LAMP-HQ: A Large-Scale Multi-Pose High-Quality Database for NIR-VIS Face Recognition

Aijing Yu$^{1,3}$, Haoxue Wu$^{2,3}$, Huaibo Huang$^{1,3}$, Zhen Lei$^{2,3}$, Ran He$^{1,3}$

$^1$CRIPAC & NLPR, CASIA $^2$CBSR & NLPR, CASIA $^3$University of Chinese Academy of Sciences

{aijing.yu, huaibo.huang}@cripac.ia.ac.cn, wuhaoxue2019@ia.ac.cn, {zlei,rhe}@nlpr.ia.ac.cn

Abstract

Near-infrared-visible (NIR-VIS) heterogeneous face recognition matches NIR to corresponding VIS face images. However, due to the sensing gap, NIR images often lose some identity information so that the recognition issue is more difficult than conventional VIS face recognition. Recently, NIR-VIS heterogeneous face recognition has attracted considerable attention in the computer vision community because of its convenience and adaptability in practical applications. Various deep learning-based methods have been proposed and substantially increased the recognition performance, but the lack of NIR-VIS training samples leads to the difficulty of the model training process. In this paper, we propose a new Large-Scale Multi-Pose High-Quality NIR-VIS database LAMP-HQ containing 56,788 NIR and 16,828 VIS images of 573 subjects with large diversities in pose, illumination, attribute, scene and accessory. We furnish a benchmark along with the protocol for NIR-VIS face recognition via generation on LAMP-HQ, including Pixel2Pixel [9], CycleGAN [31], and ADFL [23]. Furthermore, we propose a novel exemplar-based variational spectral attention network to produce high-fidelity VIS images from NIR data. A spectral conditional attention module is introduced to reduce the domain gap between NIR and VIS data and then improve the performance of NIR-VIS heterogeneous face recognition on various databases including the LAMP-HQ.

1. Introduction

Conventional face recognition under controlled visual light has been one of the most studied directions in the community of computer vision in recent years [15]. It has been applied in various fields, even achieving better performance than humans in most cases. Recently, more attention has been focus on heterogeneous face recognition such as sketch to photo [24], near-infrared to visible [13] and cross resolutions [1]. Due to the insensitivity to illumination [32], near-infrared (NIR) devices are widely used in monitoring and security systems. This leads to the issue of NIR-VIS heterogeneous face recognition (HFR), where the NIR images captured under near-infrared lighting are always matched to the registered VIS images. Because it is harder to match face images across different spectral, massive efforts have been invested in the community to address this issue [16, 6]. With the development of deep neural
network, several deep learning-based methods [16, 19, 26] have been suggested to improve the performance of HFR.

However, challenges still exist in the following three aspects: (1) **Sensing gap.** NIR and VIS face images are captured under different illuminations by different devices, which leads to different textural and geometric appearances between images in probe and gallery. Therefore, it is ineffective to some degree to directly employ traditional face recognition methods in NIR-VIS face recognition [16, 19].

(2) **Pose variations.** Most NIR face images captured by NIR sensors contain various poses with diverse angles, expressions and accessories, leading to incomplete face information. However, VIS face images are often frontal in the corresponding recognition database. Large discrepancies result in obstacles to the process of recognition. (3) **Small-scale dataset.** Due to the widespread use of the Internet, it is relatively easy to obtain a great collection of VIS face images. However, NIR face images are often captured by special NIR sensors so that it is still expensive and time consuming to collect a large-scale NIR-VIS face image database.

In this paper, we propose a LAMP-HQ database to alleviate the issues mentioned above. The comparisons with existing other databases are summed up in Table 1. The main advantages of our new database lie in the following aspects: (1) **Large-scale.** LAMP-HQ contains 56,788 NIR and 16,828 VIS images of 573 subjects with 3 races (containing Asian, white and black), 3 expressions, 5 scenes, 3 angles, broad age distribution (ranging from 6 to 70) and different accessories. (2) **Multi-scenes.** Different from other previous databases, we collect face images under 5 illumination scenes, including indoor natural light, indoor strong light (with fluorescent), indoor dim light (drawing the curtains), outdoor natural light and outdoor backlight. (3) **Multi-poses.** We capture VIS and NIR images with 3 yaw angles (0°, ±45°); on this basis, we capture more images from the side and bottom view, especially to NIR. We also acquire the solution of closing eyes and smiling (including grin and smile). (4) **High-resolutions.** VIS images are captured by the Canon-7D, and NIR images are captured by the AuthenMetric-CE31SE, which leads to high-resolution images (5184*3456 and 960*720, respectively). (5) **Accessory.** We employ 15 types of glasses as accessories to rich the diversity of our database. Headdresses and earrings are also preserved to further increase the complexity of the database. To the best of our knowledge, LAMP-HQ is the largest-scale NIR-VIS database containing various races, ages, angles, expressions, scenes, illuminations and accessories. We will release the new database in the near future. In addition, we provide an effective benchmark on a few state-of-the-art methods, including Pixel2Pixel [9], CycleGAN [31], and ADFL [23].

