

Received April 21, 2019, accepted May 21, 2019, date of publication June 5, 2019, date of current version June 21, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2920855

Heterogeneous Face Recognition Based on Multiple Deep Networks With Scatter Loss and Diversity Combination

WEIPENG HU, HAIFENG HU^{ID}, (Member, IEEE), AND XINLONG LU

School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, China

Corresponding author: Haifeng Hu (huhaif@mail.sysu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61673402, Grant 61273270, and Grant 60802069, in part by the National Key R&D Program of China under Grant 2018YFB1601101, in part by the Natural Science Foundation of Guangdong Province under Grant 2017A030311029, in part by the Science and Technology Program of Guangzhou of China under Grant 201704020180, and in part by the Fundamental Research Funds for the Central Universities of China.

ABSTRACT Due to the gap between sensing patterns of different domains and a lack of sufficient training sample, heterogeneous face recognition (HFR) is still a challenging issue in the computer vision community. In this paper, we propose a novel method called multiple deep networks with scatter loss and diversity combination (MDNDC) for solving the HFR problem. As we know, the performance of deep models is affected by data, network structure, and loss function, so we devote much effort to improve the HFR performance from all these three aspects. First, to reduce the intra-class variations and increase the inter-class variations, the scatter loss (SL) is used as an objective function that can bridge the modality gap while preserving the identity information. Second, we design a multiple deep networks (MDN) structure for feature extraction and propose a joint decision strategy called diversity combination (DC) to adaptively adjust the weights of each deep network and make a joint classification decision. Finally, instead of using only one publicly available dataset, we make full use of multiple datasets to train the networks, which can further improve the HFR performance. The extensive experiments are carried out on two challenging NIR-VIS HFR datasets, CASIA NIR-VIS 2.0 and Oulu-CASIA NIR-VIS, demonstrating the superiority of the proposed method.

INDEX TERMS Heterogeneous face recognition, multiple deep networks, scatter loss, diversity combination.

I. INTRODUCTION

Heterogeneous Face Recognition (HFR) is a challenging problem in face recognition community, referring to matching a probe face image with the face image gallery taken from alternate imaging modality. The major difficulties of HFR lie in the large discrepancies among different image modalities, such as identity related information, modality related information, face variations (e.g., lightings, poses, and expressions), etc.

In recent decades, various approaches have been proposed to address the challenging HFR problem, which can be classified into four categories: synthesis-based model [36], [43], coupled subspace learning [1], [47], feature representation [15], [23] and deep learning methods [11], [18], [28], [35]. Synthesis-based methods try to synthesize an

image from its original modality by dealing with the modality difference at the stage of image preprocessing. Coupled subspace learning based methods map multimodal data into a common feature space to eliminate the large discrepancies of cross-modality image pairs. The feature representation methods aim to reduce the modality discrepancy at the feature extraction level.

Recently, deep convolutional neural networks (CNNs) [11], [18], [28], [39], [42] have achieved outstanding performance for HFR tasks, and they map the face image into an embedding feature vector such that distance among features of the same person is relatively small but that of different individuals can be very large. Various heterogeneous face recognition approaches using deep CNNs embedding differ mainly in three following aspects of property [6], [19], [22]. The first property is the training data employed to train the network. The identity number of publicly available training data, such as VGG-Face [26], VGG2-Face [27],

The associate editor coordinating the review of this manuscript and approving it for publication was Siddhartha Bhattacharyya.

CAISA WebFace [16] and MS-Celeb-1M [44], Oulu-CASIA NIR-VIS [33] and CASIA NIR-VIS 2.0 [32], ranges from several thousand to half million. Saxena and Verbeek [31] study different aspects of leveraging a CNN pre-trained on CAISA WebFace for heterogeneous face recognition. Wu *et al.* [40] propose a Light CNN model which is trained on MS-Celeb-1M, and improve the best rank-1 accuracy from 95.82% to 96.72% in CASIA NIR-VIS 2.0 dataset. The second property is the network architecture and settings. To our best knowledge, high capacity deep convolutional neural networks, such as Inception-ResNet [12], can obtain better performance compared to VGG network [26] and Google Inception V1 network [13], [17]. The third property is the design of the loss functions [7], [20]. On the one hand, the classification-based methods [26], [46] suffer from massive GPU memory consumption on the classification layer when the identity number increases to million orders of magnitude, and they prefer balanced and sufficient training data for each identity [22]. On the other hand, the Triplet loss [17] utilizes pair training strategy, which minimizes the distance between an anchor and a positive sample and maximizes the distance between the anchor and a negative sample from a different identity. However, the training procedure of the Triplet loss [17] is tricky due to the selection of effective training samples.

As is well known that all the aforementioned factors have a high-to-low influence on the performance of HFR models. In this paper, we devote much effort to improving deep HFR from all three properties above, and here we state our contributions in the following aspects.

Network: The individual features extracted by a single deep network are less discriminative and robust compared to the fusion features of multiple parallel deep networks. Therefore, we design a Multiple Deep Networks (MDN) structure for feature extraction, and propose a joint decision strategy called Diversity Combination (DC) to adaptively adjust weights of each deep network and make a joint classification decision.

Data: As is well known that the data has a great influence on the performance of face recognition models. Therefore, instead of using only one publicly available dataset, we make full use of multiple datasets VGG2-Face, CAISA WebFace and MS-Celeb-1M to train our Multiple Deep Networks. Through plenty of experiments, we find that HFR can be further improved through multiple datasets.

Loss: The Scatter Loss (SL) is used to learn discriminative features for robust heterogeneous face recognition [8]. It shows that SL not only has a clearer mathematical expression of the inter-class and intra-class constraint but also outperforms the baseline methods (e.g. Centre loss [45] and Triplet loss [17]) in HFR tasks.

Performance: The proposed MDNDC achieves superior performance on the largest public heterogeneous face benchmark datasets CASIA NIR-VIS 2.0 [32] and Oulu-CASIA NIR-VIS [33].

The remainder of this paper is organized as follows. Section 2 provides a brief review of related works on HFR. Section 3 presents the formulation of our MDNDC model. Section 4 evaluates the performance of our method using two benchmark datasets. And we conclude the paper in Section 5.

II. RELATED WORK

HFR is widely applied in video surveillance systems and user authentication applications nowadays. Therefore, the demand for robust matching between cross-modality data has increased sharply and attracted considerable attention. Here, we briefly describe the recent works on this related subject and generally categorize these works into four classes: synthesis-based model, coupled subspace learning, feature representation and deep learning methods.

Synthesis-based methods [21], [36], [41], [43] aim at transforming face images from one modality into another via image synthesis, and they attempt to address the modality difference at the image preprocessing stage. Song *et al.* [43] propose a Real-Time Exemplar-Based Face Sketch Synthesis method that can effectively produce face sketches in real-time speed, in which Markov Random Field based methods can be formulated as the baseline to improve the performance. To synthesize the normal face from the infrared input, Li *et al.* [21] propose a learning-based framework which exploit the local linearity in both the image manifolds and image spatial domain, and they further apply the Markov random field model to the predicted normal face to improve the hallucination result. Tang and Wang [41] propose a sketch synthesis method based on separate transformation of photo texture and shape, and design a Bayesian classifier to recognize the probing real sketch from the synthesized pseudo sketches. To reduce the modality gap at the feature extraction stage, Zhang *et al.* [36] propose an inter-modality face recognition approach in which a new face descriptor based on coupled information-theoretic encoding is used to capture discriminative local face structures and effectively match photos and sketches. However, the synthesis process is actually more difficult than recognition and the performance of these methods heavily depends on the fidelity of the synthesized images [5].

Coupled subspace learning based methods [1], [25], [37], [47] map multimodal data into a common feature space to eliminate the large discrepancies of cross-modality image pairs. Huang *et al.* [37] propose a regularized discriminative spectral regression method for heterogeneous face matching, and introduce two regularization terms that can effectively make use of the class information in the training set. Kan *et al.* [25] propose a Multi-view Discriminant Analysis (MvDA) method that can learn single unified discriminant common space in which the between-class variations from both inter-view and intra-view are maximized, while the within-class variations from both inter-view and intra-view are minimized. Sharma and Jacobs [1] propose Partial Least Squares (PLS) to match probe and gallery images by linearly

projecting them into an intermediate space where images with the same identity are highly correlated. The main problem of coupled subspace learning models is that the projection procedure always causes information loss and thus decreases the recognition performance [5].

