Pose-preserving Cross-spectral Face Hallucination

Junchi Yu[1], **Jie Cao**[1,2], **Yi Li**[1,2], **Xiaofei Jia**[4] and **Ran He**[1,2,3*]
[1]NLPR&CRIPAC Institute of Automation, Chinese Academy of Sciences, China
[2]University of Chinese Academy of Sciences, China
[3]Center for Excellence in Brain Science and Intelligence Technology, CAS, China
[4]Central Media Technology Institute, Huawei Technology Co., Ltd., China
junchiyu2019@ia.ac.cn, {jie.cao, yi.li}@cripac.ia.ac.cn, jiaxiaofei2@huawei.com, rhe@nlpr.ia.ac.cn

Abstract

To narrow the inherent sensing gap in heterogeneous face recognition (HFR), recent methods have resorted to generative models and explored the "recognition via generation" framework. Even though, it remains a very challenging task to synthesize photo-realistic visible faces (VIS) from near-infrared (NIR) images especially when paired training data are unavailable. We present an approach to avert the data misalignment problem and faithfully preserve pose, expression and identity information during cross-spectral face hallucination. At the pixel level, we introduce an unsupervised attention mechanism to warping that is jointly learned with the generator to derive pixel-wise correspondence from unaligned data. At the image level, an auxiliary generator is employed to facilitate the learning of mapping from NIR to VIS domain. At the domain level, we first apply the mutual information constraint to explicitly measure the correlation between domains and thus benefit synthesis. Extensive experiments on three heterogeneous face datasets demonstrate that our approach not only outperforms current state-of-the-art HFR methods but also produce visually appealing results at a high resolution (256×256).

1 Introduction

Different sensors deployed under various circumstances may lead to changes in image illumination, which poses a huge challenge to face recognition systems. For instance, near-infrared images (NIR) produce better performance under low-lighting circumstances thus providing an effective and low-cost solution to applications in large-scale scenarios (e.g. monitoring, surveillance, and security). On the other hand, visible images are easier to access and usually used for system registration and enrollment. The illumination variation makes it difficult to match NIR images and VIS images, which has drawn much attention in computer vision.

Due to the flourishing of deep learning, many heterogeneous face recognition (HFR) methods resort to deep convolutional neural networks to learn domain invariant features [Kristan *et al.*, 2014] or project images from different domains into a common subspace [Peng *et al.*, 2019; He *et al.*, 2018]. Recent work focuses on "recognition via generation", which means synthesizing VIS images from NIR ones for recognition[Lezama *et al.*, 2017; Song *et al.*, 2018]. In this way, there is no need to modify the recognizer for NIR images. And the synthesized VIS images can be further used for suspect testification or other forensic scenarios. Although this thought tries to bridge the sensing gap between NIR and VIS, there still remains some challenges.

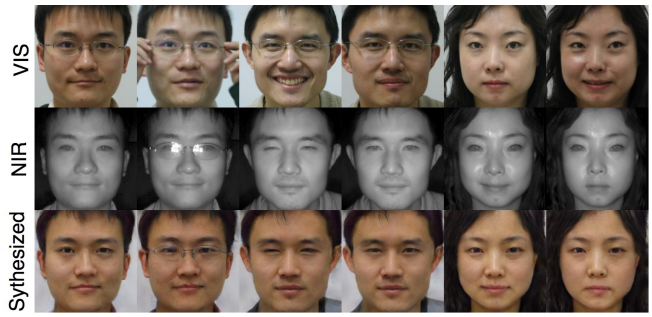

Figure 1: Most NIR and VIS images are unaligned at pixel level. There is large variation in poses and expressions. Data misalignment poses great challenges for cross-spectral face hallucination and heterogeneous face recognition (HFR).

One of the challenges comes from data misalignment ,since it's time-consuming and expensive to obtain well-aligned paired images from different sensors (as shown in Figure 1). Even though CycleGAN [Zhu *et al.*, 2017] is proposed to solve unpaired image-to-image translation with a cycle-consistency loss, it is still inadequate to generate VIS images with fine details at a high resolution. The fact is that the output size of the most existing works is often no larger than 128×128 [Riggan *et al.*, 2016; Zhang *et al.*, 2017]. Another challenge is that by formulating the mapping as an one-to-one problem like Pixel2Pixel [Isola *et al.*, 2017], the misalignment may result in the variance of poses and expressions between the input and output. Besides, in [Song *et al.*, 2018], the cycle-consistency loss is turned out to be insufficient to preserve adequate geometrical and identity information in the synthesized VIS images.

