Hierarchical Deep CNN Feature Set-Based Representation Learning for Robust Cross-Resolution Face Recognition

Guangwei Gao, *Member, IEEE*, Yi Yu, *Member, IEEE*, Jian Yang, *Member, IEEE*, Guo-Jun Qi, *Senior Member, IEEE*, and Meng Yang, *Senior Member, IEEE*

***Abstract*— Cross-resolution face recognition (CRFR), which is important in intelligent surveillance and biometric forensics, refers to the problem of matching a low-resolution (LR) probe face image against high-resolution (HR) gallery face images. Existing shallow learning-based and deep learning-based methods focus on mapping the HR-LR face pairs into a joint feature space where the resolution discrepancy is mitigated. However, little works consider how to extract and utilize the intermediate discriminative features from the noisy LR query faces to further mitigate the resolution discrepancy due to the resolution limitations. In this study, we desire to fully exploit the multi-level deep convolutional neural network (CNN) feature set for robust CRFR. In particular, our contributions are threefold. (i) To learn more robust and discriminative features, we desire to adaptively fuse the contextual features from different layers. (ii) To fully exploit these contextual features, we design a feature set-based representation learning (FSRL) scheme to collaboratively represent the hierarchical features for more accurate recognition. Moreover, FSRL utilizes the primitive form of feature maps to keep the latent structural information, especially in noisy cases. (iii) To further promote the recognition performance, we desire to fuse the hierarchical recognition outputs from different stages. Meanwhile, the discriminability from different scales can also be fully integrated. By exploiting these advantages, the efficiency of the proposed method can be delivered. Experimental results on several face datasets have verified the superiority of the presented algorithm to the other competitive CRFR approaches.**

***Index Terms*— Face recognition, representation learning, feature set, hierarchical fusion.**

I. Introduction

DURING the past few decades, the noise robust face recognition (FR) problem has been a vibrant topic due to the increasing demands in law enforcement and biometric applications [1]–[5]. Promising performance has been achieved under controlled conditions where the acquired face region contains sufficient discriminative information [6]–[12]. Nevertheless, in real surveillance scenes, the desired unambiguous high-resolution (HR) face images may not be always available because of the large distances between cameras and subjects. This results in captured faces that are usually of low-resolution (LR) with too much noise in poses and illumination conditions. Fig. 1(a) demonstrates some real examples of low-resolution faces. The primary challenge is how to match an observed noisy LR probe against those HR candidates from a face image gallery. In this case, the conventional feature extraction and metric learning methods cannot be directly used due to the existence of semantic resolution discrepancy in LR and HR image space.

Recently, we have witnessed some advanced methods investigating the use of deep neural networks for the cross-resolution face recognition (CRFR) problem [13]–[19]. Most of these deep architectures explore pre-trained models or train deep architectures in a feed-forward way to extract features (see traditional deep learning method in Fig. 1(b)). Usually convolutional layers are applied successively with various kernel sizes to capture the local salient features, and pooling layers are adopted to reduce the size of the extracted feature maps with the larger sizes of receptive fields. The final output of the fully connected layers is a high dimensional vector, which is used to represent the features of LR and HR face samples for the recognition task.

Due to the characteristics of LR images, the performance of the CRFR problem is affected by two factors – how to learn

Manuscript received August 27, 2020; revised November 10, 2020; accepted November 21, 2020. Date of publication December 3, 2020; date of current version May 5, 2022. This work was supported in part by the National Key Research and Development Program of China under Project 2018AAA0100102 and Project 2018AAA0100100; in part by the National Natural Science Foundation of China under Grant 61972212, Grant 61772568, and Grant 61833011; in part by the Six Talent Peaks Project in Jiangsu Province under Grant RJFW-011; in part by the Natural Science Foundation of Jiangsu Province under Grant BK20190089; and in part by the Fundamental Research Funds for the Central Universities under Grant 18lgzd15. This article was recommended by Associate Editor Z. Wang. *(Corresponding author: Guangwei Gao.)*

Guangwei Gao is with the Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, Nanjing 210023, China, and also with the Digital Content and Media Sciences Research Division, National Institute of Informatics, Tokyo 101-8430, Japan (e-mail: csggao@gmail.com).

Yi Yu is with the Digital Content and Media Sciences Research Division, National Institute of Informatics, Tokyo 101-8430, Japan (e-mail: yiyu@nii.ac.jp).

Jian Yang is with the School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: csjyang@njust.edu.cn).

Guo-Jun Qi is with the Department of Computer Science, University of Central Florida, Orlando, FL 32816 USA (e-mail: guojunq@gmail.com).

Meng Yang is with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China, and also with the Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, Sun Yat-sen University, Guangzhou 510006, China (e-mail: yangmengpolyu@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2020.3042178.

Digital Object Identifier 10.1109/TCSVT.2020.3042178

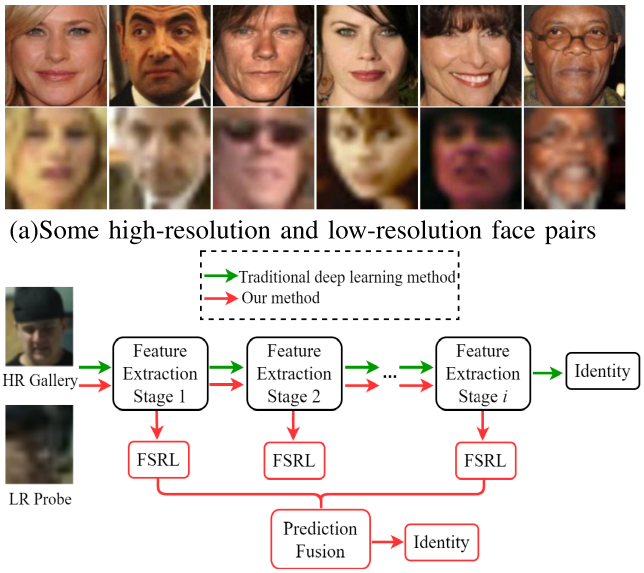


Fig. 1. Significant novelties lie in (i) intermediate FSRL is exploited to mitigate the resolution discrepancy, and (ii) hierarchical predictions from different stages are fused to boost the recognition performance.

more efficient feature representations and how to exploit them for the face recognition task. Carefully designed networks can extract representative and discriminative features for the recognition task. However, in previous methods, the discriminability of the learned representation is not fully studied across multiple latent feature extraction stages, which can provide complementary information for the final recognition. Therefore, in this article, we present to fully explore multi-level deep convolutional neural network (CNN) features through a set representation for the CRFR (Fig. 1). First, we learn multi-scale features in different stages and utilize a simple yet efficient approach to adaptively fuse them. Then, for the resultant hierarchical features, we develop a novel feature set-based representation learning (termed as FSRL) to fully explore these features for more accurate recognition. In addition, based on the observations that features from different stages contain distinct information, we propose to fuse these hierarchical recognition outputs on various scales to further improve their performance. Experiments demonstrate the effectiveness of the presented algorithm in various application scenarios.

We organize the rest of this article as follows. In Section II, we introduce two categories of the relevant works, and the proposed method is presented in Section III. The experimental results and analysis are given in Section IV. Finally, we conclude this article in Section V.

II. Related Work

We briefly introduce the previous relevant works on CRFR in this section. To recognize an LR probe face with limited details, researchers have concentrated on two main approaches, super-resolution methods that recognize faces in the synthesized HR domain space and resolution-robust mapping methods where face samples with different resolutions are matched in a unified feature space.

A. Super-Resolution Reconstruction Algorithms

Super-resolution (SR) algorithms have been investigated during last decades [20], [21]. They first super-resolved the desired HR face samples from the acquired LR one, and then perform similarity metric learning in the same resolution space by means of classical HR image recognition technologies. The authors of [22], [23] presented to obtain the super-resolved face images and remove the noise simultaneously. With the help of carefully designed representation learning strategy, an efficient face image super-resolution method was presented in [24]. To fully utilize the model based prior, a deep CNN denoiser together with multi-layer neighbor embedding method was proposed in [25]. A component generation and enhancement method was proposed in [26]. They firstly obtained the basic facial structure by several parallel CNNs and then predicted the fine grained facial structures by a component enhancement algorithm. To recover identity information when generating HR images, the authors of [27] designed a super-identity CNN model. A siamese generative adversarial network (GAN) was proposed in [28] for identity-preserving face image SR. Similarly, the authors of [29] recently designed a cascaded super-resolution framework together with identity priors to achieve superior performance. In [30], several adaptive kernel mappings were trained to predict the useful high-frequency feature from the given LR input.

