Adversarial Discriminative Heterogeneous Face Recognition

Lingxiao Song, Man Zhang, Xiang Wu, Ran He[*]

National Laboratory of Pattern Recognition, CASIA

Center for Research on Intelligent Perception and Computing, CASIA

Center for Excellence in Brain Science and Intelligence Technology, CAS

Abstract

The gap between sensing patterns of different face modalities remains a challenging problem in heterogeneous face recognition (HFR). This paper proposes an adversarial discriminative feature learning framework to close the sensing gap via adversarial learning on both raw-pixel space and compact feature space. This framework integrates cross-spectral face hallucination and discriminative feature learning into an end-to-end adversarial network. In the pixel space, we make use of generative adversarial networks to perform cross-spectral face hallucination. An elaborate two-path model is introduced to alleviate the lack of paired images, which gives consideration to both global structures and local textures. In the feature space, an adversarial loss and a high-order variance discrepancy loss are employed to measure the global and local discrepancy between two heterogeneous distributions respectively. These two losses enhance domain-invariant feature learning and modality independent noise removing. Experimental results on three NIR-VIS databases show that our proposed approach outperforms state-of-the-art HFR methods, without requiring of complex network or large-scale training dataset.

Figure 1: The proposed adversarial discriminative HFR framework. Adversarial learning is employed on both raw-pixel space and compact feature space.

Introduction

Face recognition research has been significantly promoted by deep learning techniques recently. But a persistent challenge remains to develop methods capable of matching heterogeneous faces that have large appearance discrepancy due to various sensing conditions. Typical heterogeneous face recognition (HFR) tasks conclude visual versus near infrared (VIS-NIR) face recognition (Yi et al. 2007; 2009), visual versus thermal infrared (VIS-TIR) face recognition (Socolinsky and Selinger 2002), face photo versus face sketch (Tang and Wang 2004; Wang and Tang 2009), face recognition across pose (Huang et al. 2017) and so on. VIS-NIR HFR is the most popular and representative task in HFR. This is because NIR imaging provides a low-cost and effective solution to acquire high-quality images under low-light scenarios. It is widely applied in surveillance systems nowadays. However, the popularization of NIR images is far from VIS images, and most face databases are enrolled in
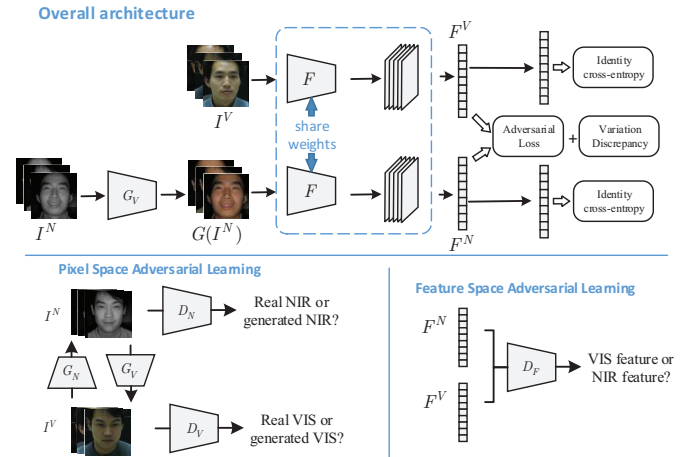
---
[*]corresponding author

VIS domain. Consequently, the demand for face matching between NIR and VIS images grows gradually.

A major challenge of HFR comes from the gap between sensing patterns of different face modalities. In practice, human face appearance is often influenced by many factors, including identities, illuminations, viewing angles, expressions and so on. Among all the factors, identity difference accounts for intra-personal differences while the rest lead to inter-personal differences. A key effort for face recognition is to alleviate intra-personal differences while enlarge inter-personal differences. Specifically, in the heterogeneous case, the noise factors that cause inter-personal differences show diverse distributions in different modalities, e.g. various spectrum sensing distribution between VIS domain and NIR domain, leading to a more complex problem to preserving the identity relevance between different modalities.

A lot of research efforts have been devoted to eliminating the sensing gap (Socolinsky and Selinger 2002; Yi et al. 2007; Li et al. 2013). One straightforward approach to cope with the sensing gap is to transform heterogeneous data onto a common comparable space (Lei et al. 2012). Another commonly used strategy is to map data from

one modality to another (Lei et al. 2008; Wang et al. 2009; Huang and Frank Wang 2013). Most of these methods only focus on minimizing the sensing gap, but not emphasize discrimination among different subjects, causing performance reduction when the number of subjects increases.

Another challenge for HFR is the lack of paired training data. General face recognition and hallucination have benefited a lot from the development of deep neural networks. However, the success of deep learning relies on large amount of labeled or paired training data to some extent. Although we can easily collect large-scale VIS images through the internet, it is hard to collect massive paired heterogeneous image data such as NIR images and TIR images. How to take the advantage of the powerful general face recognition to boost HFR and cross-spectral face hallucination is worth studying.