The feature representation methods aim to reduce the modality discrepancy at the feature extraction level [2], [4], [15], [23]. Klare *et al.* [4] present a framework called Local Feature-based Discriminant Analysis (LFDA) in which they individually represent both sketches and photos using SIFT feature descriptors and Multiscale Local Binary Patterns (MLBP). Alex *et al.* [2] propose a face descriptor called Local Difference of Gaussian Binary Pattern (LDoGBP) to encode the DoG representation of the image into a binary pattern, which offers a higher recognition rate with very low computational complexity. Lu *et al.* [23] propose a simultaneous local binary feature learning and encoding (SLBFLE) approach for HFR which jointly learns binary codes and the codebook for local face patches so that discriminative information can be obtained by using a one-stage feature learning and encoding procedure. Gong *et al.* [15] propose a new feature descriptor called common encoding model which is able to capture common discriminant information, such that the large modality gap can be significantly reduced at the feature extraction stage. Feature extraction methods reduce the modality variations when converting heterogeneous images to features, and are often applied along with subspace learning methods.

Deep learning methods attempted to transfer the knowledge learned on a large scale visible dataset to heterogeneous dataset. Lezama *et al.* [18] propose a cross-spectral hallucination and low-rank embedding method for HFR, in which cross-spectral hallucination produces VIS faces from NIR images through a deep learning approach and Low-rank embedding restores a low-rank structure for faces deep features across both NIR and VIS spectrum. Liu *et al.* [42] present a deep Transfer NIR-VIS heterogeneous face recognition network (TRIVET) for HFR, and come up with two types of NIR-VIS triplet loss to reduce intra-class NIR-VIS variations, and meanwhile augment positive training sample pairs. Wu *et al.* [39] propose a Coupled Deep Learning (CDL) approach to address the VIS-NIR heterogeneous matching problem, and introduce a trace norm to the softmax loss of convolutional neural networks, which enhances the relevance of projection matrices of NIR and VIS images so that the difference between modalities is reduced and identity information is reserved. Saxena and Verbeek [31] study different aspects of leveraging a CNN pre-trained on visible spectrum images for heterogeneous face recognition, including extracting features from different CNN layers, fine-tuning the CNN, and using various forms of metric learning. He *et al.* [28] develop an effective deep neural network architecture to learn modality invariant representation, and divide the high level layer into two orthogonal subspaces that contain modality-invariant identity information and modality-variant spectrum information respectively. Wu *et al.* [40] present a light CNN framework to learn a

compact embedding on the large-scale face data with massive noisy labels, and introduce a new activation function Max-Feature-Map (MFM) which can not only separate noisy signals and informative signals but also plays a role of feature selection. He *et al.* [29] propose a Wasserstein convolutional neural network (WCNN) approach for learning invariant features between the two different modalities, where Wasserstein distance is introduced into the shared layer to measure the dissimilarity between heterogeneous feature distributions and correlation prior is introduced on the fully-connected WCNN layers to reduce the size of the parameter space. Experimental results suggest that deep learning methods have the potential to outperform traditional HFR methods. However, because of the high intra-class variations between heterogeneous data and the over-fitting problems, HFR task is still challenging for deep learning methods [28], [29].

The Scatter Loss was first proposed in our early work [8]. In addition to providing more in-depth analysis and extensive experiments, the main difference between this work and our previous work [8] is the introduction of MDN structure, multiple datasets training trick and Diversity Combination. In our work, we compare the performance of different loss functions. The results show that Scatter Loss is more effective in HFR task. To overcome the over-fitting problem and improve HFR performance, we adopt multiple datasets VGG2-Face, CAISA WebFace and MS-Celeb-1M to pre-train Multiple Deep Networks and then fine-tune the network on HFR datasets. To further reduce modality variations and extracting robust features, we propose Diversity Combination for joint decision of MDN network. The codes have been released in the project page¹.

III. THE PROPOSED MDNDC APPROACH

This section details the proposed MDNDC model, whose flowchart is shown in Fig.1. In our model, multiple parallel deep networks are used for feature extraction with Scatter Loss, and each deep network is trained by one of the VGG2-Face, CAISA WebFace and MS-Celeb-1M datasets. At the same time, DC fusion strategy is implemented to adaptively adjust the network weights and make a joint classification decision.

A. MULTIPLE DEEP NETWORKS

As we know, multiple networks can improve the performance of HFR. In our work, we adopt three networks to build the architecture of MDN, each of which is trained with different datasets, so an appropriate fusion strategy can make the fusion feature of three networks more discriminant than one of the single network does. In our model, we adopt Inception-ResNet-v1 [12] as the backbone network, because it is a delicately designed network and can obtain better performance compared to VGG network [24], [26] and Google Inception V1 network [13], [17]. And the three backbone networks are

¹<https://github.com/HWPengSYSU1993/MDNDC-model>

arranged in parallel to build our Multiple Deep Networks structure.

We assume that the training data are denoted as $\{x_i, l_i\} \ i \in \{1, 2, \dots, n\}$, where x_i is the image data and l_i is the image label. The backbone network embedding is represented by $f^k(x_i) \in \mathbb{R}^d \ k \in \{1, 2, 3\}$, where k represents the index of the backbone network. It embeds an image x_i into a d -dimensional Euclidean space. Additionally, we constrain this embedding $f^k(x_i)$ to stay in the d -dimensional hypersphere, i.e. $\|f^k(x_i)\|_2 = 1$.

B. SCATTER LOSS

As we already described, the Softmax-based loss [26], [46] suffers from massive GPU memory consumption on the classification layer with the identity number reaching the million orders of magnitude, and it prefers balanced and sufficient training data for each identity. Moreover, the Triplet loss [17] minimizes the distance between an anchor and a positive sample and also maximizes the distance between the anchor and a negative sample from a different identity. Nevertheless, the training procedure of the Triplet loss [17] is tricky due to the selection of effective training samples. References [8] and [9] maximize distance between the classes and minimize distance within the class to obtain discriminative features. In our work, we adopt the SL loss [8] to train the backbone network, because it has the following three advantages. Firstly, the SL does not contain the classification layer [46], which reduces the number of network parameters and the GPU memory consumption. Secondly, SL can still obtain good performance when the samples in each batch are randomly selected. Finally, compared with triplet loss, the SL loss is more robust to noise samples and more stable during training process. The SL objective function can be expressed as follows:

$$L = \min(S_W + S_B) \quad (1)$$

where

$$S_W = \frac{1}{n} \sum_{r=1}^c \sum_{l_i=r} \|f(x_i) - m_r\|_2^2 \quad (2)$$

$$S_B = \frac{1}{c^2 - c} \sum_{i,j=1, i < j} [\alpha - \|m_i - m_j\|_2^2]_+ \quad (3)$$

$$m_r = \frac{1}{n_r} \sum_{k=1}^n \mathbf{I}(l_k = r) f(x_k) \quad (4)$$

$$m_i = \frac{1}{n_i} \sum_{k=1}^n \mathbf{I}(l_k = i) f(x_k) \quad (5)$$

$$m_j = \frac{1}{n_j} \sum_{k=1}^n \mathbf{I}(l_k = j) f(x_k) \quad (6)$$

where m_i , m_j and m_r are the mean value of the i^{th} , j^{th} , r^{th} class in a mini-batch, n_i , n_j and n_r are the number of samples of the i^{th} , j^{th} , r^{th} class in a mini-batch. $\mathbf{I}()$ is the indicator function with value of 1 or 0, $f(x_i)$ is the backbone network

embedding-layer features for the i^{th} sample, c is the number of class and n is the number of samples in a mini-batch, S_W represents the distance within the class while S_B represents the distance between the classes, and α is a margin value between classes. In the training stage, m is calculated in each mini-batch according to Eq. (4-6). In the testing stage, $f(x_i)$ is the extracting features of input image x_i for recognition.

The gradient descent following the chain rule is employed to optimize Scatter Loss in CNN. Specifically, we firstly calculate the loss of objective in Eq. (1), and then propagate the loss so as to compute the gradient of each layer [8], [10]. Finally we employ gradient descent to update the whole network. The gradient L w.r.t. $f(x_i)$ can be calculated as:

$$\frac{\partial L}{\partial f(x_i)} = \frac{\partial L}{\partial S_W} \frac{\partial S_W}{\partial f(x_i)} + \frac{\partial L}{\partial S_B} \frac{\partial S_B}{\partial f(x_i)} \quad (7)$$

where

$$\begin{aligned} \frac{\partial L}{\partial S_W} \frac{\partial S_W}{\partial f(x_i)} &= \frac{1}{n} (2 - \frac{2}{n_i}) (f(x_i) - m_i) \\ \frac{\partial L}{\partial S_B} \frac{\partial S_B}{\partial f(x_i)} &= \sum_{j=1, j \neq i}^c \delta_j \\ \delta_j &= \begin{cases} 0, & \text{if } \|m_i - m_j\|_2^2 > \alpha \\ \frac{2}{n_i} (m_j - m_i), & \text{otherwise} \end{cases} \end{aligned}$$

where n_i is the number of samples in the i^{th} class. The optimization process of network is carried out using mini-batch, each of which contains several individuals, and each individual contains the same number of samples.