*Corresponding Author

To address these problems, we propose a simple yet effective Pose-preserving Cross-spectral Face Hallucination (PCFH) method based on generative adversarial network (GAN) [Goodfellow *et al.*, 2014]. Given a NIR image, PCFH can synthesize a VIS image while preserving its original identity, poses, and expressions. It employs a multi-path generator with an attention warping module. Global and local paths guarantee fine-grained realistic details. An auxiliary generator is employed to facilitate the learning of mapping from NIR to VIS. The attention warping module first captures the structural difference between NIR and VIS images and achieve pixel-wise alignment by the 2D global inverse distance weighed warping [Ruprecht and Muller, 1995]. This nonlinear transformation alleviates the misalignment of images caused by the sensing gap. Considering that 2D warping introduces local deformation in the facial area, we propose an attention map learning to guide our model to pay more attention to the undeformed area. Therefore, the generator is guided to synthesize visually pleasing VIS images with poses and expressions preserved.

Moreover, inspired by the unsupervised representation learning, we also propose a mutual information constraint (MIC). Similar to [Zhang *et al.*, 2018], we argue that we can decrease the uncertainty about the domain information of a sample by giving more samples in the same domain, which is measured via mutual information mathematically. We maximize the mutual information of the synthetic VIS images and the real VIS images, and minimize that of the synthetic VIS images and the NIR ones. Since it's intractable to compute the mutual information, we resort to [Brakel and Bengio, 2017] and maximize an optimization problem to estimate the mutual information. We trained our network on the challenging CASIA NIR-VIS 2.0 database and verified it on CASIA NIR-VIS 2.0, BUAA and OULU. Extensive experiments show that our PCFH can greatly facilitate heterogeneous face recognition with high-quality generated images (256×256).

In summary, the main contributions of this work are in fourfold:

1. We propose PCFH to solve the data misalignment in heterogeneous face recognition (HFR). An auxiliary generator is employed to facilitate image synthesis and achieve global structural consistency.

2. At pixel level, we propose an attention warping module to derive pixel correspondence between pixel-wise unaligned data. The attention map is learned in an unsupervised way to provide superior supervision.

3. At domain level, we first apply the mutual information constraint to explicitly measure the correlation between domains and thus benefit synthesis.

4. Extensive experiments on three datasets including the challenging CASIA NIR-VIS 2.0, BUAA and OULU show that our method not only outperforms current state-of-the-art HFR methods but also produces visually appealing results at a high resolution(256×256).

2 Related Work

Cross domain image synthesis is to transfer images from one domain into another. Most face recognition algorithms are trained on easily acquired VIS images. Therefore, synthesizing VIS images from NIR is proposed to bridge the gap between different sensors, which is referred to as "recognition via generation". [Wang *et al.*, 2009] proposes a framework for transforming images from one domain to another before heterogeneous face matching. [Tang and Wang, 2003] synthesizes sketches to recognize for face photo retrieval.

Recently, deep learning methods have been booming and have achieved impressive progress in image processing. Based on this progress, [Lezama *et al.*, 2017] cooperates cross-spectral hallucination cooperated with low-rank embedding to maintain identity information. Generative adversarial network(GAN) [Goodfellow *et al.*, 2014] is a recently proposed generative model. It consists of a generator and a discriminator, both are trained adversarially. The generator synthesizes fake data to fool the discriminator, while the latter one discriminates its input as real or fake. Inspired by this, [Zhang *et al.*, 2017; Song *et al.*, 2018] propose two-path GAN to alleviate the lack of NIR images and synthesize facial images with fine details. However, it is still challenging to synthesize heterogeneous face images with high resolution greater than 128×128 with finer details.

Mutual information is widely adopted in unsupervised representation learning. [Hjelm *et al.*, 2019; Brakel and Bengio, 2017] maximize the input and the output of a neural network, in order to learn good representation. Recently, mutual information begins to facilitate the adversarial training. [Zhang *et al.*, 2018] employs mutual information as a measure of informativeness and diversity between the query and the response. In this work, mutual information is added to the objective as a regularizer term. However, mutual information is notoriously intractable to compute. [Zhang *et al.*, 2018] proposes Variational Information Maximization Objective (VIMO) as an alternative to directly compute mutual information. And [Zhang *et al.*, 2018; Brakel and Bengio, 2017; Belghazi *et al.*, 2018] resort to a neural network to estimate the mutual information.

