Orthogonal Modality Disentanglement and Representation Alignment Network for NIR-VIS Face Recognition

Weipeng Hu and Haifeng Hu, *Member, IEEE*

***Abstract*— Near-infrared and visual (NIR-VIS) face matching, as the most typical task in Heterogeneous Face Recognition (HFR), has attracted increasing attention in recent years. However, due to the large within-class discrepancies, including domain differences and residual discrepancies (i.e., lighting, expressions, occlusion, blurry, pose, etc), this is still a difficult task. Conventional NIR-VIS FR methods only focus on reducing the modality gap between cross-domain images, while neglecting to eliminate the residual variations. To better solve the above problems, this paper proposes a novel Orthogonal Modality Disentanglement and Representation Alignment (OMDRA) approach, which consists of three key components, including Modality-Invariant (MI) loss, Orthogonal Modality Disentanglement (OMD) and Deep Representation Alignment (DRA). Firstly, the MI loss is designed to learn modality-invariant and identity-discriminative representation, by increasing between-class separability and within-class compactness between NIR and VIS heterogeneous data. Secondly, the high-level Hybrid Facial Feature (HFF) layer of the backbone network is projected into two subspaces: the modality-related and identity-related subspaces. The OMD is designed to decouple modal information via an adversarial process, and we further impose Orthogonal Representation Decorrelation (ORD) to the OMD to decrease the correlation between identity representations and domain representations, as well as enhancing their representation capabilities. Finally, the DRA aims to eliminate the residual variations by performing a high-level representation alignment between non-neutral face and neutral face, which can effectively guides the network to learn discriminative and residual-invariant face representation. The joint scheme enables the disentanglement of modality variations, elimination of residual discrepancies, and the purification of identity information. Extensive experiments on challenging cross-domain databases indicate that our OMDRA method is superior to the state-of-the-art methods.**

***Index Terms*— NIR-VIS face recognition, cross-modality face matching, orthogonal modality disentanglement, deep representation alignment.**

I. Introduction

HETEROGENEOUS Face Recognition (HFR), as an important branch in Face Recognition (FR), refers to match facial images from different domains such as near-infrared (NIR) to visual (VIS) and sketch to photo [1]–[5]. Traditional face sensors can capture high-quality visual face images under normal illumination conditions, but they cannot work properly in dark nights or indoor low-illumination environments. Near-infrared sensors can be used to cope with these low-illumination environments by capturing high-quality NIR images [6]–[9]. Nowadays more and more surveillance cameras can capture NIR images under low illumination conditions, and NIR-VIS face matching has broad application prospects in the field of multimedia data retrieval and criminal tracking. In NIR-VIS FR task, VIS images and NIR images are collected from visible and infrared imaging device, resulting in large modality gap between the two different domain images. In addition, heterogeneous face images are rare on the Internet, making it hard to annotate large-scale cross-domain databases.

To solve the heterogeneous face recognition problem, many works mainly focus on invariant descriptor extraction [10], [11] and projection subspace modeling [12], [13]. Recently, the Convolutional Neural Networks (CNNs) [14]–[17] has achieved a landmark success in object detection, object classification, action recognition, etc., and has become a general method of computer vision community. Inspired by the success of CNN technology, more and more CNN-based methods have been developed to address the NIR-VIS FR issue. To learn identity-related features, [18] and [19] pre-train and then fine-tune a deep network structure using large-scale VIS database and HFR database, respectively. *Peng et al. [20]* present a locally linear re-ranking method to explore valuable information from the initial ranking result for HFR task. *Pereira et al. [21]* introduce a modality specific structure to extract low-level features for cross-modal images, and reduce the domain gaps by sharing high-layer network structure. *Wu et al. [22]* introduce a trace norm as a relevance constraint to enforce the relevance between different domains, and adopt a cross-modal ranking term to preserve the within-personal similarity in different modalities. To decouple domain-related information, [7] maps the high-level feature layer to two orthogonal subspaces, including modality-unrelated representation and modality-related representation. *He et al. [6]* take advantages of methods in [22] and [7], and further introduce a Wasserstein distance to enhance the similarity between NIR distribution and VIS distribution. Works in [23] and [24] aim to explore domain specific structure (i.e., mutual component analysis module and residual compensation module) to reduce the modality gap.

Manuscript received May 10, 2021; revised August 2, 2021; accepted August 13, 2021. Date of publication August 16, 2021; date of current version June 6, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62076262, Grant 61673402, and Grant 61273270, and in part by the Science and Technology Program of Huizhou of China under Grant 2020SC0702002. This article was recommended by Associate Editor G.-J. Qi. *(Corresponding author: Haifeng Hu.)*

The authors are with the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, China (e-mail: huwp5@mail2.sysu.edu.cn; huhaif@mail.sysu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2021.3105411.

Digital Object Identifier 10.1109/TCSVT.2021.3105411

1051-8215 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Korea Institute of Science and Technology. Downloaded on January 27,2023 at 07:15:41 UTC from IEEE Xplore. Restrictions apply.

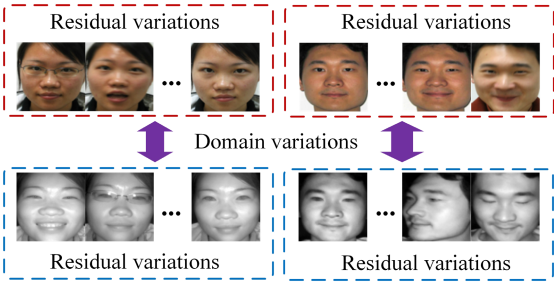


Fig. 1. Large intra-class discrepancies between NIR image and VIS image can be reflected in both residual variations (i.e., pose, expressions, lighting, occlusion and blurry) and domain variations.

However, some existing HFR approaches [18]–[20] are only intended to extract identity-related features, while ignoring the decoupling of modality-related representations. Although some recent HFR methods [6], [7], [21], [23], [25] take both identity and spectrum information into consideration, how to effectively eliminate modal differences is still a challenging problem. Moreover, as shown in Fig. 1, in addition to modality information, the residual variations (i.e., lighting, expressions, occlusion, blurry, pose, etc) can also cause large within-class variations [9], [26]. Notably, both the modality discrepancies and residual variations should be considered in network modeling, but most methods only focus on domain-specific structure modeling while disregarding residual-invariant representation modeling. Consequently, the residual-related information will be introduced to feature representation, which may reduce the HFR performance.

To deal with the aforementioned issues, this paper proposes a novel approach called Orthogonal Modality Disentanglement and Representation Alignment (OMDRA) network for NIR-VIS matching task. The OMDRA employs an end-to-end network structure to extract modality-invariant, residual-independent and identity-discriminative feature representations. The flowchart of OMDRA model is illustrated in Fig. 2, and it mainly contains three key components, including Modality-Invariant (MI) loss, Orthogonal Modality Disentanglement (OMD) and Deep Representation Alignment (DRA). Firstly, considering that traditional objective loss (such as triplet loss, softmax loss) may lack domain adaptation for different modalities, we design MI loss to increase the separability between inter-class cross-domain images and compactness between intra-class cross-modal images. Secondly, the OMD is introduced to reduce modality gaps between cross-domain images and enhance representation capabilities. Specifically, the high-level Hybrid Facial Feature (HFF) layer of the network is divided into two subspaces that contain a modality-dependent subspace and an identity-dependent subspace, respectively. The OMD aims to eliminate modality differences via a min-max adversarial process, and we further introduce an Orthogonal Representation Decorrelation (ORD) to the OMD structure to prevent redundancy of feature representation and reduce the correlation between the identity representation and modality representation. Finally, worthy of mention is that residual variations may increase the within-class distance. With the purpose of decouple residual variations, the DRA performs alignment between non-neutral face (i.e., pose, expressions, lighting, occlusion and blurry) and neutral face, and guide the network to learn residual-independent face representation. In short, the major contributions of this paper are listed below:

- The OMDRA model naturally combines MI, OMD and DRA into an end-to-end framework, and the joint scheme enables disentanglement of modality discrepancies, elimination of residual differences, and the purification of identity information.
- The MI loss is introduced to learn identity-discriminative representation, by increasing inter-class cross-domain distension and reducing intra-class cross-modal separability.
- We divide the high-level HFF layer into two parts, including modality-dependent layer and identity-dependent layer, and introduce OMD structure comprising domain adversarial disentanglement and orthogonal representation decorrelation to effectively disentangle modality discrepancy and enhance representation capabilities, respectively.
- The DRA is presented to learn residual-invariant face representation, which enforces a high-level feature alignment on non-neutral face to learn a discriminative face representation.
- Experimental results on multiple challenging databases, such as CASIA NIR-VIS 2.0 [27], Oulu-CASIA NIR-VIS [28] and BUAA NIR-VIS [29], show that the OMDRA approach is superior to the state-of-the-art methods.

The rest of the paper is organized as follows. Section 2 provides a brief reviews of prior works on HFR task. Section 3 describes the details of our OMDRA approach. The experimental results are reported in Section 4. Finally, we summarize conclusions in Section 5.

II. Related Work

Over the past decade, researchers have made great efforts to solve NIR-VIS HFR problem. In general, current HFR approaches can be classified into three groups, including domain-invariant representation, synthesis-based method, and subspace-based method.

Domain-invariant representation methods seek discriminative hand-crafted features that only contain identity-related information and disregard domain information [10], [11], [30]–[33]. In [30], the LBP is used to learn facial hand-crafted features, which involves a two-step process, including feature encoding and feature mapping. Considering that some important information may be disregarded during the binarization process in [30], to solve this problem, *Lu et al. [10]* present an unsupervised representation learning method by exploiting both specific and common information from cross-domain images, which integrates feature learning and encoding into one-stage processing. *Zhang et al. [31]* learn descriptor representations through coupled information-theoretic encoding, and explore the mutual information between cross-domain images

Authorized licensed use limited to: Korea Institute of Science and Technology. Downloaded on January 27,2023 at 07:15:41 UTC from IEEE Xplore. Restrictions apply.

