Multi-Margin based Decorrelation Learning for Heterogeneous Face Recognition

Bing Cao[1,2], **Nannan Wang**[1,3*], **Xinbo Gao**[1,2], **Jie Li**[1,2] and **Zhifeng Li**[4]
[1]State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an, China
[2]School of Electronic Engineering, Xidian University, Xi'an, China
[3]School of Telecommunications Engineering, Xidian University, Xi'an, China
[4]Tencent AI Lab, Shenzhen, China
bcao@stu.xidian.edu.cn, nnwang@xidian.edu.cn, {leejie,xbgao}@mail.xidian.edu.cn,
michaelzfli@tencent.com

Abstract

Heterogeneous face recognition (HFR) refers to matching face images acquired from different domains with wide applications in security scenarios. This paper presents a deep neural network approach namely Multi-Margin based Decorrelation Learning (MMDL) to extract decorrelation representations in a hyperspherical space for cross-domain face images. The proposed framework can be divided into two components: heterogeneous representation network and decorrelation representation learning. First, we employ a large scale of accessible visual face images to train heterogeneous representation network. The decorrelation layer projects the output of the first component into decorrelation latent subspace and obtains decorrelation representation. In addition, we design a multi-margin loss (MML), which consists of tetrad margin loss (TML) and heterogeneous angular margin loss (HAML), to constrain the proposed framework. Experimental results on two challenging heterogeneous face databases show that our approach achieves superior performance on both verification and recognition tasks, comparing with state-of-the-art methods.

1 Introduction

Heterogeneous face images refer to facial images acquired from different domains, such as visual (VIS) photo, near infrared (NIR) image, thermal infrared image, sketch and images with different resolutions, etc. In recent years, a great deal of efforts have been taken to heterogeneous face recognition, i.e. matching VIS face photos with cross-domain face images. However, different from impressive progress made in traditional face recognition, it is still a challenging problem for HFR.

Among these HFR scenarios, matching VIS face photos with NIR face images is the most straightforward and efficient solution to handle the extreme lighting conditions, which is of wide applications ranging from personal authorization to law enforcement. The lacking of sufficient training data in different domains and the significant cross-domain discrepancy are two most obstacles to train a robust model for NIR-VIS face recognition.

During the last decade, many large-scale VIS face datasets are available, which provide sufficient data to train convolution neural networks (CNN) for traditional face recognition and enormously improve the recognition performance. However, all the NIR face datasets are in small scale, which is not sufficient to train an effective CNN model without overfitting.

To address the other obstacle of significant cross-domain discrepancy, existing methods can be grouped into three categories. Synthesis-based methods transform the cross-domain face images to the same domain. Feature-based methods learn the invariant feature representation for the same identity in different domains. Subspace-based methods project the cross-domain face images to the same domain for HFR. However, these methods can not remove the discrepancy completely and the accuracy is not satisfactory.

In this paper, we propose a novel Multi-Margin based Decorrelation Learning (MMDL) framework to tackle the two aforementioned obstacles. The proposed framework contains two components: heterogeneous representation network and decorrelation representation learning. For the first obstacle, we employ a large scale of accessible visual face images to train heterogeneous representation network, which consists of input layers, output layers and four residual groups. We utilize this network to extract feature representation in a hyperspherical space that is robust to intra-class invariance and inter-class variance in VIS domain. Then, transfer learning is employed to improve the adaptation of this network to cross-domain face images. We impose a decorrelation layer to this network and design a multi-margin loss (MML) to fine-tune it. MML is utilized to minimize the cross-domain intra-class distance and further maximize cross-domain inter-class distance in the hyperspherical space. MML consists of the tetrad margin loss (TML) and the heterogeneous angular margin loss (HAML).

Our main contributions are summarized as follows:

- We design an effective end-to-end framework to extract invariant representation for both NIR and VIS face images. A decorrelation layer is imposed to heterogeneous representation networks to estimate the decorrelation latent subspace, which results in the two networks share

*Corresponding author: Nannan Wang (nnwang@xidian.edu.cn)

the same parameters.

- The multi-margin loss is designed to constrain the proposed framework. MML contains two components: TML and HAML, which are effective to minimize cross-domain intra-class distance and further maximize cross-domain inter-class distance.
- We propose an alternative optimization to fine-tune the heterogeneous representation networks and decorrelation representation learning, which improves the performance of the proposed framework.
- Experimental results on two challenging HFR databases illustrate that the proposed framework achieves superior performance, comparing with state-of-the-art methods. In addition, we conduct ablation study to demonstrate the effectiveness of various parts of the proposed approach.