3 Method

3.1 Overview

The aim of PCFH is to synthesize VIS images conditioned on input NIR images and further benefit heterogeneous face recognition (HFR). Generative adversarial network (GAN) [Goodfellow *et al.*, 2014] is proposed as a powerful method for generative task. Using cycle-consistency loss, CycleGAN [Zhu *et al.*, 2017] is proposed to solve the unpaired image-to-image translation. Although Cyclegan is presented as a universal unpaired image-to-image translation method, [Song *et al.*, 2018] showed that it's inappropriate to directly employ CycleGAN for HFR. The reason is that the cycle loss is not strong enough to retain abundant pixel level information as necessary. Despite the lack of explicit pixel-wise correspondence, we can still derive superior supervision from unaligned data in HFR. To this end, we propose Pose-preserving Cross-spectral Face Hallucination (PCFH). The method consists of a multi-path generator G_A to learn a mapping from NIR to VIS domain, an auxiliary generator G_B to learn a mapping in reverse, a discriminator D_A trained with G_A, G_B

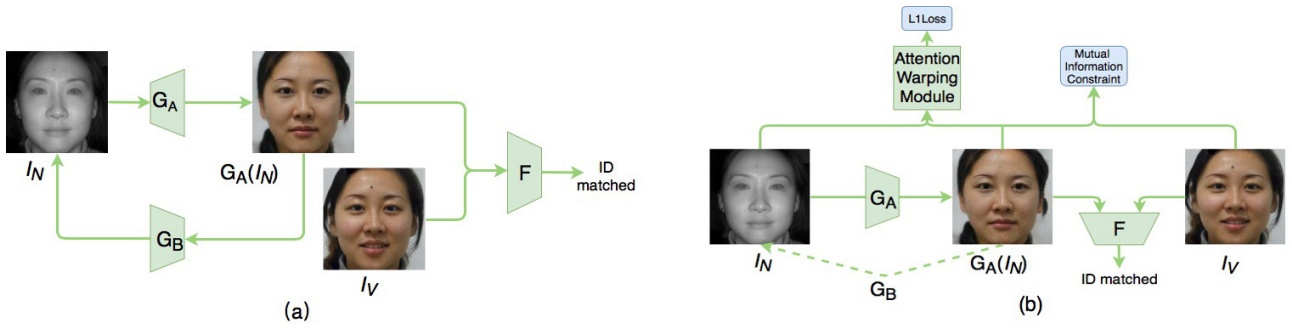

Figure 2: The proposed PCFH method. (a) is the base method. (b) is our PCFH. PCFH takes NIR images as input and outputs high-resolution VIS images with original poses and expressions well preserved. Input image: The CASIA NIR-VIS 2.0 face database.

in adversarial learning, an attention warping module (AWM) to cope with the unpaired data, a mutual information estimator N_E to estimate the MIC, and a face recognizer F. The schematic diagram is shown in Figure 2.

3.2 Adversarial Cross-spectral Hallucination

In the following, we first elaborate the basic version of our method, as presented in Figure 2 (a). Different from CycleGAN, we take G_B as an auxiliary generator which aim to facilitate the training of G_A. Only the translating from NIR to VIS and back to NIR is involved since we only focus on generating VIS images from NIR ones, and it can decrease computational cost. In the training process, G_A endeavors to synthesize VIS and NIR images while D_A discriminates whether the image is real or fake. Similiar to [Huang *et al.*, 2017], to achieve global and local detail synthesis, we add local patches on eyes, noses, and mouths individually in G_A and add them to the global patch in a max-out way. The adversarial loss is:

$$\mathcal{L}_{ad} = -\mathbb{E}_{I_N \sim p_N} D_A(G_A(I_N)) \tag{1}$$

where I_N denotes NIR images.

G_B is an auxiliary generator, since we are not concerned with the generation from VIS to NIR. This generator is to maintain the consistency of the global structure between the synthesized image and the original image by a cycle-consistency loss. It takes the synthesized VIS image as input and the output is supposed to be identical to the original NIR, which is formulated as:

$$\mathcal{L}_{rec} = \lambda_1 \mathbb{E}_{I_N \sim p_N} |G_B(G_A(I_N)) - I_N)|_1 \tag{2}$$

In order to keep identity information during the hallucination, we use a pretrained Lightcnn [Wu *et al.*, 2018] (denoted as F) as a feature extractor. We take the output vector of the second last fully connected layer as the identity feature. inspired by the perceptual loss [Johnson *et al.*, 2016], the identity preserving loss is as followed:

$$\mathcal{L}_{ID} = \lambda_2 \mathbb{E}_{I_N \sim p_N} |F(I_V) - F(G_A(I_N))|_1 \tag{3}$$

