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

Aijing Yu1,3,4 · Haoxue Wu2,3,4 · Huaibo Huang1,3,4 · Zhen Lei2,3,4 · Ran He1,3,4

Received: 15 March 2020 / Accepted: 5 January 2021 / Published online: 12 February 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC part of Springer Nature 2021

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 NIR-VIS 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 Pixel2-Pixel, CycleGAN, ADFL, PCFH, and PACH. 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.

Keywords Heterogeneous face recognition · Near infrared-visible matching · Database · Variational spectral attention · Spectral conditional attention

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 (Ouyang et al. 2016). It has been applied in various fields, even achieving better performance than humans in most cases. Recently, more attention has been focus on heterogeneous recognition issue, such as sketch to photo (Tang and Wang 2002; Wang et al. 2014), near-infrared to visible (Li et al. 2013; Xiao et al. 2013), polarimetric thermal to visible (Di et al. 2018), and cross resolutions (Biswas et al. 2011). Due to the insensitivity to illumination (Zhu et al. 2014), 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

Communicated by Vishal Patel.

Ran He
rhe@nlpr.ia.ac.cn
Aijing Yu
aijing.yu@cripac.ia.ac.cn
Haoxue Wu
wuhaoxue2019@ia.ac.cn
Huaibo Huang
huaibo.huang@cripac.ia.ac.cn
Zhen Lei
zlei@nlpr.ia.ac.cn

1 Center for Research on Intelligent Perception and Computing, CASIA, Beijing, China
2 Center for Biometrics and Security Research, CASIA, Beijing, China
3 National Laboratory of Pattern Recognition, CASIA, Beijing, China
4 School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
spectral, massive efforts have been invested in the community to address this issue (Reale et al. 2016; He et al. 2018). With the development of deep neural network, several deep learning-based methods (Reale et al. 2016; Saxena and Verbeek 2016; Wu et al. 2019) 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 various 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 (Reale et al. 2016; Saxena and Verbeek 2016). (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 faces are often frontal in the corresponding recognition database. Large discrepancies result in obstacles to the process of recognition (Li et al. 2006a; b). (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, Caucasian and African), 3 expressions, 5 scenes, 3 angles, broad age distribution (ranging from 6 to 70) and different accessories. (2) Multi-scenes. Different from 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 Canon-7D, and NIR images are captured by 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 (Isola et al. 2017), CycleGAN (Zhu et al. 2017), ADFL (Song et al. 2018), PCFH (Yu et al. 2019), and PACH (Duan et al. 2020).

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. Generally speaking, the task of cross spectral synthesis is an ill-posed problem, meaning that there exist many VIS solutions for one NIR image. This fact makes it difficult to produce a high-fidelity VIS image directly from a single NIR input. To address this issue, we propose VSANet to reduce the ambiguity of NIR-VIS hallucination by using a VIS exemplar to guide the synthesis of spectral style in the VIS outputs. As illustrated in Fig. 7, 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 VIS exemplar to guide the synthesis of spectral style in the VIS outputs. As illustrated in Fig. 7, 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 addi-
| Database                          | No. of image | Thermal image size | VIS image size | Subjects | Year |
|-----------------------------------|--------------|--------------------|----------------|----------|------|
| NIR-VIS 2.0                       | 17,580       | $640 \times 480$  | $640 \times 480$ | 725      | 2013 |
| BUAA                              | 2700         | –                  | –              | 150      | 2012 |
| Oulu                              | 7680         | –                  | –              | 80       | 2009 |
| ND-NIVL                           | 24,605       | $720 \times 480$  | $721 \times 480$ | 574      | 2015 |
| LDHF-DB                           | 800          | $5184 \times 3456$| $5184 \times 3456$ | 100      | 2012 |
| LAMP-HQ                           | 73,616       | $960 \times 720$  | $5184 \times 3456$ | 573      | 2020 |
| ARL (volume 1)                    | 960          | $640 \times 480$  | $659 \times 494$ | 60       | 2016 |
| ARL (volume 2)                    | 1766         | $640 \times 480$  | $659 \times 494$ | 111      | 2019 |
| ARL (volume 3)                    | 1936         | $640 \times 480$  | $659 \times 494$ | 121      | 2020 |
| Visible and thermal paired face database | 4200     | $160 \times 120$  | $1920 \times 1080$ | 50       | 2018 |

Table 1 Comparisons with existing heterogeneous facial databases

There are three most commonly used NIR-VIS databases to evaluate the recognition performance in the community. The CASIA NIR-VIS 2.0 Face Database (Li et al. 2013) 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. Each image is randomly captured, so the NIR and VIS images of one person are unpaired. 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 2500 VIS and 6100 NIR images of approximately 360 subjects in the training fold and the other images from the remaining 358 subjects are employed for the testing.

There are three most commonly used NIR-VIS and Thermal-VIS databases.

The contributions of our paper 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 (Isola et al. 2017), CycleGAN (Zhu et al. 2017), and ADFL (Song et al. 2018), PCFH (Yu et al. 2019), and PACH (Duan et al. 2020). 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 Heterogeneous Face Databases