2 Related Work

Matching NIR-VIS face images has become an important challenge in biometrics identification and great efforts have been made by researchers to solve this problem in the past decade. Existing HFR methods can be mainly grouped into three categories: synthesis-based methods, feature-based methods and subspace-based methods.

Synthesis-based methods try to synthesize heterogeneous face images from source domain to target domain and compare them in the same domain. These methods are designed to reduce the discrepancy between heterogeneous images in pixel-level. [Cao *et al.*, 2019] take multiple synthesized pseudo face images to improve the recognition accuracy. Generative adversarial network (GAN) was employed to synthesize heterogeneous face images in ADFL [Song *et al.*, 2018]. Though, synthesized methods can reduce the discrepancy between heterogeneous face images in pixel-level, the discriminative details are lost seriously, which affects the final recognition performance.

Feature-based methods aim at extracting invariant feature representation for heterogeneous face images. These methods are designed to reduce the cross-domain discrepancy in feature-level. To alleviate overfitting, a Coupled Deep Learning (CDL) [Wu *et al.*, 2017] introduced nuclear norm constraint on fully connected layer and proposed a cross-modal ranking to reduce cross-domain discrepancy. [He *et al.*, 2018] utilized Wasserstein distance to decrease the domain gap and acquire cross-domain invariant representation. Due to the great discrepancy between heterogeneous face images, it is hard to extract cross-domain invariant feature representation.

Subspace-based methods attempt to minimize the cross-domain discrepancy by project cross-domain face features onto a common subspace. In this subspace, heterogeneous face images can be measured directly. [Yi *et al.*, 2015] employed Restricted Boltzmann Machines (RBMs) to learn the cross-domain representation for NIR and VIS face images. The relationship of cross-domain face images was employed in [Kan *et al.*, 2016] to develop a multi-view discriminant analysis (MvDA) for HFR. However, it is inevitable to loss valid information in the process of projection, which seriously affects the final performance. Different from existing methods, our MMDL framework takes advantages of both feature-based methods and subspace-based methods.

3 Proposed Methods

Our framework contains two key components: heterogeneous representation networks and decorrelation representation learning as shown in Figure. 1. The first component extracts the low-dimensional feature representation of NIR and VIS face images. Then, the second component reduces the correlation of these low-dimensional representations and obtain the decorrelation representations that can be measured by cosine distance. In addition, we design the multi-margin loss (MML), which contains tetrad margin loss (TML) and heterogeneous angular margin loss (HAML), to optimize the proposed framework. In this section, we will detail the proposed multi-margin based decorrelation learning (MMDL) framework and the corresponding optimization scheme.

3.1 Decorrelation Representation Learning

Let Φ donates heterogeneous representation network. For heterogeneous face images, different samples of the same identity share the same invariant feature representation. The proposed network aims at extracting the invariant feature representation. Thus, the parameters W^H of Φ are learned from both NIR and VIS face images. For a NIR image x^N and VIS image x^V, the feature representations $y^i \in \mathbb{R}^n$ $(i \in \{N, V\})$ that are extracted from Φ can be donated as

$$y^i = \Phi(x^i, W^H) \quad (i \in \{N, V\}), \tag{1}$$

where $\Phi(\cdot)$ is the forward computation process of heterogeneous representation network. H donates the heterogeneous representation network. N and V represent the NIR domain and VIS domain respectively.

Decorrelation Representation

As demonstrated in previous work [Chen *et al.*, 2017], a face image x can be represented by identity information and variations that contains lighting, pose and expression. For HFR, spectrum information is also a kind of variations. As these variations of different samples are correlated [He *et al.*, 2018], it is hard to learn a discriminative model and achieve satisfactory performance on matching NIR and VIS face images. Therefore, we impose a decorrelation layer D to heterogeneous representation network Φ, in order to project feature representations y^i $(i \in \{N, V\})$ to the decorrelation latent subspace. Then, the outputs of decorrelation layer are the decorrelation representations of heterogeneous face images, which can be donated as follows:

$$z^i = (W^D)^T y^i \quad (i \in \{N, V\}), \tag{2}$$

where $z^i \in \mathbb{R}^q$ $(i \in \{N, V\})$ donates the decorrelation representations of NIR and VIS face images respectively. $W^D \in \mathbb{R}^{n \times q}$ represents the parameters of decorrelation layer and D donates the decorrelation layer. Therefore, we turn the decorrelation representation issue into estimating the parameter W^D of decorrelation layer.

