Data Augmentation-Based Joint Learning for Heterogeneous Face Recognition

Bing Cao, Nannan Wang, *Member, IEEE*, Jie Li, and Xinbo Gao, *Senior Member, IEEE*

***Abstract*— Heterogeneous face recognition (HFR) is the process of matching face images captured from different sources. HFR plays an important role in security scenarios. However, HFR remains a challenging problem due to the considerable discrepancies (i.e., shape, style, and color) between cross-modality images. Conventional HFR methods utilize only the information involved in heterogeneous face images, which is not effective because of the substantial differences between heterogeneous face images. To better address this issue, this paper proposes a data augmentation-based joint learning (DA-JL) approach. The proposed method mutually transforms the cross-modality differences by incorporating synthesized images into the learning process. The aggregated data augments the intraclass scale, which provides more discriminative information. However, this method also reduces the interclass diversity (i.e., discriminative information). We develop the DA-JL model to balance this dilemma. Finally, we obtain the similarity score between heterogeneous face image pairs through the log-likelihood ratio. Extensive experiments on a viewed sketch database, forensic sketch database, near-infrared image database, thermal-infrared image database, low-resolution photo database, and image with occlusion database illustrate that the proposed method achieves superior performance in comparison with the state-of-the-art methods.**

***Index Terms*— Data augmentation, forensic sketch, heterogeneous face recognition (HFR), infrared image, joint learning, viewed sketch.**

Manuscript received December 29, 2017; revised June 7, 2018 and September 21, 2018; accepted September 25, 2018. Date of publication October 25, 2018; date of current version May 23, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61432014, Grant 61876142, Grant U1605252, Grant 61772402, Grant 61671339, and Grant 61501339, in part by the National Key Research and Development Program of China under Grant 2016QY01W0200, in part by the Key Industrial Innovation Chain in Industrial Domain under Grant 2016KTZDGY04-02, in part by the National High-Level Talents Special Support Program of China under Grant CS31117200001, in part by the Young Elite Scientists Sponsorship Program by CAST under Grant 2016QNRC001, in part by the Natural Science Basic Research Plan in the Shaanxi Province of China under Grant 2017JM6085 and Grant 2017JQ6007, in part by the Young Talent Fund of the University Association for Science and Technology in Shaanxi, China, in part by the Fundamental Research Funds for the Central Universities under Grant XJS17086, in part by the Innovation Fund of Xidian University, in part by CCF-Tencent Open Fund, and in part by the Xidian University-Intellifusion Joint Innovation Laboratory of Artificial Intelligence. *(Corresponding author: Xinbo Gao.)*

B. Cao, J. Li, and X. Gao are with the State Key Laboratory of Integrated Services Networks, School of Electronic Engineering, Xidian University, Xi'an 710071, China (e-mail: bcaoxidian@gmail.com; leejie@mail.xidian.edu.cn; xbgao@mail.xidian.edu.cn).

N. Wang is with the State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, Xi'an 710071, China (e-mail: nnwang@xidian.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2018.2872675

I. Introduction

FACE images acquired from different sources, such as general cameras [visual (VIS) photos], near-infrared (NIR) cameras (NIR images), thermal-infrared (TIR) cameras (TIR images), and sketch artists (sketch images), are referred to as coming from different modalities. The process of matching face images from different modalities is called heterogeneous face recognition (HFR), which is an important issue in security scenarios. For instance, facial sketches are widely used by law enforcement agencies to assist in the identification and apprehension of suspects involved in criminal activities [1], [2]. Since NIR cameras provide an efficient and straightforward solution to improve face recognition performance in extreme lighting conditions [3], they are widely used to handle complicated illumination conditions in video surveillance and other security areas [4].

Conventional face recognition methods achieve poor performance when identifying a face sketch, an NIR face image, or a TIR face image from VIS face photos directly due to considerable appearance variations among heterogeneous face images. Existing HFR methods can be roughly grouped into three categories: invariant feature extraction-based methods, common space learning-based methods, and synthesis-based methods.

Invariant feature extraction-based methods [5]–[12] extract invariant features to represent heterogeneous face images to reduce the large variations among heterogeneous face images at the feature level. However, because of the high computational complexity and limited discriminability, these encoded feature descriptors require substantial time and perform poorly in recognition tasks. Common space learning-based methods [2], [13]–[19] project face images from different modalities into a common space to minimize the discrepancies. Heterogeneous face images can be matched directly within this subspace, but discriminative information is inevitably lost during the projection procedure, which decreases recognition performance. Synthesis-based methods [20]–[25] train a set of reconstruction coefficients to transform heterogeneous face images into homogeneous face images to reduce the discrepancies between heterogeneous face images at the image level. These homogeneous face images can be directly applied to conventional face recognition approaches. However, due to the considerable differences among heterogeneous face images, it is difficult for synthesis-based methods to reduce differences in shape (e.g., spectacle-frame, double-fold eyelid, and facial outline), which results in poor recognition performance in HFR scenarios.

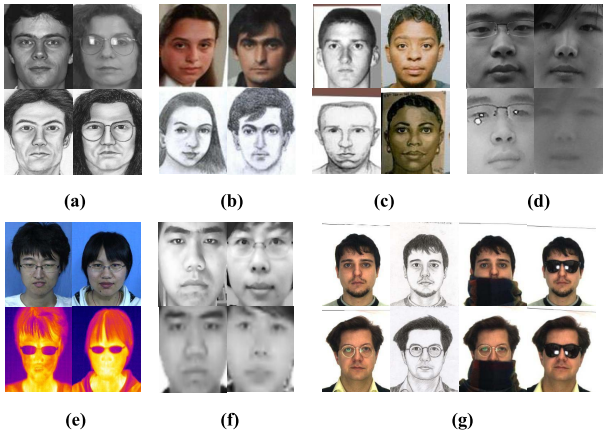

Fig. 1. Some sample images from heterogeneous face databases. (a) CUFSF database. (b) IIIT-D database. (c) Forensic Sketch database. (d) CUHK VIS-NIR database. (e) USTC-NVIE database. (f) NJU-ID database. (g) AR Face database and AR Sketch database.

An effective joint formulation method [26] for conventional face recognition, which enhanced the original Bayesian face model and achieved a promising result, was recently proposed. This method considers intrapersonal variations and interpersonal invariants over image pairs and requires a sufficient number of intraclass face images. However, most heterogeneous face databases provide a few intraclass face images, and the differences between heterogeneous face images are cross-modality. Therefore, the performance of this method on heterogeneous face databases is poor.

This paper proposes a novel data augmentation-based joint learning (DA-JL) approach for HFR. Each identity contains only a pair of heterogeneous face images, which are augmented by synthesis models. The proposed method considers the latent information from synthesized face images. The latent information refers to the shape and textures of the synthesized face images. Note that the synthesized images and the corresponding original face images together provide intraclass information of the same shape in different modalities. When extracting intraclass information, we add the synthesized face images to the training set. Therefore, we can acquire more effective intraclass information from the enlarged training set. However, this addition introduces redundant variables into the interclass information, which influences the recognition performance. To balance this dilemma, we design a DA-JL model to extract more effective intraclass information without losing interclass information in the training phase. Two covariance matrices that represent two types of information are jointly optimized with the original images and synthesized images. Finally, the log-likelihood ratio statistic is calculated as the similarity score of two input heterogeneous face images. We evaluate our method on six databases: the CUHK Face Sketch FERET (CUFSF) database [7], Forensic Sketch database [4], CUHK VIS-NIR database [12], Natural Visible and Infrared Facial Expression (USTC-NVIE) database [27], Nanjing University ID Card Face (NJU-ID) data set [28], and AR Face database [29]. Fig. 1 shows some samples images from these databases.

The contributions of this paper are summarized as follows.

1) We first utilize the latent information in synthesized images to extract more effective intraclass information and design a valid strategy to obtain more information from the limited databases.
2) A DA-JL model is developed to jointly optimize the intraclass and interclass information without losing effective interclass information.
3) AlexNet, ResNet, DenseNet, and LightCNN are investigated for image representation. We find that LightCNN achieves the best performance. Therefore, we modify our previous model by substituting VGG with LightCNN for feature representation and achieve superior performance in six HFR scenarios.

Preliminary work has been published in [30]. Compared with the preliminary version, this paper makes four major extensions and improvements. First, three new HFR scenarios that contain substantial discrepancies are introduced to verify the validity of the proposed approach. Second, we improve the early version of AJL-HFR [30] and achieve better performance than the state-of-the-art methods. Third, comparison experiments are conducted on multiple databases. Finally, diagrams and detailed explanations are provided to illustrate the proposed method. The rest of this paper is organized as follows. In Section II, representative HFR methods are briefly reviewed. In Section III, the proposed DA-JL approach is presented in detail. We provide experimental results and analysis in Section IV, and summarize this paper in Section V.

II. Related Work

In this section, we briefly review three categories of representative HFR methods: invariant feature-based methods, common space learning-based methods, and synthesis-based methods.

Invariant feature extraction-based methods focus on extracting the invariant features from face images in different modalities. Klare *et al.* [6] used scale-invariant feature transform (SIFT) [31] and multiscale local binary pattern [32] to represent face images in different modalities and proposed a local feature-based discriminant analysis (LFDA) framework based on [33] for matching heterogeneous face images. Zhang *et al.* [7] designed a coupled information-theoretic encoding (CITE) feature descriptor to reduce the modality discrepancy between sketches and photos. Galoogahi and Sim [8] proposed a local radon binary pattern (LRBP) face descriptor to project face images into radon space and utilized local binary patterns for encoding. A local difference of Gaussian binary pattern (LDoGBP) was proposed to recognize cross-modality face images in [9]. Shao and Fu [11] utilized a two-step cross-modality learning method to extract the invariant features from heterogeneous face images. Recently, a general encoding feature discriminant approach was proposed in [12] to extract common features from an encoding space projected by heterogeneous face images. In summary, feature-based methods are aimed at minimizing the modality discrepancy for heterogeneous face images. However, due to the high computational complexity and limited discriminability, the accuracies of these methods require improvement.