B. Discriminative Feature Learning Methods

Resolution-robust algorithms just adopt a couple mappings to meanwhile embed the LR input and related HR pairs into a unified feature space for similarity metric learning. The main challenge of these coupled mapping methods is to design a reasonable discriminant criterion based on some manifold assumptions. A couple of discriminant subspace works have been proposed on the basis of the linear discriminant analysis [31]–[34]. Multidimensional scaling (MDS) [35], [36] method firstly applies facial landmark localization to the LR inputs and then embeds the LR and HR pairs into a unified metric space where their distances approximate the ones in the HR space. To ensure discriminability, two discriminative multidimensional scaling (MDS) methods were presented in [37] to take full advantage of both intra-class and inter-class distance to project the coupled LR and HR faces into a unified space where their large distance gap is mitigated. In [38], [39], multi-resolution face samples were involved simultaneously to extract resolution invariant features for better recognition. Recently, many deep CNN based models have been developed. For example, the robust partially coupled networks were established in [40] to simultaneously achieve feature enhancement and recognition. Motivated by the pioneer work in [41], the authors of [42] applied deep coupled residual network to embed the LR and HR face pairs into a unified space. To investigate the scale-adaptive LR recognition problem, a cascaded SR GAN framework was proposed in [43]. Aghdam *et al.* [14] reported a deep CNN model for LR face recognition, where various training resolutions are used for feature extraction. In [44], the authors introduced a GAN pre-training architecture to further enhance the accuracy of

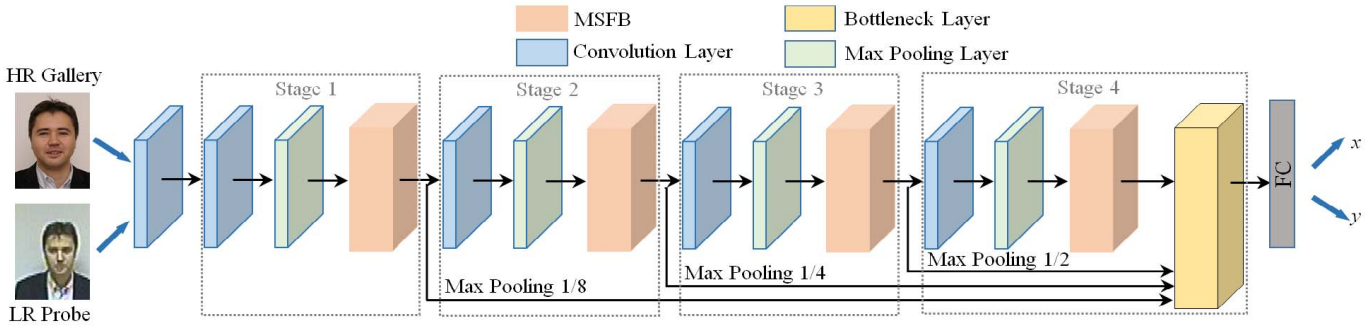


Fig. 2. Flowchart of our proposed feature extraction network (FEN), which can be divided into four stages each representing a feature set. The outputs respectively calculated from four MSFBs are fused by a bottleneck layer. Accordingly, the output from this bottleneck layer is formulated to represent a more discriminative visual feature of LR and HR face images.

several deep learning-based approaches, and a semi-supervised local GAN [45] was also presented to impose the label consistency prior that showed better performance by exploring unlabeled data. The authors of [46] presented a two-stream CNN method based on selective knowledge distillation to identify LR faces with low computational cost. An adversarial training of deep networks has also been proposed to extract the most discriminative features from the generated hard triplets [47]. The contextual information can also be incorporated into the discriminative features through hierarchically gated deep networks [48]. Feature matching between similar images by considering the discriminative spatial contexts has also been studied in literature [49]. Shu *et al.* [50] proposed fine-grained dictionaries to achieve better recognition accuracy, which is also related to the proposed CRFR approach.

Distinguishing from the existing competitive CRFR approaches, in our method, different intermediate features are learned in different stages and fused by a bottleneck layer to achieve a more discriminative feature with more local salient context information. Moreover, a feature set-based representation learning scheme is designed to collaboratively represent these extracted hierarchical features for better recognition. Meanwhile, the discriminability in different scales are federated to further boost the recognition accuracy.

III. Proposed Approach

The challenging issue in CRFR is how to extract discriminative and resolution-invariant features from the pair of LR and HR face images. To this end, in this work, multi-level deep CNN feature sets are output from different stages to investigate discriminative capability of intermediate features. Additionally, an interesting feature set-based representation learning approach is developed to mitigate the resolution discrepancy. The hierarchical recognition results calculated from the CNN feature set of different stages are fused to boost the recognition performance.

A. Feature Extraction Network

Network Architecture. Fig. 2 details the flowchart of the proposed feature extraction network (FEN), which is a Resnet-like CNN [41]. The network employs the CNN to extract discriminative and meaningful features shared by different resolutions. The LR faces are generated as follows: we first downsample the original HR faces by a scale factor s, and then upsample the LR faces to the original size by interpolation.

The convolution layer has a kernel size of 3×3 with stride and padding all setting to 1, while the max pooling is performed with a kernel size of 3×3 and a stride of 2. We add ReLU nonlinear activation after each convolution layer. The number of channels for the feature map in each convolution layer is 32, and a fully connection layer has 512 outputs as the last layer.

Following [51], we use multi-scale feature extraction block (MSFB) to extract the face image features at various scales, as shown in Fig. 3. MSFB uses two different branches with different kernel sizes. We formulate the operation in the MSFB as follows:

$$\begin{aligned}
\boldsymbol{M}_1 &= \sigma\left(w^1_{3\times3} * \boldsymbol{S}_{n-1} + b^1\right), \\
\boldsymbol{N}_1 &= \sigma\left(w^1_{5\times5} * \boldsymbol{S}_{n-1} + b^1\right), \\
\boldsymbol{M}_2 &= \sigma\left(w^2_{3\times3} * [\boldsymbol{M}_1, \boldsymbol{N}_1] + b^2\right), \\
\boldsymbol{N}_2 &= \sigma\left(w^2_{5\times5} * [\boldsymbol{N}_1, \boldsymbol{M}_1] + b^2\right), \\
\boldsymbol{M}' &= w^3_{1\times1} * [\boldsymbol{M}_2, \boldsymbol{N}_2] + b^3,
\end{aligned} \tag{1}$$

where $\sigma(x) = \max(0, x)$ denotes the ReLU operation, and the symbol $[\boldsymbol{M}_1, \boldsymbol{N}_1]$, $[\boldsymbol{N}_1, \boldsymbol{M}_1]$, $[\boldsymbol{M}_2, \boldsymbol{N}_2]$ stand for the concatenation. It should be noted that the input and the output of the first and second convolution layers in the MSFB possess the same number of feature maps. We apply an 1×1 convolution layer to reduce the number of feature maps to 32 in the MSFB.

In the experiment, we find that the output of each MSFB may contain distinct features. Therefore, we want to explore these contextual features from various stages. A simple yet effective feature fusion strategy is used – all the output features from the foregoing MSFB are sent to the end of the network. To adaptively fuse these contextual features, a bottleneck layer composed of a convolution layer with a kernel size of 1×1 is utilized.

The fusion strategy is defined as:

$$\boldsymbol{F} = w * [\boldsymbol{S}_1(8), \boldsymbol{S}_2(4), \boldsymbol{S}_3(2), \boldsymbol{S}_4] + b, \tag{2}$$

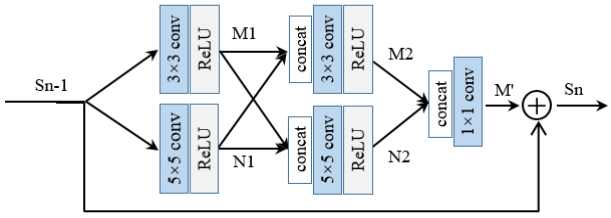


Fig. 3. Multi-scale feature extraction block (MSFB).

where $\boldsymbol{S}_i (i = 1, 2, 3, 4)$ denotes the output of the ith MSFB, and the numbers (8,4, and 2) in the parentheses denote the stride of the max pooling operation.

Training Loss: Let $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ denote the extracted feature vectors by the proposed FEN from the ith HR face and its LR counterpart, respectively. During the training of FEN, we first devote to maximizing inter-class distance to learn discriminative identity features in the respective HR and LR feature spaces. To this end, the following softmax loss is used:

$$L_s = -\sum_{i=1}^{m} \log \frac{e^{\boldsymbol{U}_{c_i}^T \boldsymbol{x}_i + a_{c_i}}}{\sum_{j=1}^{n} e^{\boldsymbol{U}_j^T \boldsymbol{x}_i + a_j}} - \sum_{i=1}^{m} \log \frac{e^{\boldsymbol{V}_{c_i}^T \boldsymbol{y}_i + b_{c_i}}}{\sum_{j=1}^{n} e^{\boldsymbol{V}_j^T \boldsymbol{y}_i + b_j}}, \tag{3}$$

where m denotes the number of the training sample pairs, n denotes the number of the object classes in the training set, c_i represents the label of the ith sample image, and $\boldsymbol{U}_j$ and $\boldsymbol{V}_j$ are the jth column of the weight matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ in the final fully connection layer, while a and b are the biases for the respective HR and LR feature spaces.

