Mutual Component Convolutional Neural Networks for Heterogeneous Face Recognition

Zhongying Deng*, Xiaojiang Peng*, Zhifeng Li, *Senior Member, IEEE* and Yu Qiao, *Senior Member, IEEE*

***Abstract*—Heterogeneous face recognition (HFR) aims to identify a person from different facial modalities such as visible and near-infrared images. The main challenges of HFR lie in the large modality discrepancy and insufficient training samples. In this paper, we propose the Mutual Component Convolutional Neural Network (MC-CNN), a modal-invariant deep learning framework, to tackle these two issues simultaneously. Our MC-CNN incorporates a generative module, i.e. the Mutual Component Analysis (MCA) [1], into modern deep convolutional neural networks by viewing MCA as a special fully-connected (FC) layer. Based on deep features, this FC layer is designed to extract modal-independent hidden factors, and is updated according to maximum likelihood analytic formulation instead of back propagation which prevents over-fitting from limited data naturally. In addition, we develop an MCA loss to update the network for modal-invariant feature learning. Extensive experiments show that our MC-CNN outperforms several fine-tuned baseline models significantly. Our methods achieve the state-of-the-art performance on CASIA NIR-VIS 2.0, CUHK NIR-VIS and IIIT-D Sketch dataset.**

***Index Terms*—Heterogeneous face recognition, Mutual Component Analysis, Mutual Component Convolutional Neural Network.**

I. Introduction

HETEROGENEOUS face recognition (HFR), also known as cross-modality face recognition, refers to matching two face images from alternative image modalities such as sketch-to-photo and infrared-to-visible. It has attracted increasing attention due to its wider range of applications in surveillance, authentication, and forensic verification. Although significant progresses have been made recently [2]–[7], HFR is still a challenging problem due to the fact that the probe images and the gallery images are acquired through different processes under different conditions and thus have significant discrepancies. Moreover, it is difficult to collect large scale heterogeneous dataset as those for general visible face recognition.

To alleviate the modal discrepancy, many previous methods [1], [5], [8] first extract hand-crafted features, e.g. SIFT and LBP, and then apply *generative models* to decouple different modalities into a common space. Though improving baseline results significantly, those methods obtain unsatisfactory performance and reach the bottleneck due to the limited representation power of hand-crafted features.

Recently, since the milestone success of Convolutional Neural Networks (CNN) in object classification [9], it has become a general method in computer vision community and achieved impressive performance in face recognition tasks [10]–[15]. Some works have utilized CNN for heterogeneous face recognition [2], [4], [6], [16], which can make full use of the discriminative power of deep neural network and learn the highly non-linear relationship between probe images and gallery images. On the HFR task, the CNN has achieved the state-of-the-art performance and surpassed those methods based on hand-crafted features, but it still has several limitations nevertheless. *First, deep CNN can easily over-fit small HFR datasets due to its massive parameters. Second, classical CNN itself is not able to extract modal-invariant features effectively.*

In this paper, to address the two issues of CNN-based HFR methods, we propose the Mutual Component Convolutional Neural Network (MC-CNN), a novel modal-invariant deep learning framework without massive data. Our MC-CNN is inspired by the Mutual Component Analysis (MCA) model in [1] and the latent factor guided CNN (LF-CNN) in [17]. MCA is a generative model for heterogeneous face recognition, which is based on a hypothesis that mutual components reflect those intrinsic modal-invariant face characteristics. Based on the HOG-MLBP features, the MCA is introduced to extract the mutual components of pairwise images from different modalities [1]. LF-CNN is used to learn age-invariant features by adding a latent identity analysis model in CNN [1]. Based on the same hypothesis, our MC-CNN elaborately embeds an MCA module in deep neural networks. First, we revisit the derivation of MCA, and formulate the computation process of mutual components as a special fully-connected layer which we called MCA layer in the paper. Second, to prevent over-fitting on small data, we update the MCA layer according to its analytic formulation instead of back propagation. Third, to overcome the limitation that MCA only extracts mutual components of an image pair, we design an MCA loss to learn modal-invariant features for each modality. The MCA

This work was supported in part by National Natural Science Foundation of China (U1613211, U1813218), and Shenzhen Research Program (JCYJ20170818164704758, JCYJ20150925163005055), and Tencent AI Lab Rhino-Bird Joint Research Program (No. JR201807) .*(corresponding author: Yu Qiao.)*

*Z. Deng and X. Peng contributed equally.

Z. Deng is with the Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, China and also with Shenzhen Key Lab of Computer Vision and Pattern Recognition, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518000, China. E-mail: zy.deng1@siat.ac.cn.

X. Peng is with Guangdong Key Lab of Computer Vision and Virtual Reality, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518000, China. E-mail: xj.peng@siat.ac.cn

Y. Qiao is with Shenzhen Key Lab of Computer Vision and Pattern Recognition, Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Sciences, Shenzhen, China, and also with SIAT-SenseTime Joint Lab. E-mail: yu.qiao@siat.ac.cn.

Z. Li is with Tencent AI Lab, Shenzhen, Guangdong Province, China. E-mail: michaelzfli@tencent.com.

loss enforces both modal components to approach the mutual components. Thus, it is able to use our MC-CNN to extract modal-invariant features for a single facial image in the test phase, which is valuable in practice since off-line feature extraction is possible. From the view of modal discrepancy, our MC-CNN decreases modal discrepancy by i) incorporating the MCA layer and ii) updating CNN features with the MCA loss.

In summary, we propose a novel modal-invariant deep neural network for heterogeneous face recognition with state-of-the-art results on several HFR datasets. Our contributions are as follows:

- By introducing an MCA layer, we propose the Mutual Component Convolutional Neural Network to extract modal-independent components for different modalities which can extend to other cross-modality tasks instead of HFR in the paper.
- We design an MCA loss for modal-invariant feature extraction of single images.
- Our methods achieve the state-of-the-art performance on three popular HFR datasets, namely 99.39% on CASIA NIR-VIS 2.0 and 99.44% on CUHK NIR-VIS and 87.4% on IIIT-D Sketch.

The rest of this paper is organized as follows. In Sec. II, we review related work on heterogeneous face recognition. In Sec. III, we revisit the mutual component analysis model. In Sec. IV, we introduce our mutual component neural network and present the detail network architecture for HFR. We show experimental results in Sec. V, and summarize our work in Sec. VI.

II. Related work

Heterogeneous learning with insufficient training examples is a challenging task [18]–[21]. Heterogeneous face recognition (HFR) can be seen as a sub-task of heterogeneous learning and has been extensively studied in the last decade [1]–[6], [22]–[28]. This section presents HFR-related works from the perspective of model characteristic, i.e. generative and discriminative.

A. Discriminative models

Discriminative models aim to learn a modal-invariant subspace or to select discriminative features from the whole feature space. In early time, discriminative models like linear discriminant analysis (LDA) [29] are applied on hand-crafted features. Lin *et al*. [30] propose a Common Discriminant Feature Extraction model to project NIR and VIS images into a common space for classification. Yi *et al*. [31] use PCA/LDA to obtain modal-shared subspace. Li *et al*. [32] apply a Hetero-Component Analysis model extended from PCA to get modal-invariant features. Klare *et al.* [25] propose a prototype representation with non-linear kernel similarities and use LDA to enhance the discriminative power of features. Liu *et al.* [24] apply the Gentle Boost classifier to select a discriminative representation based on the Light Source Invariant Features (LSIFs). Huang *et al.* [28] use a discriminative spectral regression (DSR) method to minimize the distance of images from the same class and maximize the distance of images from the different class. Some other works [26], [27] adopt general domain adaptation approaches to learn a domain-independent representation for HFR.

In recent years, as the most popular discriminative model, deep convolutional neural networks (CNN) have been applied to heterogeneous face recognition and improved the performance of HFR task in a large margin. Wu *et al.* [2] use a coupled deep learning (CDL) approach to seek a shared feature space. Sarfraz *et al.* [4] use a deep neural network to learn the highly non-linear relationship between two modalities. Saxena *et al.* [6] extend the features from a pre-trained CNN on visible spectrum by reducing the discrepancies between the different modalities. Reale *et al.* [16] use a CNN to learn the relationship between the entirety of cross-modal face images and takes some measures to prevent over-fitting. Hu *et al.* [7] propose a data generation method to alleviate over-fitting and train CNN models on both real and synthetic images. Hu *et al.* [33] develop a tensor-based framework to fuse face recognition features (FRF) and facial attribute features (FAF) to enhance face recognition performance. Liu *et al.* [34] present a deep Transfer NIR-VIS heterogeneous facE recognition neTwork (TRIVET) for NIR-VIS face recognition where they first use unpaired VIS image to train a CNN model and then transfer these models to NIR-VIS domain by fine-tuning CNN with two types of NIR-VIS triplet loss. Our work differs from these CNN-based methods in that we embed a generative model into the CNN model to utilize the advantages of generative model to deal with the problems on the small HFR datasets.

B. Generative models

In general, generative models employ some prior knowledge and have good model interpretability. Shi *et al.* [3] propose a heterogeneous joint Bayesian (HJB) formulation. The HJB extends the original joint Bayesian to heterogeneous images by applying two different Gaussian models. Tiago *et al.* [22] extend the Inter-Session Variability model from speaker authentication to HFR which mainly uses Gaussian Mixture Models to model cross-modal features. Shaikh *et al.* [35] employ a tied factor analysis (TFA) model and a bagging method to cope with the small training data problem. Li *et al.* [1] propose a Mutual Component Analysis (MCA) model to extract the mutual components of different modalities. Some other works [5], [23] apply graphical models (e.g. Markov models) to capture the spatial relation of heterogeneous faces. Our work differs from these generative models in that i) we train a generative model and a discriminative model (i.e. CNN) jointly, which takes full use of the feature learning power of CNN, ii) we introduce a MCA loss to make it possible for modal-invariant feature extraction of single images.