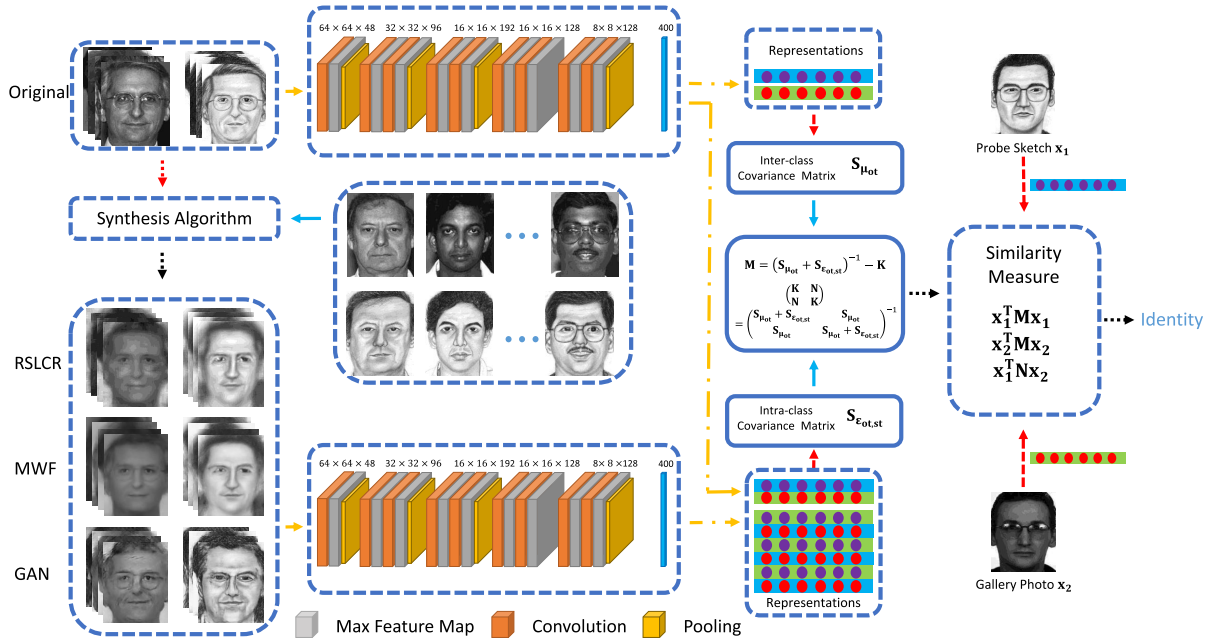

Fig. 2. Framework of the proposed DA-JL method for HFR.

Common space learning-based methods aim to minimize the intermodality discrepancy by projecting heterogeneous face features into a common subspace. A general discriminant feature extraction method was first proposed in [13] to transform the features in different modalities to a common space. Yi *et al.* [14] proposed the canonical correlation analysis approach to match NIR with VIS face images, and the approach was extended in [15]. Lei *et al.* [16] applied a coupled spectral regression method [34] to project cross-modality face images into a common space. A learning-based model (LCKS-CSR) was designed by Lei *et al.* [35] to enhance the discriminative ability of feature descriptors to improve recognition performance. Sharma and Jacobs [17] proposed a partial least squares (PLS) approach to obtain a common linear subspace. A novel HFR framework based on a nonlinear kernel was applied by Klare and Jain [2] to represent heterogeneous face images. Kan *et al.* [18] utilized the relationship of cross-modality face images to develop a multiview discriminant analysis (MvDA) approach. However, discriminative information is inevitably lost in the projection procedure, which leads to unsatisfactory performance in HFR.

Synthesis-based methods train a set of reconstruction coefficients from homogeneous images to transform heterogeneous face images to homogeneous face images. Considering the local structural information, the synthesized images of local linear embedding (LLE) model [20] provide local structural information but loses many details. Wang and Tang [21] employed the Markov random field (MRF) model to preserve the local information, but artifacts and face deformations appeared in the synthesized results. To overcome the drawbacks of insufficient candidate patches for MRF model-based methods, the Markov weight field (MWF) model [22] was proposed to decrease artifacts and face deformations. The sparse feature selection (SFS) method [24] filters high-frequency information caused by the linear combination of image patches. They further compensate for the lost high-frequency information by sparse-representation-based enhancement (SRE) [23] and support vector regression (SVR) [24]. However, blurring effects still exist in the synthesized images. Wang *et al.* [36] developed a transductive approach to transform face photos and sketches for HFR. Peng *et al.* [37] utilized multiple representations to improve the robustness of a synthesis model for different illuminations. Wang *et al.* [25] employed offline random sampling and locality constraint (RSLCR) to accelerate the face sketch synthesis process. Inspired by [38], we developed generative adversarial networks (GANs), which retain the local and detailed information, to synthesize face images. However, the shape of the synthesized images was not sufficiently clear. Although the modality discrepancies are reduced by synthesis-based models, the synthesis procedure consumes considerable time, which slows HFR. In addition, image synthesis is a difficult problem. Therefore, the performance of synthesis-based methods for HFR is unsatisfactory. *However, the synthesized face sketches reflect different information in photos, which is the basic motivation of our proposed approach.*

III. Data Augmentation-Based Joint Learning

In this section, we introduce the proposed DA-JL framework for HFR in detail, as shown in Fig. 2. Without loss of generality and for ease of representation, we describe our approach in the face sketch-photo recognition scenario, which can be generalized to other HFR scenarios. First, we introduce our motivation. Then, we demonstrate how to derive and optimize the proposed model.

A. Motivation

In Fig. 3, the first column is the ground-truth photo-sketch pair, followed by the synthesized photo-sketch pairs

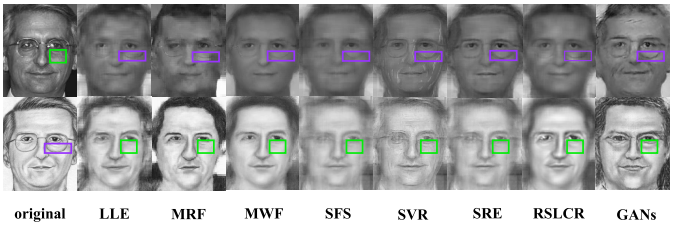

Fig. 3. Original photo-sketch pairs and the corresponding synthesized photo-sketch pairs. The names of the synthesis methods are presented under the image pairs.

generated by LLE [20], MRF [21], MWF [22], SFS [24], SVR [24], SRE [23], RSLCR [36], and GANs [38]. The shape information of the same subject in the photo domain and the sketch domain is different. Thus, the shape information is asymmetric, since we cannot obtain exactly the same shape information of different modalities. Although these discrepancies seriously affect the final recognition performance, we find some interesting phenomenon from these differences.

Given an input sketch as the test image, although we utilize the linear combination of training photo patches to synthesize a photo [39], the shapes of the synthesized photos are more similar to the original sketches than to the ground-truth photos and vice versa. Therefore, we utilize all the synthesized sketches and synthesized photos together with the ground-truth sketch-photo pairs to enlarge the training data set and provide more discriminative information for face recognition. However, the quantity of selected synthesis images should not be too large, because these images contain both valid and redundant information. When too many synthesized images are used, the redundant information can outweigh the valid information. In addition, the interclass information is obtained from the intraclass averages of all classes. If synthesized images are added to the training set, the effectiveness of the interclass information could be reduced. To balance this dilemma of valid and redundant information, we propose the DA-JL framework for HFR. The details are introduced in the following.

B. Model Derivation

As shown in Fig. 2, we need to generate pairs of synthesized sketches and synthesized photos to construct the training database to train the DA-JL model. We generated three synthesized sketch-photo pairs via the RSLCR, MWF, and GANs methods. These three methods are chosen from three different categories of face sketch synthesis methods. The rationale behind this selection strategy is presented in Section IV. We extract CNN features from the synthesized image pairs and original image pair to represent each subject. The DA-JL model (the notation **M** and **N** in Fig. 2) is calculated from the interclass covariance matrix and intraclass covariance matrix. In the recognition phase, the similarity between the input query image and the gallery image is calculated as a log-likelihood ratio based on the trained DA-JL model.