Generally speaking, heterogeneous face databases include sketch-photo, NIR-VIS, Thermal-VIS, et al. In this subsection, we mainly introduce some commonly used NIR-VIS and Thermal-VIS databases.
are selected at random so that there are 48 NIR and 48 VIS images. In addition to the commonly used databases mentioned above, there are some other databases for reference. The Near Infrared Visible Light Database (ND-NIVL) (Bernhard et al. 2015) contains a total of 2341 VIS images and 22,264 NIR images from 574 subjects. 1 VIS image and around 10 NIR images are captured per session per subject. All the images are acquired during both the fall and spring semesters. Long Distance Heterogeneous Face Database (LDHF-DB) (Maeng et al. 2012) collects indoor heterogeneous images at distance of 1 meter and outdoor heterogeneous images at distances of 60 meters, 100 meters, and 150 meters. A total of 100 subjects (70 males and 30 females) participate in the collection process. The CASIA HFB Face Database (Li et al. 2009) consists of 100 subjects (57 males and 43 females). For each subject, there are 4 VIS, 4 NIR, and 1 to 2 3D face images. In 2010, the CASIA HFB database was expanded to 5,097 images from 202 subjects. Different from the Near-infrared collection devices which generate infrared light actively, thermal imaging systems detect the heat energy from the skin. The Thermal and Visible Paired Face Database (Mallat and Dugelay 2018) contains 2100 paired thermal-visible images from 50 subjects with different face and environment variations. Each subject participates in two acquisition sessions whose time interval is 3 to 4 months. In addition to conventional thermal-visible face databases, a polarimetric thermal database named ARL Database (Hu et al. 2016) is proposed to preserve more geometric and textural facial details (Gurton et al. 2014). The first version of the ARL Database contains 60 consented subjects in total, 31 of which were collected in the fall of 2014 and others were collected in the spring of 2016. The capture devices are set at distances of 2.5 m, 5 m, and 7.5 m. At each range, two 10 second videos are collected from the participants, one is in the natural state, the other is counting loud numerically from one upwards. In the work of Zhang et al. (2019b), the ARL database is expanded to 111 subjects at distances of only 2.5 m. In the work of Di et al. (2020), the ARL database is expanded to 121 subjects with more off-pose images (excludes extreme poses). Therefore, three types of protocols are presented for later research. The first protocol is evaluated on the first version named Volume 1, in which 240 sample pairs from 30 subjects are for training and the other 240 sample pairs are for testing. In the second protocol, 680 sample pairs of 85 subjects are used for training and 208 sample pairs of 26 subjects are employed for testing. All the results are evaluated on five random splits in the above two protocols. For the third protocol, Images of 96 subjects are employed as the training set and others are for the testing. All the results are evaluated on five random splits in three protocols.

2.2 Heterogeneous Face Recognition

Heterogeneous face recognition (HFR) problem has attracted increasing attention in the recent years (Sarfraz and Stiefelhagen 2015; Riggan et al. 2016; Liu et al. 2016; Zhang et al. 2018; Di et al. 2019). NIR-VIS face recognition has been one of the most representative and studied issues in the research field. In this section, we review various recent advances of HFR from three aspects (Ouyang et al. 2016; Zhu et al. 2017): image synthesis, latent subspace, and domain-invariant features.

Image synthesis methods synthesize images from one domain to another and then match the heterogeneous images in the same spectral domain. Method in (Tang and Wang 2003) firstly attempts to synthesize a sketch photo from a visual face image. Markov random fields are adopted in (Wang and Tang 2008), which could transfer the sketch to face photo in a multi-scale way. Lei et al. (2008) utilize canonical correlation analysis (CCA) to synthesize 3D face from the NIR one. The work of Wang et al. (2009) designs an analysis-by-synthesis framework for heterogeneous face mapping. The dictionary learning is employed in (Wang et al. 2012; Huang et al. 2013; Juefei-Xu et al. 2015) during the process of reconstruction. Zhang et al. (2018) propose a dual-transfer synthesis framework which divide the transfer process into the intra- and inter-domain.

Benefiting from the essential advanced in deep learning, recent there are various impressive works based on deep networks. A VIS feature estimation and VIS image reconstruction two-step procedure is introduced in (Riggan et al. 2016) to transfer polarimetric thermal images to visual domain. Lezama et al. (2017) introduce a cross-spectral hallucination and low-rank embedding to generate heterogeneous images in a patch-based way. Method in (Di et al. 2018) employs the attributes extracted from the visible image in the synthesis progress. The work of Huang et al. (2017) utilize a global and local perception GAN (Goodfellow et al. 2014) to deal with the face rotation issue. Song et al. (2018) utilize a Cycle-GAN (Zhu et al. 2017) 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. Zhang et al. (2019b) employ the GAN model to synthesize visual images from polarimetric thermal domain. A self-attention mechanism is used in (Di et al. 2019) to guide the generating of visual images.

Latent subspace methods aim to map images of two different domains into a common latent, where the features of heterogeneous images can be matched. Lin and Tang (2006) propose the Common Discriminant Feature Extraction (CDFE) to extract the discriminant and locality feature. The coupled discriminant analysis is proposed in (Lei et al. 2012) followed by (Huang et al. 2012), which brings up a regularized discriminative spectral regression method. Wang
et al. (2015) propose a baseline method of NIR-VIS HFR by adopting the feature selection. Yi et al. (2015) employ Restricted Boltzmann Machines to learn a locally shared feature to reduce the heterogeneity around every facial point. Kan et al. (2015) propose the multi-view discriminant analysis to reduce the domain discrepancy. In order to preserve the common information of images in different domains, the work of Li et al. (2016) adds the mutual component analysis to the process of mapping. Jin et al. (2017) use the extreme learning machine (ELM) integrated with the multi-task clustering for cross-spectral feature learning. He et al. (2018) fabricate a hierarchical network to learn both modality-invariant feature subspace and modality-related spectrum subspace. An orthogonal dictionary alignment is proposed in (Mudunuri et al. 2018) to deal with the pose and low-resolution issues of NIR images.