To address the above two issues, this paper proposes an adversarial discriminative feature learning framework for H-FR by introducing adversarial learning on both raw-pixel space and compact feature space. Figure 1 is the pipeline of our approach. Cross-spectral face hallucination and discriminative feature learning are simultaneously considered in this network. In the pixel space, we make use of generative adversarial networks (GAN) as a sub-network to perform cross-spectral face hallucination. An elaborate two-path model is introduced in this sub-network to alleviate the lack of paired images, which gives consideration to both global structures and local textures and results in a better visual result. In the feature space, an adversarial loss and a high-order variance discrepancy loss are employed to measure the global and local discrepancy between two heterogeneous feature distributions respectively. These two losses enhance domain-invariant feature learning and modality independent noise removing. Moreover, we implement all these global and local information in an end-to-end adversarial network, resulting in relatively compact 256 dimensional features. Experimental results show that our proposed adversarial approach not only outperforms state-of-the-art HFR methods but also can generate photo-realistic VIS images from NIR images, without requiring of complex network or large-scale training dataset. The results also suggest that the joint hallucination and feature learning is helpful to reduce the sensing gap.

The main contributions are summarized as follows,

- A cross-spectral face hallucination framework is embedded as a sub-network in adversarial learning based on GAN. A two-path architecture is presented to cope with the absence of well aligned image pairs and improve face image quality.

- An adversarial discriminative feature learning strategy is presented to seek domain-invariant features. It aims at eliminating the heterogeneities in compact feature space and reducing the discrepancy between different modalities in terms of both local and global distributions.

- Extensive experimental evaluations on three challenging HFR databases demonstrate the superiority of the proposed adversarial method, especially taking feature dimension and visual quality into consideration.

Related Work

What makes heterogeneous face recognition different from general face recognition is that we need to place data from different domains to the same space, only by which the measurement between heterogeneous data can make sense.

A kind of approaches uses data synthesis to map data from one modality into another. Thus the similarity relationship of heterogeneous data from different domain can be measured. In (Liu et al. 2005), a local geometry preserving based nonlinear method is proposed to generate pseudo-sketch from face photo. In (Lei et al. 2008), they propose a canonical correlation analysis (CCA) based multi-variate mapping algorithm to reconstruct 3D model from a single 2D NIR image. In (Wang and Tang 2009), multi-scale Markov Random Fields (MRF) models are extend to synthesize sketch drawing from given face photo and vice versa. In (Wang et al. 2009), a cross-spectrum face mapping method is proposed to transform NIR and VIS data to another type. Many works (Wang et al. 2012; Juefei-Xu, Pal, and Savvides 2015) resort to coupled or joint dictionary learning to reconstruct face images and then perform face recognition. However, large amount of pairwise multi-view data are essential for these methods based on data synthesis, making it very difficult to collect training images. In (Lezama, Qiu, and Sapiro 2016), they design a patch mining strategy to collect aligned image patches, and then produce VIS faces from NIR images through a deep learning approach.

Another kind of methods deals with heterogeneous data by projecting them to a common latent space respectively, or learn modality-invariant features that are robust to domain transfer. In (Lin and Tang 2006), Common Discriminant Feature Extraction (CDFE) is proposed to transform data to a common feature space, which takes both inter-modality discriminant information and intra-modality local consistency into consideration. (Liao et al. 2009) use DoG filtering as preprocessing for illumination normalization, and then employ Multi-block LBP (MB-LBP) to encode NIR as well as VIS images. (Klare and Jain 2010) further combine HoG features to LBP descriptors, and utilize sparse representation to improve recognition accuracy. (Goswami et al. 2011) incorporate a series of preprocessing methods to do normalization, then combine Local Binary Pattern Histogram (LBPH) representation with LDA to extract robust features. In (Zhang, Wang, and Tang 2011), a coupled information-theoretic projection method is proposed to reduce the modality gap by maximizing the mutual information between photos and sketches in the quantized feature spaces. In (Lei et al. 2012), a coupled discriminant analysis method is suggested that involves the locality information in kernel space. In (Huang et al. 2013), a regularized discriminative spectral regression (DSR) method is developed to map heterogeneous data into the same latent space. In (Hou, Yang, and Wang 2014), a domain adaptive self-taught learning approach is developed to derive a common subspace. In (Zhu et al. 2014), Log-DoG filtering is involved with local encoding and uniform feature normalization to reduce heterogeneities between VIS and NIR images. (Shao and Fu 2017) propose a hierarchical hyperlingual-words (Hwords) to cap-

ture high-level semantics across different modalities, and a distance metric through the hierarchical structure of Hwords is presented accordingly.

Recently, many works attempt to address the cross-modal matching problem by deep learning methods benefitting from the development of deep learning. In (Yi, Lei, and Li 2015) , Restricted Boltzmann Machines (RBMs) is used to learn a shared representation between different modalities. In (Liu et al. 2016), the triplet loss is applied to reduce intra-class variations among different modalities as well as augment the number of training sample pairs. (Kan, Shan, and Chen 2016) develop a multi-view deep network that is made up of view-specific sub-network and common sub-network, in which the view-specific sub-network attempts to remove view-specific variations while the common sub-network seeks for common representation shared by all views. In (He et al. 2017), subspace learning and invariant feature extraction are combined into CNNs. This method obtains the state-of-the-art HFR result on CASIA NIR-VIS 2.0 database.