We also encode the images into YCbCr space, which is similar to [Song *et al.*, 2018]. The luminance component Y is employed to maintain a global structure consistency. We further find this term can also stabilize the training, which is formulated as:

$$\mathcal{L}_Y = \lambda_3 \mathbb{E}_{I_N \sim p_N} |Y(G_A(I_N)) - Y(I_N)|_1 \tag{4}$$

To sum up, the generator in the basic version receives four types of losses:

$$\mathcal{L}_G = \mathcal{L}_{ad} + \mathcal{L}_{rec} + \mathcal{L}_{ID} + \mathcal{L}_Y \tag{5}$$

3.3 Attention Warping Module

In reality, it's time-consuming and expensive to acquire pixel-wise aligned NIR-VIS dataset. Most NIR and VIS images are not obtained simultaneously, thus leading to pose and expression variance. This variance brings a great challenge to faithfully synthesize pose and expression-preserving VIS images from NIR ones. To utilize pixel correspondence and achieve precise pixel-wise supervision with unaligned data, we present the attention warping module (AWM), which is shown in Figure 3. It should be noted that warping is not straightforwardly inserted in our method. An unsupervised attention learning mechanism is proposed to alleviate the side effect caused by warping.

This module employs a 2D global inverse distance weighed warping [Ruprecht and Muller, 1995]. Due to the misalignment of data, we first warp the VIS images into NIR ones with a 2D global inverse distance weighed warping. It fixes the points by the correspondence of a set of control points and uses a bilinear interpolation to generate warped images. The control points are selected as 68 facial landmarks, which is well studied [Bulat and Tzimiropoulos, 2017]. After the nonlinear transformation, the warped VIS images achieve global structural alignment in the facial area, which is expected to serve as "pixel-aligned ground truth". Therefore, it greatly facilitates cross-spectral face hallucination. This process can be formulated as:

$$I_{Warp} = W(I_N, I_V), \tag{6}$$

where I_{Warp} denotes the warped VIS image, W is the warping module.

However, directly treating the warped VIS image as the "pixel-aligned ground truth" may be inappropriate. Considering the large pose and expression variations in VIS and

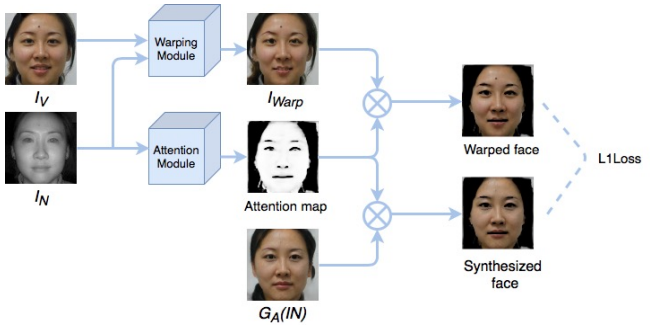

Figure 3: The details of the Attention Warping Module. This module only provides pixel-wise supervision on the "attention areas". Noted that the distortion around eyes and mouths in the warped image are paid "less attention" or even totally covered.

NIR images, there are many artifacts such as the distortion of eyes, mouth, noses, and background in the warped VIS image. These artifacts and distortion can misguide the generator by inaccurate pixel level supervision. To address this problem, inspired by the attention mechanism of human and its application in computer vision[Youssef *et al.*, 2018], we add an attention module to generate attention maps in order to adaptively decide which parts of the warped VIS images can be regarded as the pixel-aligned ground truth while ignoring the deformed region. This module is trained with the generator in an unsupervised way. The loss function is formulated as:

$$\begin{aligned}\mathcal{L}_A =& \alpha_1 \mathbb{E}_{I_N \sim p_N} |A(G_A(I_N)) - I_{Warp}|_1 \\ &+ \alpha_2 rank(Att) + \alpha_3 |Att|_1\end{aligned} \tag{7}$$

where A denotes the attention module, Att is the attention map generated by A, α_1 to α_3 are assigned as 3 ,0.003 and 1×10^{-6} respectively.

