning multi-channel deep feature representations for face recognition, in: Feature Extraction: Modern Questions and Challenges, pp. 60–71.

Cheng, G., Zhou, P., Han, J., 2018. Duplex metric learning for image set classification. IEEE Transactions on Image Processing 27, 281–292.

Cho, Y., Saul, L.K., 2009. Kernel methods for deep learning, in: Advances in neural information processing systems, pp. 342–350.

Choi, J., Sharma, A., Medioni, G., 2013. Comparing strategies for 3d face recognition from a 3d sensor, in: RO-MAN, IEEE. pp. 19–24.

Choi, Y., Kim, H.I., Ro, Y.M., 2016. Two-step learning of deep convolutional neural network for discriminative face recognition under varying illumination. Electronic Imaging 2016, 1–5.

Chowdhury, A.R., Lin, T.Y., Maji, S., Learned-Miller, E., 2015. Face identification with bilinear cnns. arXiv preprint arXiv: 1506.01342 .

Chowdhury, A.R., Lin, T.Y., Maji, S., Learned-Miller, E., 2016. One-to-many face recognition with bilinear cnns, in: Applications of Computer Vision, Winter Conf. on, IEEE. pp. 1–9.

Clevert, D.A., Unterthiner, T., Hochreiter, S., 2015. Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289 .

Crosswhite, N., Byrne, J., Parkhi, O.M., Stauffer, C., Cao, Q., Zisserman, A., 2016. Template adaptation for face verification and identification. arXiv preprint arXiv:1603.03958 .

Cui, J., Zhang, H., Han, H., Shan, S., Chen, X., 2018a. Improving 2d face recognition via discriminative face depth estimation, in: 2018 International Conference on Biometrics (ICB), IEEE. pp. 140–147.

Cui, Y., Song, Y., Sun, C., Howard, A., Belongie, S., 2018b. Large scale fine-grained categorization and domain-specific transfer learning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4109–4118.

Cui, Z., Li, W., Xu, D., Shan, S., Chen, X., 2013. Fusing robust face region descriptors via multiple metric learning for face recognition in the wild, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 3554–3561.

Dai, J., Li, Y., He, K., Sun, J., 2016. R-fcn: Object detection via region-based fully convolutional networks, in: Advances in neural information processing systems, pp. 379–387.

Dantcheva, A., Chen, C., Ross, A., 2012. Can facial cosmetics affect the matching accuracy of face recognition systems?, in: 2012 IEEE Fifth international conference on biometrics: theory, applications and systems (BTAS), IEEE. pp. 391–398.

Dapogny, A., Bailly, K., Cord, M., 2019. Decafa: Deep convolutional cascade for face alignment in the wild. arXiv preprint arXiv:1904.02549 .

DeepGlint, 2019. Trillion pairs testing faceset. http://trillionpairs.deepglint.com/overview .

Demirkus, M., Clark, J.J., Arbel, T., 2014. Robust semi-automatic head pose labeling for real-world face video sequences. Multimedia Tools and Applications 70, 495–523.

Deng, J., Cheng, S., Xue, N., Zhou, Y., Zafeiriou, S., 2018a. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7093–7102.

Deng, J., Guo, J., Zafeiriou, S., 2018b. Arcface: Additive angular margin loss for deep face recognition. arXiv preprint arXiv:1801.07698 .

Deng, J., Zhou, Y., Zafeiriou, S., 2017a. Marginal loss for deep face recognition, in: Proceedings of IEEE Intl. Conf. on Computer Vision and Pattern Recognition, Faces in-the-wild Workshop/Challenge.

Deng, W., Chen, B., Fang, Y., Hu, J., 2017b. Deep correlation feature learning for face verification in the wild. IEEE Signal Processing Letters 24, 1877–1881.

Deng, W., Hu, J., Zhang, N., Chen, B., Guo, J., 2017c. Fine-grained face verification: Fglfw database, baselines, and human-dcmn partnership. Pattern Recognition 66, 63–73.

Deng, Z., Peng, X., Li, Z., Qiao, Y., 2019. Mutual component convolutional neural networks for heterogeneous face recognition. IEEE Transactions on Image Processing .

Dhamecha, T.I., Verma, P., Shah, M., Singh, R., Vatsa, M., 2015. Annotated crowd video face database, in: Biometrics, 2015 Intl. Conf. on, IEEE. pp. 106–112.

Di, X., Zhang, H., Patel, V.M., 2018. Polarimetric thermal to visible face verification via attribute preserved synthesis, in: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS), IEEE. pp. 1–10.

Ding, C., Tao, D., 2015. Robust face recognition via multimodal deep face representation. IEEE trans. on Multimedia 17, 2049–2058.

Ding, C., Tao, D., 2016. A comprehensive survey on pose-invariant face recognition. ACM trans. on intelligent systems and technology 7, 37.

Ding, C., Tao, D., 2018. Trunk-branch ensemble convolutional neural networks for video-based face recognition. IEEE transactions on pattern analysis and machine intelligence 40, 1002–1014.

Ding, Y., Cheng, Y., Cheng, X., Li, B., You, X., Yuan, X., 2017. Noise-resistant network: a deep-learning method for face recognition under noise. EURASIP Journal on Image and Video Processing 2017, 43.

Dong, B., An, Z., Lin, J., Deng, W., 2017a. Attention-based template adaptation for face verification, in: Automatic Face & Gesture Recognition, Intl. Conf. on, IEEE. pp. 941–946.

Dong, Q., Gong, S., Zhu, X., 2017b. Class rectification hard mining for imbalanced deep learning, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1851–1860.

Dong, X., Yan, Y., Ouyang, W., Yang, Y., 2018a. Style aggregated network for facial landmark detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 379–388.

Dong, X., Yu, S.I., Weng, X., Wei, S.E., Yang, Y., Sheikh, Y., 2018b. Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 360–368.

Dong, Z., Jia, S., Zhang, C., Pei, M., 2016. Input aggregated network for face video representation. arXiv preprint arXiv:1603.06655 .

Du Terrail, J.O., Jurie, F., 2017. On the use of deep neural networks for the detection of small vehicles in ortho-images, in: 2017 IEEE International Conference on Image Processing (ICIP), IEEE. pp. 4212–4216.

Duchon, J., 1977. Splines minimizing rotation-invariant semi-norms in sobolev spaces, in: Constructive theory of functions of several variables. Springer, pp. 85–100.

Duong, C.N., Luu, K., Quach, K.G., Bui, T.D., et al., 2015. Beyond principal components: Deep boltzmann machines for face modeling., in: CVPR, p. 12.

Eigen, D., Fergus, R., 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture, in: Proceedings of the IEEE international conference on computer vision, pp. 2650–2658.

Essex, U., 2015. Faces96 database. https://cswww.essex.ac.uk/mv/allfaces/faces96.html .

Faltemier, T.C., Bowyer, K.W., Flynn, P.J., 2007. Using a multi-instance enrollment representation to improve 3d face recognition, in: 2007 First IEEE International Conference on Biometrics: Theory, Applications, and Systems, IEEE. pp. 1–6.

Fan, H., Cao, Z., Jiang, Y., Yin, Q., Doudou, C., 2014. Learning deep face representation. arXiv preprint arXiv:1403.2802 .

FG-NET, 2007. Fg-net aging database. http://webmail.cycollege.ac.cy/alanitis/fgnetaging/ , 147–166.

Fragoso, V., Ramanan, D., 2018. Bayesian embeddings for long-tailed datasets. International Conference on Learning Representations .

Fu, T.C., Chiu, W.C., Wang, Y.C.F., 2017. Learning guided convolutional neural networks for cross-resolution face recognition, in: Machine Learning for Signal Processing (MLSP), 2017 IEEE 27th International Workshop on, IEEE. pp. 1–5.

Fu, Z.P., Zhang, Y.N., Hou, H.Y., 2014. Survey of deep learning in face recognition, in: Orange Technologies, Intl. Conf. on, IEEE. pp. 5–8.

Galea, C., Farrugia, R.A., 2016. A large-scale software-generated face composite sketch database, in: 2016 International Conference of the Biometrics Special Interest Group (BIOSIG), IEEE. pp. 1–5.

Galea, C., Farrugia, R.A., 2017. Forensic face photo-sketch recognition using a deep learning-based architecture. IEEE Signal Processing Letters 24, 1586–1590.

Galea, C., Farrugia, R.A., 2018. Matching software-generated sketches to face photographs with a very deep cnn, morphed faces, and transfer learning. IEEE Transactions on Information Forensics and Security 13, 1421–1431.

Gan, Y., Yang, T., He, C., 2014. A deep graph embedding network model for face recognition, in: Signal Processing, Intl. Conf. on, IEEE. pp. 1268–1271.

Gao, S., Zhang, Y., Jia, K., Lu, J., Zhang, Y., 2015. Single sample face recognition via learning deep supervised autoencoders. trans. on Information Forensics and Security 10, 2108–2118.