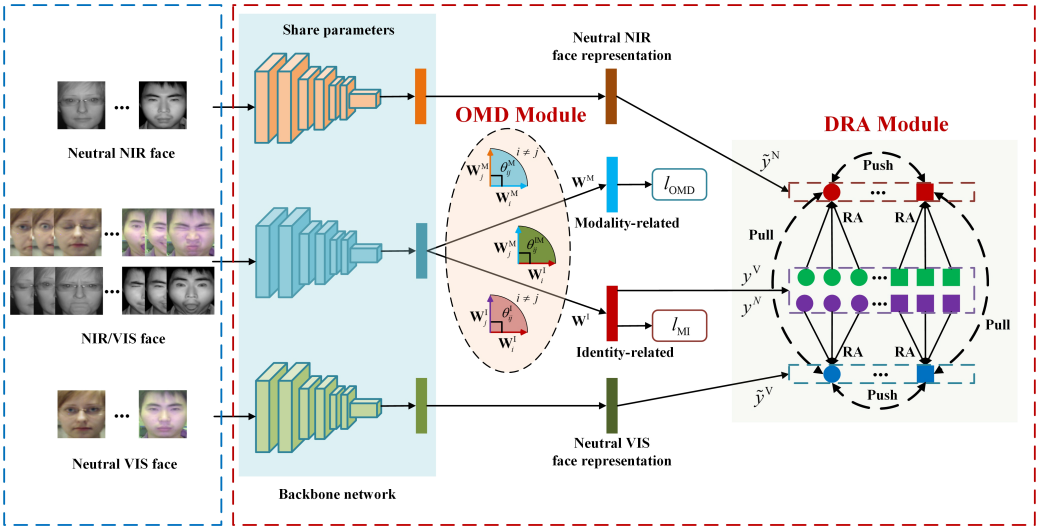


Fig. 2. The flowchart of OMDRA model. The OMDRA network contains three key components, including Modality-Invariant (MI) loss, Orthogonal Modality Disentanglement (OMD) and Deep Representation Alignment (DRA). The joint formulation leads to the disentanglement of modality information, elimination of residual variations, and the purification of identity information.

with a coupled information-theoretic projection forest. *Klare et al. [32]* represent heterogeneous data using SIFT and multiscale LBP hand-crafted features, and adopt a linear discriminative classifier for distance matching. Since conventional descriptors such as LBP, HOG and SIFT fail to extract the common distinguishing information from cross-modality data, *Gong et al. [33]* present a general encoding descriptor to capture common discriminant information and learn robust feature representations. The domain-invariant approaches rely on strong human prior knowledge to devise feature descriptors, and thus they may fail to fully exploit potentially useful information and make them have poor adaptive learning capabilities.

Synthesis-based methods try to synthesize a cross-spectral face hallucination, and then use traditional recognition approaches to deal with the same domain face images [34]–[41]. *Zhang et al. [34]* use a multiscale MRF method to convert sketch images from visual images, which contains three steps, including shape priors, patch matching, and compatibility terms. However, multiple representations used for image synthesis may lack of adaptive ability in [34]. To address these problems, *Peng et al. [35]* adopt Markov network to learn multiple representations and combine these representations for cross-domain image synthesis. *Song et al. [36]* use an adversarial discriminative feature learning framework to bridge the modality gaps between cross-spectrum images. In [36], the cycle-consistency loss cannot retain sufficient identity and geometric information in the synthesized image. To deal with this problem, Pose-preserving Cross-spectral Face Hallucination (PCFH) method is developed in [37], which preserves fine-grained details by introducing both global and local paths. Considering that the synthesized image in [37] lacks personal textures and violates variations in reality, Pose Agnostic Cross-spectral Hallucination (PACH) approach is presented in [38], which consists of two subnetworks including texture prior synthetic network and facial alignment network. *He et al. [39]* present an adversarial Crosss-pectral Face Completion (CFC) to generate high-resolution face images, which contains two core units: pose-related component and texture-related component.

Subspace-based methods attempt to project heterogeneous-domain images into a compact feature subspace, in which similarity between different domain images can be measured [6], [12], [13], [20], [22]–[24], [42], [43]. Discriminative spectral regression approach is developed in [12] to map heterogeneous-domain data into a common subspace. In [13], a Multi-view Discriminant Analysis (MvDA) approach is developed to map multiple views into a unified common space, in which the between-class discrepancies are maximized and within-class discrepancies are minimized. *Shu* and *Tang et al. seek domain-dependent features and shared cross-domain features to describe heterogeneous data [44], [45]. In [18] and [19], a deep network structure is used to extract identity-related representations for NIR and VIS images. Peng et al. [20]* fully exploit the useful initial ranking information through a Locally Linear Re-Ranking (LLRe-Rank) trick. In order to eliminate the modality discrepancies, trace norm is introduced to Coupled Deep Learning (CDL) network to enhance the relevance between heterogeneous face images [22], and orthogonal constraint is introduced to Invariant Deep Representation (IDR) model to make domain features and identity features irrelevant to each other [7]. Inspired by the methods in [22] and [7], *He et al. [6]* further use a Wasserstein distance to enhance similarity between NIR-domain distribution and VIS-domain distribution. In [23], a Mutual Component CNNs (MC-CNN) is presented to eliminate modality discrepancy. *Deng et al. [24]* present a two-paths Residual Compensation Networks (RCN) for HFR, and introduce a modality discrepancy loss and residual compensation block to the network architecture.

However, synthesis-based method [1], [2], [26], [36], [40], [41] is difficult to generate a photo-realistic synthesized image,

Authorized licensed use limited to: Korea Institute of Science and Technology. Downloaded on January 27,2023 at 07:15:41 UTC from IEEE Xplore. Restrictions apply.

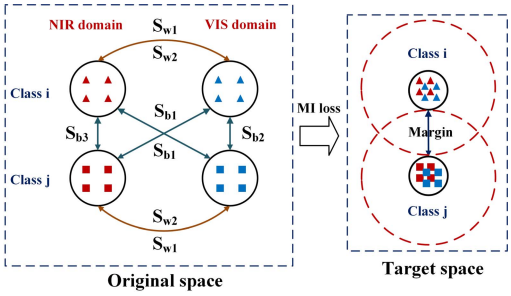


Fig. 3. Illustration of the MI objective function.Triangle and square represent the ith and the jth class, respectively. Red triangles and squares represent NIR images, and blue triangles and squares represent VIS images. The MI objective function imposes three inter-class forces (e.g. S_{b1}, S_{b2} and S_{b3}) and two intra-class forces (e.g. S_{w1} and S_{w2}) on the cross-modal representations, which learn domain-independent and identity-discriminative representations.

and the recognition system will increase extra computational cost during face synthesis procedure. Although many existing subspace-based methods attempt to bridge the domain variations, how to effectively decouple modal variations is still a challenging problem. In addition, large within-class variations between heterogeneous images can also be reflected in residual variations (i.e., lighting, expressions, occlusion, blurry, pose, etc), and most current methods disregard the elimination of residual-related variations. To solve the aforementioned problems, our OMDRA takes both modality variations and residual variations into consideration, which naturally combines MI, OMD and DRA into a unified framework to learn modal- and residual-independent representations.

III. The Proposed OMDRA Model

With the purpose of decoupling modality variations and reducing residual discrepancies, we present the OMDRA model which consists of three key components, including MI, OMD and DRA. Fig. 2 shows the architecture of the OMDRA model. In this section, we will details the three components of the OMDRA network.

A. Problem Formulation

Based on label information in NIR-VIS HFR dataset, the identity dataset and modality dataset can be represented as $D^I = \{(x_1, l_1^I), (x_2, l_2^I), \ldots, (x_q, l_q^I)\}$ and $D^M = \{(x_1, l_1^M), (x_2, l_2^M), \ldots, (x_p, l_p^M)\}$, respectively, where x_i denotes the heterogeneous image, and $l_i^I \in \{1, 2, \ldots, c\}$ and $l_i^M \in \{VIS, NIR\}$ represent identity-related and modality-related labels, respectively. In OMDRA network, Θ represents the parameters in the backbone network, and $f_I(\cdot)$ and $f_M(\cdot)$ denote the learning features in identity-dependent layer and modality-dependent layer, respectively.