For training data $X^i = \{x^i_1, x^i_2, \cdots, x^i_m\}$ $(i \in \{N, V\})$, the feature representations $Y^i = \{y^i_1, y^i_2, \cdots, y^i_m\}$ $(i \in$

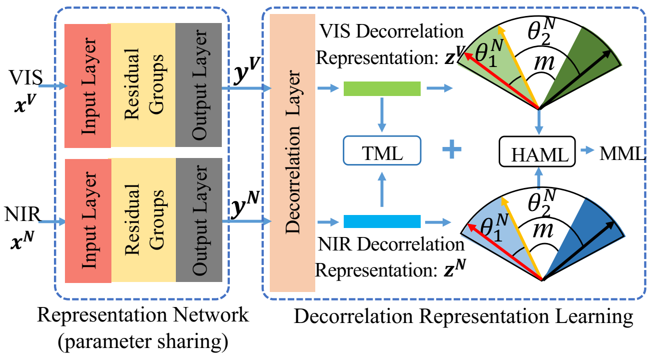


Figure 1: An illustration of the proposed MMDL framework. The heterogeneous representation network is used to extract the feature representation of the input cross-domain face images, the decorrelation layer is utilized to reduce the correlation of variations between the feature representation extracted by heterogenous representation network. With the online tetrad samples selection, the tetrad margin loss (TML) is computed by the decorrelation representation. The heterogeneous angular margin loss is introduced to joint optimize the proposed framework with TML. We can directly measure the similarity by decorrelation representation.

$\{N, V\}$) and decorrelation representations $Z^i = \{z_1^i, z_2^i, \cdots, z_m^i\}$ $(i \in \{N, V\})$ can be obtained by feedforward computation.

Suppose that, the correlation of decorrelation representations is minimized, $Z_j^i = \{z_{j,1}^i, z_{j,2}^i, \cdots, z_{j,q}^i\}$ should be in the coordinates with standard orthogonal basis $\{w_k, w_2, \cdots, w_q\}$, and $z_{j,k}^i = w_k^T y_j^i$. If we reconstruct y_j^i by z_j^i, we can obtain $\hat{y} = \sum_{k=1}^q z_{j,k}^i w_k$. Meanwhile, as we utilize cosine distance to measure the similarity of two input cross-domain face images, the angle θ between $\hat{y}^i$ and $\hat{y}^i$ should be minimize in order to avoid discriminative information loss. Considering the properties of Cosine function, our objective function is as follows:

$$\max \sum_{j=1}^m \cos\theta_j = \sum_{j=1}^m \frac{y_j^i \cdot \hat{y}_j^i}{\| y_j^i \| \| \hat{y}_j^i \|} \quad (i \in \{N, V\}) \\ \propto \max \sum_{j=1}^m (\cos\theta_j)^2 \tag{3}$$

Therefore, our objective function turns into

$$\begin{aligned} \min \sum_{j=1}^m -(\cos\theta_j)^2 &= \sum_{j=1}^m -(\frac{y_j^i \cdot \hat{y}_j^i}{\| y_j^i \| \| \hat{y}_j^i \|})^2 \\ &= \sum_{j=1}^m -\frac{((y_j^i)^T \cdot \sum_{l=1}^q z_{j,l}^i w_k)^2}{(y_j^i)^T y_j^i (\sum_{l=1}^q z_{j,l}^i w_k)^2} \\ &= \sum_{j=1}^m -(w_k)^T (\sum_{l=1}^q \frac{y_j^i (y_j^i)^T}{(y_j^i)^T y_j^i}) w_k \\ &\propto -tr(W^T \frac{YY^T}{Y^T Y} W) \\ &s.t. \quad W^T W = I \end{aligned} \tag{4}$$

We can obtain W by the eigenvalue and eigenvector of $\frac{YY^T}{Y^TY}$ via the lagrangian multiplier, singular value decomposition and $z^i = W^T y^i$. Therefore, the parameters W^D of decorrelation layer can be estimated by the W.

Multi-Margin Loss

To constrain the proposed framework, we proposed a multi-margin loss, which contains the tetrad margin loss (TML) and the heterogeneous angular margin loss (HAML).

As demonstrated in [Schroff *et al.*, 2015], triplet loss is effective to improve the accuracy of traditional face recognition. However, different from traditional face recognition, HFR matches face images from different modalities. It is meaningless to constrain the distance of face images from the same domain. Therefore, the contribution of triplet loss to improve the performance on HFR is minimal. Considering the limitation of triplet loss, we propose tetrad margin loss.