Conventionally, attention map is easy to converge to 0 or 1, which means no pixel supervision is introduced or all pixel are used for supervision. [Youssef *et al.*, 2018] use a threshold to control the generation of attention map. However, this value is hard to select. Too high or too low threshold will affect the generation of attention map. We use the loss L_A to prevent threshold selection. The first term in Eq(7) prevent the attention map converging to 0 and make sure every element are close to 1. The second term is a low-rank constraint to ensure learning of structure such as face. The third term prevents all elements from converging to 1. Therefore, our attention map is differential and continuous from 0 to 1, which can effectively provide precise supervision and can be trained in an unsupervised way.

By adding the attention module, we only calculate the pixel-wise loss between the warped VIS images and the generated VIS images on the region of interest, thus preventing the effects of the deformed region. This loss is formulated as:

$$\mathcal{L}_{AWM} = \lambda_4 \mathbb{E}_{I_N \sim p_N} |A(G_A(I_N)) - A(I_{Warp})|_1 \tag{8}$$

3.4 Mutual Information Constraint

Our work is based on GAN, which implicitly matches the distribution of the generated data to real data. Mutual information is a measure of the correlation between two variables. Intuitively, given a variable, mutual information measures to what extend the uncertainty of another variable decreases. In unsupervised representation [Belghazi *et al.*, 2018; Hjelm *et al.*, 2019] maximize the mutual information between the input and the output of an encoder in order to learn a good representation. To achieve informativeness and diversity of the query, [Zhang *et al.*, 2018] resorts to mutual information as a measure of correlation between the query and the response, that is, given a response, the uncertainty of a query decreases. Similarly, we argue that mutual information can be used as a proxy to measure domain correlation between variables. Similarly, we argue that we can decrease the uncertainty about the domain information of a sample by giving another sample in the same domain, which is measured via mutual information mathematically. In the "recognition via generation" scenario, we expect the generated VIS images to reside in the same domain as the real VIS images, and different domain from NIR images, thus bridging the sensing gap between NIR and VIS. Therefore, we propose the mutual information constraint (MIC), which maximizes the mutual information of the synthetic VIS images and the real VIS images and minimizes that of the synthetic VIS images and the NIR images. To the best of our knowledge, this is the the first time that mutual information is used to measure domain correlation, which is written as:

$$\mathcal{L}_{MIC} = \lambda_5 (I(G_A(I_N), I_V) - I(G_A(I_N), I_N)) \tag{9}$$

where $I(a, b)$ denotes the mutual information between a and b. However, it's intractable to compute the mutual information, we resort to [Brakel and Bengio, 2017] which maximizes an optimization problem as an alternative to estimate the mutual information using a neural network.

$$\max \mathbb{E}_{z \sim p} N_E(z) - \mathbb{E}_{z \sim q} N_E(z), \tag{10}$$

where p denotes the joint distribution of variables and q is the product of all marginal distribution. N_E is a neural network collaboratively trained with the generator.

4 Experiment

4.1 Datasets and Protocols

The CASIA NIR-VIS 2.0 face database [Li *et al.*, 2013] contains 725 subjects, and it's the largest and the most challenging NIR-VIS cross spectral face dataset. The images in this dataset vary in poses and expressions. In our experiment, we use two different protocols, the recognition protocol and the generation protocol. We follow the recognition protocol defined in [Huang *et al.*, 2017]. Since there are half the number of NIR images in the training test in every fold, which contains all the identities. For the CASIA NIR-VIS 2.0 database, we define the generation protocol that we only show visual results on the first fold for simplicity, which contains 6010 NIR images and 2547 VIS images of 354 identities. In this case, the testing set consists of over 6000 images that do not appear in the training set. The Rank-1 accuracy and the ROC curve is reported.

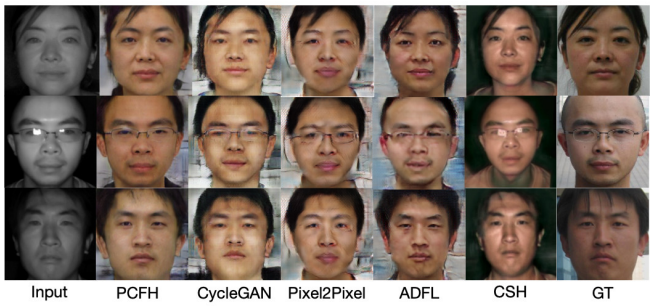

Figure 4: Comparion of visual effects between different methods.

For the BUAA-VisNir database and the Oulu-CASIA NIR-VIS database, the training sets in these two databases are not used. We directly employ our model trained on the first fold of CASIA NIR-VIS 2.0 and test our model on the testing sets in the two databases for evaluation.