To further address the challenges in HFR, we propose a novel exemplar-based variational spectral attention network (VSANet) to transfer NIR images to VIS images that are more efficient for recognition. As illustrated in Fig. 3, VSANet contains three sub-modules, i.e., a spectral variational autoencoder (SVAE), a cross-spectral UNet (CSU), and a spectral conditional attention (SCA) module. The spectral variational autoencoder models the spectral style of VIS data using a variational representation that approximately matches a prior distribution. The cross-spectral UNet is utilized to generate high-fidelity VIS images from the input NIR data along with the referenced spectral style. In addition, a spectral conditional attention module is presented to accurately transfer both the global and local spectral information accurately, where the global and local spectral styles are represented by the extracted latent and feature maps of the middle layers in SVAE. For spectral styles that can be sampled from the posterior of VIS data or the prior, the proposed method is capable of producing a VIS image from an NIR input with or without a reference VIS image.
The contributions of our paper are lie in 3-fold:

1) We release a new LAMP-HQ database to add to the research progress of HFR. The database contains 56,788 NIR and 16,828 VIS images of 573 subjects with a wide variety of poses, illumination, attributes, environments and accessories. We believe that the new LAMP-HQ database can significantly advance NIR-VIS face analysis, similar to a lamp lighting up the dark.

2) We provide a comprehensive qualitative and quantitative efficient benchmark of several state-of-the-art methods for NIR-VIS heterogeneous face recognition (HFR), including Pixel2Pixel [9], CycleGAN [31], and ADFL [23]. The performance on the LAMP-HQ reflects its challenges and difficulties.

3) We propose a novel exemplar-based variational spectral attention network (VSANet), including three modules to learn and transfer the spectral style to the generated VIS data from the NIR input. The spectral conditional attention mechanism makes it able to guide the generation of VIS data using both global and local spectral information.

2. Related Work

2.1. NIR-VIS Databases

There are three most commonly used NIR-VIS databases to evaluate the recognition performance in the community. (1) The CASIA NIR-VIS 2.0 Face Database [13] is the largest and most challenging public NIR-VIS database in existence. It consists of 725 subjects, each of which has 1-22 VIS and 5-50 NIR images with large diversities in illumination, expression, distance and pose. There are two views of protocols designed in the database: one is adopted to fine-tune the super-parameter and the other is utilized in the process of training and testing. The protocol in the View 2 includes 10-fold experiments. There are approximately 2,500 VIS and 6,100 NIR images of approximately 360 subjects in the training fold and the other images from the remaining 358 subjects are employed for the testing. (2) The BUAA-VisNir face database [7] is often used to test the performance of domain adaptation [21]. It contains 150 subjects with 9 paired VIS and NIS images synchronously captured by a single multi-spectral camera. All images of every subject are collected under 9 poses and expressions (including neutral-frontal, left-rotation, right-rotation, tilt-up, tilt-down, happiness, anger, sorrow and surprise). The training set includes 900 images of 50 subjects, and the testing set contains the remaining 1,800 images of 100 subjects. (3) The Oulu-CASIA NIR-VIS database [2] includes 80 identities, 30 of them are from the CASIA, and the rest are from the Oulu University. All the NIS and VIS images are collected under three types of illumination environments (normal indoor, weak and dark) and all the subjects are shot with six various expressions (anger, disgust, fear, happiness, sadness and surprise). Based on the protocol designed in [20], 10 subjects of the Oulu University and 30 subjects of the CASIA are singled out to build a database. Eight images from each expression are selected at random so that there are 48 NIR and 48 VIS images.

2.2. Heterogeneous Face Recognition

Heterogeneous face recognition (HFR) problem has attracted increasing attention in the recent years [18, 17, 14, 30]. NIR-VIS face recognition has been one of the most representative and studied issues in the research field. Due to the recent advances in deep learning, recognition performance has achieved considerable progress. HFR methods can be divided into three categories [15, 31]: (1) image synthesis, (2) latent subspace and (3) domain-invariant features. For the image synthesis method, Lezama et al. [12] introduce a cross-spectral hallucination and low-rank embedding to generate heterogeneous images in a patch-based way. Benefiting from the employment of Generative Adversarial Network (GAN) [4], Song et al. [23] utilize a CycleGAN [31] to integrate cross-spectral face hallucination and discriminative feature learning on both raw-pixel space and compact feature space and then improve the performance of HFR. As to latent subspace method, Yi et al. [27] employ Restricted Boltzmann Machines (RBM) to learn a locally shared feature to reduce the heterogeneity around every facial point. He et al. [6] fabricate a hierarchical network to learn both modality-invariant feature subspace and modality-related spectrum subspace. Considering the domain-invariant features method, [19] discusses various metric learning strategies to increase the HFR performance based on the pre-trained VIS CNN. [5] employs a two-level CNN to learn domain-invariant identity representation and modality-related spectrum representation.
3. Database Description

In this section, we introduce the database at length, including the process of data collection and data cleaning. A training and evaluation protocol is defined at the end of this section.

3.1. Data Collection and Cleaning

We use Canon-7D and AuthenMetric-CE31S to acquire VIS and NIR images respectively. We prepare 5 types of illumination scenes during shooting, including indoor natural light, indoor strong light (with fluorescent), indoor dim light (drawing the curtains), outdoor natural light and outdoor back light. The outdoor scenes will be reduced to one type when raining. Each subject has 6 different pictures, containing 3 expressions (neutral, closing eyes and smile or grin), 1 accessory (glasses) and 3 poses with 3 yaw angles (0°, ±45°) in per illumination scene. Some sample images are showed in Fig. 1. In addition, to advance NIR-VIS face recognition in the wild, we design a device to capture more side- and bottom-viewed NIR data, which is closer to real-world application. The detailed collecting device is shown in the Supplemental Materials D. Fig. 2 shows some examples of these views. The photographic distance and height are not strictly regulated to increase the complexity of the database. Finally, we capture 573 × 6 × 5 = 17,190 (participants × attributes × illuminations) VIS and 573 × 6 × 5 × 4 = 68,760 (participants × attributes × illuminations × lens) NIR face images. Then, we manually check all the facial images and remove those blurred or incomplete images. Finally, we preserve 16,828 VIS and 56,788 NIR images of 573 subjects, each subject contains 29 VIS and 99 NIR images on average.