III. Revisit of mutual component analysis

In this section, we revisit the Mutual Component Analysis (MCA) model introduced in [1]. The MCA model is a generative model for extracting the mutual components of two different modalities, and it provides superior performance on several HFR datasets than other generative models such as

canonical correlation analysis [1]. In the following, we first give the notations used in the paper, and more details can be found in [1].
Superscript k indicates the index of modality k.
Subscript i indicates the index of the i-th observation.
N: The number of the training observations for each modality.
D: The dimension of image features.
d: The dimension of hidden factor $\vec{y}$.
K: The number of modalities (Typically, $K = 2$ in sketch-VIS or NIR-VIS face recognition).
M^k: $D \times N$ matrix with N training image features (columns) in modality k.
$\vec{m}_i^k$: The i-th $D \times 1$ column vector in the training set M^k.
M_i: The i-th image features in all modalities, $M_i = \{\vec{m}_i^k | k = 1, ..., K\}$.
$\vec{y}_i$: The $d \times 1$ vector generates the observations M_i, with prior standard normal distribution.
$\vec{\beta}^k$: The estimated $D \times 1$ mean vector for observations in modality k.
I_d: The $d \times d$ identity matrix.

The goal of MCA is to learn a mapping function $\vec{F}^k$ so that we can infer the mutual component of the two different modalities. Suppose there are two sets of image features $M^1 = \{\vec{m}_1^1, \vec{m}_2^1, ..., \vec{m}_N^1\}$ and $M^2 = \{\vec{m}_1^2, \vec{m}_2^2, ..., \vec{m}_N^2\}$ of N pairs of subjects consisting of $D \times 1$ column vectors. They should be associated with the same set of mutual components $H = \{\vec{h}_1, \vec{h}_2, ..., \vec{h}_N\}$ consisting of the $d \times 1$ column vectors. The relationship between H and M can be modeled by a transmission channel denoted by a transmission function F. For k-th modality, there is a function from d-dimensional space to D-dimensional space $\vec{F}^k : R^{d \times 1} \rightarrow R^{D \times 1}$,

$$\vec{F}^k(x) = [\vec{F}_1^k(\vec{x}), \vec{F}_2^k(\vec{x}), ..., \vec{F}_D^k(\vec{x})]^T, \tag{1}$$

such that

$$\vec{m}_i^k = F^k(\vec{h}_i). \tag{2}$$

The MCA uses the prior knowledge: $H \sim N(\vec{\mu}, \Sigma)$ which is equivalent to $\vec{h}_i = \vec{\mu} + V\vec{y}_i$ where $\Sigma = VV^T, \vec{y}_i \sim N(\vec{0}, I)$. Then, Eq. (2) is equivalent to

$$\vec{m}_i^k = F^k(\vec{\mu} + V\vec{y}_i). \tag{3}$$

Note that the Eq.(3) can be approximated by a linear generative model. Denote two $D \times 1$ vectors $\vec{\beta}^k = \vec{F}^k(\vec{\mu})$ and $\vec{x}_i = V\vec{y}_i$. $\mathcal{J}_k(\vec{\mu})$ is the $D \times d$ Jacobian matrix with the element $\mathcal{J}_k(p, q) = \frac{\partial F_p^k(\vec{x})}{\partial x_q} |_{\vec{x}=\vec{\mu}}$, representing the partial derivative of the p-th component function in Eq. (1) w.r.t the q-th variable of $\vec{x}$. By using Gaussian random noise to model the high-order derivatives of the function $\vec{F}^k$, the MCA can be simplified as a linear generative model,

$$\vec{m}_i^k = \vec{\beta}^k + T^k\vec{y}_i + C^k\vec{z}, \tag{4}$$

where $T^k = \mathcal{J}_k(\vec{\mu})V$ is a $D \times d$ matrix that derives from the generative model, C^k is a $D \times D$ matrix with each column representing the bases of the noise, and $\vec{z}$ is the $D \times 1$ random vector whose components are independent standard normal random variables.

To estimate the model parameters, [1] proposes a Maximum A Posteriori (MAP)-based Expectation Maximization (EM) algorithm. Given modality-dependent observation $\vec{m}_i^k$, the aim is to infer the posterior distribution $P(\vec{y}_i \mid \vec{m}_i^k)$ of the hidden factor $\vec{y}_i$. With the EM algorithm (detailed derivation is in [1]), we get the conditional moment of hidden factor $\vec{y}_i$ as follows,

$$\vec{E}_{M_i}^1 = \frac{1}{K}\sum_{k=1}^{K} \vec{E}_{M_i}^{1,k}, \tag{5}$$

where the first moment $\vec{E}_{M_i}^{1,k}$ of hidden factor is given by

$$\vec{E}_{M_i}^{1,k} = (L^k)^{-1}(T^k)^T(\Sigma_0^k)^{-1}(\vec{m}_i^k - \vec{\beta}^k) \tag{6}$$

and $L^k = I_d + (T^k)^T(\Sigma^k)^{-1}T^k$. In Eq. (6), Σ_0^k is the initial estimation of Σ^k where $\Sigma^k = C^k(C^k)^T$. And $\vec{E}_{M_i}^1$ in Eq. (5) is used to estimate hidden factor which can be seen as mutual component of images in different modalities.

IV. Mutual component convolutional neural network

Inspired by [17], we aim to obtain modal-invariant features by incorporating the MCA module into a general deep neural network. In this section, we first show how to formulate the MCA as a fully-connected layer in a CNN model with a MCA loss, and then we present the used network architecture.

A. Key formulations of mutual component neural network

The MCA layer. A fully-connected layer is formulated by $f(\vec{x}) = W\vec{x} + \vec{b}$, where $\vec{x}$ is the input feature vector and $W, \vec{b}$ are learnable parameters. From Eq. (6), we find that MCA can be regarded as a FC layer. Let

$$W^k = (L^k)^{-1}(T^k)^T(\Sigma_0^k)^{-1}, \tag{7}$$

$$\vec{b}^k = -W^k\vec{\beta}^k \tag{8}$$

then we can rewrite Eq. (6) as:

$$\vec{E}_{M_i}^{1,k} = W^k\vec{m}_i^k + \vec{b}^k. \tag{9}$$

If we regard the feature vector $\vec{m}_i^k$ as the input feature $\vec{x}$ of the FC layer, then $\vec{E}_{M_i}^{1,k}$ in Eq. (9) is identical to the output of a FC layer. We regard this FC layer as the MCA layer. With all k modalities, we can get the conditional moments of the hidden factor in Eq. (5) which is the expected mutual component vector for a pair of facial images.

The MCA loss. The original MCA takes as input image pairs of same identities for mutual component feature extraction. However, feature extraction from single images is more practical for computer vision tasks since we always want to do off-line feature extraction in real-world applications. Therefore, we propose the MCA loss which enforces each modal-dependent component, i.e. $\vec{E}_{M_i}^{1,k}$ to approach the mutual component, i.e. $\vec{E}_{M_i}^1$. In other words, we hope the following equation can hold:

$$\vec{E}_{M_i}^{1,1} = \vec{E}_{M_i}^{1,2} = \vec{E}_{M_i}^1, \tag{10}$$

and then we can rewrite Eq. (5) as follows,

$$\vec{E}_{M_i}^1 = \frac{1}{2}(\vec{E}_{M_i}^{1,1} + \vec{E}_{M_i}^{1,2}) = \vec{E}_{M_i}^{1,1} = \vec{E}_{M_i}^{1,2}. \tag{11}$$

Eq. (11) shows an ideal case that the modal discrepancy is completely eliminated. However, there is no evidence that Eq. (11) can always be true. Inspired by [13], we turn to minimizing the distance between $\vec{E}^{1,k}_{M_i}$ and $\vec{E}^{1}_{M_i}$. Thus, the MCA loss can be formulated as,

$$\mathcal{L}_{mca} = \frac{1}{2NK}\sum_{i=1}^{N}\sum_{k=1}^{K}||\vec{E}^{1,k}_{M_i} - \vec{E}^{1}_{M_i}||_2^2, \tag{12}$$

where K is the number of modalities and N is the number of training images in modality k. Note that we input image pairs to the network and the total image number is $N \times K$.

We define a deep neural network with the MCA layer and the MCA loss as the Mutual Component Convolutional Neural Network (MC-CNN). There are several key points that should be noted. First, we use the prior knowledge of MCA model, and update the MCA layer with Eq. (7) and Eq. (8). The use of prior knowledge is able to alleviate over-fitting. Second, we compute the mutual component vector $\vec{E}^{1}_{M_i}$ only in training phase since the identity of an image pair is available. In test phase, we obtain $\vec{E}^{1,k}_{M_i}$ for the k-th modality. Finally, the MCA loss is used to optimize the whole network (except for the MCA layer) so that the modal discrepancy of input features $\vec{m}^k_i$ can be further reduced.

We compare our MCA layer to classical FC layer, our MC-CNN to CNN+MCA, and discuss how the MCA layer alleviate over-fitting as follows.

MCA layer vs. classical FC layer. The MCA layer is different from classical FC layer of CNN. The $W^k, \vec{b}^k$ are not optimized by SGD but by MCA algorithm, and these parameters are frozen when training other parameters in CNN so that the prior knowledge can be preserved. In this way, we can utilize MCA to reduce modal discrepancy, and also to put stronger constraint on FC layer to alleviate over-fitting.

MC-CNN vs. CNN+MCA. Firstly, our MC-CNN is optimized in an iterative way, i.e. SGD and EM. While CNN+MCA only takes CNN features as the input of the MCA. Secondly, we can NOT obtain the modal-invariant deep features for a single image by directly combining MCA and CNN. This is because MCA takes as input an image pair of the same identity from different modalities and CNN itself is not able to extract modal-invariant features effectively. By taking the proposed transformation and MCA loss, our MC-CNN can extract modal-invariant deep features for a single image in the testing stage, which is valuable in practice since off-line feature extraction is possible. Thirdly, we find that directly combining MCA and CNN achieves inferior performance. This observation drives us to seek a better method to embed MCA to CNN, i.e. MC-CNN.