B. The Proposed MI Loss

Traditional objective loss (e.g. Triplet loss [46], Contrastive loss [47] and Softmax loss) does not impose any domain-related constraint between heterogeneous face images, resulting in insufficient domain adaptability and difficulty in bridging the spectrum variations. To address this problem, the modality-invariant loss is used to constrain the deep network, which contains intra-class cross-domain constraint and inter-class cross-domain constraint, taking the following form:

$$l_I = \kappa_1 S_{w1} + \kappa_2 S_{w2} + \kappa_3 (S_{b1} + S_{b2} + S_{b3}) \tag{1}$$

where

$$S_{w1} = \frac{1}{c} \sum_{i=1}^{c} \|m_i^V - m_i^N\|_2^2$$

$$S_{w2} = \frac{1}{c} \sum_{i=1}^{c} \|\varsigma_i^V - \varsigma_i^N\|_2^2$$

$$S_{b1} = \frac{1}{c} \sum_{i=1}^{c} \sum_{j=i+1}^{c} [\alpha - \|m_i^V - m_j^N\|_2^2]_+$$

$$S_{b2} = \frac{1}{c} \sum_{i=1}^{c} \sum_{j=i+1}^{c} [\alpha - \|m_i^V - m_j^V\|_2^2]_+$$

$$S_{b3} = \frac{1}{c} \sum_{i=1}^{c} \sum_{j=i+1}^{c} [\alpha - \|m_i^N - m_j^N\|_2^2]_+$$

where S_{w1} and S_{w2} constitute the intra-class cross-domain constraint, and S_{b1}, S_{b2} and S_{b3} denote the inter-class cross-domain constraint. S_w and S_b are calculated in mini-batch manner. $\mathbf{I}(\cdot)$ represents the indicator function, and $[a]_+$ is equivalent to $max(a, 0)$ operation. $m_r^V = \frac{1}{n_r^V} \sum_{k=1}^{n} \mathbf{I}(l_k^I = r)\mathbf{I}(l_k^M = VIS) f_I(x_k)$ denotes the mean vector of the rth class of VIS data, and $m_r^N = \frac{1}{n_r^N} \sum_{k=1}^{n} \mathbf{I}(l_k^I = r)\mathbf{I}(l_k^M = NIR) f_I(x_k)$ indicates the mean vector of the rth class of NIR data. $\varsigma_r^V = \frac{1}{n_r^V} \sum_{k=1}^{n} \mathbf{I}(l_k^I = r)\mathbf{I}(l_k^M = VIS)(f_I(x_k) - m_r^V)^2$ and $\varsigma_r^N = \frac{1}{n_r^N} \sum_{k=1}^{n} \mathbf{I}(l_k^I = r)\mathbf{I}(l_k^M = NIR)(f_I(x_k) - m_r^N)^2$ represent the variance of the rth class of VIS images and NIR images [6], [26]. c denotes the number of classes, and n_r^V and n_r^N indicate the number of samples in the rth class of VIS images and NIR images, repsectively. α is the margin value, and κ is the weighting coefficient of each part.

The MI loss aims to learn domain-independent and identity-discriminative representations. As shown in Fig. 3, the MI objective function imposes three inter-class forces (e.g. S_{b1}, S_{b2} and S_{b3}) and two intra-class forces (e.g. S_{w1} and S_{w2}) on the cross-modal representations, which can effectively narrow the distribution difference of intra-class cross-view data and increase the separability of inter-class cross-view data.

The gradient descent is employed to optimize Eq. (1), and the gradient l_I w.r.t. $f_I(x_i)$ can be expressed as:

$$\begin{aligned}\frac{\partial l_I}{\partial f_I(x_i)} = & \frac{\partial l_I}{\partial S_{w1}} \frac{\partial S_{w1}}{\partial f_I(x_i)} + \frac{\partial l_I}{\partial S_{w2}} \frac{\partial S_{w2}}{\partial f_I(x_i)} \\ & + \frac{\partial l_I}{\partial S_{b1}} \frac{\partial S_{b1}}{\partial f_I(x_i)} + \frac{\partial l_I}{\partial S_{b2}} \frac{\partial S_{b2}}{\partial f_I(x_i)} + \frac{\partial l_I}{\partial S_{b3}} \frac{\partial S_{b3}}{\partial f_I(x_i)}\end{aligned} \tag{2}$$

The MI loss is adopted to optimize our deep model in mini-batch manner, which can effectively bridge the domain gaps as well as making learning feature discriminative.

Authorized licensed use limited to: Korea Institute of Science and Technology. Downloaded on January 27,2023 at 07:15:41 UTC from IEEE Xplore. Restrictions apply.

C. Orthogonal Modality Disentanglement

Due to the recent success of CNN technology in various visual fields [14], [16], [17], such as object classification and object detection, many HFR approaches directly use an end-to-end network to learn identity-related representations for cross-view images [18]–[20]. Worthy of mention is that domain difference of heterogeneous data may lead to large within-class variations. However, these methods [18], [19] neglect to eliminate domain-dependent information, resulting in poor HFR performance. Recently, many works [6], [7], [21], [23], [25] aim to design a domain specific module to decouple modality-related information. Although these methods can decouple modal gap of cross-domain images to a certain extent, how to effectively bridge the domain variations is still a challenging problem. To address this problem, we present the orthogonal modality disentanglement to effectively disentangle spectrum variations. Specifically, the high-level Hybrid Facial Feature (HFF) layer of the network is divided into two parts: the identity-related layer and modality-related layer, taking the following form:

$$\begin{aligned} y_k^I &= f_I(x_k) = W^{I^T}\gamma \quad (k = 1, 2, \ldots) \\ y_k^M &= f_M(x_k) = W^{M^T}\gamma \quad (k = 1, 2, \ldots) \end{aligned} \tag{3}$$

where $W^I \in \mathbb{R}^{d_0 \times d_1}$ and $W^M \in \mathbb{R}^{d_0 \times d_1}$ are mapping matrices used for subspace decomposition. $y_k^I \in \mathbb{R}^{d_1}$, $y_k^M \in \mathbb{R}^{d_1}$ and $\gamma \in \mathbb{R}^{d_0}$ denote identity features, domain features and hybrid facial feature, respectively. To prevent redundancy of feature representation and alleviate over-fitting, two orthogonal self-representation constraints are further imposed to the projection matrices, which can be expressed as follow:

$$\begin{aligned} cos(\theta_{ij}^I) &= \frac{W_i^{I^T} W_j^I}{\|W_i^I\|\|W_j^I\|} = 0 \quad (i \neq j) \\ cos(\theta_{ij}^M) &= \frac{W_i^{M^T} W_j^M}{\|W_i^M\|\|W_j^M\|} = 0 \quad (i \neq j) \end{aligned} \tag{4}$$

where θ_{ij}^I denotes angle between mapping vectors $W_i^I \in \mathbb{R}^{d_0}$ and $W_j^I \in \mathbb{R}^{d_0}$, and θ_{ij}^M represents angle between projecting vectors $W_i^M \in \mathbb{R}^{d_0}$ and $W_j^M \in \mathbb{R}^{d_0}$. In order to make the domain features and identity features irrelevant to each other [7], [43], orthogonal mutual-representation constraint is introduced into decomposed subspace, taking the following form:

$$cos(\theta_{ij}^{IM}) = \frac{W_i^{I^T} W_j^M}{\|W_i^I\|\|W_j^M\|} = 0 \quad (i, j = 1, 2, \ldots, d1) \tag{5}$$

where θ_{ij}^{IM} indicates angle between mapping vectors W_i^I and W_j^M. As illustrated in Fig. 2, the identity-related layer and modality-related layer aims to learn identity-dependent representations and domain-dependent representations, respectively, and orthogonal representation decorrelation (i.e., two orthogonal self-representation constraints as well as an orthogonal mutual-representation constraint) is imposed to these two layers to enhance representation capabilities, which can be expressed as follow:

$$\begin{aligned} l_I(y_i^I; \Theta, W^I) &= \kappa_1 S_{w1} + \kappa_2 S_{w2} + \kappa_3(S_{b1} + S_{b2} + S_{b3}) \\ l_M(y_i^M; \Theta, W^M) &= \frac{1}{n}\sum_{p}^{\{N,V\}}\sum_{i=1}^{n^p} \mathbf{I}(l_i^M = p)\|y_i^M - m^p\|_2^2 \\ &\quad + \frac{1}{2}[\eta - \|m^V - m^N\|_2^2]_+ \\ &\quad s.t.\ \frac{W_i^{I^T} W_j^I}{\|W_i^I\|\|W_j^I\|} = 0 \\ &\quad \frac{W_i^{M^T} W_j^M}{\|W_i^M\|\|W_j^M\|} = 0 \\ &\quad \frac{W_i^{I^T} W_j^M}{\|W_i^I\|\|W_j^M\|} = 0 \end{aligned} \tag{6}$$

where l_I and l_M denote identity-related objective loss and domain-related objective function, respectively. $n = n^N + n^V$, n^N and n^V are the number of heterogeneous images, NIR images and VIS images in a mini batch, and $m^V = \frac{1}{n^V}\sum_{k=1}^{n^V}\mathbf{I}(l_k^M = VIS)y_k^M$ and $m^N = \frac{1}{n^N}\sum_{k=1}^{n^N}\mathbf{I}(l_k^M = NIR)y_k^M$ denote the mean vector of VIS data and NIR data, respectively.

According to lagrange multipliers, the objective loss in the identity-dependent subspace can be reformulated as an unconstrained problem:

$$\begin{aligned} l_{MI}(y_i^I; \Theta, W^I) &= \kappa_1 S_{w1} + \kappa_2 S_{w2} + \kappa_3(S_{b1} + S_{b2} + S_{b3}) \\ &\quad + \sum_{i=1}^{d_1}\sum_{j=i+1}^{d_1} \mu_{ij}^I \left\|\frac{W_i^{I^T} W_j^I}{\|W_i^I\|\|W_j^I\|}\right\|_F^2 \\ &\quad + \sum_{i=1}^{d_1}\sum_{j=1}^{d_1} \mu_{ij} \left\|\frac{W_i^{I^T} W_j^M}{\|W_i^I\|\|W_j^M\|}\right\|_F^2 \end{aligned} \tag{7}$$

where $\|\cdot\|_F^2$ represents Frobenius norm, and μ_{ij}^I and μ_{ij} denote the lagrange multipliers. The objective function $l_{MI}(y_i^I; \Theta, W^I)$ is optimized using the identity-related dataset D^I. Similarly, the objective function of the domain-dependent subspace can be rewritten as an unconstrained problem, taking the following form:

$$\begin{aligned} l_{OMD}(y_i^M; \Theta, W^M) &= \frac{1}{n}\sum_{p}^{\{N,V\}}\sum_{i=1}^{n^p} \mathbf{I}(l_i^M = p)\|y_i^M - m^p\|_2^2 \\ &\quad + \frac{1}{2}[\eta - \|m^V - m^N\|_2^2]_+ \\ &\quad + \sum_{i=1}^{d_1}\sum_{j=i+1}^{d_1} \mu_{ij}^M \left\|\frac{W_i^{M^T} W_j^M}{\|W_i^M\|\|W_j^M\|}\right\|_F^2 \\ &\quad + \sum_{i=1}^{d_1}\sum_{j=1}^{d_1} \mu_{ij} \left\|\frac{W_i^{I^T} W_j^M}{\|W_i^I\|\|W_j^M\|}\right\|_F^2 \end{aligned} \tag{8}$$

where μ_{ij}^M is the lagrange multiplier. The modality dataset D^M is used to train $l_{OMD}(y_i^M; \Theta, W^M)$. Worthy of mention is that we not only expect to separate spectrum information in the domain-dependent subspace, but also to decouple the

Authorized licensed use limited to: Korea Institute of Science and Technology. Downloaded on January 27,2023 at 07:15:41 UTC from IEEE Xplore. Restrictions apply.