Meanwhile, we aim to reduce the intra-class difference between an individual face sample and its center of the same identity in the feature space. The center loss [6] is written as

$$L_c = \sum_{i=1}^{m} \left\| \boldsymbol{x}_i - \boldsymbol{z}_{c_i}^x \right\|_2^2 + \sum_{i=1}^{m} \left\| \boldsymbol{y}_i - \boldsymbol{z}_{c_i}^y \right\|_2^2, \tag{4}$$

where $\boldsymbol{z}_{c_i}^x$ and $\boldsymbol{z}_{c_i}^y$ are the centers of the HR and the LR features corresponding to the c_ith class, respectively.

As shown in Figure 2, the critical challenge of the CRFR comes from the limited distinct features in the observed LR face images. Fortunately, the HR training samples can be utilized to guide the extraction of discriminative features from the LR faces. For the CRFR task, the features of LR face images should be as closed as possible to their HR counterparts. For the sake of simplicity, we have the following Euclidean loss

$$L_e = \sum_{i=1}^{m} \left\| \boldsymbol{x}_i - \boldsymbol{y}_i \right\|_2^2. \tag{5}$$

By considering the previous three effective losses, the loss of the proposed method can be written as

$$L_{FEN} = L_s + \theta_1 L_c + \theta_2 L_e. \tag{6}$$

where θ_1 and θ_2 are two balancing hyper-parameters to control the contributions of the center loss and the Euclidean loss. In this fashion, the proposed method takes into account both the discriminative and representative ability of the learned features, making the CRFR more expressive in the learned feature space.

B. Feature Set-Based Representation Learning

In previous methods, the tail features (e.g., $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ in the aforementioned section) extracted by the trained network are usually used to train the classifiers directly for the recognition task. However, the extracted features from the MSFBs are not fully explored to their full potentials. We will elaborate in this section on how we can utilize these multi-level features to mitigate the resolution discrepancy for better recognition performance.

1) Vector Set-Based Collaborative Representation: In this part, we use a vector set to represent a face image. The features extracted by FEN from a LR query face image in a specific stage is denoted as $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_i, \ldots, \boldsymbol{y}_{n_a}\} \in \Re^{d \times n_a}$ (where each column of $\boldsymbol{Y}$ is a reshaped feature map, n_a denotes the number of feature maps in a query stage, and d is the size of the reshaped feature map).

Denote by $\boldsymbol{X}_k$ the features extracted from the kth ($k = 1, 2, \ldots, K$) HR gallery face image in the same stage. Let $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k, \ldots, \boldsymbol{X}_K] \in \Re^{d \times n_b}$ be the concatenation of the features from all the HR gallery faces, and n_b denotes the total number of the resultant feature maps.

For the query feature set $\boldsymbol{Y}$, its l_p-norm regularized hull can be defined as

$$H(\boldsymbol{Y}) = \left\{ \sum_{i=1}^{n_a} \alpha_i \boldsymbol{y}_i \mid \|\boldsymbol{\alpha}\|_{l_p} < \delta, \quad \text{s.t.} \sum \alpha_i = 1 \right\} \tag{7}$$

where $\boldsymbol{\alpha}$ is the coefficient vector. Then, we can define the representation of the hull $\boldsymbol{Y\alpha}$ over the gallery feature set $\boldsymbol{X}$ as follows:

$$\begin{aligned} &\min_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \|\boldsymbol{Y\alpha} - \boldsymbol{X\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\alpha}\|_{l_p} + \lambda_2 \|\boldsymbol{\beta}\|_{l_p} \\ &\text{s.t.} \sum \alpha_i = 1, \end{aligned} \tag{8}$$

where $\boldsymbol{\beta}$ is the representation vector, the constraint $\sum \alpha_i = 1$ is used to prevent the trivial solution $\boldsymbol{\alpha} = \boldsymbol{\beta} = 0$, and λ_1 and λ_2 are hyper-parameters to balance between the regularization terms on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively.

Either l_1-norm or l_2-norm could be explored to constrain the vector norm for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. For the sake of efficiency and effectiveness, we use l_2-norm here. In this case, Eq. (8) will have a closed-form solution. The Lagrangian function (8) can be denoted as

$$\begin{aligned} &L(\boldsymbol{z}, \boldsymbol{\varphi}) \\ &= \|\boldsymbol{Y\alpha} - \boldsymbol{X\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\alpha}\|_2^2 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 + \boldsymbol{\varphi}(\boldsymbol{e\alpha} - 1) \\ &= \|\boldsymbol{Az}\|_2^2 + \boldsymbol{z}^T \boldsymbol{Bz} + \boldsymbol{\varphi} \left(\boldsymbol{d}^T \boldsymbol{z} - 1 \right), \end{aligned} \tag{9}$$

where $\boldsymbol{e}$ is an all-one row vector, $\boldsymbol{d} = [\boldsymbol{e}\ 0]^T$, and

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix}, \boldsymbol{A} = \begin{bmatrix} \boldsymbol{Y} & -\boldsymbol{X} \end{bmatrix}, \boldsymbol{B} = \begin{bmatrix} \lambda_1 \boldsymbol{I} & 0 \\ 0 & \lambda_2 \boldsymbol{I} \end{bmatrix}. \tag{10}$$

By taking the derivative of the Langarian function wrt the multiplier $\boldsymbol{\varphi}$ and the decision variable z, and equating the results to zero, we obtain

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{\varphi}} &= \boldsymbol{d}^T z - 1 = 0\\ \frac{\partial L}{\partial z} &= \boldsymbol{A}^T \boldsymbol{A} z + \boldsymbol{B} z + \boldsymbol{\varphi} \boldsymbol{d} = 0.\end{aligned} \tag{11}$$

Then, we can obtain the closed solution to Eq. (9):

$$\hat{z} = \begin{bmatrix}\hat{\boldsymbol{\alpha}}\\ \hat{\boldsymbol{\beta}}\end{bmatrix} = z_0 / \boldsymbol{d}^T z_0, \tag{12}$$

where $z_0 = \left(\boldsymbol{A}^T \boldsymbol{A} + \boldsymbol{B}\right)^{-1} \boldsymbol{d}$.

2) Matrix Set-Based Collaborative Representation: Contrary to the previous section where each feature map is treated as a vector, here we adopt the original matrix form of the feature map to represent a face image. Existing works [52] have revealed that nuclear norm constraint could be more suitable to keep the 2D structure of a feature map. The features extracted from a LR query face image and all the HR gallery faces in a certain stage are denoted by $\boldsymbol{Y} = \left\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_i, \ldots, \boldsymbol{Y}_{n_a}\right\} \in \Re^{p \times q \times n_a}$ and $\boldsymbol{X} = \left[\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k, \ldots, \boldsymbol{X}_{n_b}\right] \in \Re^{p \times q \times n_b}$, respectively.

Then, we can define the representation of the hall $\boldsymbol{Y}$ over the corresponding gallery feature set $\boldsymbol{X}$ by

$$\begin{aligned}&\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \|\boldsymbol{Y}(\boldsymbol{\alpha}) - \boldsymbol{X}(\boldsymbol{\beta})\|_* + \lambda_1 \|\boldsymbol{\alpha}\|_2^2 + \lambda_2 \|\boldsymbol{\beta}\|_2^2\\ &\text{s.t. } \sum \alpha_i = 1,\end{aligned} \tag{13}$$

where $\|\cdot\|_*$ denotes the nuclear norm of a matrix, $\boldsymbol{Y}(\boldsymbol{\alpha}) = \alpha_1 \boldsymbol{Y}_1 +, \ldots, +\alpha_{n_a} \boldsymbol{Y}_{n_a}$, and $\boldsymbol{X}(\boldsymbol{\beta}) = \beta_1 \boldsymbol{X}_1 +, \ldots, +\beta_{n_b} \boldsymbol{X}_{n_b}$.

For convenience, Eq. (13) can be rewritten as

$$\begin{aligned}&\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \|\boldsymbol{E}\|_* + \lambda_1 \|\boldsymbol{\alpha}\|_2^2 + \lambda_2 \|\boldsymbol{\beta}\|_2^2\\ &\text{s.t. } \boldsymbol{Y}(\boldsymbol{\alpha}) - \boldsymbol{X}(\boldsymbol{\beta}) = \boldsymbol{E},\ \sum \alpha_i = 1.\end{aligned} \tag{14}$$

The alternating minimization method (ADMM) is then adopted to solve this optimization problem with the following augmented Lagrangian function:

$$\begin{aligned}L = \|\boldsymbol{E}\|_* + \lambda_1 \|\boldsymbol{\alpha}\|_2^2 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 + \langle \boldsymbol{Z}, \boldsymbol{Y}(\boldsymbol{\alpha}) - \boldsymbol{X}(\boldsymbol{\beta}) - \boldsymbol{E}\rangle\\ + \langle \boldsymbol{\gamma}, \boldsymbol{e}\boldsymbol{\alpha} - 1\rangle + \frac{\mu}{2}\left(\|\boldsymbol{Y}(\boldsymbol{\alpha}) - \boldsymbol{X}(\boldsymbol{\beta}) - \boldsymbol{E}\|_2^2 + \|\boldsymbol{e}\boldsymbol{\alpha} - 1\|_2^2\right),\end{aligned} \tag{15}$$

where $\langle \cdot, \cdot \rangle$ is the inner product, and $\boldsymbol{Z}$ and $\boldsymbol{\gamma}$ are the auxiliary Lagrange multipliers, with a positive penalty constant $\mu > 0$.