Alleviate over-fitting with prior knowledge. The over-fitting problem in MCA has been relieved compared with CNN, partly because MCA uses prior knowledge. We consider the following regularizations/constraints as the prior knowledge in MCA. (1) MCA incorporates the rank constraint on $T^k = S + U^k$ to reduce the model complexity (details are shown in Section 2.2 of [1]). In Eq.(7), we can see that the weight W^k is a function of T^k. By embedding MCA into FC layer of CNN, we implicitly incorporate the rank constraint on the weights of FC layer to prevent over-fitting. (2) As shown in Eq.(8), the bias $\vec{b}^k$ is defined as a function of the weights W^k, which is also can be seen as a regularization. (3) The prior knowledge in the MCA layer, i.e. $H \sim N(\vec{\mu}, \Sigma)$, is with stronger hypothesis than CNN and acts as another regularization. (4) The MCA loss is also an additional regularization. These regularizations/constraints are seen as prior knowledge in MCA and useful to alleviate over-fitting.

B. The architecture of MC-CNN

The architecture of our MC-CNN is shown in Figure 1. The MC-CNN takes as input an image pair of the same identity from different modalities. It then processes them with two branches of the ResNet-41 [36] to extract deep features. The architecture of the used ResNet-41 is illustrated in the right of Figure 1. It consists of 41 convolutional layers and 4 max pooling layers. Specially, it includes four groups of ResNet-41 blocks with the number of blocks as 1, 2, 5, 10, and the number of channels as 64, 128, 256, 512, respectively. After the ResNet-41, feature maps of $512 \times 7 \times 6$ dimensionality are reshaped to vector $\vec{m}^k_i$ and further fed into the MCA layer with a batch normalization (BN) layer [37]. Finally, two expected d-dimensional feature vectors (i.e. $\vec{E}^{1,1}_{M_i}$ and $\vec{E}^{1,2}_{M_i}$) are generated.

Training and testing. We pre-train the ResNet-41 on a large visible face image dataset. For training MC-CNN, we need to learn two kinds of parameters, namely the parameters $(W^k, \vec{b}^k)$ of the MCA layer and the convolutional parameters θ_c of ResNet-41. For $(W^k, \vec{b}^k)$, we first reshape the last feature map f^{conv} of ResNet-41 to a vector $\vec{m}^k_i$ for each branch, and then compute the parameters of MCA module $(T^k, \Sigma^k, \vec{\beta}^k)$ as in [1], and finally compute $(W^k, \vec{b}^k)$ according to Eq. (7) and Eq. (8). Note $(W^k, \vec{b}^k)$ are exactly the parameters of the FC layer in Figure 1. For the training of θ_c, we first freeze the MCA layer $(W^k, \vec{b}^k)$, and then feed f^{conv} into the MCA layer with a BN layer and a dropout layer [39] to obtain two d-dimensional feature vectors for loss computation, and finally we use SGD to update θ_c. Specially, the total loss is defined as follows,

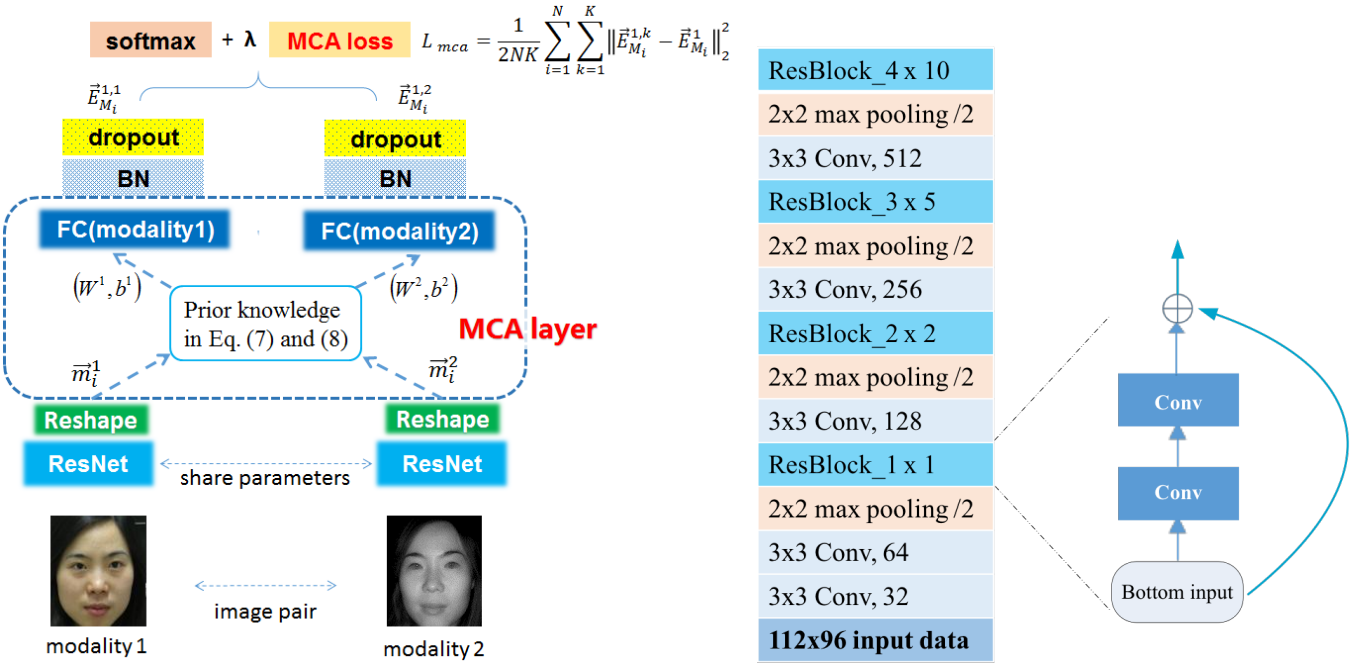
$$\mathcal{L} = \mathcal{L}_s + \lambda\mathcal{L}_{mca}, \tag{13}$$

where the first term is the cross-entropy loss of the softmax classifier, and the second term is the MCA loss in Eq. (12) and λ is a hyper parameter to trade off the two terms. We briefly summarize the training phase of MC-CNN in Algorithm 1.

For testing, a single face image of a certain modality is fed into the corresponding branch of the trained MC-CNN. The output feature vector $f = BN(\vec{E}^{1,k}_{M_i})$ from the BN layer is the expected modal-invariant feature vector. Then, we can use the modal-invariant features to compute the similarity, e.g. cosine similarity. For face identification, we adopt nearest neighbor classifier. For face verification, we use a threshold to decide whether two face images are of the same person or not.

V. Experiments

In this section, we first present the details of preprocessing and the pre-trained model and describe the implementation details of MC-CNN. We then conduct extensive experiments



(a) MC-CNN (b) ResNet-41

Fig. 1. The architecture of the proposed MC-CNN. The backbone model of MC-CNN is a deep ResNet model illustrated in (b). An image pair of the same identity from different modalities are put into the shared ResNet, and then the feature maps are reshaped into feature vectors which are further fed into the MCA layer. Note that we only need one branch for feature extraction in the test phase. In (b), all convolution layers are with 3*3 kernel size and stride of 1 and followed by PReLU [38] activation function. The number of channels of convolution layers gradually increase from 32 to 512. And 2*2 max pooling with stride of 2 is used.

Algorithm 1 Training Procedure of MC-CNN.

Input: Training data $\{\gamma_i^k\}$ (image pairs) with identity label.
Output: The parameters of FC layer $(W^k, \vec{b}^k)$ and other learnable parameters of CNN θ_c.
Initialization: Initialize all parameters of CNN θ_{all} using a pre-trained CNN model.

1: **While** not converge **do**:
2: Input image pairs to CNN and get convolutional features f^{conv}
3: Reshape f^{conv} to feature vector $\vec{m}_i^k$ and input $\vec{m}_i^k$ to the MCA model
4: Use EM algorithm to calculate the parameters of MCA $\theta_m = (T^k, \Sigma_0^k, \vec{\beta}^k)$.
5: Update $(W^k, \vec{b}^k)$ via Eq. (7) and Eq. (8).
6: Fix $(W^k, \vec{b}^k)$, compute the loss in Eq. (13) and update θ_c by SGD.
7: **End while**

on CASIA NIR-VIS 2.0 [32], CUHK VIS-NIR, IIIT-D Sketch [40], and CUHK Face Sketch (CUFS) [41]. We further evaluate the CNN architecture on CASIA NIR-VIS 2.0 and IIIT-D Sketch, then evaluate hype parameters of MC-CNN on CASIA NIR-VIS 2.0, and finally compare to the state of the art.

A. Experimental settings

1) Preprocessing: We first use MTCNN [42] with default setting to detect faces and obtain the five landmarks (nose, two eyes, mouse corners) of each face. We then apply these landmarks and similarity transform to align and crop faces. Each face is cropped to 112×96. Figure 2 shows some cropped examples on CASIA NIR-VIS 2.0, CUHK NIR-VIS, CUFS and IIIT-D Sketch. To normalize the images, each pixel ([0,255]) in RGB channels is subtracted by 127.5 and then divided by 128.

2) Pre-trained models: We use a pre-trained model to initialize the shared part (i.e. ResNet-41) of our MC-CNN. This model is trained on a large dataset which is the combination of CASIA-WebFace [43], CACD2000 [44], Celebrity+ [45], MSRA-CFW [46], MS-Celeb-1M [47]. For the MS-Celeb-1M dataset, we use the cleaned version provided by [48]. We initialize all the weights with a normal Gaussian distribution and the cross-entropy loss and the center loss [13] with SGD are adopted to train the model. The initial learning rate is set to 0.05 and is reduced by half if the loss stops to decrease. The model is trained on two GPUs with batch size of 256, and takes about 16 days to converge.