The Oulu-CASIA NIR-VIS database [Chen *et al.*, 2009] contains 80 subjects with six typical expressions including anger, disgust, fear, happiness, sadness, surprise and three illuminations including normal indoor, weak light and dark. The Rank-1 accuracy and the ROC curve are reported.

The BUAA-VISNIR face database [Huang *et al.*, 2012] consists of 150 subjects with 13 VIS-NIR pairs and 14 VIS images varying in illumination. The NIR and VIS images of the same identity are acquired simultaneously with different poses and expressions by a multi-spectral camera. The Rank-1 accuracy and the ROC curve are used as evaluation criteria.

4.2 Implementation Details

Our end-to-end network is trained on the CASIA NIR-VIS 2.0 face database on an NVIDIA Titan XP GPU. For 144×144 images, 32×32, 40×32 and 28×52 patches are cropped around two eyes, noses and mouths respectively. 52×52, 60×52 and 36×72 patches are cropped around these areas in 256×256 images. The VIS images are warped according to the corresponding landmarks of NIR and VIS images before fed into our network. The hyperparameters from λ_1 to λ_5 are assigned as 2.5×10^{-5}, 0.002, 0.1, 2.5×10^{-5}, 0.1 respectively.

A pretrained Light CNN-29 which consists of 29 convolutional layers is employed as a feature extractor to calculate identity preserving loss and also the baseline recognizer. We verify the performance of our model in the "recognition via generation" fashion. Following [Yin *et al.*, 2017], we define our distance metric as the average of the original image pair distance and the generated image pair distance. We find that utilizing both the original and the transferred NIR faces effectively improves the recognition performance.

4.3 Experiment Results

In Table 1, we compare the quantitative performance of different generative models on the 1-fold of the challenging CASIA NIR-VIS 2.0 database. It shows that our method outperforms CycleGAN and Pixel2Pixel. It also further improves the performance of Light CNN. The unpaired data makes it difficult to directly utilize the pixel-wise supervision for cross spectral face hallucination, thus leading to the unsatisfying performance of Pix2Pix. Although CycleGAN is proposed to address the unpaired image-to-image translation, it struggles to bridge the gap between the NIR and VIS images due to the large pose and expression variation. In conclusion, our method has a huge improvement over two other generative methods in HFR.

Method	Rank-1	VR@FAR=1%	VR@FAR=0.1%
Pixel2Pixel	22.13	39.22	14.45
CycleGAN	87.23	93.92	79.41
Light CNN	96.84	99.10	94.68
PCFH w/o AWM	96.89	99.07	94.96
PCFH w/o MIC	97.66	99.22	96.49
PCFH	**98.50**	**99.58**	**97.32**

Table 1: The comparison of Rank-1 accuracy (%) and verification rate (%) on the CASIA NIR-VIS 2.0 database. (1-fold)

Method	Rank-1	VR@FAR=1%	VR@FAR=0.1%
VGG	62.1±1.88	71.0±1.25	39.7±2.85
TRIVET	95.7±0.52	98.1±0.31	91.0±1.26
Light CNN	96.7±0.23	98.5±0.64	94.8±0.43
IDR-128	97.3±0.43	98.9±0.29	95.7±0.73
ADFL	98.2±0.34	99.1±0.15	97.2±0.48
PCFH	**98.80±0.26**	**99.57±0.08**	**97.68±0.26**

Table 2: The comparison of Rank-1 accuracy (%) and verification rate (%) on the CASIA NIR-VIS 2.0 database. (10-fold)

Comparison with five deep learning models including TRIVET [Liu *et al.*, 2016], IDR [He *et al.*, 2017], ADFL, VGG [Parkhi *et al.*, 2015], and Light CNN [Wu *et al.*, 2018] further indicate the superiority of our method. In Table 2, the performance of the VIS CNN method such as VGG is unsatisfying due to the large sensing gap between the NIR and VIS images. It can be observed that the proposed method outperforms the other five deep learning models. It is notable that the performance of ADFL is the closest to ours. However, in the testing period, ADFL finetunes the recognizer with the synthesized images while we directly match the generated VIS images with the VIS gallery. We show visual effects of the synthesized images in Figure 4. Considering of different protocols, we only compare 3 subjects in the first fold of CASIA NIR-VIS face database. The results of CycleGAN and Pixel2Pixel are not satisfying. ADFL fails to achieve fine-grained texture , poses and expressions consistency. The results of CSH are acquired in [Lezama *et al.*, 2017]. The background information are lost. The poses and expressions are not consistent with the original NIR ones.