**domain-invariant feature** methods extract the features related to common identity information of heterogeneous face images. In traditional methods, hand-crafted feature descriptors are utilized, i.e., Local Binary Patterns (LBP), (Scale-invariant Feature Transform) SIFT, Histograms of Oriented Gradients (HOG), Gabor filters, and Difference of Gaussian (DoG) (Klare et al. 2010; Goswami et al. 2011; Klare and Jain 2012; Zhu et al. 2014). Zhang et al. (2011) introduce a novel feature information based on mutual information of heterogeneous images. Huang et al. (2012b) use a pair of heterogeneous face databases as generic training databases, and find the common local geometrical information of heterogeneous face samples and generic training samples. Gong et al. (2017) transform the facial pixels into an encoded face space, where the features could be matched directly.

In addition, deep learning based schemes have been recently developed. Liu et al. (2016) utilize the ordinal activation function to select discriminative features and employ two types of triple loss in the process of transform. Saxena and Verbeek (2016) discuss various metric learning strategies to increase the HFR performance based on the pre-trained VIS CNN. The work of Reale et al. (2016) suggests using deeper networks to learn the features of cross-modal face images and proposes two novel network structures with small convolutional filters. He et al. (2017) employ a two-level CNN to learn domain-invariant identity representation and modality-related spectrum representation. Sarfraz and Stiefelhagen (2017) exploit a deep neural network to learn a non-linear mapping of images in different domains and preserve the identity information meanwhile. To learn information from raw facial patches directly, Peng et al. (2019) propose a deep local descriptor learning framework with a novel cross-modality enumeration loss.

### 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 Sensors

We use Canon-7D and AuthenMetric-CE31SE to acquire VIS and NIR images respectively. In this subsection, we introduce the near-infrared image collector in detail. The dual-camera module CE31SE from AuthenMetric, which contains both VIS and NIR lens, is specially designed for face recognition applications. We merely use the NIR shot which supports pixel array dimensions of 960 × 720 in the process of collecting. A 850 nm narrow band filter is equipped into the NIR shot. The collection can be controlled by the software provided by the company, and all photos captured are automatically saved on the computer used for operation.

#### 3.2 Data Collection and Cleaning

The face images were collected in the winter of 2019 in Henan Province and Tianjin, China. The participants are made up of local residents and international college students, and all the subjects agree to make their photos public. 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 backlight. The outdoor scenes will be reduced to one type when raining. Figure 2 shows the indoor and outdoor collection setup in detail. 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. These six variations are defined as attributes, and every attribute in the LAMP-HQ contains 10,339 NIR and 2817 VIS images respectively. The histograms of age, and Illumination variances are presented in Figs. 5 and 6. Some sample images are showed in Fig. 1. In addition, to advance NIR-VIS face recognition in the wild, we design a device containing 4 NIR shots to capture more side- and bottom-viewed NIR data, which is closer to real-world application. The relative positions of the 4 cameras are shown in the Fig. 3, and all the shots capture photos simultaneously. Figure 4 shows some examples of these views. Diverse image samples are provided at the end of this manuscript. The photographic distance and height are not strictly regulated to increase the complexity of the database. Finally, we capture $573 \times 6 \times 5 = 17,190$ ($p \times a \times i$) VIS and $573 \times 6 \times 5 \times 4 = 68,760$ ($p \times a \times i \times l$) NIR face images, where $p, a, i$ and $l$ denote the number of participants, attributes, illuminations and lens. 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. The procedure of pre-processing is as follows: (1) We first detect the location of faces and landmarks using the method in (Bulat and Tzimiropoulos 2017). (2) All the face images are rotated relying on the facial landmarks to make sure that the eye centers are on a horizontal line. (3) A box which contains the entire head of the subject is determined based on the distance from the center of the eyes to the month, and the images are cropped and aligned according to the size of the box. (4) All the aligned images are resized to the pixel array dimensions of $256 \times 256$.

### 3.3 Protocol

To construct the uniform database partition, we provide a fair evaluation protocol for our LAMP-HQ. We randomly divide the database into a training set and a testing set, and ensure that the number of subjects in each set accounts for about 50% of the total number of subjects. Note that there is no overlap between the training set and the test set. This random operation is repeated ten times, resulting in a total of ten-fold experiments in the protocol. In the testing set, VIS images are used as gallery (only one frontal neutral image of each subject) and NIR images are collected into the probe set to simulate the situation in real applications. The parameters are merely allowed to be tuned in the first fold experiment, and are fixed in the remaining nine fold. The Rank-1 recognition rate along with the mean accuracy and standard deviation of 10-fold experiments is used to evaluate the performance. Besides, the ROC curve generated by the first fold experiment result is necessary. Every VIS image is matched to all the NIR images. The images from the same identity are defined as positive samples, and the images from different identities are
negative samples. For example, the first fold of the database contains 27,263 positive and 7,415,536 negative samples.

4 Method

In this section, we propose a variational spectral attention network (VSANet) for NIR-VIS image translation. As illustrated in Fig. 7, 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 (Zhu et al. 2017). 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) (Kingma and Welling 2014) 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. 7, SVAE consists of two subnetworks: an inference network \(E\) that maps VIS data \(x_{vis}\) to the latent \(z_{vis}\), which approximates a prior \(p_{vis}(z_{vis})\), and a generative network \(G\) that samples VIS data \(\hat{x}_{vis}\) 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 \(\hat{x}_{vis}\), and the second regularizes the posterior \(q_{\phi}(z_{vis}|x_{vis})\) to match the prior \(p(z_{vis})\). Optimizing such ELBO, the spectral representation \(z_{vis}\) can be sampled from either the posterior \(q_{\phi}(z_{vis}|x_{vis})\) or the prior \(p(z_{vis})\), which leads to the capability of the proposed method to translate NIR data to the VIS domain with or without VIS references.