3) Implementation details of MC-CNN: The first step for training MC-CNN is to generate image pairs. Given a person with N_1 VIS images $\{\mathrm{I}_i^1; i = 1, 2, ..., N_1\}$ and N_2 NIR images $\{\mathrm{I}_j^2; j = 1, 2, ..., N_2\}$, we generate $N_1 \times N_2$ image pairs by considering all possible pairs. With this strategy, we generate about 440,000 image pairs on CASIA NIR-VIS 2.0, 1,438 on CUHK NIR-VIS, 1195 from CUFSF [49] on IIIT-D Sketch and 306 on CUFS for training. On both NIR-VIS and Sketch datasets, we learn the ResNet-41 and the MCA layer alternately. For training the ResNet-41, we update the

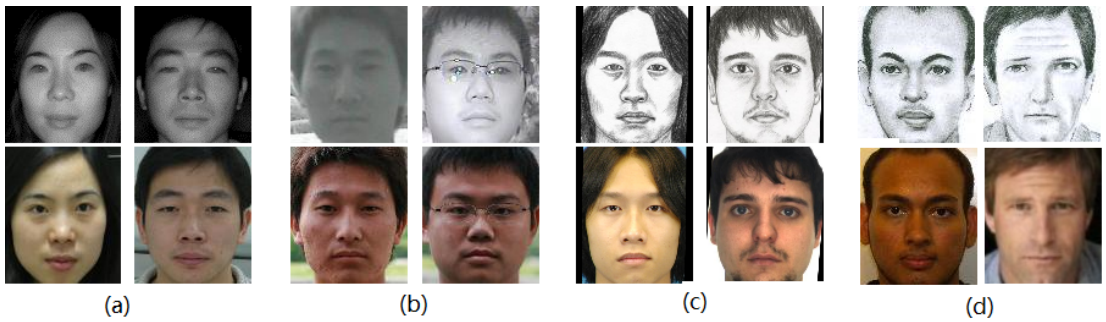


Fig. 2. Cropped face images from (a) CASIA NIR-VIS 2.0, (b) CUHK VIS-NIR, (c) CUFS and (d) IIIT-D Sketch.

top 2 layers of ResNet-41 on NIR-VIS datasets and set the initial learning rate to 0.001. On Sketch datasets, we set the initial learning rate to 0.01, and update several more bottom layers because the discrepancy of low level features may be larger than NIR-VIS data. The learning rate is divided by 10 at 8k, 13k and 17k iterations. At the $20k$-th iterations, we update the MCA layer and then reset the learning rate to the initial learning rate to continue updating the ResNet-41. These two stages are conducted alternately until convergence. The batch size is set to 128, i.e. 64 image pairs and dropout ratio is set to 0.2. And we use the images and its horizontally flipped counterpart to train our models. All experiments are implemented based on the Caffe toolbox [50] and its Matlab interface.

B. Experiments on CASIA NIR-VIS 2.0

The CASIA NIR-VIS 2.0 dataset is the largest HFR face database across NIR and VIS spectrum so far. It contains 17,580 images of 725 subjects which exhibit intra-personal variations such as pose and expression. This database includes two views: view 1 for parameter tuning and view 2 for performance evaluation. During test, the gallery and probe images are VIS and NIR images respectively, simulating the scenario of face recognition in the dark environment. The rank-1 identification rate and verification rate(VR)@false acceptance rate (FAR) including the mean accuracy and standard deviation of 10 folds are reported. Following the protocols, we train MC-CNN on view 1 and report results on view 2.

Table I shows the comparison of different methods on CASIA NIR-VIS 2.0. We provide five baselines in the first five rows. The first one comes from the pre-trained ResNet-41 where face representations are the output of a 512-D FC layer. A test on the LFW dataset [52] with this model obtains 99.68% in accuracy which demonstrates its good performance for visible faces. However, the pre-trained model achieves 88.02% in rank-1 accuracy only which indicates there is a gap between VIS and NIR images. The second row shows a straightforward transfer learning method by fine-tuning the top 2 layers of ResNet-41 and the 512-D FC layer of the pre-trained model. Specifically, we combine both modal images for each subject, and fine-tune the model with multi-class cross-entropy loss. Batch normalization and dropout are also used for fine-tuning. The fine-tuned model improves the original one by 9.32%. We argue that the fine-tuning operation mainly reduces modal discrepancy by the subject identities. We also fine-tune ResNet-41 by replacing the 512-D FC layer with two FC layers and it slightly degrades the performance. This may be because fine-tuning with two FC layers can make the over-fitting problem more serious.

TABLE I
RANK-1 ACCURACY (%) ON CASIA NIR-VIS 2.0.

Model	Hyper parameter	Rank-1
Pre-trained ResNet-41	-	88.02 ± 1.36
Fine-tuned ResNet-41 with FC	$d = 512$	97.35 ± 0.40
Fine-tuned ResNet-41 with two FC layers	$d = 512$	97.13 ± 0.35
Pre-trained ResNet-41 with FC + MCA*	$d = 512$	91.96 ± 0.97
Fine-tuned ResNet-41 with FC + MCA*	$d = 512$	94.82 ± 0.78
Fine-tuned ResNet-41 (FC) + CenterLoss [13]	$d = 512, \lambda = 0.04$	97.84 ± 0.37
Fine-tuned ResNet-41 (FC) + Euclidean loss	$d = 512, \lambda = 0.01$	98.92 ± 0.27
Fine-tuned ResNet-41 (FC) + AM-softmax [51]	$d = 512, s = 30, m = 0.15$	97.38 ± 0.46
MC-CNN	$d = 512, \lambda = 0.1$	$\mathbf{99.22 \pm 0.20}$

*It conducts classification on mutual components of image pairs, and can not extract modal-invariant features for individual images.

Since our MC-CNN learns the parameters of MCA module and the ResNet-41 jointly, we conduct another two experiments which use deep CNN features and the MCA model separately for fair comparison. Based on the features of the pre-trained ResNet-41, the MCA model achieves 91.96% (the 4th row) which is comparable to the fine-tuned model. This indicates that i) the MCA model can explicitly decrease modal discrepancy without fine-tuning and ii) the prior knowledge used in MCA is credible. The MCA model with fine-tuned features even gets worse performance than the fine-tuned model and only obtains 94.82% (the 5th row), which indicates that directly combining fine-tuning operation and the MCA model is not a good choice.

Considering the loss function, we have two terms, namely the cross-entropy loss and the MCA loss. For fair comparison, we implement the popular center loss [13] and Euclidean loss methods which also have two loss terms, i.e. the cross-entropy loss and the Euclidean/center loss. The center loss term aims

to reduce the distance between every sample (regardless of modality) and its class center within the same identity. Based on the same pre-trained model, we test several trade-off factors empirically and show the best one in the 6th row of Table I. This achieves 97.84% which is slightly better than the one without the center loss and is similar with the 2nd baseline. From the 7th row, we can observe that Euclidean loss obtains better performance than the 2nd baseline and the center loss. We also compare with another widely used loss function in face recognition, i.e. AM-softmax [51]. We empirically tune the margin (m) and scale (s) parameters of AM-softmax and show the best model in the 8th row. Though AM-softmax gets subtle improvement over the 2nd baseline, it fails to surpass center loss. Compared to the center loss, Euclidean loss and AM-softmax methods, our MC-CNN shows superior performance.

Figure 3 shows the receiver operating characteristic (ROC) curves of MC-CNN and different baseline methods in Table I. From Figure 3, we can also get similar conclusions as Table I except that adding MCA to the pre-trained or fine-tuned ResNet-41 can not improve the verification accuracy. This observation shows that directly combining CNN with MCA is not a good choice. Overall, MC-CNN works best in the verification task on CASIA NIR-VIS 2.0.

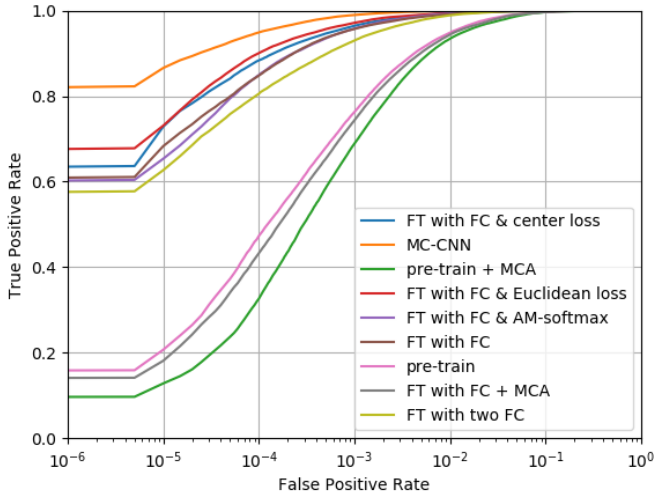


Fig. 3. ROC curves of MC-CNN and different baseline methods. 'pre-train' means the pre-trained ResNet-41 and 'FT' means fine-tuned ResNet-41.