C. DIVERSITY COMBINATION STRATEGY

In this section, we propose Diversity Combination as the fusion strategy for multiple parallel deep networks. As illustrated in Fig. 1, the DC method aims at adaptively learning a weight parameter η^k ($k = 1, 2, 3$) for each backbone network and making classification decisions. We assume that the testing set contains gallery data $\{x_i^G, l_i^G\}$ and probe data $\{x_i^P, \text{none}\}$, where l_i^G is the label of the i^{th} gallery sample. Our goal is to infer the label of probe sample x_i^P . The DC method calculates the distance $\{\bar{d}_{i1}, \bar{d}_{i2}, \dots, \bar{d}_{ic_g}\}$ of a probe sample x_i^P between all galleries $\{x_1^G, x_2^G, \dots, x_{c_g}^G\}$ and returns the closest distance $\bar{d}_{i\theta}$:

$$\bar{d}_{i\theta} = \min\{\bar{d}_{i1}, \bar{d}_{i2}, \dots, \bar{d}_{ic_g}\} \quad (8)$$

where

$$\begin{aligned} \bar{d}_{ij} &= \eta^1 d_{ij}^1 + \eta^2 d_{ij}^2 + \eta^3 d_{ij}^3 \quad (j = 1, 2, \dots, c_g) \\ d_{ij}^k &= D(f^k(x_i^P), f^k(x_j^G)) \quad (k = 1, 2, 3; j = 1, 2, \dots, c_g) \end{aligned}$$

where c_g is the number of samples in gallery set, η^k ($k = 1, 2, 3$) is the corresponding weight of the k^{th} backbone network, $f^k(\cdot)$ is the embedding-layer features of the k^{th} backbone network, $D()$ is used to calculate the distance between two samples (such as Euclidean distance and Cosine distance), d_{ij}^k ($k = 1, 2, 3$) is the distance between the i^{th} probe

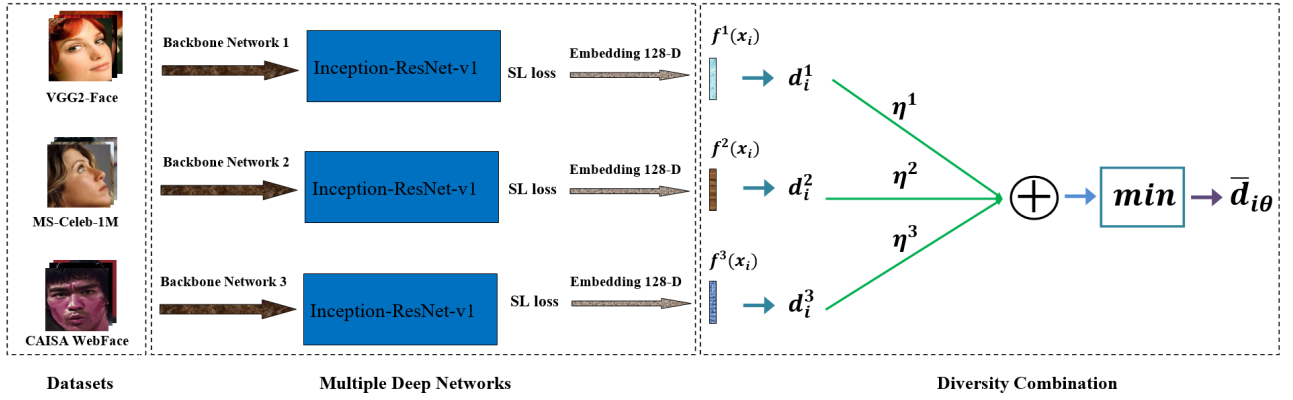


FIGURE 1. The MDNDC model includes three backbone network (Inception-ResNet-v1), and each network is trained by one of the VGG2-Face, CAISA WebFace and MS-Celeb-1M datasets. The proposed DC is used to adaptively adjust weights of each deep network and make a joint classification decision.

and the j^{th} gallery on the k^{th} backbone network, and $\bar{d}_{ij}$ is the weighted distance between the i^{th} probe and the j^{th} gallery calculated by DC. In particular, the identity of the probe x_i^P is determined by the closest distance $\bar{d}_{i\theta}$, i.e., $l_i^P = l_\theta^G$.

In our applications, η^k is obtained through the supervised manner in the training stage. We assume that the training set contains VIS data $\{x_i^V, l_i^V\}$ and NIR data $\{x_i^N, l_i^N\}$. The distance $\{d_{i1}^k, d_{i2}^k, \dots, d_{ic_v}^k\}$ between a NIR x_i^N image and all VIS images $\{x_1^V, x_2^V, \dots, x_{c_v}^V\}$ can be expressed as:

$$d_{ij}^k = D(f^k(x_i^N), f^k(x_j^V)) \quad (k = 1, 2, 3; j = 1, 2, \dots, c_v) \quad (9)$$

where c_v is the number of samples in VIS data. Here we denote that $d_i^k = [d_{i1}^k, d_{i2}^k, \dots, d_{ic_v}^k]$. We sort d_i^k element-wise, and obtain $r_i^k = [r_{i1}^k, r_{i2}^k, \dots, r_{ic_v}^k]$, $r_{ij}^k \in \{1, 2, \dots, c_v\}$. Then we get:

$$s_i^k = r_{ij}^k, \quad \text{if } l_i^N = l_j^V \quad (10)$$

where l_i^N and l_j^V are the label of the i^{th} NIR sample and the j^{th} VIS sample, respectively. s_i^k indicates that the similarity between NIR and VIS with same identity. In our method, the better generalization and recognition ability the backbone network has, the larger weights η^k should be. Therefore, the weight distribution can be obtained by the following formula:

$$\begin{aligned} \arg \min_{\lambda_i^1, \lambda_i^2, \lambda_i^3} \sum_{i=1}^{c_n} \|\lambda_i^1 s_i^1 + \lambda_i^2 s_i^2 + \lambda_i^3 s_i^3\|_2^2 \\ \text{s.t. } \lambda_i^1 + \lambda_i^2 + \lambda_i^3 = 1 \quad (i = 1, 2, \dots, c_n) \\ \text{s.t. } \lambda_i^k \geq 0 \quad (i = 1, 2, \dots, c_n; k = 1, 2, 3) \end{aligned} \quad (11)$$

where c_n is the number of samples in NIR data, and λ_i^k denotes a rewards to the k^{th} backbone network for each NIR sample x_i^N . The weights distribution η^k of each backbone network can be obtained by:

$$\eta^k = \frac{(\delta^k)^\tau}{\sum_i (\delta^i)^\tau} \quad (k = 1, 2, 3) \quad (12)$$

where

$$\delta^k = \frac{1}{c_n} \sum_i \lambda_i^k \quad (k = 1, 2, 3)$$

where τ is a scale factor used to scale the difference between η^1, η^2 and η^3 . In the experiment, we set $\tau = 3$.

In particular, we use Joint Bayesian [14] to calculate the distance $D()$ between two samples. Note that the training procedure for backbone network and DC are independent. Specifically, we first train the three backbone networks, and then following with the DC.

D. OPTIMIZATION FOR DIVERSITY COMBINATION

In order to find the optimal solution for the weights distribution η^k , we apply the Lagrangian multiplier to solve the constrained convex optimization problem. Firstly, the optimization problem in Eq. (11) can be split into c_n convex optimization problems, each of which has the following form:

$$\begin{aligned} \arg \min_{\lambda_i^1, \lambda_i^2, \lambda_i^3} \|\lambda_i^1 s_i^1 + \lambda_i^2 s_i^2 + \lambda_i^3 s_i^3\|_2^2 \\ \text{s.t. } \lambda_i^1 + \lambda_i^2 + \lambda_i^3 = 1 \\ \text{s.t. } \lambda_i^1, \lambda_i^2, \lambda_i^3 \geq 0 \end{aligned} \quad (13)$$

According to the Lagrangian multiplier, Eq. (13) can be transformed to an unconstrained optimization problem:

$$\begin{aligned} L(\lambda_i, \alpha, \beta) = \|\lambda_i^1 s_i^1 + \lambda_i^2 s_i^2 + \lambda_i^3 s_i^3\|_2^2 \\ + \alpha \left(\sum_{k=1}^3 \lambda_i^k - \frac{1}{3} \right) + \sum_{k=1}^3 \beta^k \lambda_i^k \end{aligned} \quad (14)$$

The feasible solution of $\lambda_i = [\lambda_i^1, \lambda_i^2, \lambda_i^3]$ needs to satisfy the KKT condition:

$$\begin{aligned} \nabla_{\lambda_i} L(\lambda_i, \alpha, \beta) &= 0 \\ \lambda_i^1 + \lambda_i^2 + \lambda_i^3 &= 1 \\ \lambda_i^1, \lambda_i^2, \lambda_i^3 &\geq 0 \\ \beta^k \lambda_i^k &= 0 \quad (k = 1, 2, 3) \\ \beta^k &\leq 0 \quad (k = 1, 2, 3) \end{aligned} \quad (15)$$

The solution of λ_i in Eq. (13) can be obtained by minimizing $L(\lambda_i, \alpha, \beta)$ under KKT conditions. Specifically, the optimization problem in Eq. (11) can be solve by solving c_n convex optimization subproblems in Eq. (13).