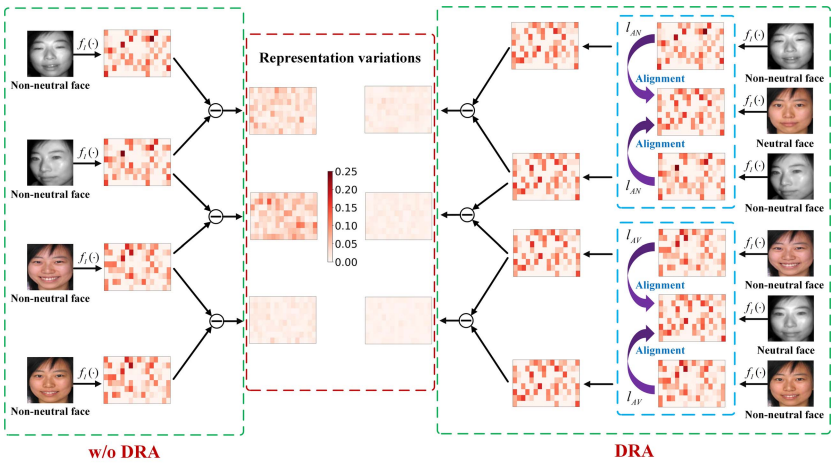


Fig. 4. The Deep Representation Alignment (DRA) performs alignment between non-neutral face and neutral face to learn residual-invariant representation (left: without DRA, right: introducing DRA). Compared with w/o DRA, the DRA can significantly reduce the residual variations.

modality-specific information from the backbone network. To attain the objective, we employ an adversarial process via a min-max two-player game [9], [42], [48], [49]. Firstly, the spectrum information in the domain-dependent layer can be learned by minimizing $l_{OMD}(y_i^M; \Theta, W^M)$:

$$W^{M*} = \arg\min_{W^M} l_{OMD}(y_i^M; \Theta, W^M) \tag{9}$$

Then, modality-specific information is disentangled from the backbone network by maximizing $l_{OMD}(y_i^M; \Theta, W^M)$:

$$\Theta^* = \arg\max_{\Theta} l_{OMD}(y_i^M; \Theta, W^M) \tag{10}$$

By introducing the OMD module, we can prevent redundancy of feature representation, reduce the correlation between the identity representation and modality representation, and decouple spectrum variations from the backbone network.

D. Deep Representation Alignment

The appearance discrepancies between heterogeneous images mainly reflect in identity, modality (VIS or NIR), and residual discrepancies (i.e., lighting, expressions, occlusion, blurry, pose, etc). Most of the current works [6], [7], [21] only attempt to eliminate the modality variations between cross-domain data. However, residual differences can also cause large within-class distance and reduce the NIR-VIS matching performance. To solve the above problem, the deep representation alignment is presented to eliminate the residual variations. As shown in Fig. 4, compared with w/o DRA (left), the DRA (right) obtains smaller intra-class representation variations. In particular, the DRA enforces a high-level features alignment between non-neutral face and neutral face to learn a discriminative face representation, and thus residual variations are significantly reduced. The high-level representation alignment can be expressed as follow:

$$l_{AN} = \frac{1}{c}\sum_{r=1}^{c}\sum_{i=1, l_i^I=r}^{n_r^N} \|y_i^N - \tilde{y}_r^V\|_2^2$$

$$l_{AV} = \frac{1}{c}\sum_{r=1}^{c}\sum_{i=1, l_i^I=r}^{n_r^V} \|y_i^V - \tilde{y}_r^N\|_2^2 \tag{11}$$

Fig. 5. The DRA reduces the variations of intra-class neutral faces and increases distances of inter-class neutral faces. PCA is applied to generate 2-D deep features.

where l_{AN} and l_{AV} indicate alignment losses for NIR and VIS images, respectively. $\tilde{y}_r^V$ and $\tilde{y}_r^N$ denote the neutral face representations of the rth class of VIS and NIR images [26], and y_i^N and y_i^V are the non-neutral face representations of NIR and VIS data. Notably, neutral face representation ($\tilde{y}_r^V$ and $\tilde{y}_r^N$) and non-neutral face representation (y_i^N and y_i^V) play the roles of "teacher" and "student". On the one hand, "students" should learn as much as possible from the "teacher", i.e., $\forall l_i^I = r$, $y_i^N \rightarrow \tilde{y}_r^V$ and $y_i^V \rightarrow \tilde{y}_r^N$, thereby facilitating the learning of residual-independent representations. On the other hand, in order to better guide "students", "teachers" should improve their own quality as much as possible. That is, as shown in Fig. 5, the DRA reduces the variations of intra-class neutral faces and increases distances of inter-class neutral faces. Specifically, the cross-domain intra-class neutral face distances are minimized, while cross-domain inter-class neutral face variations are maximized, which can be expressed as:

$$l_{AW} = \frac{1}{c}\sum_{r=1}^{c} \|\tilde{y}_r^V - \tilde{y}_r^N\|_2^2$$

$$l_{AB} = \frac{1}{c}\sum_{i=1}^{c}\sum_{j=1, j\neq i}^{c} [\xi - \|(\tilde{y}_i^V + \tilde{y}_i^N) - (\tilde{y}_j^V + \tilde{y}_j^N)\|_2^2]_+ \tag{12}$$

where l_{AW} and l_{AB} represent the cross-modal within-class variations loss and cross-modal between-class variations loss.

Authorized licensed use limited to: Korea Institute of Science and Technology. Downloaded on January 27,2023 at 07:15:41 UTC from IEEE Xplore. Restrictions apply.

ξ is the between-class margin value. According to Eq.(11) and Eq.(12), the objective loss in DRA can be expressed as:

$$l_{DRA} = \rho_1 l_{AW} + \rho_2 l_{AB} + \rho_3 l_{AN} + \rho_4 l_{AV} \tag{13}$$

where ρ is the trade-off coefficients for each part. Obviously, the first two items l_{AW} and l_{AB} aim to strengthen the discrimination of neutral face representation, while the l_{AN} and l_{AV} intend to guide the network to learn residual-invariant face representation.

Algorithm 1 OMDRA Model

Input: Training sets D^I and D^M, iterations number $iter$.
Output: Network parameters Θ, W^I.
1: Initialize Θ and W^I by pre-training network with MS-Celeb-1M database.
2: **for** $k = 1, 2, \ldots, iter$ **do**
3: Update Θ, W^I by minimizing l_{MI};
4: Update W^M by minimizing l_{OMD};
5: Update Θ by maximizing l_{OMD};
6: Update Θ, W^I by minimizing l_{DRA};
7: **end for**
8: **Return** Θ, W^I

E. Optimization Algorithm

The OMDRA contains three main components, i.e., MI, OMD and DRA. The proposed method naturally combines these three components into a unified framework. The final objective function in OMDRA can be written as:

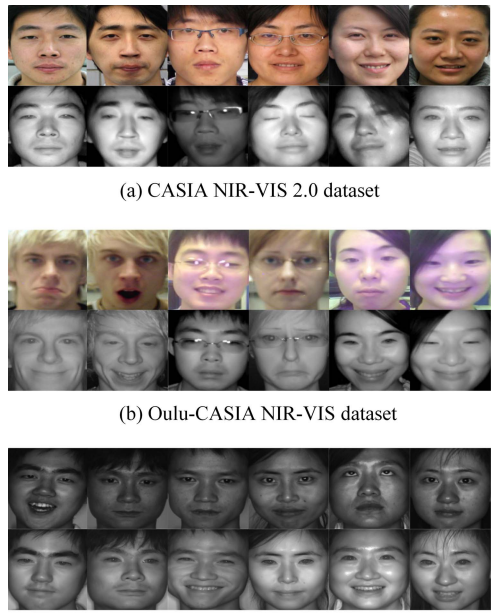
$$\mathcal{L} = l_{MI} + l_{OMD} + l_{DRA} \tag{14}$$

The parameters in the OMDRA network include Θ, W^I and W^M, and the optimization details of the model are explained in Algorithm 1. The network parameters in OMDRA are updated by loop iteration manner. The joint scheme leads to the disentanglement of modality differences, elimination of residual variations, and the purification of identity information.

F. Implementation Details

We evaluate the network in Tensorflow platform with a GPU of TITAN Xp. The Inception-ResNet-v1 [46], [50] is used as the backbone network. According to the settings in [50] and [9], we set d_0 and d_1 to 1792 and 128, respectively. The MS-Celeb-1M database [51] is cropped to 160×160 using MTCNN method [52], which is adopted to pre-train the backbone network using Softmax loss with accuracy of 99.4% on LFW [53].