Tetrad Margin Loss (TML) is designed to increase the distance between inter-class cross-domain images and decrease the distance between intra-class cross-domain images. In order to accelerate convergence, we also take within-domain negative pairs into consideration. The designed online tetrad samples selection strategy is employed to select groups of four heterogeneous decorrelation representations $\{z_j^N, z_j^V, z_k^N, z_l^V\}$ in a mini-batch as tetrad tuples, where $\{z_j^N, z_j^V\}$ shares the same identity. z_k^N donates the closest NIR representations to z_j^V from another identity and z_l^V donates the closest VIS representations to z_j^N from another identity. Due to we utilize cosine distance to measure the distance between different images. As shown in Figure. 1. The proposed TML can be computed as follows:

$$\begin{aligned} \mathcal{L}_{TML}(z_j^N, z_j^V, z_l^V) = &\sum_i^b (\frac{z_j^N \cdot z_j^V}{\|z_j^N\| \|z_j^V\|} - \frac{z_j^N \cdot z_l^V}{\|z_j^N\| \|z_l^V\|} + \alpha_1) \\ &+ \sum_i^b (\frac{z_j^N \cdot z_j^V}{\|z_j^N\| \|z_j^V\|} - \frac{z_j^V \cdot z_l^V}{\|z_j^V\| \|z_l^V\|} + \alpha_2), \end{aligned} \tag{5}$$

$$\begin{aligned} \mathcal{L}_{TML}(z_j^N, z_j^V, z_k^N, z_l^V) = &\mathcal{L}_{TML}(z_j^N, z_j^V, z_l^V) \\ &+ \mathcal{L}_{TML}(z_j^V, z_j^N, z_k^N), \end{aligned} \tag{6}$$

where b donates the number of tetrad tuples in a mini-batch. α_1 and α_2 are the tetrad margin. As shown in Figure. 1, TML is designed to decrease the cosine distance between intra-class cross-domain face images and increase the cosine distance between inter-class face images.

Heterogeneous angular margin loss (HAML) is inspired by [Liu *et al.*, 2017] and [Deng *et al.*, 2018], which is developed from Softmax loss. Softmax loss is widely used in classification tasks, which is presented as follows:

$$\mathcal{L}_s = -\frac{1}{b} \sum_{j=1}^b \log \frac{e^{(W_{c_j}^F)^T z_j^i}}{\sum_{v=1}^n e^{(W_v^F)^T z_j^i}} \quad (i \in \{N, V\}), \tag{7}$$

where z_j belongs to the c_j-th class. W_v^F donates the v-th column vector of the weights $W^F \in \mathbb{R}^{q \times c}$ in the last fully connected layer. The number of class is c. The target logit [Pereyra *et al.*, 2017] can be transformed to

$$(W_v^F)^T z_j = \|W_v^F\| \|z_j^i\| cos\theta_v^i \quad (i \in \{N, V\}), \tag{8}$$

We fix $\|W_v^F\| = 1$ and $\|z_j^i\| = s$ by L2 normalisation, where s is a constant. As we omit these constants, Eq.8 can be reformulated as $(W_v^F)^T z_j^i = cos\theta_v^i$. Then, all the feature representations are distributed in a hypersphere. The similarity of two face images is determined by the angle between the corresponding feature representations. According to [Deng *et al.*, 2018], an angular margin m is added within $\cos\theta$, the HAML can be defined as follows:

$$\begin{aligned}\mathcal{L}_{HAML} = &-\frac{\lambda_N}{b}\sum_{j=1}^{b}\log\frac{e^{s(\cos(\theta_{c_i}^N+m_1))}}{e^{s(\cos(\theta_{c_i}^N+m_1))}+\sum_{v=1,v\neq c_i}^{n} e^{s\cos\theta_v^N}}\\ &-\frac{\lambda_V}{b}\sum_{j=1}^{b}\log\frac{e^{s(\cos(\theta_{c_i}^V+m_2))}}{e^{s(\cos(\theta_{c_i}^V+m_2))}+\sum_{v=1,v\neq c_i}^{n} e^{s\cos\theta_v^V}},\\ &s.t. \quad \lambda_N + \lambda_V = 1,\end{aligned} \tag{9}$$

where λ_N and λ_V represent the trade-off parameters of loss learned from NIR domain VIS domain. As the basic network is pre-trained by VIS face images, we assign greater weight to loss learned from NIR data. The multi-margin loss can be donated as follows:

$$\mathcal{L}_{MML} = \lambda_1 \mathcal{L}_{TML} + \lambda_2 \mathcal{L}_{HAML} \tag{10}$$

where λ_1 and λ_2 are the trade-off parameters of tetrad margin loss and heterogeneous angular margin loss.