C. Experiments on CUHK NIR-VIS

The CUHK VIS-NIR face dataset consists of 2,876 different persons with each one having a pair of infrared and visible facial images. Following [1], we use 1,438 pairs of infrared and visible facial images as the training set and the remaining 1,438 pairs as the testing set. The results of different baselines and our MC-CNN are shown in Table II. In rank-1 accuracy, the pre-trained model gets 92.8%, and the fine-tuned model improves it to 97.98%. With the features of pre-trained model, the MCA model achieves 97.77% (the 3rd row) which is comparable to the fine-tuned model. These results confirm our observations on the CASIA NIR-VIS 2.0 dataset that the MCA model can explicitly decrease modal discrepancy without fine-tuning. The MCA model with fine-tuned features obtains 98.54% (the 4th row) which shows that fine-tuning operation and the MCA model are complementary on CUHK NIR-VIS. This observation is different from that on CASIA NIR-VIS 2.0. This may be because CUHK NIR-VIS includes more identities (2,876 vs. 725) and more identities can lead to a harder task for a fine-tuned ResNet-41. In this case, MCA may be helpful for the fine-tuned model to get a better performance. For this reason, more parameter in networks can also contribute to a better performance of the fine-tuned model, i.e. when fine-tuning ResNet-41 by replacing the 512-D FC layer with two FC layers, the rank-1 accuracy improves to 98.82%. With joint training for MCA layer and the MCA loss, our MC-CNN outperforms all the other models with a rank-1 accuracy of 99.44%. It is worth noting that the result of our MC-CNN is based on features of single images while the 3rd and 4th baselines are based on features of image pairs.

TABLE II
RANK-1 ACCURACY (%) ON CUHK NIR-VIS FACE DATASET.

Model	Hyper parameter	Rank-1
Pre-trained ResNet-41	-	92.80
Fine-tuned ResNet-41 with FC	$d = 512$	97.98
Pre-trained ResNet-41 with FC + MCA*	$d = 512$	97.77
Fine-tuned ResNet-41 with FC + MCA*	$d = 512$	98.54
Fine-tuned ResNet-41 with two FC layers	$d = 512$	98.82
MC-CNN	$d = 512, \lambda = 0.1$	**99.44**

*It conducts classification on mutual components of image pairs, and can not extract modal-invariant features for individual images.

D. Experiments on IIIT-D Sketch

To illustrate the effectiveness of our MC-CNN, we further test on a more challenging IIIT-D Sketch dataset which is with larger modal discrepancy. The viewed sketch-photo subset of IIIT-D Sketch dataset comprises 238 sketch-photo image pairs. The sketches are drawn by a professional sketch artist for digital images collected from the FG-NET aging database (67 pairs), Labeled Faces in Wild (LFW) database (99 pairs) and IIIT-D student & staff database (72 pairs). On this dataset, we follow the same training and test protocols of [2]: training on CUFSF [49] dataset and then rank-1 accuracy probe-gallery face identification testing on IIIT-D Sketch dataset is reported. When training MC-CNN on CUFSF dataset, we make all layers below 'ResBlock_2x2' and several layers in 'ResBlock_2x2', 'ResBlock_3x5' learnable.

In Table III, we compare our MC-CNN with other baseline methods on IIIT-D Sketch dataset. From Table III, we have the following observations: (1) The pre-trained model only gets a poor rank-1 accuracy of 43.28%, indicating that the modal discrepancy of IIIT-D Sketch dataset is much larger than NIR-VIS images. (2) Fine-tuning ResNet-41 with a classical FC layer (the 2nd row) can improve the performance to 71.43%

TABLE III
RANK-1 ACCURACY (%) ON IIIT-D SKETCH FACE DATASET.

Model	Hyper parameter	Rank-1
Pre-trained ResNet-41	-	43.28
Fine-tuned ResNet-41 with FC	$d = 512$	71.43
Fine-tuned ResNet-41 with two FC layers	$d = 512$	61.34
Fine-tuned ResNet-41 with two FC layers + MCA loss	$d = 512, \lambda = 0.15$	72.27
MC-CNN	$d = 512, \lambda = 0.15$	**84.45**

but the result is hardly satisfying. This may be because the model can easily over-fit such a small scale dataset. (3) Fine-tuning ResNet-41 by replacing the 512-D FC layer with two FC layers gets the rank-1 accuracy of 61.34% (3rd row), even worse than fine-tuned ResNet-41 with FC (2nd row, 71.43%). We owe the degradation of the performance to the fact that fine-tuning with two FC layers can make the problem of over-fitting more serious. From the comparison, we can infer that FC layer does have a great impact on the performance when training a deep CNN model with very limited training samples. This observation also motivates us to treat FC layers as a MCA layer and update the parameters of FC layers by MCA instead of back propagation (BP) to alleviate over-fitting. (4) Fine-tuning ResNet-41 with two FC layers and MCA loss can get the rank-1 accuracy of 72.27% (4th row) which is much better than the model without MCA loss (61.34%, 3rd rows). It proves the effectiveness of our proposed MCA loss in reducing modal discrepancy. (5) Our MC-CNN achieves the best performance of 84.45% among those baseline methods. The significant improvement (12.18%) over the method of fine-tuning with two FC layers and MCA loss (72.27%, 4th row) illustrates the fact that exploiting the prior knowledge of MCA, e.g. the generative hypothesis and rank constraint on weights, to update 'MCA layer' (FC layers) can alleviate over-fitting. The reason may be that the number of parameters in FC layers dominates in a CNN model and the FC layer trained by BP would get higher risk of over-fitting due to the small training dataset used.

From the above analyses, we can see that our proposed MCA layer and MCA loss can effectively alleviate over-fitting and reduce the modal discrepancy, which makes the MC-CNN surpass the other baseline models significantly.

E. Experiments on CUFS

The CUFS dataset includes 606 subjects, totally 1,212 face images. We also train models on 306 subjects and test on the rest 300 subjects as [41], [49], [53]. Here, we only learn the parameters of top 2 layers of ResNet41.

Table IV shows the rank-1 accuracy of different baseline methods. From Table IV, we can see that the pre-trained model (81%, 1st row) can hardly work well on CUFS, and fine-tuning with FC layer (93%, 2nd row) works better than the pre-trained model as well as fine-tuning with two FC layers (92.67%, 3rd row). Fine-tuning with two FC layers and MCA loss also improves the rank-1 accuracy, which proves the effectiveness of MCA loss. Moreover, MC-CNN can achieve 97.33% rank-1 accuracy which outperform the pre-trained model and fine-tuned ResNet-41 with two FC layers significantly. These results are consistent with the observations on the IIIT-D Sketch dataset. What's more, our MC-CNN also works better than fine-tuning with center loss.

TABLE IV
RANK-1 ACCURACY (%) ON CUFS FACE DATASET.

Model	Hyper parameter	Rank-1
Pre-trained ResNet-41	-	81
Fine-tuned ResNet-41 with FC	$d = 512$	93
Fine-tuned ResNet-41 with two FC layers	$d = 512$	92.67
Fine-tuned ResNet-41 with two FC layers + CenterLoss	$d = 512, \lambda = 0.04$	94.67
Fine-tuned ResNet-41 with two FC layers + MCA loss	$d = 512, \lambda = 0.2$	94.67
MC-CNN	$d = 512, \lambda = 0.6$	**97.33**

F. Exploratory experiments

1) Effect of CNN architectures: We also train two CNN models with 10 and 21 convolutional layers as our backbone models. These two backbone models are shown in Figure 4 and trained with the same settings as described in Sec. V-A2. Table V shows the accuracy of different CNN architectures on LFW and we can observe that CNN with deeper architectures can get better performance on LFW.

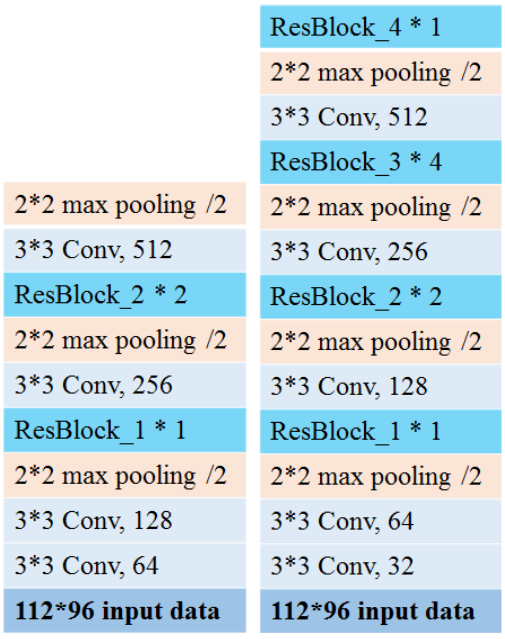


Fig. 4. CNN architectures. Left: Architecture of ResNet-10. Right: Architecture of ResNet-21

Figure 5 shows the accuracy of different CNN architectures on different heterogeneous face datasets. From Figure 5, we have following observations. Firstly, the pre-trained model with deeper architectures does not always perform better on

TABLE V
ACCURACY OF DIFFERENT CNN ARCHITECTURES ON LFW.

Model	ResNet-10	ResNet-21	ResNet-41
Accuracy(%)	99.48	99.58	99.68

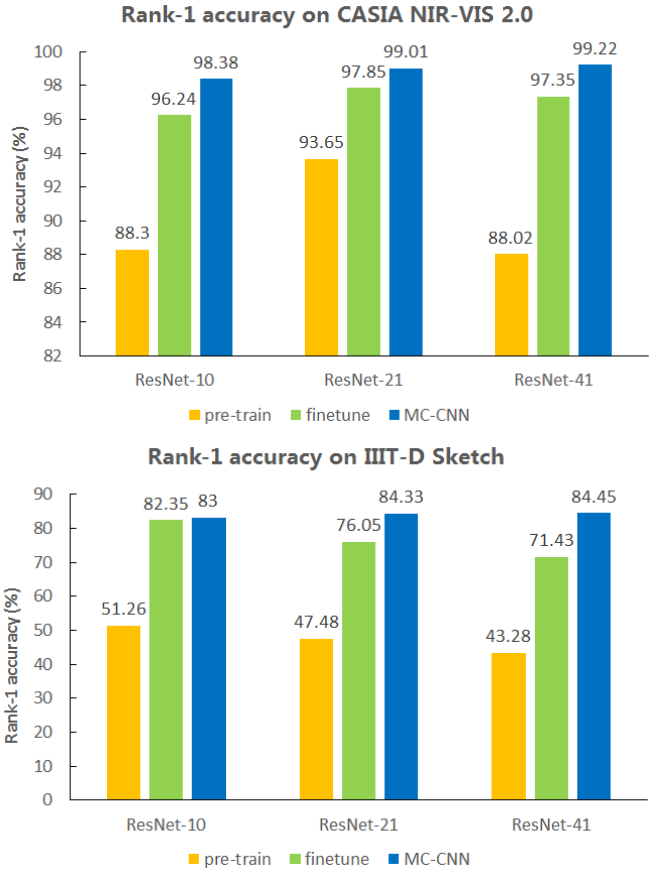


Fig. 5. Rank-1 accuracy of different CNN architectures on CASIA NIR-VIS 2.0 and IIIT-D Sketch datasets. Top: Rank-1 accuracy on CASIA NIR-VIS 2.0. Bottom: Rank-1 accuracy on IIIT-D Sketch.