As mentioned before, our work is also related to the famous adversarial learning. GAN (Goodfellow et al. 2014) has achieved great success in many computer vision applications including image style transfer (Zhu et al. 2017; Isola et al. 2017), image generation (Shrivastava et al. 2017; Huang et al. 2017) , saliency detection (Hu, Zhao, and Tan 2017) and object detection (Li et al. 2017; Wang, Shrivastava, and Gupta 2017). Adversarial learning provides a simple yet efficient way to fit target distribution via the min-max two-player game between generator and discriminator. Motivated by this, we introduce adversarial learning in NIR-VIS face hallucination and domain-invariant feature learning, aiming at closing the sensing gap of heterogeneous data in pixel space and feature space simultaneously.

The Proposed Approach

In this section, we present a novel framework for the cross-modal face matching problem based on adversarial discriminative learning. We first introduce the overall architecture, and then describe the cross-spectral face hallucination and the adversarial discriminative feature learning separately.

Overall Architecture

The goal of this paper is to design a framework that enables learning of domain-invariant feature representations for images from different modalities, i.e. VIS face images I^V and NIR face images I^N.

We can easily get numerous VIS face images for training thanks to the prosperous of social network. In most circumstances, face recognition approaches are trained with VIS face images, which cannot achieve full performance when handling with NIR images. Besides, it is necessary to archive all processed images for most face recognition systems in real-world applications. However, NIR face images are much harder to distinguish by humans comparing with VIS faces. A feasible way is to convert NIR face images into VIS spectrum. Thus, we employ a GAN to perform cross-spectral face hallucination, aiming at better fitting the VIS-

based face models as well as producing VIS-like images that are friendly to human eyes.

However, we find that it is insufficient that only transferring NIR images into VIS spectrum in NIR-VIS HFR. A reasonable explanation is that NIR images are distinct with VIS images not just on imaging spectrum. For example, NIR face images often have darker or blurrier outlines due to the distance limit of the near-infrared illumination. The special way of imaging for NIR images makes the noise factors that cause inter-personal differences show diverse distributions compared to the VIS images. Hence, an adversarial discriminative feature learning strategy is proposed in our approach to reduce heterogeneities between VIS and NIR images.

To summarize, the proposed approach consists of two key components (shown in Fig. 1): cross-spectral face hallucination and adversarial discriminative feature learning. These two components try to eliminate the gap between different modalities in raw-pixel space and compact feature space respectively.

Cross-spectral Face Hallucination

The outstanding performance of GAN in fitting data distribution has significantly promoted many computer vision applications such as image style transfer (Zhu et al. 2017; Isola et al. 2017). Motivated by its remarkable success, we employ GAN to perform the cross-spectral face hallucination that converting NIR face images into VIS spectrum.

A major challenge in NIR-VIS image converting is that image pairs are not aligned accurately in most databases. Even though we can align images based on facial landmarks, the pose and facial expression of the same subject still vary quite a lot. Therefore, we build our cross-spectral face hallucination models based on the CycleGAN framework (Zhu et al. 2017), which can handle unpaired image translation tasks. As illustrated in Fig. 1, a pair of generators $G_V : I^N \rightarrow I^V$ and $G_N : I^V \rightarrow I^N$ are introduced to achieve opposite transformation, with which we can construct mapping cycles between VIS and NIR domain. Associated with these two generators, D_V and D_N aim to distinguish between real images I and generated images $G(I)$ correspondingly.

Generators and discriminators are trained alternatively toward adversarial goals, following the pioneering work of (Goodfellow et al. 2014). The adversarial losses for generator and discriminator are shown in Eq. 1 and Eq. 2 respectively.

$$L_{G-adv} = -\mathbb{E}_{I \sim P(I)} \log D\left(G\left(I\right)\right), \quad (1)$$

$$L_{D-adv} = \mathbb{E}_{I' \sim P(I')} \log(1 - D(I^{'})) \\ + \mathbb{E}_{I \sim P(I)} \log D\left(G\left(I\right)\right), \quad (2)$$

where I and $I^{'}$ are images from different modalities.

In the CycleGAN framework, an extra cycle consistency loss L_{cyc} is introduced to guarantee consistency between input images and the reconstructed images, e.g. I^N vs. $G_N(G_V(I^N))$ and I^V vs. $G_V(G_N(I^V))$. L_{cyc} is calculated as

$$L_{cyc} = E_{I \sim P(I)} \| I - F\left(G\left(I\right)\right) \|_1, \quad (3)$$
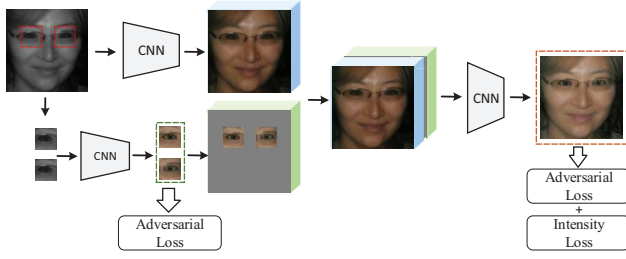
Figure 2: The proposed two-path architecture used in cross-spectral face hallucination.

where F is the opposite generator to G. In our cross-spectral face hallucination case, if G is used to transfer VIS faces into NIR spectrum, then F is used to transfer NIR faces into VIS spectrum.