Table 2. The poses, scenes, races, ages, glasses and expressions in LAMP-HQ.

| Poses | NIR: Pitch ∈ [0°, 10°]; Yaw ∈ [−45°, 45°] | VIS: Pitch = 0°; Yaw = −45°, 0°, 45° |
| Scenes | indoor: natural light, strong light, dim light, outdoor: natural light, back light |
| Races | Asian, white, black |
| Ages | 6-70 |
| Glasses Kinds | sunglasses: 2, round glasses: 8, square glasses: 5 |
| Expressions | neutral, closing eyes, smile |

3.2. Protocol

To construct the uniform dataset partition, we provide a fair evaluation protocol for our LAMP-HQ. The subjects in our database are randomly divided into training set with 300 subjects and testing set with 273 subjects.

In the training set, there are 29,525 NIR and 8,798 VIS images, and their paths are listed in ’nir_train.txt’ and ’vis_train.txt’, respectively. In the testing set, VIS images are used for the gallery set (only one frontal neutral image of each subject), while the NIR images, totally 17,163 images, are collected into the probe set. Similarly, ’probe.txt’ and ’gallery.txt’ record their path. All images are saved in jpg format.

The typical usage of the protocol is to train a generative model on the training set. Then, a fixed VIS face recognition model is applied to match VIS images in the gallery set and translated VIS images from NIR images in the probe set. The rank1 recognition rate and verification rate are used to evaluate the final performance.

4. Method

In this section, we propose a variational spectral attention network (VSANet) for NIR-VIS image translation. As illustrated in Fig. 3, VSANet contains three parts: a spectral variational autoencoder (SVAE) to learn variational spectral representation \(z_{vis}\) from the reference VIS image \(x_{vis}\), a cross-spectral UNet (CSU) to translate the input NIR image \(x_{nir}\) to its corresponding VIS version \(y_{vis}\) along with the reference representation \(z_{vis}\), and a spectral conditional attention (SCA) module to guide the combination of the content of \(x_{nir}\) and the spectral style of \(x_{vis}\). The cross-spectral UNet utilizes the same network architecture as the generator in CycleGAN [31]. The spectral variational autoencoder and spectral conditional attention module are elaborated together with the optimization objective functions in the following.

4.1. Spectral Variational Autoencoder

Variational autoencoder (VAE) [11] is one of the most popular generative models that can learn precise manifold representations in an unsupervised way. We design a spectral VAE (SVAE) to learn variational spectral representation \(z_{vis}\) from the reference VIS image \(x_{vis}\). As shown in Fig. 3, SVAE consists of two subnetworks: an inference network \(E\) that maps VIS data \(x_{vis}\) to the latent \(z_{vis}\), which approximates a prior \(p(z_{vis})\), and a generative network \(G\) that samples VIS data \(x_{vis}\) from \(z_{vis}\). The object of SVAE is to maximize the variational lower bound (or evidence lower bound, ELBO) of \(p_{\theta}(x_{vis})\):

\[
\log p_{\theta}(x_{vis})\geq E_{q_{\phi}(z_{vis}|x_{vis})}\log p_{\theta}(x_{vis}|z_{vis}) - D_{KL}(q_{\phi}(z_{vis}|x_{vis})||p(z_{vis})),
\]

where the first term on the right denotes the reconstruction accuracy for the output \(x_{vis}\), and the second regularizes the posterior \(q_{\phi}(z_{vis}|x_{vis})\) to match the prior \(p(z_{vis})\). Optimization such ELBO, the spectral representation \(z_{vis}\) can be sampled from either the posterior \(q_{\phi}(z_{vis}|x_{vis})\) or the
prior \( p(z_{\text{vis}}) \), which leads to the capability of the proposed method to translate NIR data to the VIS domain with or without VIS references.

### 4.2. Spectral Conditional Attention

As illustrated in Fig. 3, a spectral conditional attention (SCA) module is designed to build a bridge between the referenced spectral information learned by SVAE and the NIR-VIS translation flow in CSU. It consists of several multi-scale SCA blocks, each of which aims to produce such features that combine the VIS spectral style and the NIR content information. Then, the fused features are injected into the translation flow at the corresponding scale in CSU.