CASIA NIR-VIS 2.0 and IIIT-D Sketch dataset. Especially on IIIT-D Sketch face dataset, the rank-1 accuracy decreases from 51.26% to 43.28% when the number of convolutional layers increases from 10 to 41. Deeper CNN gets poorer performance because it may tend to over-fit the VIS face recognition task more easily. Secondly, fine-tuning the pre-trained model can improve the accuracy on both datasets while deeper architectures also fail to show its superior learning capability, e.g. ResNet-21 of 97.85% vs. ResNet-41 of 97.35% on CASIA NIR-VIS 2.0 and ResNet-10 of 82.35% vs. ResNet-21 of 76.05% on IIIT-D Sketch. Thirdly, the rank-1 accuracy of MC-CNN with different architectures consistently outdoes the pre-trained and fine-tuned models both on CASIA NIR-VIS 2.0 and IIIT-D Sketch datasets. What's more, different from pre-trained or fine-tuned models, MC-CNN with deeper architectures can get better performance on these two datasets, which illustrates that MC-CNN can take full advantage of the superior learning capability of deeper architectures.

While CDL [2] adopts Lightened CNN-9, we use a backbone model with similar convolutional layers, i.e. ResNet-10, to make a comparison. MC-CNN of ResNet-10 gets 2.14% improvement over fine-tuned ResNet-10 (98.38% vs. 96.24%) and 10.08% improvement over the pre-trained one (98.38% vs. 88.30%) on CASIA NIR-VIS 2.0 while CDL gets 1.47% over its fine-tuned Lightened CNN-9 model (98.62% vs. 97.15%) and 6.74% over its pre-trained one (98.62% vs. 91.88%). On IIIT-D Sketch, MC-CNN of ResNet-10 gets 31.74% over the pre-trained model (83% vs. 51.26%) while CDL gets 1.28% over the pre-trained Lightened CNN-9 (85.35% vs. 84.07%). It is easily to see that the gain from MC-CNN is much more significant than CDL on IIIT-D Sketch, which shows the superior capability of our MC-CNN to deal with the heterogeneous face recognition task.

2) Effect of λ and d: There are two hyper parameters in our MC-CNN model, namely the dimension of the MCA layer d and the trade-off factor λ for the MCA loss. We use ResNet-41 as the backbone model of MC-CNN and evaluate λ, d on the CASIA NIR-VIS 2.0 dataset. For the evaluation of λ, we fix d as the default value. The top of Figure 6 shows the results of varied λ, i.e. 0, 0.05, 0.1, 0.2, 0.4, and 1. Enlarging λ from 0 to 0.1 improves performance significantly, 99.22% at $\lambda = 0.1$ vs. 92.6% at $\lambda = 0$, while a higher value degrades performance consistently. This can be explained by that the MCA loss is designed to reduce the intra-class modal discrepancy and a higher ratio might be harmful for the inter-class discriminability. For the evaluation of d, we fix λ to 0.1 and increase d from 32 to 1024. The results are shown in the bottom of Figure 6. From the evaluation, we observe that the dimension of MCA layer has limited impact on the final performance. We achieve the best accuracy at $d = 64$, i.e. 99.39%.

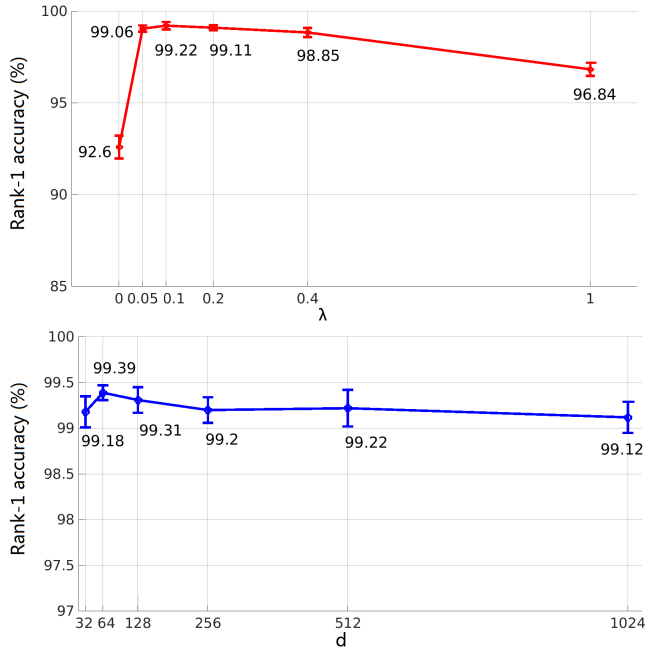


Fig. 6. Evaluation of λ and d on CASIA NIR-VIS 2.0. Top: the Rank-1 accuracy of varied λ with $d = 512$. Bottom: the Rank-1 accuracy of varied d with $\lambda = 0.1$.

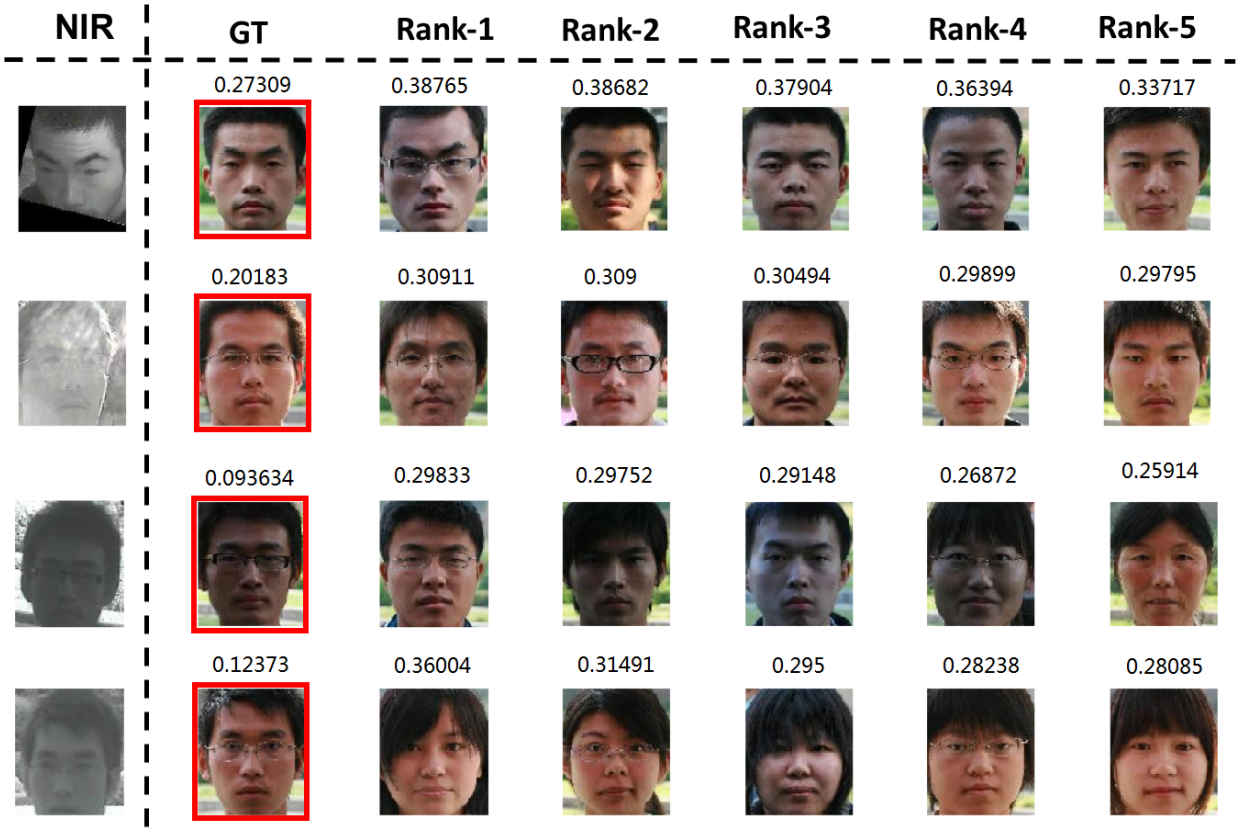


Fig. 7. False alarms on CUHK NIR-VIS. *There are totally 8 false alarms on CUHK NIR-VIS, and four of them are due to the false detection of the MTCNN face detector.

G. Visualization

To examine the limitations of our MC-CNN, we visualize some false alarms from the best models on both CUHK NIR-VIS and CASIA NIR-VIS 2.0, see Figure 7 and 8. We illustrate the probe NIR face images with their top 5 matching visible faces and their groundtruth visible faces. For the CUHK NIR-VIS dataset, we totally have eight false alarms in Rank-1, and we find four of them are caused by the false detection of the used face detector. With the rest false cases in Figure 7, we observe that those false alarms mainly come from extreme illumination in the wild (see row 2 and 3), false alignment (see row 4), and pose in pitch direction (see row 1). Actually, these issues are indeed the difficulties of general face recognition. For the CASIA NIR-VIS 2.0 dataset which is collected with indoor environment, we conclude the main difficulties are large pose in pitch direction, wearing glasses, and varied facial expression from Figure 8. We do not find any false alarms that caused by modal discrepancy after checking all of them. This indicates that our MC-CNN deals with modal discrepancy effectively and casts the key challenges for further research.