3.2 Optimization

An alternative optimization strategy for the proposed MMDL framework is introduced in this subsection. The parameter W^H of heterogeneous representation network is pre-trained by large-scale VIS face images. First, we fix the parameter W^H of heterogeneous representation network and extract the feature representation Y of training data X by Eq. 1. The parameter W^D of decorrelation layer is estimated by feature representation Y according to Eq. 3 and Eq. 4. Second, we fix the parameter W^D in the process of learning decorrelation representations Z by Eq. 2. Then, we utilize Z to compute multi-margin loss (MML) by Eq. 10 to optimize the parameter W^H of heterogeneous representation network. Finally, the parameter W^H and W^D are fixed, we can obtain the final feature representations of input cross-domain face images by the proposed framework and measure the similarity of them by cosine distance. We summarize the optimization details in Algorithm 1.

Algorithm 1 Multi-Margin Decorrelation for Heterogeneous Face Recognition

Require: Training NIR face images x^N, training VIS face images x^V, learning rate r, batch size b, the trade-off parameter λ.
Ensure: The parameter W^H of heterogeneous representation network and the parameter W^D of decorrelation layer.

```
Pre-train the parameter W^H of heterogeneous representation network.
Fix W^H and extract the feature representation Y by Eq. 1.
Estimate the parameter W^D of decorrelation layer by Y.

for t=1,...,T do
    Fix the W^D, compute the decorrelation representation Z.
    Compute loss L_MML by Eq. 10.
    Update W^H via back-propagation.
    Fix W^H, update W^D by Eq. 3 and Eq. 4.
end for
return W^H and W^D.
```

4 Experiments

In this section, we evaluate the proposed framework against some state-of-the-art methods, systemically. We conduct experiments on two popular heterogeneous face databases: CASIA NIR-VIS 2.0 Database [Li *et al.*, 2013], Oulu-CASIA NIR-VIS Database [Chen *et al.*, 2009]. Some cropped samples are shown in Figure. 2.

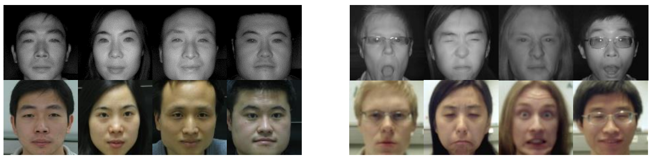
4.1 Databases and Protocols

The CASIA NIR-VIS 2.0 Face Database is the most challenging and the largest NIR-VIS database with large intra-class cross-domain variations, *i.e.* lighting, expression, pose. There are totally 725 subjects, each has 22 VIS face images and 50 NIR face images at most. We follow the partition protocols in [He *et al.*, 2018] and evaluate the proposed method on this database with 10-fold experiments. In the training phase, there are about 6100 NIR face images and 2500 VIS face images share 360 identities in each protocol. In the testing phase, there are about 6100 NIR face images in the probe set and 358 VIS face images in the gallery set. The similarity matrix of probe set and gallery set is 6100×358, computed by cosine distance. We compare the proposed method on this database at rank-1 recognition accuracy and verification rate (VR)@false accept rate (FAR) $= 0.1\%$.

The Oulu-CASIA NIR-VIS Database consists of 80 subjects with 6 expressions (*i.e.* anger, disgust, fear, happiness, sadness and surprise). We follow the protocols in [He *et al.*, 2018] and select 20 identities as the training set. For each expression, we randomly select 8 pairs of NIR-VIS face images, resulting in 96 cross-domain face images (48 pairs of NIR-VIS face images) for one identity. 20 identities are randomly selected from the remaining 60 identities as the testing set. The VIS face images in testing set are used as gallery and the corresponding NIR face images are used as probe. The similarities of all the NIR face images in the probe set and all the VIS face images in the gallery set are computed by cosine distance, which is a 960×960 similarity matrix. The rank-1 recognition accuracy and VR @ FAR $= 0.1\%$ are reported to evaluate the performance of the proposed method on this database.

Decorrelation Layer	Multi-Margin Loss		CASIA NIR-VIS 2.0		Oulu-CASIA NIR-VIS	
	TML	HAML	Rank-1	FAR=0.1%	Rank-1	FAR=0.1%
-	-	-	99.0	98.1	100	94.2
-	-	√	99.4	98.5	100	95.6
-	√	√	99.5	98.9	100	95.7
√	√	√	**99.9**	**99.4**	100	**97.2**

Table 1: The ablation study for the prposed MMDL framework. We explore the improvements benefiting from different components of the proposed framework.



(a) (b)

Figure 2: Some NIR-VIS face images. (a) the CASIA NIR-VIS 2.0 face database; (b) the Oulu-CASIA NIR-VIS face database. These samples are all aligned by five landmarks and cropped to the size of 112×112. The first row contains NIR face images and the second row contains the corresponding VIS face images.