Gao, W., Cao, B., Shan, S., Chen, X., Zhou, D., Zhang, X., Zhao, D., 2008. The cas-peal large-scale chinese face database and baseline evaluations. IEEE trans. on Systems, Man, and Cybernetics-Part A: Systems and Humans 38, 149–161.

Gecer, B., Balntas, V., Kim, T.K., 2017. Learning deep convolutional embeddings for face representation using joint sample-and set-based supervision, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1665–1672.

Gecer, B., Bhattarai, B., Kittler, J., Kim, T.K., 2018. Semi-supervised adversarial learning to generate photorealistic face images of new identities from 3d morphable model, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 217–234.

Georghiades, A.S., Belhumeur, P.N., Kriegman, D.J., 2001. From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE trans. on pattern analysis and machine intelligence 23, 643–660.

Ghiass, R.S., Arandjelović, O., Bendada, A., Maldague, X., 2014. Infrared face recognition: A comprehensive review of methodologies and databases. Pattern Recognition 47, 2807–2824.

Ghosh, S., Keshari, R., Singh, R., Vatsa, M., 2016. Face identification from low resolution near-infrared images, in: Image Processing, Intl. Conf. on, IEEE. pp. 938–942.

Goh, R., Liu, L., Liu, X., Chen, T., 2005. The cmu face in action (fia) database, in: Intl. Workshop on Analysis and Modeling of Faces and Gestures, Springer. pp. 255–263.

Gong, D., Li, Z., Huang, W., Li, X., Tao, D., 2017. Heterogeneous face recognition: A common encoding feature discriminant approach. IEEE trans. on Image Processing 26, 2079–2089.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Advances in neural information processing systems, pp. 2672–2680.

Goodfellow, I.J., Warde-Farley, D., Mirza, M., Courville, A., Bengio, Y., 2013. Maxout networks. arXiv preprint arXiv:1302.4389 .

Goswami, D., Chan, C.H., Windridge, D., Kittler, J., 2011. Evaluation of face recognition system in heterogeneous environments (visible vs nir), in: Computer Vision Workshops, Intl. Conf. on, IEEE. pp. 2160–2167.

Goswami, G., Bharadwaj, S., Vatsa, M., Singh, R., 2013. On rgb-d face recognition using kinect, in: 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), IEEE. pp. 1–6.

Goswami, G., Bhardwaj, R., Singh, R., Vatsa, M., 2014. Mdlface: Memorability augmented deep learning for video face recognition, in: Biometrics, Intl. Joint Conf. on, IEEE. pp. 1–7.

Goswami, G., Vatsa, M., Singh, R., 2017. Face verification via learned representation on feature-rich video frames. trans. on Information Forensics and Security 12, 1686–1698.

Grgic, M., Delac, K., Grgic, S., 2011. Scface–surveillance cameras face database. Multimedia tools and applications 51, 863–879.

Grm, K., Dobrisek, S., Struc, V., 2016. Deep pair-wise similarity learning for face recognition, in: Biometrics and Forensics, Intl. Workshop on, IEEE. pp. 1–6.

Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S., 2010. Multi-pie. Image and Vision Computing 28, 807–813.

Grother, P., Ngan, M., 2017. The ijb-a face identification challenge performance report, in: Technical report. National Institute of Standards and Technology.

Gruber, I., Hlaváč, M., Železnỳ, M., Karpov, A., 2017. Facing face recognition with resnet: Round one, in: Intl. Conf. on Interactive Collaborative Robotics, Springer. pp. 67–74.

Grundström, J., 2015. Face verification and open-set identification for real-time video applications. Master's Theses in Mathematical Sciences .

Gunther, M., Cruz, S., Rudd, E.M., Boult, T.E., 2017. Toward open-set face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 71–80.

Günther, M., Hu, P., Herrmann, C., Chan, C.H., Jiang, M., Yang, S., Dhamija, A.R., Ramanan, D., Beyerer, J., Kittler, J., et al., 2017. Unconstrained face detection and open-set face recognition challenge. arXiv preprint arXiv:1708.02337 .

Guo, G., 2014. Heterogeneous face recognition: An emerging topic in biometrics. Intel Technology Journal 18, 80–97.

Guo, G., Li, S.Z., Chan, K., 2000. Face recognition by support vector machines, in: Automatic Face and Gesture Recognition, Proceedings. Int'l Conf. on, IEEE. pp. 196–201.

Guo, G., Mu, G., Ricanek, K., 2010. Cross-age face recognition on a very large database: The performance versus age intervals and improvement using soft biometric traits, in: Pattern Recognition, Intl. Conf. on, IEEE. pp. 3392–3395.

Guo, G., Wen, L., Yan, S., 2014. Face authentication with makeup changes. IEEE trans. on Circuits and Systems for Video Technology 24, 814–825.

Guo, G., Zhang, N., 2018. What is the challenge for deep learning in unconstrained face recognition?, in: Automatic Face and Gesture Recognition, 2018 13th IEEE International Conference on, IEEE. pp. 436–442.

Guo, G.D., Zhang, H.J., 2001. Boosting for fast face recognition, in: Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, Proceedings. IEEE ICCV Workshop on, IEEE. pp. 96–100.

Guo, Y., Zhang, L., 2017. One-shot face recognition by promoting underrepresented classes. arXiv preprint arXiv:1707.05574 .

Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J., 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition, in: European Conf. on Computer Vision, Springer. pp. 87–102.

Gupta, S., Castleman, K.R., Markey, M.K., Bovik, A.C., 2010. Texas 3d face recognition database, in: 2010 IEEE Southwest Symposium on Image Analysis & Interpretation (SSIAI), IEEE. pp. 97–100.

Hadsell, R., Chopra, S., LeCun, Y., 2006. Dimensionality reduction by learning an invariant mapping, in: Computer vision and pattern recognition, computer society conf. on, IEEE. pp. 1735–1742.

Han, C., Shan, S., Kan, M., Wu, S., Chen, X., 2018. Face recognition with contrastive convolution, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 118–134.

Han, H., Klare, B.F., Bonnen, K., Jain, A.K., 2013. Matching composite sketches to face photos: A component-based approach. IEEE Transactions on Information Forensics and Security 8, 191–204.

Hasnat, A., Bohné, J., Gentric, S., Chen, L., 2017. Deepvisage: Making face recognition simple yet with powerful generalization skills. arXiv preprint arXiv:1703.08388 .

Hassner, T., Masi, I., Kim, J., Choi, J., Harel, S., Natarajan, P., Medioni, G., 2016. Pooling faces: template based face recognition with pooled face images, in: Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition Workshops, pp. 59–67.

Hayat, M., Bennamoun, M., An, S., 2014. Learning non-linear reconstruction models for image set classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1907–1914.

Hayat, M., Bennamoun, M., An, S., 2015. Deep reconstruction models for image set classification. IEEE transactions on pattern analysis and machine intelligence 37, 713–727.

Hayat, M., Khan, S.H., Werghi, N., Goecke, R., 2017. Joint registration and representation learning for unconstrained face identification., in: CVPR, pp. 1551–1560.

Haykin, S., 2009. Neural networks and learning machines. volume 3. Pearson education Upper Saddle River.

He, H., Garcia, E.A., 2008. Learning from imbalanced data. IEEE Transactions on Knowledge & Data Engineering , 1263–1284.

He, K., Zhang, X., Ren, S., Sun, J., 2015a. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE Intl. Conf. on computer vision, pp. 1026–1034.

He, K., Zhang, X., Ren, S., Sun, J., 2016a. Deep residual learning for image recognition, in: Proceedings of the IEEE Conf. on computer vision and pattern recognition, pp. 770–778.

He, L., Li, H., Zhang, Q., Sun, Z., 2018a. Dynamic feature learning for partial face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7054–7063.

He, L., Li, H., Zhang, Q., Sun, Z., He, Z., 2016b. Multiscale representation for partial face recognition under near infrared illumination, in: 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS), IEEE. pp. 1–7.

He, R., Cai, Y., Tan, T., Davis, L., 2015b. Learning predictable binary codes for face indexing. Pattern Recognition 48, 3160–3168.

He, R., Cao, J., Song, L., Sun, Z., Tan, T., 2019. Cross-spectral face completion for nir-vis heterogeneous face recognition. arXiv preprint arXiv: 1902.03565 .

He, R., Wu, X., Sun, Z., Tan, T., 2017. Learning invariant deep representation for nir-vis face recognition, in: Thirty-First AAAI Conference on Artificial Intelligence.

He, R., Wu, X., Sun, Z., Tan, T., 2018b. Wasserstein cnn: Learning invariant features for nir-vis face recognition. IEEE transactions on pattern analysis and machine intelligence .

Herrmann, C., Willersinn, D., Beyerer, J., 2017. Residual vs. inception vs. classical networks for low-resolution face recognition, in: Scandinavian Conf. on Image Analysis, Springer. pp. 377–388.

Hsieh, H.L., Hsu, W., Chen, Y.Y., 2017. Multi-task learning for face identification and attribute estimation, in: Acoustics, Speech and Signal Processing, Intl. Conf. on, IEEE. pp. 2981–2985.