Tables 3 and 4 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 \(q_{\phi}(z_{vis}|x_{vis}) = \mathcal{N}(z; \mu, \sigma^2)\). 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 4, to guide the NIR-VIS translation.

4.2 Spectral Conditional Attention

As illustrated in Fig. 7, 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 \(y_{vis}\) both globally and locally. One is the spectral representation \(z_{vis} \in \mathbb{R}^C_z\) sampled from \(q_{\phi}(z_{vis}|x_{vis})\) or \(p(z_{vis})\), and the other is the feature \(F_{vis}^{(i)} \in \mathbb{R}^{C_j \times H_j \times W_j}\) from the \(i\)th layer of \(G\) in SVAE, i.e.,
Fig. 7 Overview of the proposed network architecture. It contains three parts: SVAE to extract the spectral representation $z_{vis}$, CSU to translate NIR data to the VIS domain, and SCA to transfer the VIS spectral style to CSU. Note that we only demonstrate the connections of one SCA block and neglect the others for a clear illustration.

Table 3 Structure of the inference network $E$ of SVAE

| Input  | Layer | Norm | Act       | Output | Output size  |
|--------|-------|------|-----------|--------|--------------|
| $x_{vis}$ | Conv3 | IN   | LeakyReLU | X0     | $32 \times 256 \times 256$ |
| $X_0$  | MaxPool | –    | –         | X0     | $32 \times 128 \times 128$ |
| $X_0$  | Conv3 | IN   | LeakyReLU | X0     | $64 \times 128 \times 128$ |
| $X_0$  | MaxPool | –    | –         | X0     | $64 \times 64 \times 64$   |
| $X_0$  | Conv3 | IN   | LeakyReLU | X0     | $128 \times 64 \times 64$  |
| $X_0$  | MaxPool | –    | –         | X0     | $128 \times 32 \times 32$  |
| $X_0$  | Conv3 | IN   | LeakyReLU | X0     | $256 \times 32 \times 32$  |
| $X_0$  | MaxPool | –    | –         | X0     | $256 \times 16 \times 16$  |
| $X_0$  | Conv3 | IN   | LeakyReLU | X0     | $512 \times 16 \times 16$  |
| $X_0$  | MaxPool | –    | –         | X0     | $512 \times 8 \times 8$    |
| $X_0$  | Conv3 | IN   | LeakyReLU | X0     | $512 \times 8 \times 8$    |
| $X_0$  | FC    | –    | –         | $\mu, \delta$ | $512, 512$ |

For $F_{vis}^{(j)} = F_{vis}^{(i)}(z_{vis})$, where $j = 1, \ldots, M$ denotes the index of the SCA block, $M$ is the number of SCA blocks. Given the content feature $F_{NIR}^{(j)} \in \mathbb{R}^{C_j \times H_j \times W_j}$ from the $k$-th layer of CSU, the fused feature $F_{fuse}^{(j)}$ can be obtained by the $j$-th SCA block $B_{sca}^{(j)}$,

$$F_{fuse}^{(j)} = B_{sca}^{(j)}(F_{NIR}^{(j)}, F_{vis}^{(j)}, z_{vis}).$$

(2)

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

As shown in Fig. 8, the SCA block can be divided into two parts: adaptive instance normalization (AdaIN) (Huang and Belongie 2017) operation and conditional attention layer. $F_{NIR}^{(j)}$ and $F_{vis}^{(j)}$ are firstly re-normalized with the spectral $z_{vis}$ in AdaIN and then the attention map is computed to model the spatial similarities between the processed spectral feature $F_{vis}^{(j)}$ and content feature $F_{NIR}^{(j)}$ in the conditional attention layer. The AdaIN operation is defined as

$$F_{t}^{(j)} = m^{(a)}(z_{vis}) \frac{F_{t}^{(j)} - \mu(F_{t}^{(j)})}{\sigma(F_{t}^{(j)})} + m^{(b)}(z_{vis}),$$

(3)

where $t \in \{vis, NIR\}$ and $m$ are an 8-layer MLP with two output vectors $m^{(i)}$ and $m^{(0)}$.

Inspired by the self-attention mechanism (Zhang et al. 2019a), we design a conditional attention layer to capture the spatial local relationships between the spectral and content features, i.e., $F_{vis}^{(j)}$ and $F_{NIR}^{(j)}$. The attention map $Att^{(j)} \in \mathbb{R}^{(H_j W_j) \times (H_j W_j)}$ is obtained by

$$Att^{(j)} = softmax(f^T(F_{vis}^{(j)}) g(F_{NIR}^{(j)})),$$

(4)
where $f$ and $g$ are $1 \times 1$ convolutions. The SCA block’s output $F_{fuse}^{(j)}$ is computed by

$$F_{fuse}^{(j)} = F_{NIR}^{(j)} + \gamma h(F_{NIR}^{(j)}) \text{Att}^{(j)},$$

(5)

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 $z_{vis}$ and $G$. Table 5 reports the network architecture of the cross-spectral UNet (CSU) injected with the spectral condition attention (SCA) blocks. CSU produces VIS facial image $y_{NIR}$ from the input NIR image $x_{NIR}$ together with the spectral guidance $z_{vis}$, $F_{vis}^{(3)}$ and $F_{vis}^{(4)}$. As shown in Table 5, 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. The bold emphasizes the connections between CSU and SCAs.

Note that the computation of the attention map can be conducted for any two images without alignment. For example, the similarity score may be high for the mouth regions across the VIS reference and the NIR source even when they are in different locations. Therefore, the attention map can guide the transfer of the spatial style information without alignment. In conclusion, both the AdaIN operation and the conditional attention layer are robust towards the misalignment between the VIS reference and NIR source. There is no need for the identity to be the same and the SCA block can guide the spectral transfer even when existing a huge spatial geometrical difference. The visual results in Fig. 9 verifies the effectiveness of our method in transferring the spectral information across unaligned faces.