Inspired by the metric learning model [26], [40], a face $\mathbf{x}$ can be approximated by interclass variation μ and intraclass variation ε, where μ represents the identity and ε represents the heterogeneous information between cross-modality face images belonging to the same identity. Similar to the preceding models [41], [42], these two components are modeled as independent zero-mean Gaussian distributions

$$\mu \sim \mathcal{N}(0, \mathbf{S}_\mu)$$
$$\varepsilon \sim \mathcal{N}(0, \mathbf{S}_\varepsilon)$$

where $\mathbf{S}_\mu$ and $\mathbf{S}_\varepsilon$ are the two covariance matrices to be trained. A face image can be represented by the two components as follows:

$$\mathbf{x} = \mu + \varepsilon. \tag{1}$$

For two input face images $\mathbf{x}_1$ and $\mathbf{x}_2$, the covariance can be written as

$$\mathbf{cov}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{cov}(\mu_1, \mu_2) + \mathbf{cov}(\varepsilon_1, \varepsilon_2).$$

$\mathbf{H_I}$ denotes that $\mathbf{x}_1$ and $\mathbf{x}_2$ belong to the same identity; otherwise, $\mathbf{H_E}$. Thus, the intraclass joint distribution $\mathbf{P}(\mathbf{x}_1,\mathbf{x}_2|\mathbf{H_I})$ is a Gaussian with zero-mean and covariance matrix

$$\sum_{\mathbf{I}} = \begin{bmatrix} \mathbf{S}_\mu + \mathbf{S}_\varepsilon & \mathbf{S}_\mu \\ \mathbf{S}_\mu & \mathbf{S}_\mu + \mathbf{S}_\varepsilon \end{bmatrix}.$$

The interclass joint distribution $\mathbf{P}(\mathbf{x}_1,\mathbf{x}_2|\mathbf{H_E})$ is a Gaussian with zero-mean and covariance matrix

$$\sum_{\mathbf{E}} = \begin{bmatrix} \mathbf{S}_\mu + \mathbf{S}_\varepsilon & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_\mu + \mathbf{S}_\varepsilon \end{bmatrix}.$$

The covariance matrix $\mathbf{S}_\mu$ is trained from the combination of the average of each class, and the covariance matrix $\mathbf{S}_\varepsilon$ is trained from all the samples of each class. However, for heterogeneous face images, the images belonging to the same subject come from different modalities. Therefore, it is not sufficient to extract intraclass information by modeling only a single pair of sketch-photo images in the training data set, which is achieved by the joint Bayesian method [26]. To obtain more useful intraclass information about heterogeneous face images, we jointly train the model relying on both the original training data set and the synthesized image pairs. Thus, the covariance matrix $\mathbf{S}_\varepsilon$ provides more information about intraclass face images. However, the synthesized image pairs are pseudoheterogeneous face image pairs. They are different from the image pairs in the training set and contribute redundant information. When we add all these synthesized image pairs (e.g., Fig. 3) to the training set and train a joint Bayesian model directly, the covariance matrices $\mathbf{S}_\mu$ and $\mathbf{S}_\varepsilon$ are contaminated by the redundant information, which seriously affects the effectiveness of the joint Bayesian model. To solve this problem and extract more useful intraclass and interclass information, we improve the model by means of a DA-JL model.

The DA-JL model first generates a certain number of sketch-photo pairs. We present a detailed illustration of how to generate synthesized sketches and photos in Section IV. Then, the intraclass and interclass covariance matrices are trained jointly. If face images are represented by **d**-dimensional

features, the intraclass covariance matrix $\mathbf{S}_{\varepsilon_{\text{ot,st}}} \in \mathbb{R}^{d\times d}$ is derived from the original training set (*denoted as* ot) and the corresponding synthesized images (*denoted as* **st**). The interclass covariance matrix $\mathbf{S}_{\mu_{\text{ot}}} \in \mathbb{R}^{d\times d}$ is derived from only the original training set. For two independent covariance matrices, the covariance matrix $\sum_{\mathbf{I}} \in \mathbb{R}^{2d\times 2d}$ of the intraclass joint distribution $\mathbf{P}(\mathbf{x}_1,\mathbf{x}_2|\mathbf{H_I})$ can be written as

$$\sum_{\mathbf{I}} = \begin{bmatrix} \mathbf{S}_{\mu_{\text{ot}}} + \mathbf{S}_{\varepsilon_{\text{ot,st}}} & \mathbf{S}_{\mu_{\text{ot}}} \\ \mathbf{S}_{\mu_{\text{ot}}} & \mathbf{S}_{\mu_{\text{ot}}} + \mathbf{S}_{\varepsilon_{\text{ot,st}}} \end{bmatrix}.$$

The covariance matrix $\sum_{\mathbf{E}} \in \mathbb{R}^{2d\times 2d}$ of the interclass joint distribution $\mathbf{P}(\mathbf{x}_1,\mathbf{x}_2|\mathbf{H_E})$ can be written as

$$\sum_{\mathbf{E}} = \begin{bmatrix} \mathbf{S}_{\mu_{\text{ot}}} + \mathbf{S}_{\varepsilon_{\text{ot,st}}} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{\mu_{\text{ot}}} + \mathbf{S}_{\varepsilon_{\text{ot,st}}} \end{bmatrix}.$$

Finally, by omitting the constant parameter, we can compute the log-likelihood ratio $r(\mathbf{x}_1, \mathbf{x}_2)$ to obtain the similarity of two input cross-modality face images in terms of the intraclass joint distribution and interclass joint distribution as

$$\begin{aligned} r(\mathbf{x}_1, \mathbf{x}_2) &= \log \frac{\mathbf{P}(\mathbf{x}_1,\mathbf{x}_2|\mathbf{H_I})}{\mathbf{P}(\mathbf{x}_1,\mathbf{x}_2|\mathbf{H_E})} \\ &= \mathbf{x}_1^T\mathbf{M}\mathbf{x}_1 + \mathbf{x}_2^T\mathbf{M}\mathbf{x}_2 - 2\mathbf{x}_1^T\mathbf{N}\mathbf{x}_2 \end{aligned} \tag{2}$$

where

$$\mathbf{M} = (\mathbf{S}_{\mu_{\text{ot}}} + \mathbf{S}_{\varepsilon_{\text{ot,st}}})^{-1} - \mathbf{K} \tag{3}$$

and $\mathbf{K} \in \mathbb{R}^{d\times d}$ satisfies

$$\begin{bmatrix} \mathbf{K} & \mathbf{N} \\ \mathbf{N} & \mathbf{K} \end{bmatrix} = \begin{bmatrix} \mathbf{S}_{\mu_{\text{ot}}} + \mathbf{S}_{\varepsilon_{\text{ot,st}}} & \mathbf{S}_{\mu_{\text{ot}}} \\ \mathbf{S}_{\mu_{\text{ot}}} & \mathbf{S}_{\mu_{\text{ot}}} + \mathbf{S}_{\varepsilon_{\text{ot,st}}} \end{bmatrix}^{-1}. \tag{4}$$

Therefore, we transform the face verification issue into estimating the two covariance matrices $\mathbf{S}_{\mu_{\text{ot}}}$ and $\mathbf{S}_{\mu_{\text{ot,st}}}$.

Suppose that there are m_i *i.i.d.* intraclass face images belonging to the ith subject. According to (1), we can represent all the samples of the same subject by

$$\mathbf{x}_i = \mathbf{Q}_i\mathbf{h}_i \tag{5}$$

where

$$\mathbf{Q}_i = \begin{bmatrix} \mathbf{I} & \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{I} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} \end{bmatrix}$$

$$\mathbf{h}_i = [\mu_{i_{\text{ot}}}; \varepsilon_{i_{\text{ot,st}}1}; \varepsilon_{i_{\text{ot,st}}2}; \cdots; \varepsilon_{i_{\text{ot,st}}m_i}]$$

and $\mathbf{I} \in \mathbb{R}^{d\times d}$ is the identity matrix.

Note that our objective function is

$$\max \prod_i \mathbf{P}(\mathbf{x}_i|\mathbf{h}_i).$$

For each identity, the interclass variation $\mu_{i_{\text{ot}}}$ can be derived from $\mathcal{N}(0, \mathbf{S}_{\mu_{\text{ot}}})$. Then, the intraclass variation $[\varepsilon_{i_{\text{ot,st}}1}; \varepsilon_{i_{\text{ot,st}}2}; \cdots; \varepsilon_{i_{\text{ot,st}}m_i}]$ can be derived from $\mathcal{N}(0, \mathbf{S}_{\mu_{\text{ot,st}}})$. Because subjects are independent, the objective function is equivalent to

$$\max \sum_i \log \mathbf{P}(\mathbf{x}_i|\mathbf{S}_{\mu_{\text{ot}}}, \mathbf{S}_{\varepsilon_{\text{ot,st}}}).$$

Considering that the distributions of the interclass and intraclass variations are Gaussian with the covariance matrix

$$\sum_{\mathbf{h}_i} = \begin{bmatrix} \mathbf{S}_{\mu_{\text{ot}}} & & & \\ & \mathbf{S}_{\varepsilon_{\text{ot,st}}} & & \\ & & \mathbf{S}_{\varepsilon_{\text{ot,st}}} & \\ & & & \mathbf{S}_{\varepsilon_{\text{ot,st}}} \end{bmatrix} \tag{6}$$

we can write the likelihood function of subject i as

$$\mathbf{P}(\mathbf{x}_i|\mathbf{S}_{\mu_{\text{ot}}}, \mathbf{S}_{\varepsilon_{\text{ot,st}}}) = \mathcal{N}\left(\mathbf{0}, \sum_{\mathbf{x_i}}\right)$$

where

$$\begin{aligned} \sum_{\mathbf{x}_i} = \mathbf{Q}_i \sum_{\mathbf{h}_i} \mathbf{Q}_i^T &= \begin{bmatrix} \mathbf{I} & \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \\ \vdots & & & & \\ \mathbf{I} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} \end{bmatrix} \\ &\times \begin{bmatrix} \mathbf{S}_{\mu_{\text{ot}}} & & & \\ & \mathbf{S}_{\varepsilon_{\text{ot,st}}} & & \\ & & \mathbf{S}_{\varepsilon_{\text{ot,st}}} & \\ & & & \mathbf{S}_{\varepsilon_{\text{ot,st}}} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{I} & \cdots & \mathbf{I} \\ \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{S}_{\mu_{\text{ot}}} + \mathbf{S}_{\varepsilon_{\text{ot,st}}} & \mathbf{S}_{\mu_{\text{ot}}} & \cdots & \mathbf{S}_{\mu_{\text{ot}}} \\ \mathbf{S}_{\mu_{\text{ot}}} & \mathbf{S}_{\mu_{\text{ot}}} + \mathbf{S}_{\varepsilon_{\text{ot,st}}} & \cdots & \mathbf{S}_{\mu_{\text{ot}}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_{\mu_{\text{ot}}} & \mathbf{S}_{\mu_{\text{ot}}} & \cdots & \mathbf{S}_{\mu_{\text{ot}}} + \mathbf{S}_{\varepsilon_{\text{ot,st}}} \end{bmatrix}. \end{aligned}$$

We develop an Expectation Maximization (EM) algorithm to jointly optimize the estimation of the two covariance matrices in Section III-C to solve this problem.