4.2 Experiments Setting

The SE-LResNet50E-IR network [Deng *et al.*, 2018] is employed as the heterogeneous representation network in the proposed framework. As shown in Figure. 1, the heterogeneous representation network contains one input layer, four residual groups and one output layers, which is pre-trained on the MS-Celeb-1M database [Guo *et al.*, 2016]. All the images are aligned and cropped to 112×112 by five facial landmarks.

We implement all the experiments in this paper by PyTorch under the environment of Python 3.7 on Ubuntu 16.04 system with i7-6700K CPU and NVIDIA TITAN Xp GPU. The dimensions of feature representation y^i and decorrelation representation z^i are set to 512. Stochastic gradient descent (SGD) is utilized for back-propagation. We set the initial learning rate to $1e^{-4}$, which is gradually reduced to $1e^{-6}$. The batch size to 16 and the trade-off parameter λ_N, λ_V, λ_1 and λ_2 are set to 0.6, 0.4, 10 and 1 respectively. The angular margin m is set to 0.9 in this paper.

4.3 Comparsions

In this subsection, we first explore the improvements benefiting from the three parts of the proposed framework: decorrelation layer part, tetrad margin loss constraint and heterogeneous angular margin loss constraint. Then, we compare the performance of the proposed framework with state-of-the-art methods.

As shown in Table 1, we present the performance of rank-1 recognition accuracy and VR@FAR=0.1%. The baseline method utilizes the heterogeneous representation network, without decorrelation layer and multi-margin loss. For CASIA NIR-VIS 2.0 face database, the baseline method achieves rank-1 recognition accuracy of 99.0% and verification rate@FAR=0.1% of 98.1%. With the constraint of heterogeneous angular margin loss, we fine-tune the baseline network and improve the performance to 99.4% recognition accuracy at rank-1 and 98.5% VR@FAR=0.1%. The proposed tetrad margin loss contributes less to rank-1 recognition accuracy, but improves the verification rate from 98.5% to 98.9%. The proposed Decorrelation layer improves the proposed framework most, from 99.5% to 99.9% of rank-1 recognition accuracy and from 98.9% to 99.4% of VR@FAR=0.1%, which is very close to 100%. It demonstrates the effectiveness of the proposed framework.

As Oulu-CASIA NIR-VIS face database is a many-to-many database, each probe NIR face image has 8 corresponding VIS face images. Therefore, it is relatively easy to achieve satisfactory performance. As shown in Table 1, the baseline method achieves 100% recognition accuracy. However, our framework still improves the verification rate. By the constraint of heterogeneous angular margin loss, the verification rate is improved from 94.2% to 95.6%. The tetrad margin loss improves the verification rate marginally. The proposed decorrelation layer improves the verification rate much from 95.7% to 97.2%.

The experimental results demonstrate that all the three components in the proposed framework improve the performance of the baseline method on recognition accuracy and verification rate for HFR.

The rank-1 recognition accuracy and verification rates on CASIA NIR-VIS 2.0 Database is shown in Table 2. We compare the proposed approach with state-of-the-art HFR methods, including traditional methods(*i.e.* KCSR [Lei and Li, 2009], KPS [Klare and Jain, 2013], KDSR [Huang *et al.*, 2013], LCFS [Wang *et al.*, 2013], Gabor+RBM [Yi *et al.*, 2015], C-DFD [Lei *et al.*, 2014], CDFL [Jin *et al.*, 2015], H2(LBP3) [Shao and Fu, 2017] and CNN-based methods(*i.e.* VGG [Parkhi *et al.*, 2015], HFR-CNN [Saxena and Verbeek, 2016], TRIVET [Liu *et al.*, 2016], IDR [He *et al.*, 2017], ADFL [Song *et al.*, 2018], CDL [Wu *et al.*, 2017], WCNN [He *et al.*, 2018]), DVR [Wu *et al.*, 2018]. For traditional state-of-the-art methods, most of these methods try to learn a common subspace or invariant hand-designed feature representation. However, the representational ability of hand-designed feature is limited and it is hard to reduce the correlation of variations for the great gap between heterogeneous face images. Therefore, the performance of traditional HFR methods is not satisfactory. Gabor+RBM achieves the best performance on rank-1 recognition accuracy and VR@FAR=0.1% in comparing with traditional methods, which are only 86.2% and 85.8%. Owing to the strong rep-

Method	Rank-1	VR@FAR=0.1%
KCSR (2009)	33.8	7.6
KPS (2013)	28.2	3.7
KDSR (2013)	37.5	9.3
LCFS (2013)	35.4	16.7
H2(LBP3) (2017)	43.8	10.1
C-DFD (2014)	65.8	46.2
CDFL (2015)	71.5	55.1
Gabor+RBM (2015)	86.2	81.3
VGG (2015)	62.1	39.7
HFR-CNN (2016)	85.9	78.0
TRIVET (2016)	95.7	91.0
IDR (2017)	97.3	95.7
CDL (2017)	98.6	98.3
ADFL (2018)	98.2	97.2
WCNN (2018)	98.4	97.6
WCNN + low-rank	98.7	98.4
DVR (2018)	99.7	**99.6**
MMDL	**99.9**	99.4