### 4.3 Loss Functions

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

$$L_{svae} = \frac{1}{2} \|x_{vis} - \hat{x}_{vis}\|^2_{F} + \frac{1}{2} \sum_{i=1}^{C} ((\mu_{i}^{l})^2 + (\sigma_{i}^{l})^2 - \log((\sigma_{i}^{l})^2) - 1).$$

(6)

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

Similar to (Johnson et al. 2016), we employ the VGG-16 network (Simonyan and Zisserman 2014) 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:

$$L_{content} = \frac{1}{C_j H_j W_j} \|U_j(y_{vis}) - U_j(x_{NIR})\|_2^2$$

(7)
where $U_j(y_{\text{vis}})$ and $U_j(x_{\text{NIR}})$ are the output features of the
$j$th layer of VGG-16, respectively. We use the relu$_3$ layer to extract the features for the content loss.

In order to preserve the spectral information learned from $x_{\text{vis}}$ in the process of generating the output $y_{\text{vis}}$, we use the style loss proposed in (Gatys et al. 2016). It is obtained by computing the distance between the Gram matrices of $x_{\text{vis}}$ and $y_{\text{vis}}$, which is defined as:

$$L_{\text{style}} = \sum_j \| G_j^U(x_{\text{vis}}) - G_j^U(y_{\text{NIR}}) \|^2_F,$$

(8)

where

$$G_j^U(x)_{c,c'} = \frac{1}{C_j H_j W_j} \sum_{h=1}^{H_j} \sum_{w=1}^{W_j} U_j(x)_{h,w,c} U_j(x)_{h,w,c'},$$

(9)

$U_j(x)$ is the output feature of the $j$th layer of VGG-16. We use the relu$_{1,2}$, relu$_2$, relu$_3$, and relu$_4$ layers to extract the features for the style loss.

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

$$L_{\text{id}} = \| F(x_{\text{vis}}^{(\text{match})}) - F(y_{\text{vis}}) \|_1,$$

(10)

where $F(x_{\text{vis}}^{(\text{match})})$ and $F(y_{\text{vis}})$ are the extracted features respectively for $x_{\text{vis}}^{(\text{match})}$ and $y_{\text{vis}}$ 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.

Generative Adversarial Network (GAN) (Goodfellow et al. 2014) 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 $y_{\text{vis}}$, where the proposed CSU serves as the generator $G$ of GAN. The two-player minimax game is,

$$\min_{G} \max_{D} V(D, G) = \mathbb{E}_{x \sim p_{\text{vis}}(x)} [\log D(x)] + \mathbb{E}_{y_{\text{vis}} \sim p_G(y_{\text{vis}})} [\log(1 - D(y_{\text{vis}}))].$$

(11)

Table 5 Structure of CSU injected with SCA blocks. The CB block stands for a block of two convolutions, followed by Instance Normalization and leakyReLU for each one.

| Input      | Layer | Output | Output size |
|------------|-------|--------|-------------|
| $X_{\text{NIR}}$ | C Block | $F_{\text{NIR}}^{(1)}$ | $64 \times 256 \times 256$ |
| $F_{\text{NIR}}^{(1)}$ | MaxPool | $F_{\text{NIR}}^{(2)}$ | $64 \times 128 \times 128$ |
| $F_{\text{NIR}}^{(2)}$ | C Block | $F_{\text{NIR}}^{(3)}$ | $128 \times 128 \times 128$ |
| $F_{\text{NIR}}^{(3)}$ | MaxPool | $F_{\text{NIR}}^{(4)}$ | $128 \times 64 \times 64$ |
| $F_{\text{NIR}}^{(4)}$ | C Block | $F_{\text{NIR}}^{(5)}$ | $256 \times 64 \times 64$ |
| $F_{\text{NIR}}^{(5)}$ | MaxPool | $F_{\text{NIR}}^{(6)}$ | $256 \times 32 \times 32$ |
| $F_{\text{NIR}}^{(6)}$ | C Block | $F_{\text{NIR}}^{(7)}$ | $512 \times 32 \times 32$ |
| $F_{\text{NIR}}^{(7)}$ | UpSample | $F_{\text{NIR}}^{(8)}$ | $1024 \times 32 \times 32$ |
| $F_{\text{NIR}}^{(8)}$ | C Block | $F_{\text{NIR}}^{(9)}$ | $256 \times 32 \times 32$ |
| $F_{\text{NIR}}^{(9)}$ | Concat | $F_{\text{NIR}}^{(10)}$ | $512 \times 64 \times 64$ |
| $F_{\text{NIR}}^{(10)}$ | C Block | $F_{\text{NIR}}^{(11)}$ | $128 \times 64 \times 64$ |
| $F_{\text{NIR}}^{(11)}$ | UpSample | $F_{\text{NIR}}^{(12)}$ | $256 \times 128 \times 128$ |
| $F_{\text{NIR}}^{(12)}$ | C Block | $F_{\text{NIR}}^{(13)}$ | $64 \times 128 \times 128$ |
| $F_{\text{NIR}}^{(13)}$ | UpSample | $F_{\text{NIR}}^{(14)}$ | $128 \times 256 \times 256$ |
| $F_{\text{NIR}}^{(14)}$ | C Block | $F_{\text{NIR}}^{(15)}$ | $64 \times 256 \times 256$ |
| $F_{\text{NIR}}^{(15)}$ | Conv3 | $F_{\text{NIR}}^{(16)}$ | $3 \times 256 \times 256$ |
| $F_{\text{NIR}}^{(16)}$ | Sigmoid | $Y_{\text{NIR}}$ | $3 \times 256 \times 256$ |
Table 6  Structure of the discriminator

| 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 | $1 \times 16 \times 16$ |

Fig. 9  The qualitative results of exemplar-based NIR-VIS translation. Three spectral styles of VIS data in the top row are transferred to three NIR images in the left column.