We further evaluate our model on the BUAA NIR-VIS Database. Table 2 shows the comparison between our model with the prior HFR methods including KDSR[Huang *et al.*, 2013], TRIVET, IDR, ADFL and Light CNN. Our method significantly outperforms the existing HFR method in terms of Rank-1 Accuracy and verification rates. Our proposed method improves the best Rank-1 Accuracy and VR@FAR=0.1 from 96.5% to 98.3% and 86.7% to 92.44% respectively. The improvement owes to our mutual infor-

Method	Rank-1	FAR=1%	FAR=0.1%
KDSR	83	86.8	69.5
TRIVET	93.9	93.0	80.9
IDR	94.3	93.4	84.7
ADFL	95.2	95.3	88.0
Light CNN	96.5	95.4	86.7
PCFH w/o AWM	97.4	97.2	89.6
PCFH w/o MIC	98.2	97.6	91.3
PCFH	**98.4**	**97.9**	**92.4**

Table 3: Rank-1 accuracy and verification rate on the BUAA NIR-VIS Database.

Method	Rank-1	FAR=1%	FAR=0.1%
KDSR	66.9	56.1	31.9
TRIVET	92.2	67.9	33.6
IDR	94.3	73.4	46.2
ADFL	95.5	83.0	60.7
Light CNN	96.9	93.7	79.4
PCFH w/o AWM	100.0	96.63	80.22
PCFH w/o MIC	100.0	97.63	85.99
PCFH	**100.0**	**97.7**	**86.6**

Table 4: Rank-1 accuracy and verification rate on the Oulu-CASIA NIR-VIS Database.

mation constraint that explicitly narrows the domain gap between NIR and VIS images. Besides, the attention warping module contributes to the performance by providing pixel-level supervision.

We also compare the proposed method with various HFR methods on the Oulu-CASIA NIR-VIS Database. It is astonishing that 100% Rank-1 Accuracy is achieved by our model, which improves the previous best Rank-1 Accuracy by 3.1%. It indicates that every synthesized VIS images (probe) and the VIS gallery are perfectly matched. To the best of our knowledge, this is the first time that 100% Rank-1 Accuracy is achieved on this dataset, which indicates the superior of the proposed method. The VR@FAR=0.1 is relatively low compared with the performance of the first two databases. The reason is that a number of images in this database are blurred, which brings difficulties to HFR. Besides, some faces in this database are incomplete. However, our method improves VR@FAR=0.1 by 7.2% compared with the previous best performance.

4.4 Ablation Study

In this section, we evaluate the effectiveness of three main components, adversarial cross-spectral face hallucination, attention warping module (AWM) and mutual information constraint (MIC). The first module cannot be removed. Therefore, we derive two variants, PCFH w/o AWM and PCFH w/o MIC of our model, in which AWM and MIC are removed respectively. In Table 1, 3 and 4, we show the quantitative comparison between the two variants and PCFH. PCFH outperforms the other variants on Rank-1 Accuracy and verification rate, which indicates the effectiveness of the two components. It should be noted that AWM contributes more to the quantitative results. Because AWM effectively alleviates the data misalignment by providing pixel-wise supervision and preserves the original poses and expressions, thus facilitating HFR. Figure 5 compares the visual effects between different variants. We also find AWM greatly boosts the visual effects of the synthesized VIS images. However, MIC makes relatively less contribution to the visual effect. Because AWM can explicitly address the data misalignment. Above all, we prove that both AWM and MIC contribute to the performance of our framework.

Figure 5: Ablation study results. We show the visual effects of the same identity with various poses and expressions. PCFH outperforms the other two variants by preserving the original poses and expressions (256×256).

5 Conclusion

In this paper, we have proposed PCFH for heterogeneous face recognition following "recognition via generation". In the method, an auxiliary generator is trained to facilitate cross-spectral face hallucination. We design an attention warping module (AWM) to alleviate the data misalignment caused by the sensing gap. This module introduces pixel-wise supervision and decreases the negative effects caused by the distortion in the warped images. The original poses and expressions are faithfully preserved in the synthesized images. Moreover, a mutual information constraint (MIC) is employed to explicitly guide the synthesis process to preserve the correlation across different domains. An identity-preserving loss is imposed to ensure realistic results with identity information well-retained. Experiments on three datasets show our method achieves great improvements on recognition accuracy and image quality over state-of-the-art HFR methods. In the future, we intend to further explore the synthesis ability of PCFH in data augmentation.