E. DATA

Instead of using only one publicly available dataset, we make full use of multiple datasets VGG2-Face, CAISA WebFace and MS-Celeb-1M to train our Multiple Deep Networks. Specifically, each backbone network is first trained by one of the above datasets, and then fine-tunes in HFR datasets such as Oulu-CASIA NIR-VIS and CASIA NIR-VIS 2.0. HFR performance can be further improved through multiple VIS datasets.

F. ALGORITHM

The training process of the MDNDC network is summarized in Algorithm 1. Firstly, we pre-train each backbone network with Softmax loss by one of the VGG2-Face, CAISA WebFace and MS-Celeb-1M datasets. Secondly, we fine-tune backbone network with Scatter Loss by HFR datasets such as Oulu-CASIA NIR-VIS and CASIA NIR-VIS 2.0. Thirdly, we train the DC to obtain weights distribution η^k in HFR dataset.

Algorithm 1 Training the MDNDC Network

Require: VGG2-Face, MS-Celeb-1M, CAISA WebFace and HFR datasets.

Ensure: The three backbone networks parameters, weights distribution η^k ($k = 1, 2, 3$).

- 1: Pre-train the first backbone network by VGG2-Face;
 - 2: Pre-train the second backbone network by MS-Celeb-1M;
 - 3: Pre-train the third backbone network by CAISA WebFace;
 - 4: Fine-tune the first backbone network with scatter loss in HFR dataset;
 - 5: Fine-tune the second backbone network with scatter loss in HFR dataset;
 - 6: Fine-tune the third backbone network with scatter loss in HFR dataset;
 - 7: Train the DC to obtain weights distribution η^k in HFR dataset.
-

IV. EXPERIMENTS EVALUATION

In this section, a number of experiments are carried out on two biometric applications in support of the following two objectives:

- Investigate the various properties of the MDNDC algorithm.
- Evaluate MDNDC on heterogeneous face recognition problem by comparing performance with some state-of-the-art methods such as Invariant Deep Representation (IDR) [28], TRIVET [42], Coupled

Simultaneous Local Binary Feature Learning and Encoding (C-SLBFLE) [23] and WCNN [29].

- Evaluate our method and some popular models on the computational costs, time cost of a single image to extract feature and the model parameters.

A. DATASETS

Two popular public datasets are used in the experiments, in which both CASIA NIR-VIS 2.0 [32] and Oulu-CASIA NIR-VIS [33] are for NIR-to-VIS face matching.

1) CASIA NIR-VIS 2.0

CASIA NIR-VIS 2.0 dataset [32] consists of 725 subjects in total. There are 1-22 VIS and 5-50 NIR face images per subject. Under the View 2 protocol [32], the evaluation is performed via the tenfold process and in each fold, 357 subjects are used for training while the remaining 358 subjects for testing. In particular, each training fold contains approximately 2,500 VIS images and 6,100 NIR images. In each testing fold, the gallery set contains 358 subjects and each of which only has one VIS image, while the probe set contain the same subjects with over 6,000 NIR images. Rank-1 accuracy and the Receiver Operating Characteristic (ROC) curve are reported according to the protocol. Fig. 2 shows some samples of cropped VIS-NIR image pairs from this dataset.



FIGURE 2. Example images of an individual in the CASIA NIR-VIS 2.0 dataset (VIS top, NIR bottom).

2) OULU-CASIA NIR-VIS DATASET

Oulu-CASIA NIR-VIS dataset [33] contains 80 subjects with variation of six kinds of expression (happiness, sadness, fear, anger, disgust and surprise). Thirty subjects are selected from CASIA and the remaining fifty subjects are selected from Oulu University. Because images are captured under different environments from two universities, the illumination condition is slightly different. Following the protocol in [34], we select a subset that contains forty subjects, and each of which consists of 48 NIR images and 48 VIS images. In particular, we select ten subjects from Oulu, and the remaining thirty subjects are from CASIA. The training fold contains twenty subjects, and it has 960 VIS images and 960 NIR images. The testing fold consists of the remaining 20 subjects, and there are 960 VIS images in the gallery set while 960 NIR images are in the probe set. We report the rank-1 accuracy and the ROC curve according to the protocol. Some example images are shown in Fig.3.

TABLE 1. Rank-1 accuracy of different loss functions on CASIA NIR-VIS 2.0 and Oulu-CASIA NIR-VIS datasets.

Dataset	Loss	Rank-1(%)	VR@FAR=1%(%)	VR@FAR=0.1%(%)
CASIA NIR-VIS 2.0	Softmax loss	84.5 $\pm$ 1.7	90.8 $\pm$ 1.4	76.2 $\pm$ 1.6
	Center loss	84.8 $\pm$ 1.7	91.3 $\pm$ 1.4	77.0 $\pm$ 1.6
	Triplet loss	97.9 $\pm$ 0.5	98.9 $\pm$ 0.5	95.6 $\pm$ 0.8
	Scatter loss	98.5 $\pm$ 0.3	99.2 $\pm$ 0.3	97.0 $\pm$ 0.5
Oulu-CASIA NIR-VIS	Softmax loss	98.0	86.2	50.4
	Center loss	98.1	86.4	50.7
	Triplet loss	94.7	74.5	51.7
	Scatter loss	98.7	87.0	63.5

**FIGURE 3.** Example images of an individual in the Oulu-CASIA NIR-VIS dataset (VIS top, NIR bottom).

B. IMPLEMENTATION DETAILS

The open source deep learning framework tensorflow [50] is employed to train the CNN model on TITAN Xp. The training process for Inception-ResNet-v1 consists of two stages: pre-training and fine-tuning. In the pre-training stage, we first adopt VIS dataset (MS-Celeb-1M, VGG2-Face or CAISA WebFace) to train the Inception-ResNet-v1 with softmax loss function. We use MTCNN [49] to normalize and crop all training images into 160×160 according to five facial points. The batch size and the epoch size are set to 90 and 1000, respectively. We adopt the dropout ratio of 0.7 for the fully connected layer. And the training epochs for three Inception-ResNet-v1 networks pre-trained using MS-Celeb-1M, VGG2-Face and CAISA WebFace data are set to 280, 300 and 100, respectively. The learning rates of the three networks are automatically changed from $1e-1$, $5e-2$ and $5e-2$ to $1e-4$, $5e-4$ and $5e-4$, respectively. In the fine-tuning stage, we adopt HFR dataset (CASIA NIR-VIS 2.0 or Oulu-CASIA NIR-VIS dataset) to train the Inception-ResNet-v1 with SL loss function. Both NIR and VIS images are normalized and cropped into 160×160 according to five facial points. The margin value α is set to 1.5. The batch size and the epoch size are set to 100 (i.e., 20 people per batch, 5 images per person) and 1000, respectively. The training epochs for the network fine-tuned using CASIA NIR-VIS 2.0 and Oulu-CASIA NIR-VIS are about 18 and 12, respectively. The learning rate decay epochs, learning rate decay factor and dropout ratio of the fully connected layer are 4, 0.95 and 0.8, respectively. The initial learning rates of three backbone networks are $1e-3$ in CASIA NIR-VIS 2.0. In Oulu-CASIA NIR-VIS, their initial learning rates are set to $1e-3$, $1e-4$ and $1e-4$, respectively. The training process of each backbone network approximately costs us 5 and 3.5 hours in CASIA NIR-VIS 2.0 and Oulu-CASIA NIR-VIS, respectively. The HFR datasets (CASIA NIR-VIS 2.0 or Oulu-CASIA NIR-VIS) are used

to train the DC to obtain the weight distribution. And the optimization time for DC in the two HFR datasets are about 10 seconds and 6 seconds, respectively. All the parameters in Inception-ResNet-v1 are updated by Adagrad gradient descent optimization algorithm [51].