We find that a single generator is hard to synthesize high quality cross-spectral images with both global structures and local details are well reconstructed. A possible explanation is that convolutional filters are shared across all the spatial locations, which are seldom suitable for recovering global and local information at the same time. Therefore, we employ a two-path architecture as shown is Fig. 2. Since the periocular regions show special correspondences between NIR images and VIS images diverse from other facial areas, we add a local path around eyes so as to precisely recover details of the periocular regions.

Because VIS images and NIR images mainly have difference in light spectrum, the structure information should be preserved after cross-spectral translations. Similar to (Lezama, Qiu, and Sapiro 2016), we choose to represent the input and output images in YCbCr space, for which the luminance component Y encode most structure information as well as identity information. An luminance-preserving term is adopted in the global path to enforce structure consistency:

$$L_{intensity} = \mathbb{E}_{I \sim P(I)} \| Y(I) - Y(G(I)) \|_1 \quad (4)$$

in which $Y(.)$ stands for the Y channel of images in YCbCr space.

To sum up, the full objective for generators G_V, G_N is:

$$L_G = L_{G-adv} + \alpha_1 L_{cyc} + \alpha_2 L_{intensity} \quad (5)$$

where α_1 and α_2 are loss weight coefficients.

Adversarial Discriminative Feature Learning

In this section, we propose a simple way to learn domain-invariant face representations using adversarial discriminative feature learning strategy. Ideal face feature extractor should be capable of alleviating the discrepancy caused by different modalities, while keeping discriminant among different subjects.

Adversarial Loss   As mentioned above, GAN has strong ability of fitting target distribution via the simple min-max two-player game. In this section, we use GAN in cross-view feature learning so as to eliminate domain discrepancy in

feature-level. As demonstrated in Fig. 1, an extra discriminator D_F is employed to act as the adversary to our feature extractor. D_F outputs a scalar value that indicates the probability of belonging to VIS feature space. The adversarial loss of our feature extractor takes the form:

$$L_{F-adv} = -\mathbb{E}_{I^N \sim P(I^N)} \log D_f \left(F \left(G_V \left(I^N \right) \right) \right) \quad (6)$$

By enforcing the fitting of NIR feature distribution to VIS feature distribution, we can remove the noise factors accounting for domain discrepancy. Since the adversarial loss is used to eliminate the discrepancy between distributions of heterogeneous data in a global view without taking local discrepancy into consideration, and distributions in each modalities consist of many sub-distributions of different subjects, the local consistency may not be well preserved.

Variance Discrepancy   Similar to the conventional domain adaptation tasks (Long et al. 2016; Zellinger et al. 2017), we want to bridge two different domains by learning domain-invariant feature representations in HFR. But HFR faces more challenges. First, HFR needs to match the same subject or instance rather than the same class, and distinguishe two different subjects that belong to the same class in most domain adaptation tasks. Second, there is no upper limit of the number of subject classes, the majority of which are not appeared in training phase. Fortunately, unlike these unsupervised domain adaptation tasks, label information in the target domain is supported in HFR, which can supervise the discriminative feature learning.

The usage of adversarial loss can only handle partial intrapersonal difference caused by modality transfer, but not the modality-independent noise factors. Considering that the feature distribution of the same subject should be as close as possible ideally, we employ the class-wise variance discrepancy (CVD) to enforce the consistency of subject-related variation with the guide of identity label information:

$$\sigma(F) = \mathbb{E}\left((F - \mathbb{E}(F))^2 \right), \quad (7)$$

$$L_{\text{CVD}} = \sum_{c=1}^{C} \mathbb{E}\left(\| \sigma\left(F_c^V \right) - \sigma\left(F_c^N \right) \|_2 \right) \quad (8)$$

where $\sigma(.)$ is the variance function, and the F_c^V, F_c^N denote feature observations belonging to the $c-$th class in VIS and NIR domain respectively.

Cross-Entropy Loss   As the adversarial loss and the variance discrepancy penalties cannot ensure the inter-class diversity which exists in both the source domain and the target domain, we further employ the common-used classification architecture to enforce the discrimination and compactness of the learned feature. Empirical error of all samples is minimized as

$$L_{cls} = \frac{1}{|N| + |V|} \sum_{i \in \{N, V\}} \mathcal{L}\left(WF_i, y_i \right) \quad (9)$$

where W is the parameter for softmax normalization, and $\mathcal{L}\left(\cdot,\cdot\right)$ is the cross-entropy loss function.

The final loss function is a weighted sum of all the losses defined above: L_{adv} to remove the modality gap, L_{CVD} to guarantee intra-class consistency, and L_{cls} to preserve identity discrimination.

$$L = L_{F-adv} + \lambda_1 L_{\text{CVD}} + \lambda_2 L_{cls} \qquad (10)$$

Experiments

In this section, we evaluate the proposed approach on three NIR-VIS databases. The databases and testing protocols are introduced firstly. Then, the implementation details is presented. Finally, comprehensive experimental analysis is conducted among the comparison with related works.