Acknowledgements

This work is partially funded by National Natural Science Foundation of China (Grant No.61622310), Beijing Natural Science Foundation (Grant No.JQ18017), Youth Innovation Promotion Association CAS(Grant No.2015190).

References

[Belghazi *et al.*, 2018] Mohamed Ishmael Belghazi, Aristide Baratin, Saj Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: Mutual information neural estimation. In *NeurIPS*, 2018.

[Brakel and Bengio, 2017] Philemon Brakel and Yoshua Bengio. Learning independent features with adversarial nets for non-linear ica. In *ICML*, 2017.

[Bulat and Tzimiropoulos, 2017] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d and 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017.

[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 *CVPR*, 2009.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[He *et al.*, 2017] Ran He, Xiang Wu, Zhenan Sun, and Tieniu Tan. Learning invariant deep representation for nir-vis face recognition. In *AAAI*, 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 Trans on PAMI*, 2018.

[Hjelm *et al.*, 2019] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019.

[Huang *et al.*, 2012] D Huang, J Sun, and Y Wang. The buaa-visnir face database instructions. *Tech. Rep. IRIP-TR-12-FR-001*, July 2012.

[Huang *et al.*, 2013] X Huang, Z Lei, M Fan, X Wang, and Li S.Z. Regularized discriminative spectral regression method for heterogeneous face matching. *IEEE Trans on IP*, 2013.

[Huang *et al.*, 2017] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *ICCV*, 2017.

[Isola *et al.*, 2017] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[Johnson *et al.*, 2016] Justin Johnson, Alahi Alexandre, and Fei-Fei Li. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[Kristan *et al.*, 2014] P. Gurton Kristan, J. Yuffa Alex, and W. Videen Gorden. Enhanced facial recognition for thermal imagery using polarimetric imaging. *Optics Letters*, 2014.

[Lezama *et al.*, 2017] Jose Lezama, Qiu Qiang, and Sapiro Guillermo. Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding. In *CVPR*, 2017.

[Li *et al.*, 2013] Stan Z. Li, Dong Yi, Zhen Lei, and Shengcai Liao. The casia nir-vis 2.0 face database. In *CVPR Workshops*, 2013.

[Liu *et al.*, 2016] Xiaoxiang Liu, Lingxiao Song, Xiang Wu, and Tieniu Tan. Transferring deep repre- sentation for nir-vis heterogeneous face recognition. In *ICB*, 2016.

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

[Peng *et al.*, 2019] Chunlei Peng, Gao Xinbo, Wang Nannan, and Li Jie. Sparse graphical representation based discriminant analysis for heterogeneous face recognition. *Signal Processing*, 2019.

[Riggan *et al.*, 2016] Benjamin S. Riggan, Nathanial J. Short, Shuowen Hu, and Heesung Kwon. Estimation of visible spectrum faces from polarimetric thermal faces. *ICBTAS*, 2016.

[Ruprecht and Muller, 1995] D. Ruprecht and H. Muller. Image warping with scattered data interpolation. *IEEE CGA*, 1995.

[Song *et al.*, 2018] Lingxiao Song, Zhang Man, Wu Xiang, and He Ran. Adversarial discriminative heterogeneous face recognition. *AAAI*, 2018.

[Tang and Wang, 2003] Xiaoou Tang and Xiaogang Wang. Face sketch synthesis and recognition. In *ICCV*, 2003.

[Wang *et al.*, 2009] Rui Wang, Jimei Yang, Dong Yi, and Stan Z. Li. An analysis-by-synthesis method for heterogeneous face biometrics. In *ICB*, 2009.

[Wu *et al.*, 2018] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *IEEE Trans on IFS*, 2018.

[Yin *et al.*, 2017] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *ICCV*, 2017.

[Youssef *et al.*, 2018] Mejjati Youssef, Alami, Richardt Christian, Tompkin James, Cosker Darren, and Kim Kwang, In. Unsupervised attention-guided image-to-image translation. In *NeurIPS*, 2018.

[Zhang *et al.*, 2017] He Zhang, Vishal M. Patel, Benjamin S. Riggan, and Shuowen Hu. Generative adversarial network-based synthesis of visible faces from polarimetrie thermal faces. *IJCB*, 2017.

[Zhang *et al.*, 2018] Yizhe Zhang, Galley Michel, Gao Jianfeng, Gan Zhe, Li Xiujun, Brockett Chris, and Dolan Bill. Generating informative and diverse conversational responses via adversarial information maximization. In *NeurIPS*, 2018.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.