C. Model Optimization

1) E-Step: The interclass variation and the intraclass variation $\mathbf{h}_i = [\mu_{i_{\text{ot}}}; \varepsilon_{i_{\text{ot,st}}1}; \varepsilon_{i_{\text{ot,st}}2}; \cdots; \varepsilon_{i_{\text{ot,st}}m_i}]$ are chosen as the hidden variables. In the tth iteration, we compute the expectations of the latent variables. According to (5), $\mathbf{P}(\mathbf{h}_i, x_i|\mathbf{S}_{\mu_{\text{ot}}}, \mathbf{S}_{\varepsilon_{\text{ot,st}}})$ can be simplified into $\mathbf{P}(\mathbf{h}_i|\mathbf{S}_{\mu_{\text{ot}}}, \mathbf{S}_{\varepsilon_{\text{ot,st}}})$. Therefore, the expected log-likelihood function can be written as

$$\sum_i \mathbf{E}_{\mathbf{P}\left(\mathbf{h}_i\middle|x_i,\mathbf{S}^{\mathbf{t}}_{\mu_{\text{ot}}},\mathbf{S}^{\mathbf{t}}_{\varepsilon_{\text{ot,st}}}\right)} \log \mathbf{P}\left(\mathbf{h}_i\middle|\mathbf{S}^{\mathbf{t+1}}_{\mu_{\text{ot}}}, \mathbf{S}^{\mathbf{t+1}}_{\varepsilon_{\text{ot,st}}}\right) \tag{7}$$

where $\mathbf{S}^{\mathbf{t}}_{\mu_{\text{ot}}}$ and $\mathbf{S}^{\mathbf{t}}_{\varepsilon_{\text{ot,st}}}$ are known in the tth iteration, and $\mathbf{S}^{\mathbf{t+1}}_{\mu_{\text{ot}}}$ and $\mathbf{S}^{\mathbf{t+1}}_{\varepsilon_{\text{ot,st}}}$ are updated in the M-step. Note that the distribution of the latent variables is Gaussian, and the expected log-likelihood is equivalent to

$$\sum_i \log\left|\sum_{h_i}\right| + \text{tr}\left(\sum_{h_i}^{-1} W(h_i)\right) \tag{8}$$

where

$$W(h_i) = \mathbf{E}_{\mathbf{P}\left(\mathbf{h}_i\middle|x_i,\mathbf{S}^{\mathbf{t}}_{\mu_{\text{ot}}},\mathbf{S}^{\mathbf{t}}_{\varepsilon_{\text{ot,st}}}\right)} \mathbf{E}^T_{\mathbf{P}\left(\mathbf{h}_i\middle|x_i,\mathbf{S}^{\mathbf{t}}_{\mu_{\text{ot}}},\mathbf{S}^{\mathbf{t}}_{\varepsilon_{\text{ot,st}}}\right)}.$$

The expectation of latent variable $\mathbf{h}_i$ can be computed by

$$\mathbf{E}_{\mathbf{P}(\mathbf{h}_i|x_i,\mathbf{S}_{\mu_{\text{ot}}},\mathbf{S}_{\varepsilon_{\text{ot,st}}})} = \sum_{\mathbf{h}_i} \mathbf{Q}_\mathbf{i}^\mathbf{T} \sum_{\mathbf{x}_i}^{-1} \mathbf{x}_i.$$

At the beginning of the E-Step, we asymmetrically initialize $\mathbf{S}_{\mu_{\text{ot}}}$ by the covariance of the mean of each interclass identity from the original training set and initialize $\mathbf{S}_{\varepsilon_{\text{ot,st}}}$ by the covariance of intraclass face images from the original training set and synthesized face images.

2) M-Step: According to (6), (8) can be simplified to

$$\sum_i \log\left|S_{\mu_{\text{ot}}}^{t+1}\right| + \text{tr}\left(\left(S_{\mu_{\text{ot}}}^{t+1}\right)^{-1} E\left[\mu_{i_{\text{ot}}}\mu_{i_{\text{ot}}}^T\right]\right) + \sum_i \sum_j \log\left|S_{\varepsilon_{\text{ot,st}}}^{t+1}\right| + \text{tr}\left(\left(S_{\varepsilon_{\text{ot,st}}}^{t+1}\right)^{-1} E\left[\varepsilon_{i_{\text{ot,st}}j}\varepsilon_{i_{\text{ot,st}}j}^T\right]\right).$$

As the latent variable $\mathbf{h}_i$ has been estimated in previous step, we can update the parameters by substituting $\mu_{i_{\text{ot}}}$ and $\varepsilon_{i_{\text{ot}}j}$ into the following equation:

$$\mathbf{S}_{\mu_{\text{ot}}}^{\mathbf{t+1}} = \frac{1}{n}\sum_i \mathbf{E}\left[\mu_{i_{\text{ot}}}\mu_{i_{\text{ot}}}^T\right]$$

$$\mathbf{S}_{\varepsilon_{\text{ot,st}}}^{\mathbf{t+1}} = \frac{\sum_i \sum_j \mathbf{E}\left[\varepsilon_{i_{\text{ot,st}}j}\varepsilon_{i_{\text{ot,st}}j}^T\right]}{\sum_i m_i}$$

where n represents the number of subjects in the training set.

The algorithm generally converges within 50 iterations; then, we can utilize (2)–(4) to compute the similarities between the probe image and gallery images. Algorithm 1 summarizes the implementation of the proposed DA-JL method for HFR.

Algorithm 1 DA-JL

Input: Training set **A**, probe image **p**, gallery dataset **G**.

***Step* 1**: Generate synthesized image pairs corresponding to training set **A** via three face sketch synthesis methods: RSLCR, MWF, GANs. Let **B** represent the set of the synthesized image pairs and the original training image pairs.

***Step* 2**: Initialize the inter-class covariance matrix $\mathbf{S}_{\mu_{\text{ot}}}$ from image pairs of training set **A** and the intra-class covariance matrix $\mathbf{S}_{\varepsilon_{\text{ot,st}}}$ from image pairs of dataset **B**.

***Step* 3**: Apply the EM strategy to jointly optimize $\mathbf{S}_{\mu_{\text{ot}}}$ and $\mathbf{S}_{\varepsilon_{\text{ot,st}}}$. Then, calculate **M** and **N** according to Eqs. (3) and (4), respectively.

***Step* 4**: Calculate the similarity of probe image **p** and each image in gallery dataset **G**. Sort the similarities in descending order.

Output: The target heterogeneous face image **t** in gallery dataset **G**.

IV. Experimental Results and Analysis

To validate the effectiveness of the proposed approach, we conduct experiments on six HFR scenarios, i.e., viewed sketches versus visible images, forensic sketches versus visible images, NIR images versus visible images, TIR images versus visible images, low-resolution images versus high-resolution images, and viewed sketches without occlusion versus visible images with occlusion. First, we investigate the effects of different combinations of face sketch synthesis models and different features on recognition performance. Then, we confirm that the proposed approach achieves superior performance compared with the state-of-the-art methods on the six aforementioned databases, as shown in Fig. 1.

A. Databases and Protocols

1) Viewed Sketches: Artists draw the viewed sketches while viewing photos. For the viewed sketches, we compare the proposed approach with the state-of-the-art methods on the CUFSF database [7], which contains 1194 sketch-photo pairs. Fig. 1(a) presents some example images from the CUFSF database. We randomly select 500 subjects as the training set, which is also utilized to generate synthesized image pairs. The remaining 694 subjects are used for testing. All the synthesized images are combined with the original training set to augment the scale of the training data.

2) Forensic Sketches: Different from viewed sketches, the forensic sketches (Forensic Sketch database [4]) are drawn according to the descriptions of eyewitnesses or victims by artists and are used for law enforcement. The difference between viewed sketches and forensic sketches is substantial. To address this problem, researchers developed 140 semiforensic sketch-photo pairs (IIIT-D Sketch database [43]) drawn by artists according to their memory of photos viewed once by the artists. The Forensic Sketch database contains 168 mugshot photos and the corresponding forensic sketches from the real world. Fig. 1(b) and (c) shows example images from the IIIT-D database and the Forensic Sketch database. Following the first partition protocols in [4], we randomly select 124 sketch-photo pairs from the IIIT-D database as the training set and use the Forensic Sketch database for testing. Due to the considerable gap between forensic sketches and photos, a valid synthesis model is difficult to train. Therefore, we introduce the CUHK AR Sketch database [21], as shown in Fig. 1(g), to obtain an augmented training set.

Following the second partition protocol, we randomly select 112 subjects from the Forensic Sketch database as the training set. The remaining 56 subjects in the Forensic Sketch database are used as the test set. To address the problem similar to that in the first partition protocol, 250 subjects from the CUFSF database are randomly selected to generate an augmented training set. The gallery sets are all extended by 10 000 photos to make the task more similar to real-word scenarios.