Table 2: Comparisons of rank-1 recognition accuracy and VR@FAR=0.1% with state-of-the-art HFR methods on the CASIA NIR-VIS 2.0 database.

resentational ability of CNN-based method, these methods achieved better performance than traditional methods. The proposed approach also achieves comparable performance of 99.9% recognition accuracy and 99.4% VR@FAR=0.1%.

For the Oulu-CASIA NIR-VIS Database, we also compare the proposed approach with state-of-the-art methods, including traditional methods (*i.e.* MPL3 [Chen *et al.*, 2009], KCSR, KPS, KDSR, H2(LBP3)) and CNN-based methods (*i.e.* TRIVET, IDR, ADFL, CDL, WCNN, DVR). For the same reason in the experiments on CASIA NIR-VIS 2.0 database, CNN-based methods achieve much better performance of rank-1 recognition accuracy on this database. However, the performance of VR@FAR=0.1% is not satisfactory for both traditional methods and CNN-based methods. The best VR@FAR=0.1% of state-of-the-art methods is achieved by DVR with 84.9%. The proposed approach achieves superior performance on both rank-1 recognition accuracy of 100% and VR@FAR=0.1% of 97.2%. It demonstrates the effectiveness of the proposed framework. The experimental results are presented in Table 3.

We also evaluate the proposed method on CUHK VIS-NIR database [Gong *et al.*, 2017] and improve the rank-1 recognition accuracy from 83.9% to 99.7%.

Method	Rank-1	VR@FAR=0.1%
MPL3 (2009)	48.9	11.4
KCSR (2009)	66.0	26.1
KPS (2013)	62.2	22.2
KDSR (2013)	66.9	31.9
H2(LBP3) (2017)	70.8	33.6
TRIVET (2016)	92.2	33.6
IDR (2017)	94.3	46.2
IDR + low-rank	95.0	50.3
CDL (2017)	94.3	53.9
ADFL (2018)	95.5	60.7
WCNN (2018)	96.4	50.9
WCNN + low-rank	98.0	54.6
DVR (2018)	100	84.9
MMDL	**100**	**97.2**

Table 3: Comparisons of rank-1 recognition accuracy and VR@FAR=0.1% with state-of-the-art HFR methods on the Oulu-CASIA NIR-VIS database.

5 Conclusion

Considering that the correlation of cross-domain variations, this paper develops a multi-margin based decorrelation learning (MMDL) method, which employs a decorrelation layer to address this problem. The heterogeneous representation network is pre-trained by large-scale VIS face images. Then, the decorrelation layer is imposed on this network. The parameters of heterogeneous representation network and decorrelation layer are optimized by an alternative optimization strategy. The multi-margin loss is proposed to constrain the network in the fine-tune process, which contains two main components: tetrad margin loss and heterogeneous angular margin loss. Finally, the similarity of decorrelation representations in the hyperspherical space can be measured by Cosine distance. Experimental results on two popular heterogeneous face recognition databases demonstrate that the proposed MMDL framework significantly leads to superior performance in comparing with state-of-the-art methods. In addition, we explore an ablation study to show the improvements acquired by different components of the proposed MMDL framework.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (under Grant 61876142, 61432014, U1605252, 61772402, and 61671339), in part by the National Key Research and Development Program of China under Grant 2016QY01W0200, in part by the National High-Level Talents Special Support Program of China under Grant CS31117200001, in part by the Young Elite Scientists Sponsorship Program by CAST under Grant 2016QNRC001, in part by the Young Talent fund of University Association for Science and Technology in Shaanxi, China, in part by the CCF-Tencent Open Research Fund (No. RAGR20180105) and Tencent AI Lab Rhino-Bird Focused Research Program (No. JR201923), and in part by the Xidian University-Intellifusion Joint Innovation Laboratory of Artificial Intelligence, in part by the Fundamental Research Funds for the Central Universities under Grant JB190117.

References

[Cao *et al.*, 2019] Bing Cao, Nannan Wang, Jie Li, and Xinbo Gao. Data augmentation-based joint learning for heterogeneous face recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 30(6):1731–1743, June 2019.

[Chen *et al.*, 2009] Jie Chen, Dong Yi, Jimei Yang, Guoying Zhao, Stan Z Li, and Matti Pietikainen. Learning mappings for face synthesis from near infrared to visual light images. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 156–163. IEEE, 2009.