Datasets and Protocols

The CASIA NIR-VIS 2.0 face database (Li et al. 2013). It is so far the largest as well as the most challenging public face database across NIR and VIS spectrum. Its challenge contains large variations of the same identity, expression, pose and distance. The database collects 725 subjects, each with 1-22 VIS and 5-50 NIR images. All images in this database are randomly gathered, and no one-to-one correspondence between NIR and VIS images. In our experiments, we follow the View 2 of the standard protocol defined in (Li et al. 2013), which is used for performance evaluation. There are 10-fold experiments in View 2, where each fold contains non-overlapped training and testing lists. There are about 6,100 NIR images and 2,500 VIS images from about 360 identities for training in each fold. In the testing phase, cross-view face verification is taken between the gallery set of 358 VIS image belonging to different subjects, and the probe set of over 6,000 NIR images from the same 358 identity. The Rank-1 identification rate and the ROC curve are used as evaluation criteria.

The BUAA-VisNir face database (Huang, Sun, and Wang 2012). This dataset is made up of 150 subjects with 40 images per subject, among which there are 13 VIS-NIR pairs and 14 VIS images in different illumination. Each VIS-NIR image pairs are captured synchronously using a single multi-spectral camera. The paired images in the BUAA-VisNir dataset vary in poses and expressions. Following the testing protocol proposed in (Shao and Fu 2017), 900 images of 50 subjects are randomly selected for training, and the other 100 subjects make up the testing set. It is worth noted that the gallery set contains only one VIS image of each subject. Therefore, a testing set of 100 VIS images and 900 NIR images are organized. We report the Rank-1 accuracy and the ROC curve according to the protocol.

The Oulu-CASIA NIR-VIS facial expression database (Chen et al. 2009). Videos of 80 subjects with six typical expressions and three different illumination conditions are captured in both NIR and VIS imaging systems in this database. We take cross-spectral face recognition experiments following the protocols in (Shao and Fu 2017), where only images from the normal indoor illumination are used. In each expression, eight face images are randomly selected such that 48 VIS images and 48 NIR
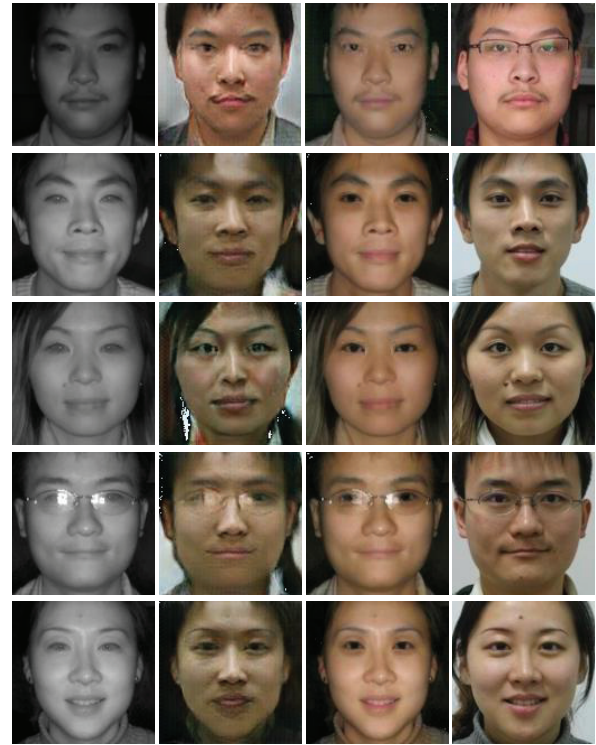

Figure 3: Results of the cross-spectral face hallucination. From left to right, the input NIR images, generated VIS images by cycleGAN, generated VIS images by the proposed cross-spectral face hallucination framework, and corresponding VIS images of the same subjects.

images of each subject are used. Based on the protocol in (Shao and Fu 2017), the training set and testing set contain 20 subjects respectively, resulting in a total of 960 gallery VIS images and 960 NIR probe images in testing phase. Similar to the above two datasets, the Rank-1 accuracy and the ROC curve are reported.

Implementation Details

Training data. Our cross-spectral hallucination network is trained on the CASIA NIR-VIS 2.0 face dataset. Note that the label annotation is not involved in the training of face hallucination module, therefore it would not affect the reliability of our following HFR tests. The feature extraction network is pre-trained on the MS-Celeb-1M dataset (Guo et al. 2016), and finetuned on each testing datasets respectively. All the face images are normalized by similarity transformation using the locations of two eyes, and then cropped to 144×144 size, of which 128×128 sized sub images are selected by random cropping in training and center cropping in testing. For the local-path, 32×32 patches are cropped around two eyes, and then flipped to the same side. As mentioned above, in the cross-spectral hallucination module, images are encoded in YCbCr space. In the feature extraction

step, grayscale images are used as input.