Then the optimal $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be solved alternatively. Specifically, by fixing others, the solution to $\boldsymbol{\alpha}$ is

$$\begin{aligned}\boldsymbol{\alpha}^{(l+1)} &= \arg\min_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}, \boldsymbol{\beta}^{(l)}, \boldsymbol{E}^{(l)}, \boldsymbol{Z}^{(l)}, \boldsymbol{\gamma}^{(l)})\\ &= \arg\min_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}) + \left\|\boldsymbol{e}\boldsymbol{\alpha} - 1 + \boldsymbol{\gamma}^{(l)}/\mu\right\|_2^2\\ &= \arg\min_{\boldsymbol{\alpha}} \|\tilde{\boldsymbol{Y}}\boldsymbol{\alpha} - \tilde{\boldsymbol{x}}\|_2^2 + \eta \|\boldsymbol{\alpha}\|_2^2,\end{aligned} \tag{16}$$

where $f(\boldsymbol{\alpha}) = \left\|\boldsymbol{Y}(\boldsymbol{\alpha}) - \boldsymbol{X}(\boldsymbol{\beta}^{(l)}) - \boldsymbol{E}^{(l)} + \boldsymbol{Z}^{(l)}/\mu\right\|_2^2 + \eta \|\boldsymbol{\alpha}\|_2^2$, $\tilde{\boldsymbol{x}} = \left[\mathrm{Vec}\left(\boldsymbol{X}(\boldsymbol{\beta}^{(l)}) + \boldsymbol{E}^{(l)} - \boldsymbol{Z}^{(l)}/\mu\right); \left(1 - \boldsymbol{\gamma}^{(l)}/\mu\right)\right]$, $\tilde{\boldsymbol{Y}} = [\boldsymbol{H};\ \boldsymbol{e}]$, $\boldsymbol{H} = \left[\mathrm{Vec}\left(\boldsymbol{Y}_1\right), \ldots, \mathrm{Vec}\left(\boldsymbol{Y}_{n_a}\right)\right]$, and $\eta = 2\lambda_1/\mu$. Thus, Eq. (16) has a closed form solution as

$$\boldsymbol{\alpha}^{(l+1)} = (\tilde{\boldsymbol{Y}}^T \tilde{\boldsymbol{Y}} + \eta \cdot \boldsymbol{I})^{-1} \tilde{\boldsymbol{Y}}^T \tilde{\boldsymbol{x}}. \tag{17}$$

Once $\boldsymbol{\alpha}^{(l+1)}$ is obtained, $\boldsymbol{\beta}^{(l+1)}$ is updated via optimizing the following minimization problem:

$$\begin{aligned}\boldsymbol{\beta}^{(l+1)} &= \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\alpha}^{(l+1)}, \boldsymbol{\beta}, \boldsymbol{E}^{(l)}, \boldsymbol{Z}^{(l)}, \boldsymbol{\gamma}^{(l)})\\ &= \arg\min_{\boldsymbol{\beta}} \|\tilde{\boldsymbol{X}}\boldsymbol{\beta} - \tilde{\boldsymbol{y}}\|_2^2 + \rho \|\boldsymbol{\beta}\|_2^2,\end{aligned} \tag{18}$$

where $\tilde{\boldsymbol{y}} = \mathrm{Vec}\left(\boldsymbol{Y}(\boldsymbol{\alpha}^{(l+1)}) - \boldsymbol{E}^{(l)} + \boldsymbol{Z}^{(l)}/\mu\right)$, $\tilde{\boldsymbol{X}} = \left[\mathrm{Vec}\left(\boldsymbol{X}_1\right), \ldots, \mathrm{Vec}\left(\boldsymbol{X}_{n_b}\right)\right]$, and $\rho = 2\lambda_2/\mu$. The closed form solution of Eq. (18) is given as

$$\boldsymbol{\beta}^{(l+1)} = (\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}} + \rho \cdot \boldsymbol{I})^{-1} \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{y}}. \tag{19}$$

By fixing other parameters, $\boldsymbol{E}^{(l+1)}$ can be solved by

$$\begin{aligned}\boldsymbol{E}^{(l+1)} &= \arg\min_{\boldsymbol{E}} L(\boldsymbol{\alpha}^{(l+1)}, \boldsymbol{\beta}^{(l+1)}, \boldsymbol{E}, \boldsymbol{Z}^{(l)}, \boldsymbol{\gamma}^{(l)})\\ &= \arg\min_{\boldsymbol{E}} \frac{1}{\mu}\|\boldsymbol{E}\|_* + \frac{1}{2}\|\boldsymbol{E} - \boldsymbol{F}\|_2^2,\end{aligned} \tag{20}$$

where $\boldsymbol{F} = \boldsymbol{Y}(\boldsymbol{\alpha}^{(l+1)}) - \boldsymbol{X}(\boldsymbol{\beta}^{(l+1)}) + \boldsymbol{Z}^{(l)}/\mu$. The solution of problem (20) could be solved by

$$\boldsymbol{E}^{(l+1)} = \boldsymbol{U}\boldsymbol{T}_{\frac{1}{\mu}}[\boldsymbol{S}]\boldsymbol{V}, \tag{21}$$

in which $\left(\boldsymbol{U}, \boldsymbol{S}, \boldsymbol{V}^T\right) = \mathrm{svd}(\boldsymbol{F})$, $\boldsymbol{T}_{\frac{1}{\mu}}[\boldsymbol{S}] = \mathrm{diag}\left(\left\{\max\left(0, s_{j,j} - \frac{1}{\mu}\right)\right\}_{1 \le j \le r}\right)$, and r denotes the rank of matrix $\boldsymbol{S}$.

Once $\boldsymbol{\alpha}^{(l+1)}$, $\boldsymbol{\beta}^{(l+1)}$ and $\boldsymbol{E}^{(l+1)}$ are obtained, the auxiliary Lagrange multipliers $\boldsymbol{Z}$ and $\boldsymbol{\gamma}$ can be updated to

$$\begin{aligned}\boldsymbol{\gamma}^{(l+1)} &= \boldsymbol{\gamma}^{(l)} + \mu\left(\boldsymbol{e}\boldsymbol{\alpha}^{(l+1)} - 1\right),\\ \boldsymbol{Z}^{(l+1)} &= \boldsymbol{Z}^{(l)} + \mu\left(\boldsymbol{Y}(\boldsymbol{\alpha}^{(l+1)}) - \boldsymbol{X}(\boldsymbol{\beta}^{(l+1)}) - \boldsymbol{E}^{(l+1)}\right).\end{aligned} \tag{22}$$

The procedure for solving Eq. (14) is summarized in **Algorithm 1**.

Algorithm 1 Solving Eq. (14) via ADMM

Input: The extracted feature set $\boldsymbol{Y} \in \Re^{p \times q \times n_a}$ from a LR query face, concatenated feature set $\boldsymbol{X} \in \Re^{p \times q \times n_b}$ from all the HR gallery faces.

Parameter: The model parameters λ_1 and λ_2, and the termination condition parameter ϵ.

Initialize: $\boldsymbol{\alpha}^0 = \boldsymbol{\beta}^0 = 0$, $\boldsymbol{\gamma}^0 = 0$, $\boldsymbol{E}^0 = \boldsymbol{Z}^0 = 0$.

while $\left\|\boldsymbol{Y}(\boldsymbol{\alpha}^{l+1}) - \boldsymbol{X}(\boldsymbol{\beta}^{l+1}) - \boldsymbol{E}^{l+1}\right\|_F^2 > \epsilon$ **do**

1: Update $\boldsymbol{\alpha}$ via Eq. (17);

2: Update $\boldsymbol{\beta}$ via Eq. (19);

3: Update $\boldsymbol{E}$ via Eq. (21);

4: Update Lagrange multipliers $\boldsymbol{Z}$ and $\boldsymbol{\gamma}$ via Eq. (14);

5: $l \leftarrow l + 1$.

end while

Output: The optimal representation vectors $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$.