H. Comparison to the state of the art

We conclude the experimental evaluation with a comparison to the state of the art in Table VI ,VII ,VIII and IX. For CASIA NIR-VIS 2.0, all the methods of [2], [6], [7], [34], [48] are based on deep neural networks. Without fine-tuning deep models on the View 1 of CASIA NIR-VIS 2.0, [34] and [2] obtain the rank-1 accuracy of 79.01% and 91.88%, respectively. [2] fine-tunes the base model by adding a relevance constraint and a cross-modal triplet loss, and improves the baseline by 6.74%. Our pre-trained ResNet-41 gets 88.02% and the MC-CNN improves it to 99.39%. Besides, MC-CNN also works well in face verification task and gets 99.27% VR@FAR=0.1%, which is much better than other state-of-the-art methods. For CUHK NIR-VIS, most of the previous methods are based on hand-crafted features. Compared to the straightforward fine-tuning method, our MC-CNN jointly learns MCA and CNN with an extra MCA loss, and it outperforms the fine-tuning one by 1.46%.

TABLE VI
COMPARISON TO THE STATE OF THE ART ON CASIA NIR-VIS 2.0 FACE DATASET.

Method	Rank-1 (%)	VR@FAR=0.1% (%)
Light CNN (2015) [48]	96.72 ± 0.23	94.77 ± 0.43
Shared, Inter+Intra (2016) [6]	85.9 ± 0.9	78
TRIVET (2016) [34]	95.74 ± 0.52	91.03 ± 1.26
G-HFR (2017) [5]	85.3 ± 0.03	-
Gabor + HJB (2017) [3]	91.65 ± 0.89	89.91 ± 0.97
CDL (2017) [2]	98.62 ± 0.20	98.32 ± 0.05
Synthetic + CNN (2018) [7]	85.05 ± 0.83	-
MC-CNN	$\mathbf{99.39 \pm 0.08}$	$\mathbf{99.27 \pm 0.14}$

Table VIII shows the comparison of different state-of-the-art methods on the more challenging IIIT-D Sketch face

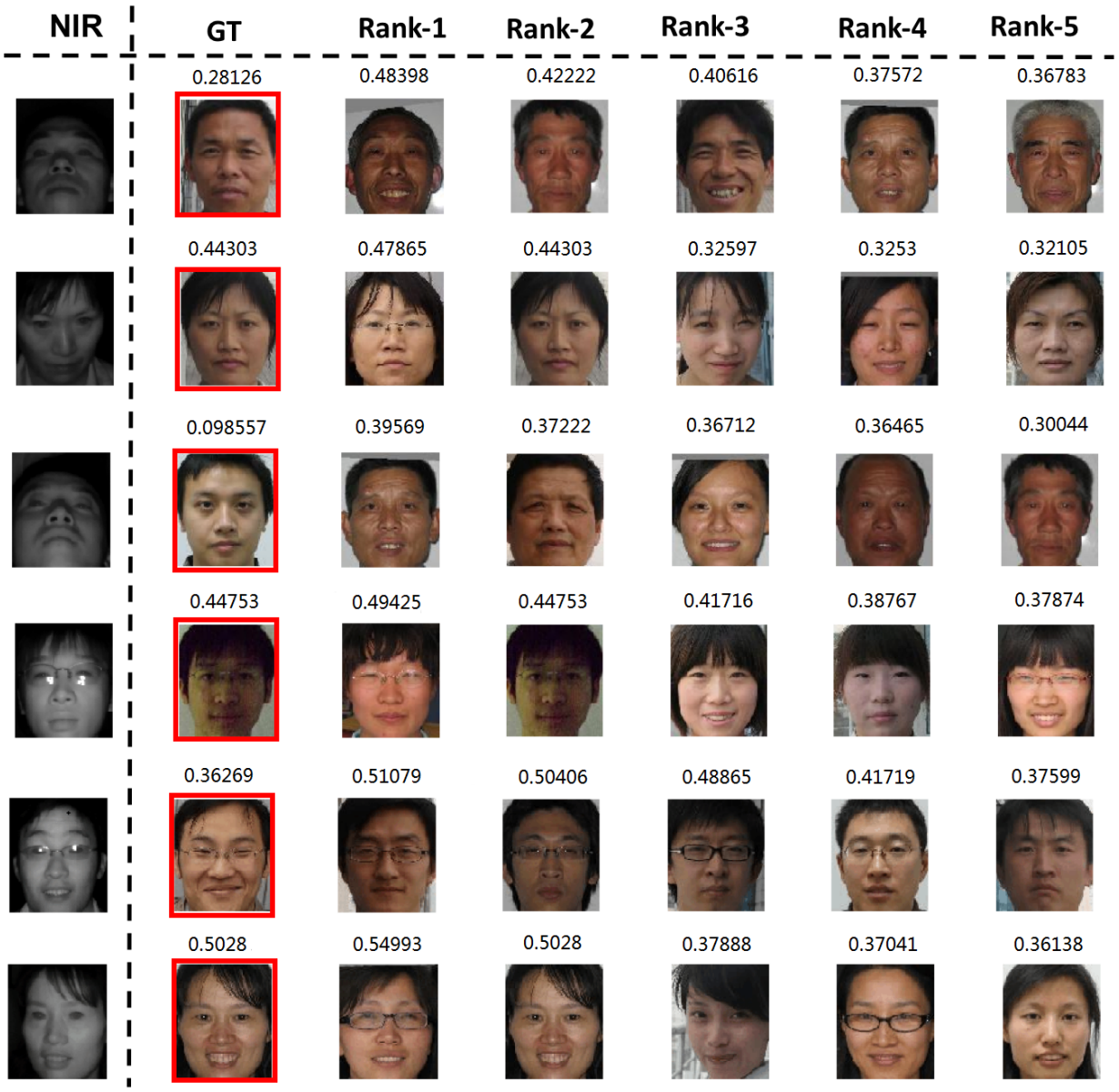


Fig. 8. Some false alarms on CASIA NIR-VIS 2.0.

TABLE VII
COMPARISON TO THE STATE OF THE ART ON CUHK NIR-VIS FACE DATASET.

Method	Rank-1 accuracy(%)
P-RS (2013) [25]	75.10
CFDA (2014) [8]	80.19
MCA (2016) [1]	86.43
CEFD (2017) [54]	83.93
Fine-tuned ResNet-41	97.98
MC-CNN	**99.44**

TABLE VIII
COMPARISON TO THE STATE OF THE ART ON IIIT-D SKETCH FACE DATASET.

Method	Rank-1 accuracy (%)
SIFT (2012) [55]	76.28
MCWLD (2012) [55]	84.24
VGG (2015) [56]	80.89
Light CNN (2015) [48]	84.07
CenterLoss (2016) [13]	84.07
CDL (2017) [2]	85.35
MC-CNN	**87.40**

dataset. From the table VIII, we can see that MC-CNN with the setting of $\lambda = 0.15, d = 128$ gets 2.05% rank-1 accuracy higher than CDL, 3.16% higher than MCWLD and 3.3% higher than CenterLoss and Light CNN.

On CUFS face dataset, we set $\lambda = 1, d = 128$ and make top 2 layers and several bottom layers learnable. Table IX shows the rank-1 accuracy of different state-of-the-art methods on CUFS. We can see from table IX that MC-CNN surpasses Sketch Synthesis [57] method and get a similar accuracy

to SIFT + LMBP [53] and CDL [2]. The similar accuracy between MC-CNN and SIFT + LMBP may be because the over-fitting problem of CNN can not be completely solved on the very small scale training dataset with only 306 image pairs. Unlike CDL which uses additional CUFSF (1195 pairs) dataset as its training set, we only use 306 pairs to train our model. It is worth noting that the training samples of CDL are about 4 times of MC-CNN and CDL only gets same accuracy as MC-CNN, which implies that MC-CNN is more efficient than CDL when dealing with the issue of training CNN on limited samples with large modal discrepancy.

TABLE IX
COMPARISON TO THE STATE OF THE ART ON CUFS FACE DATASET.

Method	Rank-1 accuracy (%)
Sketch Synthesis (2003) [57]	96.30
SIFT + LMBP (2011) [53]	99.47
CDL (2017) [2]	99.47
MC-CNN	**99.47**

VI. CONCLUSION

This paper introduces a Mutual Component Convolutional Neural Network (MC-CNN) for heterogenous face recognition, which takes full advantage of a generative model (i.e. MCA) and a CNN model. This MC-CNN embeds the MCA module into the CNN by viewing MCA as a special full-connected layer, termed as MCA layer. We alternately optimize the MCA layer and the CNN model with a novel MCA loss. Compared to the original MCA model, our method can extract modal-invariant features for single images. Extensive experiments on CASIA NIR-VIS 2.0, CUHK NIR-VIS, IIIT-D Sketch and CUFS show that our MC-CNN not only improves the pre-trained CNN model significantly but also outperforms i) the fine-tuned CNN model and ii) the MCA model based on CNN features. Our MC-CNN achieves the state-of-the-art performance on CASIA NIR-VIS 2.0, CUHK NIR-VIS, IIIT-D Sketch datasets.

REFERENCES

[1] Z. Li, D. Gong, Q. Li, D. Tao, and X. Li, "Mutual component analysis for heterogeneous face recognition," *ACM Transactions on Intelligent Systems and Technology*, vol. 7, no. 3, p. 28, 2016.