Hu, G., Hua, Y., Yuan, Y., Zhang, Z., Lu, Z., Mukherjee, S.S., Hospedales, T.M., Robertson, N.M., Yang, Y., 2017a. Attribute-enhanced face recognition with neural tensor fusion networks, in: Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition, pp. 3744–3753.

Hu, J., Lu, J., Tan, Y.P., 2014. Discriminative deep metric learning for face verification in the wild, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1875–1882.

Hu, L., Kan, M., Shan, S., Song, X., Chen, X., 2017b. Ldf-net: Learning a displacement field network for face recognition across pose, in: Automatic Face & Gesture Recognition, Intl. Conf. on, IEEE. pp. 9–16.

Hu, S., Short, N.J., Riggan, B.S., Gordon, C., Gurton, K.P., Thielke, M., Gurram, P., Chan, A.L., 2016. A polarimetric thermal database for face recognition research, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 119–126.

Hu, Y., Wu, X., He, R., 2017c. Attention-set based metric learning for video face recognition. arXiv preprint arXiv:1704.03805 .

Huang, C., Li, Y., Change Loy, C., Tang, X., 2016a. Learning deep representation for imbalanced classification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5375–5384.

Huang, D., Sun, J., Wang, Y., 2012a. The buaa-visnir face database instructions. School Comput. Sci. Eng., Beihang Univ., Beijing, China, Tech. Rep. IRIP-TR-12-FR-001 .

Huang, G.B., Lee, H., Learned-Miller, E., 2012b. Learning hierarchical representations for face verification with convolutional deep belief networks, in: Computer Vision and Pattern Recognition, Conf. on, IEEE. pp. 2518–2525.

Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E., 2007. Labeled faces

in the wild: A database for studying face recognition in unconstrained environments. Technical Report. Technical Report 07-49, University of Massachusetts, Amherst.

Huang, J., Yuan, C., 2015. Weighted-pcanet for face recognition, in: Intl. Conf. on Neural Information Processing, Springer. pp. 246–254.

Huang, R., Liu, C., Li, G., Zhou, J., 2016b. Adaptive deep supervised autoencoder based image reconstruction for face recognition. Mathematical Problems in Engineering 2016.

Huang, R., Zhang, S., Li, T., He, R., 2017. 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, pp. 2439–2448.

Huang, Z., Shan, S., Wang, R., Zhang, H., Lao, S., Kuerban, A., Chen, X., 2015. A benchmark and comparative study of video-based face recognition on cox face database. IEEE trans. on Image Processing 24, 5967–5981.

Huang, Z., Shan, S., Zhang, H., Lao, S., Kuerban, A., Chen, X., 2012c. Benchmarking still-to-video face recognition via partial and local linear discriminant analysis on cox-s2v dataset, in: Asian Conf. on Computer Vision, Springer. pp. 589–600.

Huo, J., Gao, Y., Shi, Y., Yang, W., Yin, H., 2016. Ensemble of sparse cross-modal metrics for heterogeneous face recognition, in: Proceedings of the 24th ACM international conference on Multimedia, ACM. pp. 1405–1414.

Iranmanesh, S.M., Dabouei, A., Kazemi, H., Nasrabadi, N.M., 2018. Deep cross polarimetric thermal-to-visible face recognition, in: 2018 International Conference on Biometrics (ICB), IEEE. pp. 166–173.

Iranmanesh, S.M., Kazemi, H., Soleymani, S., Dabouei, A., Nasrabadi, N.M., 2019. Deep sketch-photo face recognition assisted by facial attributes, in: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS), IEEE. pp. 1–10.

Jhuang, D.H., Lin, D.T., Tsai, C.H., 2016. Face verification with three-dimensional point cloud by using deep belief networks, in: Pattern Recognition, Intl. Conf. on, IEEE. pp. 1430–1435.

Jones, M., Kobori, H., 2017. Improving face verification and person re-identification accuracy using hyperplane similarity, in: Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1555–1563.

Kan, M., Shan, S., Chang, H., Chen, X., 2014. Stacked progressive auto-encoders (spae) for face recognition across poses, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1883–1890.

Kan, M., Shan, S., Chen, X., 2016. Multi-view deep network for cross-view classification, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 4847–4855.

Kang, B.N., Kim, Y., Kim, D., 2017. Deep convolutional neural network using triplets of faces, deep ensemble, and score-level fusion for face recognition, in: Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition Workshops, pp. 109–116.

Kang, B.N., Kim, Y., Kim, D., 2018. Pairwise relational networks for face recognition, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 628–645.

Kazemi, H., Soleymani, S., Dabouei, A., Iranmanesh, M., Nasrabadi, N.M., 2018. Attribute-centered loss for soft-biometrics guided face sketch-photo recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 499–507.

Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E., 2016. The megaface benchmark: 1 million faces for recognition at scale, in: Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition, pp. 4873–4882.

Kevin, X.C.P.J.F., Bowyer, W., 2003. Visible-light and infrared face recognition, in: Workshop on Multimodal User Authentication, Citeseer. p. 48.

Kim, D., Hernandez, M., Choi, J., Medioni, G., 2017. Deep 3d face identification, in: 2017 IEEE International Joint Conference on Biometrics (IJCB), IEEE. pp. 133–142.

Kim, K., Yang, Z., Masi, I., Nevatia, R., Medioni, G., 2018. Face and body association for video-based face recognition, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE. pp. 39–48.

Kim, M., Kumar, S., Pavlovic, V., Rowley, H., 2008. Face tracking and recognition with visual constraints in real-world videos, in: Computer Vision and Pattern Recognition, Conf. on, IEEE. pp. 1–8.

Klare, B.F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Jain, A.K., 2015. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a, in: Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1931–1939.

Klum, S.J., Han, H., Klare, B.F., Jain, A.K., 2014. The facesketchid system: Matching facial composites to mugshots. IEEE Transactions on Information Forensics and Security 9, 2248–2263.

Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H., 2012. Large scale metric learning from equivalence constraints, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conf. on, IEEE. pp. 2288–2295.

Kong, S.G., Heo, J., Abidi, B.R., Paik, J., Abidi, M.A., 2005. Recent advances in visual and infrared face recognitiona review. Computer Vision and Image Understanding 97, 103–135.

Kowalski, M., Naruniec, J., Trzcinski, T., 2017. Deep alignment network: A convolutional neural network for robust face alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 88–97.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.

Kumar, A., Chellappa, R., 2018. Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 430–439.

Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K., 2009. Attribute and simile classifiers for face verification, in: Computer Vision, Intl. Conf. on, IEEE. pp. 365–372.

Lahasan, B., Lutfi, S.L., San-Segundo, R., 2017. A survey on techniques to handle face recognition challenges: occlusion, single sample per subject and expression. Artificial Intelligence Review , 1–31.

Lawrence, S., Giles, C.L., Tsoi, A.C., Back, A.D., 1997. Face recognition: A convolutional neural-network approach. IEEE trans. on neural networks 8, 98–113.

Learned-Miller, E., Huang, G.B., RoyChowdhury, A., Li, H., Hua, G., 2016. Labeled faces in the wild: A survey, in: Advances in face detection and facial image analysis. Springer, pp. 189–248.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278–2324.

Lee, K.C., Ho, J., Yang, M.H., Kriegman, D., 2003. Video-based face recognition using probabilistic appearance manifolds, in: Computer vision and pattern recognition, computer society Conf. on, IEEE. pp. I–I.

Lee, Y.C., Chen, J., Tseng, C.W., Lai, S.H., 2016. Accurate and robust face recognition from rgb-d images with a deep learning approach., in: BMVC.

Leng, B., Yu, K., Jingyan, Q., 2017. Data augmentation for unbalanced face recognition training sets. Neurocomputing 235, 10–14.

Lezama, J., Qiu, Q., Sapiro, G., 2017. Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding, in: Conf. on Computer Vision and Pattern Recognition, IEEE. pp. 6807–6816.

Li, B.Y., Mian, A.S., Liu, W., Krishna, A., 2013a. Using kinect for face recognition under varying poses, expressions, illumination and disguise, in: Applications of Computer Vision, Workshop on, IEEE. pp. 186–192.

Li, H., Hu, H., Yip, C., 2018a. Age-related factor guided joint task modeling convolutional neural network for cross-age face recognition. IEEE Transactions on Information Forensics and Security 13, 2383–2392.

Li, H., Hua, G., 2015. Hierarchical-pep model for real-world face recognition, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 4055–4064.

Li, S., Xing, J., Niu, Z., Shan, S., Yan, S., 2015a. Shape driven kernel adaptation in convolutional neural network for robust facial traits recognition, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 222–230.

Li, S., Yi, D., Lei, Z., Liao, S., 2013b. The casia nir-vis 2.0 face database, in: Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition Workshops, pp. 348–353.