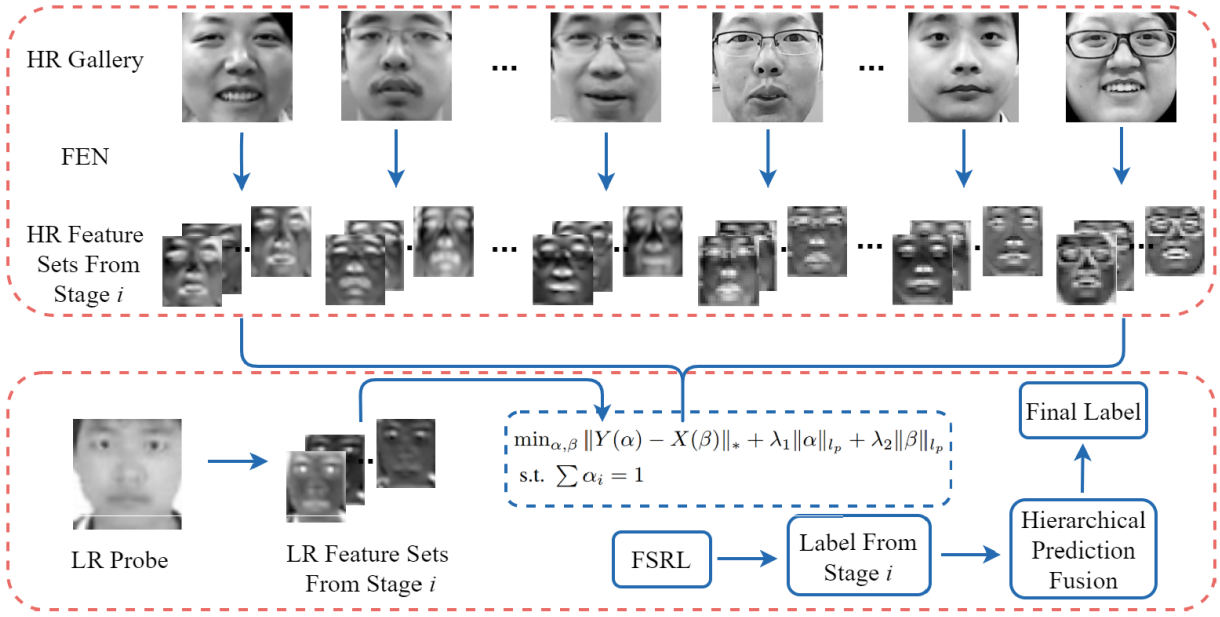


Fig. 4. The proposed HFSRL scheme for CRFR process. The FEN is used to extract discriminative feature sets. First, multi-scale features are extracted in each stage. Then, based on these hierarchical features, FSRL scheme is designed to fully exploit these deep CNN features for more accurate recognition. Last, these hierarchical recognition outputs are fused to further promote the recognition performance.

C. Hierarchical Prediction Fusion

It is well known that the features obtained from different layers contain distinct information. The features learned from the shallow layers contain the low level information such as edges and corners, while the features with rich semantics can be extracted from the deeper layers. Fully exploring the discriminative abilities of such hierarchical features is essential to the recognition tasks [53].

Suppose that we have obtained the representation vectors $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ via solving the aforementioned feature set-based representation learning problem. We can rewrite $\hat{\boldsymbol{\beta}}$ as $\hat{\boldsymbol{\beta}} = \left[\hat{\boldsymbol{\beta}}_1; \ldots; \hat{\boldsymbol{\beta}}_c; \ldots; \hat{\boldsymbol{\beta}}_C\right]$, where each $\hat{\boldsymbol{\beta}}_c$ denotes the sub-vector of the coefficients corresponding to the cth class. Then the regularized representation residual of hall $\boldsymbol{Y}(\hat{\boldsymbol{\alpha}})$ over each class $\boldsymbol{X}_c$ can be denoted by

$$r_c = \left\|\boldsymbol{Y}(\hat{\boldsymbol{\alpha}}) - \boldsymbol{X}_c(\hat{\boldsymbol{\beta}}_c)\right\|_2^2 / \left\|\hat{\boldsymbol{\beta}}_c\right\|_2^2. \tag{23}$$

Then the class label of the query feature set $\boldsymbol{Y}$ is $\text{Identity}(\boldsymbol{Y}) = \arg\min_c \{r_c\}$.

Now the problem boils down to how to fuse the hierarchical outputs from different stages (scales) to achieve a better performance. With the help of a given dataset $\boldsymbol{T} = \{(\boldsymbol{x}_i, z_i)\}$ $(i = 1, 2, \ldots, n)$ and s scales (in our model, the output of the ith stage is treated as the ith scale due to the use of a pooling operation), a decision matrix can be defined as follows:

$$d_{ij} = \begin{cases} +1, & \text{if } h_{ij} = z_i \\ -1, & \text{if } h_{ij} \neq z_i, \end{cases} \tag{24}$$

where z_i is the real label for sample $\boldsymbol{x}_i$ while $h_{ij}(j = 1, 2, \ldots, s)$ represents the predicted label of $\boldsymbol{x}_i$ on the jth scale.

In order to obtain the best recognition result from different stages of scales, we define the following objective function:

$$\begin{aligned} &\min_{\boldsymbol{\sigma}} \|\boldsymbol{e}_1 - \boldsymbol{D}\boldsymbol{\sigma}\|_2^2 + \tau\|\boldsymbol{\sigma}\|_1 \\ &\text{s.t.} \sum \sigma_i = 1, \sigma_i > 0, \end{aligned} \tag{25}$$

where $\boldsymbol{\sigma}$ is the scale weight, τ is the regularization parameter, and $\boldsymbol{e}_1 = [1, \ldots, 1]^T$ has a length of s. Eq. (25) can be rewritten as

$$\begin{aligned} &\min_{\boldsymbol{\sigma}} \|\hat{\boldsymbol{e}} - \hat{\boldsymbol{D}}\boldsymbol{\sigma}\|_2^2 + \tau\|\boldsymbol{\sigma}\|_1 \\ &\text{s.t. } \sigma_i > 0, i = 1, 2, \cdots, s, \end{aligned} \tag{26}$$

where $\hat{\boldsymbol{e}} = [\boldsymbol{e}_1; 1]$, $\hat{\boldsymbol{D}} = [\boldsymbol{D}; \boldsymbol{e}_1]$. The solution of problem (26) can be easily obtained by the widely used l_1_ls solver [54]. Once the optimal scale weights are obtained, the fused prediction can be formulated as $\text{Identity}(\boldsymbol{x}_i) = \arg\max_k \left\{\sum_j \sigma_j | h_{ij} = k\right\}$. The overall evaluation process is given in Fig. 4.

IV. Experimental Results and Analysis

In this part, we implement tests to validate the efficiency of our model. Following previous work, we use the CASIA-Webface [55] to train our FEN. The detected faces are normalized and resized to have a size of 112×96. In the next, we firstly depict the datasets and the experimental settings, and then perform comparisons between our proposed approach and several competitive CRFR approaches. We implement our model with PyTorch on the popular NVIDIA Titan Xp GPU.

A. Datasets and Settings

Experiments are performed on three well-known face datasets: UCCS (UnConstrained College Students) [56],

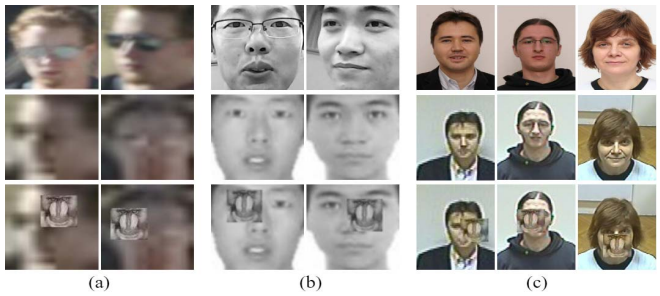


Fig. 5. Example face samples from the (a) UCCS dataset, (b) NJU-ID dataset, and (c) SCface dataset. Each column lists three images with the same identity from two respective resolutions, where image samples in the first row have HR while in the second (third) row have LR without (with) block occlusion.

NJU-ID (Nanjing University ID Card Face) [57] and SCface (Surveillance Cameras Face) [58]. Some HR-LR images pairs from these datasets are listed in Fig. 5. We detail the three datasets in the next text.

1) UCCS Dataset: The UCCS dataset collects face images of college students. The distance between the HR surveillance camera and the objects is about 100 to 150 meters. The images captured in large standoff distance and unconstrained surveillance settings make the recognition problem more difficult. Face images from 1,732 labeled persons are used, where blur, occlusion and bad illumination are existed. Following the experimental protocol in [40], we choose the top 180 subjects on the basis of the number of images. In this experiment, we separate the images of each subject according to a ratio of 1:4 to form the probe and gallery sets. The gallery face samples are reshaped to have a size of 112×96 as the HR sets, while the probe face samples are first down-sampled to 14×12 pixels and then resized to 112×96 pixels to form the LR sets. The same size face samples in CASIA-WebFace dataset are applied for training the FEN.

2) NJU-ID Dataset: The NJU-ID dataset includes face samples from 256 persons. A non-contact IC chip is embedded in the card. The ID card used here refers to the second generation of resident ID cards in China. Due to the storage limitations of the ID card, the stored images natively have low resolution. For each person, there are one HR camera image captured from a digital camera and one LR card image. The ID card image has a size of 102×126, while the camera image has a size of 640×480. All the card and camera images are resized to have a size of 112×96. To make the problem more challenging, we further down-sample the ID card images to 28×24 to form the LR query images.

3) SCface Dataset: The SCface dataset uses five video surveillance cameras with various qualities to collect uncontrolled indoor face images from 130 subjects. This dataset can be regarded as a real-world LR dataset. For each person, there is one frontal mugshot face sample captured by a digital camera and 15 images (five images at each distance) taken by five real surveillance cameras with different qualities within three distances (1.0m, 2.6m and 4.2m, respectively). In this experiment, 50 out of 130 persons are randomly picked to fine-tune the FEN while the rest for test. The CASIA-WebFace images with size of 112×96 are take as the HR images while those of 7×6, 10×8 and 16×14 are taken as LR images to train the FEN at three distances.