During the fine-tuning stage, the NIR-VIS database is cropped to 160×160 using MTCNN, and is adopted to fine-tune the deep model. The batch size is 96, and the total number of training batches are about 2,700, 1,500 and 6,000 in BUAA NIR-VIS, Oulu-CASIA NIR-VIS and CASIA NIR-VIS 2.0, respectively. We set α, κ_1, κ_2, κ_3 and η to 1.3, 1.0, 0.1, 1.2 and 1.5, respectively. The hyper parameters of ρ_1, ρ_2, ρ_3, ρ_4, ξ are set to 0.1, 0.5, 1.0, 1.0, 5.0 in Oulu-CASIA NIR-VIS, 0.2, 2.0, 0.5, 0.5, 5.1 in CASIA NIR-VIS 2.0, and 1.0, 0.2, 0.5, 0.5, 8.0 in BUAA NIR-VIS, respectively. According to the experimental results, the hyper parameters set above can achieve better NIR-VIS matching performance.



(a) CASIA NIR-VIS 2.0 dataset

(b) Oulu-CASIA NIR-VIS dataset

(c) BUAA NIR-VIS dataset

Fig. 6. Cropped VIS and NIR facial images from the three datasets (VIS top, NIR bottom).

IV. Experiments Evaluation

In this section, extensive experiments are conducted on multiple cross-domain databases (such as BUAA NIR-VIS, Oulu-CASIA NIR-VIS and CASIA NIR-VIS 2.0) to illustrate the effectiveness of the OMDRA model. First, we investigate the effects of different components, including MI, OMD and DRA. Then, we compare of our method with other state-of-the-art methods, such as CFC [39], IDR [7], PACH [38], PCFH [37], WCNN [6], DSU-Nets [21], MC-CNN [23] and RCN [24], to demonstrate the superiority of OMDRA network.

A. Datasets and Protocols

CASIA NIR-VIS 2.0 Database [27] contains large within-class discrepancies such as spectrum, expressions, occlusion, pose and lighting, which contains 725 categories and 17,580 images in total. To make a fair comparison between the OMDRA approach and other approaches, we follow the testing protocol in [27] with ten-fold experiments. According to the protocol, the training set has approximately 360 individuals, and the testing set consists of the rest 358 categories. Notably, there are multiple NIR samples per subject in the probe, and the gallery set only contains one VIS sample per individual. Examples of cropped NIR and VIS face data are shown in Fig. 6 (a).

Oulu-CASIA NIR-VIS Database [28] is a difficult cross-domain dataset with illumination variations (i.e., normal indoor, dark and weak) and expression variations (i.e., happiness, sadness, disgust, surprise, anger and fear). To make a fair comparison with other approaches, we follow the testing

Authorized licensed use limited to: Korea Institute of Science and Technology. Downloaded on January 27,2023 at 07:15:41 UTC from IEEE Xplore. Restrictions apply.

TABLE I
PERFORMANCE OF DIFFERENT LOSS FUNCTIONS ON CASIA NIR-VIS 2.0

Loss function	VR@FAR=0.1%(%)	VR@FAR=0.01%(%)
Softmax loss	92.6	81.8
Center loss	94.8	86.1
Contrastive loss	96.6	87.5
Triplet loss	95.1	84.4
MI loss	**99.2**	**97.4**

protocol in [54]. In this protocol, there are 20 individuals in each testing and training set, and each individual consists of 48 VIS samples and 48 NIR samples. Examples of cropped NIR and VIS face data are shown in Fig. 6 (b).

BUAA NIR-VIS Database [29] contains expression and pose discrepancies, including sorrow, surprise, anger, happiness, tilt-down, tilt-up, right-rotation, left-rotation and neutral-frontal. There are total of 150 subjects in this dataset, and each subject has 9 VIS samples and 9 NIR samples. Following the testing protocol in [6], [36], we randomly select 50 individuals as the training set, and the testing set consists of the rest 100 individuals. In particular, the gallery set only has one VIS sample per subject, while the probe set has 9 NIR samples per individual. Examples of cropped NIR and VIS face data are shown in Fig. 6 (c).

B. Various Properties Analysis of the OMDRA on NIR-VIS Databases

Various experiments are conducted to evaluate the effectiveness of each part (i.e., MI, OMD and DRA) of the OMDRA network. In particular, there are ten-fold experiments in CASIA NIR-VIS 2.0 dataset, and the first fold is adopted to verify various properties of our method.

1) Effectiveness of MI Loss: In order to valid the effect of MI objective loss, we compare the MI loss with other objective functions such as Softmax loss, Center loss, Triplet loss [46] and Contrastive loss [47]. The backbone network is first pre-trained using Softmax loss in the MS-Celeb-1M dataset, and then is fine-tuned with different objective functions. The learning rates of Softmax loss, Center loss, MI loss, Triplet loss and Contrastive loss are 0.005, 0.005, 0.001, 0.001 and 0.001, and the number of training batches for these five losses are 4,200, 4,200, 3,600, 4,200 and 3,600. The performance of various losses is shown in Table I. Compared with Center loss, the MI loss get better performance on VR@FAR $= 0.1\%$, VR@FAR $= 0.01\%$ and rank-1 accuracy, and improves performance by 4.4%, 11.3% and 3.4%, respectively. Compared with Triplet loss and Contrastive loss, the MI loss obtains better performance with VR@FAR $= 0.1\%$ and VR@FAR $= 0.01\%$ of 99.2% and 97.4%, respectively. These results show that the MI loss can lean domain-invariant and identity-discriminative representations, because domain-related constraints (S_{w1}, S_{w2}, S_{b1}, S_{b2} and S_{b3}) are imposed to the network to increase domain adaptability.

2) Ablation Study: To valid the effect of each component in OMDRA network, we carry out ablation study in multiple NIR-VIS databases as shown in Table II-IV. The backbone network pre-trained with Softmax loss in MS-Celeb-1M dataset acts as the BaseLine (BL) method. BL+fine-tuning denotes that the backbone network is further fine-tuned with Triplet loss in NIR-VIS dataset.

TABLE II
ABLATION STUDY ON CASIA NIR-VIS 2.0 DATASET

Method	VR@FAR=0.1%(%)	VR@FAR=0.01%(%)
BL	18.5	3.9
BL+fine-tuning	95.1	84.4
BL+MI	99.2	97.4
BL+OMD (w/o MI)	95.3	87.8
BL+DRA	99.3	97.5
BL+MI+OMD	99.4	97.8
BL+MI+OMD+DRA	**99.6**	**98.0**

TABLE III
ABLATION STUDY ON OULU-CASIA NIR-VIS DATASET

Method	VR@FAR=1%(%)	VR@FAR=0.1%(%)
BL	19.4	3.2
BL+fine-tuning	96.6	83.5
BL+MI	98.1	88.2
BL+OMD (w/o MI)	96.1	87.8
BL+DRA	97.3	89.5
BL+MI+OMD	98.2	89.5
BL+MI+OMD+DRA	**98.5**	**92.2**

TABLE IV
ABLATION STUDY ON BUAA NIR-VIS DATASET

Method	VR@FAR=0.1%(%)	VR@FAR=0.01%(%)
BL	66.4	42.5
BL+fine-tuning	96.0	88.0
BL+MI	99.5	97.6
BL+OMD (w/o MI)	98.5	96.2
BL+DRA	99.5	98.8
BL+MI+OMD	99.6	99.3
BL+MI+OMD+DRA	**99.7**	**99.4**

According to the results in CASIA NIR-VIS 2.0 (Table II), the BL+fine-tuning obtains better performance than BL method on both VR@FAR $= 0.1\%$ and VR@FAR $= 0.01\%$, which indicates that fine-tuning in NIR-VIS dataset can effectively decouple the domain discrepancies between cross-modality data. When the MI loss is introduced to the network (BL+MI), the BL+MI model obtains better performance than BL and BL+fine-tuning on face verification task. These results indicate that MI loss can effectively train the network to learn spectrum-independent and identity-related features. BL+OMD (w/o MI) performs better than BL and BL+fine-tuning, with VR@FAR $= 0.01\%$ of 87.8% and VR@FAR $= 0.1\%$ of 95.3%, respectively. Worthy of mention is that we only use the OMD module with modality dataset D^M (without using the identity dataset D^I) to disentangle domain variations, persuasively comfirming the OMD can reduce domain variations and enhance representation capabilities. Compared with BL+fine-tuning, BL+DRA achieves better performance on both VR@FAR $= 0.1\%$ and VR@FAR $= 0.01\%$, and improves performance by 4.2% and 13.1%, respectively. These results show that it is necessary to reduce the residual-dependent information, and DRA can guide the network to learn residual-invariant face representations. When we introduce the OMD module to the BL+MI network, BL+MI+OMD has better performance than BL+MI

Authorized licensed use limited to: Korea Institute of Science and Technology. Downloaded on January 27,2023 at 07:15:41 UTC from IEEE Xplore. Restrictions apply.

TABLE V
ANALYSIS OF ORD IN OMDRA ON CASIA NIR-VIS 2.0, OULU-CASIA NIR-VIS AND BUAA NIR-VIS DATASETS

Method	CASIA NIR-VIS 2.0		Oulu-CASIA NIR-VIS		BUAA NIR-VIS	
	VR@FAR=0.1%	VR@FAR=0.01%	VR@FAR=0.1%	VR@FAR=0.01%	VR@FAR=0.1%	VR@FAR=0.01%
BL+MI+OMD (w/o ORD)	99.3	97.5	88.7	80.0	99.5	99.1
BL+MI+OMD	99.4	97.8	89.5	81.7	99.6	99.3

TABLE VI
THE PERFORMANCE OF OMD VARIANTS USING DIFFERENT REPRESENTATIVE FEATURES ON CASIA NIR-VIS 2.0

Method	Identity-related features (y^I)		Modality-related features (y^M)	
	rank-1	VR@FAR=0.1%	rank-1	VR@FAR=0.1%
BL+OMD (w/o Adv)	84.1	39.3	2.8	0.5
BL+OMD	96.8	95.3	94.1	91.2

with VR@FAR = 0.01% of 97.8% and VR@FAR = 0.1% of 99.4%, respectively. It indicates that OMD can decouple the modality-specific information via adversarial process and extract discriminative representations through the three orthogonal representation constraints. Finally, we impose the DRA to the BL+MI+OMD network, the BL+MI+OMD+DRA obtains the best performance on VR@FAR = 0.1%, and VR@FAR = 0.01% of 99.6% and 98.0%, respectively, which indicates that enforcing feature alignment on non-neutral face can effectively reduce residual-related information.