[Chen *et al.*, 2017] Dong Chen, Xudong Cao, David Wipf, Fang Wen, and Jian Sun. An efficient joint formulation for bayesian face verification. *IEEE Transactions on pattern analysis and machine intelligence*, 39(1):32–46, 2017.

[Deng *et al.*, 2018] Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *arXiv:1801.07698*, 2018.

[Gong *et al.*, 2017] Dihong Gong, Zhifeng Li, Weilin Huang, Xuelong Li, and Dacheng Tao. Heterogeneous face recognition: A common encoding feature discriminant approach. *IEEE Trans. Image Processing*, 26(5):2079–2089, 2017.

[Guo *et al.*, 2016] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pages 87–102. Springer, 2016.

[He *et al.*, 2017] Ran He, Xiang Wu, Zhenan Sun, and Tieniu Tan. Learning invariant deep representation for nir-vis face recognition. In *AAAI*, volume 4, page 7, 2017.

[He *et al.*, 2018] Ran He, Xiang Wu, Zhenan Sun, and Tieniu Tan. Wasserstein cnn: Learning invariant features for nir-vis face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[Huang *et al.*, 2013] Xiangsheng Huang, Zhen Lei, Mingyu Fan, Xiao Wang, and Stan Z Li. Regularized discriminative spectral regression method for heterogeneous face matching. *IEEE Transactions on Image Processing*, 22(1):353–362, 2013.

[Jin *et al.*, 2015] Yi Jin, Jiwen Lu, and Qiuqi Ruan. Coupled discriminative feature learning for heterogeneous face recognition. *IEEE Transactions on Information Forensics and Security*, 10(3):640–652, 2015.

[Kan *et al.*, 2016] Meina Kan, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. Multi-view discriminant analysis. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):188–194, 2016.

[Klare and Jain, 2013] Brendan F Klare and Anil K Jain. Heterogeneous face recognition using kernel prototype similarities. *IEEE transactions on pattern analysis and machine intelligence*, 35(6):1410–1422, 2013.

[Lei and Li, 2009] Zhen Lei and Stan Z Li. Coupled spectral regression for matching heterogeneous faces. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1123–1128. IEEE, 2009.

[Lei *et al.*, 2014] Zhen Lei, Matti Pietikäinen, and Stan Z Li. Learning discriminant face descriptor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):289–302, 2014.

[Li *et al.*, 2013] Stan Li, Dong Yi, Zhen Lei, and Shengcai Liao. The casia nir-vis 2.0 face database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 348–353, 2013.

[Liu *et al.*, 2016] Xiaoxiang Liu, Lingxiao Song, Xiang Wu, and Tieniu Tan. Transferring deep representation for nir-vis heterogeneous face recognition. In *Biometrics (ICB), 2016 International Conference on*, pages 1–8. IEEE, 2016.

[Liu *et al.*, 2017] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 1, 2017.

[Parkhi *et al.*, 2015] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *BMVC*, volume 1, page 6, 2015.

[Pereyra *et al.*, 2017] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

[Saxena and Verbeek, 2016] Shreyas Saxena and Jakob Verbeek. Heterogeneous face recognition with cnns. In *European Conference on Computer Vision*, pages 483–491. Springer, 2016.

[Schroff *et al.*, 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[Shao and Fu, 2017] Ming Shao and Yun Fu. Cross-modality feature learning through generic hierarchical hyperlingual-words. *IEEE transactions on neural networks and learning systems*, 28(2):451–463, 2017.

[Song *et al.*, 2018] Lingxiao Song, Man Zhang, Xiang Wu, and Ran He. Adversarial discriminative heterogeneous face recognition. In *AAAI Conference on Artificial Intelligence*, pages 7355–7362, 2018.

[Wang *et al.*, 2013] Kaiye Wang, Ran He, Wei Wang, Liang Wang, and Tieniu Tan. Learning coupled feature spaces for cross-modal matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2088–2095, 2013.

[Wu *et al.*, 2017] Xiang Wu, Lingxiao Song, Ran He, and Tieniu Tan. Coupled deep learning for heterogeneous face recognition. *arXiv preprint arXiv:1704.02450*, 2017.

[Wu *et al.*, 2018] Xiang Wu, Huaibo Huang, Vishal M Patel, Ran He, and Zhenan Sun. Disentangled variational representation for heterogeneous face recognition. *arXiv preprint arXiv:1809.01936*, 2018.

[Yi *et al.*, 2015] Dong Yi, Zhen Lei, and Stan Z Li. Shared representation learning for heterogenous face recognition. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–7. IEEE, 2015.