In particular, the CASIA NIR-VIS 2.0 dataset contains ten sub-experiments [32]. In each sub-experiment, each backbone network is fine-tuned in the training fold. When the training of MND network is completed, we then train DC using the training fold. The training details for three backbone networks and DC are mentioned above. The mean accuracy and standard deviation of ten sub-experiments are evaluated as performance of the algorithm.

C. EMPIRICAL STUDIES OF THE MDNDC PROPERTIES ON TWO NIR-VIS DATASETS

The following properties of MDNDC are studied on the CASIA NIR-VIS 2.0 and Oulu-CASIA NIR-VIS datasets: The effects of the Scatter Loss, the effects of our Multiple Deep Networks structure, and the performance of DC strategy.

1) THE EFFECTS OF SCATTER LOSS

We compare the performance of Inception-ResNet-v1 network with different loss functions. Table 1 shows the rank-1 accuracy and $VR@FAR = 0.1\%$ of different objective functions. In our work, we first pre-train the Inception-ResNet-v1 network in MS-Celeb-1M with Softmax loss, and then fine-tune the network with different loss functions including Softmax, Center loss, Triplet loss and Scatter loss in NIR-VIS dataset. Empirically, when the network is trained with Softmax loss and Center loss, we set the initial learning rates, learning rate decay epochs and learning rate decay factor to 0.001, 4 and 0.96, respectively. We adopt the truncated normal initializer with standard deviation 0.1 to initialize the classification layer parameters. In particular, the center loss factor and center update rate are set to 0.01 and 0.6, respectively. Same as paper [8], we directly optimize the entire Inception-ResNet-v1 network parameters using softmax or center loss in the fine-tune stage. In Scatter loss, we randomly select classes and samples (each batch contains 20 people, and each of which has 5 samples). For a fair comparison, the triplet is also randomly selected in Triplet loss (each batch contains 90 triplets). The network is trained

TABLE 2. Rank-1 accuracy on CASIA NIR-VIS 2.0 and Oulu-CASIA NIR-VIS datasets.

Fine-tune dataset	Pre-train dataset	Network	Fusion method	Rank-1(%)	VR@FAR=1%(%)	VR@FAR=0.1%(%)
CASIA NIR-VIS 2.0	MS-Celeb-1M	Backbone	No	98.5 $\pm$ 0.3	99.2 $\pm$ 0.3	97.0 $\pm$ 0.5
	VGG2-Face	Backbone	No	95.7 $\pm$ 0.5	98.4 $\pm$ 0.4	92.3 $\pm$ 0.8
	CASIA Webface	Backbone	No	92.3 $\pm$ 0.7	96.5 $\pm$ 0.7	88.4 $\pm$ 1.1
	MS-Celeb-1M and VGG2-Face	Backbone	No	98.3 $\pm$ 0.4	99.2 $\pm$ 0.3	96.5 $\pm$ 0.5
	MS-Celeb-1M and CASIA Webface	Backbone	No	98.4 $\pm$ 0.3	99.2 $\pm$ 0.3	96.8 $\pm$ 0.5
	All three	Backbone	No	97.7 $\pm$ 0.4	99.0 $\pm$ 0.4	95.5 $\pm$ 0.7
	All three	MDN	addition	97.9 $\pm$ 0.4	99.4 $\pm$ 0.3	95.9 $\pm$ 0.7
	All three	MDN	concat	98.2 $\pm$ 0.4	99.3 $\pm$ 0.3	96.7 $\pm$ 0.5
	All three	MDN	DC	98.9 $\pm$ 0.3	99.6 $\pm$ 0.1	97.6 $\pm$ 0.4
Oulu-CASIA NIR-VIS	MS-Celeb-1M	Backbone	No	98.7	87.0	63.5
	VGG2-Face	Backbone	No	98.1	77.7	39.3
	CASIA Webface	Backbone	No	96.0	69.0	26.1
	MS-Celeb-1M and VGG2-Face	Backbone	No	98.3	83.3	50.3
	MS-Celeb-1M and CASIA Webface	Backbone	No	98.1	81.6	48.6
	All three	Backbone	No	98.4	83.4	54.5
	All three	MDN	addition	98.7	86.3	62.2
	All three	MDN	concat	99.3	81.3	38.2
	All three	MDN	DC	99.8	88.1	65.3

using four different loss, and we stop training the model until the loss value stabilizes. The training epochs of the four loss function (Softmax, Center loss, Triplet loss and Scatter loss) in CASIA NIR-VIS 2.0 are 40, 40, 22 and 18 respectively. And training epochs of them in Oulu-CASIA NIR-VIS are 32, 32, 16 and 12, respectively. Finally, we use Joint Bayesian method [14] as the classifier.

In CASIA NIR-VIS 2.0, the Softmax loss performs worst with rank-1 accuracy of 84.5% and VR@FAR = 0.1% of 76.2%. This is because Softmax loss introduces more parameters on the classification layer, and prefers balanced and sufficient training data for each identity. The center loss performs better than Softmax loss with rank-1 accuracy of 84.8% and VR@FAR = 0.1% of 77.0%, because the center loss introduces intra-class distance constraint to minimize distance within the class. Compared with Softmax loss and Center loss, the Triplet loss does not contain the the classification layer and thus has less number of parameters, which can avoid the over-fitting problem on small-scale HFR datasets. Therefore, the triplet loss performs better than Softmax loss and Center loss with rank-1 accuracy of 97.9% and VR@FAR = 0.1% of 95.6%. Above all, the Scatter loss performs best around the four objective functions with rank-1 accuracy of 98.5% and VR@FAR = 0.1% of 97.0%, because both intra-class and the inter-class constraints are based on class center, which can reduce the effects of noise samples.

In Oulu-CASIA NIR-VIS, the Triplet loss has low rank-1 accuracy of 94.7% and VR@FAR = 1% of 74.5%, respectively. It indicates that the Triplet loss can not well solve the HFR problem with insufficient training sample (less than 2,000). Softmax loss and center loss have high rank-1 accuracy of 98.0% and 98.1%, and VR@FAR = 1% of 86.2% and 86.4%, because the training set only contains 20 classes, which may introduce few parameters in classification layer. Compared with Triplet loss, Softmax loss and center loss, the SL loss has highest rank-1 accuracy of 98.7%, VR@FAR = 1% of 87.0% and VR@FAR = 0.1% of 63.5%.

The results show that SL loss can effectively train network in HFR task.

Compared with Softmax loss, center loss and Triplet loss, the SL loss achieves the best performance on both CASIA NIR-VIS 2.0 and Oulu-CASIA NIR-VIS. The results indicate that the SL can reserve the discriminative information and improve the HFR performance. Therefore, we adopt the SL loss to train three backbone networks to boost the NIR-VIS recognition performance.

2) THE EFFECTS OF OUR MULTIPLE DEEP NETWORKS STRUCTURE AND DC STRATEGY

The performance of different training dataset, network structure and fusion strategy are listed in Table 2. Here, we pre-train the Inception-ResNet-v1 network (backbone network) with different datasets including VGG2-Face, MS-Celeb-1M and CASIA Webface, and then fine-tune the network in CASIA NIR-VIS 2.0 or Oulu-CASIA NIR-VIS datasets with SL loss. When pre-training the backbone network with two VIS datasets (MS-Celeb-1M and VGG2-Face, MS-Celeb-1M and CASIA WebFace), the network is first trained with MS-Celeb-1M and then further trained with other VIS dataset. When we train the backbone network with three VIS datasets (MS-Celeb-1M, VGG2-Face and CASIA WebFace), we first pre-train it with MS-Celeb-1M, then with CASIA WebFace and finally with VGG2-Face. We adopt different fusion method to fuse the MDN network, including addition, concat and DC. In addition to MDN network with DC, all the other methods use Joint Bayesian method [14] as the classifier.

In CASIA NIR-VIS 2.0, the backbone network pre-trained with MS-Celeb-1M dataset performs best around three VIS datasets with rank-1 accuracy of 98.5% and VR@FAR = 0.1% of 97.0% when using one VIS dataset, because the dataset contains more samples with large face variation. The VGG2-Face dataset performs better than CASIA WebFace with rank-1 accuracy of 95.7% and VR@FAR = 0.1% of 92.3%, because the dataset does not contain noisy samples

and each class has more samples (around 300). Although we adopt two VIS datasets to pre-train the network, the performance of the network is worse than that trained using a single VIS dataset MS-Celeb-1M. For example, when we train the network with MS-Celeb-1M and VGG2-Face, the performance of the network dropped slightly with rank-1 accuracy of 98.3% and VR@FAR = 0.1% of 96.5%. And when we train the backbone network with MS-Celeb-1M and CASIA WebFace, the network achieves rank-1 accuracy of 98.4% and VR@FAR = 0.1% of 96.8%. In particular, when adopting three VIS datasets to train the network, the network performance is even worse with rank-1 accuracy of 97.7% and VR@FAR = 0.1% of 95.5%. The results indicate that training the backbone network with multiple VIS datasets can not boost the HFR performance. Moreover, we adopt three VIS datasets to train the MDN network with different fusion method. The MDN network using addition or concat fusion method performs even worse than a single backbone network trained with MS-Celeb-1M, and their rank-1 accuracy are 97.9% and 98.2%, respectively. While MDN network using DC method achieves the best rank-1 accuracy of 98.9% and VR@FAR = 0.1% of 97.6%. The results indicates that the fusion methods have influence on HFR performance, and simple concat or addition fusion method can not improve the HFR performance. The success of the DC strategy lies in the fact that it can take full advantage of each backbone network and it aims to learn a weight parameter adaptively for each backbone network.