3) Near-Infrared Images: NIR cameras can capture the reflected infrared waves of objects without visible light and are therefore robust to lighting conditions. Matching NIR face images with VIS face images is a common technique in security scenarios where the circumambient illumination is poor. For NIR images, we compare the proposed approach with the baseline method and the state-of-the-art methods on the CUHK NIR-VIS database [12], which contains 2800 subjects. Some example images are presented in Fig. 1(d). Each identity in the CUHK NIR-VIS database has one NIR image and a corresponding visible photo. Following the same partition protocol in [12], the database is randomly divided into two

halves without overlap: one half for training and the other half for testing. The training set is augmented by synthesis models following the same strategy of data augmentation used for the CUFSF database.

4) Thermal-Infrared Images: The TIR images are captured by TIR cameras, which produce face images from the TIR waves emitted by the human body. TIR face images are widely used in scenarios with occlusions (i.e., glasses, hats, and masks) on people's faces. Due to substantial discrepancies between TIR face images and visible face images, as shown in Fig. 1(e), matching TIR face images with VIS face images is much more difficult than matching NIR face images with VIS face images. For the TIR images, we compare the proposed DA-JL approach with the state-of-the-art methods on the USTC-NVIE database [27]. This database contains 215 subjects with variations in illumination, disguise, and expression. We select subjects that have both TIR and visible images. To make this database more difficult, we select one TIR-VIS pair for each subject. The final TIR database consists of 129 subjects, each of which has one TIR image and a corresponding visible photo. We randomly select 86 subjects as the training set, and the remaining 43 subjects are utilized for testing. We adopt the same data augmentation strategy as that used for the CUFSF database.

5) Low-Resolution Photos: The NJU-ID database [28] contains low-resolution photos and high-resolution photos of 256 subjects. Each person has one low-resolution photo and one high-resolution photo, as shown in Fig. 1(f). The low-resolution photos, with a resolution of 102×126, are from the second generation of resident ID cards of China, whereas the high-resolution photos are captured by a digital camera at a resolution of 640×480. We randomly select 100 pairs as the training set. We take the low-resolution photos of the remaining 156 pairs for testing, and the corresponding high-resolution photos compose the gallery set. The augmentation for generating low-resolution and high-resolution images is the same as the previous face sketch-photo synthesis process.

6) Gallery Augmentation: Heterogeneous face images are widely used in real-world face retrieval scenarios. However, the scales of the Forensic Sketch database, USTC-NVIE database, and NJU-ID database are relatively small. Therefore, we extend the galleries to make these tasks more similar to real-word HFR scenarios. The 10 000 face photos used to extend the galleries are randomly selected from the FERET database (2722 photos) [44], XM2VTS database (1180 photos) [45], CAS-PEAL database (3098 photos) [46], and Labeled Faces in the Wild-a (LFW-a) database (3000 photos) [47]. The photos include a total of 5329 persons.

7) Robustness Validation: To evaluate the robustness of the proposed approach to occlusion, we collect two heterogeneous occlusion face databases, as shown in Fig. 1(g). The sketches are from the CUHK AR Sketch database [21], and photos with glasses or a scarf are from the AR Face database [29], which is also used to evaluate the robustness of traditional face recognition in [48] and [49]. The data set contains a total of 123 subjects. Each subject has one sketch and one photo with glasses or a scarf. To make this task more challenging, we select 48 sketch-photo pairs from the CUHK AR Sketch database [21] as the training set. The remaining 75 sketches from the CUHK AR Sketch database constitute the probe set. The gallery set consists of 75 photos from the AR Face database with the same identities as those in the probe set.

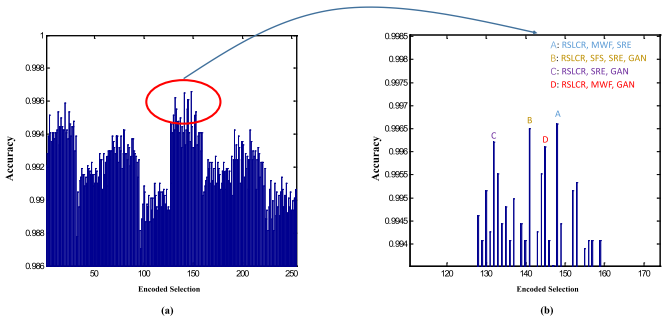

Fig. 4. (a) Recognition results on the CUFSF database corresponding to different combinations of eight synthesis methods. (b) Several of the highest recognition rates from (a).

B. Experimental Settings

The face images are aligned based on the centers of the two eyes and cropped to 250×200. The image patches used in the synthesis methods are 10×10 with 50% overlap between adjacent patches. All experiments are conducted on a Windows 7 operating system with an i7-4790 3.6-GHz CPU with MATLAB R2016b software. The proposed method does not consume substantial time for training and testing. For example, for 500 pairs of images in the training set, 32.52 s on average are required to train the proposed DA-JL model. In the test phase, the proposed method can identify 9.63 probe images from the gallery of 694 images in 1 s, on average. We repeat all the experiments 20 times for the proposed method and the baseline method by randomly partitioning the databases. Then, we report the average accuracies and standard deviations.

1) Discussion on the Combinations of Different Face Sketch Synthesis Methods: The combination of the synthesis methods whose generated image pairs have better perceptual quality does not guarantee better performance, because complementary information extracted from different synthesized image pairs is more important for recognition. Distinct categories of synthesis methods usually provide different types of complementary information. Face sketch-photo synthesis methods are generally categorized into three groups [50]: subspace learning-based methods (i.e., LLE [20], SFS [24], SVR [24], SRE [23], and RSLCR [25]), Bayesian inference-based methods (i.e., MWF [22] and MRF [21]), and deep learning-based methods (i.e., GANs [38]). For each category, we choose the method that performs best in terms of SSIM scores [51]. Finally, we choose RSLCR, MWF, and GANs to generate the synthesized images. In addition, we select some number (from 1 to 8) of synthesized image pairs for joint training and traverse all the combinations. Each experiment is repeated 10 times, and the experimental results are shown in Fig. 4. The best four combinations are {RSLCR, MWF, SRE}, {RSLCR, SFS, SRE, GANs}, {RSLCR, SRE, GANs}, and {RSLCR, MWF, GANs}. Our choice is in the top four combinations, which means that the proposed selection strategy is effective.

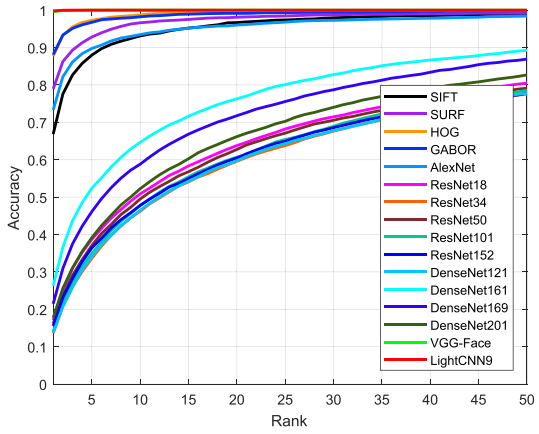

Fig. 5. Performance comparison with different features on the CUFSF database.

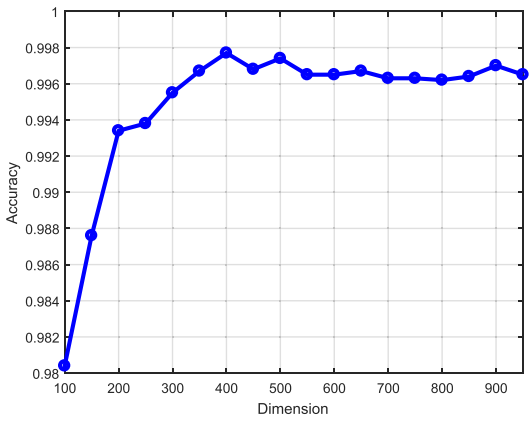

Fig. 6. Accuracies of different feature dimensions on the CUFSF database.

2) Feature Exploration: We choose the best feature descriptor for the proposed framework from two types of feature descriptors: local feature descriptors (SIFT [31], HOG [52], speeded up robust features (SURF) [53], and GABOR [54]) and deep feature representations (VGG-Face [55], Alexnet [56], ResNet [57], DenseNet [58], and LightCNN [59]). For SIFT, we employ an open-source library, i.e., VLfeat. We take the center of each image patch as the interest point and apply the default parameter settings to obtain the standard 128-D vector. For GABOR, we use the library function in the MATLAB 2016b software. The wavelengths are set to 5, 10, 15, 20, and 25, and the orientations are set to 0, 22.5, 45, 67.5, 90, 112.5, 135, and 157.5. The center pixel of each image patch extracted from the GABOR filter consists the GABOR feature. For SURF, we use the library function in the MATLAB R2016b software, which uses the standard SURF-64 version. The interest point is manually set as the center of each image patch. We choose the default parameter settings and obtain the 64-D vector as the SURF feature descriptor. For deep feature representations, we utilize the output of the last pooling layer in the networks. The pretrained models of AlexNet, ResNet, and DenseNet, implemented in the Pytorch environment [60], are used to extract the corresponding features. The VGG-Face model of VLfeat [61] is used to extract the VGG feature, and the LightCNN-9 model is employed to extract the LightCNN feature. Fig. 5 presents a performance comparison of the different features. AlexNet, ResNet, and DenseNet are pretrained on the ImageNet database [62], which is not highly compatible for the face recognition task. LightCNN and VGG-Face are pretrained on face images, which are more compatible for the HFR task. However, LightCNN is pretrained on the MS-Celeb-1M database [63], which contains many more face samples than the database used in the VGG-Face [59]. In addition, the Max-Feature-Map operation used in LightCNN has better generalization ability than ReLU used in VGG-Face. The LightCNN feature considerably outperforms the other features. Hence, we represent images as LightCNN features for the proposed approach in the following experiments.