Li, S.Z., Lei, Z., Ao, M., 2009. The hfb face database for heterogeneous face biometrics research, in: Computer Vision and Pattern Recognition Workshops, Computer Society Conf. on, IEEE. pp. 1–8.

Li, Y., Song, L., Wu, X., He, R., Tan, T., 2018b. Anti-makeup: Learning a bi-level adversarial network for makeup-invariant face verification, in: Thirty-Second AAAI Conference on Artificial Intelligence.

Li, Y., Wang, G., Lin, L., Chang, H., 2015b. A deep joint learning approach for age invariant face verification, in: CCF Chinese Conf. on Computer Vision, Springer. pp. 296–305.

Li, Y., Wang, G., Nie, L., Wang, Q., Tan, W., 2018c. Distance metric optimization driven convolutional neural network for age invariant face recognition. Pattern Recognition 75, 51–62.

Lin, L., Wang, G., Zuo, W., Feng, X., Zhang, L., 2017a. Cross-domain visual matching via generalized similarity measure and feature learning. trans. on

pattern analysis and machine intelligence 39, 1089–1102.

Lin, M., Fan, X., 2011. Low resolution face recognition with pose variations using deep belief networks, in: Image and Signal Processing, Intl. Congress on, IEEE. pp. 1522–1526.

Lin, S.H., Kung, S.Y., Lin, L.J., 1997. Face recognition/detection by probabilistic decision-based neural network. IEEE trans. on neural networks 8, 114–132.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017b. Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988.

Lin, T.Y., RoyChowdhury, A., Maji, S., 2015. Bilinear cnn models for fine-grained visual recognition, in: Proceedings of the IEEE Intl. Conf. on Computer Vision, pp. 1449–1457.

Liou, C.Y., Cheng, W.C., Liou, J.W., Liou, D.R., 2014. Autoencoder for words. Neurocomputing 139, 84–96.

Liu, D., Wang, N., Peng, C., Li, J., Gao, X., 2018a. Deep attribute guided representation for heterogeneous face recognition., in: IJCAI, pp. 835–841.

Liu, H., He, F., Zhao, Q., Fei, X., 2017a. Matching depth to rgb for boosting face verification, in: Chinese Conf. on Biometric Recognition, Springer. pp. 127–134.

Liu, J., Deng, Y., Bai, T., Wei, Z., Huang, C., 2015. Targeting ultimate accuracy: Face recognition via deep embedding. arXiv preprint arXiv:1506.07310 .

Liu, J., Fang, C., Wu, C., 2016a. A fusion face recognition approach based on 7-layer deep learning neural network. Journal of Electrical and Computer Engineering 2016.

Liu, L., Zhang, L., Liu, H., Yan, S., 2014. Toward large-population face identification in unconstrained videos. IEEE trans. on Circuits and Systems for Video Technology 24, 1874–1884.

Liu, W., Lin, R., Liu, Z., Liu, L., Yu, Z., Dai, B., Song, L., 2018b. Learning towards minimum hyperspherical energy, in: Advances in Neural Information Processing Systems, pp. 6222–6233.

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L., 2017b. Sphereface: Deep hypersphere embedding for face recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 212–220.

Liu, W., Wen, Y., Yu, Z., Yang, M., 2016b. Large-margin softmax loss for convolutional neural networks., in: ICML, pp. 507–516.

Liu, X., Kan, M., Wu, W., Shan, S., Chen, X., 2017c. Viplfacenet: an open source deep face recognition sdk. Frontiers of Computer Science 11, 208–218.

Liu, X., Song, L., Wu, X., Tan, T., 2016c. Transferring deep representation for nir-vis heterogeneous face recognition, in: Biometrics, Intl. Conf. on, IEEE. pp. 1–8.

Liu, X., Vijaya Kumar, B., Yang, C., Tang, Q., You, J., 2018c. Dependency-aware attention control for unconstrained face recognition with image sets, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 548–565.

Liu, Y., Li, H., Wang, X., 2017d. Rethinking feature discrimination and polymerization for large-scale recognition. arXiv preprint arXiv:1710.00870 .

Liu, Y., Shi, P., Peng, B., Yan, H., Zhou, Y., Han, B., Zheng, Y., Lin, C., Jiang, J., Fan, Y., et al., 2018d. iqiyi-vid: A large dataset for multi-modal person identification. arXiv preprint arXiv:1811.07548 .

Liu, Y., Wei, F., Shao, J., Sheng, L., Yan, J., Wang, X., 2018e. Exploring disentangled feature representation beyond face identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2080–2089.

Liu, Y., Yan, J., Ouyang, W., 2017e. Quality aware network for set to set recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5790–5799.

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X., 2019. Large-scale long-tailed recognition in an open world. arXiv preprint arXiv:1904.05160 .

Lu, B., Zheng, J., Chen, J.C., Chellappa, R., 2017a. Pose-robust face verification by exploiting competing tasks, in: Applications of Computer Vision, Winter Conf. on, IEEE. pp. 1124–1132.

Lu, C., Tang, X., 2015. Surpassing human-level face verification performance on lfw with gaussianface., in: AAAI, pp. 3811–3819.

Lu, J., Liong, V.E., Wang, G., Moulin, P., 2015a. Joint feature learning for face recognition. IEEE trans. on Information Forensics and Security 10, 1371–1383.

Lu, J., Wang, G., Deng, W., Moulin, P., Zhou, J., 2015b. Multi-manifold deep metric learning for image set classification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1137–1145.

Lu, J., Wang, G., Zhou, J., 2017b. Simultaneous feature and dictionary learning for image set based face recognition. IEEE Transactions on Image Processing 26, 4042–4054.

Lu, X., Yang, Y., Zhang, W., Wang, Q., Wang, Y., 2017c. Face verification with multi-task and multi-scale feature fusion. Entropy 19, 228.

Lu, Z., Jiang, X., Kot, A., 2018. Deep coupled resnet for low-resolution face recognition. IEEE Signal Processing Letters 25, 526–530.

Lumini, A., Nanni, L., Ghidoni, S., 2016. Deep features combined with handcrafted features for face recognition. Intl. Journal of Computer Research 23, 123.

Lv, J.J., Shao, X.H., Huang, J.S., Zhou, X.D., Zhou, X., 2017. Data augmentation for face recognition. Neurocomputing 230, 184–196.

Maas, A.L., Hannun, A.Y., Ng, A.Y., 2013. Rectifier nonlinearities improve neural network acoustic models, in: Proc. ICML.

Maeng, H., Liao, S., Kang, D., Lee, S.W., Jain, A.K., 2012. Nighttime face recognition at long distance: Cross-distance and cross-spectral matching, in: Asian Conf. on Computer Vision, Springer. pp. 708–721.

Mandal, B., Lim, R.Y., Dai, P., Sayed, M.R., Li, L., Lim, J.H., 2016. Trends in machine and human face recognition, in: Advances in Face Detection and Facial Image Analysis. Springer, pp. 145–187.

Martinez, A., Benavente, R., 2007. The ar face database, 1998. Computer Vision Center, Technical Report 3, 5.

Masi, I., Chang, F.J., Choi, J., Harel, S., Kim, J., Kim, K., Leksut, J., Rawls, S., Wu, Y., Hassner, T., et al., 2019a. Learning pose-aware models for pose-invariant face recognition in the wild. IEEE transactions on pattern analysis and machine intelligence 41, 379–393.

Masi, I., Rawls, S., Medioni, G., Natarajan, P., 2016a. Pose-aware face recognition in the wild, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 4838–4846.

Masi, I., Trn, A.T., Hassner, T., Leksut, J.T., Medioni, G., 2016b. Do we really need to collect millions of faces for effective face recognition?, in: European Conf. on Computer Vision, Springer. pp. 579–596.

Masi, I., Trn, A.T., Hassner, T., Sahin, G., Medioni, G., 2019b. Face-specific data augmentation for unconstrained face recognition. International Journal of Computer Vision , 1–26.

Maze, B., Adams, J., Duncan, J.A., Kalka, N., Miller, T., Otto, C., Jain, A.K., Niggel, W.T., Anderson, J., Cheney, J., Grother, P., 2018. Iarpa janus benchmark-c: Face dataset and protocol., in: 2018 IEEE International Conference on Biometrics (ICB).

Meier, L., Van De Geer, S., Bühlmann, P., 2008. The group lasso for logistic regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70, 53–71.

Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G., 1999. Xm2vtsdb: The extended m2vts database, in: Second international conference on audio and video-based biometric person authentication, pp. 965–966.

Miao, X., Zhen, X., Liu, X., Deng, C., Athitsos, V., Huang, H., 2018. Direct shape regression networks for end-to-end face alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5040–5049.

Min, R., Kose, N., Dugelay, J.L., 2014. Kinectfacedb: A kinect database for face recognition. IEEE Transactions on Systems, Man, and Cybernetics: Systems 44, 1534–1548.

Mittal, P., Jain, A., Goswami, G., Vatsa, M., Singh, R., 2017. Composite sketch recognition using saliency and attribute feedback. Information Fusion 33, 86–99.