We employ two types of referenced VIS features to control the spectral style of the output \( y_{\text{vis}} \) both globally and locally. One is the spectral representation \( z_{\text{vis}} \in \mathbb{R}^{C_z} \) sampled from \( q_{\phi}(z_{\text{vis}}|x_{\text{vis}}) \) or \( p(z_{\text{vis}}) \), and the other is the feature \( F^{(j)}_{\text{vis}} \in \mathbb{R}^{C_i \times H_j \times W_j} \) from the \( i \)th layer of \( G \) in SVAE, i.e., \( F^{(j)}_{\text{vis}} = F^{(i)}(z_{\text{vis}}) \), where \( j = 1, \ldots, M \) denotes the index of the SCA block, \( M \) is the number of SCA blocks. Given the content feature \( F^{(j)}_{\text{nir}} \in \mathbb{R}^{C_j \times H_j \times W_j} \) from the \( k \)th layer of CSU, the fused feature \( F^{(j)}_{\text{fuse}} \) can be obtained by the \( j \)th SCA block \( B^{(j)}_{\text{sca}} \),

\[
F^{(j)}_{\text{fuse}} = B^{(j)}_{\text{sca}}(F^{(j)}_{\text{nir}}, F^{(j)}_{\text{vis}}, z_{\text{vis}}).
\] (2)

For \( z_{\text{vis}} \) contains the whole information to generate a VIS image \( G \) and \( F^{(j)}_{\text{vis}} \) contains spatial style features of the size \( H_i \times W_i \), they are expected to guide the global and local spectral styles, respectively.

As shown in Fig. 4, the SCA block can be divided into two parts: the adaptive instance normalization (AdaIN) \([8]\) operation and the conditional attention layer. \( F^{(j)}_{\text{nir}} \) and \( F^{(j)}_{\text{vis}} \) are firstly re-normalized with the spectral \( z_{\text{vis}} \) in AdaIN and then the attention map is computed to model the spatial similarities between the processed spectral feature \( F^{(j)}_{\text{vis}} \) and content feature \( F^{(j)}_{\text{nir}} \) in the conditional attention layer. The AdaIN operation is defined as

\[
F^{(j)}_t = m^{(s)}(z_{\text{vis}}) \frac{F^{(j)}_t - \mu(F^{(j)}_t)}{\sigma(F^{(j)}_t)} + m^{(b)}(z_{\text{vis}}),
\] (3)
where \( t \in \{ \text{vis}, \text{nir} \} \) and \( m \) are an 8-layer MLP with two output vectors \( m^{(s)} \) and \( m^{(b)} \).

Inspired by the self-attention mechanism [29], we design a conditional attention layer to capture the spatial local relationships between the spectral and content features, i.e., \( F'_{\text{vis}}(j) \) and \( F'_{\text{nir}}(j) \). The attention map \( \text{Att}^{(j)} \in \mathbb{R}^{(H_sW_s) \times (H_sW_s)} \) is obtained by

\[
\text{Att}^{(j)} = \text{softmax}(f^T(F'_{\text{vis}}(j))g(F'_{\text{nir}}(j)) ,
\]

where \( f \) and \( g \) are \( 1 \times 1 \) convolutions. The SCA block’s output \( F'_{\text{fuse}}(j) \) is computed by

\[
F'_{\text{fuse}}(j) = F'_{\text{nir}}(j) + \gamma h(F'_{\text{nir}}(j)) \text{Att}^{(j)},
\]

where \( h \) is a \( 1 \times 1 \) convolution and \( \gamma \) is a learned parameter.

In this way, the SCA block can learn the spectral information by considering both the global and local spectral styles contained in \( z_{\text{vis}} \) and \( G \). More details about SCA are given in the Supplemental Materials B.

### 4.3. Loss Functions

In SVAE, the posterior \( q_\phi(z_{\text{vis}}|x_{\text{vis}}) \) is set to be a centered isotropic multivariate Gaussian, i.e., \( q_\phi(z_{\text{vis}}|x_{\text{vis}}) = \mathcal{N}(z_{\text{vis}}; \mu_{\text{vis}}, \sigma^2_{\text{vis}}) \), where \( \mu_{\text{vis}} \) and \( \sigma_{\text{vis}} \) are the output vectors of \( E \). The prior \( p(z_{\text{vis}}) \) is set to be a simple Gaussian, i.e., \( p(z_{\text{vis}}) = \mathcal{N}(0, I) \). \( z_{\text{vis}} \) for \( G \) is sampled from \( \mathcal{N}(z_{\text{vis}}; \mu_{\text{vis}}, \sigma^2_{\text{vis}}) \) using a reparameterization trick, i.e., \( z_{\text{vis}} = \mu_{\text{vis}} + \epsilon \odot \sigma_{\text{vis}} \), where \( \epsilon \sim \mathcal{N}(0, I) \). The negative version of the two terms in Eq. (1) can be defined as

\[
L_{\text{svae}} = \frac{1}{2} \| x_{\text{vis}} - \hat{x}_{\text{vis}} \|^2 + \frac{1}{2} \sum_{i=1}^z (\mu_{\text{vis}}^i)^2 + (\sigma_{\text{vis}}^i)^2 - \log((\mu_{\text{vis}}^i)^2) - 1.
\]

For the optimization of CSU and SCA, we utilize four losses, i.e., the content loss \( L_{\text{content}} \), the style loss \( L_{\text{style}} \), the identity-preserving loss \( L_{\text{id}} \) and the adversarial loss \( L_{\text{adv}} \), to produce high-fidelity VIS facial images. Among them, \( L_{\text{content}} \) and \( L_{\text{style}} \) are employed to combine the content of NIR data and the style of VIS data, respectively; \( L_{\text{id}} \) is used to preserve the identity information and \( L_{\text{adv}} \) is used to improve the image quality of the output image. The total loss is defined as

\[
L_{\text{total}} = L_{\text{content}} + \lambda_1 L_{\text{style}} + \lambda_2 L_{\text{id}} + \lambda_3 L_{\text{adv}},
\]

where \( \lambda_1, \lambda_2 \) and \( \lambda_3 \) are trade-off parameters. Because these losses are used popularly in the earlier works [23, 28], we elaborate their formulations in the Supplemental Materials A due to space limitations.