In addition, we explore the effectiveness of different feature dimensions for the proposed framework. The dimensions of the features are reduced by PCA, and the accuracies of different dimensions of LightCNN features are shown in Fig. 6. The best dimensionality is 400, which is much smaller than the feature dimension of our early version AJL-HFR [30].

TABLE I

RANK-1 RECOGNITION ACCURACIES OF THE STATE-OF-THE-ART APPROACHES AND OUR METHOD ON THE CUFSF DATABASE

Algorithms	Rank-1 Recognition Accuracy
LRBP [8]	91.12%
LDoGBP [9]	91.04%
G-HFR [4]	96.04%
PLS [17]	51.00%
MvDA [18]	55.50%
MRF [21]	46.03%
MWF [22]	74.15%
RSLCR [25]	75.94%
VGG [55]	45.82%
SeetaFace [64]	16.57%
JB [26]	98.43%
AJL-HFR [30]	**99.61%**
DA-JL	**99.77%**

C. Experimental Results

1) Results on Viewed Sketches: In this experiment, we compare the proposed approach with the state-of-the-art methods on the CUFSF database. Under the same partition protocol, we evaluate the rank-1 recognition accuracies. The comparison of rank-1 recognition accuracies is shown in Table I. For the invariant feature extraction-based methods LRBP [8], LDoGBP [9], and G-HFR [4], the cross-modality invariant feature is extracted for recognition. The best rank-1 accuracy of these three methods is 96.04% for G-HFR [4], which consumes excessive time in the feature extraction phase. For the common space learning-based methods PLS [17] and MvDA [18], images from different modalities are projected to a discriminant common space. These two methods are applicable for multiple heterogeneous scenarios, but the rank-1 accuracies are both below 60%. For the synthesis-based methods MRF [21], MWF [22], and RSLCR [25], heterogeneous face images are transformed to homogeneous

TABLE II

RANK-50 RECOGNITION ACCURACIES OF THE STATE-OF-THE-ART APPROACHES AND OUR METHOD ON THE IIIT-D DATABASE AND THE FORENSIC SKETCH DATABASE

Database	Algorithms	Rank-50 Recognition Accuracy
IIIT-D Sketch Database	MCWLD [43]	28.52%
	VGG [55]	22.62%
	G-HFR [4]	30.36%
	JB [26]	38.51%
	AJL-HFR [30]	**66.99%**
	DA-JL	**67.86%**
Forensic Sketch Database	P-RS [2]	20.80%
	VGG [55]	24.46%
	G-HFR [4]	31.96%
	JB [26]	31.07%
	AJL-HFR [30]	**72.86%**
	DA-JL	**73.75%**

TABLE III

RANK-1 RECOGNITION ACCURACIES OF THE STATE-OF-THE-ART APPROACHES AND OUR METHOD ON THE CUHK VIS-NIR DATABASE

Algorithms	Rank-1 Accuracy
LFDA [6]	69.22%
CITE [7]	72.53%
CEFD [12]	83.93%
LCKS-CSR [35]	71.21%
P-RS [2]	72.93%
MWF [22]	74.89%
RSLCR [25]	66.82%
VGG [55]	62.91%
SeetaFace [64]	69.50%
JB [26]	98.71%
AJL-HFR [30]	**99.05%**
DA-JL	**99.33%**

face images by a set of reconstruction coefficients. However, the synthesized photos are similar to the ground-truth sketches in shape, as explained above, which degrades the performance. For the deep learning methods VGG [55] and SeetaFace [64], the models are trained on visible photos, and the performance is poor, which further illustrates that general face recognition methods are not suitable for face sketch-photo recognition. The baseline method JB [26] achieves a rank-1 accuracy of $98.43 \pm 0.32\%$. Our early version AJL-HFR [30] achieves a rank-1 accuracy of $99.61 \pm 0.19\%$, which is superior to the state-of-the-art methods and the baseline method. Moreover, DA-JL further improves the rank-1 accuracy to $99.77 \pm 0.17\%$ with lower feature dimension.

2) Results on Forensic Sketches: Due to the substantial discrepancies between forensic sketches and mugshot photos, the Forensic Sketch database is much more difficult than the CUFSF database. The task is even more difficult when the artists' experience is different or the eyewitnesses' memory is distorted. The rank-1 recognition accuracies of the state-of-the-art methods are all below 20%, which is minimally beneficial to law enforcement purposes. Therefore, the state-of-the-art methods evaluate the recognition accuracies on the Forensic Sketch database at rank-50. The rank-50 recognition accuracies of the baseline method, state-of-the-art methods, and the proposed approach on the IIIT-D Sketch database and the Forensic Sketch database are shown in Table II.

For the IIIT-D database, we compare the proposed approach with the baseline methods JB [26] and the state-of-the-art methods Bhatt *et al.* [43], Peng *et al.* [4], and VGG [55]. To address the considerable discrepancies between forensic sketches and mugshot photos, Bhatt *et al.* [43] proposed a multiscale circular Webers local descriptor (MCWLD) to train the model on the IIIT-D database and achieved better performance than that of a model trained on viewed sketch databases. Following the partition protocol in [4], we first train the proposed method on IIIT-D by randomly selecting 124 subjects and test the method on the Forensic Sketch database. For the Forensic Sketch database collected from real-world forensic sketches and mugshot photos, we compare the proposed approach with the baseline methods and the state-of-the-art methods of Klare and Jain [2] and Peng *et al.* [4]. Klare and Jain [2] proposed a prototype random subspaces (P-RS) method that is trained and tested on the Forensic Sketch database. We follow the partition protocol in [4]: 112 subjects are randomly selected from the Forensic Sketch database to train the proposed approach and 56 forensic sketches are taken as the probe set. The gallery set is enlarged from 56 photos to 10 056 photos by adding 10 000 photos. The experimental results are presented in Table II.

G-HFR achieves 30.36% and 31.96% recognition accuracies at rank-50 on the IIIT-D database and Forensic Sketch database, respectively, which is the best performance for the state-of-the-art methods. The baseline method JB [26] achieves rank-50 accuracies of $38.51 \pm 2.96\%$ and $31.07 \pm 5.96\%$. However, the performance of our AJL-HFR achieves accuracies of $66.99 \pm 1.79\%$ and $72.86 \pm 2.35\%$, which are superior to those of the baseline method and the state-of-the-art methods. The proposed DA-JL further improves the performance to $67.86 \pm 1.25\%$ and $73.75 \pm 2.07\%$.

3) Results on Near-Infrared Images: We compare the proposed approach with the state-of-the-art methods on the CUHK VIS-NIR database, which contains 2800 subjects. Each subject has one NIR image and a corresponding visible image. The experimental results presented in Table III are obtained under the same partition protocol as that in [12]. For the invariant feature extraction-based methods, we choose three methods, namely, LFDA [6], CITE [7], and CEFD [12], to present in Table III. Klare *et al.* [6] designed an LFDA method to use the local feature descriptors to represent cross-modality face images and achieved a rank-1 recognition accuracy of 69.22%. Zhang *et al.* [7] utilized an encoding feature descriptor (CITE) to learn the invariant feature for heterogeneous face images and achieved a rank-1 recognition accuracy of 72.53%. Gong *et al.* [12] extracted the invariant feature by a common encoding feature discriminant approach and achieved a rank-1 recognition accuracy of 83.93%, which is the best performance among the state-of-the-art methods. For the common space learning-based methods, we choose two state-of-the-art methods, namely, LCKS-CSR [16] and

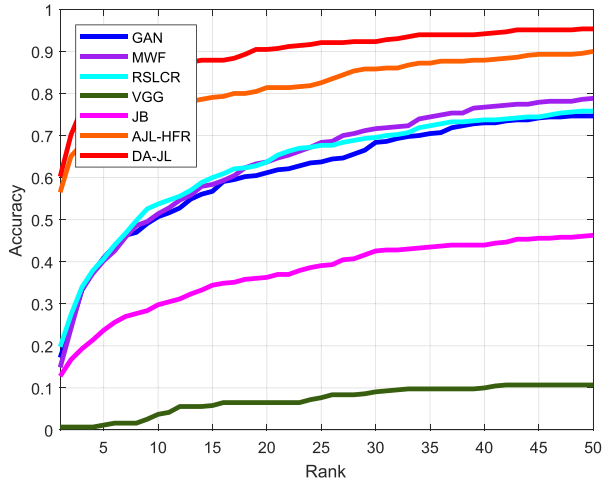

Fig. 7. Results on the USTC-NVIE database.

P-RS [2], as shown in Table III. LCKS-CSR is a learning-based model used to enhance the discriminative ability of the cross-modality feature descriptor and achieves a 71.21% rank-1 recognition accuracy; P-RS [2] achieves a similar rank-1 recognition accuracy. For the synthesis-based methods MWF [22] and RSLCR [25], face images from different modalities are transformed to homogeneous face images by a set of reconstruction coefficients learned from the training set. Then, null-space linear discriminant analysis [65] is employed to conduct the face recognition experiments in homogeneous face images. MWF [22] and RSLCR [25] achieve rank-1 recognition accuracies of 74.89% and 66.82%, respectively. For the deep learning methods VGG [55] and SeetaFace [64], the models are trained on visible photos, and the rank-1 recognition accuracies are below 70%. The proposed DA-JL achieves a rank-1 accuracy of $99.33 \pm 0.17\%$, which is superior to the baseline method JB [26], with an accuracy of $98.71 \pm 0.33\%$, and AJL-HFR [30], with an accuracy of $99.05 \pm 0.32\%$.