Mittal, P., Vatsa, M., Singh, R., 2015. Composite sketch recognition via deep network-a transfer learning approach, in: 2015 International Conference on Biometrics (ICB), IEEE. pp. 251–256.

Moghaddam, B., Jebara, T., Pentland, A., 2000. Bayesian face recognition. Pattern Recognition 33, 1771–1782.

Mohammadzade, H., Hatzinakos, D., 2013. Iterative closest normal point for 3d face recognition. trans. on pattern analysis and machine intelligence 35, 381–397.

Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S., 2017. Agedb: the first manually collected, in-the-wild age database. . .

Murtaza, M., Sharif, M., Raza, M., Shah, J.H., 2013. Analysis of face recognition under varying facial expression: a survey. Int. Arab J. Inf. Technol. 10, 378–388.

Nagpal, S., Singh, M., Singh, R., Vatsa, M., 2015. Regularized deep learning for face recognition with weight variations. IEEE Access 3, 3010–3018.

Nech, A., Kemelmacher-Shlizerman, I., 2016. Megaface 2: 672,057 identities for face recognition, in: .

Ng, H.W., Winkler, S., 2014. A data-driven approach to cleaning large face datasets, in: Image Processing, Intl. Conf. on, IEEE. pp. 343–347.

Nikisins, O., Nasrollahi, K., Greitans, M., Moeslund, T.B., 2014. Rgb-dt based face recognition, in: Pattern Recognition, Intl. Conf. on, IEEE. pp. 1716–1721.

Oh, S.K., Yoo, S.H., Pedrycz, W., 2013. Design of face recognition algorithm using pca-lda combined for hybrid data pre-processing and polynomial-based rbf neural networks: Design and its application. Expert Systems with Applications 40, 1451–1466.

Oh Song, H., Xiang, Y., Jegelka, S., Savarese, S., 2016. Deep metric learning via lifted structured feature embedding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4004–4012.

Ouyang, S., Hospedales, T., Song, Y.Z., Li, X., Loy, C.C., Wang, X., 2016a. A survey on heterogeneous face recognition: Sketch, infra-red, 3d and low-resolution. Image and Vision Computing 56, 28–48.

Ouyang, S., Hospedales, T.M., Song, Y.Z., Li, X., 2016b. Forgetmenot: Memory-aware forensic facial sketch matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5571–5579.

Ouyang, W., Wang, X., Zhang, C., Yang, X., 2016c. Factors in finetuning deep model for object detection with long-tail distribution, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 864–873.

Parchami, M., Bashbaghi, S., Granger, E., 2017a. Video-based face recognition using ensemble of haar-like deep convolutional neural networks, in: Neural Networks, Intl. Joint Conf. on, IEEE. pp. 4625–4632.

Parchami, M., Bashbaghi, S., Granger, E., Sayed, S., 2017b. Using deep autoencoders to learn robust domain-invariant representations for still-to-video face recognition, in: Advanced Video and Signal Based Surveillance, Intl. Conf. on, IEEE. pp. 1–6.

Park, B.J., Oh, S.K., Kim, H.K., 2008. Design of polynomial neural network classifier for pattern classification with two classes. Journal of Electrical Engineering and Technology 3, 108–114.

Park, S., Yu, J., Jeon, M., 2017. Learning feature representation for face verification, in: Advanced Video and Signal Based Surveillance, Intl. Conf. on, IEEE. pp. 1–6.

Parkhi, O.M., Vedaldi, A., Zisserman, A., et al., 2015. Deep face recognition., in: BMVC, p. 6.

Pathirage, C.S.N., Li, L., Liu, W., 2016. Discriminant auto encoders for face recognition with expression and pose variations, in: Pattern Recognition, Intl. Conf. on, IEEE. pp. 3512–3517.

Pathirage, C.S.N., Li, L., Liu, W., Zhang, M., 2015. Stacked face de-noising auto encoders for expression-robust face recognition, in: Digital Image Computing: Techniques and Applications, Intl. Conf. on, IEEE. pp. 1–8.

Patil, H., Kothari, A., Bhurchandi, K., 2015. 3-d face recognition: features, databases, algorithms and challenges. Artificial Intelligence Review 44, 393–441.

Peng, C., Gao, X., Wang, N., Li, J., 2018. Face recognition from multiple stylistic sketches: Scenarios, datasets, and evaluation. Pattern Recognition 84, 262–272.

Peng, C., Wang, N., Li, J., Gao, X., 2019. Dlface: Deep local descriptor for cross-modality face recognition. Pattern Recognition 90, 161–171.

Peng, X., Ratha, N., Pankanti, S., 2016. Learning face recognition from limited training data using deep neural networks, in: Pattern Recognition, Intl. Conf. on, IEEE. pp. 1442–1447.

Peng, X., Yu, X., Sohn, K., Metaxas, D.N., Chandraker, M., 2017. Reconstruction-based disentanglement for pose-invariant face recognition, in: Proceedings of the IEEE international conference on computer vision, pp. 1623–1632.

Phillips, P.J., 2010. Face recognition grand challenge (frgc). https://www.nist.gov/programs-projects/face-recognition-grand-challenge-frgc .

Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W., 2005. Overview of the face recognition grand challenge, in: Computer vision and pattern recognition, computer society Conf. on, IEEE. pp. 947–954.

Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J., 2000. The feret evaluation methodology for face-recognition algorithms. IEEE trans. on pattern analysis and machine intelligence 22, 1090–1104.

Pinto, N., Stone, Z., Zickler, T., Cox, D., 2011. Scaling up biologically-inspired computer vision: A case study in unconstrained face recognition on facebook, in: Computer Vision and Pattern Recognition Workshops, Computer Society Conf. on, IEEE. pp. 35–42.

Qi, C., Su, F., 2017. Contrastive-center loss for deep neural networks. arXiv preprint arXiv:1707.07391 .

Qi, X., Zhang, L., 2018. Face recognition via centralized coordinate learningn, in: arXiv preprint arXiv:1801.05678.

Ranjan, R., Castillo, C.D., Chellappa, R., 2017. L2-constrained softmax loss for discriminative face verification. arXiv preprint arXiv:1703.09507 .

Ranjan, R., Sankaranarayanan, S., Castillo, C.D., Chellappa, R., 2016. An all-in-one convolutional neural network for face analysis. arXiv preprint arXiv:1611.00851 .

Rao, Y., Lin, J., Lu, J., Zhou, J., 2017a. Learning discriminative aggregation network for video-based face recognition, in: Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition, pp. 3781–3790.

Rao, Y., Lu, J., Zhou, J., 2017b. Attention-aware deep reinforcement learning for video face recognition, in: Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition, pp. 3931–3940.

Reale, C., Lee, H., Kwon, H., Chellappa, R., 2017. Deep network shrinkage applied to cross-spectrum face recognition, in: Automatic Face & Gesture Recognition, Intl. Conf. on, IEEE. pp. 897–903.

Reale, C., Nasrabadi, N.M., Kwon, H., Chellappa, R., 2016. Seeing the forest from the trees: A holistic approach to near-infrared heterogeneous face recognition, in: Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition Workshops, pp. 54–62.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 779–788.

Ren, S., Cao, X., Wei, Y., Sun, J., 2014. Face alignment at 3000 fps via regressing local binary features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1685–1692.

Ricanek, K., Tesafaye, T., 2006. Morph: A longitudinal image database of normal adult age-progression, in: Automatic Face and Gesture Recognition, Intl. Conf. on, IEEE. pp. 341–345.

Riggan, B.S., Reale, C., Nasrabadi, N.M., 2015. Coupled auto-associative neural networks for heterogeneous face recognition. IEEE Access 3, 1620–1632.

Riggan, B.S., Short, N.J., Hu, S., 2016a. Optimal feature learning and discriminative framework for polarimetric thermal to visible face recognition, in: Applications of Computer Vision, Winter Conf. on, IEEE. pp. 1–7.

Riggan, B.S., Short, N.J., Hu, S., 2018. Thermal to visible synthesis of face images using multiple regions, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE. pp. 30–38.

Riggan, B.S., Short, N.J., Hu, S., Kwon, H., 2016b. Estimation of visible spectrum faces from polarimetric thermal faces, in: 2016 IEEE 8th international conference on biometrics theory, applications and systems (BTAS), IEEE. pp. 1–7.

Rothe, R., Timofte, R., Van Gool, L., 2015. Dex: Deep expectation of apparent age from a single image, in: Proceedings of IEEE Intl. Conf. on Computer Vision Workshops, pp. 10–15.

Rusu, R.B., Cousins, S., 2011. 3d is here: Point cloud library (pcl), in: Robotics and automation, Intl. Conf. on, IEEE. pp. 1–4.

Salakhutdinov, R., Hinton, G., 2009. Deep boltzmann machines, in: Artificial Intelligence and Statistics, pp. 448–455.

Sankaran, N., Tulyakov, S., Setlur, S., Govindaraju, V., 2018. Metadata-based feature aggregation network for face recognition, in: 2018 International Conference on Biometrics (ICB), IEEE. pp. 118–123.