In Oulu-CASIA NIR-VIS, when we pre-train the backbone network using a single VIS dataset, the network trained with MS-Celeb-1M achieves the best performance around three VIS datasets with rank-1 accuracy of 98.7% and VR@FAR = 0.1% of 63.5%. The VGG2-Face dataset performs better than CASIA WebFace with rank-1 accuracy of 98.1% and VR@FAR = 0.1% of 39.3%. The results show that different VIS datasets have different performance in HFR task, because each VIS dataset contains different number of samples and the face variations in different datasets vary a lot. In particular, when we adopt multiple VIS datasets to pre-train the backbone network, the network performs even worse than it trained using one VIS dataset MS-Celeb-1M. For example, the network trained using both MS-Celeb-1M and VGG2-Face has rank-1 accuracy of 98.3% and VR@FAR = 0.1% of 50.3%. And the network trained using MS-Celeb-1M and CASIA WebFace has even worse performance with rank-1 accuracy of 98.1% and VR@FAR = 0.1% of 48.6%. Moreover, compared with the network trained with MS-Celeb-1M, the rank-1 accuracy and VR@FAR = 0.1% of the network trained using three VIS datasets decrease by 0.3% and 9.0%, respectively. The results indicate that training a backbone network using multiple VIS datasets can not improve the HFR performance. We adopt three datasets to train the MND network with different fusion method. Compared to the backbone network trained using MS-Celeb-1M, the MDN network using addition fusion method has worse performance with VR@FAR = 0.1% of

62.2%. Although the MDN network using concat method has higher rank-1 accuracy than the backbone network trained using MS-Celeb-1M, it has worse performance in face verification with VR@FAR = 0.1% of 38.2%. In particular, our DC method achieves the best performance on both rank-1 accuracy of 99.8% and VR@FAR = 0.1% of 65.3%. It indicates that the DC method can effectively combine three networks and further improve the HFR performance.

D. COMPARISON WITH STATE-OF-THE-ART METHODS

In HFR experiments, we evaluate the MDNDC algorithm on CASIA NIR-VIS 2.0 and Oulu-CASIA NIR-VIS dataset by comparing the matching performance with other state-of-the-art methods.

1) CASIA NIR-VIS 2.0 DATASET

We compare the performance of MDNDC with some approaches including Common Encoding Feature Discriminant (CEFD) [15], Multi-View Discriminant Analysis (MvDA) [25], Kernel Coupled Spectral Regression (KCSR) [52], Kernel Discriminative Spectral Regression (KDSR) [37], H2-LBP3 [34], C-SLBFLE [23], VGG [26], SeetaFace [38], HFR-CNN [31], HFR-Net [53], TRIVET [42], IDR [28], Light CNN-9 [55], Light CNN-29 [40], Adversarial Discriminative Feature Learning (ADFL) [54] and WCNN [29]. Table 3 shows rank-1 accuracy and the verification rates of different NIR-VIS face recognition methods. And Fig. 4 further plots the ROC curves of the proposed method and some of its competitors. The two classical VIS deep methods: VGG and SeetaFace, have poor performance in NIR-VIS HFR with rank-1 accuracy of 62.1% and 68.0%, respectively, because they cannot well solve the over-fitting problem in small-scale NIR-VIS HFR dataset and fail to reduce the modality gap between the different domains.

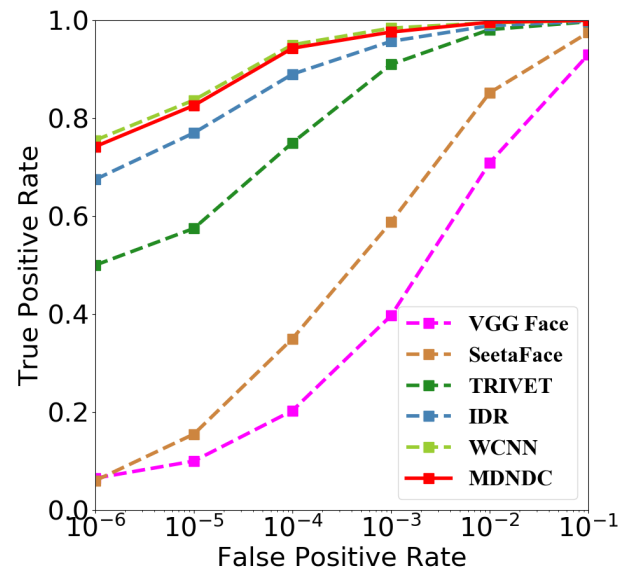


FIGURE 4. ROC curves of different methods on CASIA NIR-VIS 2.0 dataset.

TABLE 3. Rank-1 accuracy and verification rate on CASIA NIR-VIS 2.0 dataset.

Method	Rank-1(%)	$VR@FAR = 1\%$ (%)	$VR@FAR = 0.1\%$ (%)	Dim
KCSR (2009)	33.8	28.5	7.6	-
KDSR (2013)	37.5	33.0	9.3	-
H2-LBP3 (2016)	43.8	36.5	10.1	-
MvDA (2016)	41.6 $\pm$ 4.1	-	19.2	-
CEFD (2017)	85.6	-	-	-
C-SLBFLE (2018)	86.9 $\pm$ 2.2	80.3	53.0	-
VGG (2015)	62.1 $\pm$ 1.9	70.9 $\pm$ 1.3	39.7 $\pm$ 2.9	4096
SeetaFace (2016)	68.0 $\pm$ 1.7	85.2 $\pm$ 1.1	58.8 $\pm$ 2.3	2048
HFR-CNN (2016)	85.9 $\pm$ 0.9	-	78.0	-
HFR-Net (2016)	87.1 $\pm$ 0.9	-	74.5	320
TRIVET (2016)	95.7 $\pm$ 0.5	98.1 $\pm$ 0.3	91.0 $\pm$ 1.3	512
Light CNN-9(2018)	91.9 $\pm$ 0.6	-	85.3 $\pm$ 1.0	256
Light CNN-29(2018)	96.7 $\pm$ 0.2	-	94.8 $\pm$ 0.4	256
IDR (2017)	97.3 $\pm$ 0.4	98.9 $\pm$ 0.3	95.7 $\pm$ 0.7	128
ADFL (2018)	98.1 $\pm$ 0.4	99.1 $\pm$ 0.2	97.1 $\pm$ 0.5	256
WCNN (2018)	98.7 $\pm$ 0.3	99.5 $\pm$ 0.1	98.4 $\pm$ 0.4	128
MDNDC	98.9 $\pm$ 0.3	99.6 $\pm$ 0.1	97.6 $\pm$ 0.4	384

TABLE 4. Rank-1 accuracy and verification rate on Oulu-CASIA NIR-VIS dataset.

Method	Rank-1(%)	$VR@FAR = 1\%$ (%)	$VR@FAR = 0.1\%$ (%)
KCSR (2009)	66.0	49.7	26.1
MPL3 (2009)	48.9	41.9	11.4
KDSR (2013)	66.9	56.1	31.9
KPS (2013)	62.2	48.3	22.2
H2-LBP3 (2016)	70.8	62.0	33.6
TRIVET (2016)	92.2	67.9	33.6
IDR (2017)	94.3	73.4	46.2
ADFL (2018)	95.5	83.0	60.7
WCNN (2018)	98.0	81.5	54.6
MDNDC	99.8	88.1	65.3

Compared with the traditional methods CEFD and C-SLBFLE, the deep learning methods HFR-Net and Light CNN-9 have limited improvements in rank-1 accuracy of 87.1% and 91.9%. The results show that more effective deep learning methods need to be designed to solve the HFR problem. Compared with TRIVET, IDR, ADFL and WCNN, our MDNDC method achieves the best performance on rank-1 accuracy of 98.9% and $VR@FAR = 1\%$ of 99.6%. And the MDNDC obtains a smaller standard deviations than those of their competitors such as IDR, Light CNN-29 and ADFL. The superior performance of the MDNDC is due to the fact that we use SL loss to effectively train the backbone network and make full use of the publicly available dataset VGG2-Face, CAISA WebFace and MS-Celeb-1M to train the MDN. In particular, the DC fusion strategy can take full advantage of the multiple deep networks and achieves a better classification performance. In addition, the WCNN obtains better performance in $VR@FAR = 0.1\%$ of 98.4%, because it introduces a low-rank constraint and a Wasserstein distance to solve the over-fitting problems and reduce the modality gap, respectively.