### 5. Experiments

We evaluate our method qualitatively and quantitatively on the proposed LAMP-HQ database. For qualitative evaluation, we show the results of synthetic VIS images from corresponding input NIR images. For quantitative evaluation, we perform cross-spectral face recognition based on original and synthesized face images. We also provide three HFR benchmarks on LAMP-HQ, including Pixel2Pixel [9], CycleGAN [31] and ADFL [23]. Both LightCNN-9 and LightCNN-29 [25] are employed as face classifiers in the experiment. To further demonstrate the effectiveness of our method and assess the difficulty of LAMP-HQ, we also conduct experiments on CASIA NIR-VIS 2.0 Face Database [13], BUAA-VisNir face database [7] and Oulu-CASIA NIR-VIS database [2], which are widely used in the HFR field.

We crop and then align the facial images of pixel-size \( 256 \times 256 \) on all the above databases. Adam optimizer is employed with a learning rate of 2e-4, the \( \beta_1 \) of 0.5 and the \( \beta_2 \) of 0.99. The batch-size is set to 8 and the model converges in approximately 60,000 iterations. The trade-off parameters \( \lambda_1, \lambda_2 \) and \( \lambda_3 \) in Eq. (7) are set empirically to be 1, 5 and 0.1, respectively.

#### 5.1. Qualitative Evaluation

We conduct two types of NIR-VIS translation experiments on LAMP-HQ. One type produces VIS data by combining the content of NIR data and the spectral style of VIS exemplar, where the spectral representation \( z_{\text{vis}} \) is sampled from the posterior \( q_\phi(z_{\text{vis}}|x_{\text{vis}}) \). It can be observed in...
Figure 6. Qualitative comparison on the LAMP-HQ dataset. From left to right, the columns are the input NIR data, the corresponding VIS data, the results of Pixel2Pixel [9], CycleGAN [31], ADFL [23], our method without SCA, and our method, respectively.

| Method         | LightCNN-9 | LightCNN-29 |
|----------------|------------|-------------|
|                | Rank-1     | FAR=1%      | FAR=0.1%   |
| original       | 89.77      | 88.46       | 71.61      |
| Pixel2Pixel    | 17.04      | 24.45       | 7.95       |
| cycleGAN       | 80.17      | 78.13       | 32.29      |
| ADFL           | 88.09      | 87.67       | 68.06      |
| ours w/o L<sub>content</sub> | 54.21 | 53.53 | 29.05 |
| ours w/o L<sub>style</sub> | 87.98 | 86.38 | 65.42 |
| ours w/o L<sub>id</sub> | 86.84 | 86.97 | 64.91 |
| ours w/o L<sub>adv</sub> | 88.53 | 88.07 | 68.18 |
| ours w/o AdaIN | 93.67 | 91.22 | 72.89 |
| ours w/o SCA  | 92.74 | 90.41 | 72.50 |
| ours           | 94.09 | 91.81 | 74.77 |

Table 3. NIR-VIS face recognition on LAMP-HQ.

The other type of NIR-VIS translation is generating VIS data from NIR data without VIS exemplar, where \( z_{vis} \) is sampled from the prior \( p(z_{vis}) = \mathcal{N}(0, I) \). We compare the performance with other SOTA methods in this case. As demonstrated in Fig. 6, the proposed method significantly...
outperforms other methods. Pixel2Pixel fails to generate visual-realistic images due to the lack of paired NIR-VIS data. CycleGAN produces artifacts in some hard cases, such as closing eyes or large poses. ADFL cannot reconstruct realistic textures in the background. The ablation version of our method without spectral conditional attention also fails to generate realistic colors and textures for the background. This demonstrates that spectral conditional attention mechanism is helpful for generating realistic local texture details. The comparison results verify the effectiveness of the proposed method for cross-spectral facial hallucination, even in extreme conditions.

5.2. NIR-VIS Face Recognition

Following the protocol described in Section 3, we conduct quantitative comparison experiments on LAMP-HQ with several deep learning methods, including Pixel2Pixel, CycleGAN and ADFL. The pre-trained LightCNN-9 and LightCNN-29 are utilized to compute three metrics i.e., Rank-1 accuracy, and verification rates when VR@FAR=1%, 0.1%, for evaluation. The results are reported in Table 3. As shown in Table 3, the proposed method achieves the best performance of NIR-VIS face recognition. Compared to the original NIR data in the probe, the synthesized VIS data can improve Rank-1 accuracy from 89.77% to 94.09% by LightCNN-9 and from 94.94% to 97.46% by LightCNN-29. This verifies that the proposed method is effective in improving the performance of NIR-VIS face recognition.

5.3. Evaluation on Other Databases

In this section, we report the evaluation results on the CASIA 2.0, Oulu and BUAA databases. For the CASIA 2.0 database, we use the standard protocol in View 1 for evaluation. For the BUAA and Oulu databases, our model trained on LAMP-HQ is directly used to evaluate the testing sets of BUAA and Oulu following the standard protocol. We compare the recently proposed H2(LBP3) [20], TRIVET [14], IDR [5], Pixel2Pixel [9], CycleGAN [31], and ADFL [23]. The quantitative results of face recognition computed by LightCNN-29 are reported in Table 4. It can be observed that the proposed method outperforms others on all three databases. Note that the results of our method on Oulu and BUAA are obtained by the model trained on LAMP-HQ, which verifies the generalization of the proposed method.