In particular, we obtain similar results on Oulu-CASIA NIR-VIS (as shown in Table III) and BUAA NIR-VIS (as shown in Table IV) datasets. Therefore, each component in OMDRA can improve the NIR-VIS FR performance, and the joint formulation can learn domain-invariant, residual-independent and identity-discriminative representations.

3) Effectiveness of ORD: Fig. 7 illustrates the correlation of matrices $W^{I^T}W^I$, $W^{M^T}W^M$ and $W^{I^T}W^M$ for different methods, including BL+MI+OMD and BL+MI+OMD without ORD (BL+MI+OMD (w/o ORD)), and a brighter color indicates a higher correlation. Compared with BL+MI+OMD (w/o ORD), the BL+MI+OMD approach has a brighter color in the diagonal elements in both $W^{I^T}W^I$ matrix (see Fig. 7 (a)) and $W^{M^T}W^M$ matrix (see Fig. 7 (b)), indicating that the two orthogonal self-representation constraints can effectively reduce the correlation in identity features and domain features, respectively. In addition, the BL+MI+OMD has darker color in $W^{I^T}W^M$ matrix than BL+MI+OMD (w/o ORD) method (refer to Fig. 7 (c)), which indicates that the orthogonal mutual-representation constraint can make the domain features and identity features irrelevant to each other.

We further analyze the performance of ORD on these NIR-VIS databases as shown in Table V. In CASIA NIR-VIS 2.0, BL+MI+OMD obtains better performance than BL+MI+OMD (w/o ORD) method, with VR@FAR = 0.1% of 99.4% and VR@FAR = 0.01% of 97.8%, respectively. In Oulu-CASIA NIR-VIS, the BL+MI+OMD performs better than BL+MI+OMD (w/o ORD) on both VR@FAR = 0.1% and VR@FAR = 0.01%, with improvement up to 0.8% and 1.7%, and similar results can be valided in BUAA NIR-VIS. Therefore, the ORD imposed to the OMDRA can enhance representation capabilities and improve the HFR performance.

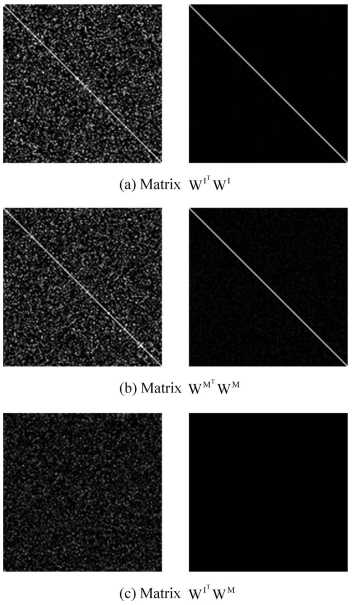


Fig. 7. The correlation illustration of matrices $W^{I^T}W^I$, $W^{M^T}W^M$ and $W^{I^T}W^M$ of different approaches (left: BL+MI+OMD (w/o ORD), right: BL+MI+OMD) on CASIA NIR-VIS 2.0. A brighter color indicates a higher correlation.

4) The Performance of OMD Variants: We analyze the performance of OMD variants using identity-related features (y^I) and modality-related features (y^M) on CASIA NIR-VIS 2.0. As shown in Table VI, BL+OMD indicates that we employ a min-max Adversarial (Adv) process (i.e., Eq. (9) and Eq. (10)) to train the network, while BL+OMD (w/o Adv) represents we only use minimization process (i.e., Eq. (9)) to train the network. When using modality-related features, BL+OMD (w/o Adv) has poor rank-1 accuracy of 2.8%, indicating that domain-specific information can be learned in modality-related layer. Compared with BL+OMD (w/o Adv), BL+OMD can effectively improve performance using modality-related features or identity-related features. The results indicate the min-max adversarial mechanism can effectively decouple spectrum variations from the backbone network.

5) Visualization Analysis: We visualize learning features $f_I(x_k)$ and representation variations on different datasets including CASIA NIR-VIS 2.0, Oulu-CASIA NIR-VIS and

Authorized licensed use limited to: Korea Institute of Science and Technology. Downloaded on January 27,2023 at 07:15:41 UTC from IEEE Xplore. Restrictions apply.

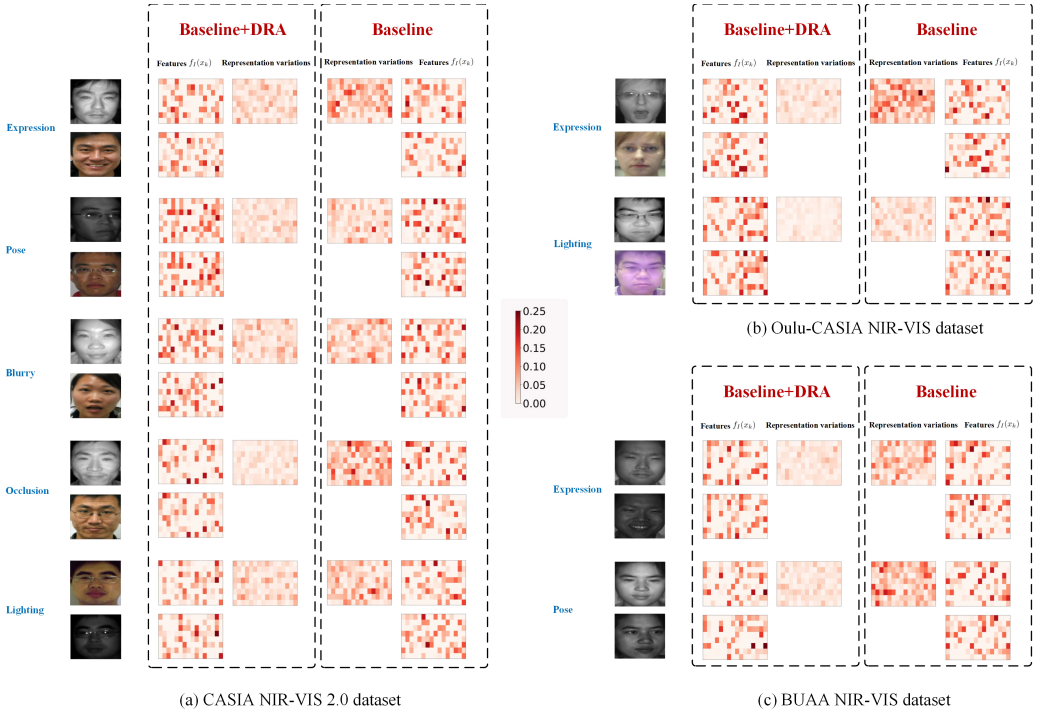


Fig. 8. Visualize learning features $f_I(x_k)$ and representation variations on different datasets, including (a) CASIA NIR-VIS 2.0, (b) Oulu-CASIA NIR-VIS, and (c) BUAA NIR-VIS. Compared with the Baseline method, the Baseline+DRA can effectively reduce the representation variations of the same individual.

TABLE VII

RANK-1 ACCURACY AND VERIFICATION RATE ON CASIA NIR-VIS 2.0 DATASET

Method	Rank-1(%)	VR@$FAR = 1\%$(%)	VR@$FAR = 0.1\%$(%)
KCSR (2009)	33.8	28.5	7.6
KPS (2013)	28.2	17.4	3.7
R-UDP (2015)	78.5 ± 1.7	-	85.8
Gabor RBM (2015)	86.2 ± 1.0	-	81.3 ± 1.8
MvDA (2016)	41.6 ± 4.1	-	19.2
H2-LBP3 (2016)	43.8	36.5	10.1
C-SLBFLE (2018)	86.9 ± 2.2	80.3	53.0
DAlign (2019)	58.4 ± 1.2	-	-
VGG (2015)	62.1 ± 1.9	70.9 ± 1.3	39.7 ± 2.9
SeetaFace (2016)	68.0 ± 1.7	85.2 ± 1.1	58.8 ± 2.3
TRIVET (2016)	95.7 ± 0.5	98.1 ± 0.3	91.0 ± 1.3
HFR-Net (2016)	87.1 ± 0.9	-	74.5
HFR-CNN (2016)	85.9 ± 0.9	-	78.0
Light CNN-9 (2015)	91.9 ± 0.6	-	85.3 ± 1.0
Light CNN-29 (2018)	96.7 ± 0.2	-	94.8 ± 0.4
IDR (2017)	97.3 ± 0.4	98.9 ± 0.3	95.7 ± 0.7
CDL (2018)	98.6 ± 0.2	-	98.3 ± 0.5
WCNN (2018)	98.7 ± 0.3	99.5 ± 0.1	98.4 ± 0.4
ADFL (2018)	98.1 ± 0.4	99.1 ± 0.2	97.1 ± 0.5
DSU-Nets (2019)	96.3 ± 0.4	-	98.4 ± 0.2
PCFH (2019)	98.8 ± 0.3	99.6 ± 0.1	97.7 ± 0.3
PACH (2019)	98.9 ± 0.2	99.6 ± 0.1	98.3 ± 0.2
RCN (2019)	99.3 ± 0.2	-	98.7 ± 0.3
MC-CNN (2019)	99.4 ± 0.1	-	99.3 ± 0.2
CFC (2020)	98.6 ± 0.1	99.2 ± 0.1	97.3 ± 0.2
OMDRA	**99.6** ± 0.1	**99.8** ± 0.1	**99.4** ± 0.2

BUAA NIR-VIS. As shown in Fig. 8, compared with the Baseline method, the Baseline+DRA can effectively reduce the representation variations of the same individual. Therefore, the DRA can learn residual-invariant face representation by enforcing a high-level feature alignment on non-neutral face images and neutral face images.