[2] X. Wu, L. Song, R. He, and T. Tan, "Coupled deep learning for heterogeneous face recognition," *CoRR*, vol. abs/1704.02450, 2017. [Online]. Available: http://arxiv.org/abs/1704.02450

[3] H. Shi, X. Wang, D. Yi, Z. Lei, X. Zhu, and S. Z. Li, "Cross-modality face recognition via heterogeneous joint bayesian," *IEEE Signal Processing Letters*, vol. 24, no. 1, pp. 81–85, Jan 2017.

[4] M. S. Sarfraz and R. Stiefelhagen, "Deep perceptual mapping for thermal to visible face recognition," *CoRR*, vol. abs/1507.02879, 2015. [Online]. Available: http://arxiv.org/abs/1507.02879

[5] C. Peng, X. Gao, N. Wang, and J. Li, "Graphical representation for heterogeneous face recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 2, pp. 301–312, 2017.

[6] S. Saxena and J. Verbeek, "Heterogeneous face recognition with CNNs," in *ECCV*, 2016, pp. 483–491.

[7] G. Hu, X. Peng, Y. Yang, T. M. Hospedales, and J. Verbeek, "Frankenstein: Learning deep face representations using small data," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 293–303, 2018.

[8] Z. F. Li, D. Gong, Y. Qiao, and D. Tao, "Common feature discriminant analysis for matching infrared face images to optical face images." *IEEE Transactions on Image Processing*, vol. 23, no. 6, pp. 2436–2445, 2014.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[10] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *CVPR*, 2014, pp. 1891–1898.

[11] Y. Chen, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *NIPS*, 2014, pp. 1988–1996.

[12] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *CVPR*, 2015, pp. 815–823.

[13] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *ECCV*, 2016, pp. 499–515.

[14] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[15] Y. Wang, D. Gong, Z. Zhou, X. Ji, H. Wang, Z. Li, W. Liu, and T. Zhang, "Orthogonal deep features decomposition for age-invariant face recognition," in *ECCV*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 764–779.

[16] C. Reale, N. M. Nasrabadi, H. Kwon, and R. Chellappa, "Seeing the forest from the trees: A holistic approach to near-infrared heterogeneous face recognition," in *CVPR Workshops*, 2016, pp. 320–328.

[17] Y. Wen, Z. Li, and Y. Qiao, "Latent factor guided convolutional neural networks for age-invariant face recognition," in *CVPR*, 2016, pp. 4893–4901.

[18] Y. Yan, Q. Wu, M. Tan, M. K. Ng, H. Min, and I. W. Tsang, "Online heterogeneous transfer by hedge ensemble of offline and online decisions," *IEEE Transactions on Neural Networks & Learning Systems*, vol. 29, no. 7, pp. 3252–3263, 2018.

[19] Y. Yan, W. Li, H. Wu, H. Min, M. Tan, and Q. Wu, "Semi-supervised optimal transport for heterogeneous domain adaptation," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, 2018, pp. 2969–2975. [Online]. Available: https://doi.org/10.24963/ijcai.2018/412

[20] Q. Wu, H. Wu, X. Zhou, M. Tan, Y. Xu, Y. Yan, and T. Hao, "Online transfer learning with multiple homogeneous or heterogeneous sources," *IEEE Transactions on Knowledge & Data Engineering*, vol. PP, no. 99, pp. 1–1, 2017.

[21] Y. Yan, W. Li, M. Ng, M. Tan, H. Wu, H. Min, and Q. Wu, "Learning discriminative correlation subspace for heterogeneous domain adaptation," in *Twenty-Sixth International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann*, 2017, pp. 3252–3258.

[22] T. d. F. Pereira and S. Marcel, "Heterogeneous face recognition using inter-session variability modelling," in *CVPR Workshops*, 2016, pp. 179–186.

[23] C. Peng, X. Gao, N. Wang, and J. Li, "Sparse graphical representation based discriminant analysis for heterogeneous face recognition," *CoRR*, vol. abs/1607.00137, 2016. [Online]. Available: http://arxiv.org/abs/1607.00137

[24] S. Liu, D. Yi, Z. Lei, and S. Z. Li, "Heterogeneous face image matching using multi-scale features," in *International Conference on Biometrics*, 2012, pp. 79–84.

[25] B. F. Klare and A. K. Jain, "Heterogeneous face recognition using kernel prototype similarities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1410–1422, Jun. 2013. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2012.229

[26] C. A. Hou, M. C. Yang, and Y. C. F. Wang, "Domain adaptive self-taught learning for heterogeneous face recognition," in *ICPR*, 2014, pp. 3068–3073.

[27] Y. H. Tsai, H. M. Hsu, C. A. Hou, and Y. C. F. Wang, "Person-specific domain adaptation with applications to heterogeneous face recognition," in *IEEE International Conference on Image Processing (ICIP)*, Oct 2014, pp. 338–342.

[28] X. Huang, Z. Lei, M. Fan, X. Wang, and S. Z. Li, "Regularized discriminative spectral regression method for heterogeneous face matching," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 353–362, 2012.

[29] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[30] D. Lin and X. Tang, "Inter-modality face recognition," in *ECCV*. Springer, 2006, pp. 13–26.
[31] D. Yi, S. Liao, Z. Lei, J. Sang, and S. Z. Li, "Partial face matching between near infrared and visual images in mbgc portal challenge," in *International Conference on Biometrics*. Springer, 2009, pp. 733–742.
[32] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The CASIA NIR-VIS 2.0 face database," in *CVPR Workshops*, 2013, pp. 348–353.
[33] G. Hu, Y. Hua, Y. Yuan, Z. Zhang, Z. Lu, S. S. Mukherjee, T. M. Hospedales, N. M. Robertson, and Y. Yang, "Attribute-enhanced face recognition with neural tensor fusion networks," in *ICCV*, 2017, pp. 3764–3773.
[34] X. Liu, L. Song, X. Wu, and T. Tan, "Transferring deep representation for nir-vis heterogeneous face recognition," in *International Conference on Biometrics (ICB)*, 2016, pp. 1–8.
[35] M. K. Shaikh, M. A. Tahir, and A. Bouridane, "Tied factor analysis using bagging for heterogeneous face recognition," in *European Workshop on Visual Information Processing (EUVIP)*, Dec 2014, pp. 1–6.
[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
[37] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *Computer Science*, 2015.
[38] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imageNet classification," pp. 1026–1034, 2015.
[39] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
[40] H. S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa, "Memetically optimized mcwld for matching sketches with digital face images," *IEEE Transactions on Information Forensics & Security*, vol. 7, no. 5, pp. 1522–1535, 2012.
[41] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 1955–67, 2009.
[42] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.
[43] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *Computer Science*, 2014.
[44] B. C. Chen, C. S. Chen, and W. H. Hsu, "Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset," *IEEE Transactions on Multimedia*, vol. 17, no. 6, pp. 804–815, 2015.
[45] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *ICCV*, 2015, pp. 3730–3738.
[46] X. Zhang, L. Zhang, X. J. Wang, and H. Y. Shum, "Finding celebrities in billions of web images," *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 995–1007, 2012.
[47] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *ECCV*, 2016, pp. 87–102.
[48] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," *Computer Science*, 2015.
[49] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *CVPR*, 2011, pp. 513–520.
[50] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.
[51] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. PP, no. 99, pp. 1–1, 2018.
[52] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.
[53] B. F. Klare, Z. Li, and A. K. Jain, "Matching forensic sketches to mug shot photos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 639–646, 2011.
[54] D. Gong, Z. Li, W. Huang, X. Li, and D. Tao, "Heterogeneous face recognition: A common encoding feature discriminant approach," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2079–2089, 2017.
[55] H. S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa, "Memetic approach for matching sketches with digital face images," 2012.
[56] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015, pp. 41.1–41.12.
[57] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition." in *ICCV*, 2003, pp. 687–694 vol.1.

Zhongying Deng is now a master student at Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, majoring in Pattern Recognition and Intelligent System. His research interests include computer vision and deep learning.

Xiaojiang Peng received his Ph.D. in school of Information Science and Technology from Southwest Jiaotong University in 2014. He currently is an Associate Professor at Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China. He was a postdoctoral researcher at Idiap Institute, Switzerland from 2016 to 2017, and was a postdoctoral researcher in LEAR Team, INRIA, France, working with Prof. Cordelia Schmid from 2015 to 2016. He serves as a reviewer for CVPR, IJCAI, FG, IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, Image and Vision Computing, IEEE Signal Processing Letter, Neurocomputing, etc. His research focus is in the area of action recognition and detection, face recognition, facial emotion analysis and deep learning.

Zhifeng Li (M'06-SM'11) is currently a principal researcher in Tencent AI Lab. He received his Ph. D. degree from the Chinese University of Hong Kong in 2006. After that, He was a postdoctoral fellow at the Chinese University of Hong Kong and Michigan State University for several years. Before joining Tencent AI Lab, he was a professor at Shenzhen Institutes of Advanced Technology, Chinese Academy of Science. His research focuses include artificial intelligence, computer vision, and human face recognition. He serves as an editorial board member of Neurocomputing.

Yu Qiao (SM' 13) received the Ph.D. degree from the University of Electro-Communications, Japan, in 2006. He was a JSPS Fellow and a Project Assistant Professor with the University of Tokyo, from 2007 to 2010. He is currently a Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. He has authored over 140 papers in journals and conference including, PAMI, IJCV, TIP, ICCV, CVPR, ECCV, and AAAI. His research interests include computer vision, deep learning, and intelligent robots. He was a recipient of the Lu Jiaxi Young Researcher Award from the Chinese Academy of Sciences in 2012. He was the first Runner-Up at the ImageNet Large Scale Visual Recognition Challenge 2015 in scene recognition and the recipient at the ActivityNet Large Scale Activity Recognition Challenge 2016 in video classification.