2) OULU-CASIA NIR-VIS DATASET

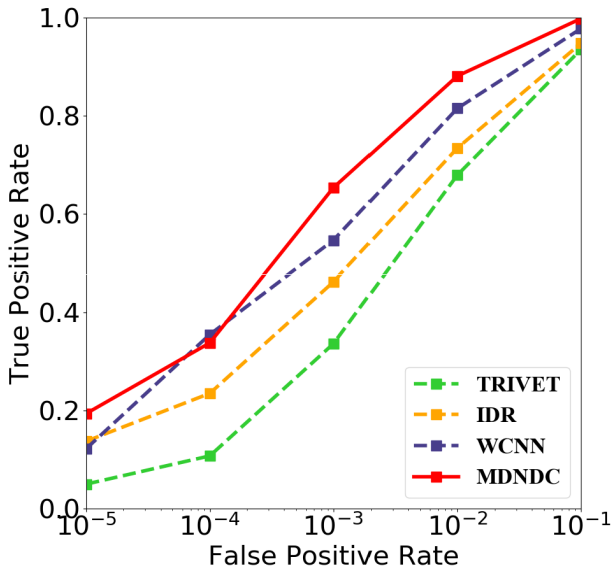
We compare the performance of MDNDC with some competitive approaches including KCSR [52], KDSR [37],

MPL3 [33], H2-LBP3 [34], Kernel Prototype Similarities (KPS) [56], IDR [28], TRIVET [42], ADFL [54] and WCNN [29]. In particular, our method does not compare with methods such as Light CNN-29, HFR-CNN, Light CNN-9, HFR-Net, C-SLBFLE and CEFD mainly because the results on Oulu-CASIA NIR-VIS dataset are not given in the published papers. Table 4 shows rank-1 accuracy and the verification rates of these HFR methods. And Fig. 5 further plots the ROC curves of the MDNDC method and some of its competitors. The following observations are recorded:

Compared with the traditional methods KCSR, KDSR, KPS, MPL3 and H2-LBP3, the deep learning methods (IDR, TRIVET, ADFL, WCNN and MDNDC) obtain better performance in terms of both rank-1 accuracy and $VR@FAR = 0.1\%$. The rank-1 accuracy of TRIVET and WCNN methods are 92.2% and 98.0%, respectively, which is higher than MPL3 and H2-LBP3 of 48.9% and 70.8%, respectively. It shows that a deep learning method can easily obtain a better performance than the traditional methods. The results indicate that the feature learned by TRIVET, ADFL, IDR, WCNN and MDNDC are more discriminative than the traditional methods. Our MDNDC method achieves the highest performance with rank-1 accuracy of 99.8% and $VR@FAR = 0.1\%$ of 65.3%. In particular, compared with WCNN, the MDNDC gains the improvement up to 1.8%

TABLE 5. The input image size, computational costs, time cost of a single image to extract feature and the model parameters of our approach compared with some popular models. The speed is tested on TITAN Xp.

Method	Input image size	Million Parameters	Billion Mult-Adds	Time cost
TRIVET (2016)	128×128	5.5	1.1	1.0ms
IDR (2017)	128×128	7.7	1.1	1.0ms
WCNN (2018)	128×128	7.7	1.1	1.0ms
Light CNN-9(2018)	128×128	5.5	1.1	1.0ms
Light CNN-29(2018)	128×128	12.6	3.8	1.3ms
VGG (2015)	224×224	138.0	15.3	-
Backbone	160×160	22.8	1.5	1.5ms
MDNDC	160×160	68.4	4.5	4.5ms

**FIGURE 5.** ROC curves of different methods on Oulu-CASIA NIR-VIS dataset.

and 10.7% in rank-1 accuracy and $VR@FAR = 0.1\%$, respectively. As we know, the Oulu-CASIA NIR-VIS dataset contains less training samples (about 2,000) than CASIA NIR-VIS 2.0 (about 8,600). This finding indicates that the MDNDC is more suitable for small-scale NIR-VIS HFR dataset than WCNN method, because the WCNN introduces two softmax layers and cannot well solve the over-fitting problem on Oulu-CASIA NIR-VIS dataset. The results also indicate that the SL loss and the DC fusion strategy can work well in small-scale NIR-VIS HFR dataset. Under the supervision of SL loss, the deep network can remove highly non-linear modality information and reserve the discriminative information, and it exploits the correlations from both inter- and intra-class data. In addition, the DC method can effectively fuse three backbone networks and boost the HFR performance. Therefore, our MDNDC method is suitable for HFR tasks.

E. COMPUTATIONAL EFFICIENCY ANALYSIS

Computational efficiency is an important issue for CNN models. To verify the computational efficiency of MDNDC model, we compare our CNN with six public CNNs,

i.e., VGG [26], TRIVET [42], IDR [28], WCNN [29], Light CNN-9 [55] and Light CNN-29 [40]. Table 5 shows the input image size, computational costs, time cost of a single image to extract feature and the model parameters of different approaches. The speed is tested on TITAN Xp. The Light CNN-9 model, TRIVET, IDR and WCNN have high computational efficiency. For example, they have only 1.1 billion Mult-Adds and less than 8 million parameters. The MDNDC contains three backbone networks (Inception-resnet-v1), so its model parameters are 3 times larger and the computational costs are 3 times higher than the backbone network. In particular, the MDNDC is 2 times smaller than that of the well-known VGG model, while the computational costs is about 3.5 times faster. However, the computational efficiency of MDNDC is lower than Light CNN-29 model, Light CNN-9 model, TRIVET, IDR and WCNN. For example, the billion Mult-Adds is about 4 times higher than TRIVET, IDR and WCNN. Although the model parameters of MDNDC are larger and its computational costs are higher than these methods (Light CNN-9 model, TRIVET, IDR and WCNN), the time cost of a single image to extract feature is 4.5 ms. The results indicate that MDNDC is potentially suitable and practical for real-time applications.

F. DISCUSSIONS

We have performed a large number of experiments on HFR task to evaluate our proposed algorithm. We can get the following observations.

- Scatter loss performs best around the four objective functions (including Softmax, Center loss, Triplet loss and Scatter loss) with rank-1 accuracy of 98.5% and $VR@FAR = 0.1\%$ of 97.0% in CASIA NIR-VIS 2.0, which shows that the Scatter Loss can learn highly discriminative features for robust heterogeneous face recognition.
- Multiple Deep Networks with DC performs better than a single backbone network, which indicates that the individual features extracted by a single deep network are less discriminative and robust compared to the DC fusion of multiple parallel deep networks.
- The backbone network trained with different datasets VGG2-Face, CAISA WebFace and MS-Celeb-1M results in differential performance, indicating that the data has a great influence on the performance of HFR

models. Instead of using only one publicly available dataset, we make full use of all three datasets to train the MDNDC. And our method performs better than a single backbone networks with rank-1 accuracy of 98.9% and $VR@FAR = 0.1\%$ of 97.6% in CASIA NIR-VIS 2.0, which shows that HFR can be further improved through multiple datasets.

- DC strategy performs better than simple concat and addition methods, and it improves rank-1 accuracy up to 0.7% and 1.0% in CASIA NIR-VIS 2.0, which shows that our DC method can effectively fuse three networks.
- The advantages of the MDNDC method is that it is a robust and effective algorithm for HFR problem. For example, it has superior performance on both CASIA NIR-VIS 2.0 and Oulu-CASIA NIR-VIS datasets with rank-1 accuracy of 98.9% and 99.8%, respectively. The performance of MDNDC is attributed to the effective SL loss, multiple deep networks structure and DC fusion strategy. In particular, if we design three backbone networks in a parallel structure, the time cost of a single image to extract feature is only 4.5 ms. The results indicate that MDNDC is potentially suitable and practical for real-time HFR applications. Therefore, the MDNDC approach has a potential application in the field of criminal investigation and multimedia information retrieval.
- One weakness of MDNDC approach is that it contains large model parameters and high computational costs. In the future, we attempt to search for a high computational efficiency backbone network to build the MDNDC, and make our MDNDC model smaller, faster and more accurate.