Moreover, we investigate the matching results between the NIR data and VIS data on CASIA NIR-VIS 2.0 dataset, as illustrated in Fig. 9, and the score below each sample denotes the matching score between the VIS gallery sample and NIR probe sample. Compared to the Baseline method, we find that our OMDRA method has a larger gap between the impostor scores (Rank-2) and the ground-truth scores

Authorized licensed use limited to: Korea Institute of Science and Technology. Downloaded on January 27,2023 at 07:15:41 UTC from IEEE Xplore. Restrictions apply.

TABLE VIII
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
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
CDL (2018)	94.3	81.6	53.9
WCNN (2018)	98.0	81.5	54.6
ADFL (2018)	95.5	83.0	60.7
PCFH (2019)	100	97.7	86.6
PACH (2019)	100	97.9	88.2
CFC (2020)	99.9	98.1	90.7
OMDRA	**100**	**98.5**	**92.2**

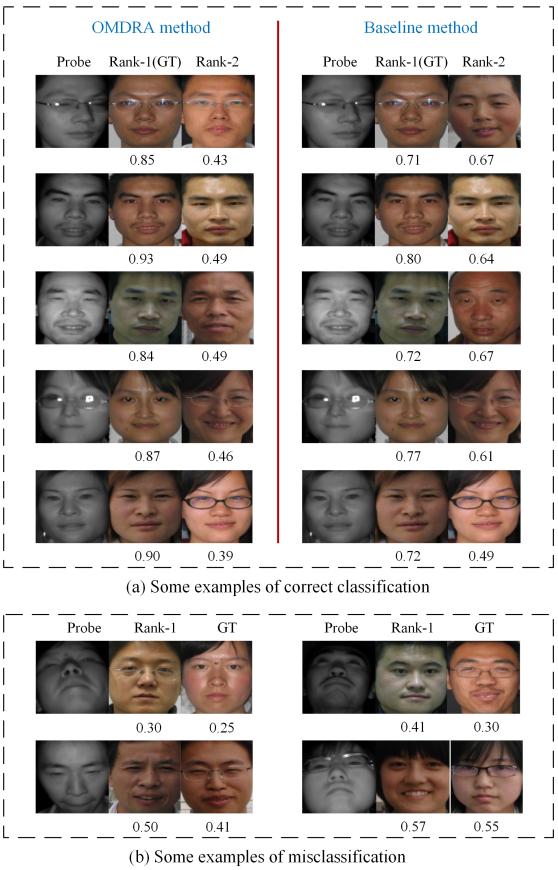


Fig. 9. Matching results of (a) correct classification by different methods (left: OMDRA method, right: Baseline method), and (b) misclassification samples by OMDRA mehtod on CASIA NIR-VIS 2.0 dataset. The score below each sample denotes the matching score between the VIS gallery sample and NIR probe sample.

('GT') in correct classification samples (refer to Fig. 9 (a)). The results indicate that the OMDRA model can well separate positive and negative samples, and our method can effectively reduce domain discrepancies, eliminate residual variations as well as retaining inherent identity information. There are totally 12 misclassification examples using OMDRA approach on CASIA NIR-VIS 2.0, and some false alarms examples are shown in Fig. 9 (b). We observe that the impostor face (Rank-1) has small similarity score, and those misclassification examples mainly come from extreme pose. Actually, the above issues are indeed the challenges of current FR task. These results show that the OMDRA network is effective for NIR-VIS FR matching.

C. Comparison With State-of-the-Art Methods

We compare the OMDRA model with other state-of-the-art methods, such as CFC [39], IDR [7], PCFH [37], PACH [38], WCNN [6], DSU-Nets [21], MC-CNN [23] and RCN [24], to verify the effectiveness of OMDRA network.

1) Performance Comparison on CASIA NIR-VIS 2.0 Database: In CASIA NIR-VIS 2.0 dataset, the OMDRA method is compared with various approaches, including KPS [3], KCSR [4], R-UDP [8], Gabor RBM [5], MvDA [13], H2-LBP3 [54], DAlign [55], C-SLBFLE [10], VGG [17], SeetaFace [56], TRIVET [19], HFR-Net [57], HFR-CNN [18], Light CNN-29 [16], Light CNN-9 [15], IDR [7], CDL [22], WCNN [6], DSU-Nets [21], ADFL [36], PACH [38], PCFH [37], RCN [24], MC-CNN [23] and CFC [39]. Table VII illustrates the performance of face recognition and face verification, and Fig. 10 (a) shows ROC curves of various approaches.

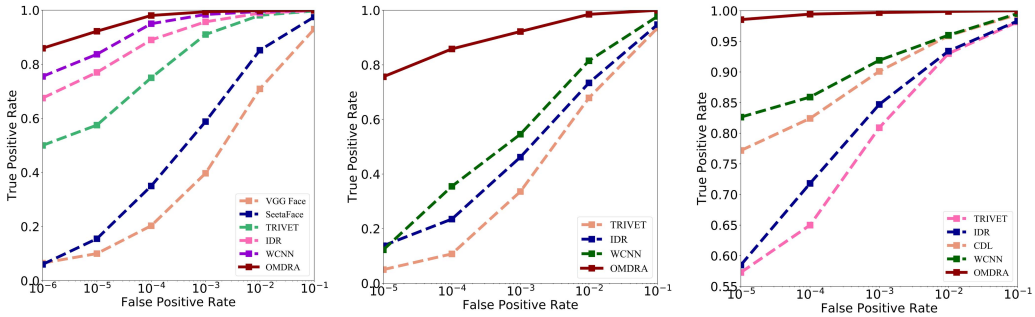
From the experimental results in CASIA NIR-VIS 2.0, the following conclusions can be drawn. Firstly, compared with the traditional approaches such as KPS, R-UDP, MvDA, C-SLBFLE and DAlign, the deep learning methods (e.g. RCN, CFC, IDR, WCNN, CDL, etc.) obtain better NIR-VIS HFR performance, indicating that CNN-based HFR approach can extract more discriminative representations for NIR-VIS FR task. Secondly, the OMDRA method outperforms the current deep learning methods, including IDR, CDL, WCNN, ADFL, PCFH, PACH, DSU-Nets, RCN, MC-CNN and CFC, with rank-1 accuracy, VR@FAR $= 1\%$ and VR@FAR $= 0.1\%$ of 99.6%, 99.8% and 99.4%, respectively. These results show that the sensing gap and residual variations can be disentangled by introducing MI, OMD and DRA into the network, and the OMDRA network can learn domain-invariant, residual-independent and identity-discriminative representations.

2) Performance Comparison on Oulu-CASIA NIR-VIS Database: In Oulu-CASIA NIR-VIS database, we compare the OMDRA approach with different competitors, including

Authorized licensed use limited to: Korea Institute of Science and Technology. Downloaded on January 27,2023 at 07:15:41 UTC from IEEE Xplore. Restrictions apply.

TABLE IX
RANK-1 ACCURACY AND VERIFICATION RATE ON BUAA NIR-VIS DATASET

Method	Rank-1(%)	VR@$FAR = 1\%$(%)	VR@$FAR = 0.1\%$(%)
MPL3 (2009)	53.2	58.1	33.3
KCSR (2009)	81.4	83.8	66.7
KPS (2013)	66.6	60.2	41.7
H2-LBP3 (2016)	88.8	88.8	73.4
TRIVET (2016)	93.9	93.0	80.9
IDR (2017)	94.3	93.4	84.7
CDL (2018)	96.9	95.9	90.1
WCNN (2018)	97.4	96.0	91.9
CFC (2020)	99.7	98.7	97.8
OMDRA	**100**	**99.9**	**99.7**



(a) CASIA NIR-VIS 2.0 ROC (b) Oulu-CASIA NIR-VIS ROC (c) BUAA NIR-VIS ROC

Fig. 10. The ROC curves on CASIA NIR-VIS 2.0, Oulu-CASIA NIR-VIS and BUAA NIR-VIS databases, respectively.

KPS [3], KCSR [4], H2-LBP3 [54], MPL3 [28], IDR [7], TRIVET [19], CDL [22], PACH [38], PCFH [37], ADFL [36], WCNN [6] and CFC [39]. Table VIII shows the results of verification rates and rank-1 accuracy, and Fig. 10 (b) shows ROC curves of various approaches.

According to the results in Oulu-CASIA NIR-VIS, we can draw the following conclusions. First of all, the OMDRA method significantly outperform the traditional approaches such as KPS, KCSR, H2-LBP3 and MPL3, which indicates that the OMDRA method can effectively eliminate the domain discrepancies between cross-modality face images. Second, compared with current CNN-based heterogeneous face matching approaches (i.e., IDR, CDL, WCNN, ADFL, PCFH, PACH and CFC), the OMDRA method achieves the best performance on both face recognition and face verification, and improves VR@FAR $= 1\%$ from 98.1% to 98.5% and VR@FAR $= 0.1\%$ from 90.7% to 92.2%. The results indicates that the OMDRA network can learn modal- and residual-independent representations, by taking both domain variations and residual variations into consideration. Therefore, the OMDRA model naturally combines MI, OMD and DRA into a unified framework, which can effectively decouple the within-class discrepancies and retain identity-related information.