Network architecture. Our cross-spectral hallucination networks take the architecture of ResNet (He et al. 2016), where the global-path is comprised of 6 residual blocks and the local-path contains 3 residual blocks. Output of the local-path is feed to the global-path before the last block. In the adversarial discriminative feature learning module, we employ the model-B of the Light CNN (Wu, He, and Sun 2015) as our basic model, which includes 9 convolution layers, 4 max-pooling and one fully-connected layer. Parameters of the convolution layers are shared across the VIS and NIR channels as shown in Fig. 1. The output feature dimension of our approach is 256, which is relatively compact comparing with other state-of-the-art face recognition networks.

Experimental Results

Face Hallucination Results Fig. 3 shows some examples generated by our cross-spectral hallucination framework. We report the results of cycleGAN (Zhu et al. 2017) for comparison. As shown in Fig. 3, the results of cycleGAN are not satisfying, which may caused by the lack of strong constraint such as the proposed $L_{intensity}$. Note that our method can accurately recover details of the VIS faces, e.p. eyeballs, mouths and hairs. Specifically, the periocular regions are well transformed to VIS-like faces in which eyeballs are distinguishable. Results in Fig. 3 demonstrate the ability of our cross-spectral hallucination framework to generate photo-realistic VIS images from NIR inputs, with both global structure and local details are well preserved.

Results on the CASIA NIR-VIS 2.0 database Table 1 shows results of the proposed approach with different settings. We report mean value and standard deviation of Rank-1 identification rate, verification rates at 1%, 0.1%, 0.01% false accept rate (VR@FAR=1%, VR@FAR=0.1%, VR@FAR=0.01%) for a detailed analysis. We evaluate the performance obtained by our method in different settings, including cross-spectral hallucination, ADFL and hallucination + ADFL. In order to validate the effectiveness of L_{adv} and L_{CVD}, we report results of removing one of them respectively. The cross-spectral hallucination brings a performance gain for about 3% in Rank-1 accuracy as well as VR@FAR=1%, addressing that the cross-spectral image transfer helps to close the sensing gap between different modalities. Obviously, significant improvements can be observed when the proposed ADFL is used. Since supervision signals are introduced in the ADFL, it has stronger capacity than cross-spectral hallucination to boost the HFR accuracy. Both the adversarial loss and the variance discrepancy help to improve the recognition performance according to results of w/o L_{adv} and w/o L_{CVD}. When the cross-spectral hallucination and the adversarial discriminative learning strategies are applied together, the best performance is obtained.

We also compare the proposed approach with both conventional and state-of-the-art deep learning based NIR-VIS face recognition methods: PCA+Sym+HCA (Li et al. 2013), learning coupled feature space (LCFS) (Jin, Lu, and Ruan 2015), coupled discriminant face descriptor(CDFD) (Jin, Lu, and Ruan 2015; Wang et al. 2013),cou-

Table 2: Experimental results for the 10-fold face verification tasks on the CASIA NIR-VIS 2.0 database.

	Rank-1	FAR=0.1%	Dim.
PCA+Sym+HCA(2013)	23.70	19.27	-
LCFS(2015)	35.40	16.74	-
CDFD(2015)	65.8	46.3	-
CDFL(2015)	71.5	55.1	1000
Gabor+RBM(2015)	86.16	81.29	14080
Recon.+UDP(2015)	78.46	85.80	-
$H^2(LBP_3)$(2016)	43.8	10.1	-
COTS+Low-rank(2017)	89.59	-	1024
IDR(2017)	97.33	95.73	**128**
Ours	**98.15**	**97.18**	256

Table 3: Experimental results on the BUAA-VisNir Database.

	Rank-1	FAR=1%	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
KDSR(2013)	83.0	86.8	69.5
$H^2(LBP_3)$(2017)	88.8	88.8	73.4
IDR(2017)	94.3	93.4	84.7
Basic model	92.0	91.5	78.9
Softmax	94.2	93.1	80.6
ADFL w/o L_{CVD}	94.8	92.2	83.9
ADFL w/o L_{adv}	94.9	94.5	87.7
ADFL	**95.2**	**95.3**	**88.0**

pled discriminant feature learning (CDFL) (Jin, Lu, and Ruan 2015), Gabor+RBM (Yi, Lei, and Li 2015), NIR-VIS reconstruction+UDP (Juefei-Xu, Pal, and Savvides 2015), COTS+Low-rank citelezama2016not and Invariant Deep Representation (IDR) (He et al. 2017). The experimental results are consolidated in Table 2. We can see that deep learning based HFR methods perform much better than conventional approaches. The proposed method improves the previous best Rank-1 accuracy and VR@FAR=0.1%, which are obtained by IDR in (He et al. 2017), from 97.33% to 98.14% and 95.73% to 97.18% respectively. All of these results suggest that our method is effective for the NIR-VIS recognition problem.

Results on the BUAA-VisNir face database We compare the proposed approach with MPL3 (Chen et al. 2009), KCSR (Lei and Li 2009), KPS (Lei and Li 2009), KD-SR (Huang et al. 2013) and H^2(LBP$_3$ (Shao and Fu 2017). The results of these comparing methods are from (Shao and Fu 2017). Table 3 shows the Rank-1 accuracy and verification rate of each method. Profit from the powerful large-scale training data, the basic model achieves really good performance that is better than most of the comparing methods. We can see that performance can be further improved when adversarial loss and variance discrepancy are introduced. Particularly, without the constraint of variance consistency, the verification rate drops dramatically at low FAR.