Sankaranarayanan, S., Alavi, A., Castillo, C.D., Chellappa, R., 2016a. Triplet probabilistic embedding for face verification and clustering, in: Biometrics Theory, Applications and Systems, Intl. Conf. on, IEEE. pp. 1–8.

Sankaranarayanan, S., Alavi, A., Chellappa, R., 2016b. Triplet similarity embedding for face verification. arXiv preprint arXiv:1602.03418 .

Saragih, J., Goecke, R., 2007. A nonlinear discriminative approach to aam fitting, in: 2007 IEEE 11th International Conference on Computer Vision, IEEE. pp. 1–8.

Sarfraz, M.S., Stiefelhagen, R., 2017. Deep perceptual mapping for cross-modal face recognition. Intl. Journal of Computer Vision 122, 426–438.

Savchenko, A.V., Belova, N.S., 2017. Maximum a posteriori estimation of distances between deep features in still-to-video face recognition. arXiv preprint arXiv:1708.07972 .

Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., Akarun, L., 2008. Bosphorus database for 3d face analysis, in: European Workshop on Biometrics and Identity Management, Springer. pp. 47–56.

Saxena, S., Verbeek, J., 2016. Heterogeneous face recognition with cnns, in: Computer Vision–ECCV Workshops, Springer. pp. 483–491.

Scherer, D., Müller, A., Behnke, S., 2010. Evaluation of pooling operations in

convolutional architectures for object recognition, in: International conference on artificial neural networks, Springer. pp. 92–101.

Schroff, F., Kalenichenko, D., Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 815–823.

Sengupta, S., Chen, J.C., Castillo, C., Patel, V.M., Chellappa, R., Jacobs, D.W., 2016. Frontal to profile face verification in the wild, in: Applications of Computer Vision, Winter Conf. on, IEEE. pp. 1–9.

Seo, J.J., Kim, H.I., Ro, Y.M., 2015. Pose-robust and discriminative feature representation by multi-task deep learning for multi-view face recognition, in: Multimedia, Intl. Symposium on, IEEE. pp. 166–171.

Sepas-Moghaddam, A., Pereira, F., Correia, P.L., 2019. Face recognition: A novel multi-level taxonomy based survey. arXiv preprint arXiv:1901.00713 .

Shao, M., Ding, Z., Fu, Y., 2015. Sparse low-rank fusion based deep features for missing modality face recognition, in: Automatic Face and Gesture Recognition, Intl. Conf. and Workshops on, IEEE. pp. 1–6.

Sharma, P., Yadav, R., Arya, K., 2016. Face recognition from video using generalized mean deep learning neural network, in: Computational and Business Intelligence, Intl. Symposium on, IEEE. pp. 195–199.

Shi, Y., Jain, A.K., 2019. Docface: Matching id document photos to selfies, in: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS), IEEE. pp. 1–8.

Shrivastava, A., Gupta, A., Girshick, R., 2016. Training region-based object detectors with online hard example mining, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 761–769.

Sim, T., Baker, S., Bsat, M., 2002. The cmu pose, illumination, and expression (pie) database, in: Automatic Face and Gesture Recognition, Proceedings. Intl. Conf. on, IEEE. pp. 53–58.

Simón, M.O., Corneanu, C., Nasrollahi, K., Nikisins, O., Escalera, S., Sun, Y., Li, H., Sun, Z., Moeslund, T.B., Greitans, M., 2016. Improved rgb-dt based face recognition. Iet Biometrics 5, 297–303.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Singh, M., Nagpal, S., Gupta, N., Gupta, S., Ghosh, S., Singh, R., Vatsa, M., 2016. Cross-spectral cross-resolution video database for face recognition, in: Biometrics Theory, Applications and Systems, Int'l Conf. on, IEEE. pp. 1–7.

Singh, M., Nagpal, S., Singh, R., Vatsa, M., 2014. On recognizing face images with weight and age variations. IEEE Access 2, 822–830.

Smirnov, E., Melnikov, A., Novoselov, S., Luckyanets, E., Lavrentyeva, G., 2017. Doppelganger mining for face representation learning, in: Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1916–1923.

Sohn, K., 2016. Improved deep metric learning with multi-class n-pair loss objective, in: Advances in Neural Information Processing Systems, pp. 1857–1865.

Sohn, K., Liu, S., Zhong, G., Yu, X., Yang, M.H., Chandraker, M., 2017. Unsupervised domain adaptation for face recognition in unlabeled videos. arXiv preprint arXiv:1708.02191 .

Song, L., Zhang, M., Wu, X., He, R., 2017. Adversarial discriminative heterogeneous face recognition. arXiv preprint arXiv:1709.03675 .

Song, L., Zhang, M., Wu, X., He, R., 2018. Adversarial discriminative heterogeneous face recognition, in: Thirty-Second AAAI Conference on Artificial Intelligence.

Srivastava, N., Salakhutdinov, R.R., 2012. Multimodal learning with deep boltzmann machines, in: Advances in neural information processing systems, pp. 2222–2230.

Stephen, B., 2015. Deep learning and face recognition: The state of the art, in: Proc. SPIE.

Sun, H., Zhen, X., Zheng, Y., Yang, G., Yin, Y., Li, S., 2017. Learning deep match kernels for image-set classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3307–3316.

Sun, Y., Chen, Y., Wang, X., Tang, X., 2014a. Deep learning face representation by joint identification-verification, in: Advances in neural information processing systems, pp. 1988–1996.

Sun, Y., Liang, D., Wang, X., Tang, X., 2015a. Deepid3: Face recognition with very deep neural networks. arXiv preprint arXiv:1502.00873 .

Sun, Y., Wang, X., Tang, X., 2013. Hybrid deep learning for face verification, in: Proceedings of the IEEE Intl. Conf. on Computer Vision, pp. 1489–1496.

Sun, Y., Wang, X., Tang, X., 2014b. Deep learning face representation from predicting 10,000 classes, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1891–1898.

Sun, Y., Wang, X., Tang, X., 2015b. Deeply learned face representations are sparse, selective, and robust, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 2892–2900.

Sun, Y., Wang, X., Tang, X., 2016. Sparsifying neural network connections for face recognition, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 4856–4864.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, in: Proceedings of the IEEE Conf. on computer vision and pattern recognition, pp. 1–9.

Tadmor, O., Wexler, Y., Rosenwein, T., Shalev-Shwartz, S., Shashua, A., 2016. Learning a metric embedding for face recognition using the multibatch method. arXiv preprint arXiv:1605.07270 .

Taigman, Y., Yang, M., Ranzato, M., Wolf, L., 2014. Closing the gap to human-level performance in face verification. deepface, in: IEEE Computer Vision and Pattern Recognition (CVPR).

Taigman, Y., Yang, M., Ranzato, M., Wolf, L., 2015. Web-scale training for face identification, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 2746–2754.

Thakare, N.M., Thakare, V., 2011. An innovative hybrid approach to construct fuzzy-neural network for 3d face recognition system, in: Hybrid Intelligent Systems, Intl. Conf. on, IEEE. pp. 463–467.

Tian, L., Fan, C., Ming, Y., 2016. Multiple scales combined principle component analysis deep learning network for face recognition. Journal of Electronic Imaging 25, 023025–023025.

Tian, L., Fan, C., Ming, Y., Jin, Y., 2015a. Stacked pca network (spcanet): an effective deep learning for face recognition, in: Digital Signal Processing, Intl. Conf. on, IEEE. pp. 1039–1043.

Tian, L., Fan, C., Ming, Y., Shi, J., 2015b. Srdanet: an efficient deep learning algorithm for face analysis, in: Intl. Conf. on Intelligent Robotics and Applications, Springer. pp. 499–510.

Toderici, G., Evangelopoulos, G., Fang, T., Theoharis, T., Kakadiaris, I.A., 2013. Uhdb11 database for 3d-2d face recognition, in: Pacific-Rim Symposium on Image and Video Technology, Springer. pp. 73–86.

Tran, L., Yin, X., Liu, X., 2017. Disentangled representation learning gan for pose-invariant face recognition, in: CVPR, p. 7.

Trigeorgis, G., Snape, P., Nicolaou, M.A., Antonakos, E., Zafeiriou, S., 2016. Mnemonic descent method: A recurrent process applied for end-to-end face alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4177–4187.

Trigueros, D.S., Meng, L., Hartnett, M., 2017. Enhancing convolutional neural networks for face recognition with occlusion maps and batch triplet loss. arXiv preprint arXiv:1707.07923 .

Turk, M., Pentland, A., 1991. Eigenfaces for recognition. Journal of cognitive neuroscience 3, 71–86.

Tzimiropoulos, G., Pantic, M., 2013. Optimization problems for fast aam fitting in-the-wild, in: Proceedings of the IEEE international conference on computer vision, pp. 593–600.

UTK, 2012. Eee otcbvs ws series bench, in: DOE University Research Program in Robotics under grant DOE-DEFG02-86NE37968, http://vciplokstate.org/pbvs/bench/.