3) Performance Comparison on BUAA NIR-VIS Database: In this database, the OMDRA method is compared with other methods, such as KPS [3], KCSR [4], H2-LBP3 [54], MPL3 [28], TRIVET [19], CDL [22], WCNN [6], IDR [7] and CFC [39]. Table IX and Fig. 10 (c) illustrate the performance of different methods. From the results, the OMDRA method outperforms the traditional methods (such as KPS, KCSR, H2-LBP3 and MPL3) and current CNN-based HFR approaches (such as IDR, TRIVET, WCNN, CDL and CFC).

TABLE X
RANK-1 ACCURACY AND VERIFICATION RATE ON TUFTS FACE DATASET

Method	Rank-1(%)	VR@$FAR = 1\%$(%)
Light CNN-29	50.2	29.8
BL	12.8	6.1
BL+fine-tuning	55.3	30.0
OMDRA (ours)	**75.2**	**52.5**

Specifically, the OMDRA improves rank-1 accuracy from 99.7% to 100%, VR@FAR $= 1\%$ from 98.7% to 99.9% and VR@FAR $= 0.1\%$ from 97.8% to 99.7%, respectively. The results reveal that the OMDRA can learn discriminative and robust representations for HFR task, persuasively comfirming the effectiveness of OMDRA model.

D. Generalization Analysis of the Proposed Methods

The previous experiments have verified the effectiveness of our method in NIR-VIS FR tasks, and we further conduct experiments on other heterogeneous face recognition tasks to verify the generality of the model. Tufts Face [58] is a challenging VIS-Thermal heterogeneous face database with expression, pose and occlusion variations. This dataset consists of 112 subjects from more than 15 countries. We randomly select 80 individuals as the training set including 1,104 VIS-Thermal pairs, and the rest 32 people are used as testing set containing 428 cross-domain pairs. Table X illustrates the performance of different approaches. Compared with BL+fine-tuning, our proposed OMDRA obtains better performance with improvement up to 19.9% on rank-1 and 22.5% on VR@FAR $= 1\%$, which indicates that the OMDRA

Authorized licensed use limited to: Korea Institute of Science and Technology. Downloaded on January 27,2023 at 07:15:41 UTC from IEEE Xplore. Restrictions apply.

is effective for VIS-Thermal face recognition task. Moreover, compared with Light CNN-29, the OMDRA achieves better performance with rank-1 of 75.2% and VR@FAR = 1% of 52.5%, indicating that our method can learn more discriminative facial representations. These experimental results verify the generalization and effectiveness of our model in other heterogeneous face tasks.

V. Conclusion

In this paper, a novel Orthogonal Modality Disentanglement and Representation Alignment network is developed for cross-domain FR task. The OMDRA network naturally integrates MI, OMD and DRA into an end-to-end framework. Specifically, the MI loss can increase intra-class compactness and inter-class separability. Then, the OMD structure comprising domain adversarial disentanglement and ORD aims to effectively disentangle modality discrepancy and enhance representation capabilities. Finally, the DRA is presented to learn residual-invariant face representation, by enforcing a high-level representation alignment between non-neutral face and neutral face. The joint scheme leads to the disentanglement of modality information, elimination of residual variations, and the purification of identity information. Experimental results on multiple challenging databases demonstrate the effectiveness of the OMDRA method over state-of-the-art approaches.

References

[1] C. Peng, X. Gao, N. Wang, and J. Li, "Superpixel-based face sketch–photo synthesis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 2, pp. 288–299, Feb. 2017.

[2] X. Gao, J. Zhong, J. Li, and C. Tian, "Face sketch synthesis algorithm based on E-HMM and selective ensemble," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 4, pp. 487–496, Apr. 2008.

[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] 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.

[5] D. Yi, Z. Lei, and S. Z. Li, "Shared representation learning for heterogenous face recognition," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, May 2015, pp. 1–7.

[6] R. He, X. Wu, Z. N. Sun, and T. N. Tan, "Wasserstein CNN: Learning invariant features for NIR-VIS face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1803–1816, Jun. 2019.

[7] 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. 2000–2006.

[8] F. J. Xu, D. K. Pal, and M. Savvides, "NIR-VIS heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2015, pp. 141–150.

[9] W. Hu and H. Hu, "Dual adversarial disentanglement and deep representation decorrelation for NIR-VIS face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 70–85, 2021.

[10] J. Lu, V. 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.

[11] M. Zhang, R. Wang, X. Gao, J. Li, and D. Tao, "Dual-transfer face sketch-photo synthesis," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 642–657, Feb. 2019.

[12] 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.

[13] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 188–194, Jan. 2012.

[14] X. Li *et al.*, "DeepSaliency: Multi-task deep neural network model for salient object detection," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3919–3930, Aug. 2016.

[15] 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: http://arxiv.org/abs/1511.02683

[16] 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.

[17] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis.*, vol. 1, no. 3, 2015, pp. 6–17.

[18] S. Saxena and J. Verbeek, "Heterogeneous face recognition with CNNs," in *Proc. Eur. Conf. Comput. Vis. Workshops*, Oct. 2016, pp. 483–491.

[19] 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.

[20] 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.

[21] T. de Freitas Pereira, A. Anjos, and S. Marcel, "Heterogeneous face recognition using domain specific units," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 7, pp. 1803–1816, Jul. 2018.

[22] X. Wu, L. X. Song, R. He, and T. N. Tan, "Coupled deep learning for heterogeneous face recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 1679–1686.

[23] Z. Deng, X. Peng, Z. Li, and Y. Qiao, "Mutual component convolutional neural networks for heterogeneous face recognition," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 3102–3114, Jun. 2019.

[24] Z. Y. Deng, X. J. Peng, and Y. Qiao, "Residual compensation networks for heterogeneous face recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8239–8246.

[25] W. Hu and H. Hu, "Adversarial disentanglement spectrum variations and cross-modality attention networks for NIR-VIS face recognition," *IEEE Trans. Multimedia*, vol. 23, pp. 145–160, 2021.

[26] W. Hu, W. Yan, and H. Hua, "Dual face alignment learning network for NIR-VIS face recognition," *IEEE Trans. Circuits Syst. Video Technol.*, early access, May 18, 2021, doi: 10.1109/TCSVT.2021.3081514.

[27] 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.

[28] J. 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.

[29] D. Huang, J. Sun, and Y. H. Wang, "The BUAA-VisNir face database instructions," Beihang Univ., Beijing, China, Tech. Rep. IRIP-TR-12-FR-001, 2012.

[30] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[31] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *Proc. CVPR*, Jun. 2011, pp. 513–520.

[32] 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.

[33] 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.

[34] W. Zhang, X. G. Wang, and X. O. Tang, "Lighting and pose robust face sketch synthesis," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 420–433.

[35] C. Peng, X. Gao, N. Wang, D. Tao, X. Li, and J. Li, "Multiple representations-based face sketch–photo synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 11, pp. 2201–2215, Nov. 2016.

[36] L. Song, M. Zhang, X. Wu, and R. He, "Adversarial discriminative heterogeneous face recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7355–7362.

[37] J. Yu, J. Cao, Y. Li, X. Jia, and R. He, "Pose-preserving cross spectral face hallucination," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 1018–1024.

[38] B. Duan, C. Fu, Y. Li, X. Song, and R. He, "Cross-spectral face hallucination via disentangling independent factors," 2019, *arXiv:1909.04365*. [Online]. Available: http://arxiv.org/abs/1909.04365

[39] R. He, J. Cao, L. Song, Z. Sun, and T. Tan, "Adversarial cross-spectral face completion for NIR-VIS face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1025–1037, May 2020.

Authorized licensed use limited to: Korea Institute of Science and Technology. Downloaded on January 27,2023 at 07:15:41 UTC from IEEE Xplore. Restrictions apply.

[40] X. Gao, N. Wang, D. Tao, and X. Li, "Face sketch-photo synthesis and retrieval using sparse representation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 8, pp. 1213–1226, Aug. 2012.
[41] N. Wang, X. Gao, L. Sun, and J. Li, "Anchored neighborhood index for face sketch synthesis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2154–2163, Sep. 2018.
[42] W. Hu and H. Hu, "Disentangled spectrum variations networks for NIR–VIS face recognition," *IEEE Trans. Multimedia*, vol. 22, no. 5, pp. 1234–1248, May 2020.
[43] 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.
[44] X. Shu, G.-J. Qi, J. Tang, and J. Wang, "Weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 35–44.
[45] J. H. Tang, X. G. Shu, and Z. Chao, "Generalized deep transfer networks for knowledge propagation in heterogeneous domains," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 12, no. 4s, pp. 1–12, 2016.
[46] 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.
[47] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. NIPS*, 2014, pp. 1988–1996.
[48] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.
[49] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: http://arxiv.org/abs/1411.1784
[50] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.
[51] Y. D. 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.
[52] 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.
[53] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 07-49, 2007.
[54] 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.
[55] 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.
[56] 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, Apr. 2017.
[57] 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 (CVPRW)*, Jun. 2016, pp. 54–62.
[58] K. Panetta *et al.*, "A comprehensive database for benchmarking imaging systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 509–520, Mar. 2020.

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 in electronics and information technology. His current research interests include computer vision, biometrics, heterogeneous face recognition, and person re-identification.

Haifeng Hu (Member, IEEE) received the Ph.D. degree from Sun Yat-sen University in 2004. He is currently a Professor with the School of Electronics and Information Technology, Sun Yat-sen University. He has published over 180 articles since 2000. His research interests include computer vision, pattern recognition, image processing, and neural computation.

Authorized licensed use limited to: Korea Institute of Science and Technology. Downloaded on January 27,2023 at 07:15:41 UTC from IEEE Xplore. Restrictions apply.