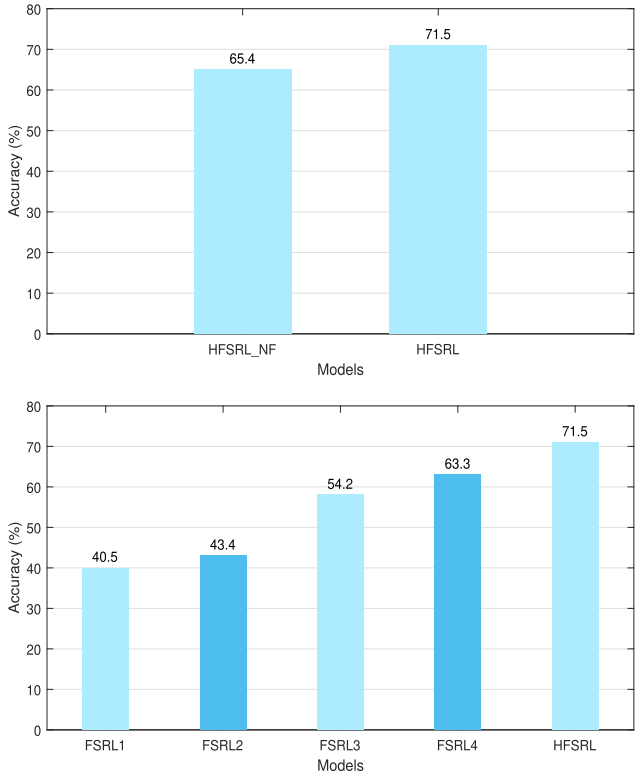


Fig. 6. Ablation study on effects of the feature fusion (top) and the hierarchical prediction fusion (bottom).

B. Ablation Study

Fig. 6 presents the ablation study on the feature fusion and hierarchical prediction fusion. In this part, for the sake of convenience, we use HFSRL to represent hierarchical vector set-based collaborative learning. Compared to HFSRL, HFSRL_NF removes the feature connections from other stages. FSRLi ($i = 1, 2, 3, 4$) indicates using the feature sets from the ith stage for representation learning. From Fig. 6, we can see that, FSRL obtains better recognition accuracy than FSRL_NF, which reveals the feature fusion strategy is useful for recognition. The reason may be that the features from other stages can carry some discriminative information from early layers to latter layers.

From Fig. 6, we can also find that the performance from different stages varies a lot. Generally, the features extracted from the lower layer have the worst performance since the semantic information revealed by the lower layer is limited. The features extracted from the higher layer achieve better performance than that in lower layer. The reason may be that the features in higher layer contain more semantic information, that is essential for recognition tasks. Moreover, our fusion method obtains the best performance, which reveals that fusing the results from latent layers can bring complementary discriminative ability for the final recognition.

TABLE I

FACE RECOGNITION INDEXES (%) OF RESPECTIVE METHODS ON THE UCCS DATASET. THE BOLDFACE INDICATES OUR METHOD

Methods	Accuracy (%)	Year
SICNN [27]	66.5	2018
SiGAN [28]	67.2	2019
PCN [40]	55.4	2016
DCR [42]	70.3	2018
DAlign [34]	71.9	2019
SKD [46]	75.2	2019
Centerloss [44]	76.4	2019
HFSRL-v	**79.5**	-
HFSRL-m	**80.8**	-

TABLE II

FACE RECOGNITION INDEXES (%) OF RESPECTIVE METHODS ON THE NJU-ID DATASET. THE BOLDFACE INDICATES OUR METHOD

Methods	Accuracy (%)	Year
SICNN [27]	62.4	2018
SiGAN [28]	62.8	2019
PCN [40]	58.5	2016
DCR [42]	63.7	2018
DAlign [34]	64.5	2019
SKD [46]	67.8	2019
Centerloss [44]	68.4	2019
HFSRL-v	**71.4**	-
HFSRL-m	**72.6**	-

TABLE III

FACE RECOGNITION INDEXES (%) OF RESPECTIVE METHODS ON THE SCFACE DATASET. THE BOLDFACE INDICATES OUR METHOD

Methods	Dist 1	Dist 2	Dist 3	Year
SICNN [27]	28.3	38.2	44.5	2018
SiGAN [28]	28.8	38.7	44.8	2019
PCN [40]	26.8	38.2	43.5	2016
DCR [42]	30.3	40.5	45.3	2018
DAlign [34]	32.4	42.7	48.7	2019
SKD [46]	38.5	48.0	54.7	2019
Centerloss [44]	40.5	51.8	57.5	2019
HFSRL-v	**44.2**	**54.3**	**59.5**	-
HFSRL-m	**45.3**	**55.3**	**60.6**	-

C. Competitive Results

We also compare our presented algorithm with two categories of advanced approaches to handle the resolution mismatching issue: one is super-resolution methods, such as SICNN [27] and SiGAN [28], together with one deep-based recognition method, i.e., DFL [6]. The other is resolution-robust methods, such as PCN [40], DCR [42], DAlign [34], SKD [46] and Centerloss [44]. For those super-resolution approaches, we adopt the CASIA-Webface dataset for training. While for resolution-robust approaches, we employ the same probe and gallery sets. We use HFSRL-v and HFSRL-m to denote the hierarchical feature set-based representation learning with vector and matrix form, respectively.

Tables I-III show the recognition results. We see that directly feeding the super-resolved faces into the classical recognition method appears to have a small contribution to final recognition since that the synthesized faces may be not optimized

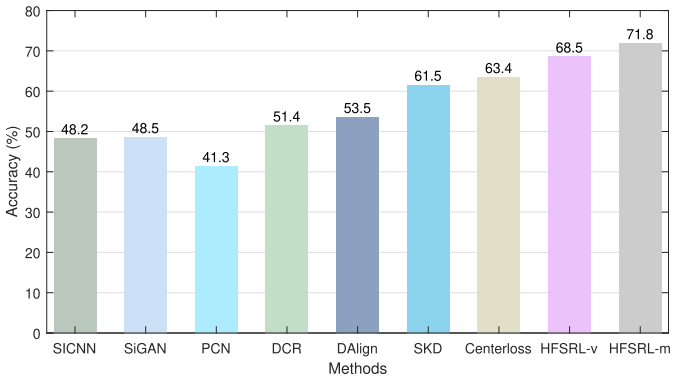


Fig. 7. Face recognition accuracy (%) of respective method on the UCCS dataset with random occlusion.

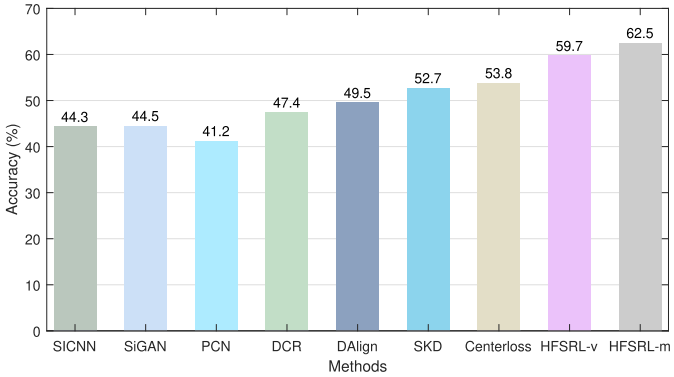


Fig. 8. Face recognition accuracy (%) of respective method on the NJU-ID dataset with random occlusion.

for recognition tasks. By comparison, the resolution-robust approaches (i.e., PCN, DCR, DAlign, SKD, and Centerloss) take the discriminability of features into account, achieving better recognition performance. The quantitative comparisons on three datasets also validate that our HFSRL approach gets the best performance among all competitive ones. By fully exploiting the multi-level deep CNN features, our proposed HFSRL can dramatically boost the recognition accuracy.

On account of the complicated and unknown imaging scenes, the effect of noise cannot be neglected in real-world applications. In this part, the observed LR query face samples are corrupted by a square "baboon" image with a random location under an occlusion standard of 20%. Some examples are displayed in Fig. 5. The recognition results of competitive approaches are given in Fig. 7-9. We can survey that the performance of all methods are reduced drastically. Our method (both HFSRL-v and HFSRL-m) can also perform better than other competitors. Particularly, by considering the latent structural information of the feature set, our proposed HFSRL-m can better reveal noise and performs better than HFSRL-v.