Vareto, R., Silva, S., Costa, F., Schwartz, W.R., 2017. Towards open-set face recognition using hashing functions, in: Biometrics (IJCB), 2017 IEEE International Joint Conference on, IEEE. pp. 634–641.

Vijayan, V., Bowyer, K.W., Flynn, P.J., Huang, D., Chen, L., Hansen, M., Ocegueda, O., Shah, S.K., Kakadiaris, I.A., 2011. Twins 3d face recognition challenge, in: 2011 International Joint Conference on Biometrics (IJCB), IEEE. pp. 1–7.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A., 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research 11, 3371–3408.

Wan, W., Zhong, Y., Li, T., Chen, J., 2018. Rethinking feature distribution for loss functions in image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9117–9126.

Wang, D., Otto, C., Jain, A.K., 2016. Face search at scale. IEEE transactions on pattern analysis and machine intelligence 39, 1122–1136.

Wang, F., Chen, L., Li, C., Huang, S., Chen, Y., Qian, C., Change Loy, C., 2018a. The devil of face recognition is in the noise, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 765–780.

Wang, F., Cheng, J., Liu, W., Liu, H., 2018b. Additive margin softmax for face

verification, in: IEEE Signal Processing Letters, pp. 926–930.
Wang, F., Xiang, X., Cheng, J., Yuille, A.L., 2017a. Normface: l_2 hypersphere embedding for face verification. arXiv preprint arXiv:1704.06369 .
Wang, G., Sun, Y., Geng, K., Li, S., Chen, W., 2017b. Deep embedding for face recognition in public video surveillance, in: Chinese Conf. on Biometric Recognition, Springer. pp. 31–39.
Wang, H., Wang, Y., Zhou, Z., Ji, X., Li, Z., Gong, D., Zhou, J., Liu, W., 2018c. Cosface: Large margin cosine loss for deep face recognition. arXiv preprint arXiv:1801.09414 .
Wang, S., Liu, Z., Lv, S., Lv, Y., Wu, G., Peng, P., Chen, F., Wang, X., 2010. A natural visible and infrared facial expression database for expression recognition and emotion inference. IEEE Transactions on Multimedia 12, 682–691.
Wang, T., Shi, P., 2009. Kernel grassmannian distances and discriminant analysis for face recognition from image sets. Pattern Recognition Letters 30, 1161–1165.
Wang, W., Wang, R., Shan, S., Chen, X., 2017c. Discriminative covariance oriented representation learning for face recognition with image sets, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 5599–5608.
Wang, W., Yang, J., Xiao, J., Li, S., Zhou, D., 2014. Face recognition based on deep learning, in: Intl. Conf. on Human Centered Computing, Springer. pp. 812–820.
Wang, X., Tang, X., 2009. Face photo-sketch synthesis and recognition. IEEE trans. on Pattern Analysis and Machine Intelligence 31, 1955–1967.
Wang, X., Zhou, Y., Kong, D., Currey, J., Li, D., Zhou, J., 2017d. Unleash the black magic in age: a multi-task deep neural network approach for cross-age face verification, in: Automatic Face & Gesture Recognition, Intl. Conf. on, IEEE. pp. 596–603.
Wang, Y., Bao, T., Ding, C., Zhu, M., 2017e. Face recognition in real-world surveillance videos with deep learning method, in: Image, Vision and Computing, Intl. Conf. on, IEEE. pp. 239–243.
Wang, Y., Gong, D., Zhou, Z., Ji, X., Wang, H., Li, Z., Liu, W., Zhang, T., 2018d. Orthogonal deep features decomposition for age-invariant face recognition, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 738–753.
Wang, Y.X., Ramanan, D., Hebert, M., 2017f. Learning to model the tail, in: Advances in Neural Information Processing Systems, pp. 7029–7039.
Wen, Y., Li, Z., Qiao, Y., 2016a. Latent factor guided convolutional neural networks for age-invariant face recognition, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 4893–4901.
Wen, Y., Zhang, K., Li, Z., Qiao, Y., 2016b. A discriminative feature learning approach for deep face recognition, in: European Conf. on Computer Vision, Springer. pp. 499–515.
Whitelam, C., Taborsky, E., Blanton, A., Maze, B., Adams, J.C., Miller, T., Kalka, N.D., Jain, A.K., Duncan, J.A., Allen, K., et al., 2017. Iarpa janus benchmark-b face dataset., in: CVPR Workshops, pp. 592–600.
Wolf, L., Hassner, T., Maoz, I., 2011. Face recognition in unconstrained videos with matched background similarity, in: Computer Vision and Pattern Recognition, Conf. on, IEEE. pp. 529–534.
Wong, Y., Chen, S., Mau, S., Sanderson, C., Lovell, B.C., 2011. Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition, in: Computer Vision and Pattern Recognition Workshops, Computer Society Conf. on, IEEE. pp. 74–81.
Wu, W., Kan, M., Liu, X., Yang, Y., Shan, S., Chen, X., 2017a. Recursive spatial transformer (rest) for alignment-free face recognition, in: Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition, pp. 3772–3780.
Wu, X., 2015. Learning robust deep face representation. arXiv preprint arXiv:1507.04844 .
Wu, X., He, R., Sun, Z., Tan, T., 2015. A light cnn for deep face representation with noisy labels. arXiv preprint arXiv:1511.02683 .
Wu, X., Huang, H., Patel, V.M., He, R., Sun, Z., 2018a. Disentangled variational representation for heterogeneous face recognition. arXiv preprint arXiv:1809.01936 .
Wu, X., Song, L., He, R., Tan, T., 2017b. Coupled deep learning for heterogeneous face recognition. arXiv preprint arXiv:1704.02450 .
Wu, X., Song, L., He, R., Tan, T., 2018b. Coupled deep learning for heterogeneous face recognition, in: Thirty-Second AAAI Conference on Artificial Intelligence.
Wu, Y., Shah, S.K., Kakadiaris, I.A., 2016. Rendering or normalization? an analysis of the 3d-aided pose-invariant face recognition, in: Identity, Security and Behavior Analysis, Intl. Conf. on, IEEE. pp. 1–8.
Wu, Y., Wang, Z., Ji, Q., 2013. Facial feature tracking under varying facial expressions and face poses based on restricted boltzmann machines, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 3452–3459.
Wu, Z., Deng, W., 2016. One-shot deep neural network for pose and illumination normalization face recognition, in: Multimedia and Expo, Intl. Conf. on, IEEE. pp. 1–6.
Xi, M., Chen, L., Polajnar, D., Tong, W., 2016. Local binary pattern network: a deep learning approach for face recognition, in: Image Processing, Intl. Conf. on, IEEE. pp. 3224–3228.
Xie, S., Tu, Z., 2015. Holistically-nested edge detection, in: Proceedings of the IEEE international conference on computer vision, pp. 1395–1403.
Xiong, C., Zhao, X., Tang, D., Jayashree, K., Yan, S., Kim, T.K., 2015. Conditional convolutional neural network for modality-aware face recognition, in: Proceedings of the IEEE Intl. Conf. on Computer Vision, pp. 3667–3675.
Xiong, L., Karlekar, J., Zhao, J., Feng, J., Pranata, S., Shen, S., 2017. A good practice towards top performance of face recognition: Transferred deep feature fusion. arXiv preprint arXiv:1704.00438 .
Xiong, X., De la Torre, F., 2013. Supervised descent method and its applications to face alignment, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 532–539.
Xu, B., Wang, N., Chen, T., Li, M., 2015. Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853 .
Xu, C., Liu, Q., Ye, M., 2017a. Age invariant face recognition and retrieval by coupled auto-encoder networks. Neurocomputing 222, 62–71.
Xu, C., Tan, T., Li, S., Wang, Y., Zhong, C., 2006. Learning effective intrinsic features to boost 3d-based face recognition, in: European Conference on Computer Vision, Springer. pp. 416–427.
Xu, X., Le, H.A., Dou, P., Wu, Y., Kakadiaris, I.A., 2017b. Evaluation of a 3d-aided pose invariant 2d face recognition system, in: Biometrics (IJCB), 2017 IEEE International Joint Conference on, IEEE. pp. 446–455.
Yalcin, M., Cevikalp, H., Serhan Yavuz, H., 2015. Towards large-scale face recognition based on videos, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 26–33.
Yang, H., Patras, I., 2015. Mirror, mirror on the wall, tell me, is the error small?, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4685–4693.
Yang, J., Ren, P., Zhang, D., Chen, D., Wen, F., Li, H., Hua, G., 2017a. Neural aggregation network for video face recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4362–4371.
Yang, M., Zhang, L., Yang, J., Zhang, D., 2010. Metaface learning for sparse representation based face recognition, in: Image Processing, Intl. Conf. on, IEEE. pp. 1601–1604.
Yang, Z., Jian, M., Bao, B., Wu, L., 2017b. Max-feature-map based light convolutional embedding networks for face verification, in: Chinese Conf. on Biometric Recognition, Springer. pp. 58–65.
Yanning, Z., Zhe, G., Zenggang, L., Hongxia, Z., Chao, Z., 2012. The npu multi-case chinese 3d face database and information processing. Chinese Journal of Electronics 21, 283–286.
Yeung, H.W.F., Li, J., Chung, Y.Y., 2017. Improved performance of face recognition using cnn with constrained triplet loss layer, in: Neural Networks, Intl. Joint Conf. on, IEEE. pp. 1948–1955.
Yi, D., Lei, Z., Li, S.Z., 2015. Shared representation learning for heterogenous face recognition, in: Automatic Face and Gesture Recognition, Intl. Conf. and Workshops on, IEEE. pp. 1–7.
Yi, D., Lei, Z., Liao, S., Li, S.Z., 2014. Learning face representation from scratch. arXiv preprint arXiv:1411.7923 .
Yim, J., Jung, H., Yoo, B., Choi, C., Park, D., Kim, J., 2015. Rotating your face using multi-task deep neural network, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 676–684.
Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J., 2006. A 3d facial expression database for facial behavior research, in: 7th international conference on automatic face and gesture recognition (FGR06), IEEE. pp. 211–216.
Yin, X., Liu, X., 2017. Multi-task convolutional neural network for pose-invariant face recognition. IEEE Transactions on Image Processing 27, 964–975.
Yin, X., Liu, X., 2018. Multi-task convolutional neural network for pose-invariant face recognition. IEEE Transactions on Image Processing 27, 964–975.
Yin, X., Yu, X., Sohn, K., Liu, X., Chandraker, M., 2017. Towards large-pose face frontalization in the wild, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 3990–3999.