Table 4. Quantitative comparisons on CASIA 2.0, BUAA, and Oulu NIR-VIS databases. The results of the compared methods are copied from the published papers.

| Method | CASIA 2.0 | BUAA | Oulu |
|--------|-----------|------|------|
|        | Rank-1 | FAR=1% | FAR=0.1% | Rank-1 | FAR=1% | FAR=0.1% | Rank-1 | FAR=1% | FAR=0.1% |
| original | 96.84 | 99.10 | 94.68 | 96.5 | 95.4 | 86.7 | 96.7 | 92.4 | 65.1 |
| H2(LBP3) | 43.8 | 36.5 | 10.1 | 88.8 | 88.8 | 73.4 | 70.8 | 62.0 | 33.6 |
| TRIVET | 95.7 ± 0.52 | 98.1 ± 0.31 | 91.0 ± 1.26 | 93.9 | 93.0 | 80.9 | 92.2 | 67.9 | 33.6 |
| IDR | 97.3 ± 0.43 | 98.9 ± 0.29 | 95.7 ± 0.73 | 94.3 | 93.4 | 84.7 | 94.3 | 73.4 | 46.2 |
| Pixel2Pixel | 22.13 | 39.22 | 14.45 | / | / | / | / | / | / |
| cycleGAN | 87.23 | 93.92 | 79.41 | / | / | / | / | / | / |
| ADFL | 98.2 ± 0.34 | 99.1 ± 0.15 | 97.2 ± 0.48 | 95.2 | 95.3 | 88.0 | 95.5 | 83.0 | 60.7 |
| ours | 99.0 | 99.9 | 98.3 | 98.0 | 98.2 | 92.5 | 99.9 | 96.8 | 82.3 |

6. Conclusion

This paper proposes a new large-scale multi-pose high-quality database for NIR-VIS heterogeneous face recognition. To the best of our knowledge, LAMP-HQ is the largest NIR-VIS database containing different illuminations, scenes, expressions, poses, and accessories. We also provide an efficient benchmark for NIR-VIS face recognition on LAMP-HQ, including Pixel2Pixel, CycleGAN, and ADFL. In addition, we propose a novel exemplar-based variational spectral attention network (VSANet) to combine the learned spectral information of referenced VIS images and the content information of input NIR images. In this way, a photo-realistic image can be generated that is helpful for cross-spectral face recognition. We hope that our LAMP-HQ database and the benchmark could make for the development of NIR-VIS face recognition.
References

[1] Soma Biswas, Kevin W Bowyer, and Patrick J Flynn. Multidimensional scaling for matching low-resolution face images. *IEEE transactions on pattern analysis and machine intelligence*, 34(10):2019–2030, 2011.

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

[3] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[5] Ran He, Xiang Wu, Zhenan Sun, and Tieniu Tan. Learning invariant deep representation for nir-vis face recognition. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[6] 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*, 41(7):1761–1773, 2018.

[7] Di Huang, Jia Sun, and Yunhong Wang. The buaa-visnir face database instructions. *School Comput. Sci. Eng., Beihang Univ., Beijing, China, Tech. Rep. IRIP-TR-12-FR-001*, 2012.

[8] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *International Conference on Computer Vision*, pages 1501–1510, 2017.

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[10] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[11] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

[12] José Lezama, Qiang Qiu, and Guillermo Sapiro. Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6628–6637, 2017.

[13] 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.

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

[15] Shuxin Ouyang, Timothy Hospedales, Yi-Zhe Song, Xuening Li, Chen Change Loy, and Xiaogang Wang. A survey on heterogeneous face recognition: Sketch, infra-red, 3d and low-resolution. *Image and Vision Computing*, 56:28–48, 2016.

[16] Christopher Reale, Nasser M Nasrabadi, Heesung Kwon, and Rama Chellappa. Seeing the forest from the trees: A holistic approach to near-infrared heterogeneous face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 54–62, 2016.

[17] Benjamin S Riggan, Nathaniel J Short, Shuowen Hu, and Heesung Kwon. Estimation of visible spectrum faces from polarimetric thermal faces. In *2016 IEEE 8th international conference on biometrics theory, applications and systems (BTAS)*, pages 1–7. IEEE, 2016.

[18] M Saquib Sarfraz and Rainer Stiefelhagen. Deep perceptual mapping for thermal to visible face recognition. *arXiv preprint arXiv:1507.02879*, 2015.

[19] Shreyas Saxena and Jakob Verbeek. Heterogeneous face recognition with cnns. In *European conference on computer vision*, pages 483–491. Springer, 2016.

[20] 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, 2016.

[21] Ming Shao, Dmitry Kit, and Yun Fu. Generalized transfer subspace learning through low-rank constraint. *International Journal of Computer Vision*, 109(1-2):74–93, 2014.

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[23] Lingxiao Song, Man Zhang, Xiang Wu, and Ran He. Adversarial discriminative heterogeneous face recognition. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[24] Xiaoou Tang and Xiaoang Wang. Face photo recognition using sketch. In *Proceedings. International Conference on Image Processing*, volume 1, pages I–I. IEEE, 2002.

[25] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face recognition with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018.

[26] Xiang Wu, HuaiHo Huang, Vishal M Patel, Ran He, and Zhenan Sun. Disentangled variational representation for heterogeneous face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9005–9012, 2019.