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,

$$L_{adv} = -\log(D(y_{vis}))$$  \hspace{1cm} (12)

The total loss is defined as

$$L_{total} = L_{content} + \lambda_1 L_{style} + \lambda_2 L_{id} + \lambda_3 L_{adv},$$  \hspace{1cm} (13)

where $\lambda_1$, $\lambda_2$ and $\lambda_3$ are trade-off parameters.

4.4 Network Architecture

Table 6 specifically reports the network architecture of the discriminator $D$ used in the computation of the adversarial loss. The discriminator $D$ and the CSU network along with the SCA module are optimized iteratively following the original GAN (Goodfellow et al. 2014).

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 five HFR benchmarks on LAMP-HQ, including Pixel2Pixel (Isola et al. 2017), CycleGAN (Zhu et al. 2017), ADFL (Song et al. 2018), PCFH (Yu et al. 2019), and PACH (Duan et al. 2020). Both LightCNN-9 and LightCNN-29 (Wu et al. 2018) 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 (Li et al. 2013), BUAA-VisNIR face database (Huang et al. 2012a) and Oulu-CASIA NIR-VIS database (Chen et al. 2009), which are widely used in the HFR field.

The model is trained end-to-end and the trade-off parameters $\lambda_1$, $\lambda_2$, and $\lambda_3$ in Eq. (13) are set empirically to be 5, 0.1 and 1, respectively. Adam optimizer is employed with a learning rate of $2 \times 10^{-4}$, 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.
Fig. 11 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 (Isola et al. 2017), CycleGAN (Zhu et al. 2017), ADFL (Song et al. 2018), our method without SCA, and our method, respectively.

Table 7 NIR-VIS face recognition on LAMP-HQ (the first fold)

| Method          | LightCNN-9 Rank-1 | FAR=1% | FAR=0.1% | LightCNN-29 Rank-1 | FAR=1% | FAR=0.1% |
|-----------------|-------------------|--------|----------|-------------------|--------|----------|
| Original        | 89.77             | 88.46  | 71.61    | 94.94             | 93.42  | 78.65    |
| Pixel2Pixel     | 17.04             | 24.45  | 7.95     | 21.47             | 27.35  | 8.48     |
| cycleGAN        | 80.17             | 78.13  | 32.29    | 87.29             | 84.85  | 59.92    |
| ADFL            | 88.09             | 87.67  | 68.06    | 95.77             | 91.49  | 71.00    |
| PCFH            | /                 | /      | /        | 96.38             | 93.39  | 76.81    |
| PACH            | /                 | /      | /        | 96.85             | 93.86  | 78.65    |
| Ours w/o $L_{content}$ | 54.21             | 53.53  | 29.05    | 63.49             | 60.62  | 37.55    |
| Ours w/o $L_{style}$ | 87.98             | 86.38  | 65.42    | 94.35             | 92.57  | 74.03    |
| Ours w/o $L_{id}$ | 86.84             | 86.97  | 64.91    | 94.12             | 92.75  | 74.03    |
| Ours w/o $L_{adv}$ | 88.53             | 88.07  | 68.18    | 95.52             | 94.40  | 79.52    |
| Ours w/o AdaN | 93.67             | 91.22  | 72.89    | 96.70             | 95.05  | 80.78    |
| Ours w/o SCA   | 92.74             | 90.41  | 72.50    | 95.72             | 94.25  | 77.26    |
| Ours w/o SVAE&SCA | /                 | /      | /        | 95.65             | 94.02  | 82.94    |
| Ours           | 94.09             | 91.81  | 74.77    | 97.84             | 95.98  | 84.36    |
| Fine-tune 1    | /                 | /      | /        | 94.23             | 93.99  | 80.74    |
| Fine-tune 2    | /                 | /      | /        | 95.81             | 92.35  | 81.46    |

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 $z_{vis}$ is sampled from the posterior $q_{\phi}(z_{vis}|x_{vis})$. It can be observed in Fig. 9 that given different VIS exemplars, the outputs of the same NIR input image have different spectral styles. This phenomenon verifies that the proposed method can extract spectral information from VIS data and transfers it to NIR data to produce photo-realistic VIS images.

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. 11, 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 Sect. 3, we conduct quantitative comparison experiments on LAMP-HQ with several deep learning methods, including Pixel2Pixel, CycleGAN, ADFL, PCFH, and PACH. The pre-trained models, i.e., 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 Tables 7 and 8. As shown in Table 7, 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 97.91% by LightCNN-29. This verifies that the proposed method is effective in improving the performance of NIR-VIS face recognition. Besides, the recognition performance also improves in the 10-fold experiments. Figure 10 presents the ROC curves of different NIR-VIS face recognition methods. It could be observed that when FPR is smaller than 0.001, the proposed method significantly outperforms its competitors.

Besides, we fine-tune the LightCNN-29 on the proposed dataset and the results are shown in the bottom two rows of Table 7. The fine-tune 1 and fine-tune 2 stand for the fine-tuning on the training NIR-VIS images and the synthesized VIS images, respectively. The data show that the results decline a little than before. The reason may be that the NIR-VIS database is small-scale compared to the training dataset employed by LightCNN, which has millions of images of more than one hundred thousand people (Wu et al. 2018). Therefore the priori knowledge learned before is easily destroyed even after fine-tuning, then influences the recognition results.