D. Speed Comparisons

In this part, we check the computational speed of competitive methods. We conduct tests with a configuration of Intel CPU @ 3.4 GHz. For the simplicity of demonstration, we only provide the comparisons on the NJU-ID dataset. The average inference time of respective methods are tabulated

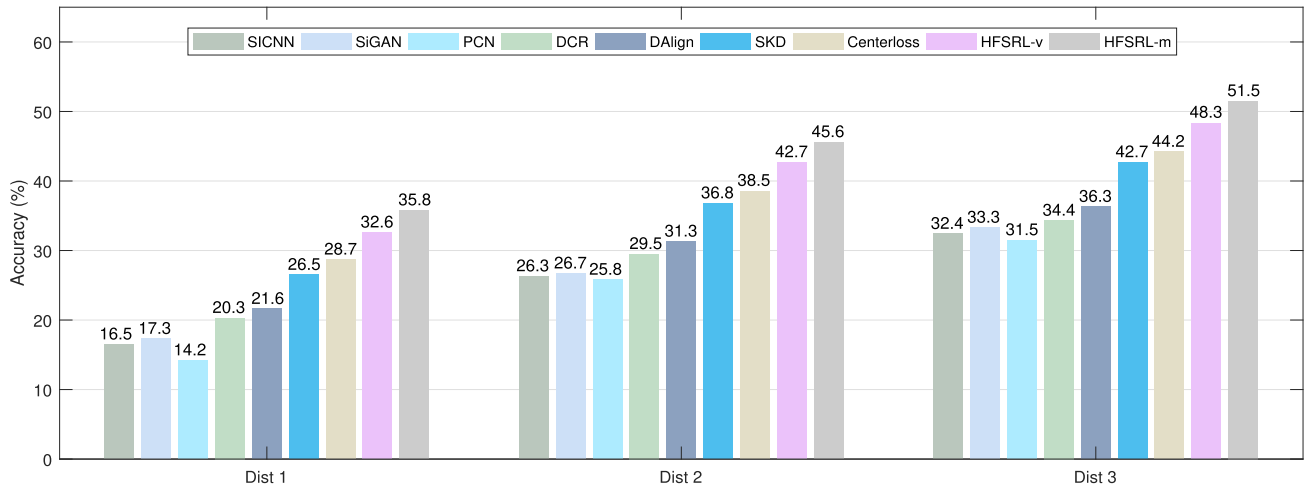


Fig. 9. Face recognition accuracy (%) of respective method on the SCface dataset with random occlusion.

TABLE IV
SPEED COMPARISONS (SECONDS) OF RESPECTIVE METHODS ON THE NJU-ID DATASET

Methods	Time (seconds)	Year
SICNN [27]	0.92	2018
SiGAN [28]	1.15	2019
PCN [40]	0.33	2016
DCR [42]	0.46	2018
DAlign [34]	0.62	2019
SKD [46]	0.25	2019
Centerloss [44]	0.53	2019
HFSRL-v	**1.62**	-
HFSRL-m	**4.50**	-

in Table IV. The two super-resolution methods, SICNN and SiGAN, cost little more time due to the extra operation of resolution enhancement. By directly performing recognition, the resolution-robust approaches, PCN, DCR, DAlign, SKD, and Centerloss, need relatively lower computational cost. Different from previous methods, which directly use the tail extracted feature vector for recognition, our proposed methods fully take the multi-level hierarchical features into account, thus cost much more computational time. Especially, HFSRL-v has closed solution and only involves a matrix inversion operation. Thus, it has comparative time consumption with other methods. HFSRL-m obtains the best performance at the cost of higher time consumption due to the iterative procedure in representation learning. In our future work, we will try our best to investigate fast and efficient ADMM to accelerate the procedure of representation learning.

V. CONCLUSION

In this work, we present to exploit multi-level deep CNN feature set to further mitigate the resolution discrepancy for better CRFR. An end-to-end feature extraction network is suggested to learn a more discriminative feature representation, which can contain more details of visual and contextual information. A feature set-based representation learning scheme is proposed to jointly represent hierarchical features. By fusing recognition results respectively generated by hierarchical features, CRFR accuracy can be improved. In addition, experimental results over three different popular face datasets with various recognition scenes have verified that the presented approach can outperform some competitive CRFR approaches.

In the future work, we will incorporate face priors such as face landmark and face parsing into the attention network to enhance the discriminability of the features. Also, we will try to adopt the graph neural networks to handle the multi-level hierarchical features for better recognition. Moreover, we will investigate the adversarial metric learning methods to robustly match the cross-resolution face image pairs.

REFERENCES

[1] J. Li *et al.*, "Robust face recognition with deep multi-view representation learning," in *Proc. ACM Multimedia Conf. (MM)*, 2016, pp. 1068–1072.
[2] C. Peng, N. Wang, J. Li, and X. Gao, "Re-ranking high-dimensional deep local representation for NIR-VIS face recognition," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4553–4565, Sep. 2019.
[3] G. Gao, Y. Yu, M. Yang, H. Chang, P. Huang, and D. Yue, "Cross-resolution face recognition with pose variations via multilayer locality-constrained structural orthogonal procrustes regression," *Inf. Sci.*, vol. 506, pp. 19–36, Jan. 2020.
[4] F. Keinert, D. Lazzaro, and S. Morigi, "A robust group-sparse representation variational method with applications to face recognition," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2785–2798, Jun. 2019.
[5] W. Deng, J. Hu, and J. Guo, "Compressive binary patterns: Designing a robust binary face descriptor with random-field eigenfilters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 758–767, Mar. 2019.
[6] Y. Wen, K. Zhang, and Z. Li, "A discriminative feature learning approach for deep face recognition," in *Proc. Comput. Vis. (ECCV)*, 2016, pp. 499–515.
[7] L. Liu, C. Xiong, H. Zhang, Z. Niu, M. Wang, and S. Yan, "Deep aging face verification with large gaps," *IEEE Trans. Multimedia*, vol. 18, no. 1, pp. 64–75, Jan. 2016.
[8] G. Gao, J. Yang, X.-Y. Jing, F. Shen, W. Yang, and D. Yue, "Learning robust and discriminative low-rank representations for face recognition with occlusion," *Pattern Recognit.*, vol. 66, pp. 129–143, Jun. 2017.
[9] C. Jing, Z. Dong, M. Pei, and Y. Jia, "Heterogeneous hashing network for face retrieval across image and video domains," *IEEE Trans. Multimedia*, vol. 21, no. 3, pp. 782–794, Mar. 2019.
[10] M. Yang, W. Wen, X. Wang, L. Shen, and G. Gao, "Adaptive convolution local and global learning for class-level joint representation of facial recognition with a single sample per data subject," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2469–2484, 2020.

[11] Z. Wang *et al.*, "Exploiting temporal and depth information for multi-frame face anti-spoofing," 2018, *arXiv:1811.05118*. [Online]. Available: http://arxiv.org/abs/1811.05118

[12] X. Zhu *et al.*, "Large-scale bisample learning on ID versus spot face recognition," *Int. J. Comput. Vis.*, vol. 127, nos. 6–7, pp. 684–700, Jun. 2019.

[13] S. P. Mudunuri, S. Sanyal, and S. Biswas, "GenLR-net: Deep framework for very low resolution face and object recognition with generalization to unseen categories," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 489–498.

[14] O. A. Aghdam, B. Bozorgtabar, H. K. Ekenel, and J.-P. Thiran, "Exploring factors for improving low resolution face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–8.

[15] M. Li, Z. Zhang, G. Xie, and J. Yu, "A deep learning approach for face hallucination guided by facial boundary responses," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 1, pp. 1–23, Apr. 2020.

[16] S. Ge, S. Zhao, X. Gao, and J. Li, "Fewer-shots and lower-resolutions: Towards ultrafast face recognition in the wild," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 229–237.

[17] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "FSRNet: End-to-end learning face super-resolution with facial priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2492–2501.

[18] X. Hu, P. Ma, Z. Mai, S. Peng, Z. Yang, and L. Wang, "Face hallucination from low quality images using definition-scalable inference," *Pattern Recognit.*, vol. 94, pp. 110–121, Oct. 2019.

[19] M. Singh, S. Nagpal, R. Singh, and M. Vatsa, "Dual directed capsule network for very low resolution image recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 340–349.

[20] H. Yu *et al.*, "Computed tomography super-resolution using convolutional neural networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3944–3948.

[21] G. Gao, Y. Yu, J. Xie, J. Yang, M. Yang, and J. Zhang, "Constructing multilayer locality-constrained matrix regression framework for noise robust face super-resolution," *Pattern Recognit.*, vol. 110, Feb. 2021, Art. no. 107539.

[22] J. Jiang, R. Hu, Z. Wang, and Z. Han, "Noise robust face hallucination via locality-constrained representation," *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1268–1281, Aug. 2014.

[23] L. Liu, S. Li, and C. L. Philip Chen, "Iterative relaxed collaborative representation with adaptive weights learning for noise robust face hallucination," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 5, pp. 1284–1295, May 2019.

[24] J. Jiang, Y. Yu, S. Tang, J. Ma, A. Aizawa, and K. Aizawa, "Context-patch face hallucination based on thresholding locality-constrained representation and reproducing learning," *IEEE Trans. Cybern.*, vol. 50, no. 1, pp. 324–337, Jan. 2020.

[25] J. Jiang, Y. Yu, J. Hu, S. Tang, and J. Ma, "Deep CNN denoiser and multi-layer neighbor component embedding for face hallucination," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 771–778.

[26] Y. Song, J. Zhang, S. He, L. Bao, and Q. Yang, "Learning to hallucinate face images via component generation and enhancement," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 4537–4543.

[27] K. Zhang *et al.*, "Super-identity convolutional neural network for face hallucination," in *Proc. ECCV*, 2018, pp. 183–198.