[27] Dong Yi, Zhen Lei, and Stan Z Li. Shared representation learning for heterogenous face recognition. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, volume 1, pages 1–7. IEEE, 2015.
[28] Junchi Yu, Jie Cao, Yi Li, Xiaofei Jia, and Ran He. Pose-preserving cross-spectral face hallucination. In Proceedings of the 28th International Joint Conference on Artificial Intelligence, pages 1018–1024. AAAI Press, 2019.

[29] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In International Conference on Machine Learning, pages 7354–7363, 2019.

[30] Mingjin Zhang, Ruxin Wang, Xinbo Gao, Jie Li, and Dacheng Tao. Dual-transfer face sketch–photo synthesis. IEEE Transactions on Image Processing, 28(2):642–657, 2018.

[31] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision, pages 2223–2232, 2017.

[32] Jun-Yong Zhu, Wei-Shi Zheng, Jian-Huang Lai, and Stan Z Li. Matching nir face to vis face using transduction. IEEE Transactions on Information Forensics and Security, 9(3):501–514, 2014.
A. Loss Functions

As shown in Eq. (7) in the submitted manuscript, four losses are used to optimize the cross-spectral U-Net (CSU) and spectral condition attention (SCA) module. They are the content loss \( L_{content} \), the style loss \( L_{style} \), the identity-preserving loss \( L_{id} \), and the adversarial loss \( L_{adv} \), the details of which are given below.

**Content Loss.** Similar to [10], we employ the VGG-16 network [22] to compute the content loss to restrict the content similarity between the input \( x_{nir} \) and the output \( y_{vis} \). The content loss is obtained by computing the Euclidean distance between the features of \( x_{vis} \) and \( y_{vis} \) extracted by VGG-16. It can be formulated as follows:

\[
L_{content} = \frac{1}{C_jH_jW_j} \| U_j(y_{vis}) - U_j(x_{nir}) \|_2^2 \tag{1}
\]

where \( U_j(y_{vis}) \) and \( U_j(x_{nir}) \) are the output features of the \( j \)th layer of VGG-16, respectively. We use the relu3_3 layer to extract the features for the content loss.

**Style Loss.** In order to preserve the spectral information learned from \( x_{vis} \), in the process of generating the output \( y_{vis} \), we use the style loss proposed in [3]. It is obtained by computing the distance between the Gram matrices of \( x_{vis} \) and \( y_{vis} \), which is defined as follows:

\[
L_{style} = \sum_j \| G_j^U(x_{vis}) - G_j^U(y_{nir}) \|_{F'}^2 \tag{2}
\]

where

\[
G_j^U(x)_{c,c'} = \frac{1}{C_jH_jW_j} \sum_{h=1}^{H_j} \sum_{w=1}^{W_j} U_j(x)_{h,w,c} U_j(x)_{h,w,c'} \tag{3}
\]

\( U_j(x) \) is the output feature of the \( j \)th layer of VGG-16. We use the relu1_2, relu2_2, relu3_3 and relu4_3 layers to extract the features for the style loss.

**Identity-preserving Loss.** As discussed in the manuscript, LAMP-HQ contains facial images of different poses for each subject. In the training phase, we introduce a frontal VIS image \( x_{vis}^{(match)} \) that has the same identity with the input NIR \( x_{nir} \). The produced image \( y_{vis} \) is expected to have the most identity characteristics with \( x_{vis}^{(match)} \). For this purpose, an identity-preserving loss is designed in a similar way with the content loss, which is formulated as:

\[
L_{id} = \| F(x_{vis}^{(match)}) - F(y_{vis}) \|_1 \tag{4}
\]

where \( F(x_{vis}^{(match)}) \) and \( F(y_{vis}) \) are the extracted features respectively for \( x_{vis}^{(match)} \) and \( y_{vis} \) by a face recognition network, i.e., a pretrained LightCNN in this paper. We utilize the feature before the last FC layer of LightCNN for the identity-preserving loss.

**Adversarial Loss.** Generative Adversarial Network (GAN) [4] is one of the most popular generative models nowadays. It can produce photo-realistic images via playing a min-max game between a generator network and a discriminator network. We utilize GAN to improve the visual quality of the produced VIS image \( y_{vis} \), where the proposed CSU serves as the generator \( G \) of GAN. The two-player minimax game is,

\[
\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{vis}(x)} [\log D(x)] + \mathbb{E}_{y_{vis} \sim p_g(y_{vis})} [\log (1 - D(y_{vis}))] \tag{5}
\]

where \( p_{vis}(x) \) represents the real distribution of VIS data and \( p_g(y_{vis}) \) represents the distribution of the generated VIS data. Then the adversarial loss can be formulated as,

\[
L_{adv} = - \log(D(y_{vis})) \tag{6}
\]

B. Network Architecture

Table 1 and Table 2 are the network architectures of the inference network \( E \) and the generator \( G \) in spectral VAE (SVAE), respectively. Given a VIS image of \( 3 \times 256 \times 256 \), \( E \) encodes it into two 512-d vectors, i.e., \( \mu \) and \( \sigma \), which forms the posterior \( q_\phi(z_{vis}|x_{vis}) = N(z; \mu, \sigma^2) \). The generator \( G \) produces the reconstruction image \( \hat{x}_{vis} \) from the spectral latent \( z_{vis} \), where \( z_{vis} \sim q_\phi(z_{vis}|x_{vis}) \). Noted that we use \( z_{vis} \) with two middle feature maps, i.e., \( F_{vis}^{(3)} \) and \( F_{vis}^{(4)} \) in Table 2, to guide the NIR-VIS translation.