4) Results on Thermal-Infrared Images: We compare the proposed approach with the state-of-the-art methods GANs [38], MWF [22], RSLCR [25], VGG [55], and SeetaFace [64] and the baseline method JB [26] on the USTC-NVIE database [27], as shown in Fig. 7. All the experiments are conducted under the same partition protocol. Similar to the evaluation benchmark of the Forensic Sketch database, we compare the proposed approach with the state-of-the-art methods in terms of the rank-50 recognition accuracy. The experimental results are shown in Table IV. Deep learning methods used for traditional face recognition perform poorly in HFR scenarios. SeetaFace is invalid in this recognition scenario, and VGG achieves only a 10.70% accuracy at rank-50. Synthesis-based methods are more effective in this scenario where the differences between heterogeneous images are large. The rank-50 accuracy of our AJL-HFR is $90.00 \pm 3.96\%$, which is superior to the synthesis-based methods and to the $46.28 \pm 12.71\%$ of JB [26]. The proposed DA-JL further improves the performance of AJL-HFR to $95.35 \pm 2.45\%$.

5) Results on Low-Resolution Photos: Fig. 8 presents the results on the NJU-ID database [28]. We compare the proposed approach with the state-of-the-art methods GANs [38], MWF [22], RSLCR [25], and VGG [55] and the baseline method JB [26]. Similar to the evaluation benchmark on the Forensic Sketch database, the rank-50 recognition accuracy is adopted to evaluate these methods, as shown in Table V. The baseline method JB achieves only a $17.95 \pm 11.69\%$ accuracy at rank-50. The AJL-HFR method achieves much better accuracy of $86.39 \pm 1.98\%$ than VGG and JB. The proposed DA-JL method further improves the recognition accuracy to $89.03 \pm 1.55\%$. Since the images in the NJU-ID database are all visible photos with different resolutions, VGG, which is pretrained on a large quantity of visible images, achieves better performance than the state-of-the-art face sketch-photo synthesis methods such as GAN, MWF, and RSLCR. Synthesis-based methods are less effective in this scenario, because the gap between low- and high-resolution images is small.

TABLE IV

RANK-50 RECOGNITION ACCURACIES OF THE STATE-OF-THE-ART APPROACHES AND OUR METHOD ON THE USTC-NVIE DATABASE

Algorithms	Rank-50 Accuracy on USTC-NVIE database
GANs [38]	74.65%
MWF [22]	79.53%
RSLCR [25]	77.91%
VGG [55]	10.70%
JB [26]	46.28%
AJL-HFR [30]	**90.00%**
DA-JL	**95.35%**

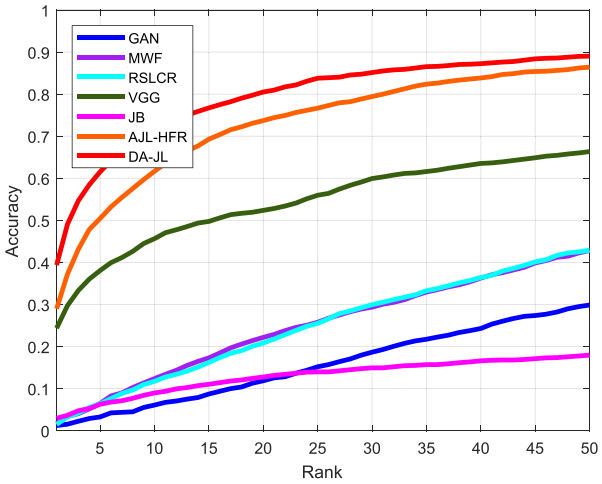

Fig. 8. Results on the NJU-ID database.

TABLE V

RANK-50 RECOGNITION ACCURACIES OF THE STATE-OF-THE-ART APPROACHES AND OUR METHOD ON THE NJU-ID DATABASE

Algorithms	Rank-50 Recognition Accuracy
GANs [38]	29.74%
MWF [22]	42.76%
RSLCR [25]	43.72%
VGG [55]	66.35%
JB [26]	17.95%
AJL-HFR [30]	**86.39%**
DA-JL	**89.03%**

TABLE VI
RANK-1 RECOGNITION ACCURACIES OF THE STATE-OF-THE-ART APPROACHES AND OUR METHOD ON THE AR DATABASE WITH OCCLUSION

Algorithms	Rank-1 Accuracy on AR with scarf	Rank-1 Accuracy on AR with glasses
GANs [38]	22.40%	39.60%
MWF [22]	41.07%	38.67%
RSLCR [25]	36.00%	42.27%
VGG [55]	39.20%	40.40%
JB [26]	29.20%	37.20%
AJL-HFR [30]	**69.33%**	**70.00%**
DA-JL	**65.33%**	**66.93%**

6) Results on Images With Occlusion: We compare the proposed approach with the state-of-the-art methods GANs [38], MWF [22], RSLCR [25], and VGG [55] and the baseline method JB [26]. The image pairs in the training set are cross-modality. However, the images in the gallery set are from another database and include a scarf or glasses. This task is difficult not only for the state-of-the-art methods but also for the deep learning-based methods. Table VI presents the results on images with occlusion. The experimental results of the state-of-the-art methods and baseline method are much lower than those of the AJL-HFR method, which achieves accuracies of $69.33 \pm 4.40\%$ and $70.00 \pm 3.51\%$ on the AR database images with a scarf and glasses, respectively. The proposed DA-JL method also achieves competitive rank-1 accuracies of $65.33 \pm 4.75\%$ and $66.93 \pm 5.47\%$. The reason that AJL-HFR achieves higher accuracy than DA-JL is that VGG contains many more convolutional kernels than LightCNN (512 kernels versus 128 kernels). More kernels usually result in better representation diversity. The generalization ability of the VGG feature is therefore better than that of the LightCNN feature, and the discriminability of the LightCNN feature is better than that of the VGG feature. Therefore, AJL-HFR performs better than DA-JL in the cross-database scenario.

D. Discussion

As claimed in our motivation, the original training set is asymmetric, because each identity has only one pair of heterogeneous face images. Thus, the existing state-of-the-art methods cannot obtain enough intraclass information to train a discriminative model. The proposed approach takes advantage of data augmentation to provide the missing shape and texture information in different modalities. Furthermore, we develop a joint learning model to reduce the influence of redundant information on interclass information. Therefore, our early version AJL-HFR achieves superior performance compared with the state-of-the-art methods. Considering the discriminability of image representations, we introduce the LightCNN model to represent images. The dimensions of the image representations are also reduced substantially. DA-JL further improves the performance of AJL-HFR on most heterogeneous face databases.

V. CONCLUSION

In this paper, we conduct multiple experiments on a viewed sketch (CUFSF) database, Forensic Sketch database, NIR image (CUHK VIS-NIR) database, TIR image (USTC-NVIE) database, low-resolution photo (NJU-ID) database, and occlusion image (AR database with occlusion) database. Each identity in the training set has only one pair of heterogeneous face images. The state-of-the-art methods train models on the data sets without data augmentation, and a valid model is difficult to be obtained, since the samples in the training set are asymmetric. The proposed DA-JL method addresses this problem by extracting more intraclass information from the augmented training set and simultaneously avoiding the loss of interclass information. The experimental results illustrate the effectiveness and superiority of the proposed DA-JL in comparison with the state-of-the-art methods and the excellent generalization ability in multiple HFR scenarios. Our future work will focus on investigating more effective synthesis models and researching the essential features of different HFR scenarios to improve the recognition performance.

REFERENCES

[1] S. Klum, H. Hu, A. K. Jain, and B. Klare, "Sketch based face recognition: Forensic vs. Composite sketches," in *Proc. Int. Conf. Biometrics*, Jun. 2013, pp. 1–8.

[2] B. F. Klare and A. K. Jain, "Heterogeneous face recognition using kernel prototype similarities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1410–1422, Jun. 2013.

[3] R. He, X. Wu, Z. Sun, and T. Tan, "Wasserstein CNN: Learning invariant features for NIR-VIS face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2018.2842770.

[4] C. Peng, X. Gao, N. Wang, and J. Li, "Graphical representation for heterogeneous face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 301–312, Feb. 2017.

[5] S. Liao, D. Yi, Z. Lei, R. Qin, and S. Z. Li, "Heterogeneous face recognition from local structures of normalized appearance," in *Proc. Int. Conf. Biometrics*. Berlin, Germany: Springer, 2009, pp. 209–218.

[6] B. F. Klare, Z. Li, and A. K. Jain, "Matching forensic sketches to mug shot photos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 639–646, Mar. 2011.

[7] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 513–520.

[8] H. K. Galoogahi and T. Sim, "Face sketch recognition by local radon binary pattern: LRBP," in *Proc. 19th IEEE Int. Conf. Image Process. (ICIP)*, Sep./Oct. 2012, pp. 1837–1840.

[9] A. T. Alex, V. K. Asari, and A. Mathew, "Local difference of Gaussian binary pattern: Robust features for face sketch recognition," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2013, pp. 1211–1216.

[10] P. Mittal, M. Vatsa, and R. Singh, "Composite sketch recognition via deep network—A transfer learning approach," in *Proc. Int. Conf. Biometrics (ICB)*, May 2015, pp. 251–256.