Table 1: Experimental results for the 10-fold face verification tasks on the CASIA NIR-VIS 2.0 database of the proposed method.

	Rank-1 acc.(%)	VR@FAR=1%(%)	VR@FAR=0.1%(%)	VR@FAR=0.01%(%)
Basic model	87.16 ± 0.45	89.65 ± 0.89	72.06 ± 1.38	48.25 ± 2.68
Softmax	95.89 ± 0.75	98.26 ± 0.48	93.25 ± 1.14	75.13 ± 3.02
ADFL w/o L_{adv}	96.56 ± 0.63	98.56 ± 0.27	95.24 ± 0.36	81.69 ± 1.77
ADFL w/o L_{CVD}	97.34 ± 0.53	98.95 ± 0.14	96.88 ± 0.40	85.83 ± 3.02
Hallucination	90.56 ± 0.86	92.95 ± 0.20	81.17 ± 0.42	62.24 ± 2.77
ADFL	97.81 ± 0.29	99.04 ± 0.21	$\mathbf{97.21} \pm 0.34$	$\mathbf{88.11} \pm 3.09$
Hallucination + ADFL	$\mathbf{98.15} \pm 0.34$	$\mathbf{99.12} \pm 0.15$	97.18 ± 0.48	87.79 ± 2.33

Table 4: Experimental results on Oulu-CASIA NIR-VIS Database.

	Rank-1	FAR=1%	FAR=0.1%
MPL3(2009)	48.9	41.9	11.4
KCSR(2009)	66.0	49.7	26.1
KPS(2013)	62.2	48.3	22.2
KDSR(2013)	66.9	56.1	31.9
$H^2(LBP_3)$(2017)	70.8	62.0	33.6
IDR(2017)	94.3	73.4	46.2
Basic model	92.2	80.3	53.1
Softmax	93.0	80.9	56.1
ADFL w/o L_{CVD}	93.1	81.2	55.0
ADFL w/o L_{adv}	92.7	**83.5**	60.6
ADFL	**95.5**	83.0	**60.7**

This phenomenon demonstrates the effectiveness of variance discrepancy in removing intra-subject variations. Finally, the proposed ADFL acquires the best performance.

Results on the Oulu-CASIA NIR-VIS facial expression database Results on the Oulu-CASIA NIR-VIS are presented in Table4, in which the results of these comparing methods are from (Shao and Fu 2017). Similar to results on the BUAA-VisNir database, our proposed ADFL further boosts the performance beyond the powerful basic model. We observe that the adversarial loss contributes little to this database since the training set of Oulu-CASIA NIR-VIS database only contains 20 subjects and is relatively small-scale. So it is easy for the powerful Light CNN to learn good feature extractor for such a small dataset with the guidance of softmax loss. Besides, the variance discrepancy still shows great capability in promoting verification rate at low FAR. These results demonstrate the superiority of our method.

Conclusions

In this paper, we focus on the VIS-NIR face verification problem. An adversarial discriminative feature learning framework is developed by introducing adversarial learning in both raw-pixel space and compact feature space. In the raw-pixel space, the powerful generative adversarial network is employed to perform cross-spectral face hallucination, using a two-path architecture that is carefully designed to alleviate the absence of paired images in NIR-VIS transfer. As for the feature space, we utilize the adversarial loss and a high-order variance discrepancy loss to measure the global and local discrepancy between feature distributions of heterogeneous data respectively. The proposed cross-spectral face hallucination and adversarial discriminative learning are embedded in an end-to-end adversarial network, resulting in a compact 256-dimensional feature representation. Experimental results on three challenging NIR-VIS face databases demonstrate the effectiveness of the proposed method in NIR-VIS face verification.

Acknowledgments

References

Chen, J.; Yi, D.; Yang, J.; Zhao, G.; Li, S. Z.; and Pietikainen, M. 2009. Learning mappings for face synthesis from near infrared to visual light images. In *CVPR*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS*.

Goswami, D.; Chan, C. H.; Windridge, D.; and Kittler, J. 2011. Evaluation of face recognition system in heterogeneous environments (visible vs nir). In *ICCVW*.

Guo, Y.; Zhang, L.; Hu, Y.; He, X.; and Gao, J. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

He, R.; Wu, X.; Sun, Z.; and Tan, T. 2017. Learning invariant deep representation for nir-vis face recognition. In *AAAI*.

Hou, C.-A.; Yang, M.-C.; and Wang, Y.-C. F. 2014. Domain adaptive self-taught learning for heterogeneous face recognition. In *ICPR*.

Hu, X.; Zhao, Xin ans Huang, K.; and Tan, T. 2017. Adversarial learning based saliency detection. In *ACPR*.

Huang, D.-A., and Frank Wang, Y.-C. 2013. Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition. In *ICCV*.

Huang, X.; Lei, Z.; Fan, M.; Wang, X.; and Li, S. Z. 2013. Regularized discriminative spectral regression method for heterogeneous face matching. *ITIP* 22(1):353–362.