[28] C.-C. Hsu, C.-W. Lin, W.-T. Su, and G. Cheung, "SiGAN: Siamese generative adversarial network for identity-preserving face hallucination," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 6225–6236, Dec. 2019.

[29] K. Grm, W. J. Scheirer, and V. Struc, "Face hallucination using cascaded super-resolution and identity priors," *IEEE Trans. Image Process.*, vol. 29, pp. 2150–2165, 2020.

[30] J. Shi and G. Zhao, "Face hallucination via coarse-to-fine recursive kernel regression structure," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2223–2236, Sep. 2019.

[31] C.-X. Ren, D.-Q. Dai, and H. Yan, "Coupled kernel embedding for low-resolution face image recognition," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3770–3783, Aug. 2012.

[32] M. Jian and K.-M. Lam, "Simultaneous hallucination and recognition of low-resolution faces based on singular value decomposition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 11, pp. 1761–1772, Nov. 2015.

[33] M. Haghighat and M. Abdel-Mottaleb, "Low resolution face recognition in surveillance systems using discriminant correlation analysis," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 912–917.

[34] S. P. Mudunuri, S. Venkataramanan, and S. Biswas, "Dictionary alignment with re-ranking for low-resolution NIR-VIS face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 4, pp. 886–896, Apr. 2019.

[35] S. Biswas, G. Aggarwal, P. J. Flynn, and K. W. Bowyer, "Pose-robust recognition of low-resolution face images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 3037–3049, Dec. 2013.

[36] S. P. Mudunuri and S. Biswas, "Low resolution face recognition across variations in pose and illumination," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 1034–1040, May 2016.

[37] F. Yang, W. Yang, R. Gao, and Q. Liao, "Discriminative multidimensional scaling for low-resolution face recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 3, pp. 388–392, Mar. 2018.

[38] X. Li, W.-S. Zheng, X. Wang, T. Xiang, and S. Gong, "Multi-scale learning for low-resolution person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3765–3773.

[39] D. Zeng, H. Chen, and Q. Zhao, "Towards resolution invariant face recognition in uncontrolled scenarios," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2016, pp. 1–8.

[40] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang, "Studying very low resolution recognition using deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4792–4800.

[41] 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. 2016, pp. 770–778.

[42] Z. Lu, X. Jiang, and A. Kot, "Deep coupled ResNet for low-resolution face recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 4, pp. 526–530, Apr. 2018.

[43] Z. Wang, M. Ye, F. Yang, X. Bai, and S. Satoh, "Cascaded SR-GAN for scale-adaptive low resolution person re-identification," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 3891–3897.

[44] P. Li, L. Prieto, D. Mery, and P. J. Flynn, "On low-resolution face recognition in the wild: Comparisons and new techniques," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 8, pp. 2000–2012, Aug. 2019.

[45] G.-J. Qi, L. Zhang, H. Hu, M. Edraki, J. Wang, and X.-S. Hua, "Global versus localized generative adversarial nets," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1517–1525.

[46] S. Ge, S. Zhao, C. Li, and J. Li, "Low-resolution face recognition in the wild via selective knowledge distillation," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 2051–2062, Apr. 2019.

[47] Y. Zhao, Z. Jin, G.-J. Qi, H. Lu, and X.-S. Hua, "An adversarial approach to hard triplet generation," in *Proc. ECCV*, 2018, pp. 501–517.

[48] G.-J. Qi, "Hierarchically gated deep networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2267–2275.

[49] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, and H.-J. Zhang, "Image classification with kernelized spatial-context," *IEEE Trans. Multimedia*, vol. 12, no. 4, pp. 278–287, Jun. 2010.

[50] X. Shu, J. Tang, G.-J. Qi, Z. Li, Y.-G. Jiang, and S. Yan, "Image classification with tailored fine-grained dictionaries," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 2, pp. 454–467, Feb. 2018.

[51] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in *Proc. ECCV*, 2018, pp. 517–532.

[52] J. Yang, L. Luo, J. Qian, Y. Tai, F. Zhang, and Y. Xu, "Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 156–171, Jan. 2017.

[53] H. Yu, X. Chen, H. Shi, T. Chen, T. S. Huang, and S. Sun, "Motion pyramid networks for accurate and efficient cardiac motion estimation," in *Proc. MICCAI*. Cham, Switzerland: Springer, 2020, pp. 436–446.

[54] K. Koh, S.-J. Kim, and S. Boyd, "An interior-point method for large-scale l1-regularized logistic regression," *J. Mach. Learn. Res.*, vol. 8, pp. 1519–1555, Jul. 2007.

[55] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014, *arXiv:1411.7923*. [Online]. Available: http://arxiv.org/abs/1411.7923

[56] A. Sapkota and T. E. Boult, "Large scale unconstrained open set face database," in *Proc. IEEE 6th Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS)*, Sep. 2013, pp. 1–8.

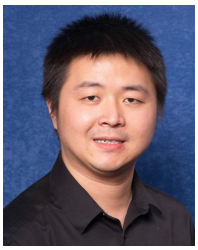
[57] J. Huo, Y. Gao, Y. Shi, W. Yang, and H. Yin, "Ensemble of sparse cross-modal metrics for heterogeneous face recognition," in *Proc. ACM Multimedia Conf. (MM)*, 2016, pp. 1405–1414.

[58] M. Grgic, K. Delac, and S. Grgic, "SCface—Surveillance cameras face database," *Multimedia Tools Appl.*, vol. 51, no. 3, pp. 863–879, Feb. 2011.

Guangwei Gao (Member, IEEE) received the B.S. degree in information and computation science from Nanjing Normal University, Nanjing, China, in 2009, and the Ph.D. degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology, Nanjing, in 2014. He was an Exchange Student of the Department of Computing, The Hong Kong Polytechnic University, in 2011 and 2013. He is currently an Associate Professor with the Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, and also a Project Researcher with the Digital Content and Media Sciences Research Division, National Institute of Informatics (NII), Japan. His research interests include pattern recognition and computer vision. He has served as a Reviewer for the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (TNNLS), the IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), the IEEE TRANSACTIONS ON MULTIMEDIA (TMM), the IEEE TRANSACTIONS ON CYBERNETICS (TCYB), *Pattern Recognition*, *Neurocomputing*, *Patter Recognition Letter*, and AAAI/ICPR/ICIP.

Yi Yu (Member, IEEE) received the Ph.D. degree in information and computer science from Nara Women's University, Japan. She was a Senior Research Fellow with the School of Computing, National University of Singapore. She is currently an Assistant Professor with the National Institute of Informatics (NII), Japan. Her research interests include large-scale multimedia data mining and pattern analysis, location-based mobile media service, and social media analysis. She and her team received the best Paper Award from the IEEE ISM 2012, the 2nd prize in Yahoo Flickr Grand Challenge 2015, were in the top winners (out of 29 teams) from ACM SIGSPATIAL GIS Cup 2013, and the Best Paper Runner-Up in APWeb-WAIM 2017, recognized as finalist of the World's FIRST 10K Best Paper Award in ICME 2017.

Jian Yang (Member, IEEE) received the Ph.D. degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology (NUST) in 2002. In 2003, he was a Post-Doctoral Researcher with the University of Zaragoza. From 2004 to 2006, he was a Post-Doctoral Fellow with the Biometrics Centre, The Hong Kong Polytechnic University. From 2006 to 2007, he was a Post-Doctoral Fellow with the Department of Computer Science, New Jersey Institute of Technology. He is currently a Chang-Jiang Professor with the School of Computer Science and Engineering, NUST. He has authored more than 100 scientific articles in pattern recognition and computer vision. His articles have been cited more than 4000 times in the Web of Science, and 9000 times in the Scholar Google. His research interests include pattern recognition, computer vision, and machine learning. He is a fellow of IAPR. He is/was an Associate Editor of *Pattern Recognition Letters*, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and *Neurocomputing*.



Guo-Jun Qi (Senior Member, IEEE) received the Ph.D. degree from the University of Illinois at Urbana–Champaign in 2013. He is currently a Faculty Member with the Department of Computer Science, University of Central Florida. His research interests include pattern recognition, machine learning, computer vision, multimedia, and data mining. He has served as a program committee member and a reviewer for many academic conferences and journals in the fields of pattern recognition, machine learning, data mining, computer vision, and multimedia. He was a recipient of the IBM Ph.D. fellowships for two times and the Microsoft Fellowship. He received the Best Paper Award at the 15th ACM International Conference on Multimedia, Augsburg, Germany, in 2007.

Meng Yang (Senior Member, IEEE) received the Ph.D. degree from The Hong Kong Polytechnic University in 2012. He was a Post-Doctoral Fellow with the Computer Vision Lab, ETH Zurich. He is currently an Associate Professor with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He has published about 90 academic articles, including 14 CVPR/ICCV/AAAI/IJCAI/ICML/ECCV articles and several IJCV, IEEE TNNLS, TIP, and TIFs journal articles. Now, his Google citation is more than 7800. His research interests include computer vision, sparse coding and dictionary learning, natural language processing, and machine learning.