[11] M. Shao and Y. Fu, "Cross-modality feature learning through generic hierarchical hyperlingual-words," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 2, pp. 451–463, Feb. 2016.

[12] D. Gong, Z. Li, W. Huang, X. Li, and D. Tao, "Heterogeneous face recognition: A common encoding feature discriminant approach," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2079–2089, May 2017.

[13] D. Lin and X. Tang, "Inter-modality face recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2006, pp. 13–26.

[14] D. Yi, R. Liu, R. Chu, Z. Lei, and S. Z. Li, "Face matching between near infrared and visible light images," in *Proc. Int. Conf. Biometrics (ICB)*, 2007, pp. 523–530.

[15] A. Li, S. Shan, X. Chen, and W. Gao, "Maximizing intra-individual correlations for face recognition across pose differences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 605–611.

[16] Z. Lei, C. Zhou, D. Yi, A. K. Jain, and S. Z. Li, "An improved coupled spectral regression for heterogeneous face recognition," in *Proc. 5th IAPR Int. Conf. Biometrics (ICB)*, Mar./Apr. 2012, pp. 7–12.

[17] A. Sharma and D. W. Jacobs, "Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 593–600.

[18] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 188–194, Jan. 2012.
[19] D. Yi, Z. Lei, and S. Z. Li, "Shared representation learning for heterogenous face recognition," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, vol. 1, May 2015, pp. 1–7.
[20] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma, "A nonlinear approach for face sketch synthesis and recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 1005–1010.
[21] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1955–1967, Nov. 2009.
[22] H. Zhou, Z. Kuang, and K.-Y. K. Wong, "Markov weight fields for face sketch synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 1091–1097.
[23] X. Gao, N. Wang, D. Tao, and X. Li, "Face sketch–photo synthesis and retrieval using sparse representation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 8, pp. 1213–1226, Aug. 2012.
[24] N. Wang, J. Li, D. Tao, X. Li, and X. Gao, "Heterogeneous image transformation," *Pattern Recognit. Lett.*, vol. 34, no. 1, pp. 77–84, 2013.
[25] N. Wang, X. Gao, and J. Li, "Random sampling for fast face sketch synthesis," *Pattern Recognit.*, vol. 76, pp. 215–227, 2018.
[26] D. Chen, X. Cao, D. Wipf, F. Wen, and J. Sun, "An efficient joint formulation for Bayesian face verification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 32–46, Jan. 2017.
[27] S. Wang *et al.*, "A natural visible and infrared facial expression database for expression recognition and emotion inference," *IEEE Trans. Multimedia*, vol. 12, no. 7, pp. 682–691, Nov. 2010.
[28] J. Huo, Y. Gao, Y. Shi, W. Yang, and H. Yin, "Heterogeneous face recognition by margin-based cross-modality metric learning," *IEEE Trans. Cybern.*, vol. 48, no. 6, pp. 1814–1826, Jun. 2018.
[29] A. M. Martinez, "The AR face database," CVC, Barcelona, Spain, Tech. Rep. #24, 1998.
[30] B. Cao, N. Wang, X. Gao, and J. Li, "Asymmetric joint learning for heterogeneous face recognition," in *Proc. AAAI*, 2018, pp. 1–8.
[31] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[32] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[33] Z. Li, D. Lin, and X. Tang, "Nonparametric discriminant analysis for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 755–761, Sep. 2009.
[34] Z. Lei and S. Z. Li, "Coupled spectral regression for matching heterogeneous faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1123–1128.
[35] Z. Lei, S. Liao, A. K. Jain, and S. Z. Li, "Coupled discriminant analysis for heterogeneous face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 6, pp. 1707–1716, Dec. 2012.
[36] N. Wang, D. Tao, X. Gao, X. Li, and J. Li, "Transductive face sketch-photo synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 9, pp. 1364–1376, Sep. 2013.
[37] C. Peng, X. Gao, N. Wang, D. Tao, X. Li, and J. Li, "Multiple representations-based face sketch–photo synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 11, pp. 2201–2215, Nov. 2016.
[38] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. (2016). "Image-to-image translation with conditional adversarial networks." [Online]. Available: https://arxiv.org/abs/1611.07004
[39] N. Wang, D. Tao, X. Gao, X. Li, and J. Li, "A comprehensive survey to face hallucination," *Int. J. Comput. Vis.*, vol. 106, no. 1, pp. 9–30, 2014.
[40] D. Wang and X. Tan, "Bayesian neighborhood component analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 7, pp. 3140–3151, Jul. 2018.
[41] B. Moghaddam, T. Jebara, and A. Petland, "Bayesian face recognition," *Pattern Recognit.*, vol. 33, no. 11, pp. 1771–1782, 2000.
[42] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2288–2295.
[43] H. S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa, "Memetic approach for matching sketches with digital face images," Indraprastha Inst. Inform. Technol.-Delhi, New Delhi, India, Tech. Rep. TR-2011-006, Oct. 2011.
[44] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.
[45] K. Messer, "XM2VTSDB: The extended M2VTS database," in *Proc. 2nd Int. Conf. Audio Video-Based Biometric Person Authentication*, 1999, pp. 72–77.
[46] W. Gao *et al.*, "The CAS-PEAL large-scale Chinese face database and baseline evaluations," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 38, no. 1, pp. 149–161, Jan. 2008.
[47] L. Wolf, T. Hassner, and Y. Taigman, "Effective unconstrained face recognition by combining multiple descriptors and learned background statistics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 1978–1990, Oct. 2011.
[48] S.-J. Wang, J. Yang, M.-F. Sun, X.-J. Peng, M.-M. Sun, and C.-G. Zhou, "Sparse tensor discriminant color space for face verification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 6, pp. 876–888, Jun. 2012.
[49] R. He, W.-S. Zheng, B.-G. Hu, and X.-W. Kong, "Two-stage nonnegative sparse representation for large-scale face recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 1, pp. 35–46, Jan. 2013.
[50] N. Wang, M. Zhu, J. Li, B. Song, and Z. Li, "Data-driven vs. Model-driven: Fast face sketch synthesis," *Neurocomputing*, vol. 257, pp. 214–221, Sep. 2017.
[51] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[52] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.
[53] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2006, pp. 404–417.
[54] D. Gabor, "Theory of communication. Part 1: The analysis of information," *J. Inst. Elect. Eng.-III, Radio Commun. Eng.*, vol. 93, no. 26, pp. 429–441, 1946.
[55] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. BMVC*, 2015, vol. 1, no. 3, p. 6.
[56] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 770–778.
[58] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2261–2269.
[59] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.
[60] P. Core Team. *Pytorch*. Accessed: May 9, 2018. [Online]. Available: https://pytorch.org/
[61] T. V. Authors, *Vlfeat.Org*. Accessed: May 9, 2018. [Online]. Available: http://www.vlfeat.org/
[62] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255.
[63] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 87–102.
[64] X. Liu, M. Kan, W. Wu, S. Shan, and X. Chen, "VIPLFaceNet: An open source deep face recognition SDK," *Frontiers Comput. Sci.*, vol. 11, no. 2, pp. 208–218, 2017.
[65] L. Chen, H.-Y. M. Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu, "A new LDA-based face recognition system which can solve the small sample size problem," *Pattern Recognit.*, vol. 33, no. 10, pp. 1713–1726, 2000.

Bing Cao received the B.Eng. degree from Hebei University, Baoding, China, in 2015. He is currently pursuing the Ph.D. degree in intelligent information processing with the School of Electronic Engineering, Xidian University, Xi'an, China.

His current research interests include computer vision and pattern recognition.

Nannan Wang (M'16) received the B.Sc. degree in information and computation science from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2009, and the Ph.D. degree in information and telecommunications engineering from Xidian University, Xi'an, in 2015.

From 2011 to 2013, he was a Visiting Ph.D. Student with the University of Technology, Sydney, NSW, Australia. He is currently with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an. He has authored over 50 papers in refereed journals and proceedings including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the *International Journal of Computer Vision*, the Association for the Advancement of Artificial Intelligence, and the International Joint Conference on Artificial Intelligence. His current research interests include computer vision, pattern recognition, and machine learning.

Jie Li received the B.Eng. degree in electronic engineering, the M.Sc. degree in signal and information processing, and the Ph.D. degree in circuit and systems from Xidian University, Xi'an, China, in 1995, 1998, and 2004, respectively.

She is currently a Professor with the School of Electronic Engineering, Xidian University. She has published around 50 technical articles in refereed journals and proceedings including the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and *Information Sciences*. Her current research interests include image processing and machine learning.

Xinbo Gao (M'02–SM'07) received the B.Eng., M.Sc., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively.

From 1997 to 1998, he was a Research Fellow with the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a Post-Doctoral Research Fellow with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong. Since 2001, he has been with the School of Electronic Engineering, Xidian University. He is currently a Cheung Kong Professor of Ministry of Education, a Professor of Pattern Recognition and Intelligent System, and the Director of the State Key Laboratory of Integrated Services Networks, Xi'an. He has authored five books and around 200 technical articles in refereed journals and proceedings, including the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, the *International Journal of Computer Vision*, and *Pattern Recognition*. His current research interests include multimedia analysis, computer vision, pattern recognition, machine learning, and wireless communications.

Dr. Gao is on the Editorial Boards of several journals, including *Signal Processing* (Elsevier) and *Neurocomputing* (Elsevier). He served as the general chair/co-chair, the program committee chair/co-chair, or a PC member for around 30 major international conferences. He is currently a fellow of the Institution of Engineering and Technology.