V. CONCLUSION

In this paper, we have developed a novel Multiple Deep Networks with scatter loss and Diversity Combination method for dealing with HFR problem. The Scatter Loss can maximize distance between the classes and minimize distance within the class to learn highly discriminative features for robust heterogeneous face recognition, and it is robust to noise samples and stable during training process. In addition, we design a Multiple Deep Networks structure for feature extraction, and propose a joint decision strategy called Diversity Combination to adaptively adjust weights of each deep network and make a joint classification decision. By plenty of experiments, we find that HFR can be further improved through multiple training datasets. Experiments on two heterogeneous face benchmark datasets CASIA NIR-VIS 2.0 and Oulu-CASIA NIR-VIS demonstrate the superiority of the MDNDC approach.

REFERENCES

- [1] A. Sharma and D. W. Jacobs, "Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 593–600.
- [2] A. T. Alex, V. K. Asari, and A. Mathew, "Local difference of Gaussian binary pattern: Robust features for face sketch recognition," in *Proc. IEEE Int. Conf. Syst. Man Cybern.*, Oct. 2013, pp. 1211–1216.
- [3] B. F. Klare and A. K. Jain, "Heterogeneous face recognition using Kernel prototype similarities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1410–1422, Jun. 2013.
- [4] B. F. Klare, Z. Li, and A. K. Jain, "Matching forensic sketches to mug shot photos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 639–646, Mar. 2011.
- [5] C. Peng, X. Gao, N. Wang, and J. Li, "Graphical representation for heterogeneous face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 301–312, Feb. 2017.
- [6] J. Zhang, S. Ding, and N. Zhang, "An overview on probability undirected graphs and their applications in image processing," *Neurocomputing*, vol. 321, pp. 156–168, Dec. 2018.
- [7] N. Zhang, S. Ding, J. Zhang, and X. Zhao, "Robust spike-and-slab deep Boltzmann machines for face denoising," *Neural Comput. Appl.*, pp. 1–13, 2018. doi: 10.1007/s00521-018-3866-6.
- [8] W. Hu and H. Hu, "Discriminant deep feature learning based on joint supervision loss and multi-layer feature fusion for heterogeneous face recognition," *Comput. Vis. Image Understand.*, vol. 184, pp. 9–21, Jul. 2019.
- [9] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.
- [10] M. Kan, S. Shan, and X. Chen, "Multi-view deep network for cross-view classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4847–4855.
- [11] C. Reale, H. Lee, and H. Kwon, "Deep heterogeneous face recognition networks based on cross-modal distillation and an equitable distance metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 226–232.
- [12] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," 2016, *arXiv:1602.07261*. [Online]. Available: <https://arxiv.org/abs/1602.07261>
- [13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
- [14] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun, "Bayesian face revisited: A joint formulation," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 566–579.
- [15] D. Gong, Z. Li, W. Huang, X. Li, and D. Tao, "Heterogeneous face recognition: A common encoding feature discriminant approach," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2079–2089, May 2017.
- [16] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014, *arXiv:1411.7923*. [Online]. Available: <https://arxiv.org/abs/1411.7923>
- [17] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.
- [18] J. Lezama, Q. Qiu, and G. Sapiro, "Not afraid of the dark: NIR-VIS face recognition via cross-spectral hallucination and low-rank embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6807–6816.
- [19] N. Zhang, S. Ding, J. Zhang, and Y. Xue, "Research on point-wise gated deep networks," *Appl. Soft Comput.*, vol. 52, pp. 1210–1221, Mar. 2017.
- [20] J. Zhang, S. Ding, N. Zhang, and Y. Xue, "Weight uncertainty in Boltzmann machine," *Cogn. Comput.*, vol. 8, no. 6, pp. 1064–1073, 2016.
- [21] J. Li, P. Hao, C. Zhang, and M. Dou, "Hallucinating faces from thermal infrared images," in *Proc. 15th IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 465–468.
- [22] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," 2018, *arXiv:1801.07698*. [Online]. Available: <https://arxiv.org/abs/1801.07698>
- [23] J. Lu, V. E. Liong, and J. Zhou, "Simultaneous local binary feature learning and encoding for homogeneous and heterogeneous face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 8, pp. 1979–1993, Aug. 2018.
- [24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: <https://arxiv.org/abs/1409.1556>
- [25] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 808–821.
- [26] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 1–12.

- [27] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," 2017, *arXiv:1710.08092*. [Online]. Available: <https://arxiv.org/abs/1710.08092>
- [28] R. He, X. Wu, Z. Sun, and T. Tan, "Learning invariant deep representation for NIR-VIS face recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 7–18.
- [29] R. He, X. Wu, Z. Sun, and T. Tan, "Wasserstein cnn: Learning invariant features for NIR-VIS face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.
- [30] S. Wang, D. Huang, Y. Wang, and Y. Tang, "2D-3D heterogeneous face recognition based on deep canonical correlation analysis," in *Proc. Chin. Conf. Biometric Recognit.*, 2017, pp. 77–85.
- [31] S. Saxena and J. Verbeek, "Heterogeneous face recognition with CNNs," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 483–491.
- [32] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The casia NIR-VIS 2.0 face database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2013, pp. 348–353.
- [33] S. Chen, D. Yi, J. Yang, G. Zhao, S. Z. Li, and M. Pietikainen, "Learning mappings for face synthesis from near infrared to visual light images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 156–163.
- [34] M. Shao and Y. Fu, "Cross-modality feature learning through generic hierarchical hyperlingual-words," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 2, pp. 451–463, Feb. 2016.
- [35] W. Zhang, Z. Shu, D. Samaras, and L. Chen, "Improving heterogeneous face recognition with conditional adversarial networks," 2017, *arXiv:1709.02848*. [Online]. Available: <https://arxiv.org/abs/1709.02848>
- [36] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 513–520.
- [37] X. Huang, Z. Lei, M. Fan, X. Wang, and S. Z. Li, "Regularized discriminative spectral regression method for heterogeneous face matching," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 353–362, Jan. 2013.
- [38] X. Liu, M. Kan, W. Wu, S. Shan, and X. Chen, "VIPLFaceNet: An open source deep face recognition SDK," *Frontiers Comput. Sci.*, vol. 11, no. 2, pp. 208–218, 2017.
- [39] X. Wu, L. Song, R. He, and T. Tan, "Coupled deep learning for heterogeneous face recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 1679–1686.
- [40] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.
- [41] X. Tang and X. Wang, "Face sketch synthesis and recognition," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Oct. 2003, pp. 687–694.
- [42] X. Liu, L. Song, X. Wu, and T. Tan, "Transferring deep representation for NIR-VIS heterogeneous face recognition," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2016, pp. 1–8.
- [43] Y. Song, L. Bao, Q. Yang, and M.-H. Yang, "Real-time exemplar-based face sketch synthesis," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 800–813.
- [44] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 87–102.
- [45] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 499–515.
- [46] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1701–1708.
- [47] Y. Tian, C. Yan, X. Bai, and J. Zhou, "Heterogeneous face recognition via Grassmannian based nearest subspace search," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 1077–1081.
- [48] Z. Lei, M. Pietikainen, and S. Z. Li, "Learning discriminant face descriptor," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 289–302, Feb. 2014.
- [49] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
- [50] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016, *arXiv:1603.04467*. [Online]. Available: <https://arxiv.org/abs/1603.04467>
- [51] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, no. 7, pp. 257–269, 2010.
- [52] Z. Lei and S. Z. Li, "Coupled spectral regression for matching heterogeneous faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1123–1128.
- [53] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPR)*, Jun. 2016, pp. 54–62.
- [54] L. Song, M. Zhang, X. Wu, and R. He, "Adversarial discriminative heterogeneous face recognition," in *Proc. 32nd Conf. Artif. Intell.*, 2018, pp. 7355–7362.
- [55] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," 2015, *arXiv:1511.02683*. [Online]. Available: <https://arxiv.org/abs/1511.02683>
- [56] B. F. Klare and A. K. Jain, "Heterogeneous face recognition using kernel prototype similarities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1410–1422, Jun. 2013.



WEIPENG HU received the M.E. degree in electronics and information technology from Sun Yat-sen University, Guangzhou, China, in 2018, where he is currently pursuing the Ph.D. degree. His current research interests include computer vision, image processing, and face recognition.



HAIFENG HU received the Ph.D. degree from Sun Yat-sen University, in 2004, where he is currently a Professor with the School of Electronics and Information Technology. He has published about 120 papers, since 2000. His research interests include computer vision, pattern recognition, image processing, and neural computation.



XINLONG LU is currently pursuing the B.E. degree in electronics and information technology with Sun Yat-sen University, Guangzhou, China. His current research interests include computer vision and face recognition.

...