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) (Shao and Fu 2016), TRIVET (Liu et al. 2016), IDR (He et al. 2017), Pixel2Pixel (Isola et al. 2017), CycleGAN (Zhu et al. 2017), and ADFL (Song et al. 2018). The quantitative results of face recognition computed by LightCNN-29 are reported in Tables 9 and 10. We also design cross experiments of the proposed method on the 4 databases above, the evaluation results are showed in Table 11. 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.

The 10-fold results in Tables 8 and 10 show that the performance of NIR-VIS face recognition on the proposed LAMP-HQ database is much lower than that on the CASIA 2.0 database. This indicates that LAMP-HQ is a more challenging NIR-VIS database for face recognition. The first fold results shown in Tables 7 and 9 also prove this point.

5.4 Ablation Study

To indicate the effects of each component of our method on the performance of HFR, we show in Table 7 the results of the ablation study. We can observe that the content loss in Eq. (13) is essential to preserve the input’s content, including identity information. The style loss, identity-preserving loss and adversarial loss seem to play a similar role in boosting the recognition performance because all of them regularize

---

**Table 8** NIR-VIS face recognition on LAMP-HQ (10-fold). The LightCNN-29 is employed as the face classifier

| Method   | Rank-1 | FAR = 1% | FAR = 0.1% |
|----------|--------|----------|------------|
| Original | 94.6 ± 0.28 | 92.5 ± 0.62 | 75.6 ± 1.92 |
| ADFL     | 95.1 ± 0.46 | 92.1 ± 0.93 | 73.3 ± 2.17 |
| PCFH     | 95.3 ± 0.48 | 92.9 ± 0.57 | 75.1 ± 1.84 |
| PACH     | 95.4 ± 0.53 | 93.1 ± 0.36 | 75.3 ± 1.67 |
| Ours     | 97.3 ± 0.16 | 94.8 ± 0.65 | 78.2 ± 2.97 |

**Table 9** NIR-VIS face recognition on CASIA 2.0 (the first fold)

| Method   | Rank-1 | FAR = 1% | FAR = 0.1% |
|----------|--------|----------|------------|
| Original | 96.8   | 99.1     | 94.7       |
| Pixel2Pixel | 22.13 | 39.22    | 14.45      |
| cycleGAN | 87.23  | 93.92    | 79.41      |
| ADFL     | 98.2   | 97.2     | /          |
| PCFH     | 98.5   | 99.6     | 97.3       |
| PACH     | 99.0   | 99.6     | 98.5       |
| Ours     | 99.1   | 99.7     | 98.1       |
Table 10  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 (10-fold) |  | BUAA |  | Oulu |  |
|--------|---------------------|---|-------|---|------|---|
|        | Rank-1  | FAR = 1% | Rank-1  | FAR = 1% | Rank-1  | FAR = 1% | FAR = 0.1% | 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 |
| 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 |
| PCFH | 98.8 ± 0.26 | 99.6 ± 0.08 | 97.7 ± 0.26 | 98.4 | 97.9 | 92.4 | 100.0 | 97.7 | 86.6 |
| PACH | 98.9 ± 0.19 | 99.6 ± 0.10 | 98.3 ± 0.21 | 98.6 | 98.0 | 93.5 | 100.0 | 97.9 | 88.2 |
| Ours | 99.2 ± 0.04 | 99.7 ± 0.01 | 98.2 ± 0.19 | 98.8 | 98.3 | 93.4 | 100.0 | 97.7 | 89.0 |

Fig. 12  The visual results of the ablation study. From left to right, the columns are the input NIR data, the corresponding VIS data, the results of several version of the proposed method, i.e., the ablation versions without $L_{content}$, $L_{id}$, $L_{style}$, $L_{adv}$, the SCA module and the AdaIN operation, respectively. The last column are the results of the proposed method.

Table 11  Cross experiment results on Oulu, BUAA, CASIA2.0, and LAMP

| Test | Oulu | BUAA | CASIA 2.0 | LAMP | Metric |
|------|------|------|----------|------|--------|
| Train |  |  |  |  |  |
| Oulu | 86.3 | 76.9 | 67.8 | 62.3 | Rank-1 |
|     | 71.5 | 72.1 | 83.1 | 79.3 | FAR = 1% |
|     | 35.9 | 48.2 | 52.8 | 47.4 | FAR = 0.1% |
| BUAA | 97.1 | 97.8 | 92.5 | 90.9 | Rank-1 |
|     | 80.6 | 97.2 | 97.9 | 81.4 | FAR = 1% |
|     | 50.6 | 92.7 | 90.6 | 76.2 | FAR = 0.1% |
| CASIA 2.0 | 100.0 | 98.8 | 99.1 | 92.4 | Rank-1 |
|     | 97.7 | 98.3 | 99.7 | 88.7 | FAR = 1% |
|     | 89.0 | 93.4 | 98.1 | 77.7 | FAR = 0.1% |
| LAMP | 100.0 | 99.5 | 99.0 | 97.8 | Rank-1 |
|     | 97.3 | 98.8 | 99.8 | 96.0 | FAR = 1% |
|     | 86.4 | 96.6 | 98.5 | 84.4 | FAR = 0.1% |
the generated data to preserve certain VIS characteristics. When the AdaIN operation or the SCA module is absent, the recognition performance drops marginally. Figure 13 shows the detailed visualization results.

6 Conclusion

This paper proposes a 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.

Acknowledgements

This work is partially funded by National Key Research and Development Program of China (Grant No. 2020AAA0140001), and National Natural Science Foundation of China (Grant Nos. 61721004, U20A20223, 62006228).
References

Bernhard J, Barr J, Bowyer KW, Flynn P (2015) Near-infrared to visible light face matching: Effectiveness of pre-processing options for commercial matchers. In: IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS), pp 1–8

Biswas, S., Bowyer, K. W., & Flynn, P. J. (2011). Multidimensional scaling for matching low-resolution face images. *IEEE transactions on pattern analysis and machine intelligence*, 34(10), 2019–2030.