Huang, R.; Zhang, S.; Li, T.; and He, R. 2017. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *ICCV*.

Huang, D.; Sun, J.; and Wang, Y. 2012. The BUAA-VisNir face database instructions. Technical Report IRIP-TR-12-FR-001, Beihang University, Beijing, China.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *CVPR*.

Jin, Y.; Lu, J.; and Ruan, Q. 2015. Coupled discriminative feature learning for heterogeneous face recognition. *TIFS* 10(3):640–652.

Juefei-Xu, F.; Pal, D. K.; and Savvides, M. 2015. Nir-vis heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction. In *CVPRW*.

Kan, M.; Shan, S.; and Chen, X. 2016. Multi-view deep network for cross-view classification. In *CVPR*.

Klare, B., and Jain, A. K. 2010. Heterogeneous face recognition: Matching nir to visible light images. In *ICPR*.

Lei, Z., and Li, S. Z. 2009. Coupled spectral regression for matching heterogeneous faces. In *CVPR*.

Lei, Z.; Bai, Q.; He, R.; and Li, S. Z. 2008. Face shape recovery from a single image using cca mapping between tensor spaces. In *CVPR*.

Lei, Z.; Liao, S.; Jain, A. K.; and Li, S. Z. 2012. Coupled discriminant analysis for heterogeneous face recognition. *TIFS* 7(6):1707–1716.

Lezama, J.; Qiu, Q.; and Sapiro, G. 2016. Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding. In *CVPR*.

Li, S.; Yi, D.; Lei, Z.; and Liao, S. 2013. The casia nir-vis 2.0 face database. In *CVPRW*.

Li, J.; Liang, X.; Wei, Y.; Xu, T.; Feng, J.; and Yan, S. 2017. Perceptual generative adversarial networks for small object detection. In *CVPR*.

Liao, S.; Yi, D.; Lei, Z.; Qin, R.; and Li, S. Z. 2009. Heterogeneous face recognition from local structures of normalized appearance. In *ICB*.

Lin, D., and Tang, X. 2006. Inter-modality face recognition. In *ECCV*.

Liu, Q.; Tang, X.; Jin, H.; Lu, H.; and Ma, S. 2005. A nonlinear approach for face sketch synthesis and recognition. In *CVPR*.

Liu, X.; Song, L.; Wu, X.; and Tan, T. 2016. Transferring deep representation for nir-vis heterogeneous face recognition. In *ICB*.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2016. Unsupervised domain adaptation with residual transfer networks. In *NIPS*.

Shao, M., and Fu, Y. 2017. Cross-modality feature learning through generic hierarchical hyperlingual-words. *TNNLS* 28(2):451–463.

Shrivastava, A.; Pfister, T.; Tuzel, O.; Susskind, J.; Wang, W.; and Webb, R. 2017. Learning from simulated and unsupervised images through adversarial training. In *CVPR*.

Socolinsky, D. A., and Selinger, A. 2002. A comparative analysis of face recognition performance with visible and thermal infrared imagery. Technical report, DTIC Document.

Tang, X., and Wang, X. 2004. Face sketch recognition. *TCSVT* 14(1):50–57.

Wang, X., and Tang, X. 2009. Face photo-sketch synthesis and recognition. *TPAMI* 31(11):1955–1967.

Wang, R.; Yang, J.; Yi, D.; and Li, S. Z. 2009. An analysis-by-synthesis method for heterogeneous face biometrics. In *ICB*.

Wang, S.; Zhang, L.; Liang, Y.; and Pan, Q. 2012. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In *CVPR*.

Wang, K.; He, R.; Wang, W.; Wang, L.; and Tan, T. 2013. Learning coupled feature spaces for cross-modal matching. In *ICCV*.

Wang, X.; Shrivastava, A.; and Gupta, A. 2017. A-fast-rcnn: Hard positive generation via adversary for object detection. In *CVPR*.

Wu, X.; He, R.; and Sun, Z. 2015. A lightened cnn for deep face representation. *arXiv preprint arXiv:1511.02683*.

Yi, D.; Liu, R.; Chu, R.; Lei, Z.; and Li, S. Z. 2007. Face matching between near infrared and visible light images. In *ICB*.

Yi, D.; Liao, S.; Lei, Z.; Sang, J.; and Li, S. Z. 2009. Partial face matching between near infrared and visual images in mbgc portal challenge. In *ICB*.

Yi, D.; Lei, Z.; and Li, S. Z. 2015. Shared representation learning for heterogenous face recognition. In *FG*.

Zellinger, W.; Grubinger, T.; Lughofer, E.; Natschläger, T.; and Saminger-Platz, S. 2017. Central moment discrepancy (cmd) for domain-invariant representation learning. In *ICLR*.

Zhang, W.; Wang, X.; and Tang, X. 2011. Coupled information-theoretic encoding for face photo-sketch recognition. In *CVPR*.

Zhu, J.-Y.; Zheng, W.-S.; Lai, J.-H.; and Li, S. Z. 2014. Matching nir face to vis face using transduction. *TIFS* 9(3):501–514.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*.