Table 3 reports the network architecture of the cross-spectral U-Net (CSU) injected with the spectral condition attention (SCA) blocks. CSU produces VIS facial image \( y_{vis} \) from the input NIR image \( x_{vis} \) together with the spectral guidance \( z_{vis} \), \( F_{vis}^{(3)} \) and \( F_{vis}^{(4)} \). As shown in Tabel 3, we use four SCA blocks to fuse the exemplar spectral information with the features learned by CSU. The fused features are employed as the input for the next layer in CSU.

Table 4 reports the network architecture of the discriminator \( D \) used in the computation of the adversarial loss. The discriminator \( D \) and and the CSU network along with the SCA module are optimized iteratively following the original GAN [4].

C. Visual Results for Ablation Study

We report the visual results of the ablation study in Fig. 1. It can be observed that the content loss is the most important factor in producing visual realistic images. The identity-preserving loss has little influence on the visual quality, but it may hamper the recognition performance. Without the style loss, the generated results cannot capture the spectral information of VIS data. The adversarial loss also influences the spectral style, and the generations
lack colorization diversity without it. The SCA module and the AdaIN operation play important roles in synthesizing photo-realistic images both for the foreground and background. The appealing performance of the full version of our method demonstrates that each part, including the losses and the designed attention modules, is beneficial in producing realistic VIS images from NIR data.

D. More Examples in LAMP-HQ

Fig.2 shows the capture device of NIR images with various angles. More examples are provided in Fig.3, Fig.4, Fig.5, Fig.6 and Fig.7 to further demonstrate the diversity of LAMP-HQ in races, ages, poses, glasses, and illuminations.
| Input | Layer | Norm | Act     | Output | Output Size |
|-------|-------|------|---------|--------|-------------|
| z<sub>vis</sub> | FC    | /    | /       | F<sup>(1)</sup> | 512×8×8  |
| F<sub>vis</sub><sup>(1)</sup> | Upsample | /    | /       | F<sup>(1)</sup> | 512×16×16 |
| F<sub>vis</sub><sup>(1)</sup> | Conv3 | IN   | LeakyReLU | F<sup>(2)</sup> | 512×16×16 |
| F<sub>vis</sub><sup>(2)</sup> | Upsample | /    | /       | F<sup>(2)</sup> | 512×32×32 |
| F<sub>vis</sub><sup>(2)</sup> | Conv3 | IN   | LeakyReLU | F<sup>(3)</sup> | 256×32×32 |
| F<sub>vis</sub><sup>(3)</sup> | Upsample | /    | /       | F<sup>(3)</sup> | 256×64×64 |
| F<sub>vis</sub><sup>(3)</sup> | Conv3 | IN   | LeakyReLU | F<sup>(4)</sup> | 128×64×64 |
| F<sub>vis</sub><sup>(4)</sup> | Upsample | /    | /       | F<sup>(4)</sup> | 128×128×128 |
| F<sub>vis</sub><sup>(4)</sup> | Conv3 | IN   | LeakyReLU | F<sup>(5)</sup> | 64×128×128 |
| F<sub>vis</sub><sup>(5)</sup> | Upsample | /    | /       | F<sup>(5)</sup> | 64×256×256 |
| F<sub>vis</sub><sup>(5)</sup> | Conv3 | IN   | LeakyReLU | F<sup>(6)</sup> | 32×256×256 |
| F<sub>vis</sub><sup>(6)</sup> | Conv3 | IN   | LeakyReLU | F<sup>(6)</sup> | 3×256×256 |
| F<sub>vis</sub><sup>(6)</sup> | /     | /    | sigmoid | X<sub>vis</sub> | 3×256×256 |

Table 2. Structure of the generator network G of SVAE. Conv3 denotes a convolution layer of kernel-size 3 × 3, stride 1 and padding 1. The input 512-d vector z<sub>vis</sub> and two middle feature maps, i.e., F<sup>(3)</sup><sub>vis</sub> of 256 × 32 × 32 size and F<sup>(4)</sup><sub>vis</sub> of 128 × 64 × 64 size, are used as the inputs for SCAs.

Figure 2. The capture device of NIR images, which includes four NIR shots.
Table 3. Structure of CSU injected with SCA blocks. Conv3 denotes a convolution layer of kernel-size 3 × 3, stride 1 and padding 1. The red color emphasizes the connections between CSU and SCAs.
| Input       | Layer | Norm | Act       | Output | Output Size   |
|-------------|-------|------|-----------|--------|---------------|
| $X_{vis}/Y_{vis}$ | Conv4 | /    | LeakyReLU | X0     | $64 \times 128 \times 128$ |
| X0          | Conv4 | IN   | LeakyReLU | X0     | $128 \times 64 \times 64$ |
| X0          | Conv4 | IN   | LeakyReLU | X0     | $256 \times 32 \times 32$ |
| X0          | Conv4 | /    | /         | out    | $1 \times 16 \times 16$   |

Table 4. Structure of the discriminator. Conv4 denotes a convolution layer of kernel-size $4 \times 4$, stride 2 and padding 0.

Figure 3. Examples of black, white and Asian facial images in LAMP-HQ.
Figure 4. Examples of age diversity in LAMP-HQ.
Figure 5. Examples of pose diversity in LAMP-HQ.
Figure 6. Examples of glasses diversity in LAMP-HQ.
Figure 7. Examples of illumination and scene diversity in LAMP-HQ.