Bulat A, Tzimiropoulos G (2017) How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In: International Conference on Computer Vision

Chen J, Yi D, Yang J, Zhao G, Li SZ, Pietikainen M (2009) Learning mappings for face synthesis from near infrared to visual light images. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 156–163

Di X, Zhang H, Patel VM (2018) Polarimetric thermal to visible face verification via attribute preserved synthesis. In: IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS), pp 1–10

Di X, Riggan BS, Hu S, Short NJ, Patel VM (2019) Polarimetric thermal to visible face verification via self-attention guided synthesis. In: 2019 International Conference on Biometrics (ICB), pp 1–8

Di X, Riggan BS, Hu S, Short NJ, Patel VM (2020) Multi-scale thermal to visible face verification via attribute guided synthesis. arXiv preprint arXiv:2004.09502

Duan B, Fu C, Li Y, Song X, He R (2020) Cross-spectral face hallucination via disentangling independent factors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7930–7938

Gatys LA, Ecker AS, Bethge M (2016) Image style transfer using convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 2414–2423

Gong, D., Li, Z., Huang, W., Li, X., & Tao, D. (2017). Heterogeneous face recognition: A common encoding feature discriminant approach. *IEEE Transactions on Image Processing*, 26(5), 2079–2089.

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems, pp 2672–2680

Goswami D, Chan CH, Windridge D, Kittler J (2011) Evaluation of face recognition system in heterogeneous environments (visible vs nir). In: IEEE International Conference on Computer Vision Workshops, pp 2160–2167

Gurton, K. P., Yuffa, A. J., & Videen, G. W. (2014). Enhanced facial recognition for thermal imagery using polarimetric imaging. *Optics letters*, 39(13), 3857–3859.

He R, Wu X, Sun Z, Tan T (2017) Learning invariant deep representation for nir-vis face recognition. In: AAAI Conference on Artificial Intelligence

He, R., Wu, X., Sun, Z., & Tan, T. (2018). Wasserstein cnp: Learning invariant features for nir-vis face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 41(7), 1761–1773.

Hu, S., Short NJ, Riggan BS, Gordon C, Gurton KP, Thielke M, Gurram P, Chan AL (2016) A polarimetric thermal database for face recognition research. In: IEEE conference on computer vision and pattern recognition workshops, pp 119–126

Huang D, Sun J, Wang Y (2012a) The buaa-visnir face database instructions. School Comput Sci Eng, Beihang Univ, Beijing, China, Tech Rep IRIP-TR-12-FR-001

Huang DA, Frank Wang YC (2013) Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition. In: IEEE International Conference on Computer Vision, pp 2496–2503

Huang L, Lu J, Tan YP (2012b) Learning modality-invariant features for heterogeneous face recognition. In: IEEE International Conference on Pattern Recognition, pp 1683–1686

Huang R, Zhang S, Li T, He R (2017) Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In: IEEE International Conference on Computer Vision, pp 2439–2448

Huang S, Belongie S (2017) Arbitrary style transfer in real-time with adaptive instance normalization. In: IEEE International Conference on Computer Vision, pp 1501–1510

Huang, X., Lei, Z., Fan, M., Wang, X., & Li, S. Z. (2012c). Regularized discriminative spectral regression method for heterogeneous face matching. *IEEE Transactions on Image Processing*, 22(1), 353–362.

Isola P, Zhu JY, Zhou T. Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 1125–1134

Jin, Y., Li, J., Lang, C., & Ruan, Q. (2017). Multi-task clustering elm for vis-nir cross-modal feature learning. *Multidimensional Systems and Signal Processing*, 28(3), 905–920.

Johnson J, Alahi A, Fei-Fei L (2016) Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision, pp 694–711

Juelfe-Xu F, Pal DK, Savvides M (2015) Nir-vis heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition workshops, pp 141–150

Kan, M., Shan, S., Zhang, H., Lao, S., & Chen, X. (2015). Multi-view discriminant analysis. *IEEE transactions on pattern analysis and machine intelligence*, 38(1), 188–194.

Kingma DP, Welling M (2014) Auto-encoding variational bayses. In: International Conference on Learning Representations

Klare, B., Li, Z., & Jain, A. K. (2010). Matching forensic sketches to mug shot photos. *IEEE transactions on pattern analysis and machine intelligence*, 33(3), 639–646.

Klare, B. F., & Jain, A. K. (2012). Heterogeneous face recognition using kernel prototype similarities. *IEEE transactions on pattern analysis and machine intelligence*, 35(6), 1410–1422.

Lei Z, Bai Q, He R, Li SZ (2008) Face shape recovery from a single image using cca mapping between tensor spaces. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 1–7

Lei, Z., Liao, S., Jain, A. K., & Li, S. Z. (2012). Coupled discriminant analysis for heterogeneous face recognition. *IEEE Transactions on Information Forensics and Security*, 7(6), 1707–1716.

Lezama J, Qiu Q, Sapiro G (2017) Not afraid of the dark: Nir-vis face recognition via cross-spectral joint dictionary learning and reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 141–150

Li, Z., Liao, S., Shan, S., Zhang, H., Lao, S., & Chen, X. (2015). Multi-view discriminant analysis. *IEEE transactions on pattern analysis and machine intelligence*, 38(1), 188–194.

Lin D, Tang X (2006) Inter-modality face recognition. In: European Conference on Computer Vision, pp 13–26