Yin, X., Yu, X., Sohn, K., Liu, X., Chandraker, M., 2018. Feature transfer learning for deep face recognition with long-tail data. arXiv preprint arXiv:1803.09014 .

Yue, L., Miao, X., Wang, P., Zhang, B., Zhen, X., Cao, X., 2018. Attentional alignment network, in: BMVC, p. 7.

Zafar, U., Ghafoor, M., Zia, T., Ahmed, G., Latif, A., Malik, K.R., Sharif, A.M., 2019. Face recognition with bayesian convolutional networks for robust surveillance systems. EURASIP Journal on Image and Video Processing 2019, 10.

Zhang, B., Zhang, L., Zhang, D., Shen, L., 2010. Directional binary code with application to polyu near-infrared face database. Pattern Recognition Letters 31, 2337–2344.

Zhang, H., Patel, V.M., Riggan, B.S., Hu, S., 2017a. Generative adversarial network-based synthesis of visible faces from polarimetrie thermal faces, in: 2017 IEEE International Joint Conference on Biometrics (IJCB), IEEE. pp. 100–107.

Zhang, J., Huang, D., Wang, Y., Sun, J., 2016a. Lock3dface: A large-scale database of low-cost kinect 3d faces, in: 2016 International Conference on Biometrics (ICB), IEEE. pp. 1–8.

Zhang, K., Zhang, Z., Li, Z., Qiao, Y., 2016b. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters 23, 1499–1503.

Zhang, N., Shelhamer, E., Gao, Y., Darrell, T., 2015a. Fine-grained pose prediction, normalization, and recognition. arXiv preprint arXiv:1511.07063 .

Zhang, T., Wiliem, A., Yang, S., Lovell, B., 2018. Tv-gan: Generative adversarial network based thermal to visible face recognition, in: 2018 International Conference on Biometrics (ICB), IEEE. pp. 174–181.

Zhang, W., Shu, Z., Samaras, D., Chen, L., 2017b. Improving heterogeneous face recognition with conditional adversarial networks. arXiv preprint arXiv:1709.02848 .

Zhang, W., Wang, X., Tang, X., 2011. Coupled information-theoretic encoding for face photo-sketch recognition, in: Computer Vision and Pattern Recognition, Conf. on, IEEE. pp. 513–520.

Zhang, X., Fang, Z., Wen, Y., Li, Z., Qiao, Y., 2017c. Range loss for deep face recognition with long-tailed training data, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 5409–5418.

Zhang, X., Zhang, L., Wang, X.J., Shum, H.Y., 2012. Finding celebrities in billions of web images. IEEE trans. on Multimedia 14, 995–1007.

Zhang, Y., Shao, M., Wong, E.K., Fu, Y., 2013. Random faces guided sparse many-to-one encoder for pose-invariant face recognition, in: Proceedings of the IEEE Intl. Conf. on Computer Vision, pp. 2416–2423.

Zhang, Z., Luo, P., Loy, C.C., Tang, X., 2015b. Learning deep representation for face alignment with auxiliary attributes. IEEE transactions on pattern analysis and machine intelligence 38, 918–930.

Zhao, J., Cheng, Y., Cheng, Y., Yang, Y., Lan, H., Zhao, F., Xiong, L., Xu, Y., Li, J., Pranata, S., et al., 2018a. Look across elapse: Disentangled representation learning and photorealistic cross-age face synthesis for age-invariant face recognition. arXiv preprint arXiv:1809.00338 .

Zhao, J., Cheng, Y., Xu, Y., Xiong, L., Li, J., Zhao, F., Jayashree, K., Pranata, S., Shen, S., Xing, J., et al., 2018b. Towards pose invariant face recognition in the wild, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2207–2216.

Zhao, J., Xiong, L., Cheng, Y., Cheng, Y., Li, J., Zhou, L., Xu, Y., Karlekar, J., Pranata, S., Shen, S., et al., 2018c. 3d-aided deep pose-invariant face recognition., in: IJCAI, p. 11.

Zhao, J., Xiong, L., Jayashree, P.K., Li, J., Zhao, F., Wang, Z., Pranata, P.S., Shen, P.S., Yan, S., Feng, J., 2017. Dual-agent gans for photorealistic and identity preserving profile face synthesis, in: Advances in Neural Information Processing Systems, pp. 66–76.

Zheng, J., Chen, J.C., Bodla, N., Patel, V.M., Chellappa, R., 2016. Vlad encoded deep convolutional features for unconstrained face verification, in: Pattern Recognition, Intl. Conf. on, IEEE. pp. 4101–4106.

Zheng, T., Deng, W., 2018. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments, in: Beijing University of Posts and Telecommunications, Tech. Rep (2018): 18-01.

Zheng, T., Deng, W., Hu, J., 2017a. Age estimation guided convolutional neural network for age-invariant face recognition, in: Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition Workshops, pp. 1–9.

Zheng, T., Deng, W., Hu, J., 2017b. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments, in: arXiv preprint arXiv:1708.08197.

Zheng, Y., Pal, D.K., Savvides, M., 2018. Ring loss: Convex feature normalization for face recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5089–5097.

Zheng, Z., Guo, G., 2016. A joint optimization scheme to combine different levels of features for face recognition with makeup changes, in: Image Processing, Intl. Conf. on, IEEE. pp. 3001–3005.

Zhou, E., Cao, Z., Yin, Q., 2015. Naive-deep face recognition: Touching the limit of lfw benchmark or not? arXiv preprint arXiv:1501.04690 .

Zhou, H., Mian, A., Wei, L., Creighton, D., Hossny, M., Nahavandi, S., 2014. Recent advances on singlemodal and multimodal face recognition: a survey. IEEE trans. on Human-Machine Systems 44, 701–716.

Zhu, S., Li, C., Change Loy, C., Tang, X., 2015a. Face alignment by coarse-to-fine shape searching, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4998–5006.

Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z., 2016. Face alignment across large poses: A 3d solution, in: Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition, pp. 146–155.

Zhu, Y., Guo, G., 2016. Exploring deep features with different distance measures for still to video face matching, in: Chinese Conf. on Biometric Recognition, Springer. pp. 158–166.

Zhu, Y., Zheng, Z., Li, Y., Mu, G., Shan, S., Guo, G., 2015b. Still to video face recognition using a heterogeneous matching approach, in: Biometrics Theory, Applications and Systems, IEEE Intl. Conf. on, pp. 1–6.

Zhu, Z., Luo, P., Wang, X., Tang, X., 2013. Deep learning identity-preserving face space, in: Proceedings of the IEEE Intl. Conf. on Computer Vision, pp. 113–120.

Zhu, Z., Luo, P., Wang, X., Tang, X., 2014a. Multi-view perceptron: a deep model for learning face identity and view representations, in: Advances in Neural Information Processing Systems, pp. 217–225.

Zhu, Z., Luo, P., Wang, X., Tang, X., 2014b. Recover canonical-view faces in the wild with deep neural networks. arXiv preprint arXiv:1404.3543 .

Zou, W., Zhu, S., Yu, K., Ng, A.Y., 2012. Deep learning of invariant features via simulated fixations in video, in: Advances in neural information processing systems, pp. 3203–3211.

Zulqarnain Gilani, S., Mian, A., 2018. Learning from millions of 3d scans for large-scale 3d face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1896–1905.